Introductory Statistics

INTRODUCTORY STATISTICS

JARED EUSEA; PHYLLIS OKWAN; RACHID BELMASROUR; STEPHAN PATTERSON; AND STEPHEN ANDRUS

ASHLEY SEGALLA AND DEBBIE SHEPHERD

LOUIS: The Louisiana Library Network

Introductory Statistics Copyright © 2024 by LOUIS: The Louisiana Library Network is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License, except where otherwise noted.

This resource was adapted from Introduction to Statistics, by Barbara Illowsky and Susan Dean, OpenStax (Web), licensed under a CC BY 4.0 License. It used a version imported into Pressbooks by the University of Hawai'i.

Additional content was added from Significant Statistics: An Introduction to Statistics by John Morgan Russell, OpenStaxCollege, OpenIntro and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License, (where noted) and videos were added from Introduction to Statistics: An Excel-Based Approach by Valerie Watts and licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License(where noted).

CONTENTS

Introduction 1

Introduction 1

About This Book 2

About the Authors 4

Authors 4

Editors/Reviewers/Contributors 4

Chapter 1: Sampling and Data

Introduction to Chapter 1: Sampling and Data 9

Collaborative Exercises 11

1.1 Definitions of Statistics, Probability, and Key Terms 16

Probability 16

Key Terms 17

1.2 Data, Sampling, and Variation in Data and Sampling 23

Sampling 35

Variation in Data 45

Variation in Samples 45

Critical Evaluation 46

References 47

1.3 Frequency, Frequency Tables, and Levels of Measurement 48
Answers and Rounding Off 48
Levels of Measurement 48
Frequency 50
References 61
1.4 Experimental Design and Ethics 62
Ethics 65
References 68
Chapter 1 Review 70
Chapter Review from 1.1 70
Chapter Review from 1.2 70
Chapter Review from 1.3 70
Chapter 1 Practice 74
Chapter 1 Homework 86
HOMEWORK from 1.2 86

Chapter 2: Descriptive Statistics

Introduction to Chapter 2: Descriptive Statistics 119
Collaborative Exercises 121
2.1 Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs 123
References 136
2.2 Histograms, Frequency Polygons, and Time Series Graphs 137
Frequency Polygons 145
Constructing a Time Series Graph 148
References 151

2.3 Measures of the Location of the Data 153
A Formula for Finding the kth Percentile 160
A Formula for Finding the Percentile of a Value in a Data Set 162
Interpreting Percentiles, Quartiles, and Median 164
References 168
2.4 Box Plots 169
References 176
2.5 Measures of the Center of the Data 177
The Law of Large Numbers and the Mean 182
Sampling Distributions and Statistic of a Sampling Distribution 182
Calculating the Mean of Grouped Frequency Tables 183
References 185
2.6 Skewness and the Mean, Median, and Mode 186
2.7 Measures of the Spread of the Data 191
The Standard Deviation 191
Sampling Variability of a Statistic 194
Standard deviation of Grouped Frequency Tables 201
Comparing Values from Different Data Sets 203
References 205
Chapter 2 Review 207
Formula Review 219
Chapter Review from 2.5 225
Formula Review 226
Formula Review 235
238
Chapter 2 Practice 239
Bringing It Together from 2.4 239
Bringing It Together from 2.5 240

Chapter 2 Homework 252
Homework from 2.1 252
Homework from 2.2 255
Homework from 2.5 272
Homework from 2.7 274

Chapter 3: Probability Topics

Introduction to Chapter 3: Probability Topics 283
Collaborative Exercises 284
3.1 Terminology 286
Definition of Probability 287
Combining Events 289
3.2 Independent and Mutually Exclusive Events 296
Independent Events 296
Mutually Exclusive Events 299
References 308
Chapter Review 308
Formula Review 308
3.3 Two Basic Rules of Probability 309
The Multiplication Rule 310
3.4 Contingency Tables 317
Contingency Tables 317
3.5 Tree and Venn Diagrams 324
Tree Diagrams 324
Venn Diagrams 333
References 337

Chapter 3 Review 338
Section 3.1 Review 338
Section 3.2 Review 338
Section 3.3 Review 339
Section 3.4 Review 339
Section 3.5 Review 339
Chapter 3 Practice 341
Chapter 3 Homework 354
Section 3.1 354
Section 3.4 373

Chapter 4: Discrete Random Variables

Introduction to Chapter 4: Discrete Random Variables 393
Collaborative Exercises 394
4.1 Probability Distribution Function (PDF) for a Discrete Random Variable 395
Characteristics and Notation 396
4.2 Measures of General Discrete Random Variables 402
The Expected Value (Mean) of a Discrete Random Variable 402
The Variance and Standard Deviation of a Discrete Random Variable 405
4.3 Binomial Distribution 409
Binomial Experiments 409
Notation for the Binomial 410
Experiments That Are Not Binomial 413
Binomial Probability Function 414
Expected Value and Standard Deviation 417
421

4.4 Geometric Distribution 422
Expected Value and Standard Deviation 425
4.5 Hypergeometric Distribution 430
Notation and Mean for the Hypergeometric 432
4.6 Poisson Distribution 436
Notation and Mean 439
Chapter 4 Review 446
Section 4.1 Review 446
Section 4.2 Review 446
Section 4.3 Review 446
Section 4.4 Review 447
Section 4.5 Review 448
Section 4.6 Review 449
Chapter 4 Practice 450
Section 4.1 450
Section 4.4 462
Chapter 4 Homework 468
Section 4.2 Homework 470
494
Section 4.5 Homework 501

Chapter 5: Continuous Random Variables

Introduction to Chapter 5: Continuous Random Variables 515
Collaborative Exercises 516
5.1 Continuous Probability Functions 518
Properties of Continuous Probability Distributions 518
Continuous Probability Functions 520

5.2 The Uniform Distribution 525
5.3 The Exponential Distribution 540
Chapter 5 Review 558
5.1 Review 558
Formula Review 558
5.2 Review 559
Formula Review 559
5.3 Review 560
Formula Review 560
Chapter 5 Practice 562
5.1 Practice 562
5.2 Practice 565
References 565
5.2 Practice 570
References 570
Chapter 5 Homework 575
5.1 Homework 575
5.2 Homework 575
5.3 Homework 581

Chapter 6: The Normal Distribution and The Central Limit Theorem

Introduction to Chapter 6a: The Normal Distribution 593
Collaborative Exercises 594
6.1 The Standard Normal Distribution 595
Z-Scores 597
The Empirical Rule 603
References 607

6.2 Using the Normal Distribution 608
Finding Normal Probabilities 608
The Standardizing Process 610
The "Un-standardizing" Process 616
Using Technology 618
References 626
Introduction to Chapter 6b: The Central Limit Theorem 628
Collaborative Exercises 629
6.3 The Central Limit Theorem for Sample Means (Averages) 631
Using the CLT 634
References 647
6.4 The Central Limit Theorem for Sums 649
References 662
6.5 The Normal Approximation to the Binomial 663
Binomial Approximation Conditions 665
The Continuity Correction 667
References 671
Chapter 6 Review 672
Chapter Review 672
Formula Review 674
Chapter 6 Practice 677
Section 6.1 Practice 677
Section 6.2 Practice 689
Section 6.3 Practice 695
Section 6.4 Practice 698
Section 6.5 Practice 706

Chapter 6 Homework 712
Section 6.1 Homework 712
Section 6.2 Homework 717
Section 6.3 Homework 731
Section 6.4 Homework 738
Section 6.5 Homework 741

Chapter 7: Confidence Intervals

Introduction to Chapter 7: Confidence Intervals 759
Collaborative Exercises 761
7.1 A Single Population Mean Using the Normal Distribution 763
The Confidence Interval (Known Standard Deviation) 763
Changing the Confidence Level or Sample Size 774
Summary: Effect of Changing the Sample Size 777
Working Backwards to Find the Error Bound or Sample Mean 777
The Sample Size n 779
7.2 A Single Population Mean using the Student t Distribution 782
Student's t-probabilities. 782
The Confidence Interval 784
7.3 A Population Proportion 792
The error bound for a proportion. 792
"Plus Four" Confidence Interval for p 800
The Sample Size n 806
Chapter 7 Review 810
7.1 Chapter Review 810
7.2 Chapter Review 811
7.3 Chapter Review 813

Chapter 7 Practice 815
7.1 Practice 815
7.2 Practice 820
7.3 Practice 825
Chapter 7 Homework 830
7.1 Homework 830
7.2 Homework 836
7.3 Homework 844

Chapter 8: Hypothesis Testing with One Sample

Introduction to Chapter 8: Hypothesis Testing with One Sample 855
Collaborative Exercises 857
8.1 Null and Alternative Hypotheses 858
Hypothesis Testing 858
8.2 Outcomes and the Type I and Type II Errors 864
Type I and Type II Errors 864
8.3 Distribution Needed for Hypothesis Testing 870
Distributions Required to Conduct a Hypothesis Test 870
Assumptions 871
8.4 Rare Events, the Sample, Decision, and Conclusion 872
Rare Events 872
Using the Sample to Test the Null Hypothesis 873
Decision and Conclusion 875

8.5 Additional Information and Full Hypothesis Test Examples 878
Hypothesis Tests for a Single Population Mean, Population Standard Deviation Known 878
Historical Note ([link]) 886
Hypothesis Tests for a Single Population Mean, Population Standard Deviation Unknown 889
Hypothesis Tests for a Single Population Proportion 892
Chapter 8 Review 906
8.1 Chapter Review 906
8.2 Chapter Review 906
8.3 Chapter Review 907
8.4 Chapter Review 908
8.5 Chapter Review 908
Chapter 8 Practice 909
8.1 Practice 909
8.2 Practice 910
8.3 Practice 912
8.4 Practice 913
8.5 Practice 916
Chapter 8 Homework 919
8.1 Homework 919
8.2 Homework 920
8.3 Homework 923
8.4 Homework 924
8.5 Homework 924
Note 925
Note 929
946

Chapter 9: Hypothesis Testing with Two Samples

Introduction to Chapter 9: Hypothesis Testing with Two Samples 949

9.1 Two Population Means with Unknown Standard Deviations 952

Hypothesis Tests for Two Population Means, Population Standard Deviations Unknown 952

9.2 Two Population Means with Known Standard Deviations 966

Hypothesis Tests for Two Population Means, Population Standard Deviations Known 966

9.3 Comparing Two Independent Population Proportions 972

Hypothesis tests for two population proportions 972

9.4 Matched or Paired Samples 983

Hypothesis Tests for Matched or Paired Samples 983

Chapter 9 Review 992

9.1 Chapter Review 992

9.2 Chapter Review 993

9.3 Chapter Review 993

9.4 Chapter Review 994

Chapter 9 Practice 995

9.1 Practice 995

9.2 Practice 999

9.3 Practice 1002

9.4 Practice 1004

Chapter 9 Homework 1007

1007

9.1 Homework 1007

9.2 Homework 1013

9.3 Homework 1015

9.4 Homework 1023

Chapter 10: Linear Regression and Correlation

Introduction to Chapter 10: Linear Regression and Correlation 1035

Collaborative Exercises 1036

10.1 Linear Equations 1037

1040

Slope and Y-Intercept of a Linear Equation 1041

1042

References 1042

10.2 Scatter Plots 1043

10.3 The Regression Equation 1048

Least Squares Criteria for Best Fit 1052

Understanding Slope 1053

The Correlation Coefficient r 1055

The Coefficient of Determination 1057

10.4 Testing the Significance of the Correlation Coefficient 1058

Third-Exam vs. Final-Exam Example: p-value method 1061

1065

Third-Exam vs. Final-Exam Example: critical value method 1065

1066

10.5 Prediction 1069

1071

10.6 Outliers 1072
Identifying Outliers 1072
1075
Numerical Identification of Outliers 1075
How Does the Outlier Affect the Best Fit Line? 1076
Numerical Identification of Outliers: Calculating s and Finding Outliers Manually 1077
1083
95% Critical Values of the Sample Correlation Coefficient Table 1084
References 1086
Chapter 10 Review 1087
Chapter Review from 10.1 1087
Formula Review 1087
Chapter Review from 10.2 1092
Chapter Review from 10.3 1093
Chapter Review from 10.5 1095
Chapter Review from 10.6 1100
Chapter 10 Homework 1102
Homework from 10.1 1102
1103
Homework from 10.2 1103
Homework from 10.3 1104
Chapter Review from 10.4 1105
Formula Review 1105
Homework from 10.4 1106
Homework from 10.5 1106
Homework from 10.6 1112
Chapter 10 Practice 1118

Chapter 11: The Chi-Square Distribution

Introduction to Chapter 11: The Chi-Square Distribution 1129

Collaborative Exercises 1130

11.1 Facts About the Chi-Square Distribution 1132

References 1133

11.2 Goodness-of-Fit Test 1134

Note 1135

11.3 Test of Independence 1150

Note 1151

11.4 Test for Homogeneity 1160

References 1165

11.5 Comparison of the Chi-Square Tests 1166

11.6 Test of a Single Variance 1168

References 1173

Chapter 11 Review 1174

Formula Review 1174

11.2 Review 1175

Formula Review 1175

11.3 Review 1179

Formula Review 1179

11.4 Review 1184

Formula Review 1184

11.5 Review 1186

11.6 Review 1186

Formula Review 1187

Chapter 11 Practice 1189

Chapter 11 Homework 1190
11.1 Homework 1190
11.3 Homework 1197
11.4 Homework 1203
11.5 Homework 1206
Bringing It Together 1207
11.6 Homework 1207

Class Group Activities/Projects

Chapter 1 Project I: Data Collection Experiment 1215
Chapter 1 Project II: Sampling Experiment 1217
Chapter 2 Project: Descriptive Statistics 1221
Chapter 3 Project: Probability Topics 1223
Chapter 4 Project I: Discrete Distribution (Lucky Dice Experiment) 1226
Chapter 4 Project II: Discrete Distribution (Playing Card Experiment) 1230
Chapter 5 Project: Continuous Distribution 1234
Chapter 6 Project I: Normal Distribution (Lap Times) 1237
Chapter 6 Project II: Normal Distribution (Pinkie Length) 1240
Chapter 6 Project III: Central Limit Theorem (Pocket Change) 1243
Chapter 6 Project IV: Central Limit Theorem (Cookie Recipes) 1247
Chapter 7 Project I: Confidence Interval (Home Costs) 1251
Chapter 7 Project II: Confidence Interval (Place of Birth) 1254
Chapter 7 Project III: Confidence Interval (Women's Heights) 1256
Chapter 8 Project: Hypothesis Testing of a Single Mean and Single Proportion 1259
Chapter 9 Project: Hypothesis Testing for Two Means and Two Proportions 1262
Chapter 10 Project I: Regression (Distance from School) 1266
Chapter 10 Project II: Regression (Textbook Cost) 1269
Chapter 10 Project III: Regression (Fuel Efficiency) 1271

Chapter 11 Project I: Chi-Square Goodness-of-Fit 1274
Chapter 11 Project II: Chi-Square Test of Independence 1279
Group and Partner Projects 1281

Glossary 1291
Summary of Adaptations 1308
Full Resource Adaptations 1308
Chapter-Specific Changes 1309

INTRODUCTION

Introduction

"Introductory *Statistics* follows scope and sequence requirements of a one-semester introduction to statistics course and is geared toward students majoring in fields other than math or engineering. The text assumes some knowledge of intermediate algebra and focuses on statistics application over theory. *Introductory Statistics* includes innovative practical applications that make the text relevant and accessible, as well as collaborative exercises, technology integration problems, and statistics labs."

Pedagogical foundation and features

- **Examples** are placed strategically throughout the text to show students the step-by-step process of interpreting and solving statistical problems. To keep the text relevant for students, the examples are drawn from a broad spectrum of practical topics, including examples about college life and learning, health and medicine, retail and business, and sports and entertainment.
- **Try It** practice problems immediately follow many examples and give students the opportunity to practice as they read the text. **They are usually based on practical and familiar topics, like the Examples themselves**.
- **Collaborative Exercises** provide an in-class scenario for students to work together to explore presented concepts.
- **Using the TI-83, 83+, 84, 84+ Calculator** shows students step-by-step instructions to input problems into their calculator.
- **The Technology Icon** indicates where the use of a TI calculator or computer software is recommended.
- **Practice, Homework, and Bringing It Together** problems give the students problems at various degrees of difficulty while also including real-world scenarios to engage students.

Coverage and scope

- Chapter 1 Sampling and Data
- Chapter 2 Descriptive Statistics
- Chapter 3 Probability Topics
- Chapter 4 Discrete Random Variables

- Chapter 5 Continuous Random Variables
- Chapter 6 The Normal Distribution and The Central Limit Theorem
- Chapter 7 Confidence Intervals
- Chapter 8 Hypothesis Testing with One Sample
- Chapter 9 Hypothesis Testing with Two Samples
- Chapter 10 Linear Regression and Correlation
- Chapter 11 The Chi-Square Distribution

Alternate sequencing

Introductory Statistics was written to fit a particular topical sequence, but it can be used flexibly to accommodate other course structures. One such potential structure, which fits reasonably well with the textbook content, is provided below. Please consider, however, that the chapters were not written to be completely independent, and that the proposed alternate sequence should be carefully considered for student preparation and textual consistency.

- Chapter 1 Sampling and Data
- Chapter 2 Descriptive Statistics
- Chapter 10 Linear Regression and Correlation
- Chapter 3 Probability Topics
- Chapter 4 Discrete Random Variables
- Chapter 5 Continuous Random Variables
- Chapter 6 The Normal Distribution and The Central Limit Theorem
- Chapter 7 Confidence Intervals
- Chapter 8 Hypothesis Testing with One Sample
- Chapter 9 Hypothesis Testing with Two Samples
- Chapter 11 The Chi-Square Distribution

This Introduction section was adapted from: "Preface." *Introductory Statistics*, 2nd ed., OpenStax, https://openstax.org/books/introductory-statistics-2e/pages/preface. Accessed 16 Dec. 2024.

About This Book

This textbook was created through Connecting the Pipeline: Libraries, OER, and Dual Enrollment from Secondary to Postsecondary, a $1.3 million project funded by LOUIS: The Louisiana Library Network and the Institute of Library and Museum Services. This project supports the extension of access to high-quality

post-secondary opportunities to high school students across Louisiana and beyond by creating materials that can be adopted for dual enrollment environments. Dual enrollment is the opportunity for a student to be enrolled in high school and college at the same time.

The cohort-developed OER course materials are released under a license that permits their free use, reuse, modification and sharing with others. This includes a corresponding course available in Moodle and Canvas that can be imported to other platforms. For access/questions, contact Affordable Learning Louisiana.

If you are adopting this textbook, we would be glad to know of your use via this brief survey.

Cover Image

The cover image is "Color Coded" by Hopeless_Wonderer and licensed under a Creative Commons Attribution-ShareAlike 2.0 Generic (CC BY-SA 2.0 DEED).

ABOUT THE AUTHORS

Authors

Jared Eusea has been teaching at River Parishes Community College since January 2011, where he has been promoted from Instructor to Associate Professor and has been named the Department Chair of the Mathematics Department (starting in Fall 2021). He has taught lecture, online (asynchronous and synchronous), and hybrid (blended) courses for Developmental Mathematics, College Algebra, Introduction to Statistics, Trigonometry, and Calculus. All of his courses are 100% OER content! He has created lecture notes following the OpenStax textbooks for faculty to use and has shared these on the OpenStax OER Commons website. In addition to teaching a full-load each semester since he started at RPCC, Jared has led committees in redesigning the math curriculum and sequence and led discussion for establishing course commonality for all mathematics courses, including course and unit objectives, textbooks, assessment policies, and a standard homework system. Jared has been instrumental in moving forward improvements in online teaching and learning at RPCC by developing Master courses that could be used for all math online courses. He has also been the chair of the committee at RPCC that reviews those courses being offered online at RPCC to check that they are designed following certain criteria consistent to the QM standards and has been involved with faculty on an internal Instructional Design Team to develop a common course template for use towards our college's goal of fully online programs. Jared is a QM Peer Reviewer and has dealt with the three main LMS platforms: Blackboard, Moodle, and Canvas, so I have seen a lot of various course designs. Lastly, since 2017 I have been a faculty participant and Principal Investigator for two system-wide eLearning Innovation Fund Grants, one Affordable Learning Grant, and a Summer Faculty cohort, where funding was provided to faculty who developed fully online OER courses to share within our system. I also participated as a reviewer in the last course design project that was led by the LOUIS team.

Phyllis Okwan

Rachid Belmasrour

Stephan Patterson

Stephen Andrus

Editors/Reviewers/Contributors

Dr. Ashley Segalla is an Instructor of Mathematics at Southeastern Louisiana University in Hammond, Louisiana. Ashley holds a Bachelor of Science in Mathematics and a Master of Business Administration, both

from Southeastern Louisiana University. She recently completed her Ph.D. in Curriculum and Instruction: Mathematics Education from Louisiana State University. Ashley has taught 7 years at the high school and college levels.

Debbie Shepherd

CHAPTER 1: SAMPLING AND DATA

INTRODUCTION TO CHAPTER 1: SAMPLING AND DATA



Figure 1.1a. Smartphone Display of COVID-19 Statistics. Since the Coronavirus is novel, statisticians must collect and translate data into digestible information to give to the public.



Figure 1.1b. We encounter statistics in our daily lives more often than we probably realize and from many different sources, like the news. (credit: David Sim)

"Statistics' ultimate goal is translating data into knowledge" – Alan Agresti & Christine Franklin

You are probably asking yourself the question, "When and where will I use statistics?" If you read any newspaper, watch television, or use the Internet, you will see statistical information. There are statistics about crime, sports, education, politics, and real estate. Typically, when you read a newspaper article or watch a television news program, you are given sample information. With this information, you may make a decision about the correctness of a statement, claim, or "fact." Statistical methods can help you make the "best educated guess."

Since you will undoubtedly be given statistical information at some point in your life, you need to know some techniques for analyzing the information thoughtfully. Think about buying a house or managing a budget. Think about your chosen profession. The fields of economics, business, psychology, education, biology, law, computer science, police science, and early childhood development require at least one course in statistics.

Included in this chapter are the basic ideas and words of probability and statistics. You will soon understand that statistics and probability work together. You will also learn how data are gathered and what "good" data can be distinguished from "bad."

Collaborative Exercises

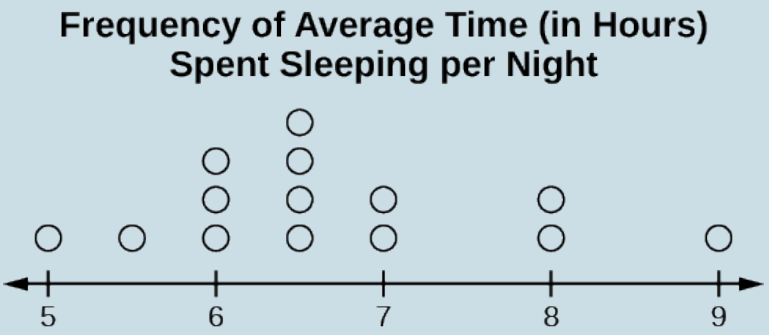
Section 1.1

Collaborative Exercise 1

In your classroom, try this exercise. Have class members write down the **average** time (in hours, to the nearest half-hour) they sleep per night. Your instructor will record the data. Then create a simple graph (called a **dot plot**) of the data. A dot plot consists of a number line and dots (or points) positioned above the number line. For example, consider the following data:

5, 5.5, 6, 6, 6, 6.5, 6.5, 6.5, 6.5, 7, 7, 8, 8, 9

The dot plot for this data would be as follows:



Does your dot plot look the same as or different from the example? Why? If you did the same example in an English class with the same number of students, do you think the results would be the same? Why or why not?

Where do your data appear to cluster? How might you interpret the clustering?

The questions above ask you to analyze and interpret your data. With this example, you have begun your study of statistics.

Collaborative Exercise 2

Do the following exercise collaboratively with up to four people per group.

Find a population, a sample, the parameter, the statistic, a variable, and data for the following study: You want to determine the average (mean) number of glasses of milk college students drink per day. Suppose yesterday, in your English class, you asked five students how many glasses of milk they drank the day before. The answers were 1, 0, 1, 3, and 4 glasses of milk.

Section 1.2

Collaborative Exercise 1

As a class, determine whether or not the following samples are representative. If they are not, discuss the reasons.

1. To find the average GPA of all students in a university, use all honor students at the university as the sample.
2. To find out the most popular cereal among young people under the age of ten, stand outside a large supermarket for three hours and speak to every twentieth child under age ten who enters the supermarket.
3. To find the average annual income of all adults in the United States, sample U.S. congressmen. Create a cluster sample by considering each state as a stratum (group). By using simple random sampling, select states to be part of the cluster. Then survey every U.S. congressman in the cluster.

4. To determine the proportion of people taking public transportation to work, survey 20 people in New York City. Conduct the survey by sitting in Central Park on a bench and interviewing every person who sits next to you.
5. To determine the average cost of a two-day stay in a hospital in Massachusetts, survey 100 hospitals across the state using simple random sampling.

Collaborative Exercise 2

Divide into groups of two, three, or four. Your instructor will give each group one six-sided die. Try this experiment twice. Roll one fair die (six-sided) 20 times. Record the number of ones, twos, threes, fours, fives, and sixes you get in Table 1 and Table 2 ("frequency" is the number of times a particular face of the die occurs):

First Experiment (20 rolls)	
Face on Die	**Frequency**
1	
2	
3	
4	
5	
6	

Table 1

Second Experiment (20 rolls)	
Face on Die	**Frequency**
1	
2	
3	
4	
5	
6	

Table 2

Did the two experiments have the same results? Probably not. If you did the experiment a third time, do you expect the results to be identical to the first or second experiment? Why or why not?

Which experiment had the correct results? They both did. The job of the statistician is to see through the variability and draw appropriate conclusions.

Section 1.3

Collaborative Exercise

In your class, have someone conduct a survey of the number of siblings (brothers and sisters) each student has. Create a frequency table. Add to it a relative frequency column and a cumulative relative frequency column. Answer the following questions:

1. What percentage of the students in your class have no siblings?

2. What percentage of the students have from one to three siblings?
3. What percentage of the students have fewer than three siblings?

Media Attributions

- Private: Chapter 1 Introduction Figure 1a © Markus Winkler (2020). "Corona death and new cases stats." adapted by Significant Statistics is licensed under a Public Domain license
- Chapter 1 Introduction Figure 1b © OpenStax Introductory Statistics is licensed under a CC BY (Attribution) license

1.1 DEFINITIONS OF STATISTICS, PROBABILITY, AND KEY TERMS

Learning Objectives

By the end of this section, the student should be able to:

- Recognize and differentiate between key terms.

The science of **statistics** deals with the collection, analysis, interpretation, and presentation of data. We see and use **data** in our everyday lives.

In this course, you will learn how to organize and summarize data. Organizing and summarizing data is called **descriptive statistics**. Two ways to summarize data are by graphing and by using numbers (for example, finding an average). After you have studied probability and probability distributions, you will use formal methods for drawing conclusions from "good" data. The formal methods are called **inferential statistics**. Statistical inference uses probability to determine how confident we can be that our conclusions are correct.

Effective interpretation of data (inference) is based on good procedures for producing data and thoughtful examination of the data. You will encounter what will seem to be too many mathematical formulas for interpreting data. The goal of statistics is not to perform numerous calculations using the formulas, but to gain an understanding of your data. The calculations can be done using a calculator or a computer. The understanding must come from you. If you can thoroughly grasp the basics of statistics, you can be more confident in the decisions you make in life.

Probability

Probability is a mathematical tool used to study randomness. It deals with the chance (the likelihood) of an

event occurring. For example, if you toss a fair coin four times, the outcomes may not be two heads and two tails. However, if you toss the same coin 4,000 times, the outcomes will be close to half heads and half tails. The expected theoretical probability of heads in any one toss is $\frac{1}{2}$ or 0.5. Even though the outcomes of a few repetitions are uncertain, there is a regular pattern of outcomes when there are many repetitions. After reading about the English statistician Karl Pearson who tossed a coin 24,000 times with a result of 12,012 heads, one of the authors tossed a coin 2,000 times. The results were 996 heads. The fraction $\frac{996}{2000}$ is equal to 0.498 which is very close to 0.5, the expected probability.

The theory of probability began with the study of games of chance such as poker. Predictions take the form of probabilities. To predict the likelihood of an earthquake, of rain, or whether you will get an A in this course, we use probabilities. Doctors use probability to determine the chance of a vaccination causing the disease the vaccination is supposed to prevent. A stockbroker uses probability to determine the rate of return on a client's investments. You might use probability to decide to buy a lottery ticket or not. In your study of statistics, you will use the power of mathematics through probability calculations to analyze and interpret your data.

Key Terms

In statistics, we generally want to study a **population**. You can think of a population as a collection of persons, things, or objects under study. To study the population, we select a **sample**. The idea of sampling is to select a portion (or subset) of the larger population and study that portion (the sample) to gain information about the population. Data are the result of sampling from a population.

Because it takes a lot of time and money to examine an entire population, sampling is a very practical technique. If you wished to compute the overall grade point average at your school, it would make sense to select a sample of students who attend the school. The data collected from the sample would be the students' grade point averages. In presidential elections, opinion poll samples of 1,000–2,000 people are taken. The opinion poll is supposed to represent the views of the people in the entire country. Manufacturers of canned carbonated drinks take samples to determine if a 16 ounce can contains 16 ounces of carbonated drink.

From the sample data, we can calculate a **statistic**. A statistic is a number that represents a property of the sample. For example, if we consider one math class to be a sample of the population of all math classes, then the average number of points earned by students in that one math class at the end of the term is an example of a statistic. The statistic is an estimate of a population **parameter**. A parameter is a number that is a property of the population. Since we considered all math classes to be the population, then the average number of points earned per student over all the math classes is an example of a parameter.

One of the main concerns in the field of statistics is how accurately a statistic estimates a parameter. The accuracy really depends on how well the sample represents the population. The sample must contain the characteristics of the population in order to be a representative sample. We are interested in both the sample

statistic and the population parameter in inferential statistics. In a later chapter, we will use the sample statistic to test the validity of the established population parameter.

A **variable**, notated by capital letters such as X and Y, is a characteristic of interest for each person or thing in a population. Variables may be **numerical** or **categorical**. **Numerical variables** take on values with equal units such as weight in pounds and time in hours. **Categorical variables** place the person or thing into a category. If we let X equal the number of points earned by one math student at the end of a term, then X is a numerical variable. If we let Y be a person's party affiliation, then some examples of Y include Republican, Democrat, and Independent. Y is a categorical variable. We could do some math with values of X (calculate the average number of points earned, for example), but it makes no sense to do math with values of Y (calculating an average party affiliation makes no sense).

Data are the actual values of the variable. They may be numbers or they may be words. **Datum** is a single value.

Two words that come up often in statistics are **mean** and **proportion**. If you were to take three exams in your math classes and obtain scores of 86, 75, and 92, you would calculate your mean score by adding the three exam scores and dividing by three (your mean score would be 84.3 to one decimal place). If, in your math class, there are 40 students and 22 are men and 18 are women, then the proportion of men students is $\frac{22}{40}$ and the proportion of women students is $\frac{18}{40}$. Mean and proportion are discussed in more detail in later chapters.

NOTE

The words "mean" and "average" are often used interchangeably. The substitution of one word for the other is common practice. The technical term is "arithmetic mean," and "average" is technically a center location. However, in practice among non-statisticians, "average" is commonly accepted for "arithmetic mean."

Example

Determine what the key terms refer to in the following study. We want to know the average (mean) amount of money first year college students spend at ABC College on school supplies that do not include books. We randomly survey 100 first year students at the college. Three of those students spent $150, $200, and $225, respectively.

Solution

The **population** is all first year students attending ABC College this term.

The **sample** could be all students enrolled in one section of a beginning statistics course at ABC College (although this sample may not represent the entire population).

The **parameter** is the average (mean) amount of money spent (excluding books) by first year college students at ABC College this term.

The **statistic** is the average (mean) amount of money spent (excluding books) by first year college students in the sample.

The **variable** could be the amount of money spent (excluding books) by one first year student. Let X = the amount of money spent (excluding books) by one first year student attending ABC College.

The **data** are the dollar amounts spent by the first year students. Examples of the data are $150, $200, and $225.

Try It

Determine what the key terms refer to in the following study. We want to know the average (mean) amount of money spent on school uniforms each year by families with children at Knoll Academy. We randomly survey 100 families with children in the school. Three of the families spent $65, $75, and $95, respectively.

Solution

The **population** is all families with children attending Knoll Academy.

The **sample** is a random selection of 100 families with children attending Knoll Academy.

The **parameter** is the average (mean) amount of money spent on school uniforms by families with children at Knoll Academy.

The **statistic** is the average (mean) amount of money spent on school uniforms by families in the sample.

The **variable** is the amount of money spent by one family. Let X = the amount of money spent on school uniforms by one family with children attending Knoll Academy.

The **data** are the dollar amounts spent by the families. Examples of the data are $65, $75, and $95.

Determine what the key terms refer to in the following study.

A study was conducted at a local college to analyze the average cumulative GPAs of students who graduated last year. Fill in the letter of the phrase that best describes each of the items below.

1._____ Population 2._____ Statistic 3._____ Parameter 4._____ Sample 5._____ Variable 6._____ Data

- a) all students who attended the college last year
- b) the cumulative GPA of one student who graduated from the college last year
- c) 3.65, 2.80, 1.50, 3.90
- d) a group of students who graduated from the college last year, randomly selected
- e) the average cumulative GPA of students who graduated from the college last year
- f) all students who graduated from the college last year
- g) the average cumulative GPA of students in the study who graduated from the college last year

Solution

1. f
2. g
3. e
4. d
5. b
6. c

Determine what the key terms refer to in the following study.

As part of a study designed to test the safety of automobiles, the National Transportation Safety Board collected and reviewed data about the effects of an automobile crash on test dummies. Here is the criterion they used:

Speed at which Cars Crashed	Location of "drive" (i.e. dummies)
35 miles/hour	Front Seat

Cars with dummies in the front seats were crashed into a wall at a speed of 35 miles per hour. We want to know the proportion of dummies in the driver's seat that would have had head injuries, if they had been actual drivers. We start with a simple random sample of 75 cars.

Solution

The **population** is all cars containing dummies in the front seat.

The **sample** is the 75 cars, selected by a simple random sample.

The **parameter** is the proportion of driver dummies (if they had been real people) who would have suffered head injuries in the population.

The **statistic** is the proportion of driver dummies (if they had been real people) who would have suffered head injuries in the sample.

The **variable** X = the number of driver dummies (if they had been real people) who would have suffered head injuries.

The **data** are either: yes, had head injury, or no, did not.

Determine what the key terms refer to in the following study.

An insurance company would like to determine the proportion of all medical doctors who have been involved in one or more malpractice lawsuits. The company selects 500 doctors at random from a professional directory and determines the number in the sample who have been involved in a malpractice lawsuit.

Solution

The **population** is all medical doctors listed in the professional directory.

The **parameter** is the proportion of medical doctors who have been involved in one or more malpractice suits in the population.

The **sample** is the 500 doctors selected at random from the professional directory.

The **statistic** is the proportion of medical doctors who have been involved in one or more malpractice suits in the sample.

The **variable** X = the number of medical doctors who have been involved in one or more malpractice suits.

The **data** are either: yes, was involved in one or more malpractice lawsuits, or no, was not.

References

The Data and Story Library, http://lib.stat.cmu.edu/DASL/Stories/CrashTestDummies.html (accessed May 1, 2013).

1.2 DATA, SAMPLING, AND VARIATION IN DATA AND SAMPLING

Learning Objectives

By the end of this section, the student should be able to:

- Classify data as qualitative or quantitative, continuous, discrete, or neither.
- Determine types of graphs appropriate for specific data.
- Apply various types of sampling methods to data collection.

Data may come from a **population** or from a **sample**. Small letters like x or y generally are used to represent data values. Most data can be put into the following categories:

- **Qualitative**
- **Quantitative**

Qualitative data are the result of categorizing or describing attributes of a population. Hair color, blood type, ethnic group, the car a person drives, and the street a person lives on are examples of qualitative data. Qualitative data are generally described by words or letters. For instance, hair color might be black, dark brown, light brown, blond, gray, or red. Blood type might be AB+, O-, or B+. Researchers often prefer to use quantitative data over qualitative data because it lends itself more easily to mathematical analysis. For example, it does not make sense to find an average hair color or blood type.

Quantitative data are always numbers. Quantitative data are the result of **counting** or **measuring** attributes of a population. Amount of money, pulse rate, weight, number of people living in your town, and number of students who take statistics are examples of quantitative data. Quantitative data may be either **discrete** or **continuous**.

All data that are the result of counting are called quantitative discrete data. These data take on only certain

numerical values. If you count the number of phone calls you receive for each day of the week, you might get values such as zero, one, two, or three.

All data that are the result of measuring are quantitative continuous data assuming that we can measure accurately. Measuring angles in radians might result in such numbers as $\frac{\pi}{6}$, $\frac{\pi}{3}$, $\frac{\pi}{2}$, π, $\frac{3\pi}{4}$, and so on. If you and your friends carry backpacks with books in them to school, the numbers of books in the backpacks are discrete data and the weights of the backpacks are continuous data.

DATA SAMPLE OF QUANTITATIVE DISCRETE DATA

The data are the number of books students carry in their backpacks. You sample five students. Two students carry three books, one student carries four books, one student carries two books, and one student carries one book. The numbers of books (three, four, two, and one) are the quantitative discrete data.

Try It

The data are the number of machines in a gym. You sample five gyms. One gym has 12 machines, one gym has 15 machines, one gym has ten machines, one gym has 22 machines, and the other gym has 20 machines. What type of data is this?

Solution
quantitative discrete data

DATA SAMPLE OF QUANTITATIVE CONTINUOUS DATA

The data are the weights of backpacks with books in them. You sample the same five students. The weights

(in pounds) of their backpacks are 6.2, 7, 6.8, 9.1, 4.3. Notice that backpacks carrying three books can have different weights. Weights are quantitative continuous data because weights are measured.

Try It

The data are the areas of lawns in square feet. You sample five houses. The areas of the lawns are 144 sq. feet, 160 sq. feet, 190 sq. feet, 180 sq. feet, and 210 sq. feet. What type of data is this?

Solution
quantitative continuous data

You go to the supermarket and purchase three cans of soup (19 ounces of tomato bisque, 14.1 ounces of lentil, and 19 ounces of Italian wedding), two packages of nuts (walnuts and peanuts), four different kinds of vegetable (broccoli, cauliflower, spinach, and carrots), and two desserts (16 ounces of Cherry Garcia ice cream and two pounds or 32 ounces of chocolate chip cookies).

Name data sets that are quantitative discrete, quantitative continuous, and qualitative.

Solution
One Possible Solution:

- The three cans of soup, two packages of nuts, four kinds of vegetables and two desserts are quantitative discrete data because you count them.
- The weights of the soups (19 ounces, 14.1 ounces, 19 ounces) are quantitative continuous data because you measure weights as precisely as possible.
- Types of soups, nuts, vegetables and desserts are qualitative data because they are categorical.

Try to identify additional data sets in this example.

The data are the colors of backpacks. Again, you sample the same five students. One student has a red backpack, two students have black backpacks, one student has a green backpack, and one student has a gray backpack. The colors red, black, black, green, and gray are qualitative data.

Try It

The data are the colors of houses. You sample five houses. The colors of the houses are white, yellow, white, red, and white. What type of data is this?

Solution
qualitative data

NOTE

You may collect data as numbers and report it categorically. For example, the quiz scores for each student are recorded throughout the term. At the end of the term, the quiz scores are reported as A, B, C, D, or F.

Try It

Work collaboratively to determine the correct data type (quantitative or qualitative). Indicate whether quantitative data are continuous or discrete. Hint: Data that are discrete often start with the words "the number of."

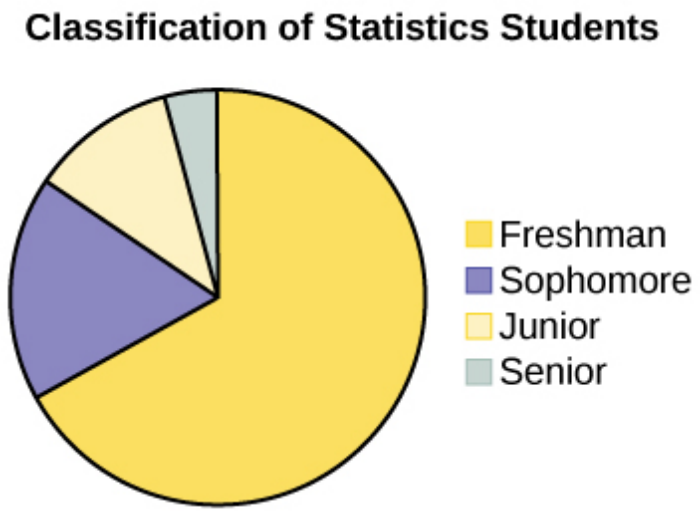
1. the number of pairs of shoes you own
2. the type of car you drive
3. where you go on vacation
4. the distance it is from your home to the nearest grocery store
5. the number of classes you take per school year.
6. the tuition for your classes
7. the type of calculator you use
8. movie ratings
9. political party preferences
10. weights of sumo wrestlers
11. amount of money (in dollars) won playing poker
12. number of correct answers on a quiz
13. people's attitudes toward the government
14. IQ scores (This may cause some discussion.)

Solution
Items 1, 5, 6, 11, and 12 are quantitative discrete; items 4, 10, and 14 are quantitative continuous; items 2, 3, 7, 8, 9, and 13 are qualitative.

Determine the correct data type (quantitative or qualitative) for the number of cars in a parking lot. Indicate whether quantitative data are continuous or discrete.

Solution
quantitative discrete

A statistics professor collects information about the classification of her students as freshmen, sophomores, juniors, or seniors. The data she collects are summarized in the pie chart [link]. What type of data does this graph show?

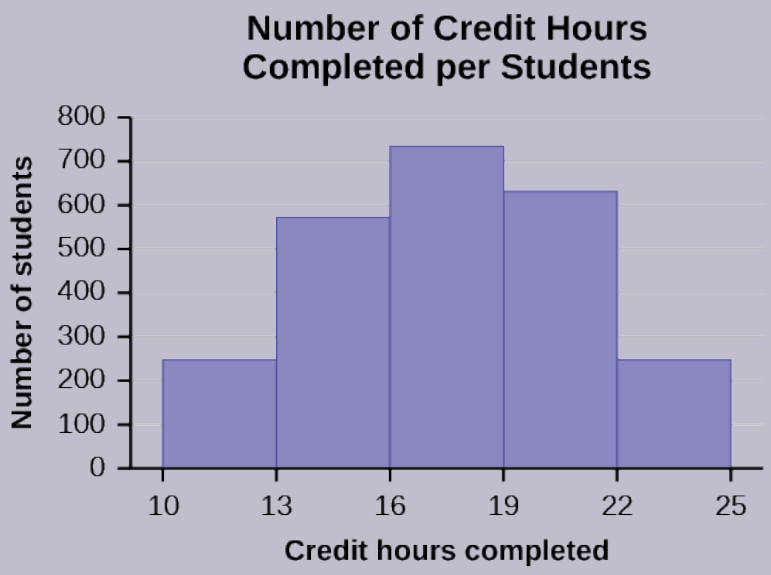


Solution

This pie chart shows the students in each year, which is **qualitative data**.

Try It

The registrar at State University keeps records of the number of credit hours students complete each semester. The data he collects are summarized in the histogram. The class boundaries are 10 to less than 13, 13 to less than 16, 16 to less than 19, 19 to less than 22, and 22 to less than 25.



What type of data does this graph show?

Solution

A histogram is used to display quantitative data: the numbers of credit hours completed. Because students can complete only a whole number of hours (no fractions of hours allowed), this data is quantitatively discrete.

Qualitative Data Discussion

Below are tables comparing the number of part-time and full-time students at De Anza College and Foothill College enrolled for the spring 2010 quarter. The tables display counts (frequencies) and percentages or proportions (relative frequencies). The percent columns make comparing the same categories in the colleges easier. Displaying percentages along with the numbers is often helpful, but it is particularly important when comparing sets of data that do not have the same totals, such as the total enrollments for both colleges in this example. Notice how much larger the percentage for part-time students at Foothill College is compared to De Anza College.

Fall Term 2007 (Census day)

De Anza College			**Foothill College**		
	Number	Percent		Number	Percent
Full-time	9,200	40.9%	Full-time	4,059	28.6%
Part-time	13,296	59.1%	Part-time	10,124	71.4%
Total	22,496	100%	Total	14,183	100%

Tables are a good way of organizing and displaying data. But graphs can be even more helpful in understanding the data. There are no strict rules concerning which graphs to use. Two graphs that are used to display qualitative data are pie charts and bar graphs.

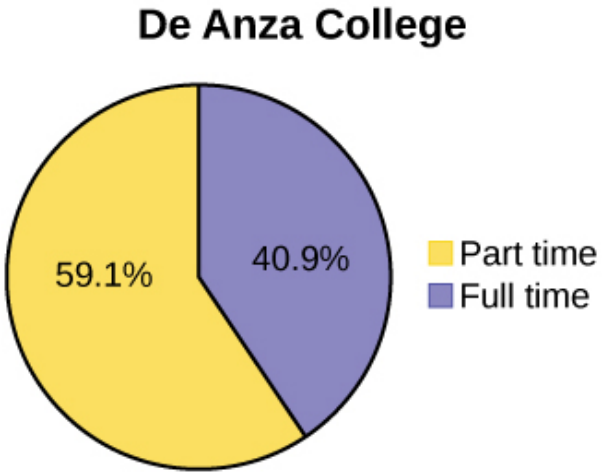
In a pie chart, categories of data are represented by wedges in a circle and are proportional in size to the percent of individuals in each category.

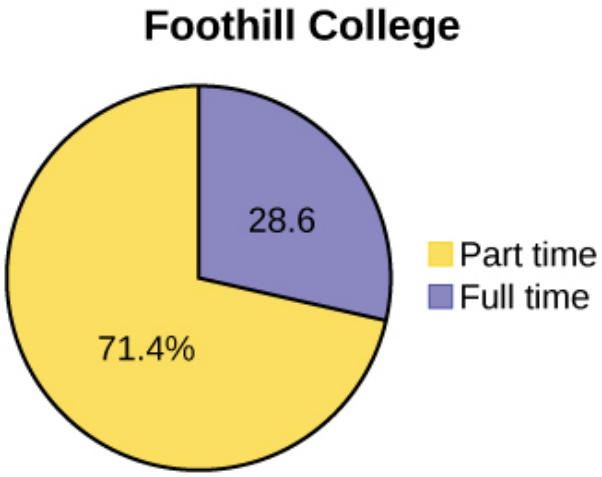
In a bar graph, the length of the bar for each category is proportional to the number or percent of individuals in each category. Bars may be vertical or horizontal.

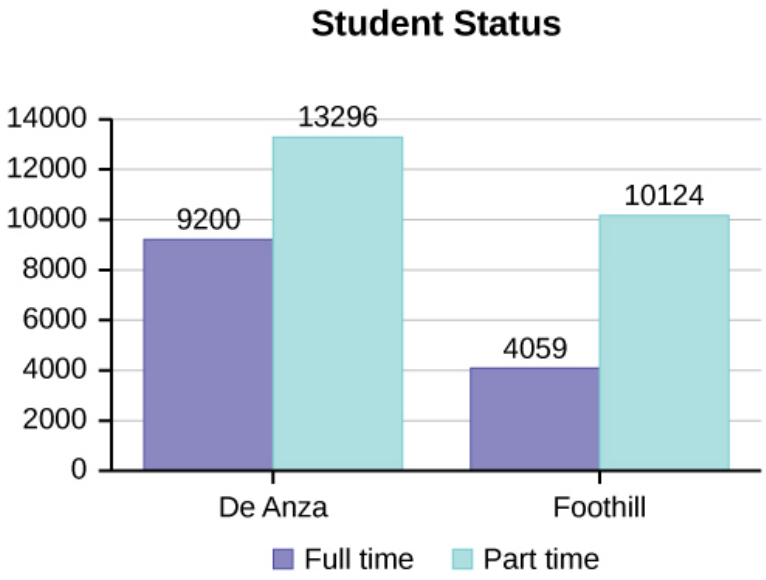
A Pareto chart consists of bars that are sorted into order by category size (largest to smallest).

Look at [link] and [link] and determine which graph (pie or bar) you think displays the comparisons better.

It is a good idea to look at a variety of graphs to see which is the most helpful in displaying the data. We might make different choices of what we think is the "best" graph depending on the data and the context. Our choice also depends on what we are using the data for.





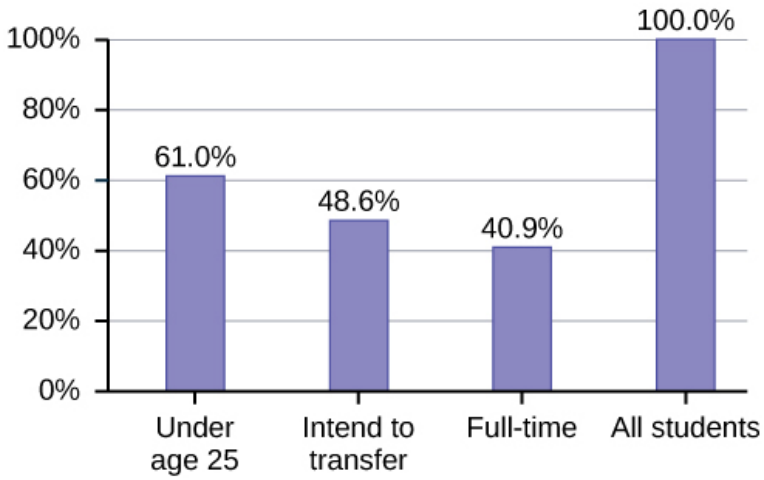


Percentages That Add to More (or Less) Than 100%

Sometimes percentages add up to be more than 100% (or less than 100%). In the graph, the percentages add to more than 100% because students can be in more than one category. A bar graph is appropriate to compare the relative size of the categories. A pie chart cannot be used. It also could not be used if the percentages added to less than 100%.

De Anza College Spring 2010

Characteristic/Category	Percent
Full-Time Students	40.9%
Students who intend to transfer to a 4-year educational institution	48.6%
Students under age 25	61.0%
TOTAL	150.5%

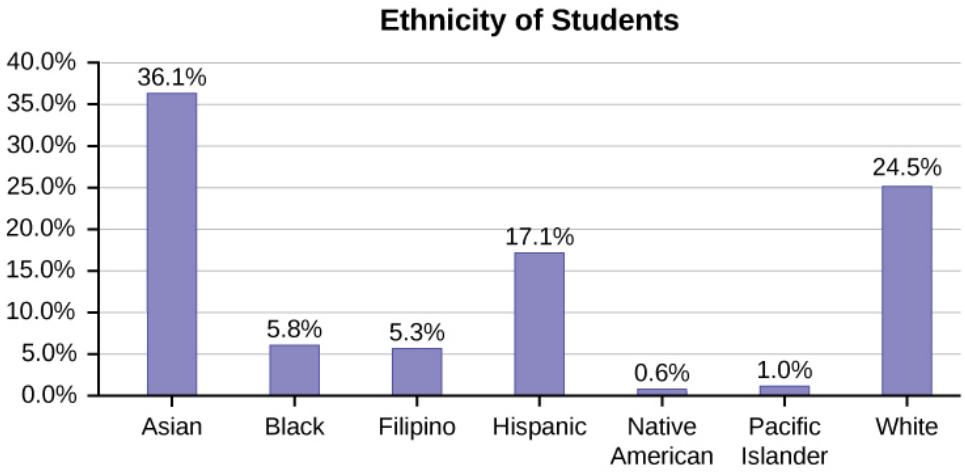


Omitting Categories/Missing Data

The table displays Ethnicity of Students but is missing the "Other/Unknown" category. This category contains people who did not feel they fit into any of the ethnicity categories or declined to respond. Notice that the frequencies do not add up to the total number of students. In this situation, create a bar graph and not a pie chart.

Ethnicity of Students at De Anza College Fall Term 2007 (Census Day)

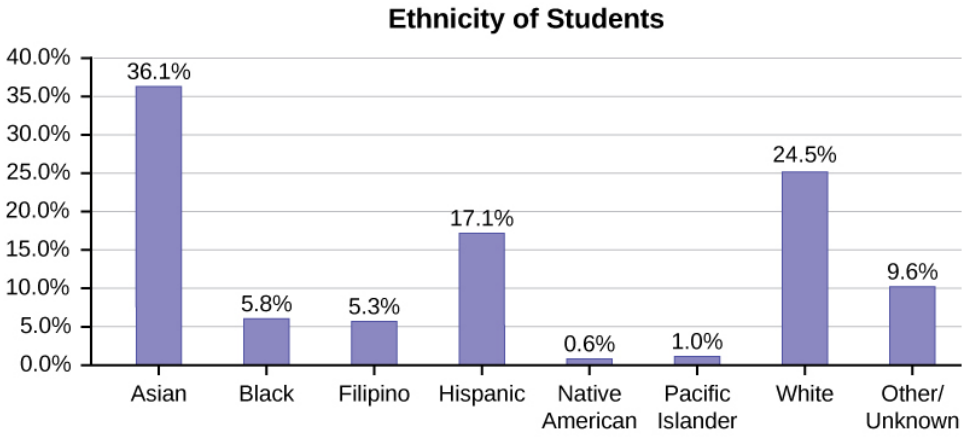
	Frequency	**Percent**
Asian	8,794	36.1%
Black	1,412	5.8%
Filipino	1,298	5.3%
Hispanic	4,180	17.1%
Native American	146	0.6%
Pacific Islander	236	1.0%
White	5,978	24.5%
TOTAL	22,044 out of 24,382	90.4% out of 100%



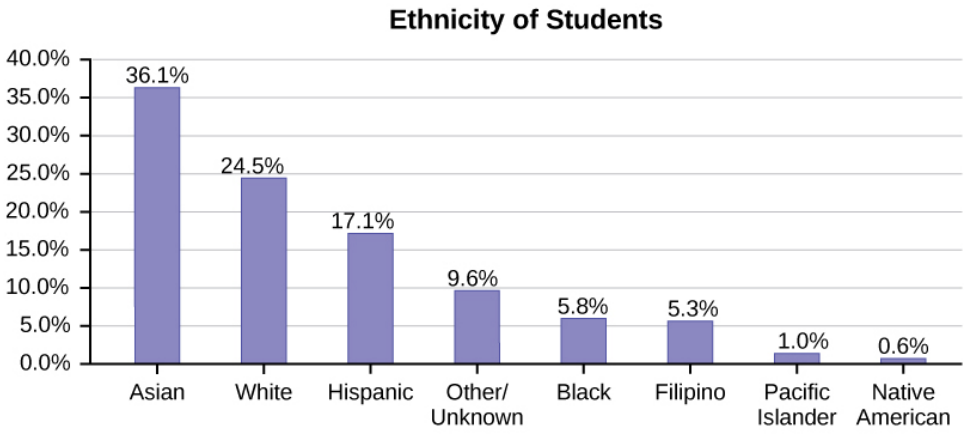
The following graph is the same as the previous graph but the "Other/Unknown" percent (9.6%) has been included. The "Other/Unknown" category is large compared to some of the other categories (Native American, 0.6%; Pacific Islander, 1.0%). This is important to know when we think about what the data are telling us.

This particular bar graph in [link] can be difficult to understand visually. The graph in [link] is a Pareto chart. The Pareto chart has the bars sorted from largest to smallest and is easier to read and interpret.

BAR GRAPH WITH OTHER/UNKNOWN CATEGORY

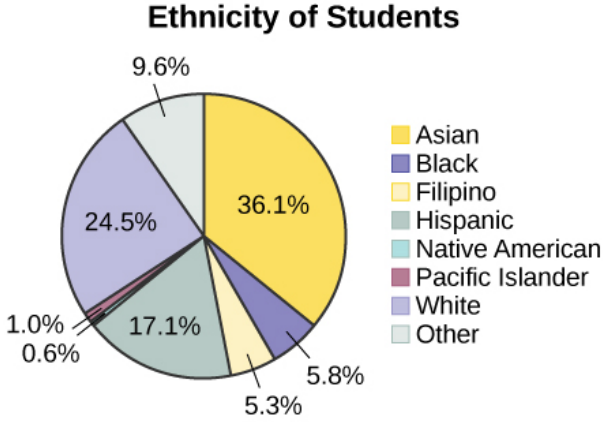


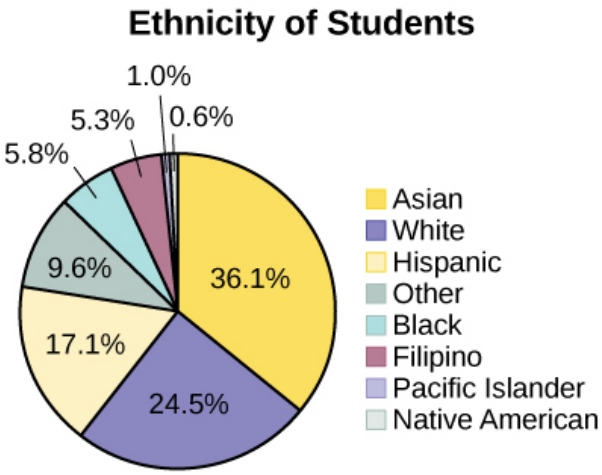
PARETO CHART WITH BARS SORTED BY SIZE



Pie Charts: No Missing Data

The following pie charts have the "Other/Unknown" category included (since the percentages must add to 100%). The chart in Table I is organized by the size of each wedge, which makes it a more visually informative graph than the unsorted, alphabetical graph in [link].





Sampling

Gathering information about an entire population often costs too much or is virtually impossible. Instead, we use a sample of the population. **A sample should have the same characteristics as the population it is representing.** Most statisticians use various methods of **random sampling** in an attempt to achieve this goal. This section will describe a few of the most common methods. There are several different methods of **random sampling**. In each form of random sampling, each member of a population initially has an equal chance of being selected for the sample. Each method has pros and cons. The easiest method to describe is called a **simple random sample**. Any group of n individuals is equally likely to be chosen by any other group of n individuals if the **simple random sampling** technique is used. In other words, each sample of the same size has an equal chance of being selected. For example, suppose Lisa wants to form a four-person study group (herself and three other people) from her pre-calculus class, which has 31 members not including Lisa. To choose a simple random sample of size three from the other members of her class, Lisa could put all 31 names in a hat, shake the hat, close her eyes, and pick out three names. A more technological way is for Lisa to first list the last names of the members of her class together with a two-digit number, as in [link]:

Class Roster

ID	Name	ID	Name	ID	Name
00	Anselmo	11	King	21	Roquero
01	Bautista	12	Legeny	22	Roth
02	Bayani	13	Lundquist	23	Rowell
03	Cheng	14	Macierz	24	Salangsang
04	Cuarismo	15	Motogawa	25	Slade
05	Cuningham	16	Okimoto	26	Stratcher
06	Fontecha	17	Patel	27	Tallai
07	Hong	18	Price	28	Tran
08	Hoobler	19	Quizon	29	Wai
09	Jiao	20	Reyes	30	Wood
10	Khan				

Lisa can use a table of random numbers (found in many statistics books and mathematical handbooks), a calculator, or a computer to generate random numbers. For this example, suppose Lisa chooses to generate random numbers from a calculator. The numbers generated are as follows:

0.94360, 0.99832, 0.14669, 0.51470, 0.40581, 0.73381, 0.04399

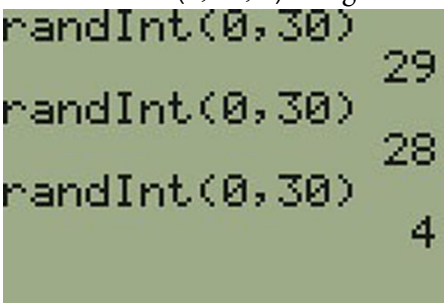
Lisa reads two-digit groups until she has chosen three class members (that is, she reads 0.94360 as the groups 94, 43, 36, 60). Each random number may only contribute one class member. If she needed to, Lisa could have generated more random numbers.

The random numbers 0.94360 and 0.99832 do not contain appropriate two digit numbers. However the third random number, 0.14669, contains 14 (the fourth random number also contains 14), the fifth random number contains 05, and the seventh random number contains 04. The two-digit number 14 corresponds to Macierz, 05 corresponds to Cuningham, and 04 corresponds to Cuarismo. Besides herself, Lisa's group will consist of Marcierz, Cuningham, and Cuarismo.

To generate random numbers:

- Press MATH.
- Arrow over to PRB.
- Press 5:randInt(Enter 0, 30).
- Press ENTER for the first random number.
- Press ENTER two more times for the other 2 random numbers. If there is a repeat press ENTER again.

Note: randInt(0, 30, 3) will generate 3 random numbers.



Besides simple random sampling, there are other forms of sampling that involve a chance process for getting the sample. **Other well-known random sampling methods are the stratified sample, the cluster sample, and the systematic sample.**

To choose a **stratified sample**, divide the population into groups called strata and then take a **proportionate** number from each stratum. For example, you could stratify (group) your college population by department and then choose a proportionate simple random sample from each stratum (each department) to get a stratified random sample. To choose a simple random sample from each department, number each member of the first department, number each member of the second department, and do the same for the remaining departments. Then use simple random sampling to choose proportionate numbers from the first department and do the same for each of the remaining departments. Those numbers picked from the first department, picked from the second department, and so on represent the members who make up the stratified sample.

To choose a **cluster sample**, divide the population into clusters (groups) and then randomly select some of the clusters. All the members from these clusters are in the cluster sample. For example, if you randomly sample four departments from your college population, the four departments make up the cluster sample. Divide your college faculty by department. The departments are the clusters. Number each department, and then choose four different numbers using simple random sampling. All members of the four departments with those numbers are the cluster sample.

To choose a **systematic sample**, randomly select a starting point and take every n^{th} piece of data from a listing of the population. For example, suppose you have to do a phone survey. Your phone book contains 20,000 residence listings. You must choose 400 names for the sample. Number the population 1–20,000 and then use a simple random sample to pick a number that represents the first name in the sample. Then choose every fiftieth name thereafter until you have a total of 400 names (you might have to go back to the beginning of your phone list). Systematic sampling is frequently chosen because it is a simple method.

A type of sampling that is non-random is convenience sampling. **Convenience sampling** involves using results that are readily available. For example, a computer software store conducts a marketing study by interviewing potential customers who happen to be in the store browsing through the available software. The results of convenience sampling may be very good in some cases and highly biased (favor certain outcomes) in others.

Sampling data should be done very carefully. Collecting data carelessly can have devastating results. Surveys mailed to households and then returned may be very biased (they may favor a certain group). It is better for the person conducting the survey to select the sample respondents.

True random sampling is done **with replacement**. That is, once a member is picked, that member goes back into the population and thus may be chosen more than once. However for practical reasons, in most populations, simple random sampling is done **without replacement**. Surveys are typically done without replacement. That is, a member of the population may be chosen only once. Most samples are taken from large populations and the sample tends to be small in comparison to the population. Since this is the case, sampling without replacement is approximately the same as sampling with replacement because the chance of picking the same individual more than once with replacement is very low.

In a college population of 10,000 people, suppose you want to pick a sample of 1,000 randomly for a survey. **For any particular sample of 1,000**, if you are sampling **with replacement**,

- the chance of picking the first person is 1,000 out of 10,000 (0.1000);
- the chance of picking a different second person for this sample is 999 out of 10,000 (0.0999);
- the chance of picking the same person again is 1 out of 10,000 (very low).

If you are sampling **without replacement**,

- the chance of picking the first person for any particular sample is 1000 out of 10,000 (0.1000);
- the chance of picking a different second person is 999 out of 9,999 (0.0999);
- you do not replace the first person before picking the next person.

Compare the fractions 999/10,000 and 999/9,999. For accuracy, carry the decimal answers to four decimal places. To four decimal places, these numbers are equivalent (0.0999).

Sampling without replacement instead of sampling with replacement becomes a mathematical issue only when the population is small. For example, if the population is 25 people, the sample is ten, and you are sampling **with replacement for any particular sample**, then the chance of picking the first person is ten out of 25, and the chance of picking a different second person is nine out of 25 (you replace the first person).

If you sample **without replacement**, then the chance of picking the first person is ten out of 25, and then the chance of picking the second person (who is different) is nine out of 24 (you do not replace the first person).

Compare the fractions 9/25 and 9/24. To four decimal places, 9/25 = 0.3600 and 9/24 = 0.3750. To four decimal places, these numbers are not equivalent.

When you analyze data, it is important to be aware of **sampling errors** and nonsampling errors. The actual process of sampling causes sampling errors. For example, the sample may not be large enough. Factors

not related to the sampling process cause **nonsampling errors.** A defective counting device can cause a nonsampling error.

In reality, a sample will never be exactly representative of the population so there will always be some sampling error. As a rule, the larger the sample, the smaller the sampling error.

In statistics, **a sampling bias** is created when a sample is collected from a population and some members of the population are not as likely to be chosen as others (remember, each member of the population should have an equally likely chance of being chosen). When a sampling bias happens, there can be incorrect conclusions drawn about the population that is being studied.

Example

A study is done to determine the average tuition that San Jose State undergraduate students pay per semester. Each student in the following samples is asked how much tuition he or she paid for the Fall semester. What is the type of sampling in each case?

1. A sample of 100 undergraduate San Jose State students is taken by organizing the students' names by classification (freshman, sophomore, junior, or senior), and then selecting 25 students from each.
2. A random number generator is used to select a student from the alphabetical listing of all undergraduate students in the Fall semester. Starting with that student, every 50th student is chosen until 75 students are included in the sample.
3. A completely random method is used to select 75 students. Each undergraduate student in the fall semester has the same probability of being chosen at any stage of the sampling process.
4. The freshman, sophomore, junior, and senior years are numbered one, two, three, and four, respectively. A random number generator is used to pick two of those years. All students in those two years are in the sample.
5. An administrative assistant is asked to stand in front of the library one Wednesday and to ask the first 100 undergraduate students he encounters what they paid for tuition the Fall

semester. Those 100 students are the sample.

Solution
1. stratified; 2. systematic; 3. simple random; 4. cluster; 5. convenience

Try It

You are going to use the random number generator to generate different types of samples from the data.

This table displays six sets of quiz scores (each quiz counts 10 points) for an elementary statistics class.

#1	#2	#3	#4	#5	#6
5	7	10	9	8	3
10	5	9	8	7	6
9	10	8	6	7	9
9	10	10	9	8	9
7	8	9	5	7	4
9	9	9	10	8	7
7	7	10	9	8	8
8	8	9	10	8	8
9	7	8	7	7	8
8	8	10	9	8	7

Instructions: Use the Random Number Generator to pick samples.

1. Create a stratified sample by column. Pick three quiz scores randomly from each column.
    - Number each row one through ten.
    - On your calculator, press Math and arrow over to PRB.

- For column 1, Press 5:randInt(and enter 1,10). Press ENTER. Record the number. Press ENTER 2 more times (even the repeats). Record these numbers. Record the three quiz scores in column one that correspond to these three numbers.
- Repeat for columns two through six.
- These 18 quiz scores are a stratified sample.

2. Create a cluster sample by picking two of the columns. Use the column numbers: one through six.
    - Press MATH and arrow over to PRB.
    - Press 5:randInt(and enter 1,6). Press ENTER. Record the number. Press ENTER and record that number.
    - The two numbers are for two of the columns.
    - The quiz scores (20 of them) in these 2 columns are the cluster sample.
3. Create a simple random sample of 15 quiz scores.
    - Use the numbering one through 60.
    - Press MATH. Arrow over to PRB. Press 5:randInt(and enter 1, 60).
    - Press ENTER 15 times and record the numbers.
    - Record the quiz scores that correspond to these numbers.
    - These 15 quiz scores are the systematic sample.
4. Create a systematic sample of 12 quiz scores.
    - Use the numbering one through 60.
    - Press MATH. Arrow over to PRB. Press 5:randInt(and enter 1, 60).
    - Press ENTER. Record the number and the first quiz score. From that number, count ten quiz scores and record that quiz score. Keep counting ten quiz scores and recording the quiz score until you have a sample of 12 quiz scores. You may wrap around (go back to the beginning).

Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

1. A soccer coach selects six players from a group of boys aged eight to ten, seven players from a group of boys aged 11 to 12, and three players from a group of boys aged 13 to 14 to form a recreational soccer team.
2. A pollster interviews all human resource personnel in five different high tech companies.
3. A high school educational researcher interviews 50 high school female teachers and 50 high school male teachers.
4. A medical researcher interviews every third cancer patient from a list of cancer patients at a local hospital.
5. A high school counselor uses a computer to generate 50 random numbers and then picks students whose names correspond to the numbers.
6. A student interviews classmates in his algebra class to determine how many pairs of jeans a student owns, on the average.

Solution

a. stratified; b. cluster; c. stratified; d. systematic; e. simple random; f. convenience

Try It

Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

A high school principal polls 50 freshmen, 50 sophomores, 50 juniors, and 50 seniors regarding policy changes for after school activities.

Solution

stratified

If we were to examine two samples representing the same population, even if we used random sampling methods for the samples, they would not be exactly the same. Just as there is variation in data, there is variation in samples. As you become accustomed to sampling, the variability will begin to seem natural.

Suppose ABC College has 10,000 part-time students (the population). We are interested in the average amount of money a part-time student spends on books in the fall term. Asking all 10,000 students is an almost impossible task.

Suppose we take two different samples.

First, we use convenience sampling and survey ten students from a first term organic chemistry class. Many of these students are taking first term calculus in addition to the organic chemistry class. The amount of money they spend on books is as follows:

$128; $87; $173; $116; $130; $204; $147; $189; $93; $153

The second sample is taken using a list of senior citizens who take P.E. classes and taking every fifth senior citizen on the list, for a total of ten senior citizens. They spend:

$50; $40; $36; $15; $50; $100; $40; $53; $22; $22

It is unlikely that any student is in both samples.

a. Do you think that either of these samples is representative of (or is characteristic of) the entire 10,000 part-time student population?

Solution
a. No. The first sample probably consists of science-oriented students. Besides the chemistry course, some of them are also taking first-term calculus. Books for these classes tend to be expensive. Most of these students are, more than likely, paying more than the average part-time student for their books. The second sample is a group of senior citizens who are, more than likely, taking courses for health and interest. The amount of money they spend on books is probably much less than the average part-time student. Both samples are biased. Also, in both cases, not all students have a chance to be in either sample.

b. Since these samples are not representative of the entire population, is it wise to use the results to describe the entire population?

Solution
b. No. For these samples, each member of the population did not have an equally likely chance of being chosen.

Now, suppose we take a third sample. We choose ten different part-time students from the disciplines of chemistry, math, English, psychology, sociology, history, nursing, physical education,

art, and early childhood development. (We assume that these are the only disciplines in which part-time students at ABC College are enrolled and that an equal number of part-time students are enrolled in each of the disciplines.) Each student is chosen using simple random sampling. Using a calculator, random numbers are generated and a student from a particular discipline is selected if he or she has a corresponding number. The students spend the following amounts:

$180; $50; $150; $85; $260; $75; $180; $200; $200; $150

c. Is the sample biased?

Solution

c. The sample is unbiased, but a larger sample would be recommended to increase the likelihood that the sample will be close to representative of the population. However, for a biased sampling technique, even a large sample runs the risk of not being representative of the population.

Students often ask if it is "good enough" to take a sample, instead of surveying the entire population. If the survey is done well, the answer is yes.

Try It

A local radio station has a fan base of 20,000 listeners. The station wants to know if its audience would prefer more music or more talk shows. Asking all 20,000 listeners is an almost impossible task.

The station uses convenience sampling and surveys the first 200 people they meet at one of the station's music concert events. 24 people said they'd prefer more talk shows, and 176 people said they'd prefer more music.

Do you think that this sample is representative of (or is characteristic of) the entire 20,000 listener population?

Solution

The sample probably consists more of people who prefer music because it is a concert event. Also, the sample represents only those who showed up to the event earlier than the majority. The

sample probably doesn't represent the entire fan base and is probably biased towards people who would prefer music.

Variation in Data

Variation is present in any set of data. For example, 16-ounce cans of beverage may contain more or less than 16 ounces of liquid. In one study, eight 16 ounce cans were measured and produced the following amount (in ounces) of beverage:

15.8, 16.1, 15.2, 14.8, 15.8, 15.9, 16.0, 15.5

Measurements of the amount of beverage in a 16-ounce can may vary because different people make the measurements or because the exact amount, 16 ounces of liquid, was not put into the cans. Manufacturers regularly run tests to determine if the amount of beverage in a 16-ounce can falls within the desired range.

Be aware that as you take data, your data may vary somewhat from the data someone else is taking for the same purpose. This is completely natural. However, if two or more of you are taking the same data and get very different results, it is time for you and the others to reevaluate your data-taking methods and your accuracy.

Variation in Samples

It was mentioned previously that two or more samples from the same population, taken randomly, and having close to the same characteristics of the population will likely be different from each other. Suppose Doreen and Jung both decide to study the average amount of time students at their college sleep each night. Doreen and Jung each take samples of 500 students. Doreen uses systematic sampling and Jung uses cluster sampling. Doreen's sample will be different from Jung's sample. Even if Doreen and Jung used the same sampling method, in all likelihood their samples would be different. Neither would be wrong, however.

Think about what contributes to making Doreen's and Jung's samples different.

If Doreen and Jung took larger samples (i.e. the number of data values is increased), their sample results (the average amount of time a student sleeps) might be closer to the actual population average. But still, their samples would be, in all likelihood, different from each other. This **variability in samples** cannot be stressed enough.

Size of a Sample

The size of a sample (often called the number of observations) is important. The examples you have seen in this book so far have been small. Samples of only a few hundred observations, or even smaller, are sufficient for many purposes. In polling, samples that are from 1,200 to 1,500 observations are considered large enough and good enough if the survey is random and is well done. You will learn why when you study confidence intervals.

Be aware that many large samples are biased. For example, call-in surveys are invariably biased, because people choose to respond or not.

Critical Evaluation

We need to evaluate the statistical studies we read about critically and analyze them before accepting the results of the studies. Common problems to be aware of include

- Problems with samples: A sample must be representative of the population. A sample that is not representative of the population is biased. Biased samples that are not representative of the population give results that are inaccurate and not valid.
- Self-selected samples: Responses only by people who choose to respond, such as call-in surveys, are often unreliable.
- Sample size issues: Samples that are too small may be unreliable. Larger samples are better, if possible. In some situations, having small samples is unavoidable and can still be used to draw conclusions. Examples: crash testing cars or medical testing for rare conditions
- Undue influence: collecting data or asking questions in a way that influences the response
- Non-response or refusal of subject to participate: The collected responses may no longer be representative of the population. Often, people with strong positive or negative opinions may answer surveys, which can affect the results.
- Causality: A relationship between two variables does not mean that one causes the other to occur. They may be related (correlated) because of their relationship through a different variable.
- Self-funded or self-interest studies: A study performed by a person or organization in order to support their claim. Is the study impartial? Read the study carefully to evaluate the work. Do not automatically assume that the study is good, but do not automatically assume the study is bad either. Evaluate it on its merits and the work done.
- Misleading use of data: improperly displayed graphs, incomplete data, or lack of context
- Confounding: When the effects of multiple factors on a response cannot be separated. Confounding makes it difficult or impossible to draw valid conclusions about the effect of each factor.

References

Gallup-Healthways Well-Being Index. http://www.well-beingindex.com/default.asp (accessed May 1, 2013).

Gallup-Healthways Well-Being Index. http://www.well-beingindex.com/methodology.asp (accessed May 1, 2013).

Gallup-Healthways Well-Being Index. http://www.gallup.com/poll/146822/gallup-healthways-index-questions.aspx (accessed May 1, 2013).

Data from http://www.bookofodds.com/Relationships-Society/Articles/A0374-How-George-Gallup-Picked-the-President

Dominic Lusinchi, "'President' Landon and the 1936 *Literary Digest* Poll: Were Automobile and Telephone Owners to Blame?" Social Science History 36, no. 1: 23-54 (2012), http://ssh.dukejournals.org/content/36/1/23.abstract (accessed May 1, 2013).

"The Literary Digest Poll," Virtual Laboratories in Probability and Statistics http://www.math.uah.edu/stat/data/LiteraryDigest.html (accessed May 1, 2013).

"Gallup Presidential Election Trial-Heat Trends, 1936–2008," Gallup Politics http://www.gallup.com/poll/110548/gallup-presidential-election-trialheat-trends-19362004.aspx#4 (accessed May 1, 2013).

The Data and Story Library, http://lib.stat.cmu.edu/DASL/Datafiles/USCrime.html (accessed May 1, 2013).

LBCC Distance Learning (DL) program data in 2010-2011, http://de.lbcc.edu/reports/2010-11/future/highlights.html#focus (accessed May 1, 2013).

Data from San Jose Mercury News

Footnotes

1. 1 lastbaldeagle. 2013. On Tax Day, House to Call for Firing Federal Workers Who Owe Back Taxes. Opinion poll posted online at: http://www.youpolls.com/details.aspx?id=12328 (accessed May 1, 2013).
2. 2 Scott Keeter et al., "Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey," Public Opinion Quarterly 70 no. 5 (2006), http://poq.oxfordjournals.org/content/70/5/759.full (accessed May 1, 2013).
3. 3 Frequently Asked Questions, Pew Research Center for the People & the Press, http://www.people-press.org/methodology/frequently-asked-questions/#dont-you-have-trouble-getting-people-to-answer-your-polls (accessed May 1, 2013).

1.3 FREQUENCY, FREQUENCY TABLES, AND LEVELS OF MEASUREMENT

Learning Objectives

By the end of this section, the student should be able to:

- Determine the level of measurements for variables.
- Create and interpret frequency tables.

Once you have a set of data, you will need to organize it so that you can analyze how frequently each datum occurs in the set. However, when calculating the frequency, you may need to round your answers so that they are as precise as possible.

Answers and Rounding Off

A simple way to round off answers is to carry your final answer one more decimal place than was present in the original data. Round off only the final answer. Do not round off any intermediate results, if possible. If it becomes necessary to round off intermediate results, carry them to at least twice as many decimal places as the final answer. For example, the average of the three quiz scores four, six, and nine is 6.3, rounded off to the nearest tenth, because the data are whole numbers. Most answers will be rounded off in this manner.

It is not necessary to reduce most fractions in this course. Especially in Probability Topics, the chapter on probability, it is more helpful to leave an answer as an unreduced fraction.

Levels of Measurement

The way a set of data is measured is called its level of measurement. Correct statistical procedures depend on a

researcher being familiar with levels of measurement. Not every statistical operation can be used with every set of data. Data can be classified into four levels of measurement. They are (from lowest to highest level):

- **Nominal scale level**
- **Ordinal scale level**
- **Interval scale level**
- **Ratio scale level**

Data that is measured using a nominal scale is **qualitative**. Categories, colors, names, labels and favorite foods along with yes or no responses are examples of nominal level data. Nominal scale data are not ordered. For example, trying to classify people according to their favorite food does not make any sense. Putting pizza first and sushi second is not meaningful.

Smartphone companies are another example of nominal scale data. Some examples are Sony, Motorola, Nokia, Samsung and Apple. This is just a list and there is no agreed upon order. Some people may favor Apple but that is a matter of opinion. Nominal scale data cannot be used in calculations.

Data that is measured using an ordinal scale is similar to nominal scale data but there is a big difference. The ordinal scale data can be ordered. An example of ordinal scale data is a list of the top five national parks in the United States. The top five national parks in the United States can be ranked from one to five but we cannot measure differences between the data.

Another example of using the ordinal scale is a cruise survey where the responses to questions about the cruise are "excellent," "good," "satisfactory," and "unsatisfactory." These responses are ordered from the most desired response to the least desired. But the differences between two pieces of data cannot be measured. Like the nominal scale data, ordinal scale data cannot be used in calculations.

Data that is measured using the interval scale is similar to ordinal level data because it has a definite ordering but there is a difference between data. The differences between interval scale data can be measured, though the data does not have a starting point.

Temperature scales like Celsius (C) and Fahrenheit (F) are measured by using the interval scale. In both temperature measurements, 40° is equal to 100° minus 60°. Differences make sense. But 0 degrees does not because, in both scales, 0 is not the absolute lowest temperature. Temperatures like -10° F and -15° C exist and are colder than 0.

Interval level data can be used in calculations, but one type of comparison cannot be done. 80° C is not four times as hot as 20° C (nor is 80° F four times as hot as 20° F). There is no meaning to the ratio of 80 to 20 (or four to one).

Data that is measured using the ratio scale takes care of the ratio problem and gives you the most information. Ratio scale data is like interval scale data, but it has a 0 point and ratios can be calculated. For example, four multiple choice statistics final exam scores are 80, 68, 20 and 92 (out of a possible 100 points). The exams are machine-graded.

The data can be put in order from lowest to highest: 20, 68, 80, 92.

The differences between the data have meaning. The score 92 is more than the score 68 by 24 points. Ratios can be calculated. The smallest score is 0. So 80 is four times 20. The score of 80 is four times better than the score of 20.

Frequency

Twenty students were asked how many hours they worked per day. Their responses, in hours, are as follows: 5, 6, 3, 3, 2, 4, 7, 5, 2, 3, 5, 6, 5, 4, 4, 3, 5, 2, 5, 3.

[link] lists the different data values in ascending order and their frequencies.

Frequency Table of Student Work Hours

DATA VALUE	FREQUENCY
2	3
3	5
4	3
5	6
6	2
7	1

A **frequency** is the number of times a value of the data occurs. According to [link], there are three students who work two hours, five students who work three hours, and so on. The sum of the values in the frequency column, 20, represents the total number of students included in the sample.

A **relative frequency** is the ratio (fraction or proportion) of the number of times a value of the data occurs in the set of all outcomes to the total number of outcomes. To find the relative frequencies, divide each frequency by the total number of students in the sample—in this case, 20. Relative frequencies can be written as fractions, percentages, or decimals.

Frequency Table of Student Work Hours with Relative Frequencies

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY
2	3	$\frac{3}{20}$ or 0.15
3	5	$\frac{5}{20}$ or 0.25
4	3	$\frac{3}{20}$ or 0.15
5	6	$\frac{6}{20}$ or 0.30
6	2	$\frac{2}{20}$ or 0.10
7	1	$\frac{1}{20}$ or 0.05

The sum of the values in the relative frequency column of [link] is $\frac{20}{20}$, or 1.

Cumulative relative frequency is the accumulation of the previous relative frequencies. To find the cumulative relative frequencies, add all the previous relative frequencies to the relative frequency for the current row, as shown in [link].

Frequency Table of Student Work Hours with Relative and Cumulative Relative Frequencies

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
2	3	$\frac{3}{20}$ or 0.15	0.15
3	5	$\frac{5}{20}$ or 0.25	$0.15 + 0.25 = 0.40$
4	3	$\frac{3}{20}$ or 0.15	$0.40 + 0.15 = 0.55$
5	6	$\frac{6}{20}$ or 0.30	$0.55 + 0.30 = 0.85$
6	2	$\frac{2}{20}$ or 0.10	$0.85 + 0.10 = 0.95$
7	1	$\frac{1}{20}$ or 0.05	$0.95 + 0.05 = 1.00$

The last entry of the cumulative relative frequency column is one, indicating that one hundred percent of the data has been accumulated.

NOTE

Because of rounding, the relative frequency column may not always sum to one, and the last entry in the cumulative relative frequency column may not be one. However, they each should be close to one.

[link] represents the heights, in inches, of a sample of 100 male semiprofessional soccer players.

Frequency Table of Soccer Player Height

HEIGHTS (INCHES)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
59.95–61.95	5	$\frac{5}{100} = 0.05$	0.05
61.95–63.95	3	$\frac{3}{100} = 0.03$	$0.05 + 0.03 = 0.08$
63.95–65.95	15	$\frac{15}{100} = 0.15$	$0.08 + 0.15 = 0.23$
65.95–67.95	40	$\frac{40}{100} = 0.40$	$0.23 + 0.40 = 0.63$
67.95–69.95	17	$\frac{17}{100} = 0.17$	$0.63 + 0.17 = 0.80$
69.95–71.95	12	$\frac{12}{100} = 0.12$	$0.80 + 0.12 = 0.92$
71.95–73.95	7	$\frac{7}{100} = 0.07$	$0.92 + 0.07 = 0.99$
73.95–75.95	1	$\frac{1}{100} = 0.01$	$0.99 + 0.01 = 1.00$
	Total = 100	**Total = 1.00**	

The data in this table have been **grouped** into the following intervals:

- 59.95 to 61.95 inches
- 61.95 to 63.95 inches
- 63.95 to 65.95 inches
- 65.95 to 67.95 inches
- 67.95 to 69.95 inches
- 69.95 to 71.95 inches
- 71.95 to 73.95 inches
- 73.95 to 75.95 inches

NOTE

This example is used again in Descriptive Statistics, where the method used to compute the intervals will be explained.

In this sample, there are **five** players whose heights fall within the interval 59.95–61.95 inches, **three** players whose heights fall within the interval 61.95–63.95 inches, **15** players whose heights fall within the interval 63.95–65.95 inches, **40** players whose heights fall within the interval 65.95–67.95 inches, **17** players whose heights fall within the interval 67.95–69.95 inches, **12** players whose heights fall within the interval 69.95–71.95, **seven** players whose heights fall within the interval 71.95–73.95, and **one** player whose heights fall within the interval 73.95–75.95. All heights fall between the endpoints of an interval and not at the endpoints.

From [link], find the percentage of heights that are less than 65.95 inches.

Solution

If you look at the first, second, and third rows, the heights are all less than 65.95 inches. There are 5 + 3 + 15 = 23 players whose heights are less than 65.95 inches. The percentage of heights less than 65.95 inches is then $\frac{23}{100}$ or 23%. This percentage is the cumulative relative frequency entry in the third row.

TRY IT

[link] shows the amount, in inches, of annual rainfall in a sample of towns.

Rainfall (Inches)	Frequency	Relative Frequency	Cumulative Relative Frequency
2.95–4.97	6	$\frac{6}{50} = 0.12$	0.12
4.97–6.99	7	$\frac{7}{50} = 0.14$	$0.12 + 0.14 = 0.26$
6.99–9.01	15	$\frac{15}{50} = 0.30$	$0.26 + 0.30 = 0.56$
9.01–11.03	8	$\frac{8}{50} = 0.16$	$0.56 + 0.16 = 0.72$
11.03–13.05	9	$\frac{9}{50} = 0.18$	$0.72 + 0.18 = 0.90$
13.05–15.07	5	$\frac{5}{50} = 0.10$	$0.90 + 0.10 = 1.00$
	Total = 50	Total = 1.00	

From [link], find the percentage of rainfall that is less than 9.01 inches.
Solution

0.56 or 56%

From [link], find the percentage of heights that fall between 61.95 and 65.95 inches.

Solution
Add the relative frequencies in the second and third rows: $0.03 + 0.15 = 0.18$ or 18%.

TRY IT

From [link], find the percentage of rainfall that is between 6.99 and 13.05 inches.
Solution

$0.30 + 0.16 + 0.18 = 0.64$ or 64%

Use the heights of the 100 male semiprofessional soccer players in [link]. Fill in the blanks and check your answers.

1. The percentage of heights that are from 67.95 to 71.95 inches is: ____.
2. The percentage of heights that are from 67.95 to 73.95 inches is: ____.
3. The percentage of heights that are more than 65.95 inches is: ____.
4. The number of players in the sample who are between 61.95 and 71.95 inches tall is: ____.
5. What kind of data are the heights?
6. Describe how you could gather this data (the heights) so that the data are characteristic of all male semiprofessional soccer players.

Remember, you **count frequencies**. To find the relative frequency, divide the frequency by the total number of data values. To find the cumulative relative frequency, add all of the previous relative frequencies to the relative frequency for the current row.

Solution

1. 29%
2. 36%
3. 77%
4. 87
5. quantitative continuous
6. get rosters from each team and choose a simple random sample from each

TRY IT

From [link], find the number of towns that have rainfall between 2.95 and 9.01 inches.

Solution

$6 + 7 + 15 = 28 \text{towns}$

Nineteen people were asked how many miles, to the nearest mile, they commute to work each day. The data are as follows:
2; 5; 7; 3; 2; 10; 18; 15; 20; 7; 10; 18; 5; 12; 13; 12; 4; 5; 10. [link] was produced:

Frequency of Commuting Distances

DATA	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
3	3	$\frac{3}{19}$	0.1579
4	1	$\frac{1}{19}$	0.2105
5	3	$\frac{3}{19}$	0.1579
7	2	$\frac{2}{19}$	0.2632
10	3	$\frac{4}{19}$	0.4737
12	2	$\frac{2}{19}$	0.7895
13	1	$\frac{1}{19}$	0.8421
15	1	$\frac{1}{19}$	0.8948
18	1	$\frac{1}{19}$	0.9474
20	1	$\frac{1}{19}$	1.0000

1. Is the table correct? If it is not correct, what is wrong?
2. True or False: Three percent of the people surveyed commute three miles. If the statement is not correct, what should it be? If the table is incorrect, make the corrections.
3. What fraction of the people surveyed commute five or seven miles?

4. What fraction of the people surveyed commute 12 miles or more? Less than 12 miles? Between five and 13 miles (not including five and 13 miles)?

Solution

1. No. The frequency column sums to 18, not 19. Not all cumulative relative frequencies are correct.
2. False. The frequency for three miles should be one; for two miles (left out), two. The cumulative relative frequency column should read: 0.1052, 0.1579, 0.2105, 0.3684, 0.4737, 0.6316, 0.7368, 0.7895, 0.8421, 0.9474, 1.0000.
3. $\frac{5}{19}$
4. $\frac{7}{19}, \frac{12}{19}, \frac{7}{19}$

TRY IT

[link] represents the amount, in inches, of annual rainfall in a sample of towns. What fraction of towns surveyed get between 11.03 and 13.05 inches of rainfall each year?

Solution

$\frac{9}{50}$

[link] contains the total number of deaths worldwide as a result of earthquakes for the period from 2000 to 2012.

Year	Total Number of Deaths
2000	231
2001	21,357
2002	11,685
2003	33,819
2004	228,802
2005	88,003
2006	6,605
2007	712
2008	88,011
2009	1,790
2010	320,120
2011	21,953
2012	768
Total	823,856

Answer the following questions.

1. What is the frequency of deaths measured from 2006 through 2009?
2. What percentage of deaths occurred after 2009?
3. What is the relative frequency of deaths that occurred in 2003 or earlier?
4. What is the percentage of deaths that occurred in 2004?
5. What kind of data are the numbers of deaths?
6. The Richter scale is used to quantify the energy produced by an earthquake. Examples of Richter scale numbers are 2.3, 4.0, 6.1, and 7.0. What kind of data are these numbers?

Solution

1. 97,118 (11.8%)
2. 41.6%
3. 67,092/823,856 or 0.081 or 8.1 %
4. 27.8%
5. Quantitative discrete
6. Quantitative continuous

TRY IT

[link] contains the total number of fatal motor vehicle traffic crashes in the United States for the period from 1994 to 2011.

Year	Total Number of Crashes	Year	Total Number of Crashes
1994	36,254	2004	38,444
1995	37,241	2005	39,252
1996	37,494	2006	38,648
1997	37,324	2007	37,435
1998	37,107	2008	34,172
1999	37,140	2009	30,862
2000	37,526	2010	30,296
2001	37,862	2011	29,757
2002	38,491	Total	653,782
2003	38,477		

Answer the following questions.

1. What is the frequency of deaths measured from 2000 through 2004?
2. What percentage of deaths occurred after 2006?
3. What is the relative frequency of deaths that occurred in 2000 or before?
4. What is the percentage of deaths that occurred in 2011?
5. What is the cumulative relative frequency for 2006? Explain what this number tells you about the data.

Solution

1. 190,800 (29.2%)
2. 24.9%
3. 260,086/653,782 or 39.8%
4. 4.6%
5. 75.1% of all fatal traffic crashes for the period from 1994 to 2011 happened from 1994 to 2006.

References

"State & County QuickFacts," U.S. Census Bureau. http://quickfacts.census.gov/qfd/download_data.html (accessed May 1, 2013).

"State & County QuickFacts: Quick, easy access to facts about people, business, and geography," U.S. Census Bureau. http://quickfacts.census.gov/qfd/index.html (accessed May 1, 2013).

"Table 5: Direct hits by mainland United States Hurricanes (1851-2004)," National Hurricane Center, http://www.nhc.noaa.gov/gifs/table5.gif (accessed May 1, 2013).

"Levels of Measurement," http://infinity.cos.edu/faculty/woodbury/stats/tutorial/Data_Levels.htm (accessed May 1, 2013).

Courtney Taylor, "Levels of Measurement," about.com, http://statistics.about.com/od/HelpandTutorials/a/Levels-Of-Measurement.htm (accessed May 1, 2013).

David Lane. "Levels of Measurement," Connexions, http://cnx.org/content/m10809/latest/ (accessed May 1, 2013).

1.4 EXPERIMENTAL DESIGN AND ETHICS

Learning Objectives

By the end of this section, the student should be able to:

- Classify variables used for conducting an experiment.
- Identify the ethical concerns that arise when conducting a study.

Does aspirin reduce the risk of heart attacks? Is one brand of fertilizer more effective at growing roses than another? Is fatigue as dangerous to a driver as the influence of alcohol? Questions like these are answered using randomized experiments. In this module, you will learn important aspects of experimental design. Proper study design ensures the production of reliable, accurate data.

The purpose of an experiment is to investigate the relationship between two variables. When one variable causes change in another, we call the first variable the **explanatory variable**. The affected variable is called the **response variable**. In a randomized experiment, the researcher manipulates values of the explanatory variable and measures the resulting changes in the response variable. The different values of the explanatory variable are called **treatments**. An experimental unit is a single object or individual to be measured.

You want to investigate the effectiveness of vitamin E in preventing disease. You recruit a group of subjects and ask them if they regularly take vitamin E. You notice that the subjects who take vitamin E exhibit better health on average than those who do not. Does this prove that vitamin E is effective in disease prevention? It does not. There are many differences between the two groups compared to vitamin E consumption. People who take vitamin E regularly often take other steps to improve their health: exercise, diet, other vitamin supplements, choosing not to smoke. Any one of these factors could be influencing health. As described, this study does not prove that vitamin E is the key to disease prevention.

Additional variables that can cloud a study are called **lurking variables**. In order to prove that the explanatory variable is causing a change in the response variable, it is necessary to isolate the explanatory variable. The researcher must design her experiment in such a way that there is only one difference between groups being compared: the planned treatments. This is accomplished by the **random assignment** of

experimental units to treatment groups. When subjects are assigned treatments randomly, all of the potential lurking variables are spread equally among the groups. At this point the only difference between groups is the one imposed by the researcher. Different outcomes measured in the response variable, therefore, must be a direct result of the different treatments. In this way, an experiment can prove a cause-and-effect connection between the explanatory and response variables.

The power of suggestion can have an important influence on the outcome of an experiment. Studies have shown that the expectation of the study participant can be as important as the actual medication. In one study of performance-enhancing drugs, researchers noted:

*Results showed that believing one had taken the substance resulted in [*performance*] times almost as fast as those associated with consuming the drug itself. In contrast, taking the drug without knowledge yielded no significant performance increment.*[1]

When participation in a study prompts a physical response from a participant, it is difficult to isolate the effects of the explanatory variable. To counter the power of suggestion, researchers set aside one treatment group as a **control group**. This group is given a placebo treatment–a treatment that cannot influence the response variable. The control group helps researchers balance the effects of being in an experiment with the effects of the active treatments. Of course, if you are participating in a study and you know that you are receiving a pill which contains no actual medication, then the power of suggestion is no longer a factor. **Blinding** in a randomized experiment preserves the power of suggestion. When a person involved in a research study is blinded, he does not know who is receiving the active treatment(s) and who is receiving the **placebo treatment**. A **double-blind experiment** is one in which both the subjects and the researchers involved with the subjects are blinded.

Example

The Smell & Taste Treatment and Research Foundation conducted a study to investigate whether smell can affect learning. Subjects completed mazes multiple times while wearing masks. They completed the pencil and paper mazes three times wearing floral-scented masks, and three times with unscented masks. Participants were assigned at random to wear the floral mask during the first three trials or during the last three trials. For each trial, researchers recorded the time it took to complete the maze and the subject's impression of the mask's scent: positive, negative, or neutral.

1. Describe the explanatory and response variables in this study.

2. What are the treatments?
3. Identify any lurking variables that could interfere with this study.
4. Is it possible to use blinding in this study?

Solution

1. The explanatory variable is scent, and the response variable is the time it takes to complete the maze.
2. There are two treatments: a floral-scented mask and an unscented mask.
3. All subjects experienced both treatments. The order of treatments was randomly assigned so there were no differences between the treatment groups. Random assignment eliminates the problem of lurking variables.
4. Subjects will clearly know whether they can smell flowers or not, so subjects cannot be blinded in this study. Researchers timing the mazes can be blinded, though. The researcher who is observing a subject will not know which mask is being worn.

A researcher wants to study the effects of birth order on personality. Explain why this study could not be conducted as a randomized experiment. What is the main problem in a study that cannot be designed as a randomized experiment?

Solution
The explanatory variable is birth order. You cannot randomly assign a person's birth order. Random assignment eliminates the impact of lurking variables. When you cannot assign subjects to treatment groups at random, there will be differences between the groups other than the explanatory variable.

TRY IT

You are concerned about the effects of texting on driving performance. Design a study to test the response time of drivers while texting and while driving only. How many seconds does it take for a driver to respond when a leading car hits the brakes?

1. Describe the explanatory and response variables in the study.
2. What are the treatments?
3. What should you consider when selecting participants?

4. Your research partner wants to divide participants randomly into two groups: one to drive without distraction and one to text and drive simultaneously. Is this a good idea? Why or why not?
5. Identify any lurking variables that could interfere with this study.
6. How can blinding be used in this study?

Solution

1. Explanatory: presence of distraction from texting; response: response time measured in seconds
2. Driving without distraction and driving while texting
3. Answers will vary. Possible responses: Do participants regularly send and receive text messages? How long has the subject been driving? What is the age of the participants? Do participants have similar texting and driving experience?
4. This is not a good plan because it compares drivers with different abilities. It would be better to assign both treatments to each participant in random order.
5. Possible responses include: texting ability, driving experience, type of phone.
6. The researchers observing the trials and recording response time could be blinded to the treatment being applied.

Ethics

The widespread misuse and misrepresentation of statistical information often gives the field a bad name. Some say that "numbers don't lie," but the people who use numbers to support their claims often do.

A recent investigation of famous social psychologist, Diederik Stapel, has led to the retraction of his articles from some of the world's top journals including *Journal of Experimental Social Psychology, Social Psychology, Basic and Applied Social Psychology, British Journal of Social Psychology,* and the magazine *Science*. Diederik Stapel is a former professor at Tilburg University in the Netherlands. Over the past two years, an extensive investigation involving three universities where Stapel has worked concluded that the psychologist is guilty of fraud on a colossal scale. Falsified data taints over 55 papers he authored and 10 Ph.D. dissertations that he supervised.

Stapel did not deny that his deceit was driven by ambition. But it was more complicated than that, he told me. He insisted that he loved social psychology but had been frustrated by the messiness of experimental data, which rarely led to clear conclusions. His lifelong obsession with elegance and order, he said, led him to concoct sexy results that journals found attractive. "It was a quest for aesthetics, for beauty—instead of the truth," he said. He

described his behavior as an addiction that drove him to carry out acts of increasingly daring fraud, like a junkie seeking a bigger and better high.[2]

The committee investigating Stapel concluded that he is guilty of several practices including:

- creating datasets, which largely confirmed the prior expectations,
- altering data in existing datasets,
- changing measuring instruments without reporting the change, and
- misrepresenting the number of experimental subjects.

Clearly, it is never acceptable to falsify data the way this researcher did. Sometimes, however, violations of ethics are not as easy to spot.

Researchers have a responsibility to verify that proper methods are being followed. The report describing the investigation of Stapel's fraud states that, "statistical flaws frequently revealed a lack of familiarity with elementary statistics."[3] Many of Stapel's co-authors should have spotted irregularities in his data. Unfortunately, they did not know very much about statistical analysis, and they simply trusted that he was collecting and reporting data properly.

Many types of statistical fraud are difficult to spot. Some researchers simply stop collecting data once they have just enough to prove what they had hoped to prove. They don't want to take the chance that a more extensive study would complicate their lives by producing data contradicting their hypothesis.

Professional organizations, like the American Statistical Association, clearly define expectations for researchers. There are even laws in the federal code about the use of research data.

When a statistical study uses human participants, as in medical studies, both ethics and the law dictate that researchers should be mindful of the safety of their research subjects. The U.S. Department of Health and Human Services oversees federal regulations of research studies with the aim of protecting participants. When a university or other research institution engages in research, it must ensure the safety of all human subjects. For this reason, research institutions establish oversight committees known as **Institutional Review Boards (IRB)**. All planned studies must be approved in advance by the IRB. Key protections that are mandated by law include the following:

- Risks to participants must be minimized and reasonable with respect to projected benefits.
- Participants must give **informed consent**. This means that the risks of participation must be clearly explained to the subjects of the study. Subjects must consent in writing, and researchers are required to keep documentation of their consent.
- Data collected from individuals must be guarded carefully to protect their privacy.

These ideas may seem fundamental, but they can be very difficult to verify in practice. Is removing a participant's name from the data record sufficient to protect privacy? Perhaps the person's identity could be

discovered from the data that remains. What happens if the study does not proceed as planned and risks arise that were not anticipated? When is informed consent really necessary? Suppose your doctor wants a blood sample to check your cholesterol level. Once the sample has been tested, you expect the lab to dispose of the remaining blood. At that point the blood becomes biological waste. Does a researcher have the right to take it for use in a study?

It is important that students of statistics take time to consider the ethical questions that arise in statistical studies. How prevalent is fraud in statistical studies? You might be surprised—and disappointed. There is a website (www.retractionwatch.com) dedicated to cataloging retractions of study articles that have been proven fraudulent. A quick glance will show that the misuse of statistics is a bigger problem than most people realize.

Vigilance against fraud requires knowledge. Learning the basic theory of statistics will empower you to analyze statistical studies critically.

Describe the unethical behavior in each example and describe how it could impact the reliability of the resulting data. Explain how the problem should be corrected.

A researcher is collecting data in a community.

1. She selects a block where she is comfortable walking because she knows many of the people living on the street.
2. No one seems to be home at four houses on her route. She does not record the addresses and does not return at a later time to try to find residents at home.
3. She skips four houses on her route because she is running late for an appointment. When she gets home, she fills in the forms by selecting random answers from other residents in the neighborhood.

Solution

1. By selecting a convenient sample, the researcher is intentionally selecting a sample that could be biased. Claiming that this sample represents the community is misleading. The researcher needs to select areas in the community at random.
2. Intentionally omitting relevant data will create bias in the sample. Suppose the researcher is gathering information about jobs and child care. By ignoring people who are not home, she may be missing data from working families that are relevant to her study. She needs to make every effort to interview all members of the target sample.
3. It is never acceptable to fake data. Even though the responses she uses are “real” responses provided by other participants, the duplication is fraudulent and can create bias in the data. She needs to work diligently to interview everyone on her route.

TRY IT

Describe the unethical behavior, if any, in each example and describe how it could impact the reliability of the resulting data. Explain how the problem should be corrected.

A study is commissioned to determine the favorite brand of fruit juice among teens in California.

1. The survey is commissioned by the seller of a popular brand of apple juice.
2. There are only two types of juice included in the study: apple juice and cranberry juice.
3. Researchers allow participants to see the brand of juice as samples are poured for a taste test.
4. Twenty-five percent of participants prefer Brand X, 33% prefer Brand Y and 42% have no preference between the two brands. Brand X references the study in a commercial saying "Most teens like Brand X as much as or more than Brand Y."

1. This is not necessarily a problem. The study should be monitored carefully, however, to ensure that the company is not pressuring researchers to return biased results.
2. If the researchers truly want to determine the favorite brand of juice, then researchers should ask teens to compare different brands of the same type of juice. Choosing a sweet juice to compare against a sharp-flavored juice will not lead to an accurate comparison of brand quality.
3. Participants could be biased by the knowledge. The results may be different from those obtained in a blind taste test.
4. The commercial tells the truth, but not the whole truth. It leads consumers to believe that Brand X was preferred by more participants than Brand Y while the opposite is true.

References

"Vitamin E and Health," Nutrition Source, Harvard School of Public Health, http://www.hsph.harvard.edu/nutritionsource/vitamin-e/ (accessed May 1, 2013).

Stan Reents. "Don't Underestimate the Power of Suggestion," athleteinme.com, http://www.athleteinme.com/ArticleView.aspx?id=1053 (accessed May 1, 2013).

Ankita Mehta. "Daily Dose of Aspirin Helps Reduce Heart Attacks: Study," International Business Times, July 21, 2011. Also available online at http://www.ibtimes.com/daily-dose-aspirin-helps-reduce-heart-attacks-study-300443 (accessed May 1, 2013).

The Data and Story Library, http://lib.stat.cmu.edu/DASL/Stories/ScentsandLearning.html (accessed May 1, 2013).

M.L. Jacskon et al., "Cognitive Components of Simulated Driving Performance: Sleep Loss Effect and

Predictors," Accident Analysis and Prevention Journal, Jan no. 50 (2013), http://www.ncbi.nlm.nih.gov/pubmed/22721550 (accessed May 1, 2013).

"Earthquake Information by Year," U.S. Geological Survey. http://earthquake.usgs.gov/earthquakes/eqarchives/year/ (accessed May 1, 2013).

"Fatality Analysis Report Systems (FARS) Encyclopedia," National Highway Traffic and Safety Administration. http://www-fars.nhtsa.dot.gov/Main/index.aspx (accessed May 1, 2013).

Data from www.businessweek.com (accessed May 1, 2013).

Data from www.forbes.com (accessed May 1, 2013).

"America's Best Small Companies," http://www.forbes.com/best-small-companies/list/ (accessed May 1, 2013).

U.S. Department of Health and Human Services, Code of Federal Regulations Title 45 Public Welfare Department of Health and Human Services Part 46 Protection of Human Subjects revised January 15, 2009. Section 46.111:Criteria for IRB Approval of Research.

"April 2013 Air Travel Consumer Report," U.S. Department of Transportation, April 11 (2013), http://www.dot.gov/airconsumer/april-2013-air-travel-consumer-report (accessed May 1, 2013).

Lori Alden, "Statistics can be Misleading," econoclass.com, http://www.econoclass.com/misleadingstats.html (accessed May 1, 2013).

Maria de los A. Medina, "Ethics in Statistics," Based on "Building an Ethics Module for Business, Science, and Engineering Students" by Jose A. Cruz-Cruz and William Frey, Connexions, http://cnx.org/content/m15555/latest/ (accessed May 1, 2013).

Footnotes

1. 1 McClung, M. Collins, D. "Because I know it will!": placebo effects of an ergogenic aid on athletic performance. Journal of Sport & Exercise Psychology. 2007 Jun. 29(3):382-94. Web. April 30, 2013.
2. 2
   Yudhijit Bhattacharjee, "The Mind of a Con Man," Magazine, New York Times, April 26, 2013. Available online at: http://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html?src=dayp&_r=2& (accessed May 1, 2013).
3. 3 "Flawed Science: The Fraudulent Research Practices of Social Psychologist Diederik Stapel," Tilburg University, November 28, 2012, http://www.tilburguniversity.edu/upload/064a10cd-bce5-4385-b9ff-05b840caeae6_120695_Rapp_nov_2012_UK_web.pdf (accessed May 1, 2013).
4. 4 Andrew Gelman, "Open Data and Open Methods," Ethics and Statistics, http://www.stat.columbia.edu/~gelman/research/published/ChanceEthics1.pdf (accessed May 1, 2013).

CHAPTER 1 REVIEW

Chapter Review from 1.1

The mathematical theory of statistics is easier to learn when you know the language. This module presents important terms that will be used throughout the text.

Chapter Review from 1.2

Data are individual items of information that come from a population or sample. Data may be classified as qualitative, quantitative continuous, or quantitative discrete.

Because it is not practical to measure the entire population in a study, researchers use samples to represent the population. A random sample is a representative group from the population chosen by using a method that gives each individual in the population an equal chance of being included in the sample. Random sampling methods include simple random sampling, stratified sampling, cluster sampling, and systematic sampling. Convenience sampling is a nonrandom method of choosing a sample that often produces biased data.

Samples that contain different individuals result in different data. This is true even when the samples are well-chosen and representative of the population. When properly selected, larger samples model the population more closely than smaller samples. There are many different potential problems that can affect the reliability of a sample. Statistical data needs to be critically analyzed, not simply accepted.

Chapter Review from 1.3

Some calculations generate numbers that are artificially precise. It is not necessary to report a value to eight decimal places when the measures that generated that value were only accurate to the nearest tenth. Round off your final answer to one more decimal place than was present in the original data. This means that if you have data measured to the nearest tenth of a unit, report the final statistic to the nearest hundredth.

In addition to rounding your answers, you can measure your data using the following four levels of measurement.

- **Nominal scale level:** data that cannot be ordered nor can it be used in calculations
- **Ordinal scale level:** data that can be ordered; the differences cannot be measured
- **Interval scale level:** data with a definite ordering but no starting point; the differences can be measured, but there is no such thing as a ratio.
- **Ratio scale level:** data with a starting point that can be ordered; the differences have meaning and ratios can be calculated.

When organizing data, it is important to know how many times a value appears. How many statistics students study five hours or more for an exam? What percent of families on our block own two pets? Frequency, relative frequency, and cumulative relative frequency are measures that answer questions like these.

Exercises

What type of measure scale is being used? Nominal, ordinal, interval or ratio.

1. High school soccer players classified by their athletic ability: Superior, Average, Above average
2. Baking temperatures for various main dishes: 350, 400, 325, 250, 300
3. The colors of crayons in a 24-crayon box
4. Social security numbers
5. Incomes measured in dollars
6. A satisfaction survey of a social website by number: 1 = very satisfied, 2 = somewhat satisfied, 3 = not satisfied
7. Political outlook: extreme left, left-of-center, right-of-center, extreme right
8. Time of day on an analog watch
9. The distance in miles to the closest grocery store
10. The dates 1066, 1492, 1644, 1947, and 1944
11. The heights of 21–65 year-old women
12. Common letter grades: A, B, C, D, and F

Solution

1. ordinal
2. interval

3. nominal
4. nominal
5. ratio
6. ordinal
7. nominal
8. interval
9. ratio
10. interval
11. ratio
12. ordinal

Chapter Review from 1.4

A poorly designed study will not produce reliable data. There are certain key components that must be included in every experiment. To eliminate lurking variables, subjects must be assigned randomly to different treatment groups. One of the groups must act as a control group, demonstrating what happens when the active treatment is not applied. Participants in the control group receive a placebo treatment that looks exactly like the active treatments but cannot influence the response variable. To preserve the integrity of the placebo, both researchers and subjects may be blinded. When a study is designed properly, the only difference between treatment groups is the one imposed by the researcher. Therefore, when groups respond differently to different treatments, the difference must be due to the influence of the explanatory variable.

"An ethics problem arises when you are considering an action that benefits you or some cause you support, hurts or reduces benefits to others, and violates some rule."[4] Ethical violations in statistics are not always easy to spot. Professional associations and federal agencies post guidelines for proper conduct. It is important that you learn basic statistical procedures so that you can recognize proper data analysis.

Exercises

Design an experiment. Identify the explanatory and response variables. Describe the population being studied and the experimental units. Explain the treatments that will be used and how they

will be assigned to the experimental units. Describe how blinding and placebos may be used to counter the power of suggestion.

Solution

Answers will vary.

Exercises

Discuss potential violations of the rule requiring informed consent.

1. Inmates in a correctional facility are offered good behavior credit in return for participation in a study.
2. A research study is designed to investigate a new children's allergy medication.
3. Participants in a study are told that the new medication being tested is highly promising, but they are not told that only a small portion of participants will receive the new medication. Others will receive placebo treatments and traditional treatments.

Solution

1. Inmates may not feel comfortable refusing participation, or may feel obligated to take advantage of the promised benefits. They may not feel truly free to refuse participation.
2. Parents can provide consent on behalf of their children, but children are not competent to provide consent for themselves.
3. All risks and benefits must be clearly outlined. Study participants must be informed of relevant aspects of the study in order to give appropriate consent.

CHAPTER 1 PRACTICE

Practice from 1.2

Exercise

Use the following information to answer the next five exercises.

Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment program. Suppose that a new AIDS antibody drug is currently under study. It is given to patients once the AIDS symptoms have revealed themselves. Of interest is the average (mean) length of time in months patients live once they start the treatment. Two researchers each follow a different set of 40 patients with AIDS from the start of treatment until their deaths. The following data (in months) are collected.

Researcher A: 3; 4; 11; 15; 16; 17; 22; 44; 37; 16; 14; 24; 25; 15; 26; 27; 33; 29; 35; 44; 13; 21; 22; 10; 12; 8; 40; 32; 26; 27; 31; 34; 29; 17; 8; 24; 18; 47; 33; 34

Researcher B: 3; 14; 11; 5; 16; 17; 28; 41; 31; 18; 14; 14; 26; 25; 21; 22; 31; 2; 35; 44; 23; 21; 21; 16; 12; 18; 41; 22; 16; 25; 33; 34; 29; 13; 18; 24; 23; 42; 33; 29

Determine what the following key terms refer to in the example for Researcher A:

- population
- sample
- parameter
- statistic
- variable

Solution

- The population is AIDS patients.
- The sample is AIDS patients sampled from researcher A and researcher B.
- The parameter is the average length of time (in months) AIDS patients live after

treatment.

- The statistic is the average length of time (in months) AIDS patients from the sample live after treatment.
- The variable is X = the length of time (in months) AIDS patients live after treatment.

Exercises

"Number of times per week" is what type of data?

a. qualitative b. quantitative discrete c. quantitative continuous

Solution

b

Exercises

Use the following information to answer the next four exercises.

A study was done to determine the age, number of times per week, and the duration (amount of time) of residents using a local park in San Antonio, Texas. The first house in the neighborhood around the park was selected randomly, and then the resident of every eighth house in the neighborhood around the park was interviewed.

The sampling method was

a. simple random b. systematic c. stratified d. cluster

Solution

b

"Duration (amount of time)" is what type of data?

a. qualitative b. quantitative discrete c. quantitative continuous

Solution
c

The colors of the houses around the park are what kind of data?

a. qualitative b. quantitative discrete c. quantitative continuous

Solution
a

The population is _____

Solution
the houses in the neighborhood around a local park in San Antonio, Texas for the given example.

Exercises

[link] contains the total number of deaths worldwide as a result of earthquakes from 2000 to 2012.

Year	Total Number of Deaths
2000	231
2001	21,357
2002	11,685
2003	33,819
2004	228,802
2005	88,003
2006	6,605
2007	712
2008	88,011
2009	1,790
2010	320,120
2011	21,953
2012	768
Total	**823,856**

Use [link] to answer the following questions.

1. What is the proportion of deaths between 2007 and 2012 (not including the years 2007 and 2012)?
2. What percent of deaths occurred before 2001?
3. What is the percentage of deaths that occurred in 2003 or after 2010?
4. What is the fraction of deaths that happened before 2012?
5. What kind of data is the number of deaths?
6. Earthquakes are quantified according to the amount of energy they produce (examples are 2.1, 5.0, 6.7). What type of data is that?
7. What contributed to the large number of deaths in 2010? In 2004? Explain.

Solution

1. 0.5242
2. 0.03%
3. 6.86%
4. $\frac{823,088}{823,856}$

5. quantitative discrete
6. quantitative continuous
7. In both years, underwater earthquakes produced massive tsunamis.

Exercises

For the following four exercises, determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

A group of test subjects is divided into twelve groups; then four of the groups are chosen at random.

Solution
cluster

A market researcher polls every tenth person who walks into a store.

Solution
systematic

The first 50 people who walk into a sporting event are polled on their television preferences.

Solution
convenience

A computer generates 100 random numbers, and 100 people whose names correspond with the numbers on the list are chosen.

Solution
simple random

Exercises

Use the following information to answer the next seven exercises.

Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment program. Suppose that a new AIDS antibody drug is currently under study. It is given to patients once the AIDS symptoms have revealed themselves. Of interest is the average (mean) length of time in months patients live once starting the treatment. Two researchers each follow a different set of 40 AIDS patients from the start of treatment until their deaths. The following data (in months) are collected.

Researcher A: 3; 4; 11; 15; 16; 17; 22; 44; 37; 16; 14; 24; 25; 15; 26; 27; 33; 29; 35; 44; 13; 21; 22; 10; 12; 8; 40; 32; 26; 27; 31; 34; 29; 17; 8; 24; 18; 47; 33; 34

Researcher B: 3; 14; 11; 5; 16; 17; 28; 41; 31; 18; 14; 14; 26; 25; 21; 22; 31; 2; 35; 44; 23; 21; 21; 16; 12; 18; 41; 22; 16; 25; 33; 34; 29; 13; 18; 24; 23; 42; 33; 29

Complete the tables using the data provided:

Researcher A

Survival Length (in months)	Frequency	Relative Frequency	Cumulative Relative Frequency
0.5–6.5			
6.5–12.5			
12.5–18.5			
18.5–24.5			
24.5–30.5			
30.5–36.5			
36.5–42.5			
42.5–48.5			

Researcher B

Survival Length (in months)	Frequency	Relative Frequency	Cumulative Relative Frequency
0.5–6.5			
6.5–12.5			
12.5–18.5			
18.5–24.5			
24.5–30.5			
30.5–36.5			
36.5-45.5			

Determine what the key term data refers to in the above example for Researcher A.

Solution
values for X, such as 3, 4, 11, and so on

List two reasons why the data may differ.

Solution
Answers will vary. Sample answer: One reason may be the average age of the individuals in the two samples. Or, perhaps the drug affects men and women differently. If the ratio of men and women aren't the same in both sample groups, then the data would differ.

Can you tell if one researcher is correct and the other one is incorrect? Why?

Solution
No, we do not have enough information to make such a claim.

Would you expect the data to be identical? Why or why not?

Solution
Since the treatment is not the same the data might be different unless neither treatment has an effect.

How might the researchers gather random data?

Solution
Take a simple random sample from each group. One way is by assigning a number to each patient and using a random number generator to randomly select patients.

Suppose that the first researcher conducted his survey by randomly choosing one state in the

nation and then randomly picking 40 patients from that state. What sampling method would that researcher have used?

Solution
He has used a simple random sample method.

Suppose that the second researcher conducted his survey by choosing 40 patients he knew. What sampling method would that researcher have used? What concerns would you have about this data set, based upon the data collection method?

Solution
This would be convenience sampling and is not random.

Exercises

Use the following data to answer the next five exercises.

Two researchers are gathering data on hours of video games played by school-aged children and young adults. They each randomly sample different groups of 150 students from the same school. They collect the following data.

Researcher A

Hours Played per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
0–2	26	0.17	0.17
2–4	30	0.20	0.37
4–6	49	0.33	0.70
6–8	25	0.17	0.87
8–10	12	0.08	0.95
10–12	8	0.05	1

Researcher B

Hours Played per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
0–2	48	0.32	0.32
2–4	51	0.34	0.66
4–6	24	0.16	0.82
6–8	12	0.08	0.90
8–10	11	0.07	0.97
10–12	4	0.03	1

Give a reason why the data may differ.

Solution
The researchers are studying different groups, so there will be some variation in the data.

Would the sample size be large enough if the population is the students in the school?

Solution
Yes, the sample size of 150 would be large enough to reflect a population of one school.

Would the sample size be large enough if the population is school-aged children and young adults in the United States?

Solution
There are many school-aged children and young adults in the United States, and the study was done at only one school, so the sample size is not large enough to reflect the population.

Researcher A concludes that most students play video games between four and six hours each week. Researcher B concludes that most students play video games between two and four hours each week. Who is correct?

Solution
Even though the specific data support each researcher's conclusions, the different results suggest that more data need to be collected before the researchers can reach a conclusion.

As part of a way to reward students for participating in the survey, the researchers gave each student a gift card to a video game store. Would this affect the data if students knew about the award before the study?

Solution

Yes, people who play games more might be more likely to participate, since they would want the gift card more than a student who does not play video games. This would leave out many students who do not play games at all and skew the data.

Exercises

Use the following data to answer the next five exercises.

A pair of studies was performed to measure the effectiveness of a new software program designed to help stroke patients regain their problem-solving skills. Patients were asked to use the software program twice a day, once in the morning and once in the evening. The studies observed 200 stroke patients recovering over a period of several weeks. The first study collected the data in [link]. The second study collected the data in [link].

Group	Showed improvement	No improvement	Deterioration
Used program	142	43	15
Did not use program	72	110	18

Group	Showed improvement	No improvement	Deterioration
Used program	105	74	19
Did not use program	89	99	12

Given what you know, which study is correct?

Solution

There is not enough information given to judge if either one is correct or incorrect.

The first study was performed by the company that designed the software program. The second study was performed by the American Medical Association. Which study is more reliable?

Solution

The second study is more reliable, because the company would be interested in showing results that favored a higher rate of improvement from patients using their software. The data may be

skewed; however, the American Medical Association is not concerned with the success of the software and so should be objective.

Both groups that performed the study concluded that the software works. Is this accurate?

Solution
The software program seems to work because the second study shows that more patients improve while using the software than not. Even though the difference is not as large as that in the first study, the results from the second study are likely more reliable and still show improvement.

The company takes the two studies as proof that their software causes mental improvement in stroke patients. Is this a fair statement?

Solution
No, the data suggest the two are correlated, but more studies need to be done to prove that using the software causes improvement in stroke patients.

Patients who used the software were also a part of an exercise program whereas patients who did not use the software were not. Does this change the validity of the conclusions from [link]?

Solution
Yes, because we cannot tell if the improvement was due to the software or the exercise; the data is confounded, and a reliable conclusion cannot be drawn. New studies should be performed.

Exercises

Is a sample size of 1,000 a reliable measure for a population of 5,000?

Solution
Yes, 1,000 represents 20% of the population and should be representative, if the population of the sample is chosen at random.

Is a sample of 500 volunteers a reliable measure for a population of 2,500?

Solution

No, even though the sample is large enough, the fact that the sample consists of volunteers makes it a self-selected sample, which is not reliable.

Exercises

A question on a survey reads: "Do you prefer the delicious taste of Brand X or the taste of Brand Y?" Is this a fair question?

Solution
No, the question is creating undue influence by adding the word "delicious" to describe Brand X. The wording may influence responses.

Exercises

Is a sample size of two representative of a population of five?

Solution
No, even though the sample is a large portion of the population, two responses are not enough to justify any conclusions. Because the population is so small, it would be better to include everyone in the population to get the most accurate data.

Is it possible for two experiments to be well run with similar sample sizes to get different data?

Solution
Yes, there will most likely be a degree of variation between any two studies, even if they are set up and run the same way. Each study may be affected differently by unknown factors such as location, mood of the subjects, or time of year.

CHAPTER 1 HOMEWORK

HOMEWORK from 1.2

Exercises

For each of the following eight exercises, identify: a. the population, b. the sample, c. the parameter, d. the statistic, e. the variable, and f. the data. Give examples where appropriate.

1. A fitness center is interested in the mean amount of time a client exercises in the center each week.

Solution
a. The population is all of the clients of the fitness center.
b. A sample of the clients that use the fitness center for a given week.
c. The average amount of time that all clients exercise in one week.
d. The average amount of time that a sample of clients exercises in one week.
e. The amount of time that a client exercises in one week.
f. values for X, such as 2, 5, and 7.5

2. Ski resorts are interested in the mean age that children take their first ski and snowboard lessons. They need this information to plan their ski classes optimally.

Solution
a. all children who take ski or snowboard lessons
b. a group of these children
c. the population mean age of children who take their first snowboard lesson
d. the sample mean age of children who take their first snowboard lesson
e. X = the age of one child who takes his or her first ski or snowboard lesson
f. values for X, such as 3, 7, and so on

3. A cardiologist is interested in the mean recovery period of her patients who have had heart attacks.

Solution
a. the cardiologist's patients
b. a group of the cardiologist's patients
c. the mean recovery period of all of the cardiologist's patients
d. the mean recovery period of the group of the cardiologist's patients
e. X = the recovery period of one patient
f. values for X, such as 10 days, 14 days, 20 days, and so on

4. Insurance companies are interested in the mean health costs each year of their clients, so that they can determine the costs of health insurance.

Solution
a. the clients of the insurance companies
b. a group of the clients
c. the mean health costs of the clients
d. the mean health costs of the sample
e. *X* = the health costs of one client
f. values for *X*, such as 34, 9, 82, and so on

5. A politician is interested in the proportion of voters in his district who think he is doing a good job.

Solution
a. all voters in the politician's district
b. a random selection of voters in the politician's district
c. the proportion of voters in this district who think this politician is doing a good job
d. the proportion of voters in this district who think this politician is doing a good job in the sample
e. X = the number of voters in the district who think this politician is doing a good job
f. Yes, he is doing a good job. No, he is not doing a good job.

6. A marriage counselor is interested in the proportion of clients she counsels who stay married.

Solution
a. all the clients of this counselor
b. a group of clients of this marriage counselor
c. the proportion of all her clients who stay married
d. the proportion of the sample of the counselor's clients who stay married
e. X = the number of couples who stay married
f. yes, no

7. Political pollsters may be interested in the proportion of people who will vote for a particular cause.

Solution
a. all voters (in a certain geographic area)
b. a random selection of all the voters
c. the proportion of voters who are interested in this particular cause
d. the proportion of voters who are interested in this particular cause in the sample
e. X = the number of voters who are interested in this particular cause
f. yes, no

8. A marketing company is interested in the proportion of people who will buy a particular product.

Solution
a. all people (maybe in a certain geographic area, such as the United States)
b. a group of the people
c. the proportion of all people who will buy the product
d. the proportion of the sample who will buy the product
e. X = the number of people who will buy it
f. buy, not buy

Exercises

Use the following information to answer the next three exercises: A Lake Tahoe Community College instructor is interested in the mean number of days Lake Tahoe Community College math students are absent from class during a quarter.

1. What is the population she is interested in?

    a. all Lake Tahoe Community College students
    b. all Lake Tahoe Community College English students
    c. all Lake Tahoe Community College students in her classes
    d. all Lake Tahoe Community College math students

Solution
d

2. Consider the following:

X = number of days a Lake Tahoe Community College math student is absent

In this case, X is an example of a:

    a. variable.
    b. population.
    c. statistic.
    d. data.

Solution
a

3. The instructor's sample produces a mean number of days absent of 3.5 days. This value is an example of a:

    a. parameter.
    b. data.

c. statistic.
d. variable.

Solution
c

Exercises

For the following exercises, identify the type of data that would be used to describe a response (quantitative discrete, quantitative continuous, or qualitative), and give an example of the data.

1. number of tickets sold to a concert

Solution
quantitative discrete, 150

2. percentage of body fat

Solution
quantitative continuous, 19.2%

3. favorite baseball team

Solution
qualitative, Oakland A's

4. time in line to buy groceries

Solution
quantitative continuous, 7.2 minutes

5. number of students enrolled at Evergreen Valley College

Solution
quantitative discrete, 11,234 students

6. most-watched television show

Solution
qualitative, Dancing with the Stars

7. brand of toothpaste

Solution
qualitative, Crest

8. distance to the closest movie theater

Solution
quantitative continuous, 8.32 miles

9. age of executives in Fortune 500 companies

Solution
quantitative continuous, 47.3 years

10. number of competing computer spreadsheet software packages

Solution
quantitative discrete, three

Exercises

Use the following information to answer the next two exercises: A study was done to determine the age, number of times per week, and the duration (amount of time) of resident use of a local park in San Jose. The first house in the neighborhood around the park was selected randomly and then every 8th house in the neighborhood around the park was interviewed.

1. "Number of times per week" is what type of data?

    a. qualitative
    b. quantitative discrete
    c. quantitative continuous

Solution
b

2. "Duration (amount of time)" is what type of data?

    a. qualitative
    b. quantitative discrete
    c. quantitative continuous

Solution
c

Exercises

Airline companies are interested in the consistency of the number of babies on each flight, so that they have adequate safety equipment. Suppose an airline conducts a survey. Over Thanksgiving

weekend, it surveys six flights from Boston to Salt Lake City to determine the number of babies on the flights. It determines the amount of safety equipment needed by the result of that study.

1. Using complete sentences, list three things wrong with the way the survey was conducted.
2. Using complete sentences, list three ways that you would improve the survey if it were to be repeated.

Solution

1. The survey was conducted using six similar flights.
   The survey would not be a true representation of the entire population of air travelers.
   Conducting the survey on a holiday weekend will not produce representative results.
2. Conduct the survey during different times of the year.
   Conduct the survey using flights to and from various locations.
   Conduct the survey on different days of the week.

Exercises

Suppose you want to determine the mean number of students per statistics class in your state. Describe a possible sampling method in three to five complete sentences. Make the description detailed.

Solution

Answers will vary. Sample Answer: Randomly choose 25 colleges in the state. Use all statistics classes from each of the chosen colleges in the sample. This can be done by listing all the colleges together with a two-digit number starting with 00 then 01, etc. The list of colleges can be found on Wikipedia. http://en.wikipedia.org/wiki/List_of_colleges_and_universities_in_California Use a random number generator to pick 25 colleges.

Exercises

Suppose you want to determine the mean number of cans of soda drunk each month by students in their twenties at your school. Describe a possible sampling method in three to five complete sentences. Make the description detailed.

Solution

Answers will vary. Sample Answer: You could use a systematic sampling method. Stop the tenth person as they leave one of the buildings on campus at 9:50 in the morning. Then stop the tenth person as they leave a different building on campus at 1:50 in the afternoon.

Exercises

List some practical difficulties involved in getting accurate results from a telephone survey.

Solution

Answers will vary. Sample Answer: Not all people have a listed phone number. Many people hang up or do not respond to phone surveys.

Exercises

List some practical difficulties involved in getting accurate results from a mailed survey.

Solution
Answers will vary. Sample Answer: Many people will not respond to mail surveys. If they do respond to the surveys, you can't be sure who is responding. In addition, mailing lists can be incomplete.

Exercises

With your classmates, brainstorm some ways you could overcome these problems if you needed to conduct a phone or mail survey.

Solution
Ask everyone to include their age then take a random sample from the data. Include in the report how the survey was conducted and why the results may not be accurate.

Exercises

The instructor takes her sample by gathering data on five randomly selected students from each Lake Tahoe Community College math class. The type of sampling she used is

a. cluster sampling
b. stratified sampling

c. simple random sampling
d. convenience sampling

Solution
b

Exercises

A study was done to determine the age, number of times per week, and the duration (amount of time) of residents using a local park in San Jose. The first house in the neighborhood around the park was selected randomly and then every eighth house in the neighborhood around the park was interviewed. The sampling method was:

a. simple random
b. systematic
c. stratified
d. cluster

Solution
b

Exercises

Name the sampling method used in each of the following situations:

1. A woman in the airport is handing out questionnaires to travelers asking them to evaluate the airport's service. She does not ask travelers who are hurrying through the airport with their hands full of luggage, but instead asks all travelers who are sitting near gates and not taking naps while they wait.
2. A teacher wants to know if her students are doing homework, so she randomly selects rows two and five and then calls on all students in row two and all students in row five to present the solutions to homework problems to the class.
3. The marketing manager for an electronics chain store wants information about the ages of its customers. Over the next two weeks, at each store location, 100 randomly selected customers are given questionnaires to fill out asking for information about age, as well as about other variables of interest.
4. The librarian at a public library wants to determine what proportion of the library users are children. The librarian has a tally sheet on which she marks whether books are checked out by an adult or a child. She records this data for every fourth patron who checks out books.
5. A political party wants to know the reaction of voters to a debate between the candidates. The day after the debate, the party's polling staff calls 1,200 randomly selected phone numbers. If a registered voter answers the phone or is available to come to the phone, that registered voter is asked whom he or she intends to vote for and whether the debate changed his or her opinion of the candidates.

Solution

1. convenience
2. cluster
3. stratified
4. systematic
5. simple random

Exercises

A "random survey" was conducted of 3,274 people of the "microprocessor generation" (people born since 1971, the year the microprocessor was invented). It was reported that 48% of those individuals surveyed stated that if they had $2,000 to spend, they would use it for computer equipment. Also, 66% of those surveyed considered themselves relatively savvy computer users.

1. Do you consider the sample size large enough for a study of this type? Why or why not?
2. Based on your "gut feeling," do you believe the percentages accurately reflect the U.S. population for those individuals born since 1971? If not, do you think the percentages of the population are actually higher or lower than the sample statistics? Why?
   [Additional information: The survey, reported by Intel Corporation, was filled out by individuals who visited the Los Angeles Convention Center to see the Smithsonian Institute's road show called "America's Smithsonian."]
3. With this additional information, do you feel that all demographic and ethnic groups were equally represented at the event? Why or why not?
4. With the additional information, comment on how accurately you think the sample statistics reflect the population parameters.

Solution

1. Yes, in polling, samples that are from 1,200 to 1,500 observations are considered large enough and good enough if the survey is random and is well done.
2. We do not have enough information to decide if this is a random sample from the U.S. population.
3. No, this is a convenience sample taken from individuals who visited an exhibition in the Angeles Convention Center. This sample is not representative of the U.S. population.
4. It is possible that the two sample statistics, 48% and 66% are larger than the true parameters in the population at large. In any event, no conclusion about the population proportions can be inferred from this convenience sample.

Exercises

The Gallup-Healthways Well-Being Index is a survey that follows trends of U.S. residents on a regular basis. There are six areas of health and wellness covered in the survey: Life Evaluation, Emotional Health, Physical Health, Healthy Behavior, Work Environment, and Basic Access. Some of the questions used to measure the Index are listed below.

Identify the type of data obtained from each question used in this survey: qualitative, quantitative discrete, or quantitative continuous.

1. Do you have any health problems that prevent you from doing any of the things people your age can normally do?
2. During the past 30 days, for about how many days did poor health keep you from doing your usual activities?
3. In the last seven days, on how many days did you exercise for 30 minutes or more?
4. Do you have health insurance coverage?

Solution

1. qualitative
2. quantitative discrete
3. quantitative discrete
4. qualitative

Exercises

In advance of the 1936 Presidential Election, a magazine titled Literary Digest released the results of an opinion poll predicting that the republican candidate Alf Landon would win by a large margin. The magazine sent postcards to approximately 10,000,000 prospective voters. These prospective voters were selected from the subscription list of the magazine, from automobile registration lists, from phone lists, and from club membership lists. Approximately 2,300,000 people returned the postcards.

1. Think about the state of the United States in 1936. Explain why a sample chosen from magazine subscription lists, automobile registration lists, phone books, and club membership lists was not representative of the population of the United States at that time.
2. What effect does the low response rate have on the reliability of the sample?
3. Are these problems examples of sampling error or nonsampling error?
4. During the same year, George Gallup conducted his own poll of 30,000 prospective voters. His researchers used a method they called "quota sampling" to obtain survey answers from specific subsets of the population. Quota sampling is an example of which sampling method described in this module?

Solution

1. The country was in the middle of the Great Depression, and many people could not afford these "luxury" items and therefore were not able to be included in the survey.
2. Samples that are too small can lead to sampling bias.
3. sampling error
4. stratified

Exercises

Crime-related and demographic statistics for 47 US states in 1960 were collected from government agencies, including the FBI's *Uniform Crime Report*. One analysis of this data found a strong connection between education and crime indicating that higher levels of education in a community correspond to higher crime rates.

Which of the potential problems with samples discussed in [link] could explain this connection?

Solution

Causality: The fact that two variables are related does not guarantee that one variable is influencing

the other. We cannot assume that crime rate impacts education level or that education level impacts crime rate.

Confounding: There are many factors that define a community other than education level and crime rate. Communities with high crime rates and high education levels may have other lurking variables that distinguish them from communities with lower crime rates and lower education levels. Because we cannot isolate these variables of interest, we cannot draw valid conclusions about the connection between education and crime. Possible lurking variables include police expenditures, unemployment levels, region, average age, and size.

Exercises

YouPolls is a website that allows anyone to create and respond to polls. One question posted April 15 asks:

"Do you feel happy paying your taxes when members of the Obama administration are allowed to ignore their tax liabilities?"[1]

As of April 25, 11 people responded to this question. Each participant answered "NO!"

Which of the potential problems with samples discussed in this module could explain this connection?

Solution
Self-Selected Samples: Only people who are interested in the topic are choosing to respond.
Sample Size Issues: A sample with only 11 participants will not accurately represent the opinions of a nation.
Undue Influence: The question is wording in a specific way to generate a specific response.
Self-Funded or Self-Interest Studies: This question was generated to support one person's claim and it was designed to get the answer that the person desires.

Exercises

A scholarly article about response rates begins with the following quote:

"Declining contact and cooperation rates in random digit dial (RDD) national telephone surveys raise serious concerns about the validity of estimates drawn from such research."[2]

The Pew Research Center for People and the Press admits:

"The percentage of people we interview – out of all we try to interview – has been declining over the past decade or more."[3]

1. What are some reasons for the decline in response rate over the past decade?
2. Explain why researchers are concerned with the impact of the declining response rate on public opinion polls.

Solution

1. Possible reasons: increased use of caller id, decreased use of landlines, increased use of private numbers, voice mail, privacy managers, hectic nature of personal schedules, decreased willingness to be interviewed
2. When a large number of people refuse to participate, then the sample may not have the same characteristics of the population. Perhaps the majority of people willing to participate are doing so because they feel strongly about the subject of the survey.

Bringing It Together

Exercises

Seven hundred and seventy-one distance learning students at Long Beach City College responded to surveys in the 2010-11 academic year. Highlights of the summary report are listed in [link].

LBCC Distance Learning Survey Results

Have computer at home	96%
Unable to come to campus for classes	65%
Age 41 or over	24%
Would like LBCC to offer more DL courses	95%
Took DL classes due to a disability	17%
Live at least 16 miles from campus	13%
Took DL courses to fulfill transfer requirements	71%

1. What percentage of the students surveyed do not have a computer at home?
2. About what percentage students in the survey live at least 16 miles from campus?
3. If the same survey were done at Great Basin College in Elko, Nevada, do you think the percentages would be the same? Why?

Solution

4%

13%

Not necessarily. Long Beach City is the seventh largest college in California, and it has an enrollment of approximately 27,000 students. On the other hand, Great Basin College has its campuses in rural northeastern Nevada, and its enrollment of about 3,500 students.

Exercises

Several online textbook retailers advertise that they have lower prices than on-campus bookstores. However, an important factor is whether the Internet retailers actually have the textbooks that students need in stock. Students need to be able to get textbooks promptly at the beginning of the

college term. If the book is not available, then a student would not be able to get the textbook at all, or might get a delayed delivery if the book is back ordered.

A college newspaper reporter is investigating textbook availability at online retailers. He decides to investigate one textbook for each of the following seven subjects: calculus, biology, chemistry, physics, statistics, geology, and general engineering. He consults textbook industry sales data and selects the most popular nationally used textbook in each of these subjects. He visits websites for a random sample of major online textbook sellers and looks up each of these seven textbooks to see if they are available in stock for quick delivery through these retailers. Based on his investigation, he writes an article in which he draws conclusions about the overall availability of all college textbooks through online textbook retailers.

Write an analysis of his study that addresses the following issues: Is his sample representative of the population of all college textbooks? Explain why or why not. Describe some possible sources of bias in this study, and how it might affect the results of the study. Give some suggestions about what could be done to improve the study.

Solution
Answers will vary. Sample answer: The sample is not representative of the population of all college textbooks. Two reasons why it is not representative are that he only sampled seven subjects and he only investigated one textbook in each subject. There are several possible sources of bias in the study. The seven subjects that he investigated are all in mathematics and the sciences; there are many subjects in the humanities, social sciences, and other subject areas (for example: literature, art, history, psychology, sociology, business) that he did not investigate at all. It may be that different subject areas exhibit different patterns of textbook availability, but his sample would not detect such results.

He also looked only at the most popular textbook in each of the subjects he investigated. The availability of the most popular textbooks may differ from the availability of other textbooks in one of two ways:

- the most popular textbooks may be more readily available online, because more new copies are printed, and more students nationwide are selling back their used copies, OR
- the most popular textbooks may be harder to find available online, because more student demand exhausts the supply more quickly.

In reality, many college students do not use the most popular textbooks in their subject, and this study gives no useful information about the situation for those less popular textbooks.

He could improve this study by:

- expanding the selection of subjects he investigates so that it is more representative of all subjects studied by college students, and
- expanding the selection of textbooks he investigates within each subject to include a mixed representation of both the most popular and less popular textbooks.

HOMEWORK from 1.3

Exercises

Fifty part-time students were asked how many courses they were taking this term. The (incomplete) results are shown below:

Part-time Student Course Loads

# of Courses	Frequency	Relative Frequency	Cumulative Relative Frequency
1	30	0.6	
2	15		
3			

1. Fill in the blanks in [link].
2. What percent of students take exactly two courses?
3. What percent of students take one or two courses?

Solution

30%
90%

Exercises

Sixty adults with gum disease were asked the number of times per week they used to floss before their diagnosis. The (incomplete) results are shown in [link].

Flossing Frequency for Adults with Gum Disease

# Flossing per Week	Frequency	Relative Frequency	Cumulative Relative Freq.
0	27	0.4500	
1	18		
3			0.9333
6	3	0.0500	
7	1	0.0167	

1. Fill in the blanks in [link].
2. What percent of adults flossed six times per week?
3. What percentage flossed at most three times per week?

Solution

1.

# Flossing per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
0	27	0.4500	0.4500
1	18	0.3000	0.7500
3	11	0.1833	0.9333
6	3	0.0500	0.9833
7	1	0.0167	1

2. 5.00%
3. 93.33%

Exercises

Nineteen immigrants to the U.S were asked how many years, to the nearest year, they have lived in the U.S. The data are as follows: 2; 5; 7; 2; 2; 10; 20; 15; 0; 7; 0; 20; 5; 12; 15; 12; 45; 10.

[link] was produced.

Frequency of Immigrant Survey Responses

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
0	2	$\frac{2}{19}$	0.1053
2	3	$\frac{3}{19}$	0.2632
4	1	$\frac{1}{19}$	0.3158
5	3	$\frac{3}{19}$	0.4737
7	2	$\frac{2}{19}$	0.5789
10	2	$\frac{2}{19}$	0.6842
12	2	$\frac{2}{19}$	0.7895
15	1	$\frac{1}{19}$	0.8421
20	1	$\frac{1}{19}$	1.0000

1. Fix the errors in [link]. Also, explain how someone might have arrived at the incorrect number(s).
2. Explain what is wrong with this statement: “47 percent of the people surveyed have lived in

the U.S. for 5 years."

3. Fix the statement in **b** to make it correct.
4. What fraction of the people surveyed have lived in the U.S. five or seven years?
5. What fraction of the people surveyed have lived in the U.S. at most 12 years?
6. What fraction of the people surveyed have lived in the U.S. fewer than 12 years?
7. What fraction of the people surveyed have lived in the U.S. from five to 20 years, inclusive?

Solution

1. The Frequencies for 15 and 20 should both be two and the Relative Frequencies should both be 2-19
2. The mistake could be due to copying the data down wrong. The Cumulative Relative Frequency for five years should be 0.4737. The mistake is due to calculating the Relative Frequency instead of the Cumulative Relative Frequency. The Cumulative Relative Frequency for 15 years should be 0.8947
3. The 47% is the Cumulative Relative Frequency, not the Relative Frequency.
   47% of the people surveyed have lived in the U.S. for five years or less.
4. $\frac{5}{19}$
5. $\frac{15}{19}$
6. $\frac{13}{19}$
7. $\frac{13}{19}$

Exercises

How much time does it take to travel to work? [link] shows the mean commute time by state for

workers at least 16 years old who are not working at home. Find the mean travel time, and round off the answer properly.

24.0	24.3	25.9	18.9	27.5	17.9	21.8	20.9	16.7	27.3
18.2	24.7	20.0	22.6	23.9	18.0	31.4	22.3	24.0	25.5
24.7	24.6	28.1	24.9	22.6	23.6	23.4	25.7	24.8	25.5
21.2	25.7	23.1	23.0	23.9	26.0	16.3	23.1	21.4	21.5
27.0	27.0	18.6	31.7	23.3	30.1	22.9	23.3	21.7	18.6

Solution

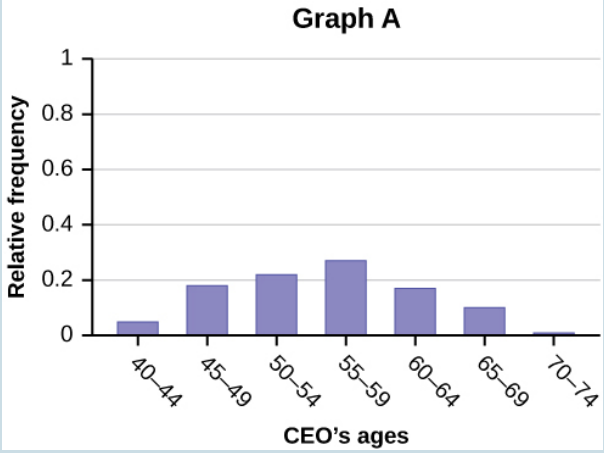
The sum of the travel times is 1,173.1. Divide the sum by 50 to calculate the mean value: 23.462. Because each state's travel time was measured to the nearest tenth, round this calculation to the nearest hundredth: 23.46.

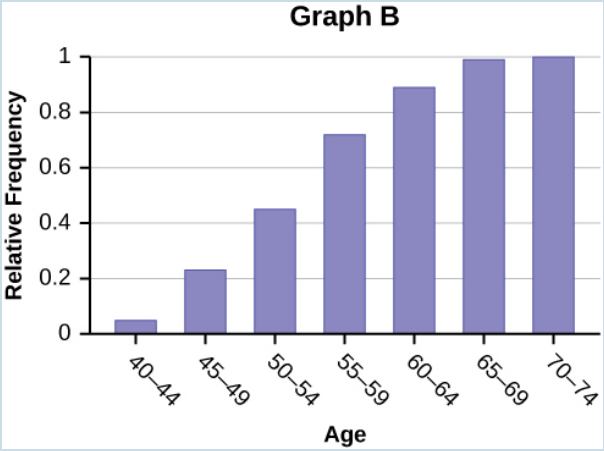
Exercises

Forbes magazine published data on the best small firms in 2012. These were firms which had been publicly traded for at least a year, have a stock price of at least $5 per share, and have reported annual revenue between $5 million and $1 billion. [link] shows the ages of the chief executive officers for the first 60 ranked firms.

Age	Frequency	Relative Frequency	Cumulative Relative Frequency
40–44	3		
45–49	11		
50–54	13		
55–59	16		
60–64	10		
65–69	6		
70–74	1		

1. What is the frequency for CEO ages between 54 and 65?
2. What percentage of CEOs are 65 years or older?
3. What is the relative frequency of ages under 50?
4. What is the cumulative relative frequency for CEOs younger than 55?
5. Which graph shows the relative frequency and which shows the cumulative relative frequency?





Solution

26 (This is the count of CEOs in the 55 to 59 and 60 to 64 categories.)

12% (number of CEOs age 65 or older ÷ total number of CEOs)

14/60; 0.23; 23%

0.45

Graph A represents the cumulative relative frequency, and Graph B shows the relative frequency.

Exercises

Use the following information to answer the next two exercises:[link] contains data on hurricanes that have made direct hits on the U.S. Between 1851 and 2004. A hurricane is given a strength category rating based on the minimum wind speed generated by the storm.

Frequency of Hurricane Direct Hits

Category	Number of Direct Hits	Relative Frequency	Cumulative Frequency
1	109	0.3993	0.3993
2	72	0.2637	0.6630
3	71	0.2601	
4	18		0.9890
5	3	0.0110	1.0000
	Total = 273		

1. What is the relative frequency of direct hits that were category 4 hurricanes?

a. 0.0768
b. 0.0659
c. 0.2601
d. Not enough information to calculate

Solution
b

2. What is the relative frequency of direct hits that were AT MOST a category 3 storm?

a. 0.3480
b. 0.9231
c. 0.2601
d. 0.3370

Solution
b

HOMEWORK from 1.4

Exercises

How does sleep deprivation affect your ability to drive? A recent study measured the effects on 19 professional drivers. Each driver participated in two experimental sessions: one after normal sleep and one after 27 hours of total sleep deprivation. The treatments were assigned in random order. In each session, performance was measured on a variety of tasks including a driving simulation.

Use key terms from this module to describe the design of this experiment.

Solution
Explanatory variable: amount of sleep

Response variable: performance measured in assigned tasks

Treatments: normal sleep and 27 hours of total sleep deprivation

Experimental Units: 19 professional drivers

Lurking variables: none – all drivers participated in both treatments

Random assignment: treatments were assigned in random order; this eliminated the effect of any "learning" that may take place during the first experimental session

Control/Placebo: completing the experimental session under normal sleep conditions

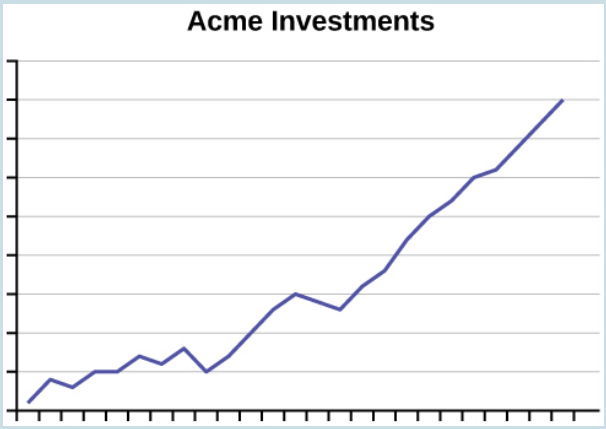
Blinding: researchers evaluating subjects' performance must not know which treatment is being applied at the time

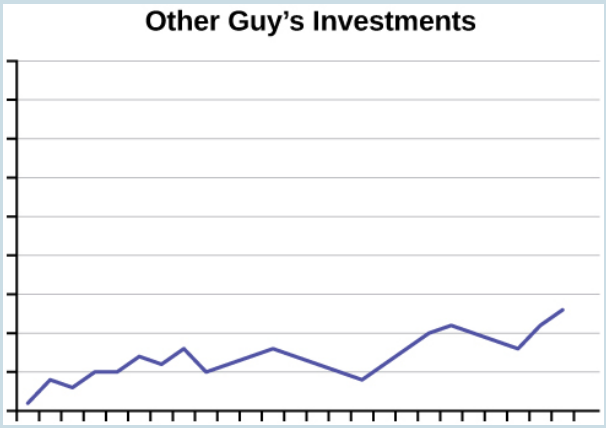
Exercises

An advertisement for Acme Investments displays the two graphs in [link] to show

the value of Acme's product in comparison with the Other Guy's product. Describe the potentially misleading visual effect of these comparison graphs. How can this be corrected?

As the graphs show, Acme consistently outperforms the Other Guys!





Solution

The graphs do not show scales of values. We do not know the period of time each graph represents; they may show data from different years. We also do not know if the vertical scales on each graph are equivalent. The scales may have been adjusted to exaggerate or minimize trends. There is no reliable information to be gleaned from these graphs, and setting them up as examples of performance is misleading.

Exercises

The graph in [link] shows the number of complaints for six different airlines as reported to the US

Department of Transportation in February 2013. Alaska, Pinnacle, and Airtran Airlines have far fewer complaints reported than American, Delta, and United. Can we conclude that American, Delta, and United are the worst airline carriers since they have the most complaints?



Solution

You cannot assume that the numbers of complaints reflect the quality of the airlines. The airlines shown with the greatest number of complaints are the ones with the most passengers. You must consider the appropriateness of methods for presenting data; in this case displaying totals is misleading.

CHAPTER 2: DESCRIPTIVE STATISTICS

INTRODUCTION TO CHAPTER 2: DESCRIPTIVE STATISTICS



Figure 2.1. When you have large amounts of data, you will need to organize it in a way that makes sense. These ballots from an election are rolled together with similar ballots to keep them organized. (credit: William Greeson)

Once you have collected data, what will you do with it? Data can be described and presented in many different formats. For example, suppose you are interested in buying a house in a particular area. You may have no clue about the house prices, so you might ask your real estate agent to give you a sample data set of prices. Looking at all the prices in the sample often is overwhelming. A better way might be to look at the median price and the variation of prices. The median and variation are just two ways that you will learn to describe data. Your agent might also provide you with a graph of the data.

In this chapter, you will study numerical and graphical ways to describe and display your data. This area of

statistics is called **"Descriptive Statistics."** You will learn how to calculate, and even more importantly, how to interpret these measurements and graphs.

A statistical graph is a tool that helps you learn about the shape or distribution of a sample or a population. A graph can be a more effective way of presenting data than a mass of numbers because we can see where data clusters and where there are only a few data values. Newspapers and the Internet use graphs to show trends and to enable readers to compare facts and figures quickly. Statisticians often graph data first to get a picture of the data. Then, more formal tools may be applied.

Some of the types of graphs that are used to summarize and organize data are the dot plot, the bar graph, the histogram, the stem-and-leaf plot, the frequency polygon (a type of broken line graph), the pie chart, and the box plot. In this chapter, we will briefly look at stem-and-leaf plots, line graphs, and bar graphs, as well as frequency polygons, and time series graphs. Our emphasis will be on histograms and box plots.

Note

This book contains instructions for constructing a histogram and a box plot for the TI-83+ and TI-84 calculators. The Texas Instruments (TI) website provides additional instructions for using these calculators.

Collaborative Exercises

Section 2.2

Collaborative Exercise

Count the money (bills and change) in your pocket or purse. Your instructor will record the amounts. As a class, construct a histogram displaying the data. Discuss how many intervals you think is appropriate. You may want to experiment with the number of intervals.

Section 2.3

Collaborative Exercise

Your instructor or a member of the class will ask everyone in class how many sweaters they own. Answer the following questions:

1. How many students were surveyed?
2. What kind of sampling did you do?
3. Construct two different histograms. For each, starting value = _____ ending value = ____.
4. Find the median, first quartile, and third quartile.
5. Construct a table of the data to find the following:
   a. the 10^{th} percentile

b. the 70^{th} percentile
c. the percentage of students who own fewer than four sweaters

Media Attributions

- Chapter 2 Introduction Figure © OpenStax Introductory Statistics is licensed under a CC BY (Attribution) license

2.1 STEM-AND-LEAF GRAPHS (STEMPLOTS), LINE GRAPHS, AND BAR GRAPHS

Learning Objectives

By the end of this section, the student should be able to:

- Display data graphically and interpret data with stemplots.
- Construct other types of graphs and interpret information displayed in: line graphs and bar graphs.

One simple graph, the **stem-and-leaf graph** or **stemplot**, comes from the field of exploratory data analysis. It is a good choice when the data sets are small. To create the plot, divide each observation of data into a stem and a leaf. The leaf consists of a **final significant digit**. For example, 23 has stem two and leaf three. The number 432 has stem 43 and leaf two. Likewise, the number 5,432 has stem 543 and leaf two. The decimal 9.3 has stem nine and leaf three. Write the stems in a vertical line from smallest to largest. Draw a vertical line to the right of the stems. Then write the leaves in increasing order next to their corresponding stem.

For Susan Dean's spring pre-calculus class, scores for the first exam were as follows (smallest to largest):

33; 42; 49; 49; 53; 55; 55; 61; 63; 67; 68; 68; 69; 69; 72; 73; 74; 78; 80; 83; 88; 88; 88; 90; 92; 94; 94; 94; 94; 96; 100

Stem-and-Leaf Graph	
Stem	**Leaf**
3	3
4	2 9 9
5	3 5 5
6	1 3 7 8 8 9 9
7	2 3 4 8
8	0 3 8 8 8
9	0 2 4 4 4 4 6
10	0

The stemplot shows that most scores fell in the 60s, 70s, 80s, and 90s. Eight out of the 31 scores or approximately 26% $\left(\frac{8}{31}\right)$ were in the 90s or 100, a fairly high number of As.

TRY IT

For the Park City basketball team, scores for the last 30 games were as follows (smallest to largest):

32; 32; 33; 34; 38; 40; 42; 42; 43; 44; 46; 47; 47; 48; 48; 48; 49; 50; 50; 51; 52; 52; 52; 53; 54; 56; 57; 57; 60; 61

Construct a stem plot for the data.

Stem	**Leaf**
3	2 2 3 4 8
4	0 2 2 3 4 6 7 7 8 8 8 9
5	0 0 1 2 2 2 3 4 6 7 7
6	0 1

The stemplot is a quick way to graph data and gives an exact picture of the data. You want to look for an overall pattern and any outliers. An outlier is an observation of data that does not fit the rest of the data. It is sometimes called an **extreme value.** When you graph an outlier, it will appear not to fit the pattern of the graph. Some outliers are due to mistakes (for example, writing down 50 instead of 500) while others may indicate that something unusual is happening. It takes some background information to explain outliers, so we will cover them in more detail later.

The data are the distances (in kilometers) from a home to local supermarkets. Create a stemplot using the data:

1.1; 1.5; 2.3; 2.5; 2.7; 3.2; 3.3; 3.3; 3.5; 3.8; 4.0; 4.2; 4.5; 4.5; 4.7; 4.8; 5.5; 5.6; 6.5; 6.7; 12.3

Do the data seem to have any concentration of values?

The leaves are to the right of the decimal.

Solution
The value 12.3 may be an outlier. Values appear to concentrate at three and four kilometers.

Stem	Leaf
1	1 5
2	3 5 7
3	2 3 3 5 8
4	0 2 5 5 7 8
5	5 6
6	5 7
7	
8	
9	
10	
11	
12	3

TRY IT

The following data show the distances (in miles) from the homes of off-campus statistics students to the college. Create a stem plot using the data and identify any outliers:

0.5; 0.7; 1.1; 1.2; 1.2; 1.3; 1.3; 1.5; 1.5; 1.7; 1.7; 1.8; 1.9; 2.0; 2.2; 2.5; 2.6; 2.8; 2.8; 2.8; 3.5; 3.8; 4.4; 4.8; 4.9; 5.2; 5.5; 5.7; 5.8; 8.0

Stem	Leaf
0	5 7
1	1 2 2 3 3 5 5 7 7 8 9
2	0 2 5 6 8 8 8
3	5 8
4	4 8 9
5	2 5 7 8
6	
7	
8	0

The value 8.0 may be an outlier. Values appear to concentrate at one and two miles.

A **side-by-side stem-and-leaf plot** allows a comparison of the two data sets in two columns. In a side-by-side stem-and-leaf plot, two sets of leaves share the same stem. The leaves are to the left and the right of the stems. [link] and [link] show the ages of presidents at their inauguration and at their death. Construct a side-by-side stem-and-leaf plot using this data.

Solution

Ages at Inauguration		Ages at Death
998777632	4	69
87777666555544444211111 0	5	366778
954421110	6	003344567778
	7	0011147889
	8	01358
	9	0033

Presidential Ages at Inauguration

President	Age	President	Age	President	Age
Washington	57	Lincoln	52	Hoover	54
J. Adams	61	A. Johnson	56	F. Roosevelt	51
Jefferson	57	Grant	46	Truman	60
Madison	57	Hayes	54	Eisenhower	62
Monroe	58	Garfield	49	Kennedy	43
J. Q. Adams	57	Arthur	51	L. Johnson	55
Jackson	61	Cleveland	47	Nixon	56
Van Buren	54	B. Harrison	55	Ford	61
W. H. Harrison	68	Cleveland	55	Carter	52
Tyler	51	McKinley	54	Reagan	69
Polk	49	T. Roosevelt	42	G.H.W. Bush	64
Taylor	64	Taft	51	Clinton	47
Fillmore	50	Wilson	56	G. W. Bush	54
Pierce	48	Harding	55	Obama	47
Buchanan	65	Coolidge	51		

Presidential Age at Death

President	Age	President	Age	President	Age
Washington	67	Lincoln	56	Hoover	90
J. Adams	90	A. Johnson	66	F. Roosevelt	63
Jefferson	83	Grant	63	Truman	88
Madison	85	Hayes	70	Eisenhower	78
Monroe	73	Garfield	49	Kennedy	46
J. Q. Adams	80	Arthur	56	L. Johnson	64
Jackson	78	Cleveland	71	Nixon	81
Van Buren	79	B. Harrison	67	Ford	93
W. H. Harrison	68	Cleveland	71	Reagan	93
Tyler	71	McKinley	58		
Polk	53	T. Roosevelt	60		
Taylor	65	Taft	72		
Fillmore	74	Wilson	67		
Pierce	64	Harding	57		
Buchanan	77	Coolidge	60		

The table shows the number of wins and losses the Atlanta Hawks have had in 42 seasons. Create a side-by-side stem-and-leaf plot of these wins and losses.

Losses	Wins	Year	Losses	Wins	Year
34	48	1968–1969	41	41	1989–1990
34	48	1969–1970	39	43	1990–1991
46	36	1970–1971	44	38	1991–1992
46	36	1971–1972	39	43	1992–1993
36	46	1972–1973	25	57	1993–1994
47	35	1973–1974	40	42	1994–1995
51	31	1974–1975	36	46	1995–1996
53	29	1975–1976	26	56	1996–1997
51	31	1976–1977	32	50	1997–1998
41	41	1977–1978	19	31	1998–1999
36	46	1978–1979	54	28	1999–2000
32	50	1979–1980	57	25	2000–2001
51	31	1980–1981	49	33	2001–2002
40	42	1981–1982	47	35	2002–2003
39	43	1982–1983	54	28	2003–2004
42	40	1983–1984	69	13	2004–2005
48	34	1984–1985	56	26	2005–2006
32	50	1985–1986	52	30	2006–2007
25	57	1986–1987	45	37	2007–2008
32	50	1987–1988	35	47	2008–2009
30	52	1988–1989	29	53	2009–2010

Solution

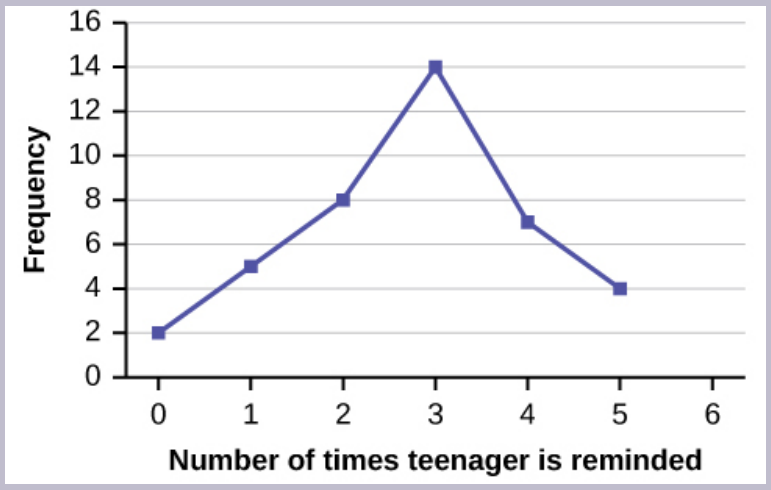
Atlanta Hawks Wins and Losses

Number of Wins		Number of Losses
3	1	9
98865	2	559
8766554311110	3	02222445666999
887666333221 10	4	0011245667789
776320000	5	111234467
	6	9

Another type of graph that is useful for specific data values is a **line graph**. In the particular line graph shown in [link], the ***x*-axis** (horizontal axis) consists of **data values** and the ***y*-axis** (vertical axis) consists of **frequency points**. The frequency points are connected using line segments.

In a survey, 40 mothers were asked how many times per week a teenager must be reminded to do his or her chores. The results are shown in [link] and in [link].

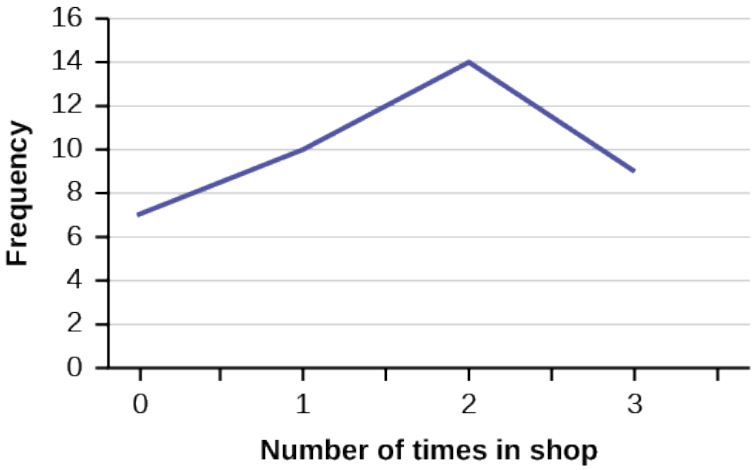
Number of times teenager is reminded	Frequency
0	2
1	5
2	8
3	14
4	7
5	4



TRY IT

In a survey, 40 people were asked how many times per year they had their car in the shop for repairs. The results are shown in [link]. Construct a line graph.

Number of times in shop	Frequency
0	7
1	10
2	14
3	9

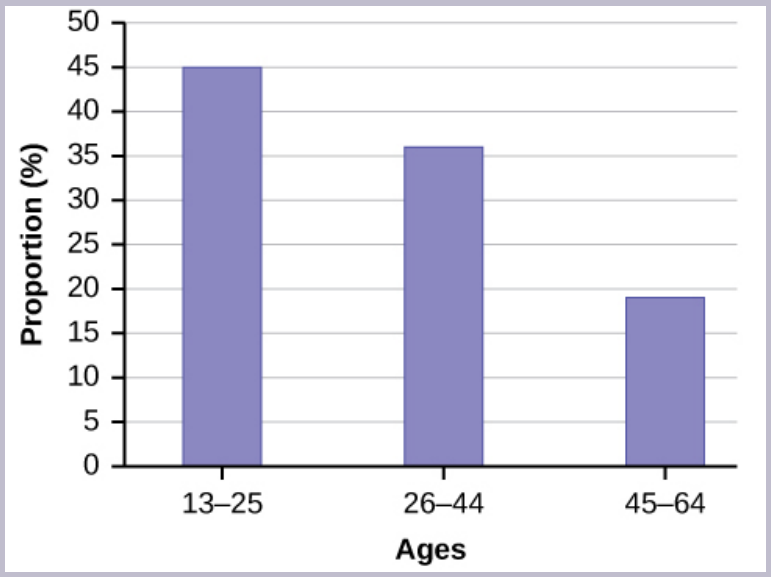


Bar graphs consist of bars that are separated from each other. The bars can be rectangles or they can be rectangular boxes (used in three-dimensional plots), and they can be vertical or horizontal. The **bar graph** shown in [link] has age groups represented on the ***x*-axis** and proportions on the ***y*-axis**.

By the end of 2011, Facebook had over 146 million users in the United States. [link] shows three age groups, the number of users in each age group, and the proportion (%) of users in each age group. Construct a bar graph using this data.

Age groups	Number of Facebook users	Proportion (%) of Facebook users
13–25	65,082,280	45%
26–44	53,300,200	36%
45–64	27,885,100	19%

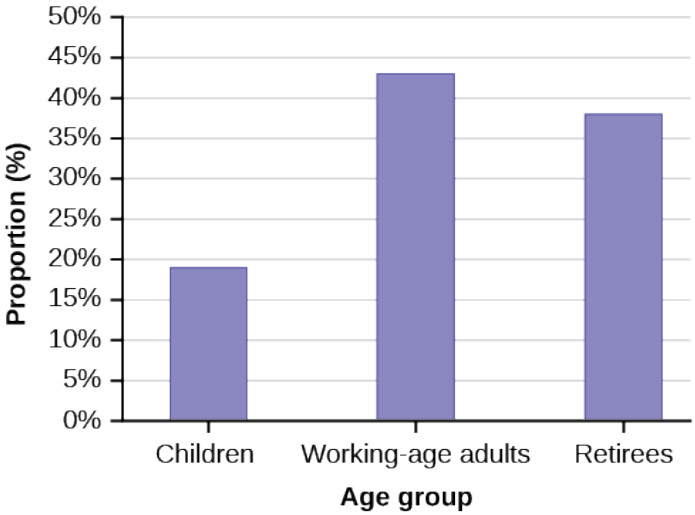
Solution



TRY IT

The population in Park City is made up of children, working-age adults, and retirees. [link] shows the three age groups, the number of people in the town from each age group, and the proportion (%) of people in each age group. Construct a bar graph showing the proportions.

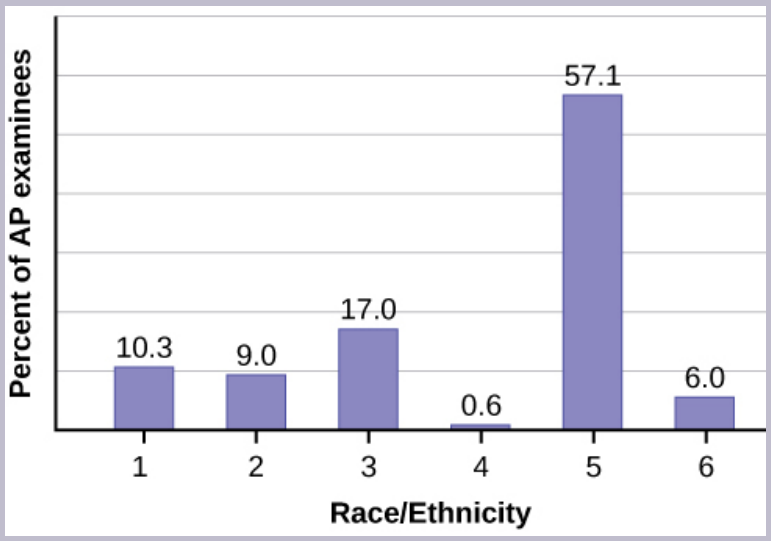
Age groups	Number of people	Proportion of population
Children	67,059	19%
Working-age adults	152,198	43%
Retirees	131,662	38%



The columns in [link] contain: the race or ethnicity of students in U.S. Public Schools for the class of 2011, percentages for the Advanced Placement examine population for that class, and percentages for the overall student population. Create a bar graph with the student race or ethnicity (qualitative data) on the *x*-axis, and the Advanced Placement examinee population percentages on the *y*-axis.

Race/Ethnicity	AP Examinee Population	Overall Student Population
1 = Asian, Asian American or Pacific Islander	10.3%	5.7%
2 = Black or African American	9.0%	14.7%
3 = Hispanic or Latino	17.0%	17.6%
4 = American Indian or Alaska Native	0.6%	1.1%
5 = White	57.1%	59.2%
6 = Not reported/other	6.0%	1.7%

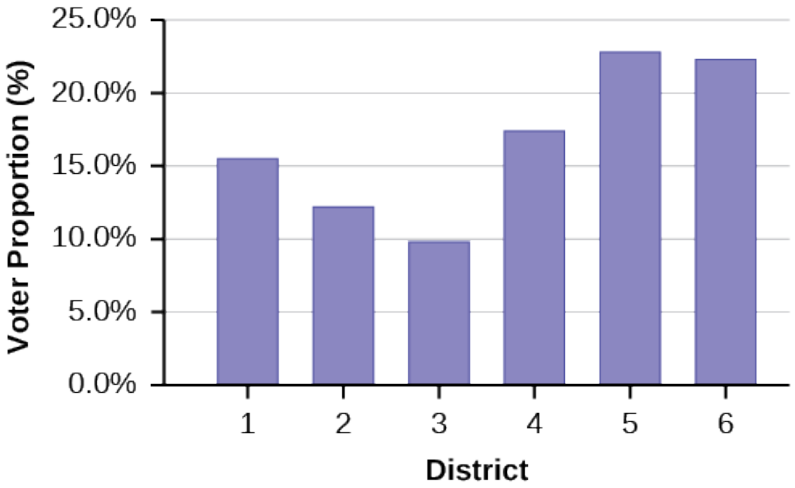
Solution



TRY IT

Park city is broken down into six voting districts. The table shows the percent of the total registered voter population that lives in each district as well as the percent total of the entire population that lives in each district. Construct a bar graph that shows the registered voter population by district.

District	Registered voter population	Overall city population
1	15.5%	19.4%
2	12.2%	15.6%
3	9.8%	9.0%
4	17.4%	18.5%
5	22.8%	20.7%
6	22.3%	16.8%



References

Burbary, Ken. *Facebook Demographics Revisited – 2001 Statistics,* 2011. Available online at http://www.kenburbary.com/2011/03/facebook-demographics-revisited-2011-statistics-2/ (accessed August 21, 2013).

"9th Annual AP Report to the Nation." CollegeBoard, 2013. Available online at http://apreport.collegeboard.org/goals-and-findings/promoting-equity (accessed September 13, 2013).

"Overweight and Obesity: Adult Obesity Facts." Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/obesity/data/adult.html (accessed September 13, 2013).

2.2 HISTOGRAMS, FREQUENCY POLYGONS, AND TIME SERIES GRAPHS

Learning Objectives

By the end of this section, the student should be able to:

- Display data graphically and interpret data with histograms.
- Construct other types of graphs and interpret information displayed in: frequency polygons and time series graphs.

For most of the work you do in this book, you will use a **histogram** to display the data. One advantage of a histogram is that it can readily display large data sets. A rule of thumb is to use a histogram when the data set consists of 100 values or more.

A histogram consists of contiguous (adjoining) boxes. It has both a horizontal axis and a vertical axis. The horizontal axis is labeled with what the data represents (for instance, distance from your home to school). The vertical axis is labeled either frequency or relative frequency (or percent frequency or probability). The graph will have the same shape with either label. The histogram (like the stemplot) can give you the shape of the data, the center, and the spread of the data.

The relative frequency is equal to the frequency for an observed value of the data divided by the total number of data values in the sample. (Remember, frequency is defined as the number of times an answer occurs.) If:

- f = **frequency**
- n = total number of data values (or the sum of the individual frequencies), and
- RF = **relative frequency**,

then:

$$\text{RF} = \frac{f}{n}$$

For example, if three students in Mr. Ahab's English class of 40 students received from 90% to 100%, then, $f = 3$, $n = 40$, and $RF = \frac{f}{n} = \frac{3}{40} = 0.075$. 7.5% of the students received 90–100%. 90–100% are quantitative measures.

To construct a histogram, first decide how many **bars** or **intervals**, also called classes, represent the data. Many histograms consist of five to 15 bars or classes for clarity. The number of bars needs to be chosen. Choose a starting point for the first interval to be less than the smallest data value. A **convenient starting point** is a lower value carried out to one more decimal place than the value with the most decimal places. For example, if the value with the most decimal places is 6.1 and this is the smallest value, a convenient starting point is 6.05 (6.1 – 0.05 = 6.05). We say that 6.05 has more precision. If the value with the most decimal places is 2.23 and the lowest value is 1.5, a convenient starting point is 1.495 (1.5 – 0.005 = 1.495). If the value with the most decimal places is 3.234 and the lowest value is 1.0, a convenient starting point is 0.9995 (1.0 – 0.0005 = 0.9995). If all the data happen to be integers and the smallest value is two, then a convenient starting point is 1.5 (2 – 0.5 = 1.5). Also, when the starting point and other boundaries are carried to one additional decimal place, no data value will fall on a boundary. The next two examples go into detail about how to construct a histogram using continuous data and how to create a histogram using discrete data.

The following data are the heights (in inches to the nearest half inch) of 100 male semiprofessional soccer players. The heights are **continuous** data, since height is measured.

60; 60.5; 61; 61; 61.5

63.5; 63.5; 63.5

64; 64; 64; 64; 64; 64; 64; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5

66; 66; 66; 66; 66; 66; 66; 66; 66; 66; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5

68; 68; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69.5; 69.5; 69.5; 69.5; 69.5

70; 70; 70; 70; 70; 70; 70.5; 70.5; 70.5; 71; 71; 71

72; 72; 72; 72.5; 72.5; 73; 73.5

74

The smallest data value is 60. Since the data with the most decimal places has one decimal (for

instance, 61.5), we want our starting point to have two decimal places. Since the numbers 0.5, 0.05, 0.005, etc. are convenient numbers, use 0.05 and subtract it from 60, the smallest value, for the convenient starting point.

60 – 0.05 = 59.95 which is more precise than, say, 61.5 by one decimal place. The starting point is, then, 59.95.

The largest value is 74, so 74 + 0.05 = 74.05 is the ending value.

Next, calculate the width of each bar or class interval. To calculate this width, subtract the starting point from the ending value and divide by the number of bars (you must choose the number of bars you desire). Suppose you choose eight bars.

$$\frac{74.05 - 59.95}{8} = 1.76$$

NOTE

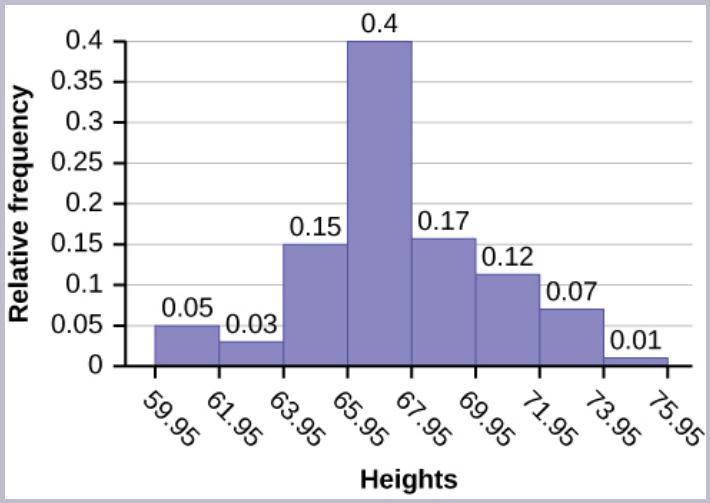
We will round up to two and make each bar or class interval two units wide. Rounding up to two is one way to prevent a value from falling on a boundary. Rounding to the next number is often necessary even if it goes against the standard rules of rounding. For this example, using 1.76 as the width would also work. A guideline that is followed by some for the width of a bar or class interval is to take the square root of the number of data values and then round to the nearest whole number, if necessary. For example, if there are 150 values of data, take the square root of 150 and round to 12 bars or intervals.

The boundaries are:

- 59.95
- 59.95 + 2 = 61.95
- 61.95 + 2 = 63.95
- 63.95 + 2 = 65.95
- 65.95 + 2 = 67.95
- 67.95 + 2 = 69.95
- 69.95 + 2 = 71.95
- 71.95 + 2 = 73.95
- 73.95 + 2 = 75.95

The heights 60 through 61.5 inches are in the interval 59.95–61.95. The heights that are 63.5 are in the interval 61.95–63.95. The heights that are 64 through 64.5 are in the interval 63.95–65.95. The heights 66 through 67.5 are in the interval 65.95–67.95. The heights 68 through 69.5 are in the interval 67.95–69.95. The heights 70 through 71 are in the interval 69.95–71.95. The heights 72 through 73.5 are in the interval 71.95–73.95. The height 74 is in the interval 73.95–75.95.

The following histogram displays the heights on the *x*-axis and relative frequency on the *y*-axis.



TRY IT

The following data are the shoe sizes of 50 male students. The sizes are continuous data since shoe size is measured. Construct a histogram and calculate the width of each bar or class interval. Suppose you choose six bars.

9; 9; 9.5; 9.5; 10; 10; 10; 10; 10; 10; 10.5; 10.5; 10.5; 10.5; 10.5; 10.5; 10.5; 10.5

11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11.5; 11.5; 11.5; 11.5; 11.5; 11.5; 11.5

12; 12; 12; 12; 12; 12; 12; 12.5; 12.5; 12.5; 12.5; 14

Smallest value: 9

Largest value: 14

Convenient starting value: 9 – 0.05 = 8.95

Convenient ending value: 14 + 0.05 = 14.05

$$\frac{14.05 - 8.95}{6} = 0.85$$

The calculations suggest using 0.85 as the width of each bar or class interval. You can also use an interval with a width equal to one.

The following data are the number of books bought by 50 part-time college students at ABC College. The number of books is **discrete data**, since books are counted.

1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1

2; 2; 2; 2; 2; 2; 2; 2; 2; 2

3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3

4; 4; 4; 4; 4; 4

5; 5; 5; 5; 5

6; 6

Eleven students buy one book. Ten students buy two books. Sixteen students buy three books. Six students buy four books. Five students buy five books. Two students buy six books.

Because the data are integers, subtract 0.5 from 1, the smallest data value and add 0.5 to 6, the largest data value. Then the starting point is 0.5 and the ending value is 6.5.

Next, calculate the width of each bar or class interval. If the data are discrete and there are not too many different values, a width that places the data values in the middle of the bar or class interval is the most convenient. Since the data consist of the numbers 1, 2, 3, 4, 5, 6, and the starting point is 0.5, a width of one places the 1 in the middle of the interval from 0.5 to 1.5, the 2 in the middle of the interval from 1.5 to 2.5, the 3 in the middle of the interval from 2.5 to 3.5, the 4 in the middle of the interval from _____ to _____, the 5 in the middle of the interval from _____ to _____, and the _____ in the middle of the interval from _____ to _____ .

Solution

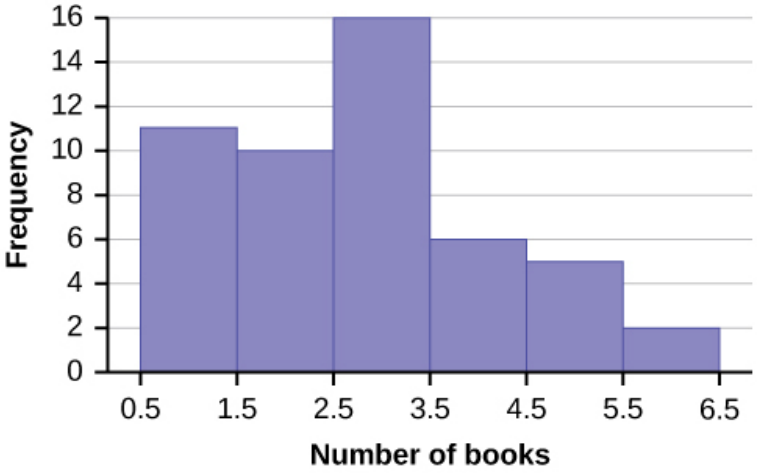
- 3.5 to 4.5
- 4.5 to 5.5
- 6
- 5.5 to 6.5

Calculate the number of bars as follows:

$$\frac{6.5 - 0.5}{\text{numberofbars}} = 1$$

where 1 is the width of a bar. Therefore, bars = 6.

The following histogram displays the number of books on the *x*-axis and the frequency on the *y*-axis.



Below are calculator instructions for entering data and for creating a customized histogram. Create the histogram for [link].

- Press Y=. Press CLEAR to delete any equations.
- Press STAT 1:EDIT. If L1 has data in it, arrow up into the name L1, press CLEAR and then arrow down. If necessary, do the same for L2.
- Into L1, enter 1, 2, 3, 4, 5, 6.
- Into L2, enter 11, 10, 16, 6, 5, 2.
- Press WINDOW. Set Xmin = .5, Xscl = (6.5 – .5)/6, Ymin = –1, Ymax = 20, Yscl = 1, Xres = 1.
- Press 2nd Y=. Start by pressing 4:Plotsoff ENTER.
- Press 2nd Y=. Press 1:Plot1. Press ENTER. Arrow down to TYPE. Arrow to the 3rd picture (histogram). Press ENTER.
- Arrow down to Xlist: Enter L1 (2nd 1). Arrow down to Freq. Enter L2 (2nd 2).
- Press GRAPH.
- Use the TRACE key and the arrow keys to examine the histogram.

TRY IT

The following data are the number of sports played by 50 student athletes. The number of sports is discrete data since sports are counted.

1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1

2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2

3; 3; 3; 3; 3; 3; 3; 3

20 student athletes play one sport. 22 student athletes play two sports. Eight student athletes play three sports.

Fill in the blanks for the following sentence. Since the data consist of the numbers 1, 2, 3, and the starting point is 0.5, a width of one places the 1 in the middle of the interval 0.5 to _____, the 2 in the middle of the interval from _____ to _____, and the 3 in the middle of the interval from _____ to _____.

1.5

1.5 to 2.5

2.5 to 3.5

Using this data set, construct a histogram.

Number of Hours My Classmates Spent Playing Video Games on Weekends				
9.95	10	2.25	16.75	0
19.5	22.5	7.5	15	12.75
5.5	11	10	20.75	17.5
23	21.9	24	23.75	18
20	15	22.9	18.8	20.5

Solution



Some values in this data set fall on boundaries for the class intervals. A value is counted in a class interval if it falls on the left boundary, but not if it falls on the right boundary. Different researchers may set up histograms for the same data in different ways. There is more than one correct way to set up a histogram.

TRY IT

The following data represent the number of employees at various restaurants in New York City. Using this data, create a histogram.

22; 35; 15; 26; 40; 28; 18; 20; 25; 34; 39; 42; 24; 22; 19; 27; 22; 34; 40; 20; 38; and 28

Use 10–19 as the first interval.

Frequency Polygons

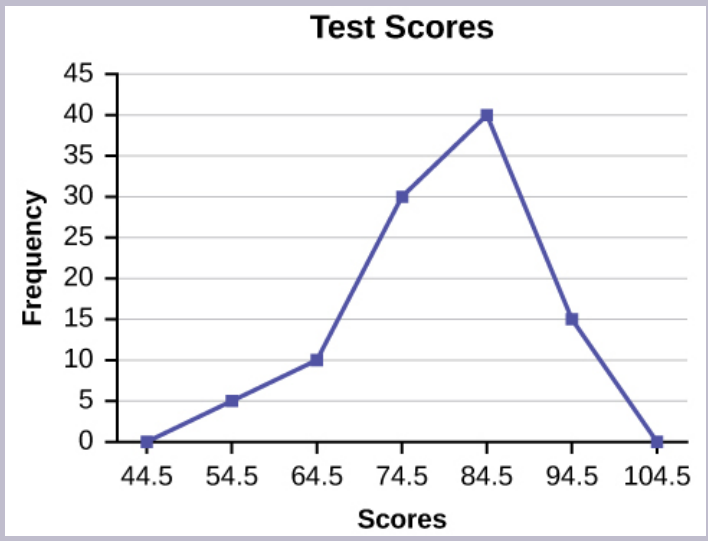
Frequency polygons are analogous to line graphs, and just as line graphs make continuous data visually easy to interpret, so too do frequency polygons.

To construct a **frequency polygon**, first examine the data and decide on the number of intervals, or class intervals, to use on the *x*-axis and *y*-axis. After choosing the appropriate ranges, begin plotting the data points. After all the points are plotted, draw line segments to connect them.

A frequency polygon was constructed from the frequency table below.

Frequency Distribution for Calculus Final Test Scores

Lower Bound	Upper Bound	Frequency	Cumulative Frequency
49.5	59.5	5	5
59.5	69.5	10	15
69.5	79.5	30	45
79.5	89.5	40	85
89.5	99.5	15	100



The first label on the *x*-axis is 44.5. This represents an interval extending from 39.5 to 49.5. Since the lowest test score is 54.5, this interval is used only to allow the graph to touch the *x*-axis. The point labeled 54.5 represents the next interval, or the first "real" interval from the table, and contains five scores. This reasoning is followed for each of the remaining intervals with the point 104.5 representing the interval from 99.5 to 109.5. Again, this interval contains no data and is only

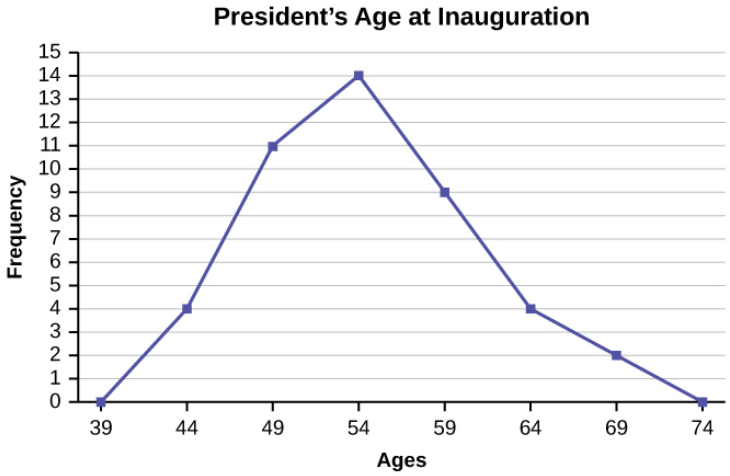
used so that the graph will touch the *x*-axis. Looking at the graph, we say that this distribution is skewed because one side of the graph does not mirror the other side.

TRY IT

Construct a frequency polygon of U.S. Presidents' ages at inauguration shown in [link].

Age at Inauguration	Frequency
41.5–46.5	4
46.5–51.5	11
51.5–56.5	14
56.5–61.5	9
61.5–66.5	4
66.5–71.5	2

The first label on the *x*-axis is 39. This represents an interval extending from 36.5 to 41.5. Since there are no ages less than 41.5, this interval is used only to allow the graph to touch the *x*-axis. The point labeled 44 represents the next interval, or the first "real" interval from the table, and contains four scores. This reasoning is followed for each of the remaining intervals with the point 74 representing the interval from 71.5 to 76.5. Again, this interval contains no data and is only used so that the graph will touch the *x*-axis. Looking at the graph, we say that this distribution is skewed because one side of the graph does not mirror the other side.



Frequency polygons are useful for comparing distributions. This is achieved by overlaying the frequency polygons drawn for different data sets.

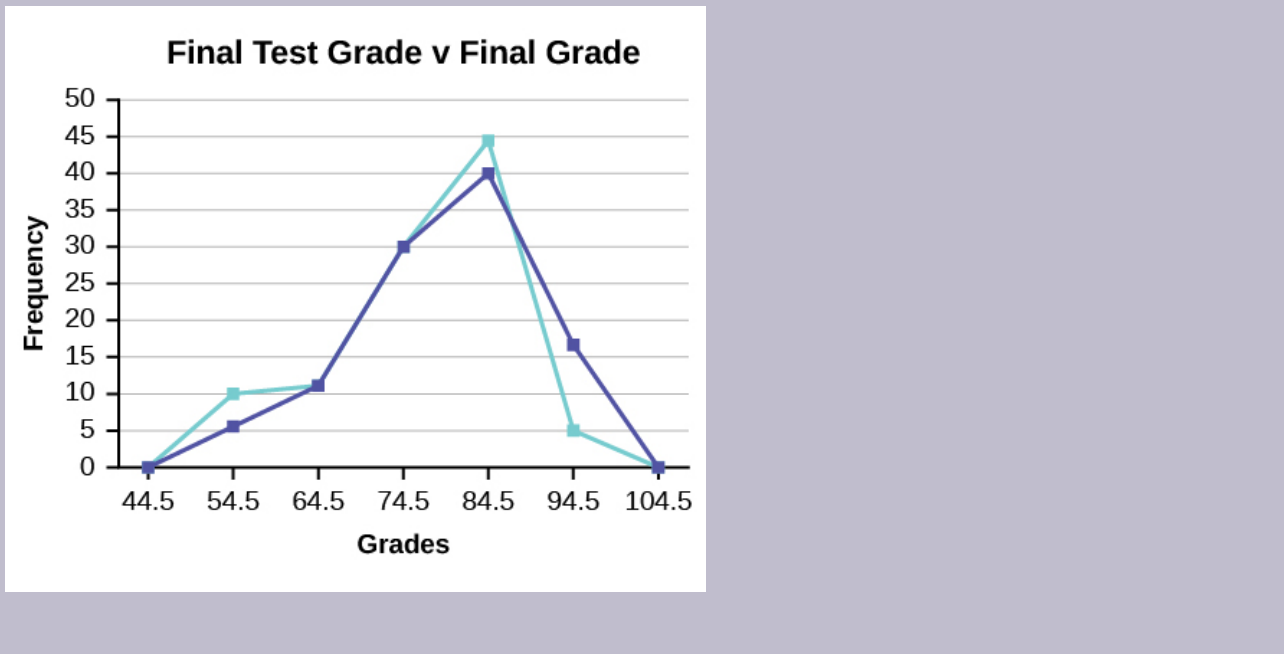
We will construct an overlay frequency polygon comparing the scores from the data below with the students' final numeric grade.

Frequency Distribution for Calculus Final Test Scores

Lower Bound	Upper Bound	Frequency	Cumulative Frequency
49.5	59.5	5	5
59.5	69.5	10	15
69.5	79.5	30	45
79.5	89.5	40	85
89.5	99.5	15	100

Frequency Distribution for Calculus Final Grades

Lower Bound	Upper Bound	Frequency	Cumulative Frequency
49.5	59.5	10	10
59.5	69.5	10	20
69.5	79.5	30	50
79.5	89.5	45	95
89.5	99.5	5	100



Suppose that we want to study the temperature range of a region for an entire month. Every day at noon we note the temperature and write this down in a log. A variety of statistical studies could be done with this data. We could find the mean or the median temperature for the month. We could construct a histogram displaying the number of days that temperatures reach a certain range of values. However, all of these methods ignore a portion of the data that we have collected.

One feature of the data that we may want to consider is that of time. Since each date is paired with the temperature reading for the day, we don't have to think of the data as being random. We can instead use the times given to impose a chronological order on the data. A graph that recognizes this ordering and displays the changing temperature as the month progresses is called a time series graph.

Constructing a Time Series Graph

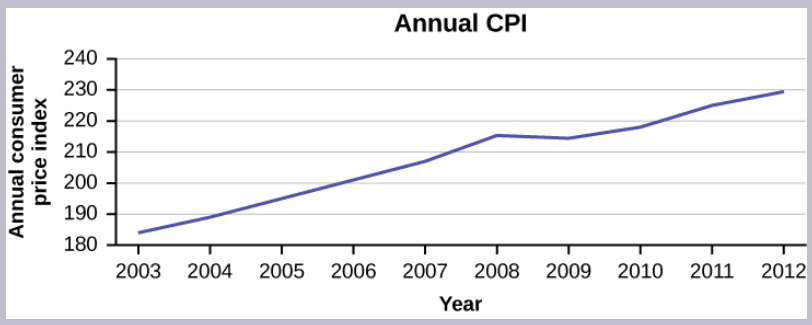
To construct a time series graph, we must look at both pieces of our paired data set. We start with a standard Cartesian coordinate system. The horizontal axis is used to plot the date or time increments, and the vertical axis is used to plot the values of the variable that we are measuring. By doing this, we make each point on the graph correspond to a date and a measured quantity. The points on the graph are typically connected by straight lines in the order in which they occur.

The following data shows the Annual Consumer Price Index, each month, for ten years. Construct a time series graph for the Annual Consumer Price Index data only.

Year	Jan	Feb	Mar	Apr	May	Jun	Jul
2003	181.7	183.1	184.2	183.8	183.5	183.7	183.9
2004	185.2	186.2	187.4	188.0	189.1	189.7	189.4
2005	190.7	191.8	193.3	194.6	194.4	194.5	195.4
2006	198.3	198.7	199.8	201.5	202.5	202.9	203.5
2007	202.416	203.499	205.352	206.686	207.949	208.352	208.299
2008	211.080	211.693	213.528	214.823	216.632	218.815	219.964
2009	211.143	212.193	212.709	213.240	213.856	215.693	215.351
2010	216.687	216.741	217.631	218.009	218.178	217.965	218.011
2011	220.223	221.309	223.467	224.906	225.964	225.722	225.922
2012	226.665	227.663	229.392	230.085	229.815	229.478	229.104

Year	Aug	Sep	Oct	Nov	Dec	Annual
2003	184.6	185.2	185.0	184.5	184.3	184.0
2004	189.5	189.9	190.9	191.0	190.3	188.9
2005	196.4	198.8	199.2	197.6	196.8	195.3
2006	203.9	202.9	201.8	201.5	201.8	201.6
2007	207.917	208.490	208.936	210.177	210.036	207.342
2008	219.086	218.783	216.573	212.425	210.228	215.303
2009	215.834	215.969	216.177	216.330	215.949	214.537
2010	218.312	218.439	218.711	218.803	219.179	218.056
2011	226.545	226.889	226.421	226.230	225.672	224.939
2012	230.379	231.407	231.317	230.221	229.601	229.594

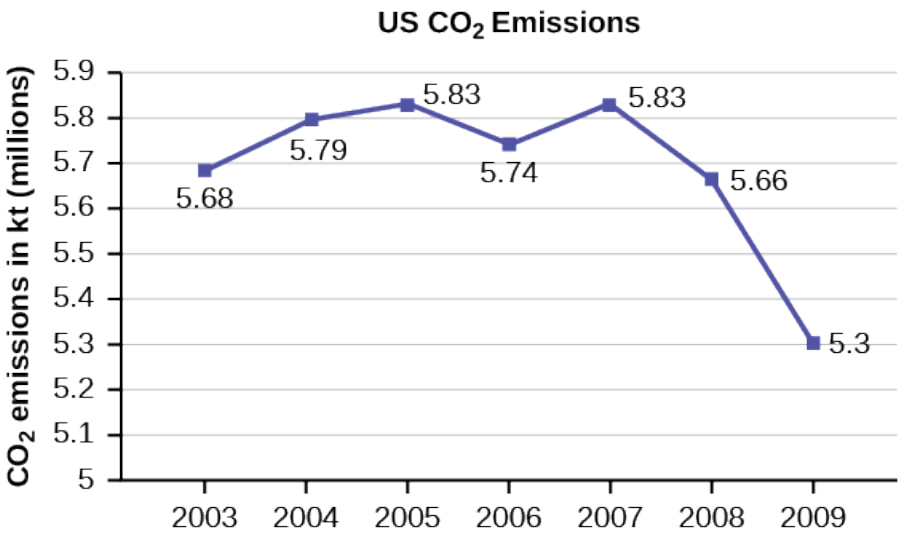
Solution



TRY IT

The following table is a portion of a data set from www.worldbank.org. Use the table to construct a time series graph for CO_2 emissions for the United States.

CO_2 Emissions			
	Ukraine	**United Kingdom**	**United States**
2003	352,259	540,640	5,681,664
2004	343,121	540,409	5,790,761
2005	339,029	541,990	5,826,394
2006	327,797	542,045	5,737,615
2007	328,357	528,631	5,828,697
2008	323,657	522,247	5,656,839
2009	272,176	474,579	5,299,563



Uses of a Time Series Graph

Time series graphs are important tools in various applications of statistics. When recording values of the same variable over an extended period of time, sometimes it is difficult to discern any trend or pattern. However, once the same data points are displayed graphically, some features jump out. Time series graphs make trends easy to spot.

References

Data on annual homicides in Detroit, 1961–73, from Gunst & Mason's book 'Regression Analysis and its Application', Marcel Dekker

"Timeline: Guide to the U.S. Presidents: Information on every president's birthplace, political party, term of office, and more." Scholastic, 2013. Available online at http://www.scholastic.com/teachers/article/timeline-guide-us-presidents (accessed April 3, 2013).

"Presidents." Fact Monster. Pearson Education, 2007. Available online at http://www.factmonster.com/ipka/A0194030.html (accessed April 3, 2013).

"Food Security Statistics." Food and Agriculture Organization of the United Nations. Available online at http://www.fao.org/economic/ess/ess-fs/en/ (accessed April 3, 2013).

"Consumer Price Index." United States Department of Labor: Bureau of Labor Statistics. Available online at http://data.bls.gov/pdq/SurveyOutputServlet (accessed April 3, 2013).

"CO_2 emissions (kt)." The World Bank, 2013. Available online at http://databank.worldbank.org/data/home.aspx (accessed April 3, 2013).

"Births Time Series Data." General Register Office For Scotland, 2013. Available online at http://www.gro-scotland.gov.uk/statistics/theme/vital-events/births/time-series.html (accessed April 3, 2013).

"Demographics: Children under the age of 5 years underweight." Indexmundi. Available online at http://www.indexmundi.com/g/r.aspx?t=50&v=2224&aml=en (accessed April 3, 2013).

Gunst, Richard, Robert Mason. *Regression Analysis and Its Application: A Data-Oriented Approach*. CRC Press: 1980.

"Overweight and Obesity: Adult Obesity Facts." Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/obesity/data/adult.html (accessed September 13, 2013).

2.3 MEASURES OF THE LOCATION OF THE DATA

Learning Objectives

By the end of this section, the student should be able to:

- Recognize, describe, and calculate the measures of location of data: quartiles and percentiles.
- Compute the five-number summary.

The common measures of location are **quartiles** and **percentiles**

Quartiles are special percentiles. The first quartile, Q_1, is the same as the 25^{th} percentile, and the third quartile, Q_3, is the same as the 75^{th} percentile. The median, M, is called both the second quartile and the 50^{th} percentile.

To calculate quartiles and percentiles, the data must be ordered from smallest to largest. Quartiles divide ordered data into quarters. Percentiles divide ordered data into hundredths. To score in the 90^{th} percentile of an exam does not mean, necessarily, that you received 90% on a test. It means that 90% of test scores are the same or less than your score and 10% of the test scores are the same or greater than your test score.

Percentiles are useful for comparing values. For this reason, universities and colleges use percentiles extensively. One instance in which colleges and universities use percentiles is when SAT results are used to determine a minimum testing score that will be used as an acceptance factor. For example, suppose Duke accepts SAT scores at or above the 75^{th} percentile. That translates into a score of at least 1220.

Percentiles are mostly used with very large populations. Therefore, if you were to say that 90% of the test scores are less (and not the same or less) than your score, it would be acceptable because removing one particular data value is not significant.

The median is a number that measures the "center" of the data. You can think of the median as the "middle value," but it does not actually have to be one of the observed values. It is a number that separates ordered data

into halves. Half the values are the same number or smaller than the median, and half the values are the same number or larger. For example, consider the following data.

1; 11.5; 6; 7.2; 4; 8; 9; 10; 6.8; 8.3; 2; 2; 10; 1

Ordered from smallest to largest:

1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

Since there are 14 observations, the median is between the seventh value, 6.8, and the eighth value, 7.2. To find the median, add the two values together and divide by two.

$$\frac{6.8 + 7.2}{2} = 7$$

The median is seven. Half of the values are smaller than seven and half of the values are larger than seven.

Quartiles are numbers that separate the data into quarters. Quartiles may or may not be part of the data. To find the quartiles, first find the median or second quartile. The first quartile, Q_1, is the middle value of the lower half of the data, and the third quartile, Q_3, is the middle value, or median, of the upper half of the data. To get the idea, consider the same data set:

1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

The median or **second quartile** is seven. The lower half of the data are 1, 1, 2, 2, 4, 6, 6.8. The middle value of the lower half is two.

1; 1; 2; 2; 4; 6; 6.8

The number two, which is part of the data, is the first quartile. One-fourth of the entire sets of values are the same as or less than two and three-fourths of the values are more than two.

The upper half of the data is 7.2, 8, 8.3, 9, 10, 10, 11.5. The middle value of the upper half is nine.

The third quartile, $Q3$, is nine. Three-fourths (75%) of the ordered data set are less than nine. One-fourth (25%) of the ordered data set are greater than nine. The third quartile is part of the data set in this example.

The **interquartile range** is a number that indicates the spread of the middle half or the middle 50% of the data. It is the difference between the third quartile (Q_3) and the first quartile (Q_1).

$IQR = Q_3 - Q_1$

The IQR can help to determine potential **outliers**. **A value is suspected to be a potential outlier if it is less than (1.5)(IQR) below the first quartile or more than (1.5)(IQR) above the third quartile**. Potential outliers always require further investigation.

NOTE

A potential outlier is a data point that is significantly different from the other data points. These special data points may be errors or some kind of abnormality or they may be a key to understanding the data.

For the following 13 real estate prices, calculate the *IQR* and determine if any prices are potential outliers. Prices are in dollars.

389,950; 230,500; 158,000; 479,000; 639,000; 114,950; 5,500,000; 387,000; 659,000; 529,000; 575,000; 488,800; 1,095,000

Solution
Order the data from smallest to largest.

114,950; 158,000; 230,500; 387,000; 389,950; 479,000; 488,800; 529,000; 575,000; 639,000; 659,000; 1,095,000; 5,500,000

M = 488,800

$Q_1 = \frac{230{,}500 + 387{,}000}{2}$ = 308,750

$Q_3 = \frac{639{,}000 + 659{,}000}{2}$ = 649,000

IQR = 649,000 – 308,750 = 340,250

(1.5)(IQR) = (1.5)(340,250) = 510,375

Q_1 – (1.5)(IQR) = 308,750 – 510,375 = –201,625

Q_3 + (1.5)(IQR) = 649,000 + 510,375 = 1,159,375

No house price is less than –201,625. However, 5,500,000 is more than 1,159,375. Therefore, 5,500,000 is a potential outlier.

TRY IT

For the following 11 salaries, calculate the *IQR* and determine if any salaries are outliers. The salaries are in dollars.

$33,000; $64,500; $28,000; $54,000; $72,000; $68,500; $69,000; $42,000; $54,000; $120,000; $40,500

Solution

Order the data from smallest to largest.

$28,000; $33,000; $40,500; $42,000; $54,000; $54,000; $64,500; $68,500; $69,000; $72,000; $120,000

Median = $54,000

Q_1 = $40,500

Q_3 = $69,000

IQR = $69,000 – $40,500 = $28,500

(1.5)(IQR) = (1.5)($28,500) = $42,750

Q_1 – (1.5)(IQR) = $40,500 – $42,750 = –$2,250

Q_3 + (1.5)(IQR) = $69,000 + $42,750 = $111,750

No salary is less than –$2,250. However, $120,000 is more than $11,750, so $120,000 is a potential outlier.

For the two data sets in the test scores example, find the following:

1. The interquartile range. Compare the two interquartile ranges.
2. Any outliers in either set.

Solution

The five number summary for the day and night classes is

	Minimum	Q_1	Median	Q_3	Maximum
Day	32	56	74.5	82.5	99
Night	25.5	78	81	89	98

1. The IQR for the day group is $Q_3 – Q_1$ = 82.5 – 56 = 26.5
   The IQR for the night group is $Q_3 – Q_1$ = 89 – 78 = 11

   The interquartile range (the spread or variability) for the day class is larger than the night class *IQR*. This suggests more variation will be found in the day class's class test scores.
2. Day class outliers are found using the IQR times 1.5 rule. So,
   - $Q_1 – IQR(1.5)$ = 56 – 26.5(1.5) = 16.25
   - $Q_3 + IQR(1.5)$ = 82.5 + 26.5(1.5) = 122.25

Since the minimum and maximum values for the day class are greater than 16.25 and less than 122.25, there are no outliers.

Night class outliers are calculated as:

- $Q_1 - IQR\,(1.5) = 78 - 11(1.5) = 61.5$
- $Q_3 + \text{IQR}(1.5) = 89 + 11(1.5) = 105.5$

For this class, any test score less than 61.5 is an outlier. Therefore, the scores of 45 and 25.5 are outliers. Since no test score is greater than 105.5, there is no upper end outlier.

TRY IT

Find the interquartile range for the following two data sets and compare them.

Test Scores for Class *A*

69; 96; 81; 79; 65; 76; 83; 99; 89; 67; 90; 77; 85; 98; 66; 91; 77; 69; 80; 94

Test Scores for Class *B*

90; 72; 80; 92; 90; 97; 92; 75; 79; 68; 70; 80; 99; 95; 78; 73; 71; 68; 95; 100

Solution

Class *A*

Order the data from smallest to largest.

65; 66; 67; 69; 69; 76; 77; 77; 79; 80; 81; 83; 85; 89; 90; 91; 94; 96; 98; 99

$$\text{Median} = \frac{80 + 81}{2} = 80.5$$

$$Q_1 = \frac{69 + 76}{2} = 72.5$$

$$Q_3 = \frac{90 + 91}{2} = 90.5$$

$IQR = 90.5 - 72.5 = 18$

Class *B*

Order the data from smallest to largest.

68; 68; 70; 71; 72; 73; 75; 78; 79; 80; 80; 90; 90; 92; 92; 95; 95; 97; 99; 100

$$\text{Median} = \frac{80 + 80}{2} = 80$$

$$Q_1 = \frac{72 + 73}{2} = 72.5$$

$$Q_3 = \frac{92 + 95}{2} = 93.5$$

$IQR = 93.5 - 72.5 = 21$

The data for Class *B* has a larger IQR, so the scores between Q_3 and Q_1 (middle 50%) for the data for Class *B* are more spread out and not clustered about the median.

Fifty statistics students were asked how much sleep they get per school night (rounded to the nearest hour). The results were:

AMOUNT OF SLEEP PER SCHOOL NIGHT (HOURS)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
4	2	0.04	0.04
5	5	0.10	0.14
6	7	0.14	0.28
7	12	0.24	0.52
8	14	0.28	0.80
9	7	0.14	0.94
10	3	0.06	1.00

Find the 28th percentile. Notice the 0.28 in the "cumulative relative frequency" column. Twenty-eight percent of 50 data values is 14 values. There are 14 values less than the 28th percentile. They include the two 4s, the five 5s, and the seven 6s. The 28th percentile is between the last six and the first seven. **The 28th percentile is 6.5.**

Find the median. Look again at the "cumulative relative frequency" column and find 0.52. The median is the 50th percentile or the second quartile. 50% of 50 is 25. There are 25 values less than the median. They include the two 4s, the five 5s, the seven 6s, and eleven of the 7s. The median or 50th percentile is between the 25th, or seven, and 26th, or seven, values. **The median is seven.**

Find the third quartile. The third quartile is the same as the 75th percentile. You can "eyeball" this answer. If you look at the "cumulative relative frequency" column, you find 0.52 and 0.80. When

you have all the fours, fives, sixes and sevens, you have 52% of the data. When you include all the 8s, you have 80% of the data. **The 75th percentile, then, must be an eight**. Another way to look at the problem is to find 75% of 50, which is 37.5, and round up to 38. The third quartile, Q_3, is the 38th value, which is an eight. You can check this answer by counting the values. (There are 37 values below the third quartile and 12 values above.)

TRY IT

Forty bus drivers were asked how many hours they spend each day running their routes (rounded to the nearest hour). Find the 65th percentile.

Amount of time spent on route (hours)	Frequency	Relative Frequency	Cumulative Relative Frequency
2	12	0.30	0.30
3	14	0.35	0.65
4	10	0.25	0.90
5	4	0.10	1.00

Solution

The 65th percentile is between the last three and the first four.

The 65th percentile is 3.5.

Using [link]:

1. Find the 80th percentile.
2. Find the 90th percentile.
3. Find the first quartile. What is another name for the first quartile?

Solution

Using the data from the frequency table, we have:

1. The 80th percentile is between the last eight and the first nine in the table (between the 40th and 41st values). Therefore, we need to take the mean of the 40th and 41st values. The 80th percentile $= \frac{8+9}{2} = 8.5$
2. The 90th percentile will be the 45th data value (location is 0.90(50) = 45) and the 45th data value is nine.
3. Q_1 is also the 25th percentile. The 25th percentile location calculation: P_{25} = 0.25(50) = 12.5 ≈ 13 the 13th data value. Thus, the 25th percentile is six.

TRY IT

Refer to the [link]. Find the third quartile. What is another name for the third quartile?
Solution

The third quartile is the 75th percentile, which is four. The 65th percentile is between three and four, and the 90th percentile is between four and 5.75. The third quartile is between 65 and 90, so it must be four.

A Formula for Finding the *k*th Percentile

If you were to do a little research, you would find several formulas for calculating the k^{th} percentile. Here is one of them.

k = the k^{th} percentile. It may or may not be part of the data.
i = the index (ranking or position of a data value)
n = the total number of data

- Order the data from smallest to largest.
- Calculate $i = \frac{k}{100}(n+1)$

- If i is an integer, then the k^{th} percentile is the data value in the i^{th} position in the ordered set of data.
- If i is not an integer, then round i up and round i down to the nearest integers. Average the two data values in these two positions in the ordered data set. This is easier to understand in an example.

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

1. Find the 70^{th} percentile.
2. Find the 83^{rd} percentile.

Solution

1.
   - $k = 70$
   - i = the index
   - $n = 29$

   $i = \frac{k}{100}(n + 1) = (\frac{70}{100})(29 + 1) = 21$. Twenty-one is an integer, and the data value in the 21^{st} position in the ordered data set is 64. The 70^{th} percentile is 64 years.

   - $k = 83^{rd}$ percentile
   - i = the index
   - $n = 29$

2. $i = \frac{k}{100}(n + 1) =)\frac{83}{100})(29 + 1) = 24.9$, which is NOT an integer. Round it down to 24 and up to 25. The age in the 24^{th} position is 71 and the age in the 25^{th} position is 72. Average 71 and 72. The 83^{rd} percentile is 71.5 years.

TRY IT

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

Calculate the 20th percentile and the 55th percentile.

Solution

$k = 20$. Index = $i = \frac{k}{100}(n+1) = \frac{20}{100}(29+1) = 6$. The age in the sixth position is 27. The 20th percentile is 27 years.

$k = 55$. Index = $i = \frac{k}{100}(n+1) = \frac{55}{100}(29+1) = 16.5$. Round down to 16 and up to 17. The age in the 16th position is 52 and the age in the 17th position is 55. The average of 52 and 55 is 53.5. The 55th percentile is 53.5 years.

NOTE

You can calculate percentiles using calculators and computers. There are a variety of online calculators.

A Formula for Finding the Percentile of a Value in a Data Set

- Order the data from smallest to largest.
- x = the number of data values counting from the bottom of the data list up to but not including the data value for which you want to find the percentile.
- y = the number of data values equal to the data value for which you want to find the percentile.
- n = the total number of data.
- Calculate $\frac{x + 0.5y}{n}(100)$. Then round to the nearest integer.

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

1. Find the percentile for 58.

2. Find the percentile for 25.

Solution

1. Counting from the bottom of the list, there are 18 data values less than 58. There is one value of 58.

   $x = 18$ and $y = 1$. $\frac{x + 0.5y}{n}(100) = \frac{18 + 0.5\,(1)}{29}(100) = 63.80$. 58 is the 64^{th} percentile.

2. Counting from the bottom of the list, there are three data values less than 25. There is one value of 25.

   $x = 3$ and $y = 1$. $\frac{x + 0.5y}{n}(100) = \frac{3 + 0.5\,(1)}{29}(100) = 12.07$. Twenty-five is the 12^{th} percentile.

TRY IT

Listed are 30 ages for Academy Award winning best actors in order from smallest to largest.

18; 21; 22; 25; 26; 27; 29; 30; 31; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

Find the percentiles for 47 and 31.

Solution

Percentile for 47: Counting from the bottom of the list, there are 15 data values less than 47. There is one value of 47.

$x = 15$ and $y = 1$. $\frac{x + 0.5y}{n}(100) = \frac{15 + 0.5\,(1)}{30}(100) = 51.89$. 47 is the 52^{nd} percentile.

Percentile for 31: Counting from the bottom of the list, there are eight data values less than 31. There are two values of 31.

$x = 8$ and $y = 2$. $\frac{x + 0.5y}{n}(100) = \frac{8 + 0.5\,(2)}{30}(100) = 27$. 31 is the 27^{th} percentile.

Interpreting Percentiles, Quartiles, and Median

A percentile indicates the relative standing of a data value when data are sorted into numerical order from smallest to largest. Percentages of data values are less than or equal to the pth percentile. For example, 15% of data values are less than or equal to the 15^{th} percentile.

- Low percentiles always correspond to lower data values.
- High percentiles always correspond to higher data values.

A percentile may or may not correspond to a value judgment about whether it is "good" or "bad." The interpretation of whether a certain percentile is "good" or "bad" depends on the context of the situation to which the data applies. In some situations, a low percentile would be considered "good"; in other contexts a high percentile might be considered "good." In many situations, there is no value judgment that applies.

Understanding how to interpret percentiles properly is important not only when describing data, but also when calculating probabilities in later chapters of this text.

GUIDELINE

When writing the interpretation of a percentile in the context of the given data, the sentence should contain the following information.

- information about the context of the situation being considered
- the data value (value of the variable) that represents the percentile
- the percent of individuals or items with data values below the percentile
- the percent of individuals or items with data values above the percentile.

On a timed math test, the first quartile for time it took to finish the exam was 35 minutes. Interpret the first quartile in the context of this situation.

Solution

- Twenty-five percent of students finished the exam in 35 minutes or less.
- Seventy-five percent of students finished the exam in 35 minutes or more.
- A low percentile could be considered good, as finishing more quickly on a timed exam is desirable. (If you take too long, you might not be able to finish.)

TRY IT

For the 100-meter dash, the third quartile for times for finishing the race was 11.5 seconds. Interpret the third quartile in the context of the situation.

Solution

Twenty-five percent of runners finished the race in 11.5 seconds or more. Seventy-five percent of runners finished the race in 11.5 seconds or less. A lower percentile is good because finishing a race more quickly is desirable.

On a 20 question math test, the 70^{th} percentile for the number of correct answers was 16. Interpret the 70^{th} percentile in the context of this situation.

Solution

- Seventy percent of students answered 16 or fewer questions correctly.
- Thirty percent of students answered 16 or more questions correctly.
- A higher percentile could be considered good, as answering more questions correctly is desirable.

TRY IT

On a 60 point written assignment, the 80^{th} percentile for the number of points earned was 49. Interpret the 80^{th} percentile in the context of this situation.
Solution

Eighty percent of students earned 49 points or fewer. Twenty percent of students earned 49 or more points. A higher percentile is good because getting more points on an assignment is desirable.

At a community college, it was found that the 30^{th} percentile of credit units that students are enrolled for is seven units. Interpret the 30^{th} percentile in the context of this situation.

Solution

- Thirty percent of students are enrolled in seven or fewer credit units.
- Seventy percent of students are enrolled in seven or more credit units.
- In this example, there is no "good" or "bad" value judgment associated with a higher or lower percentile. Students attend community college for varied reasons and needs, and their course load varies according to their needs.

TRY IT

During a season, the 40^{th} percentile for points scored per player in a game is eight. Interpret the 40^{th} percentile in the context of this situation.
Solution

Forty percent of players scored eight points or fewer. Sixty percent of players scored eight points or more. A higher percentile is good because getting more points in a basketball game is desirable.

Sharpe Middle School is applying for a grant that will be used to add fitness equipment to the gym. The principal surveyed 15 anonymous students to determine how many minutes a day the students spend exercising. The results from the 15 anonymous students are shown.

0 minutes; 40 minutes; 60 minutes; 30 minutes; 60 minutes

10 minutes; 45 minutes; 30 minutes; 300 minutes; 90 minutes;

30 minutes; 120 minutes; 60 minutes; 0 minutes; 20 minutes

Determine the following five values.

- Min = 0
- $Q_1 = 20$
- Med = 40
- $Q_3 = 60$
- Max = 300

If you were the principal, would you be justified in purchasing new fitness equipment? Since 75% of the students exercise for 60 minutes or less daily, and since the *IQR* is 40 minutes (60 – 20 = 40), we know that half of the students surveyed exercise between 20 minutes and 60 minutes daily. This seems a reasonable amount of time spent exercising, so the principal would be justified in purchasing the new equipment.

However, the principal needs to be careful. The value 300 appears to be a potential outlier.

$Q_3 + 1.5(IQR) = 60 + (1.5)(40) = 120$.

The value 300 is greater than 120 so it is a potential outlier. If we delete it and calculate the five values, we get the following values:

- Min = 0
- $Q_1 = 20$
- Med=35
- $Q_3 = 60$
- Max = 120

We still have 75% of the students exercising for 60 minutes or less daily and half of the students exercising between 20 and 60 minutes a day. However, 15 students is a small sample and the principal should survey more students to be sure of his survey results.

References

Cauchon, Dennis, Paul Overberg. "Census data shows minorities now a majority of U.S. births." USA Today, 2012. Available online at http://usatoday30.usatoday.com/news/nation/story/2012-05-17/minority-birthscensus/55029100/1 (accessed April 3, 2013).

Data from the United States Department of Commerce: United States Census Bureau. Available online at http://www.census.gov/ (accessed April 3, 2013).

"1990 Census." United States Department of Commerce: United States Census Bureau. Available online at http://www.census.gov/main/www/cen1990.html (accessed April 3, 2013).

Data from *San Jose Mercury News*.

Data from *Time Magazine*; survey by Yankelovich Partners, Inc.

2.4 BOX PLOTS

Learning Objectives

By the end of this section, the student should be able to:

- Display data graphically and interpret box plots.

Box plots (also called box-and-whisker plots or box-whisker plots) give a good graphical image of the concentration of the data. They also show how far the extreme values are from most of the data. A **box plot** is constructed from five values: the minimum value, the first quartile, the median, the third quartile, and the maximum value. We use these values to compare how close other data values are to them.

To construct a box plot, use a horizontal or vertical number line and a rectangular box. The smallest and largest data values label the endpoints of the axis. The **first quartile** marks one end of the box and the third quartile marks the other end of the box. Approximately **the middle 50 percent of the data fall inside the box.** The "whiskers" extend from the ends of the box to the smallest and largest data values. The median or second quartile can be between the first and third quartiles, or it can be one, or the other, or both. The box plot gives a good, quick picture of the data.

NOTE

You may encounter box-and-whisker plots that have dots marking outlier values. In those cases, the whiskers are not extending to the minimum and maximum values.

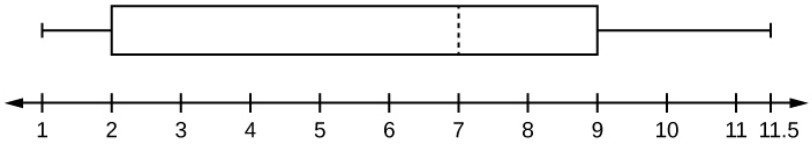
Consider, again, this dataset.

1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

The first quartile is two, the median is seven, and the third quartile is nine. The smallest value is one, and the largest value is 11.5. The following image shows the constructed box plot.

NOTE

See the calculator instructions on the TI web site or in the appendix.



The two whiskers extend from the first quartile to the smallest value and from the third quartile to the largest value. The median is shown with a dashed line.

NOTE

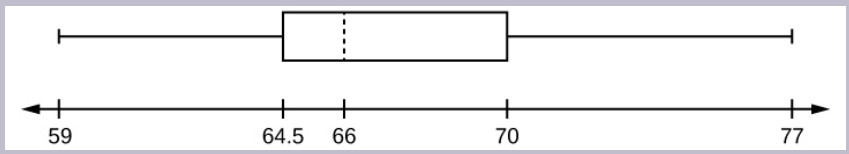
It is important to start a box plot with a **scaled number line**. Otherwise the box plot may not be useful.

The following data are the heights of 40 students in a statistics class.

59; 60; 61; 62; 62; 63; 63; 64; 64; 64; 65; 65; 65; 65; 65; 65; 65; 65; 65; 66;
66; 67; 67; 68; 68; 69; 70; 70; 70; 70; 70; 71; 71; 72; 72; 73; 74; 74; 75; 77

Construct a box plot with the following properties; the calculator instructions for the minimum and maximum values as well as the quartiles follow the example.

- Minimum value = 59
- Maximum value = 77
- $Q1$: First quartile = 64.5
- $Q2$: Second quartile or median= 66
- $Q3$: Third quartile = 70



1. Each quarter has approximately 25% of the data.
2. The spreads of the four quarters are 64.5 – 59 = 5.5 (first quarter), 66 – 64.5 = 1.5 (second quarter), 70 – 66 = 4 (third quarter), and 77 – 70 = 7 (fourth quarter). So, the second quarter has the smallest spread and the fourth quarter has the largest spread.
3. Range = maximum value – the minimum value = 77 – 59 = 18
4. Interquartile Range: $IQR = Q3 - Q1 = 70 - 64.5 = 5.5$.
5. The interval 59–65 has more than 25% of the data so it has more data in it than the interval 66 through 70 which has 25% of the data.
6. The middle 50% (middle half) of the data has a range of 5.5 inches.

To find the minimum, maximum, and quartiles:

Enter data into the list editor (Press STAT 1:EDIT). If you need to clear the list, arrow up to the name L1, press CLEAR, and then arrow down.

Put the data values into the list L1.

Press STAT and arrow to CALC. Press 1:1-VarStats. Enter L1.

Press ENTER.

Use the down and up arrow keys to scroll.

Smallest value = 59.

Largest value = 77.

Q_1: First quartile = 64.5.

Q_2: Second quartile or median = 66.

Q_3: Third quartile = 70.

To construct the box plot:

Press 4:Plotsoff. Press ENTER.

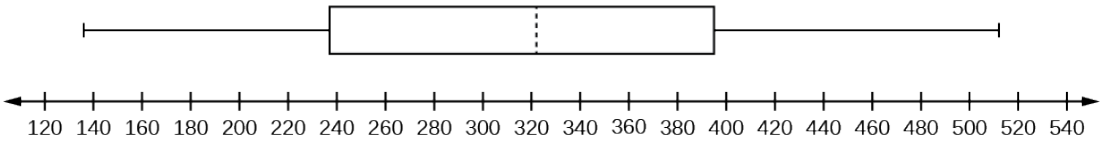
Arrow down and then use the right arrow key to go to the fifth picture, which is the box plot. Press ENTER.

Arrow down to Xlist: Press 2nd 1 for L1
Arrow down to Freq: Press ALPHA. Press 1.
Press Zoom. Press 9: ZoomStat.
Press TRACE, and use the arrow keys to examine the box plot.

TRY IT

The following data are the number of pages in 40 books on a shelf. Construct a box plot using a graphing calculator, and state the interquartile range.

136; 140; 178; 190; 205; 215; 217; 218; 232; 234; 240; 255; 270; 275; 290; 301; 303; 315; 317; 318; 326; 333; 343; 349; 360; 369; 377; 388; 391; 392; 398; 400; 402; 405; 408; 422; 429; 450; 475; 512
Solution



$IQR = 158$

For some sets of data, some of the largest value, smallest value, first quartile, median, and third quartile may be the same. For instance, you might have a data set in which the median and the third quartile are the same. In this case, the diagram would not have a dotted line inside the box displaying the median. The right side of the box would display both the third quartile and the median. For example, if the smallest value and the first quartile were both one, the median and the third quartile were both five, and the largest value was seven, the box plot would look like:



In this case, at least 25% of the values are equal to one. Twenty-five percent of the values are between one and

five, inclusive. At least 25% of the values are equal to five. The top 25% of the values fall between five and seven, inclusive.

Test scores for a college statistics class held during the day are:

99; 56; 78; 55.5; 32; 90; 80; 81; 56; 59; 45; 77; 84.5; 84; 70; 72; 68; 32; 79; 90

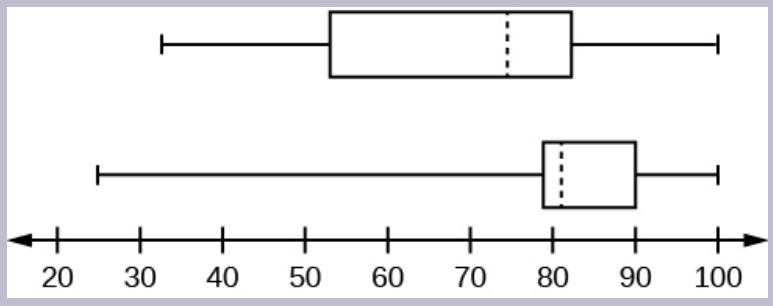
Test scores for a college statistics class held during the evening are:

98; 78; 68; 83; 81; 89; 88; 76; 65; 45; 98; 90; 80; 84.5; 85; 79; 78; 98; 90; 79; 81; 25.5

1. Find the smallest and largest values, the median, and the first and third quartile for the day class.
2. Find the smallest and largest values, the median, and the first and third quartile for the night class.
3. For each data set, what percentage of the data is between the smallest value and the first quartile? the first quartile and the median? the median and the third quartile? the third quartile and the largest value? What percentage of the data is between the first quartile and the largest value?
4. Create a box plot for each set of data. Use one number line for both box plots.
5. Which box plot has the widest spread for the middle 50% of the data (the data between the first and third quartiles)? What does this mean for that set of data in comparison to the other set of data?

Solution

1. The values are given below:
   Min = 32; Q_1 = 56; Median = 74.5; Q_3 = 82.5; Max = 99
2. The values are given below:
   Min = 25.5; Q_1 = 78; Median = 81; Q_3 = 89; Max = 98
3. Day class: There are six data values ranging from 32 to 56: 30%. There are six data values ranging from 56 to 74.5: 30%. There are five data values ranging from 74.5 to 82.5: 25%. There are five data values ranging from 82.5 to 99: 25%. There are 16 data values between the first quartile, 56, and the largest value, 99: 75%. Night class

4. 

5. The first data set has the wider spread for the middle 50% of the data. The *IQR* for the first data set is greater than the *IQR* for the second set. This means that there is more variability in the middle 50% of the first data set.

TRY IT

The following data set shows the heights in inches for the boys in a class of 40 students.

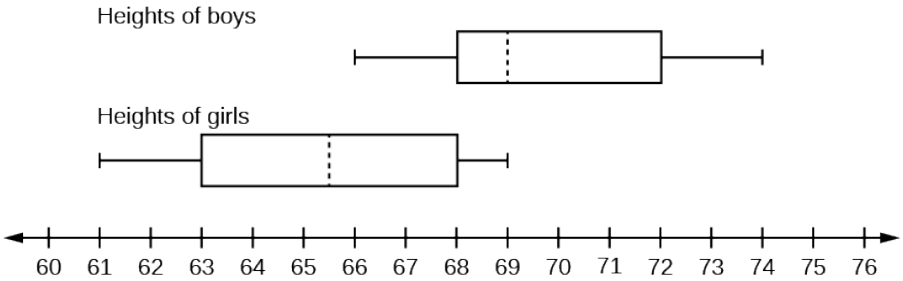
66; 66; 67; 67; 68; 68; 68; 68; 68; 69; 69; 69; 70; 71; 72; 72; 72; 73; 73; 74

The following data set shows the heights in inches for the girls in a class of 40 students.

61; 61; 62; 62; 63; 63; 63; 65; 65; 65; 66; 66; 66; 67; 68; 68; 68; 69; 69; 69

Construct a box plot using a graphing calculator for each data set, and state which box plot has the wider spread for the middle 50% of the data.

Solution



IQR for the boys = 4

IQR for the girls = 5

The box plot for the heights of the girls has the wider spread for the middle 50% of the data.

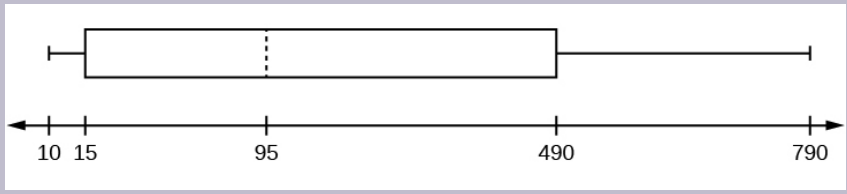
Graph a box-and-whisker plot for the data values shown.

10; 10; 10; 15; 35; 75; 90; 95; 100; 175; 420; 490; 515; 515; 790

The five numbers used to create a box-and-whisker plot are:

- Min: 10
- Q_1: 15
- Med: 95
- Q_3: 490
- Max: 790

The following graph shows the box-and-whisker plot.



TRY IT

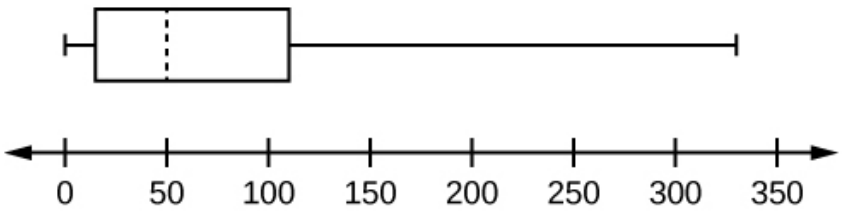
Follow the steps you used to graph a box-and-whisker plot for the data values shown.

0, 5, 5, 15, 30, 30, 45, 50, 50, 60, 75, 110, 140, 240, 330

Solution

The data are in order from least to greatest. There are 15 values, so the eighth number in order is the median: 50. There are seven data values written to the left of the median and 7 values to the right. The five values that are used to create the box plot are:

- Min: 0
- Q_1: 15
- Med: 50
- Q_3: 110
- Max: 330



References

Data from *West Magazine*.

2.5 MEASURES OF THE CENTER OF THE DATA

Learning Objectives

By the end of this section, the student should be able to:

- Recognize, describe, and calculate the measures of the center of data: mean, median, and mode.
- Compute the mean, median, and mode of grouped frequency.

The "center" of a data set is also a way of describing location.

The two most widely used measures of the "center" of the data are the **mean** (average) and the **median**. To calculate the **mean weight** of 50 people, add the 50 weights together and divide by 50. To find the **median weight** of the 50 people, order the data and find the number that splits the data into two equal parts. The median is generally a better measure of the center when there are extreme values or outliers because it is not affected by the precise numerical values of the outliers. The mean is the most common measure of the center.

NOTE

The words "mean" and "average" are often used interchangeably. The substitution of one word for the other is common practice. The technical term is "arithmetic mean" and "average" is technically a center location. However, in practice among non-statisticians, "average" is commonly accepted for "arithmetic mean."

When each value in the data set is not unique, the mean can be calculated by multiplying each distinct value by its frequency and then dividing the sum by the total number of data values. The letter used to represent the **sample mean** is an x with a bar over it (pronounced "x bar"): $\overline{x}$.

The Greek letter μ (pronounced "mew") represents the **population mean**. One of the requirements for the **sample mean** to be a good estimate of the **population mean** is for the sample taken to be truly random.

To see that both ways of calculating the mean are the same, consider the sample:

1; 1; 1; 2; 2; 3; 4; 4; 4; 4; 4

$$\overline{x} = \frac{1+1+1+2+2+3+4+4+4+4+4}{11} = 2.7$$

$$\overline{x} = \frac{3\,(1)+2\,(2)+1\,(3)+5\,(4)}{11} = 2.7$$

In the second example, the frequencies are 3, 2, 1, 5, for the values of 1, 2, 3, and 4, respectively.

You can quickly find the location of the median by using the expression $\frac{n+1}{2}$.

The letter n is the total number of data values in the sample. If n is an odd number, the median is the middle value of the ordered data (ordered smallest to largest). If n is an even number, the median is equal to the two middle values added together and divided by two after the data has been ordered. For example, if the total number of data values is 97, then $\frac{n+1}{2} = \frac{97+1}{2} = 49$. The median is the 49^{th} value in the ordered data. If the total number of data values is 100, then $\frac{n+1}{2} = \frac{100+1}{2} = 50.5$. The median occurs midway between the 50^{th} and 51^{st} values. The location of the median and the value of the median are **not** the same. The upper case letter M is often used to represent the median. The next example illustrates the location of the median and the value of the median.

AIDS data indicating the number of months a patient with AIDS lives after taking a new antibody drug are as follows (smallest to largest):

3; 4; 8; 8; 10; 11; 12; 13; 14; 15; 15; 16; 16; 17; 17; 18; 21; 22; 22; 24; 24; 25; 26; 26; 27; 27; 29; 29; 31; 32; 33; 33; 34; 34; 35; 37; 40; 44; 44; 47;

Calculate the mean and the median.

Solution

The calculation for the mean is:

$$\overline{x} = \frac{[3 + 4 + (8)(2) + 10 + 11 + 12 + 13 + 14 + (15)(2) + (16)(2) + ... + 35 + 37 + 40 + (44)(2) + 47]}{40} = 23.6$$

To find the median, *M*, first use the formula for the location. The location is:

$$\frac{n+1}{2} = \frac{40+1}{2} = 20.5$$

Starting at the smallest value, the median is located between the 20th and 21st values (the two 24s):

3; 4; 8; 8; 10; 11; 12; 13; 14; 15; 15; 16; 16; 17; 17; 18; 21; 22; 22; 24; 24; 25; 26; 26; 27; 27; 29; 29; 31; 32; 33; 33; 34; 34; 35; 37; 40; 44; 44; 47;

$$M = \frac{24 + 24}{2} = 24$$

To find the mean and the median:

Clear list L1. Press STAT 4:ClrList. Enter 2nd 1 for list L1. Press ENTER.

Enter data into the list editor. Press STAT 1:EDIT.

Put the data values into list L1.

Press STAT and arrow to CALC. Press 1:1-VarStats. Press 2nd 1 for L1 and then ENTER.

Press the down and up arrow keys to scroll.

$\overline{x}$ = 23.6, M = 24

TRY IT

The following data show the number of months patients typically wait on a transplant list before getting surgery. The data are ordered from smallest to largest. Calculate the mean and median.

3, 4, 5, 7, 7, 7, 7, 8, 8, 9, 9, 10, 10, 10, 10, 10, 11, 12, 12, 13, 14, 14, 15, 15, 17, 17, 18, 19, 19, 19, 21, 21, 22, 22, 23, 24, 24, 24, 24

Solution

Mean: 3 + 4 + 5 + 7 + 7 + 7 + 7 + 8 + 8 + 9 + 9 + 10 + 10 + 10 + 10 + 10 + 11 + 12 + 12 + 13 + 14 + 14 + 15 + 15 + 17 + 17 + 18 + 19 + 19 + 19 + 21 + 21 + 22 + 22 + 23 + 24 + 24 + 24 = 544

$$\frac{544}{39} = 13.95$$

Median: Starting at the smallest value, the median is the 20th term, which is 13.

Suppose that in a small town of 50 people, one person earns $ 5,000,000 per year and the other 49 each earn $ 30,000. Which is the better measure of the "center": the mean or the median?

Solution

$$\overline{x} = \frac{5,000,000 + 49\,(30,000)}{50} = 129,400$$

M = 30,000

(There are 49 people who earn $ 30,000 and one person who earns $ 5,000,000.)

The median is a better measure of the "center" than the mean because 49 of the values are 30,000 and one is 5,000,000. The 5,000,000 is an outlier. The 30,000 gives us a better sense of the middle of the data.

TRY IT

In a sample of 61 households, one house is worth $2,500,000. Half of the rest are worth $280,000, and all the others are worth $315,000. Which is the better measure of the "center": the mean or the median?
Solution

The median is the better measure of the "center" than the mean because 30 of the values are $280,000, 30 values are $315,000, and one is $2,500,000. The $2,500,000 is an outlier. The median of $315,000 gives us a better sense of the middle of the data.

Another measure of the center is the mode. The mode is the most frequent value. There can be more than one **mode** in a data set as long as those values have the same frequency and that frequency is the highest. A data set with two modes is called bimodal.

Statistics exam scores for 20 students are as follows:

50, 53, 59, 59, 63, 63, 72, 72, 72, 72, 72, 76, 78, 81, 83, 84, 84, 84, 90, 93

Find the mode.

Solution
The most frequent score is 72, which occurs five times. Mode = 72.

TRY IT

The number of books checked out from the library from 25 students are as follows:
0, 0, 0, 1, 2, 3, 3, 4, 4, 5, 5, 7, 7, 7, 7, 8, 8, 8, 9, 10, 10, 11, 11, 12, 12

Find the mode.
Solution

The most frequent number of books is 7, which occurs four times. Mode = 7.

Five real estate exam scores are 430, 430, 480, 480, 495. The data set is bimodal because the scores 430 and 480 each occur twice.

When is the mode the best measure of the "center"? Consider a weight loss program that advertises a mean weight loss of six pounds the first week of the program. The mode might indicate that most people lose two pounds the first week, making the program less appealing.

NOTE

The mode can be calculated for qualitative data as well as for quantitative data. For example, if the data set is red, red, red, green, green, yellow, purple, black, blue, the mode is red.

Statistical software will easily calculate the mean, the median, and the mode. Some graphing calculators can also make these calculations. In the real world, people make these calculations using software.

TRY IT

Consider the annual earnings of workers at a factory. The mode is $25,000 and occurs 150 times out of 301. The median is $50,000 and the mean is $47,500. What would be the best measure of the "center?"
Solution

Because $25,000 occurs nearly half the time, the mode would be the best measure of the center because the median and mean don't represent what most people make at the factory.

The Law of Large Numbers and the Mean

The Law of Large Numbers says that if you take samples of larger and larger size from any population, then the mean $\overline{x}$ of the sample is very likely to get closer and closer to μ. This is discussed in more detail later in the text.

Sampling Distributions and Statistic of a Sampling Distribution

You can think of a sampling distribution as a **relative frequency distribution** with a great many samples. (See **Sampling and Data** for a review of relative frequency.) Suppose thirty randomly selected students were asked the number of movies they watched the previous week. The results are in the **relative frequency table** shown below.

# of movies	Relative Frequency
0	$\frac{5}{30}$
1	$\frac{15}{30}$
2	$\frac{6}{30}$
3	$\frac{3}{30}$
4	$\frac{1}{30}$

If you let the number of samples get very large (say, 300 million or more), the relative frequency table becomes a relative frequency distribution.

A **statistic** is a number calculated from a sample. Statistic examples include the mean, the median and the mode as well as others. The sample mean $\overline{x}$ is an example of a statistic which estimates the population mean μ.

Calculating the Mean of Grouped Frequency Tables

When only grouped data is available, you do not know the individual data values (we only know intervals and interval frequencies); therefore, you cannot compute an exact mean for the data set. What we must do is estimate the actual mean by calculating the mean of a **frequency table**. A frequency table is a data representation in which grouped data is displayed along with the corresponding frequencies. To calculate the mean from a grouped frequency table we can apply the basic definition of mean: *mean* = $\frac{data\ sum}{number\ of\ data\ values}$ We simply need to modify the definition to fit within the restrictions of a frequency table.

Since we do not know the individual data values we can instead find the **midpoint** of each interval. The midpoint is $\frac{lower\ boundary + upper\ boundary}{2}$. We can now modify the mean definition to be $Mean\ of\ Frequency\ Table = \frac{\sum fm}{\sum f}$ where f = the frequency of the interval and m = the midpoint of the interval.

A frequency table displaying professor Blount's last statistic test is shown. Find the best estimate of the class mean.

Grade Interval	Number of Students
50–56.5	1
56.5–62.5	0
62.5–68.5	4
68.5–74.5	4
74.5–80.5	2
80.5–86.5	3
86.5–92.5	4
92.5–98.5	1

Solution

- Find the midpoints for all intervals

Grade Interval	Midpoint
50–56.5	53.25
56.5–62.5	59.5
62.5–68.5	65.5
68.5–74.5	71.5
74.5–80.5	77.5
80.5–86.5	83.5
86.5–92.5	89.5
92.5–98.5	95.5

- Calculate the sum of the product of each interval frequency and midpoint. $\sum fm$

  $53.25\,(1) + 59.5\,(0) + 65.5\,(4) + 71.5\,(4) + 77.5\,(2) + 83.5\,(3) + 89.5\,(4) + 95.5\,(1) = 1460.25$

- $\mu = \dfrac{\sum fm}{\sum f} = \dfrac{1460.25}{19} = 76.86$

TRY IT

Maris conducted a study on the effect that playing video games has on memory recall. As part of her study, she compiled the following data:

Hours Teenagers Spend on Video Games	Number of Teenagers
0–3.5	3
3.5–7.5	7
7.5–11.5	12
11.5–15.5	7
15.5–19.5	9

What is the best estimate for the mean number of hours spent playing video games?
Solution

Find the midpoint of each interval, multiply by the corresponding number of teenagers, add the results and then divide by the total number of teenagers

The midpoints are 1.75, 5.5, 9.5, 13.5,17.5.

Mean = (1.75)(3) + (5.5)(7) + (9.5)(12) + (13.5)(7) + (17.5)(9) = 409.75

References

Data from The World Bank, available online at http://www.worldbank.org (accessed April 3, 2013).

"Demographics: Obesity – adult prevalence rate." Indexmundi. Available online at http://www.indexmundi.com/g/r.aspx?t=50&v=2228&l=en (accessed April 3, 2013).

2.6 SKEWNESS AND THE MEAN, MEDIAN, AND MODE

Learning Objectives

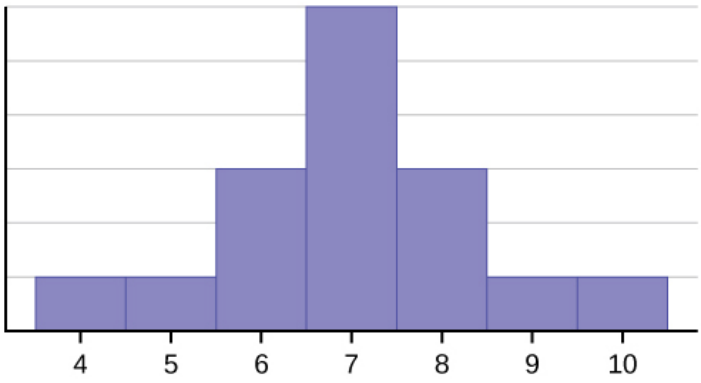
By the end of this section, the student should be able to:

- Recognize basic distribution shapes.

Consider the following data set.

4; 5; 6; 6; 6; 7; 7; 7; 7; 7; 7; 8; 8; 8; 9; 10

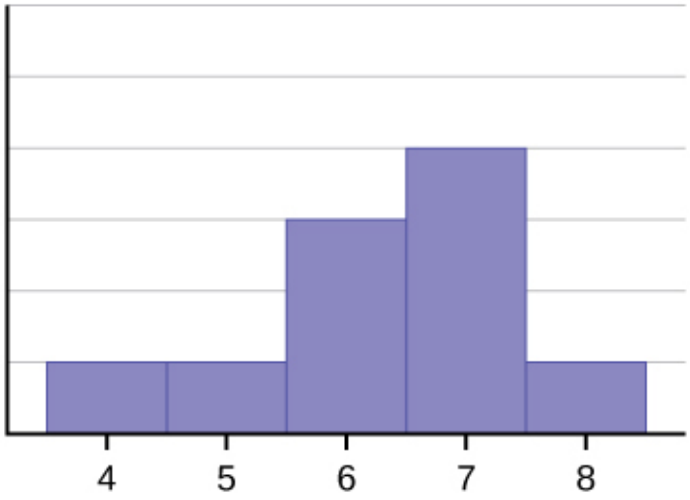
This data set can be represented by the following histogram. Each interval has width one, and each value is located in the middle of an interval.



The histogram displays a **symmetrical** distribution of data. A distribution is symmetrical if a vertical line can be drawn at some point in the histogram such that the shape to the left and the right of the vertical line are mirror images of each other. The mean, the median, and the mode are each seven for these data. **In a perfectly symmetrical distribution, the mean and the median are the same.** This example has one mode

(unimodal), and the mode is the same as the mean and median. In a symmetrical distribution that has two modes (bimodal), the two modes would be different from the mean and median.

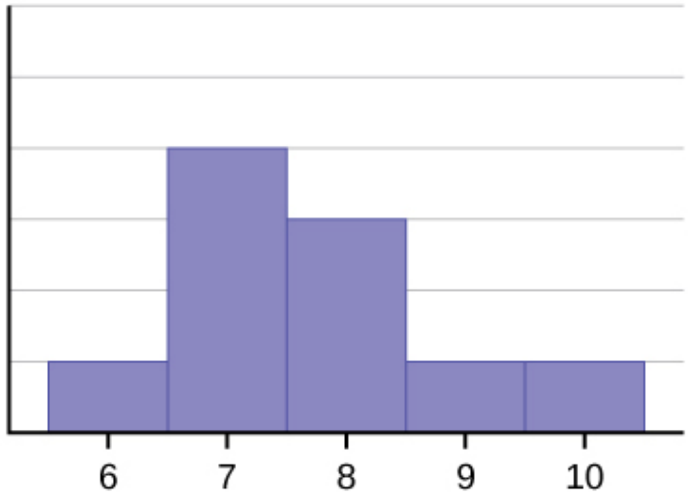
The histogram for the data:
4; 5; 6; 6; 6; 7; 7; 7; 7; 8 is not symmetrical. The right-hand side seems "chopped off" compared to the left side. A distribution of this type is called **skewed to the left** because it is pulled out to the left.



The mean is 6.3, the median is 6.5, and the mode is seven. **Notice that the mean is less than the median, and they are both less than the mode.** The mean and the median both reflect the skewing, but the mean reflects it more so.

The histogram for the data:

6; 7; 7; 7; 7; 8; 8; 8; 9; 10, is also not symmetrical. It is **skewed to the right**.



The mean is 7.7, the median is 7.5, and the mode is seven. Of the three statistics, **the mean is the largest, while the mode is the smallest**. Again, the mean reflects the skewing the most.

To summarize, generally if the distribution of data is skewed to the left, the mean is less than the median,

which is often less than the mode. If the distribution of data is skewed to the right, the mode is often less than the median, which is less than the mean.

Skewness and symmetry become important when we discuss probability distributions in later chapters.

Statistics are used to compare and sometimes identify authors. The following lists shows a simple random sample that compares the letter counts for three authors.

Terry: 7; 9; 3; 3; 3; 4; 1; 3; 2; 2

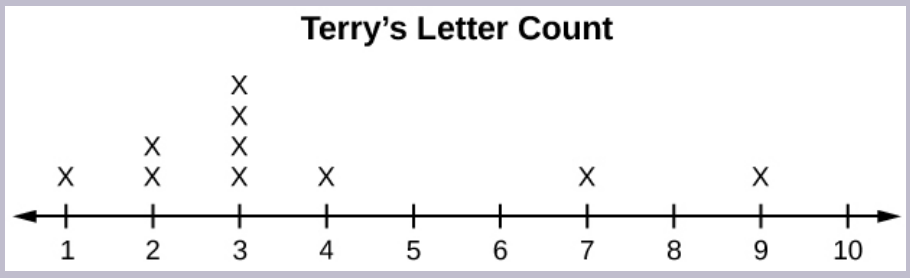
Davis: 3; 3; 3; 4; 1; 4; 3; 2; 3; 1

Maris: 2; 3; 4; 4; 4; 6; 6; 6; 8; 3

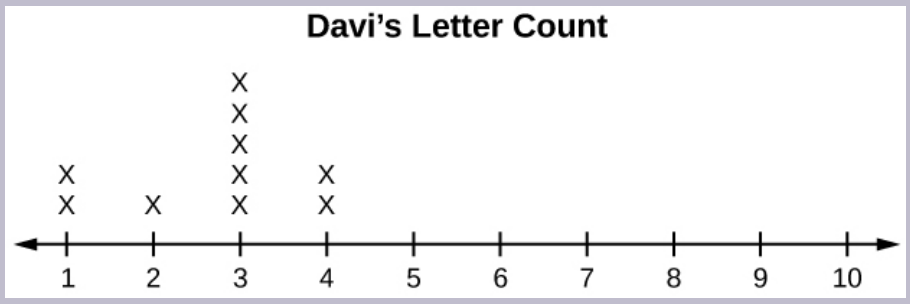
1. Make a dot plot for the three authors and compare the shapes.
2. Calculate the mean for each.
3. Calculate the median for each.
4. Describe any pattern you notice between the shape and the measures of center.

Solution

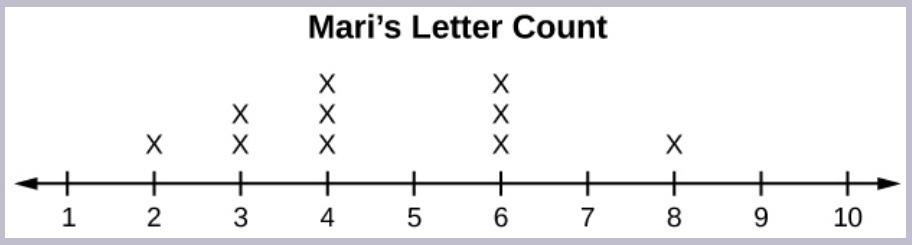
1. Terry's distribution has a right (positive) skew.



Davis' distribution has a left (negative) skew



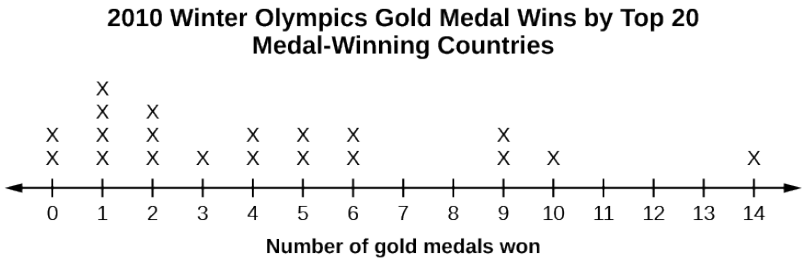
Maris' distribution is symmetrically shaped.



2. Terry's mean is 3.7, Davis' mean is 2.7, Maris' mean is 4.6.
3. Terry's median is three, Davis' median is three. Maris' median is four.
4. It appears that the median is always closest to the high point (the mode), while the mean tends to be farther out on the tail. In a symmetrical distribution, the mean and the median are both centrally located close to the high point of the distribution.

TRY IT

Discuss the mean, median, and mode for each of the following problems. Is there a pattern between the shape and measure of the center?

1. 

2.

The Ages Former U.S Presidents Died

4	6 9
5	3 6 7 7 7 8
6	0 0 3 3 4 4 5 6 7 7 7 8
7	0 1 1 2 3 4 7 8 8 9
8	0 1 3 5 8
9	0 0 3 3

Key: 8|0 means 80.

3.



Solution

1. mean = 4.25, median = 3.5, mode = 1; The mean > median > mode which indicates skewness to the right. (data are 0, 1, 2, 3, 4, 5, 6, 9, 10, 14 and respective frequencies are 2, 4, 3, 1, 2, 2, 2, 2, 1, 1)
2. mean = 70.1 , median = 68, mode = 57, 67 bimodal; the mean and median are close but there is a little skewness to the right which is influenced by the data being bimodal (data are 46, 49, 53, 56, 57, 57, 57, 58, 60, 60, 63, 63, 64, 64, 65, 66, 67, 67, 67, 68, 70, 71, 71, 72, 73, 74, 77, 78, 78, 79, 80, 81, 83, 85, 88, 90, 90 93, 93).
3. These are estimates: mean =16.095, median = 17.495, mode = 22.495 (there may be no mode); The mean < median < mode which indicates skewness to the left. (data are the midpoints of the intervals: 2.495, 7.495, 12.495, 17.495, 22.495 and respective frequencies are 2, 3, 4, 7, 9).

2.7 MEASURES OF THE SPREAD OF THE DATA

Learning Objectives

By the end of this section, the student should be able to:

- Recognize, describe, and calculate the measures of the spread of data: variance, standard deviation, and range.
- Use the mean and the standard deviation to calculate z-scores and the empirical rule.

An important characteristic of any set of data is the **variation** in the data. In some data sets, the data values are concentrated closely near the mean; in other data sets, the data values are more widely spread out from the mean. The most common measure of variation, or spread, is the standard deviation. The standard deviation is a number that measures how far data values are from their mean.

The Standard Deviation

- provides a numerical measure of the overall amount of variation in a data set, and
- can be used to determine whether a particular data value is close to or far from the mean.

The standard deviation provides a measure of the overall variation in a data set

The **standard deviation** is always positive or zero. The standard deviation is small when the data are all

concentrated close to the mean, exhibiting little variation or spread. The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation.

Suppose that we are studying the amount of time customers wait in line at the checkout at supermarket *A* and supermarket *B*. the average wait time at both supermarkets is five minutes. At supermarket *A*, the standard deviation for the wait time is two minutes; at supermarket *B* the standard deviation for the wait time is four minutes.

Because supermarket *B* has a higher standard deviation, we know that there is more variation in the wait times at supermarket *B*. Overall, wait times at supermarket *B* are more spread out from the average; wait times at supermarket *A* are more concentrated near the average.

The standard deviation can be used to determine whether a data value is close to or far from the mean.

Suppose that Rosa and Binh both shop at supermarket *A*. Rosa waits at the checkout counter for seven minutes and Binh waits for one minute. At supermarket *A*, the mean waiting time is five minutes and the standard deviation is two minutes. The standard deviation can be used to determine whether a data value is close to or far from the mean.

Rosa waits for seven minutes:

- Seven is two minutes longer than the average of five; two minutes is equal to one standard deviation.
- Rosa's wait time of seven minutes is **two minutes longer than the average** of five minutes.
- Rosa's wait time of seven minutes is **one standard deviation above the average** of five minutes.

Binh waits for one minute.

- One is four minutes less than the average of five; four minutes is equal to two standard deviations.
- Binh's wait time of one minute is **four minutes less than the average** of five minutes.
- Binh's wait time of one minute is **two standard deviations below the average** of five minutes.
- A data value that is two standard deviations from the average is just on the borderline for what many statisticians would consider to be far from the average. Considering data to be far from the mean if it is more than two standard deviations away is more of an approximate "rule of thumb" than a rigid rule. In general, the shape of the distribution of the data affects how much of the data is further away than two standard deviations. (You will learn more about this in later chapters.)

The number line may help you understand standard deviation. If we were to put five and seven on a number line, seven is to the right of five. We say, then, that seven is **one** standard deviation to the **right** of five because $5 + (1)(2) = 7$.

If one were also part of the data set, then one is **two** standard deviations to the **left** of five because $5 + (-2)(2) = 1$.



- In general, a **value = mean + (#ofSTDEV)(standard deviation)**
- where #ofSTDEVs = the number of standard deviations
- #ofSTDEV does not need to be an integer
- One is **two standard deviations less than the mean** of five because: $1 = 5 + (-2)(2)$.

The equation **value = mean + (# of STDEVs)(standard deviation)** can be expressed for a sample and for a population.
sample: $x = \overline{x} + (\#\text{ of STDEV}) \cdot (s)$
Population: $x = \mu + (\#\text{ of STDEV}) \cdot (\sigma)$

The lower case letter *s* represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation.

The symbol $\overline{x}$ is the sample mean and the Greek symbol
μ is the population mean.

Calculating the Standard Deviation

If *x* is a number, then the difference "*x* – mean" is called its **deviation**. In a data set, there are as many deviations as there are items in the data set. The deviations are used to calculate the standard deviation. If the numbers belong to a population, in symbols a deviation is $x - \mu$. For sample data, in symbols a deviation is $x - \overline{x}$.

The procedure to calculate the standard deviation depends on whether the numbers are the entire population or are data from a sample. The calculations are similar, but not identical. Therefore the symbol used to represent the standard deviation depends on whether it is calculated from a population or a sample. The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation. If the sample has the same characteristics as the population, then s should be a good estimate of σ.

To calculate the standard deviation, we need to calculate the variance first. The variance is the **average of the squares of the deviations** (the $x - \overline{x}$ values for a sample, or the $x - \mu$ values for a population). The symbol σ^2 represents the population variance; the population standard deviation σ is the square root of the population variance. The symbol s^2 represents the sample variance; the sample standard deviation *s* is the square root of the sample variance. You can think of the standard deviation as a special average of the deviations.

If the numbers come from a census of the entire **population** and not a sample, when we calculate the

average of the squared deviations to find the variance, we divide by N, the number of items in the population. If the data are from a **sample** rather than a population, when we calculate the average of the squared deviations, we divide by $\boldsymbol{n-1}$, one less than the number of items in the sample.

Formulas for the Sample Standard Deviation

- $s=\sqrt{\frac{\Sigma(x-\overline{x})^2}{n-1}}$ or $s=\sqrt{\frac{\Sigma f(x-\overline{x})^2}{n-1}}$
- For the sample standard deviation, the denominator is $\boldsymbol{n-1}$, that is the sample size MINUS 1.

Formulas for the Population Standard Deviation

- $\sigma=\sqrt{\frac{\Sigma(x-\mu)^2}{N}}$ or $\sigma=\sqrt{\frac{\Sigma f(x-\mu)^2}{N}}$
- For the population standard deviation, the denominator is N, the number of items in the population.

In these formulas, f represents the frequency with which a value appears. For example, if a value appears once, f is one. If a value appears three times in the data set or population, f is three.

Sampling Variability of a Statistic

The statistic of a sampling distribution was discussed in Descriptive Statistics: Measuring the Center of the Data. How much the statistic varies from one sample to another is known as the sampling variability of a statistic. You typically measure the sampling variability of a statistic by its standard error. The **standard error of the mean** is an example of a standard error. It is a special standard deviation and is known as the standard deviation of the sampling distribution of the mean. You will cover the standard error of the mean in the chapter The Central Limit Theorem (not now). The notation for the standard error of the mean is $\frac{\sigma}{\sqrt{n}}$ where σ is the standard deviation of the population and n is the size of the sample.

NOTE

In practice, USE A CALCULATOR OR COMPUTER SOFTWARE TO CALCULATE THE STANDARD DEVIATION. If you are using a TI-83, 83+, 84+ calculator, you need to select the appropriate standard deviation σ_x or s_x from the summary statistics. We will concentrate on using and interpreting the information that the standard deviation gives us. However you should study the following step-by-step example to help you understand how the standard deviation measures variation from the mean. (The calculator instructions appear at the end of this example.)

In a fifth grade class, the teacher was interested in the average age and the sample standard deviation of the ages of her students. The following data are the ages for a SAMPLE of $n = 20$ fifth grade students. The ages are rounded to the nearest half year:

9; 9.5; 9.5; 10; 10; 10; 10; 10.5; 10.5; 10.5; 10.5; 11; 11; 11; 11; 11; 11; 11.5; 11.5; 11.5;

$$\overline{x} = \frac{9 + 9.5(2) + 10(4) + 10.5(4) + 11(6) + 11.5(3)}{20} = 10.525$$

The average age is 10.53 years, rounded to two places.

The variance may be calculated by using a table. Then the standard deviation is calculated by taking the square root of the variance. We will explain the parts of the table after calculating *s*.

Data	**Freq.**	**Deviations**	***Deviations*^{2}**	**(Freq.)(*Deviations*^{2})**
x	f	$(x - \overline{x})$	$(x - \overline{x})^2$	$(f)(x - \overline{x})^2$
9	1	9 – 10.525 = –1.525	$(-1.525)^2 = 2.325625$	1 × 2.325625 = 2.325625
9.5	2	9.5 – 10.525 = –1.025	$(-1.025)^2 = 1.050625$	2 × 1.050625 = 2.101250
10	4	10 – 10.525 = –0.525	$(-0.525)^2 = 0.275625$	4 × 0.275625 = 1.1025
10.5	4	10.5 – 10.525 = –0.025	$(-0.025)^2 = 0.000625$	4 × 0.000625 = 0.0025
11	6	11 – 10.525 = 0.475	$(0.475)^2 = 0.225625$	6 × 0.225625 = 1.35375
11.5	3	11.5 – 10.525 = 0.975	$(0.975)^2 = 0.950625$	3 × 0.950625 = 2.851875
				The total is 9.7375

The sample variance, s^2, is equal to the sum of the last column (9.7375) divided by the total number of data values minus one (20 – 1):

$$s^2 = \frac{9.7375}{20 - 1} = 0.5125$$

The **sample standard deviation** *s* is equal to the square root of the sample variance:

$s = \sqrt{0.5125} = 0.715891$, which is rounded to two decimal places, $s = 0.72$.

Typically, you do the calculation for the standard deviation on your calculator or computer. The intermediate results are not rounded. This is done for accuracy.

For the following problems, recall that **value = mean + (#ofSTDEVs)(standard deviation)**. Verify the mean and standard deviation on a calculator or computer.
For a sample: $x = \overline{x} + (\#\text{ofSTDEVs})(\text{s})$
For a population: $x = \mu + (\#\text{ofSTDEVs})(\sigma)$
For this example, use $x = \overline{x} + (\#\text{ofSTDEVs})(\text{s})$ because the data is from a sample

1. Verify the mean and standard deviation on your calculator or computer.
2. Find the value that is one standard deviation above the mean. Find $(\overline{x} + 1s)$.
3. Find the value that is two standard deviations below the mean. Find $(\overline{x} - 2s)$.
4. Find the values that are 1.5 standard deviations **from** (below and above) the mean.

Solution

1.
   - Clear lists L1 and L2. Press STAT 4:ClrList. Enter 2nd 1 for L1, the comma (,), and 2nd 2 for L2.
   - Enter data into the list editor. Press STAT 1:EDIT. If necessary, clear the lists by arrowing up into the name. Press CLEAR and arrow down.
   - Put the data values (9, 9.5, 10, 10.5, 11, 11.5) into list L1 and the frequencies (1, 2, 4, 4, 6, 3) into list L2. Use the arrow keys to move around.
   - Press STAT and arrow to CALC. Press 1:1-VarStats and enter L1 (2nd 1), L2 (2nd 2). Do not forget the comma. Press ENTER.
   - $\overline{x} = 10.525$
   - Use Sx because this is sample data (not a population): $\text{Sx} = 0.715891$
2. $(\overline{x} + 1s) = 10.53 + (1)(0.72) = 11.25$
3. $(\overline{x} - 2s) = 10.53 - (2)(0.72) = 9.09$
4. One standard deviation below the mean is $(\overline{x} - 1.5s) = 10.53 - (1.5)(0.72) = 9.45$.
   One standard deviation below the mean is
   $(\overline{x} + 1.5s) = 10.53 + (1.5)(0.72) = 11.61$.

TRY IT

On a baseball team, the ages of each of the players are as follows:

21; 21; 22; 23; 24; 24; 25; 25; 28; 29; 29; 31; 32; 33; 33; 34; 35; 36; 36; 36; 36; 38; 38; 38; 40

Use your calculator or computer to find the mean and standard deviation. Then find the value that is two standard deviations above the mean.

Mean: $\overline{x} = 30.68$

Standard Deviation: $s = 6.09$

Two standard deviations above the mean: $(\overline{x} + 2s) = 30.68 + (2)(6.09) = 42.86$

Explanation of the standard deviation calculation shown in the table

The deviations show how spread out the data are about the mean. The data value 11.5 is farther from the mean than is the data value 11 which is indicated by the deviations 0.97 and 0.47. A positive deviation occurs when the data value is greater than the mean, whereas a negative deviation occurs when the data value is less than the mean. The deviation is –1.525 for the data value nine. **If you add the deviations, the sum is always zero**. (For [link], there are $n = 20$ deviations.) So you cannot simply add the deviations to get the spread of the data. By squaring the deviations, you make them positive numbers, and the sum will also be positive. The variance, then, is the "average" squared deviation.

The variance is a squared measure and does not have the same units as the data. Taking the square root solves the problem. The standard deviation measures the spread in the same units as the data.

Notice that instead of dividing by $n = 20$, the calculation is divided by $n-1 = 20-1 = 19$ because the data is a sample. For the **sample** variance, we divide by the sample size minus one $(n-1)$. Why not divide by n? The answer has to do with the population variance. **The sample variance is an estimate of the population variance.** Based on the theoretical mathematics that lies behind these calculations, dividing by $(n-1)$ gives a better estimate of the population variance.

NOTE

Your concentration should be on what the standard deviation tells us about the data. The standard deviation

is a number which measures how far the data are spread from the mean. Let a calculator or computer do the arithmetic.

The standard deviation, s or σ, is either zero or larger than zero. When the standard deviation is zero, there is no spread; that is, all the data values are equal to each other. The standard deviation is small when the data are all concentrated close to the mean, and is larger when the data values show more variation from the mean. When the standard deviation is a lot larger than zero, the data values are very spread out about the mean; outliers can make s or σ very large.

The standard deviation, when first presented, can seem unclear. By graphing your data, you can get a better "feel" for the deviations and the standard deviation. You will find that in symmetrical distributions, the standard deviation can be very helpful but in skewed distributions, the standard deviation may not be much help. The reason is that the two sides of a skewed distribution have different spreads. In a skewed distribution, it is better to look at the first quartile, the median, the third quartile, the smallest value, and the largest value. Because numbers can be confusing, **always graph your data**. Display your data in a histogram or a box plot.

Use the following data (first exam scores) from Susan Dean's spring pre-calculus class:

33; 42; 49; 49; 53; 55; 55; 61; 63; 67; 68; 68; 69; 69; 72; 73; 74; 78; 80; 83; 88; 88; 88; 90; 92; 94; 94; 94; 94; 96; 100

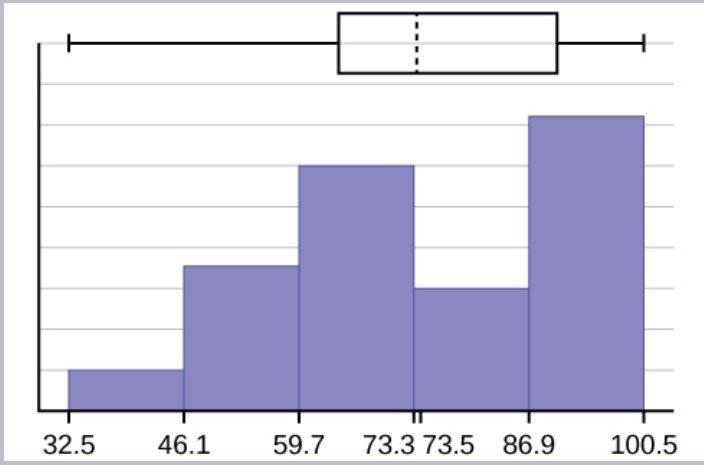
1. Create a chart containing the data, frequencies, relative frequencies, and cumulative relative frequencies to three decimal places.
2. Calculate the following to one decimal place using a TI-83+ or TI-84 calculator:
    1. The sample mean
    2. The sample standard deviation
    3. The median
    4. The first quartile
    5. The third quartile
    6. *IQR*
3. Construct a box plot and a histogram on the same set of axes. Make comments about the box plot, the histogram, and the chart.

Solution

1. See [link]
2. The solutions are below:

1. The sample mean = 73.5
2. The sample standard deviation = 17.9
3. The median = 73
4. The first quartile = 61
5. The third quartile = 90
6. *IQR* = 90 – 61 = 29

3. The *x*-axis goes from 32.5 to 100.5; *y*-axis goes from –2.4 to 15 for the histogram. The number of intervals is five, so the width of an interval is (100.5 – 32.5) divided by five, is equal to 13.6. Endpoints of the intervals are as follows: the starting point is 32.5, 32.5 + 13.6 = 46.1, 46.1 + 13.6 = 59.7, 59.7 + 13.6 = 73.3, 73.3 + 13.6 = 86.9, 86.9 + 13.6 = 100.5 = the ending value; No data values fall on an interval boundary.



The long left whisker in the box plot is reflected in the left side of the histogram. The spread of the exam scores in the lower 50% is greater ($73 - 33 = 40$) than the spread in the upper 50% ($100 - 73 = 27$). The histogram, box plot, and chart all reflect this. There are a substantial number of A and B grades (80s, 90s, and 100). The histogram clearly shows this. The box plot shows us that the middle 50% of the exam scores (*IQR* = 29) are Ds, Cs, and Bs. The box plot also shows us that the lower 25% of the exam scores are Ds and Fs.

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
33	1	0.032	0.032
42	1	0.032	0.064
49	2	0.065	0.129
53	1	0.032	0.161
55	2	0.065	0.226
61	1	0.032	0.258
63	1	0.032	0.29
67	1	0.032	0.322
68	2	0.065	0.387
69	2	0.065	0.452
72	1	0.032	0.484
73	1	0.032	0.516
74	1	0.032	0.548
78	1	0.032	0.580
80	1	0.032	0.612
83	1	0.032	0.644
88	3	0.097	0.741
90	1	0.032	0.773
92	1	0.032	0.805
94	4	0.129	0.934
96	1	0.032	0.966
100	1	0.032	0.998 (Why isn't this value 1?)

TRY IT

The following data show the different types of pet food stores in the area carry.

6; 6; 6; 6; 7; 7; 7; 7; 7; 8; 9; 9; 9; 9; 9; 10; 10; 10; 10; 10; 11; 11; 11; 11; 12; 12; 12; 12; 12; 12;

Calculate the sample mean and the sample standard deviation to one decimal place using a TI-83+ or TI-84 calculator.

$\overline{x} = 9.3$

s = 2.2

Standard deviation of Grouped Frequency Tables

Recall that for grouped data we do not know individual data values, so we cannot describe the typical value of the data with precision. In other words, we cannot find the exact mean, median, or mode. We can, however, determine the best estimate of the measures of center by finding the mean of the grouped data with the formula:

$$Mean\ of\ Frequency\ Table = \frac{\sum fm}{\sum f} \quad \text{where} \quad f = \text{interval frequencies} \quad \text{and}$$

$m = \text{interval midpoints}$.

Just as we could not find the exact mean, neither can we find the exact standard deviation. Remember that standard deviation describes numerically the expected deviation a data value has from the mean. In simple English, the standard deviation allows us to compare how "unusual" individual data is compared to the mean.

Find the standard deviation for the data in [link].

Class	Frequency, f	Midpoint, m	m^2	$\overline{x}$	fm^2	Standard Deviation
0–2	1	1	1	7.58	1	3.5
3–5	6	4	16	7.58	96	3.5
6–8	10	7	49	7.58	490	3.5
9–11	7	10	100	7.58	700	3.5
12–14	0	13	169	7.58	0	3.5
15–17	2	16	256	7.58	512	3.5

For this data set, we have the mean, $\overline{x} = 7.8$ and the standard deviation, $s_x = 3.5$. This means that a randomly selected data value would be expected to be 3.5 units from the mean. If we look at the first class, we see that the class midpoint is equal to one. This is almost two full standard deviations from the mean since $7.58 - 3.5 - 3.5 = 0.58$. While

the formula for calculating the standard deviation is not complicated, $s_x = \sqrt{\frac{f(m-\overline{x})^2}{n-1}}$, where $s_x = \textbf{sample standard deviation}$ and $\overline{x} = \textbf{sample mean}$, the calculations are tedious. It is usually best to use technology when performing the calculations.

TRY IT

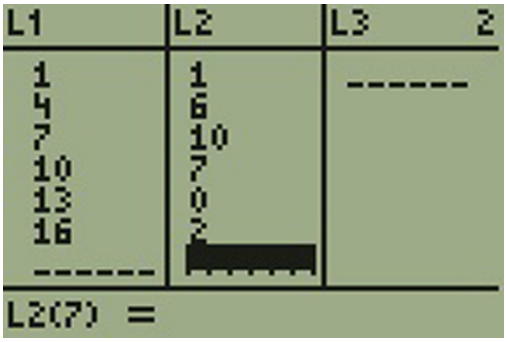
Find the standard deviation for the data from the previous example

Class	Frequency, f
0–2	1
3–5	6
6–8	10
9–11	7
12–14	0
15–17	2

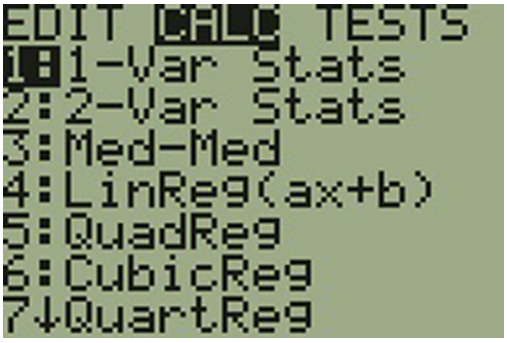
First, press the **STAT** key and select **1:Edit**



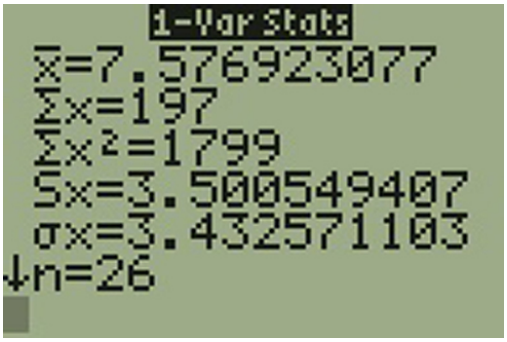
Input the midpoint values into **L1** and the frequencies into **L2**



Select **STAT**, **CALC**, and **1: 1-Var Stats**



Select **2nd** then **1** then , **2nd** then **2 Enter**



You will see displayed both a population standard deviation, σ_x, and the sample standard deviation, s_x.

Comparing Values from Different Data Sets

The standard deviation is useful when comparing data values that come from different data sets. If the data sets have different means and standard deviations, then comparing the data values directly can be misleading.

- For each data value, calculate how many standard deviations away from its mean the value is.
- Use the formula: $\text{value} = \text{mean} + (\#\text{ofSTDEVs})(\text{standard deviation})$; solve for #ofSTDEVs.

- $\#\text{ofSTDEVs} = \dfrac{\text{value} - \text{mean}}{\text{standard deviation}}$
- Compare the results of this calculation.

#ofSTDEVs is often called a "z-score"; we can use the symbol z. In symbols, the formulas become:

Sample	$x = \overline{x} + zs$	$z = \dfrac{x - \overline{x}}{s}$
Population	$x = \mu + z\sigma$	$z = \dfrac{x - \mu}{\sigma}$

Two students, John and Ali, from different high schools, wanted to find out who had the highest GPA when compared to his school. Which student had the highest GPA when compared to his school?

Student	GPA	School Mean GPA	School Standard Deviation
John	2.85	3.0	0.7
Ali	77	80	10

Solution

For each student, determine how many standard deviations (#ofSTDEVs) his GPA is away from the average, for his school. Pay careful attention to signs when comparing and interpreting the answer.

$$z = \#\text{ofSTDEVs} = \frac{\text{value} - \text{mean}}{\text{standard deviation}} = \frac{x - \mu}{\sigma}$$

For John, $z = \#\text{ofSTDEVs} = \dfrac{2.85 - 3.0}{0.7} = -0.21$

For Ali, $z = \#\text{ofSTDEVs} = \dfrac{77 - 80}{10} = -0.3$

John has the better GPA when compared to his school because his GPA is 0.21 standard deviations **below** his school's mean while Ali's GPA is 0.3 standard deviations **below** his school's mean.

John's z-score of –0.21 is higher than Ali's z-score of –0.3. For GPA, higher values are better, so we conclude that John has the better GPA when compared to his school.

TRY IT

Two swimmers, Angie and Beth, from different teams, wanted to find out who had the fastest time for the 50 meter freestyle when compared to her team. Which swimmer had the fastest time when compared to her team?

Swimmer	Time (seconds)	Team Mean Time	Team Standard Deviation
Angie	26.2	27.2	0.8
Beth	27.3	30.1	1.4

For Angie: $z = \frac{26.2 - 27.2}{0.8} = -1.25$

For Beth: $z = \frac{27.3 - 30.1}{1.4} = -2$

The following lists give a few facts that provide a little more insight into what the standard deviation tells us about the distribution of the data.
For ANY data set, no matter what the distribution of the data is:

- At least 75% of the data is within two standard deviations of the mean.
- At least 89% of the data is within three standard deviations of the mean.
- At least 95% of the data is within 4.5 standard deviations of the mean.
- This is known as Chebyshev's Rule.

For data having a distribution that is BELL-SHAPED and SYMMETRIC:

- Approximately 68% of the data is within one standard deviation of the mean.
- Approximately 95% of the data is within two standard deviations of the mean.
- More than 99% of the data is within three standard deviations of the mean.
- This is known as the Empirical Rule.
- It is important to note that this rule only applies when the shape of the distribution of the data is bell-shaped and symmetric. We will learn more about this when studying the "Normal" or "Gaussian" probability distribution in later chapters.

References

Data from Microsoft Bookshelf.

King, Bill. "Graphically Speaking." Institutional Research, Lake Tahoe Community College. Available online at http://www.ltcc.edu/web/about/institutional-research (accessed April 3, 2013).

CHAPTER 2 REVIEW

Chapter Review from 2.1

A **stem-and-leaf plot** is a way to plot data and look at the distribution. In a stem-and-leaf plot, all data values within a class are visible. The advantage in a stem-and-leaf plot is that all values are listed, unlike a histogram, which gives classes of data values. A **line graph** is often used to represent a set of data values in which a quantity varies with time. These graphs are useful for finding trends. That is, finding a general pattern in data sets including temperature, sales, employment, company profit or cost over a period of time. A **bar graph** is a chart that uses either horizontal or vertical bars to show comparisons among categories. One axis of the chart shows the specific categories being compared, and the other axis represents a discrete value. Some bar graphs present bars clustered in groups of more than one (grouped bar graphs), and others show the bars divided into subparts to show cumulative effect (stacked bar graphs). Bar graphs are especially useful when categorical data is being used.

Exercises

For each of the following data sets, create a stem plot and identify any outliers.

The miles per gallon rating for 30 cars are shown below (lowest to highest).

19, 19, 19, 20, 21, 21, 25, 25, 25, 26, 26, 28, 29, 31, 31, 32, 32, 33, 34, 35, 36, 37, 37, 38, 38, 38, 38, 41, 43, 43

Solution

Stem	Leaf
1	999
2	0115556689
3	11223456778888
4	133

2. The height in feet of 25 trees is shown below (lowest to highest).

25, 27, 33, 34, 34, 34, 35, 37, 37, 38, 39, 39, 39, 40, 41, 45, 46, 47, 49, 50, 50, 53, 53, 54, 54

3. The data are the prices of different laptops at an electronics store. Round each value to the nearest ten.

249, 249, 260, 265, 265, 280, 299, 299, 309, 319, 325, 326, 350, 350, 350, 365, 369, 389, 409, 459, 489, 559, 569, 570, 610

Solution

Stem	Leaf
2	5 5 6 7 7 8
3	0 0 1 2 3 3 5 5 5 7 7 9
4	1 6 9
5	6 7 7
6	1

4. The data are daily high temperatures in a town for one month.

61, 61, 62, 64, 66, 67, 67, 67, 68, 69, 70, 70, 70, 71, 71, 72, 74, 74, 74, 75, 75, 75, 76, 76, 77, 78, 78, 79, 79, 95

Exercises

For the next three exercises, use the data to construct a line graph.

In a survey, 40 people were asked how many times they visited a store before making a major purchase. The results are shown in [link].

Number of times in store	Frequency
1	4
2	10
3	16
4	6
5	4

Solution



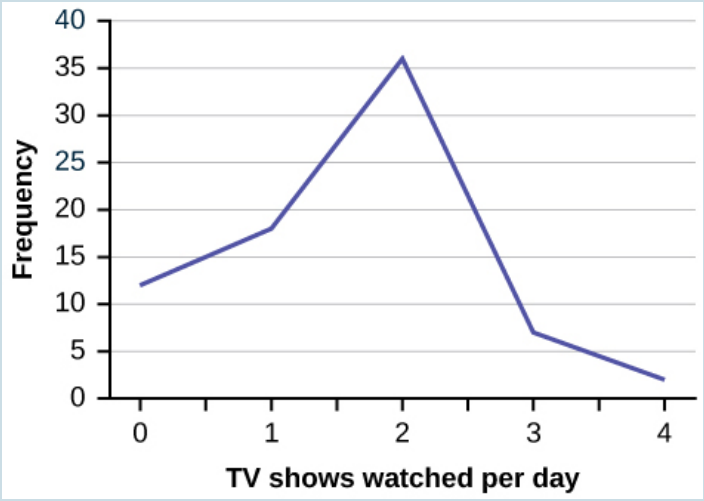
2. In a survey, several people were asked how many years it has been since they purchased a mattress. The results are shown in [link].

Years since last purchase	Frequency
0	2
1	8
2	13
3	22
4	16
5	9

3. Several children were asked how many TV shows they watch each day. The results of the survey are shown in [link].

Number of TV Shows	Frequency
0	12
1	18
2	36
3	7
4	2

Solution



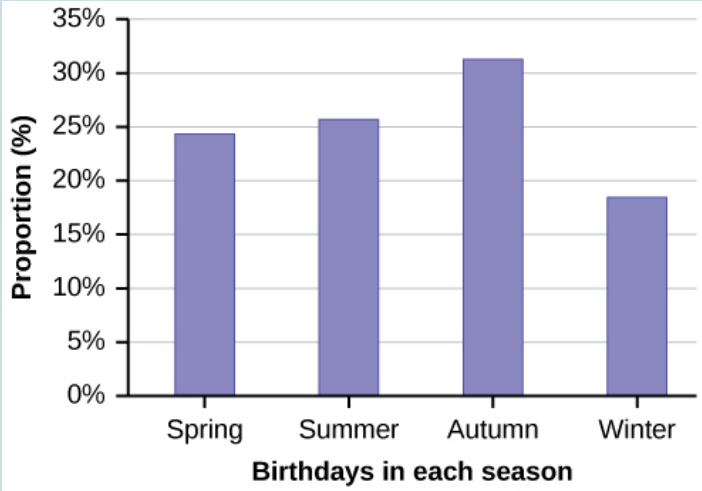
Exercises

The students in Ms. Ramirez's math class have birthdays in each of the four seasons. [link] shows the four seasons, the number of students who have birthdays in each season, and the percentage (%) of students in each group. Construct a bar graph showing the number of students.

Seasons	Number of students	Proportion of population
Spring	8	24%
Summer	9	26%
Autumn	11	32%
Winter	6	18%

Using the data from Mrs. Ramirez's math class supplied in [link], construct a bar graph showing the percentages.

Solution



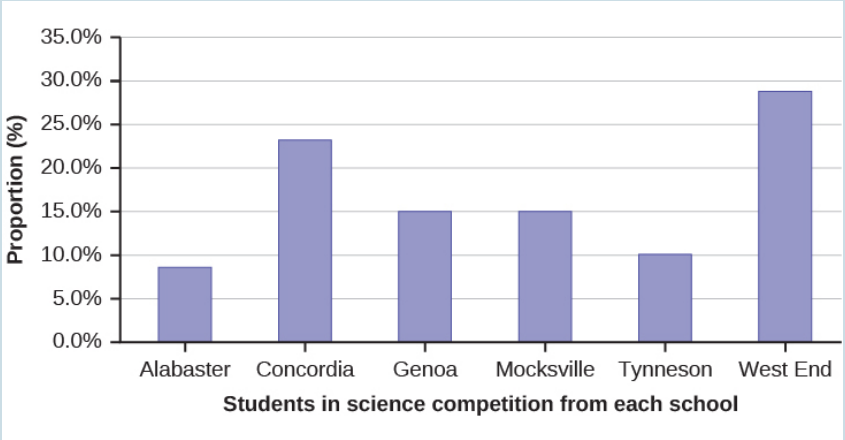
Exercises

David County has six high schools. Each school sent students to participate in a county-wide science competition. [link] shows the percentage breakdown of competitors from each school, and the percentage of the entire student population of the county that goes to each school. Construct a bar graph that shows the population percentage of competitors from each school.

High School	Science competition population	Overall student population
Alabaster	28.9%	8.6%
Concordia	7.6%	23.2%
Genoa	12.1%	15.0%
Mocksville	18.5%	14.3%
Tynneson	24.2%	10.1%
West End	8.7%	28.8%

Use the data from the David County science competition supplied in [link]. Construct a bar graph that shows the county-wide population percentage of students at each school.

Solution



Chapter Review from 2.2

A **histogram** is a graphic version of a frequency distribution. The graph consists of bars of equal width drawn adjacent to each other. The horizontal scale represents classes of quantitative data values and the vertical scale represents frequencies. The heights of the bars correspond to frequency values. Histograms are typically used for large, continuous, quantitative data sets. A frequency polygon can also be used when graphing large data sets with data points that repeat. The data usually goes on the *y*-axis with the frequency being graphed on the *x*-axis. Time series graphs can be helpful when looking at large amounts of data for one variable over a period of time.

Exercises

Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars. Complete the table.

Data Value (# cars)	Frequency	Relative Frequency	Cumulative Relative Frequency

Solution

Data Value (# cars)	Frequency	Relative Frequency	Cumulative Relative Frequency
3	14	0.22	0.22
4	19	0.29	0.51
5	12	0.18	0.69
6	9	0.14	0.83
7	11	0.17	1

What does the frequency column in the table sum to? Why?

Solution
65

What does the relative frequency column in the table sum to? Why?

What is the difference between relative frequency and frequency for each data value in the table?

Solution
The relative frequency shows the *proportion* of data points that have each value. The frequency tells the *number* of data points that have each value.

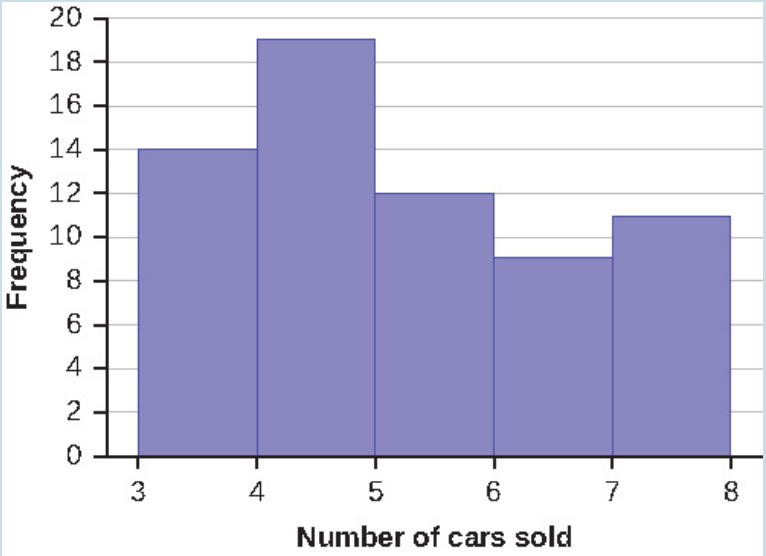
What is the difference between cumulative relative frequency and relative frequency for each data value?

To construct the histogram for the data in the table, determine appropriate minimum and maximum x and y values and the scaling. Sketch the histogram. Label the horizontal and vertical axes with words. Include numerical scaling.



Solution

Answers will vary. One possible histogram is shown:



Exercises

Construct a frequency polygon for the following:

1.

Pulse Rates for Women	Frequency
60–69	12
70–79	14
80–89	11
90–99	1
100–109	1
110–119	0
120–129	1

2.

Actual Speed in a 30 MPH Zone	Frequency
42–45	25
46–49	14
50–53	7
54–57	3
58–61	1

3.

Tar (mg) in Nonfiltered Cigarettes	Frequency
10–13	1
14–17	0
18–21	15
22–25	7
26–29	2

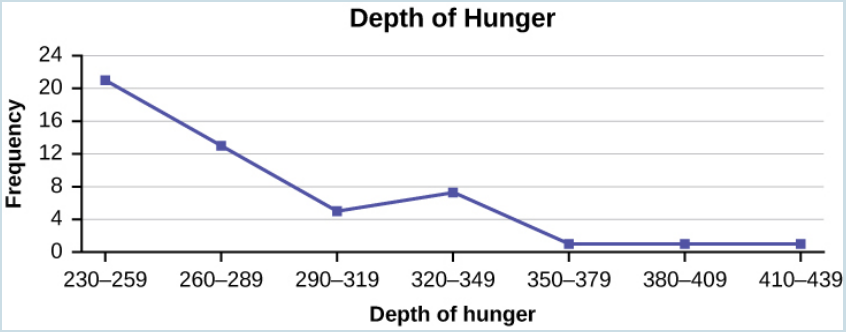
Exercises

Construct a frequency polygon from the frequency distribution for the 50 highest ranked countries for depth of hunger.

Depth of Hunger	Frequency
230–259	21
260–289	13
290–319	5
320–349	7
350–379	1
380–409	1
410–439	1

Solution

Find the midpoint for each class. These will be graphed on the *x*-axis. The frequency values will be graphed on the *y*-axis values.



Exercises

Use the two frequency tables to compare the life expectancy of men and women from 20 randomly selected countries. Include an overlaid frequency polygon and discuss the shapes of the distributions, the center, the spread, and any outliers. What can we conclude about the life expectancy of women compared to men?

Life Expectancy at Birth – Women	Frequency
49–55	3
56–62	3
63–69	1
70–76	3
77–83	8
84–90	2

Life Expectancy at Birth – Men	Frequency
49–55	3
56–62	3
63–69	1
70–76	1
77–83	7
84–90	5

Exercises

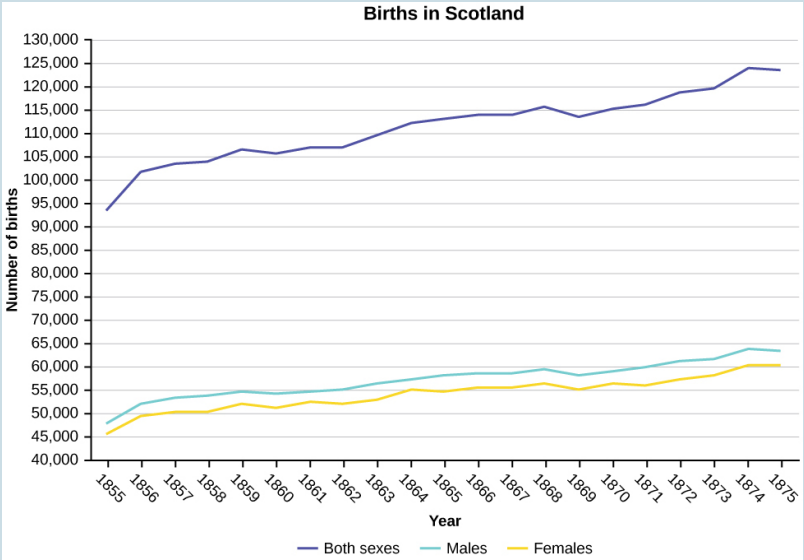
Construct a times series graph for (a) the number of male births, (b) the number of female births, and (c) the total number of births.

Sex/Year	1855	1856	1857	1858	1859	1860	1861
Female	45,545	49,582	50,257	50,324	51,915	51,220	52,403
Male	47,804	52,239	53,158	53,694	54,628	54,409	54,606
Total	93,349	101,821	103,415	104,018	106,543	105,629	107,009

Sex/Year	1862	1863	1864	1865	1866	1867	1868	1869
Female	51,812	53,115	54,959	54,850	55,307	55,527	56,292	55,033
Male	55,257	56,226	57,374	58,220	58,360	58,517	59,222	58,321
Total	107,069	109,341	112,333	113,070	113,667	114,044	115,514	113,354

Sex/Year	1871	1870	1872	1871	1872	1827	1874	1875
Female	56,099	56,431	57,472	56,099	57,472	58,233	60,109	60,146
Male	60,029	58,959	61,293	60,029	61,293	61,467	63,602	63,432
Total	116,128	115,390	118,765	116,128	118,765	119,700	123,711	123,578

Solution



Exercises

The following data sets list full time police per 100,000 citizens along with homicides per 100,000 citizens for the city of Detroit, Michigan during the period from 1961 to 1973.

Year	1961	1962	1963	1964	1965	1966	1967
Police	260.35	269.8	272.04	272.96	272.51	261.34	268.89
Homicides	8.6	8.9	8.52	8.89	13.07	14.57	21.36

Year	1968	1969	1970	1971	1972	1973
Police	295.99	319.87	341.43	356.59	376.69	390.19
Homicides	28.03	31.49	37.39	46.26	47.24	52.33

1. Construct a double time series graph using a common *x*-axis for both sets of data.
2. Which variable increased the fastest? Explain.
3. Did Detroit's increase in police officers have an impact on the murder rate? Explain.

Chapter Review from 2.3

The values that divide a rank-ordered set of data into 100 equal parts are called percentiles. Percentiles are used to compare and interpret data. For example, an observation at the 50^{th} percentile would be greater than 50 percent of the other observations in the set. Quartiles divide data into quarters. The first quartile (Q_1) is the 25^{th} percentile, the second quartile (Q_2 or median) is 50^{th} percentile, and the third quartile (Q_3) is the 75^{th} percentile. The interquartile range, or *IQR*, is the range of the middle 50 percent of the data values. The *IQR* is found by subtracting Q_1 from Q_3, and can help determine outliers by using the following two expressions.

- $Q_3 + IQR(1.5)$
- $Q_1 - IQR(1.5)$

Formula Review

$$i = \left(\frac{k}{100}\right)(n+1)$$

where i = the ranking or position of a data value,
k = the kth percentile,
n = total number of data.

Expression for finding the percentile of a data value: $\left(\frac{x + 0.5y}{n}\right)(100)$, where

x = the number of values counting from the bottom of the data list up to but not including the data value for which you want to find the percentile,

y = the number of data values equal to the data value for which you want to find the percentile,

n = total number of data

Exercises

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

1. Find the 40^{th} percentile.
2. Find the 78^{th} percentile.

Solution

1. The 40^{th} percentile is 37 years.
2. The 78^{th} percentile is 70 years.

Exercises

Listed are 32 ages for Academy Award winning best actors *in order from smallest to largest.*

18; 18; 21; 22; 25; 26; 27; 29; 30; 31; 31; 33; 36; 37; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

1. Find the percentile of 37.
2. Find the percentile of 72.

Jesse was ranked 37^{th} in his graduating class of 180 students. At what percentile is Jesse's ranking?

Solution

Jesse graduated 37th out of a class of 180 students. There are 180 – 37 = 143 students ranked below Jesse. There is one rank of 37.

$x = 143$ and $y = 1$. $\frac{x + 0.5y}{n}(100) = \frac{143 + 0.5(1)}{180}(100) = 79.72$. Jesse's rank of 37 puts him at the 80th percentile.

Exercises

1. For runners in a race, a low time means a faster run. The winners in a race have the shortest running times. Is it more desirable to have a finish time with a high or a low percentile when running a race?
2. The 20th percentile of run times in a particular race is 5.2 minutes. Write a sentence interpreting the 20th percentile in the context of the situation.
3. A bicyclist in the 90th percentile of a bicycle race completed the race in 1 hour and 12 minutes. Is he among the fastest or slowest cyclists in the race? Write a sentence interpreting the 90th percentile in the context of the situation.
4. For runners in a race, a higher percentile means a faster run. Is it more desirable to have a speed with a high or a low percentile when running a race?
5. The 40th percentile of speeds in a particular race is 7.5 miles per hour. Write a sentence interpreting the 40th percentile in the context of the situation.

Solution

1. For runners in a race it is more desirable to have a high percentile for speed. A high percentile means a higher speed which is faster.

5. 40% of runners ran at speeds of 7.5 miles per hour or less (slower). 60% of runners ran at speeds of 7.5 miles per hour or more (faster).

Exercises

On an exam, would it be more desirable to earn a grade with a high or low percentile? Explain.

Exercises

Mina is waiting in line at the Department of Motor Vehicles (DMV). Her wait time of 32 minutes is the 85^{th} percentile of wait times. Is that good or bad? Write a sentence interpreting the 85^{th} percentile in the context of this situation.

Solution

When waiting in line at the DMV, the 85^{th} percentile would be a long wait time compared to the other people waiting. 85% of people had shorter wait times than Mina. In this context, Mina would prefer a wait time corresponding to a lower percentile. 85% of people at the DMV waited 32 minutes or less. 15% of people at the DMV waited 32 minutes or longer.

Exercises

In a survey collecting data about the salaries earned by recent college graduates, Li found that her salary was in the 78^{th} percentile. Should Li be pleased or upset by this result? Explain.

Exercises

In a study collecting data about the repair costs of damage to automobiles in a certain type of crash tests, a certain model of car had \$1,700 in damage and was in the 90^{th} percentile. Should the manufacturer and the consumer be pleased or upset by this result? Explain and write a sentence that interprets the 90^{th} percentile in the context of this problem.

Solution

The manufacturer and the consumer would be upset. This is a large repair cost for the damages, compared to the other cars in the sample. INTERPRETATION: 90% of the crash tested cars had damage repair costs of \$1700 or less; only 10% had damage repair costs of \$1700 or more.

Exercises

The University of California has two criteria used to set admission standards for freshman to be admitted to a college in the UC system:

1. Students' GPAs and scores on standardized tests (SATs and ACTs) are entered into a formula that calculates an "admissions index" score. The admissions index score is used to set eligibility standards intended to meet the goal of admitting the top 12% of high school students in the state. In this context, what percentile does the top 12% represent?
2. Students whose GPAs are at or above the 96^{th} percentile of all students at their high school are eligible (called eligible in the local context), even if they are not in the top 12% of all students in the state. What percentage of students from each high school are "eligible in the local context"?

Exercises

Suppose that you are buying a house. You and your realtor have determined that the most expensive house you can afford is the 34th percentile. The 34th percentile of housing prices is $240,000 in the town you want to move to. In this town, can you afford 34% of the houses or 66% of the houses?

Solution
You can afford 34% of houses. 66% of the houses are too expensive for your budget.
INTERPRETATION: 34% of houses cost $240,000 or less. 66% of houses cost $240,000 or more.

Exercises

Use [link] to calculate the following values:

First quartile = _____

Second quartile = median = 50th percentile = _____

Solution
4

Third quartile = _____

Interquartile range (*IQR*) = _____ – _____ = _____

Solution
6 – 4 = 2

10th percentile = _____

70th percentile = _____

Solution
6

Chapter Review from 2.4

Box plots are a type of graph that can help visually organize data. To graph a box plot the following data points must be calculated: the minimum value, the first quartile, the median, the third quartile, and the maximum value. Once the box plot is graphed, you can display and compare distributions of data.

Exercises

Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars.

Construct a box plot below. Use a ruler to measure and scale accurately.

Looking at your box plot, does it appear that the data are concentrated together, spread out evenly, or concentrated in some areas, but not in others? How can you tell?

Solution
More than 25% of salespersons sell four cars in a typical week. You can see this concentration in the box plot because the first quartile is equal to the median. The top 25% and the bottom 25% are spread out evenly; the whiskers have the same length.

Chapter Review from 2.5

The mean and the median can be calculated to help you find the "center" of a data set. The mean is the best estimate for the actual data set, but the median is the best measurement when a data set contains several outliers or extreme values. The mode will tell you the most frequently occurring datum (or data) in your data set. The mean, median, and mode are extremely helpful when you need to analyze your data, but if your data set consists of ranges which lack specific values, the mean may seem impossible to calculate. However, the mean can be approximated if you add the lower boundary with the upper boundary and divide by two to find the midpoint of each interval. Multiply each midpoint by the number of values found in the corresponding range. Divide the sum of these values by the total number of data values in the set.

Formula Review

$\mu = \dfrac{\sum fm}{\sum f}$ Where f = interval frequencies and m = interval midpoints.

Exercises

Find the mean for the following frequency tables.

1.

Grade	Frequency
49.5–59.5	2
59.5–69.5	3
69.5–79.5	8
79.5–89.5	12
89.5–99.5	5

2.

Daily Low Temperature	Frequency
49.5–59.5	53
59.5–69.5	32
69.5–79.5	15
79.5–89.5	1
89.5–99.5	0

3.

Points per Game	Frequency
49.5–59.5	14
59.5–69.5	32
69.5–79.5	15
79.5–89.5	23
89.5–99.5	2

Exercises

Use the following information to answer the next three exercises:

The following data show the lengths of boats moored in a marina. The data are ordered from smallest to largest:
16; 17; 19; 20; 20; 21; 23; 24; 25; 25; 25; 26; 26; 27; 27; 27; 28; 29; 30; 32; 33; 33; 34; 35; 37; 39; 40

Calculate the mean.

Solution
Mean:

$$16 + 17 + 19 + 20 + 20 + 21 + 23 + 24 + 25 + 25 + 25 + 26 + 26 + 27 + 27 + 27 + 28 + 29 + 30 + 32 + 33 + 33 + 34 + 35 + 37 + 39 + 40 = 738$$

$$\frac{738}{27} = 27.33$$

Identify the median.

Identify the mode.

Solution
The most frequent lengths are 25 and 27, which occur three times. Mode = 25, 27

Exercises

Use the following information to answer the next three exercises:

Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars. Calculate the following:

sample mean = $\overline{x}$ = _____

median = _____

Solution
4

mode = _____

Chapter Review from 2.6

Looking at the distribution of data can reveal a lot about the relationship between the mean, the median, and the mode. There are three types of distributions. A **right (or positive) skewed** distribution has a shape like [link]. A **left (or negative) skewed** distribution has a shape like [link]. A **symmetrical** distribution looks like [link].

Exercises

Use the following information to answer the next three exercises:

State whether the data are symmetrical, skewed to the left, or skewed to the right.

1; 1; 1; 2; 2; 2; 2; 3; 3; 3; 3; 3; 3; 3; 3; 4; 4; 4; 5; 5

Solution
The data are symmetrical. The median is 3 and the mean is 2.85. They are close, and the mode lies close to the middle of the data, so the data are symmetrical.

16; 17; 19; 22; 22; 22; 22; 22; 23

87; 87; 87; 87; 87; 88; 89; 89; 90; 91

Solution
The data are skewed right. The median is 87.5 and the mean is 88.2. Even though they are close,

the mode lies to the left of the middle of the data, and there are many more instances of 87 than any other number, so the data are skewed right.

Exercises

When the data are skewed left, what is the typical relationship between the mean and median?

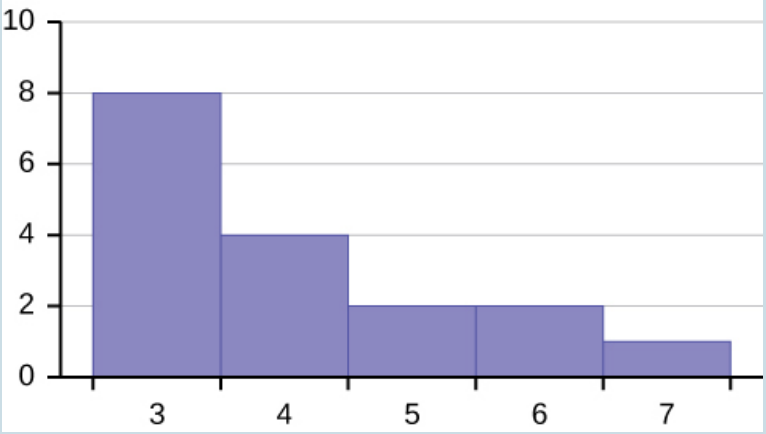
When the data are symmetrical, what is the typical relationship between the mean and median?

Solution
When the data are symmetrical, the mean and median are close or the same.

What word describes a distribution that has two modes?

Exercises

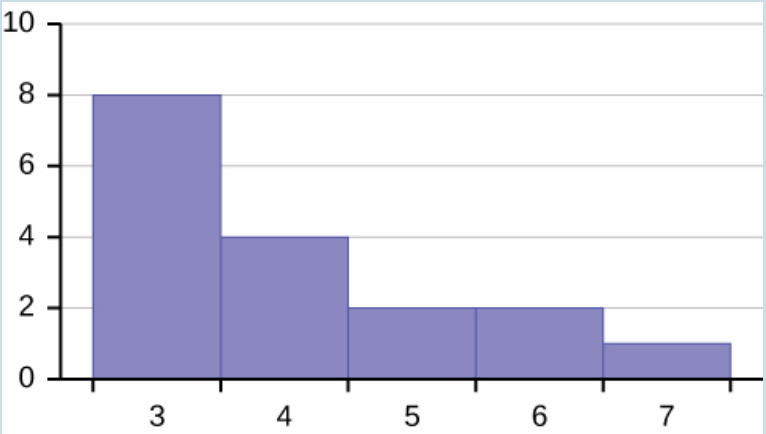
Describe the shape of this distribution.



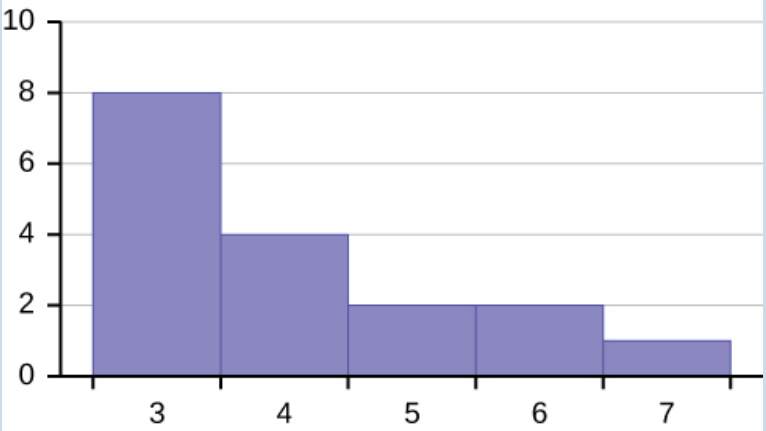
Solution

The distribution is skewed right because it looks pulled out to the right.

Describe the relationship between the mode and the median of this distribution.



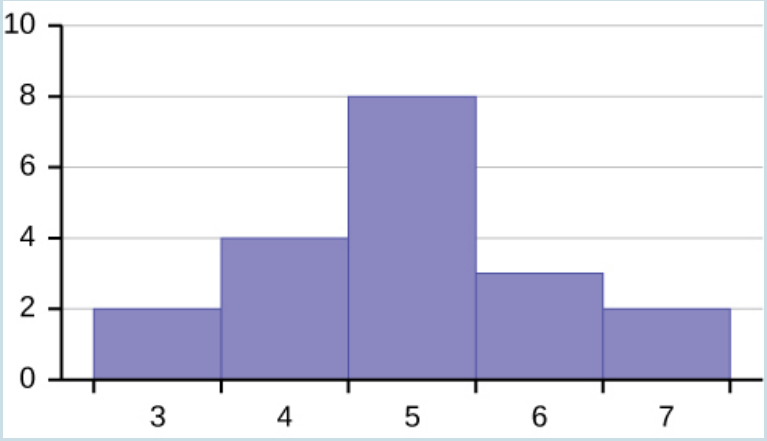
Describe the relationship between the mean and the median of this distribution.



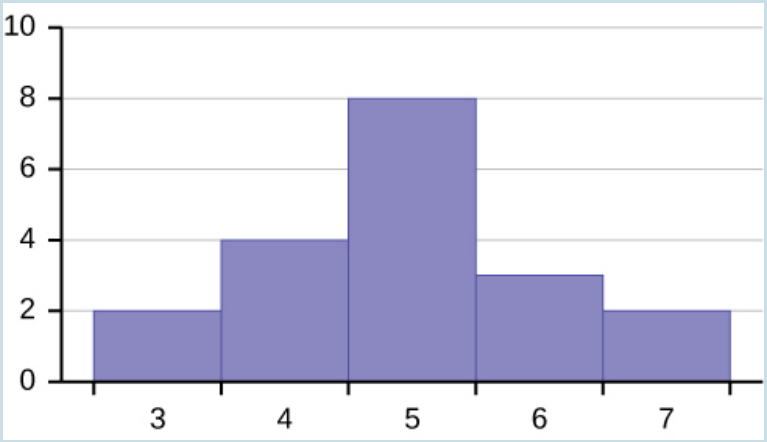
Solution

The mean is 4.1 and is slightly greater than the median, which is four.

Describe the shape of this distribution.



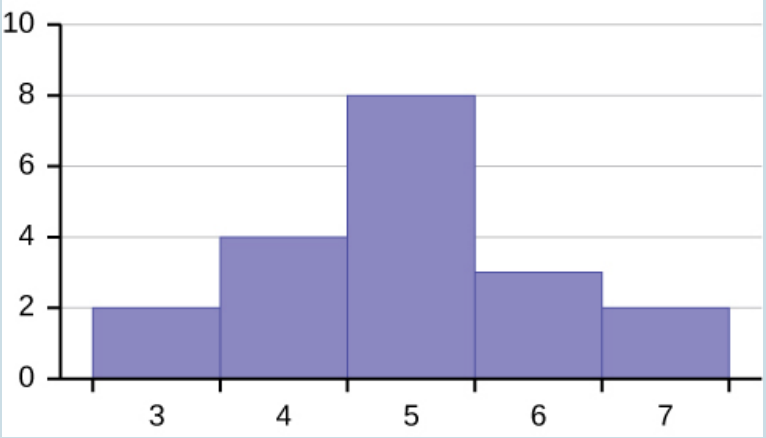
Describe the relationship between the mode and the median of this distribution.



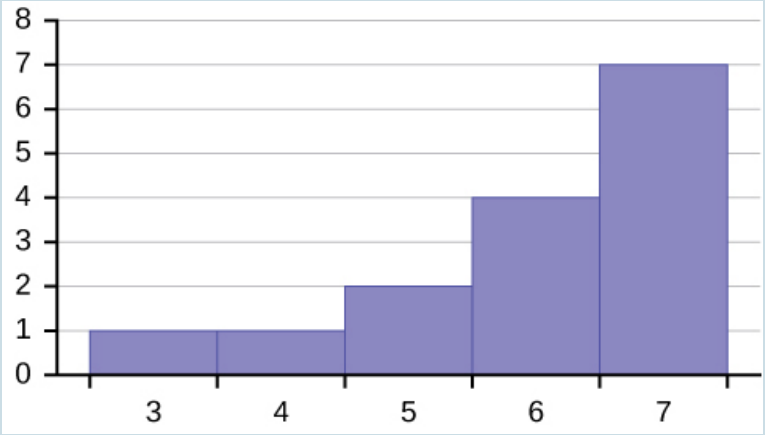
Solution

The mode and the median are the same. In this case, they are both five.

Are the mean and the median the exact same in this distribution? Why or why not?



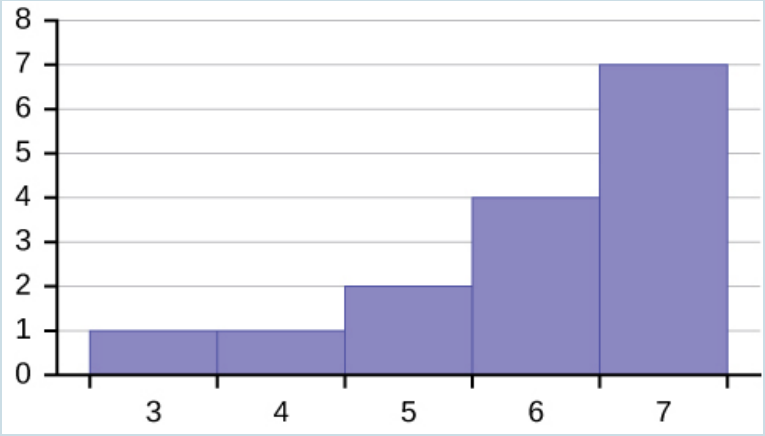
Describe the shape of this distribution.



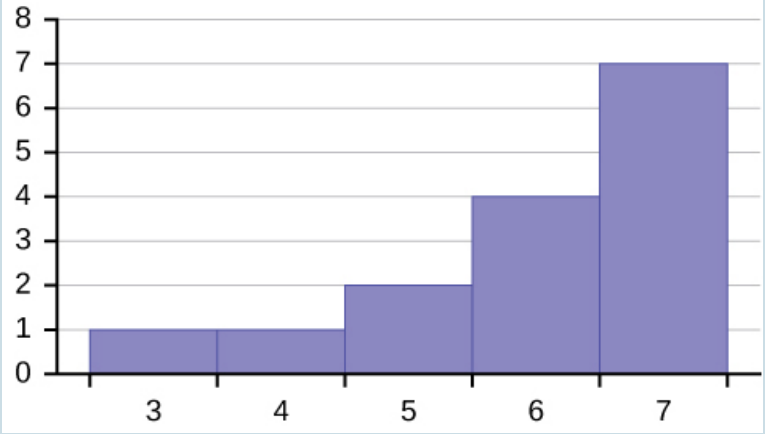
Solution

The distribution is skewed left because it looks pulled out to the left.

Describe the relationship between the mode and the median of this distribution.



Describe the relationship between the mean and the median of this distribution.



Solution

The mean and the median are both six.

Exercises

The mean and median for the data are the same.

3; 4; 5; 5; 6; 6; 6; 6; 7; 7; 7; 7; 7; 7; 7

Is the data perfectly symmetrical? Why or why not?

Exercises

Which is the greatest, the mean, the mode, or the median of the data set?

11; 11 ;12 ;12 ;12 ;12 ;13 ;15 ;17 ;22 ;22 ;22

Solution
The mode is 12, the median is 13.5, and the mean is 15.1. The mean is the largest.

Exercises

Which is the least, the mean, the mode, and the median of the data set?

56; 56; 56; 58; 59; 60; 62; 64; 64; 65; 67

Of the three measures, which tends to reflect skewing the most, the mean, the mode, or the median? Why?

Solution
The mean tends to reflect skewing the most because it is affected the most by outliers.

In a perfectly symmetrical distribution, when would the mode be different from the mean and median?

Chapter Review from 2.7

The standard deviation can help you calculate the spread of data. There are different equations to use if you are calculating the standard deviation of a sample or of a population.

- The Standard Deviation allows us to compare individual data or classes to the data set mean numerically.
- $s = \sqrt{\frac{\sum (x - \overline{x})^2}{n - 1}}$ or $s = \sqrt{\frac{\sum f(x - \overline{x})^2}{n - 1}}$ is the formula for calculating the standard deviation of a sample. To calculate the standard deviation of a population, we would use the population mean, μ, and the formula $\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\sum f(x - \mu)^2}{N}}$.

Formula Review

Sample Standard Deviation

- $s = \sqrt{\frac{\Sigma(x - \overline{x})^2}{n - 1}}$ or $s = \sqrt{\frac{\Sigma f(x - \overline{x})^2}{n - 1}}$
- For the sample standard deviation, the denominator is $\boldsymbol{n - 1}$, that is the sample size MINUS 1.

Population Standard Deviation

- $\sigma = \sqrt{\frac{\Sigma(x - \mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\Sigma f(x - \mu)^2}{N}}$
- For the population standard deviation, the denominator is N, the number of items in the population.

Exercises

Use the following information to answer the next two exercises: The following data are the distances between 20 retail stores and a large distribution center. The distances are in miles.

29; 37; 38; 40; 58; 67; 68; 69; 76; 86; 87; 95; 96; 96; 99; 106; 112; 127; 145; 150

Use a graphing calculator or computer to find the standard deviation and round to the nearest tenth.

Solution
s = 34.5

Find the value that is one standard deviation below the mean.

Exercises

Two baseball players, Fredo and Karl, on different teams wanted to find out who had the higher batting average when compared to his team. Which baseball player had the higher batting average when compared to his team?

Baseball Player	Batting Average	Team Batting Average	Team Standard Deviation
Fredo	0.158	0.166	0.012
Karl	0.177	0.189	0.015

Solution

For Fredo: $z = \frac{0.158 - 0.166}{0.012} = -0.67$

For Karl: $z = z = \frac{0.177 - 0.189}{0.015} = -0.8$

Fredo's z-score of –0.67 is higher than Karl's z-score of –0.8. For batting average, higher values are better, so Fredo has a better batting average compared to his team.

Exercises

Use [link] to find the value that is three standard deviations:

- above the mean
- below the mean

Exercises

Find the standard deviation for the following frequency tables using the formula. Check the calculations with the TI 83/84.

1.

Grade	Frequency
49.5–59.5	2
59.5–69.5	3
69.5–79.5	8
79.5–89.5	12
89.5–99.5	5

2.

Daily Low Temperature	Frequency
49.5–59.5	53
59.5–69.5	32
69.5–79.5	15
79.5–89.5	1
89.5–99.5	0

3.

Points per Game	Frequency
49.5–59.5	14
59.5–69.5	32
69.5–79.5	15
79.5–89.5	23
89.5–99.5	2

Solution

1. $s_x = \sqrt{\frac{\sum fm^2}{n} - \overline{x}^2} = \sqrt{\frac{193157.45}{30} - 79.5^2} = 10.88$

2. $s_x = \sqrt{\frac{\sum fm^2}{n} - \overline{x}^2} = \sqrt{\frac{380945.3}{101} - 60.94^2} = 7.62$
3. $s_x = \sqrt{\frac{\sum fm^2}{n} - \overline{x}^2} = \sqrt{\frac{440051.5}{86} - 70.66^2} = 11.14$

CHAPTER 2 PRACTICE

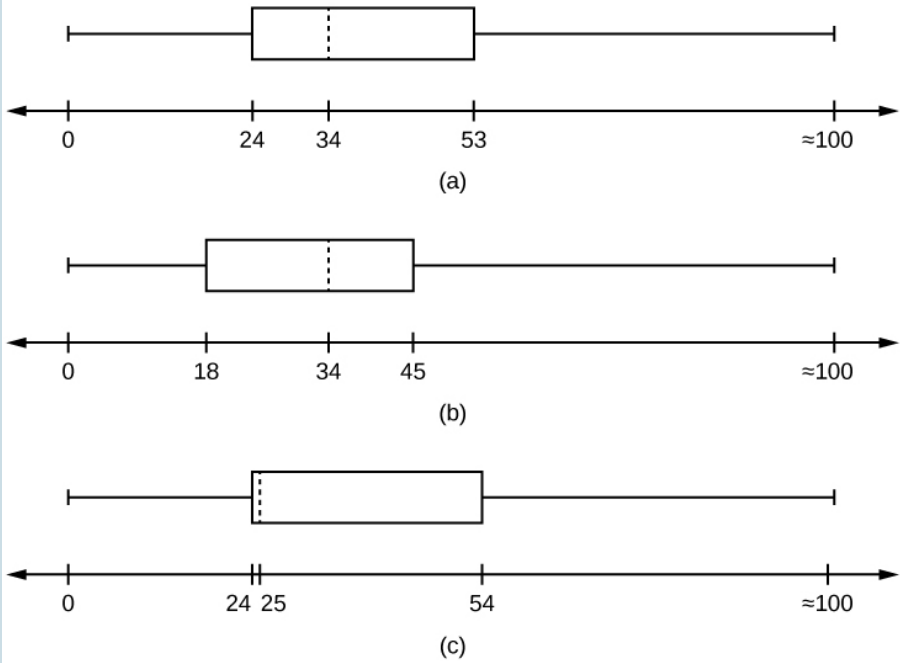
Bringing It Together from 2.4

Exercises

Santa Clara County, CA, has approximately 27,873 Japanese Americans. Their ages are as follows:

Age Group	Percentage of Community
0–17	18.9
18–24	8.0
25–34	22.8
35–44	15.0
45–54	13.1
55–64	11.9
65+	10.3

1. Construct a histogram of the Japanese American community in Santa Clara County, CA. The bars will **not** be the same width for this example. Why not? What impact does this have on the reliability of the graph?
2. What percentage of the community is under age 35?
3. Which box plot most resembles the information above?



Solution

1. For graph, check student's solution.
2. 49.7% of the community is under the age of 35.
3. Based on the information in the table, graph (a) most closely represents the data.

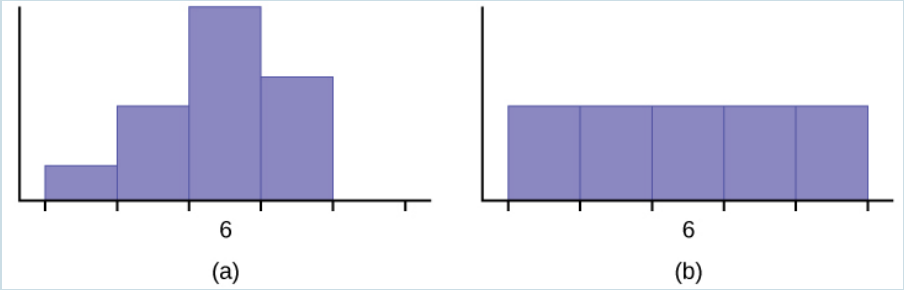
Bringing It Together from 2.5

Exercises

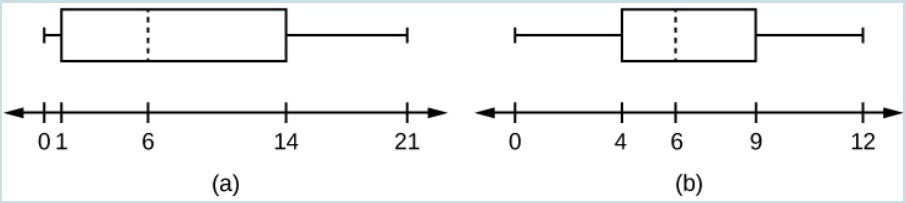
Javier and Ercilia are supervisors at a shopping mall. Each was given the task of estimating the mean distance that shoppers live from the mall. They each randomly surveyed 100 shoppers. The samples yielded the following information.

	Javier	Ercilia
$\overline{x}$	6.0 miles	6.0 miles
s	4.0 miles	7.0 miles

1. How can you determine which survey was correct?
2. Explain what the difference in the results of the surveys implies about the data.
3. If the two histograms depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know?



4. If the two box plots depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know?



Exercises

Use the following information to answer the next three exercises:

We are interested in the number of years students in a particular elementary statistics class have lived in California. The information in the following table is from the entire section.

Number of years	Frequency	Number of years	Frequency
7	1	22	1
14	3	23	1
15	1	26	1
18	1	40	2
19	4	42	2
20	3		
			Total = 20

What is the *IQR*?

a. 8
b. 11
c. 15
d. 35

Solution
a

What is the mode?

a. 19
b. 19.5
c. 14 and 20
d. 22.65

Is this a sample or the entire population?

a. sample
b. entire population
c. neither

Solution
b

Bringing It Together from 2.7

Exercises

Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows:

# of movies	Frequency
0	5
1	9
2	6
3	4
4	1

1. Find the sample mean $\overline{x}$.
2. Find the approximate sample standard deviation, *s*.

Solution

1. 1.48
2. 1.12

Exercises

Forty randomly selected students were asked the number of pairs of sneakers they owned. Let *X* = the number of pairs of sneakers owned. The results are as follows:

X	Frequency
1	2
2	5
3	8
4	12
5	12
6	0
7	1

1. Find the sample mean $\overline{x}$
2. Find the sample standard deviation, s
3. Construct a histogram of the data.
4. Complete the columns of the chart.
5. Find the first quartile.
6. Find the median.
7. Find the third quartile.
8. Construct a box plot of the data.
9. What percent of the students owned at least five pairs?
10. Find the 40^{th} percentile.
11. Find the 90^{th} percentile.
12. Construct a line graph of the data
13. Construct a stemplot of the data

Exercises

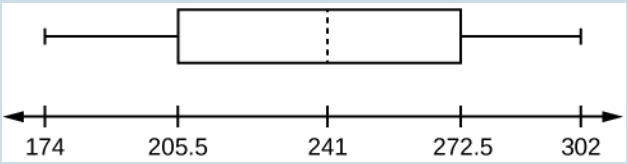
Following are the published weights (in pounds) of all of the team members of the San Francisco 49ers from a previous year.

177; 205; 210; 210; 232; 205; 185; 185; 178; 210; 206; 212; 184; 174; 185; 242; 188; 212; 215; 247; 241; 223;

220; 260; 245; 259; 278; 270; 280; 295; 275; 285; 290; 272; 273; 280; 285; 286; 200; 215; 185; 230; 250; 241; 190; 260; 250; 302; 265; 290; 276; 228; 265

1. Organize the data from smallest to largest value.
2. Find the median.
3. Find the first quartile.
4. Find the third quartile.
5. Construct a box plot of the data.
6. The middle 50% of the weights are from _____ to _____.
7. If our population were all professional football players, would the above data be a sample of weights or the population of weights? Why?
8. If our population included every team member who ever played for the San Francisco 49ers, would the above data be a sample of weights or the population of weights? Why?
9. Assume the population was the San Francisco 49ers. Find:
    1. the population mean, μ.
    2. the population standard deviation, σ.
    3. the weight that is two standard deviations below the mean.
    4. When Steve Young, quarterback, played football, he weighed 205 pounds. How many standard deviations above or below the mean was he?
10. That same year, the mean weight for the Dallas Cowboys was 240.08 pounds with a standard deviation of 44.38 pounds. Emmitt Smith weighed in at 209 pounds. With respect to his team, who was lighter, Smith or Young? How did you determine your answer?

Solution

1. 174; 177; 178; 184; 185; 185; 185; 185; 188; 190; 200; 205; 205; 206; 210; 210; 210; 212; 212; 215; 215; 220; 223; 228; 230; 232; 241; 241; 242; 245; 247; 250; 250; 259; 260; 260; 265; 265; 270; 272; 273; 275; 276; 278; 280; 280; 285; 285; 286; 290; 290; 295; 302
2. 241
3. 205.5
4. 272.5
5. 

6. 205.5, 272.5
7. sample

8. population
9. The values are:
    1. 236.34
    2. 37.50
    3. 161.34
    4. 0.84 std. dev. below the mean
10. Young

Exercises

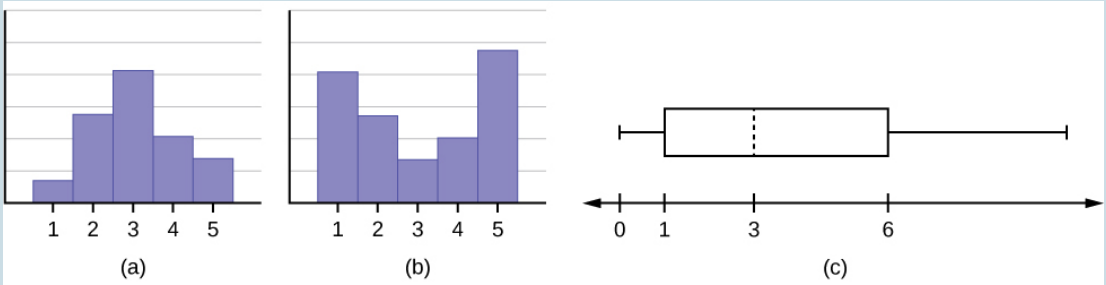
One hundred teachers attended a seminar on mathematical problem solving. The attitudes of a representative sample of 12 of the teachers were measured before and after the seminar. A positive number for change in attitude indicates that a teacher's attitude toward math became more positive. The 12 change scores are as follows:

3; 8; –1; 2; 0; 5; –3; 1; –1; 6; 5; –2

1. What is the mean change score?
2. What is the standard deviation for this population?
3. What is the median change score?
4. Find the change score that is 2.2 standard deviations below the mean.

Exercises

Refer to [link] determine which of the following are true and which are false. Explain your solution to each part in complete sentences.



1. The medians for all three graphs are the same.
2. We cannot determine if any of the means for the three graphs is different.
3. The standard deviation for graph b is larger than the standard deviation for graph a.
4. We cannot determine if any of the third quartiles for the three graphs is different.

Solution

1. True
2. True
3. True
4. False

Exercises

In a recent issue of the IEEE SPECTRUM, 84 engineering conferences were announced. Four conferences lasted two days. Thirty-six lasted three days. Eighteen lasted four days. Nineteen lasted five days. Four lasted six days. One lasted seven days. One lasted eight days. One lasted nine days. Let X = the length (in days) of an engineering conference.

1. Organize the data in a chart.
2. Find the median, the first quartile, and the third quartile.
3. Find the 65^{th} percentile.

4. Find the 10^{th} percentile.
5. Construct a box plot of the data.
6. The middle 50% of the conferences last from _____ days to _____ days.
7. Calculate the sample mean of days of engineering conferences.
8. Calculate the sample standard deviation of days of engineering conferences.
9. Find the mode.
10. If you were planning an engineering conference, which would you choose as the length of the conference: mean; median; or mode? Explain why you made that choice.
11. Give two reasons why you think that three to five days seem to be popular lengths of engineering conferences.

Exercises

A survey of enrollment at 35 community colleges across the United States yielded the following figures:

6414; 1550; 2109; 9350; 21828; 4300; 5944; 5722; 2825; 2044; 5481; 5200; 5853; 2750; 10012; 6357; 27000; 9414; 7681; 3200; 17500; 9200; 7380; 18314; 6557; 13713; 17768; 7493; 2771; 2861; 1263; 7285; 28165; 5080; 11622

1. Organize the data into a chart with five intervals of equal width. Label the two columns "Enrollment" and "Frequency."
2. Construct a histogram of the data.
3. If you were to build a new community college, which piece of information would be more valuable: the mode or the mean?
4. Calculate the sample mean.
5. Calculate the sample standard deviation.
6. A school with an enrollment of 8,000 would be how many standard deviations away from the mean?

Solution

1.

Enrollment	Frequency
1000-5000	10
5000-10000	16
10000-15000	3
15000-20000	3
20000-25000	1
25000-30000	2

2. Check student's solution.
3. mode
4. 8628.74
5. 6943.88
6. –0.09

Exercises

Use the following information to answer the next two exercises. X = the number of days per week that 100 clients use a particular exercise facility.

x	Frequency
0	3
1	12
2	33
3	28
4	11
5	9
6	4

The 80^{th} percentile is _____

a. 5
b. 80
c. 3
d. 4

The number that is 1.5 standard deviations BELOW the mean is approximately _____

a. 0.7
b. 4.8
c. –2.8
d. Cannot be determined

Solution
a

Exercises

Suppose that a publisher conducted a survey asking adult consumers the number of fiction paperback books they had purchased in the previous month. The results are summarized in the [link].

# of books	Freq.	Rel. Freq.
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

1. Are there any outliers in the data? Use an appropriate numerical test involving the *IQR* to identify outliers, if any, and clearly state your conclusion.
2. If a data value is identified as an outlier, what should be done about it?
3. Are any data values further than two standard deviations away from the mean? In some situations, statisticians may use this criteria to identify data values that are unusual, compared to the other data values. (Note that this criteria is most appropriate to use for data that is mound-shaped and symmetric, rather than for skewed data.)
4. Do parts a and c of this problem give the same answer?
5. Examine the shape of the data. Which part, a or c, of this question gives a more appropriate result for this data?
6. Based on the shape of the data which is the most appropriate measure of center for this data: mean, median, or mode?

CHAPTER 2 HOMEWORK

Homework from 2.1

Exercises

Student grades on a chemistry exam were: 77, 78, 76, 81, 86, 51, 79, 82, 84, 99

1. Construct a stem-and-leaf plot of the data.
2. Are there any potential outliers? If so, which scores are they? Why do you consider them outliers?

Exercises

[link] contains the 2010 obesity rates in U.S. states and Washington, DC.

State	Percent (%)	State	Percent (%)	State	Percent (%)
Alabama	32.2	Kentucky	31.3	North Dakota	27.2
Alaska	24.5	Louisiana	31.0	Ohio	29.2
Arizona	24.3	Maine	26.8	Oklahoma	30.4
Arkansas	30.1	Maryland	27.1	Oregon	26.8
California	24.0	Massachusetts	23.0	Pennsylvania	28.6
Colorado	21.0	Michigan	30.9	Rhode Island	25.5
Connecticut	22.5	Minnesota	24.8	South Carolina	31.5
Delaware	28.0	Mississippi	34.0	South Dakota	27.3
Washington, DC	22.2	Missouri	30.5	Tennessee	30.8
Florida	26.6	Montana	23.0	Texas	31.0
Georgia	29.6	Nebraska	26.9	Utah	22.5
Hawaii	22.7	Nevada	22.4	Vermont	23.2
Idaho	26.5	New Hampshire	25.0	Virginia	26.0
Illinois	28.2	New Jersey	23.8	Washington	25.5
Indiana	29.6	New Mexico	25.1	West Virginia	32.5
Iowa	28.4	New York	23.9	Wisconsin	26.3
Kansas	29.4	North Carolina	27.8	Wyoming	25.1

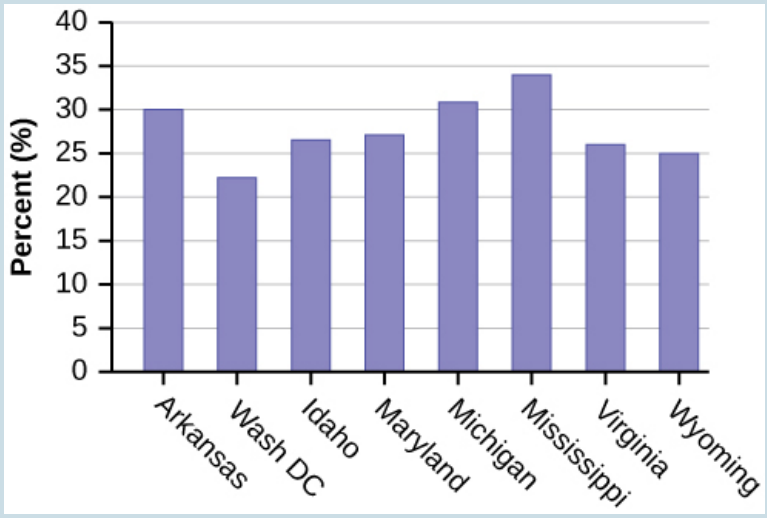
1. Use a random number generator to randomly pick eight states. Construct a bar graph of the obesity rates of those eight states.
2. Construct a bar graph for all the states beginning with the letter "A."
3. Construct a bar graph for all the states beginning with the letter "M."

Solution

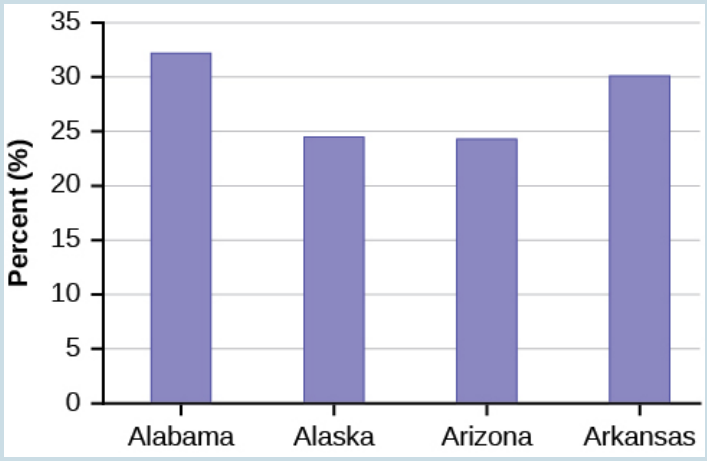
1. Example solution for using the random number generator for the TI-84+ to generate a simple random sample of 8 states. Instructions are as follows.
    - Number the entries in the table 1–51 (Includes Washington, DC; Numbered vertically)
    - Press MATH
    - Arrow over to PRB
    - Press 5:randInt(
    - Enter 51,1,8)

Eight numbers are generated (use the right arrow key to scroll through the numbers). The numbers correspond to the numbered states (for this example: {47 21 9 23 51 13 25 4}. If any numbers are repeated, generate a different number by using 5:randInt(51,1)). Here, the states (and Washington DC) are {Arkansas, Washington DC, Idaho, Maryland, Michigan, Mississippi, Virginia, Wyoming}.

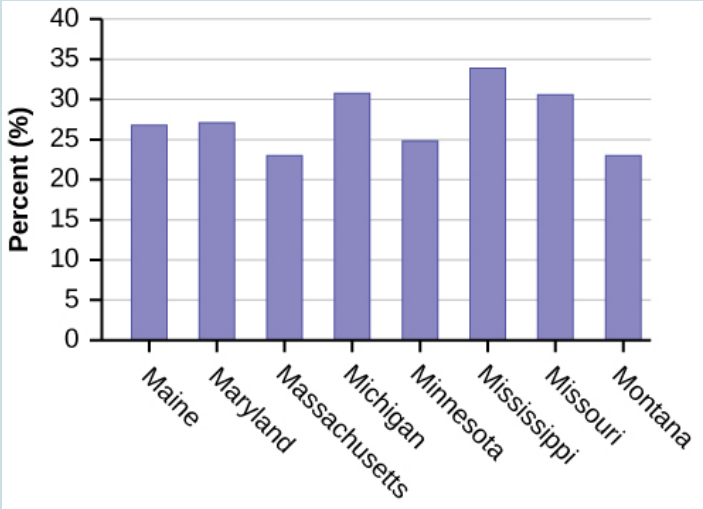
Corresponding percentages are {30.1, 22.2, 26.5, 27.1, 30.9, 34.0, 26.0, 25.1}.



2.



3.



Homework from 2.2

Exercises

Suppose that three book publishers were interested in the number of fiction paperbacks adult consumers purchase per month. Each publisher conducted a survey. In the survey, adult consumers were asked the number of fiction paperbacks they had purchased the previous month. The results are as follows:

Publisher A

# of books	Freq.	Rel. Freq.
0	10	
1	12	
2	16	
3	12	
4	8	
5	6	
6	2	
8	2	

Publisher B

# of books	Freq.	Rel. Freq.
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

Publisher C

# of books	Freq.	Rel. Freq.
0–1	20	
2–3	35	
4–5	12	
6–7	2	
8–9	1	

1. Find the relative frequencies for each survey. Write them in the charts.
2. Using either a graphing calculator, computer, or by hand, use the frequency column to construct a histogram for each publisher's survey. For Publishers A and B, make bar widths

of one. For Publisher C, make bar widths of two.

3. In complete sentences, give two reasons why the graphs for Publishers A and B are not identical.
4. Would you have expected the graph for Publisher C to look like the other two graphs? Why or why not?
5. Make new histograms for Publisher A and Publisher B. This time, make bar widths of two.
6. Now, compare the graph for Publisher C to the new graphs for Publishers A and B. Are the graphs more similar or more different? Explain your answer.

Exercises

Often, cruise ships conduct all on-board transactions, with the exception of gambling, on a cashless basis. At the end of the cruise, guests pay one bill that covers all onboard transactions. Suppose that 60 single travelers and 70 couples were surveyed as to their on-board bills for a seven-day cruise from Los Angeles to the Mexican Riviera. Following is a summary of the bills for each group.

	Singles	
Amount($)	**Frequency**	**Rel. Frequency**
51–100	5	
101–150	10	
151–200	15	
201–250	15	
251–300	10	
301–350	5	

Couples		
Amount($)	**Frequency**	**Rel. Frequency**
100–150	5	
201–250	5	
251–300	5	
301–350	5	
351–400	10	
401–450	10	
451–500	10	
501–550	10	
551–600	5	
601–650	5	

1. Fill in the relative frequency for each group.
2. Construct a histogram for the singles group. Scale the *x*-axis by $50 widths. Use relative frequency on the *y*-axis.
3. Construct a histogram for the couples group. Scale the *x*-axis by $50 widths. Use relative frequency on the *y*-axis.
4. Compare the two graphs:
    1. List two similarities between the graphs.
    2. List two differences between the graphs.
    3. Overall, are the graphs more similar or different?
5. Construct a new graph for the couples by hand. Since each couple is paying for two individuals, instead of scaling the *x*-axis by $50, scale it by $100. Use relative frequency on the *y*-axis.
6. Compare the graph for the singles with the new graph for the couples:
    1. List two similarities between the graphs.
    2. Overall, are the graphs more similar or different?
7. How did scaling the couples graph differently change the way you compared it to the singles graph?
8. Based on the graphs, do you think that individuals spend the same amount, more or less, as singles as they do person by person as a couple? Explain why in one or two complete

sentences.

Solution

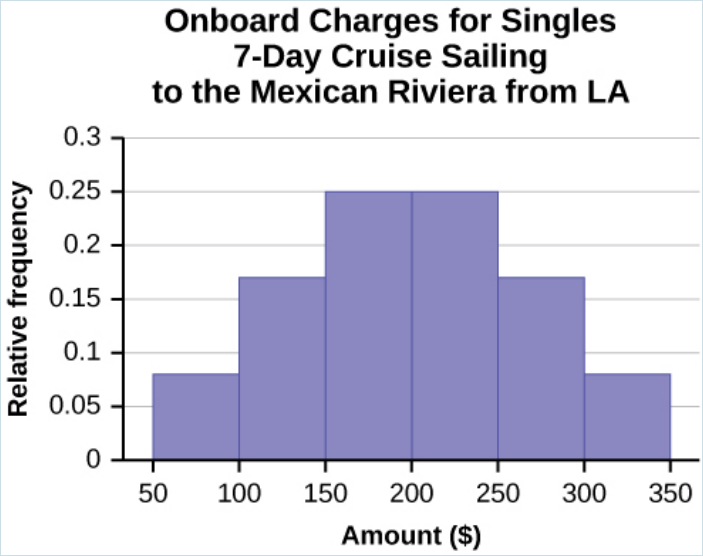
Singles

Amount($)	Frequency	Relative Frequency
51–100	5	0.08
101–150	10	0.17
151–200	15	0.25
201–250	15	0.25
251–300	10	0.17
301–350	5	0.08

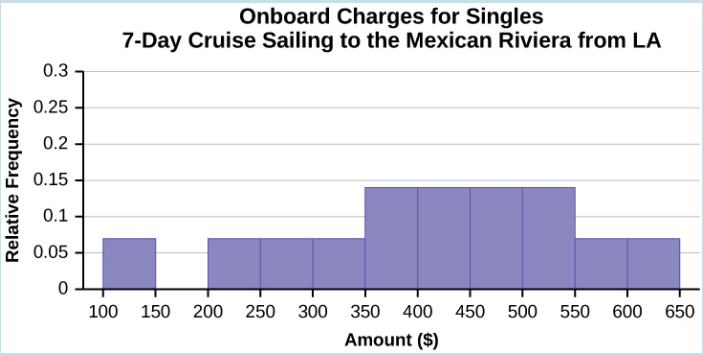
Couples

Amount($)	Frequency	Relative Frequency
100–150	5	0.07
201–250	5	0.07
251–300	5	0.07
301–350	5	0.07
351–400	10	0.14
401–450	10	0.14
451–500	10	0.14
501–550	10	0.14
551–600	5	0.07
601–650	5	0.07

1. See [link] and [link].
2. In the following histogram data values that fall on the right boundary are counted in the class interval, while values that fall on the left boundary are not counted (with the exception of the first interval where both boundary values are included).



3. In the following histogram, the data values that fall on the right boundary are counted in the class interval, while values that fall on the left boundary are not counted (with the exception of the first interval where values on both boundaries are included).



4. Compare the two graphs:

    1. Answers may vary. Possible answers include:

        - Both graphs have a single peak.
        - Both graphs use class intervals with width equal to $50.

    2. Answers may vary. Possible answers include:

        - The couples graph has a class interval with no values.
        - It takes almost twice as many class intervals to display the data for couples.

    3. Answers may vary. Possible answers include: The graphs are more similar than different because the overall patterns for the graphs are the same.

5. Check student's solution.
6. Compare the graph for the Singles with the new graph for the Couples:

    - Both graphs have a single peak.
    - Both graphs display 6 class intervals.
    - Both graphs show the same general pattern.

2. Answers may vary. Possible answers include: Although the width of the class intervals for couples is double that of the class intervals for singles, the graphs are more similar than they are different.
3. Answers may vary. Possible answers include: You are able to compare the graphs interval by interval. It is easier to compare the overall patterns with the new scale on the Couples graph. Because a couple represents two individuals, the new scale leads to a more accurate comparison.
4. Answers may vary. Possible answers include: Based on the histograms, it seems that spending does not vary much from singles to individuals who are part of a couple. The overall patterns are the same. The range of spending for couples is approximately double the range for individuals.

Exercises

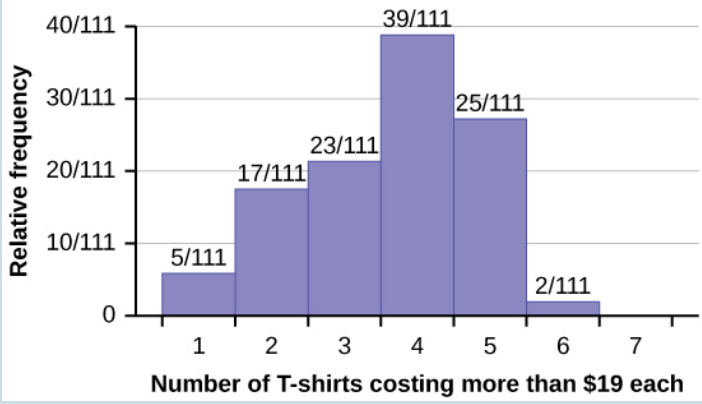
Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows.

# of movies	Frequency	Relative Frequency	Cumulative Relative Frequency
0	5		
1	9		
2	6		
3	4		
4	1		

1. Construct a histogram of the data.
2. Complete the columns of the chart.

Exercises

Use the following information to answer the next two exercises: Suppose one hundred eleven people who shopped in a special T-shirt store were asked the number of T-shirts they own costing more than $19 each.



The percentage of people who own at most three T-shirts costing more than $19 each is approximately:

1. 21
2. 59
3. 41
4. Cannot be determined

Solution
c

If the data were collected by asking the first 111 people who entered the store, then the type of sampling is:

1. cluster
2. simple random
3. stratified
4. convenience

Exercises

Following are the 2010 obesity rates by U.S. states and Washington, DC.

State	Percent (%)	State	Percent (%)	State	Percent (%)
Alabama	32.2	Kentucky	31.3	North Dakota	27.2
Alaska	24.5	Louisiana	31.0	Ohio	29.2
Arizona	24.3	Maine	26.8	Oklahoma	30.4
Arkansas	30.1	Maryland	27.1	Oregon	26.8
California	24.0	Massachusetts	23.0	Pennsylvania	28.6
Colorado	21.0	Michigan	30.9	Rhode Island	25.5
Connecticut	22.5	Minnesota	24.8	South Carolina	31.5
Delaware	28.0	Mississippi	34.0	South Dakota	27.3
Washington, DC	22.2	Missouri	30.5	Tennessee	30.8
Florida	26.6	Montana	23.0	Texas	31.0
Georgia	29.6	Nebraska	26.9	Utah	22.5
Hawaii	22.7	Nevada	22.4	Vermont	23.2
Idaho	26.5	New Hampshire	25.0	Virginia	26.0
Illinois	28.2	New Jersey	23.8	Washington	25.5
Indiana	29.6	New Mexico	25.1	West Virginia	32.5
Iowa	28.4	New York	23.9	Wisconsin	26.3
Kansas	29.4	North Carolina	27.8	Wyoming	25.1

Construct a bar graph of obesity rates of your state and the four states closest to your state. Hint: Label the *x*-axis with the states.

Solution

Answers will vary.

Homework from 2.3

The median age for U.S. blacks currently is 30.9 years; for U.S. whites it is 42.3 years.

Based upon this information, give two reasons why the black median age could be lower than the white median age.

Does the lower median age for blacks necessarily mean that blacks die younger than whites? Why or why not?

How might it be possible for blacks and whites to die at approximately the same age, but for the median age for whites to be higher?

Exercises

Six hundred adult Americans were asked by telephone poll, "What do you think constitutes a middle-class income?" The results are in [link]. Also, include left endpoint, but not the right endpoint.

Salary ($)	Relative Frequency
< 20,000	0.02
20,000–25,000	0.09
25,000–30,000	0.19
30,000–40,000	0.26
40,000–50,000	0.18
50,000–75,000	0.17
75,000–99,999	0.02
100,000+	0.01

1. What percentage of the survey answered "not sure"?
2. What percentage think that middle-class is from $25,000 to $50,000?
3. Construct a histogram of the data.
    1. Should all bars have the same width, based on the data? Why or why not?
    2. How should the <20,000 and the 100,000+ intervals be handled? Why?
4. Find the 40^{th} and 80^{th} percentiles
5. Construct a bar graph of the data

Solution

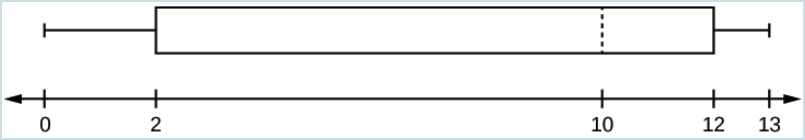
1. 1 – (0.02+0.09+0.19+0.26+0.18+0.17+0.02+0.01) = 0.06
2. 0.19+0.26+0.18 = 0.63
3. Check student's solution.
4. 40^{th} percentile will fall between 30,000 and 40,000

   80^{th} percentile will fall between 50,000 and 75,000

5. Check student's solution.

Exercises

Given the following box plot:

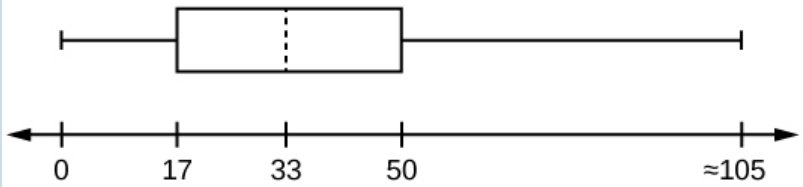


1. which quarter has the smallest spread of data? What is that spread?
2. which quarter has the largest spread of data? What is that spread?
3. find the interquartile range (*IQR*).
4. are there more data in the interval 5–10 or in the interval 10–13? How do you know this?
5. which interval has the fewest data in it? How do you know this?

    1. 0–2
    2. 2–4
    3. 10–12
    4. 12–13
    5. need more information

Exercises

The following box plot shows the U.S. population for 1990, the latest available year.



1. Are there fewer or more children (age 17 and under) than senior citizens (age 65 and over)? How do you know?
2. 12.6% are age 65 and over. Approximately what percentage of the population are working age adults (above age 17 to age 65)?

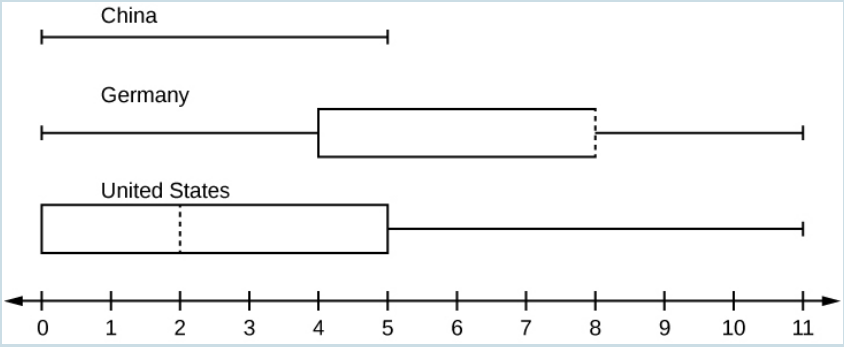
Solution

1. more children; the left whisker shows that 25% of the population are children 17 and younger. The right whisker shows that 25% of the population are adults 50 and older, so adults 65 and over represent less than 25%.
2. 62.4%

Homework from 2.4

Exercises

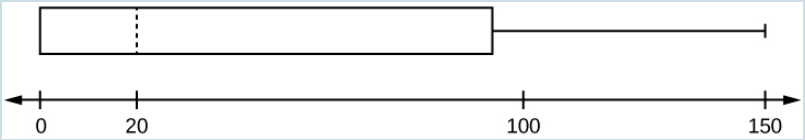
In a survey of 20-year-olds in China, Germany, and the United States, people were asked the number of foreign countries they had visited in their lifetime. The following box plots display the results.



1. In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected.
2. Have more Americans or more Germans surveyed been to over eight foreign countries?
3. Compare the three box plots. What do they imply about the foreign travel of 20-year-old residents of the three countries when compared to each other?

Exercises

Given the following box plot, answer the questions.



1. Think of an example (in words) where the data might fit into the above box plot. In 2–5 sentences, write down the example.
2. What does it mean to have the first and second quartiles so close together, while the second to third quartiles are far apart?

Solution

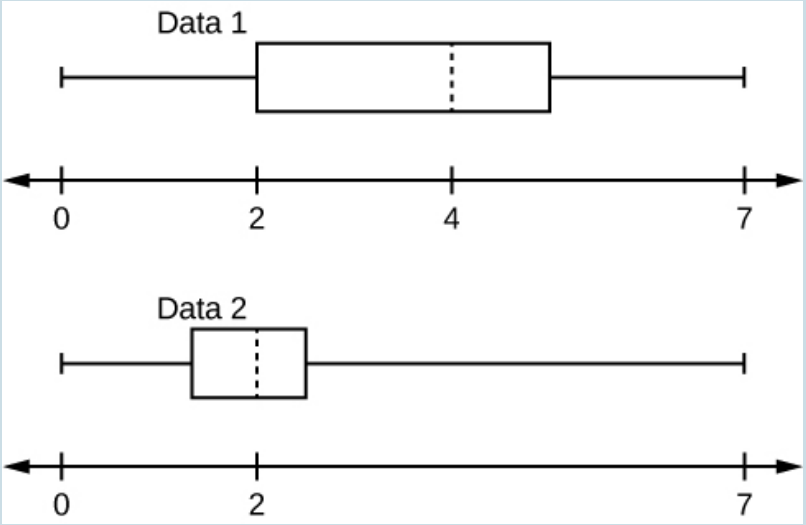
1. Answers will vary. Possible answer: State University conducted a survey to see how involved its students are in community service. The box plot shows the number of community service

hours logged by participants over the past year.

2. Because the first and second quartiles are close, the data in this quarter is very similar. There is not much variation in the values. The data in the third quarter is much more variable or spread out. This is clear because the second quartile is so far away from the third quartile.

Exercises

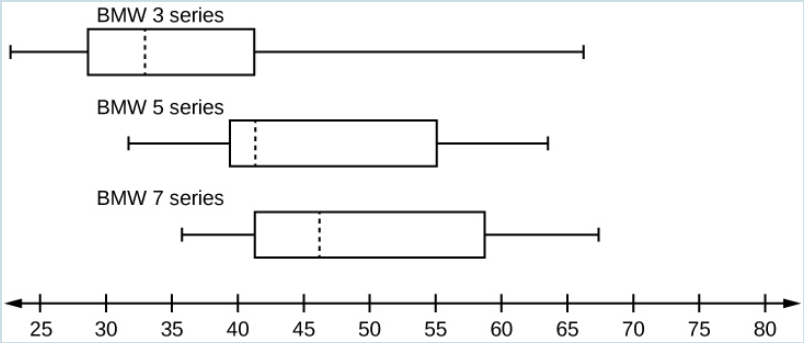
Given the following box plots, answer the questions.



1. In complete sentences, explain why each statement is false.
    1. **Data 1** has more data values above two than **Data 2** has above two.
    2. The data sets cannot have the same mode.
    3. For **Data 1**, there are more data values below four than there are above four.
2. For which group, Data 1 or Data 2, is the value of "7" more likely to be an outlier? Explain why in complete sentences.

Exercises

A survey was conducted of 130 purchasers of new BMW 3 series cars, 130 purchasers of new BMW 5 series cars, and 130 purchasers of new BMW 7 series cars. In it, people were asked the age they were when they purchased their car. The following box plots display the results.



1. In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected for that car series.
2. Which group is most likely to have an outlier? Explain how you determined that.
3. Compare the three box plots. What do they imply about the age of purchasing a BMW from the series when compared to each other?
4. Look at the BMW 5 series. Which quarter has the smallest spread of data? What is the spread?
5. Look at the BMW 5 series. Which quarter has the largest spread of data? What is the spread?
6. Look at the BMW 5 series. Estimate the interquartile range (IQR).
7. Look at the BMW 5 series. Are there more data in the interval 31 to 38 or in the interval 45 to 55? How do you know this?
8. Look at the BMW 5 series. Which interval has the fewest data in it? How do you know this?

    1. 31–35
    2. 38–41
    3. 41–64

Solution

1. Each box plot is spread out more in the greater values. Each plot is **skewed** to the right, so

the ages of the top 50% of buyers are more variable than the ages of the lower 50%.

2. The BMW 3 series is most likely to have an outlier. It has the longest whisker.
3. Comparing the median ages, younger people tend to buy the BMW 3 series, while older people tend to buy the BMW 7 series. However, this is not a rule, because there is so much variability in each data set.
4. The second quarter has the smallest spread. There seems to be only a three-year difference between the first quartile and the median.
5. The third quarter has the largest spread. There seems to be approximately a 14-year difference between the median and the third quartile.
6. *IQR* ~ 17 years
7. There is not enough information to tell. Each interval lies within a quarter, so we cannot tell exactly where the data in that quarter is concentrated.
8. The interval from 31 to 35 years has the fewest data values. Twenty-five percent of the values fall in the interval 38 to 41, and 25% fall between 41 and 64. Since 25% of values fall between 31 and 38, we know that fewer than 25% fall between 31 and 35.

Exercises

Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows:

# of movies	Frequency
0	5
1	9
2	6
3	4
4	1

Construct a box plot of the data.

Homework from 2.5

Exercises

The most obese countries in the world have obesity rates that range from 11.4% to 74.6%. This data is summarized in the following table.

Percent of Population Obese	Number of Countries
11.4–20.45	29
20.45–29.45	13
29.45–38.45	4
38.45–47.45	0
47.45–56.45	2
56.45–65.45	1
65.45–74.45	0
74.45–83.45	1

1. What is the best estimate of the average obesity percentage for these countries?
2. The United States has an average obesity rate of 33.9%. Is this rate above average or below?
3. How does the United States compare to other countries?

Exercises

[link] gives the percent of children under five considered to be underweight. What is the best estimate for the mean percentage of underweight children?

Percent of Underweight Children	Number of Countries
16–21.45	23
21.45–26.9	4
26.9–32.35	9
32.35–37.8	7
37.8–43.25	6
43.25–48.7	1

Solution

The mean percentage, $\overline{x} = \frac{1328.65}{50} = 26.75$

Homework from 2.6

Exercises

The median age of the U.S. population in 1980 was 30.0 years. In 1991, the median age was 33.1 years.

1. What does it mean for the median age to rise?
2. Give two reasons why the median age could rise.
3. For the median age to rise, is the actual number of children less in 1991 than it was in 1980?

Why or why not?

Homework from 2.7

Exercises

Use the following information to answer the next nine exercises:

The population parameters below describe the full-time equivalent number of students (FTES) each year at Lake Tahoe Community College from 1976–1977 through 2004–2005.

- μ = 1000 FTES
- median = 1,014 FTES
- σ = 474 FTES
- first quartile = 528.5 FTES
- third quartile = 1,447.5 FTES
- n = 29 years

A sample of 11 years is taken. About how many are expected to have a FTES of 1014 or above? Explain how you determined your answer.

Solution
The median value is the middle value in the ordered list of data values. The median value of a set of 11 will be the 6th number in order. Six years will have totals at or below the median.

75% of all years have an FTES:

at or below: _____
at or above: _____

The population standard deviation = _____

Solution

474 FTES

What percentage of the FTES was from 528.5 to 1447.5? How do you know?

What is the *IQR*? What does the *IQR* represent?

Solution
919

How many standard deviations away from the mean is the median?

Additional Information: The population FTES for 2005–2006 through 2010–2011 was given in an updated report. The data are reported here.

Year	2005–6	2006–7	2007–8	2008–9	2009–10	2010–11
Total FTES	1,585	1,690	1,735	1,935	2,021	1,890

Calculate the mean, median, standard deviation, the first quartile, the third quartile and the *IQR*. Round to one decimal place.

Solution

- mean = 1,809.3
- median = 1,812.5
- standard deviation = 151.2
- first quartile = 1,690
- third quartile = 1,935
- *IQR* = 245

Construct a box plot for the FTES for 2005–2006 through 2010–2011 and a box plot for the FTES for 1976–1977 through 2004–2005.

Compare the *IQR* for the FTES for 1976–77 through 2004–2005 with the *IQR* for the FTES for 2005-2006 through 2010–2011. Why do you suppose the *IQR*s are so different?

Solution
Hint: Think about the number of years covered by each time period and what happened to higher education during those periods.

Exercises

Three students were applying to the same graduate school. They came from schools with different grading systems. Which student had the best GPA when compared to other students at his school? Explain how you determined your answer.

Student	GPA	School Average GPA	School Standard Deviation
Thuy	2.7	3.2	0.8
Vichet	87	75	20
Kamala	8.6	8	0.4

Exercises

A music school has budgeted to purchase three musical instruments. They plan to purchase a piano costing $3,000, a guitar costing $550, and a drum set costing $600. The mean cost for a piano is $4,000 with a standard deviation of $2,500. The mean cost for a guitar is $500 with a standard deviation of $200. The mean cost for drums is $700 with a standard deviation of $100. Which cost is the lowest, when compared to other instruments of the same type? Which cost is the highest when compared to other instruments of the same type. Justify your answer.

Solution

For pianos, the cost of the piano is 0.4 standard deviations BELOW the mean. For guitars, the cost of the guitar is 0.25 standard deviations ABOVE the mean. For drums, the cost of the drum set is 1.0 standard deviations BELOW the mean. Of the three, the drums cost the lowest in comparison to

the cost of other instruments of the same type. The guitar costs the most in comparison to the cost of other instruments of the same type.

Exercises

An elementary school class ran one mile with a mean of 11 minutes and a standard deviation of three minutes. Rachel, a student in the class, ran one mile in eight minutes. A junior high school class ran one mile with a mean of nine minutes and a standard deviation of two minutes. Kenji, a student in the class, ran 1 mile in 8.5 minutes. A high school class ran one mile with a mean of seven minutes and a standard deviation of four minutes. Nedda, a student in the class, ran one mile in eight minutes.

1. Why is Kenji considered a better runner than Nedda, even though Nedda ran faster than him?
2. Who is the fastest runner with respect to his or her class? Explain why.

Exercises

The most obese countries in the world have obesity rates that range from 11.4% to 74.6%. This data is summarized in Table 14.

Percent of Population Obese	Number of Countries
11.4–20.45	29
20.45–29.45	13
29.45–38.45	4
38.45–47.45	0
47.45–56.45	2
56.45–65.45	1
65.45–74.45	0
74.45–83.45	1

What is the best estimate of the average obesity percentage for these countries? What is the standard deviation for the listed obesity rates? The United States has an average obesity rate of 33.9%. Is this rate above average or below? How "unusual" is the United States' obesity rate compared to the average rate? Explain.

Solution

- $\overline{x} = 23.32$
- Using the TI 83/84, we obtain a standard deviation of: $s_x = 12.95.$
- The obesity rate of the United States is 10.58% higher than the average obesity rate.
- Since the standard deviation is 12.95, we see that 23.32 + 12.95 = 36.27 is the obesity percentage that is one standard deviation from the mean. The United States obesity rate is slightly less than one standard deviation from the mean. Therefore, we can assume that the United States, while 34% obese, does not have an unusually high percentage of obese people.

Exercises

[link] gives the percent of children under five considered to be underweight.

Percent of Underweight Children	Number of Countries
16–21.45	23
21.45–26.9	4
26.9–32.35	9
32.35–37.8	7
37.8–43.25	6
43.25–48.7	1

What is the best estimate for the mean percentage of underweight children? What is the standard deviation? Which interval(s) could be considered unusual? Explain.

CHAPTER 3: PROBABILITY TOPICS

INTRODUCTION TO CHAPTER 3: PROBABILITY TOPICS



Figure 3.1. Meteor showers are rare, but the probability of them occurring can be calculated. (credit: Navicore/flickr)

It is often necessary to "guess" about the outcome of an event in order to make a decision. Politicians study polls to guess their likelihood of winning an election. Teachers choose a particular course of study based on what they think students can comprehend. Doctors choose the treatments needed for various diseases based on their assessment of likely results. In a casino, people choose to play games because of the belief that the likelihood of winning is good. You may have chosen your course of study based on the probable availability of jobs.

You have, more than likely, used probability. In fact, you probably have an intuitive sense of probability. Probability deals with the chance of an event occurring. Whenever you weigh the odds of whether or not to do your homework or to study for an exam, you are using probability. In this chapter, you will learn how to solve probability problems using a systematic approach.

Collaborative Exercises

Section 3.1

Collaborative Exercise

Your instructor will survey your class. Count the number of students in the class today.

- Raise your hand if you have any change in your pocket or purse. Record the number of raised hands.
- Raise your hand if you rode a bus within the past month. Record the number of raised hands.
- Raise your hand if you answered "yes" to BOTH of the first two questions. Record the number of raised hands.

Use the class data as estimates of the following probabilities. *P*(change) means the probability that a randomly chosen person in your class has change in their pocket or purse. *P*(bus) means the probability that a randomly chosen person in your class rode a bus within the last month and so on. Discuss your answers.

- Find *P*(change).
- Find *P*(bus).
- Find *P*(change AND bus). Find the probability that a randomly chosen student in your class has change in his/her pocket or purse and rode a bus within the last month.
- Find *P*(change|bus). Find the probability that a randomly chosen student has change given that he or she rode a bus within the last month. Count all the students that rode a bus. From the group of students who rode a bus, count those who have change. The probability is equal to those who have change and rode a bus divided by those who rode a bus.

Media Attributions

- Chapter 3 Introduction Figure © OpenStax Introductory Statistics is licensed under a CC BY (Attribution) license

3.1 TERMINOLOGY

Learning Objectives

By the end of this section, you should be able to:

- Define and use common probability terms
- Compute probabilities for common experiments with equally likely outcomes
- Combine events and compute the resulting probabilities
- Compute conditional probabilities

Probability is a measure that is associated with how certain we are of outcomes of a particular experiment or activity. An experiment is a planned operation carried out under controlled conditions. If the result is not predetermined, then the experiment is said to be a **probability experiment**. Flipping one fair coin twice is an example of an **experiment**.

A result of an experiment is called an outcome. The sample space of an experiment is the set of all possible outcomes. Three ways to represent a **sample space** are: to list the possible outcomes, to create a tree diagram, or to create a Venn diagram. The uppercase letter S is used to denote the sample space. For example, if you flip a fair coin once, the possible outcomes are H = heads and T = tails, so the sample space is $S = \{H, T\}$.

An **event** is a set containing any combination of outcomes. Upper case letters like A and B represent events. For example, if the experiment is to flip one fair coin twice, event A might be getting at most one head.

Example

Consider the sample space S that contains the whole numbers (starting at one) less than 20.

Let event A = the even numbers and event B = numbers greater than 13.

List the elements of S, A, and B.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=685#h5p-207

Definition of Probability

The probability of any **outcome** is the long-term relative frequency of that outcome. **Probabilities are between zero and one, inclusive** (that is, zero and one and all numbers between these values). The probability of an event A is written $P(A)$. If an event A cannot occur, then $P(A) = 0$. If the event is guaranteed to occur, $P(A) = 1$. If $P(A) = \frac{1}{2} = 0.5$, the event A is **equally likely** to occur or not to occur. For example, if you flip one fair coin repeatedly (from 20 to 2,000 to 20,000 times) the relative frequency of heads approaches 0.5 (the probability of heads).

Equally likely means that each outcome of an experiment occurs with equal probability. For example, if you toss a fair, six-sided die, each face (1, 2, 3, 4, 5, or 6) is as likely to occur as any other face. If you toss a fair coin, a Head (H) and a Tail (T) are equally likely to occur. If you randomly guess the answer to a true/false question on an exam, you are equally likely to select a correct answer or an incorrect answer.

To calculate the probability of an event A when all outcomes in the sample space are equally likely, count the number of outcomes for event A and divide by the total number of outcomes in the sample space.

Example

If you toss a fair dime and a fair nickel, the sample space is $\{HH, TH, HT, TT\}$. The sample space has four outcomes. Let $A =$ getting exactly one Head. There are two outcomes that meet this condition, so $A = \{HT, TH\}$, so $P(A) = \frac{2}{4} = 0.5$.

The Law of Large Numbers

An important characteristic of probability experiments, known as the **law of large numbers**, states that as the number of repetitions of an experiment is increased, the relative frequency obtained in the experiment tends to become closer and closer to the theoretical probability. Even though the outcomes do not happen according to any set pattern or order, overall, the long-term observed relative frequency will approach the theoretical probability. (The word empirical is often used instead of the word observed.)

You toss a coin and record the result. What is the probability that the result is heads? If you flip a coin two times, does probability tell you that these flips will result in one heads and one tail? You might toss a fair coin ten times and record nine heads. Probability does not describe the short-term results of an experiment, rather it gives information about what can be expected in the long term. To demonstrate this, Karl Pearson once tossed a fair coin 24,000 times! He recorded the results of each toss, obtaining heads 12,012 times. In his experiment, Pearson illustrated the Law of Large Numbers.

Example

Suppose you roll one fair six-sided die, with the numbers {1, 2, 3, 4, 5, 6} on its faces. Let event *E* = rolling a number that is at least five.

There are two outcomes {5, 6}, so $P(E) = \frac{2}{6} = \frac{1}{3}$.

If you were to roll the die only a few times, you would not be surprised if your observed results did not match the probability. But if you were to roll the die a very large number of times, you would expect that, overall, $\frac{2}{6}$ of the rolls would result in an outcome of "at least five". You would not

expect exactly $\frac{2}{6}$. The long-term relative frequency of obtaining this result would approach the theoretical probability of $\frac{2}{6}$ as the number of repetitions grows larger and larger.

It is important to realize that in many situations, the outcomes are not equally likely. A coin or die may be unfair, or **biased**. Two math professors in Europe had their statistics students test the Belgian one Euro coin and discovered that in 250 trials, a head was obtained 56% of the time and a tail was obtained 44% of the time. The data seem to show that the coin is not a fair coin; more repetitions would be helpful to draw a more accurate conclusion about such bias. Some dice may be biased. Look at the dice in a game you have at home; the spots on each face are usually small holes carved out and then painted to make the spots visible. Your dice may or may not be biased; it is possible that the outcomes may be affected by the slight weight differences due to the different numbers of holes in the faces. Gambling casinos make a lot of money depending on outcomes from rolling dice, so casino dice are made differently to eliminate bias. Casino dice have flat faces; the holes are completely filled with paint having the same density as the material that the dice are made out of so that each face is equally likely to occur. Later we will learn techniques to use to work with probabilities for events that are not equally likely.

Combining Events

Often we are not just interested in a single event, but multiple events which may or may not happen at the same time. In order to find probabilities relating to multiple events, we first look at how to combine two events into a new one.

First, consider an **"AND" Event**: An outcome is in the event A AND B if the outcome is in both A and B at the same time. In set notation, this is called the **intersection** of two events, denoted $A \cap B$, and the set $A \cap B$ contains the outcomes that are in both events A AND B.

Example

Let $A = \{1, 2, 3, 4, 5\}$ and $B = \{4, 5, 6, 7, 8\}$.

Then A AND $B = A \cap B = \{4, 5\}$.

Next, consider an **"OR" Event**: An outcome is in the event A OR B if the outcome is in A or is in B or is in both A and B. In set notation, this is called the **union** of two events, denoted $A \cup B$, and the set $A \cup B$ contains the outcomes that are in either event A or B or both.

Example

Let $A = \{1, 2, 3, 4, 5\}$ and $B = \{4, 5, 6, 7, 8\}$.

Then A OR $B = A \cup B = \{1, 2, 3, 4, 5, 6, 7, 8\}$. Notice that 4 and 5 are NOT listed twice; duplicates are not listed in set notation.

Third, the **complement** of event A is denoted A' (read "A prime") and consists of all outcomes that are **NOT** in A. As a result, $P(A) + P(A') = 1$.

Example

Let $S = \{1, 2, 3, 4, 5, 6\}$, like when rolling a fair six-sided die, and let $A = \{1, 2, 3, 4\}$. Compute $P(A)$ and $P(A')$.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=685#h5p-208

Finally, the **conditional probability** of A given B, written $P(A|B)$, is the probability that event A will occur given that the event B has already occurred. A conditional reduces the sample space: we calculate the probability of A from the reduced sample space containing the outcomes where B has occurred. Assuming $P(B)$ is greater than zero, the formula to calculate $P(A|B)$is

$$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)}, \text{ or equivalently,}$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Example

Suppose we toss one fair, six-sided die. The sample space is $S = \{1, 2, 3, 4, 5, 6\}$. Let $A =$ "face is 2 or 3" and $B =$ "face is even." Calculate $P(A|B)$ two ways, first finding the smaller sample space, and then using the formula.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=685#h5p-209

It is important to read each problem carefully to think about and understand what the events are. Understanding the wording is the first very important step in solving probability problems. Reread the problem several times if necessary. Clearly identify the event of interest. Determine whether there is a condition stated in the wording that would indicate that the probability is conditional; carefully identify the condition, if any.

Example

Consider the sample space S that contains the whole numbers (starting at one) less than 20. Let event A = the even numbers and event B = numbers greater than 13. Compute each set or probability.

1. $P(A) =$ _____
2. $P(B) =$ _____
3. $A \cap B =$ _____
4. $A \cup B =$ _____
5. $P(A \cap B) =$ _____
6. $P(A \cup B) =$ _____
7. $A' =$ _____
8. $P(A') =$ _____
9. $P(A) + P(A') =$ _____
10. $P(A|B) =$ _____
11. $P(B|A) =$ _____

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=685#h5p-210

Your turn!

The sample space *S* is the ordered pairs of two whole numbers, the first from one to three and the second from one to four (Example: (1, 4)). Let event *A* = the sum is even and event *B* = the first number is prime.

1. Enumerate the set S.
2. Enumerate the set A.
3. Enumerate the set B.
4. Compute $P(A)$.
5. Compute $P(B)$.
6. Enumerate the set $A \cap B$.
7. Enumerate the set $A \cup B$.
8. Compute $P(A \cap B)$.
9. Compute $P(A \cup B)$.
10. Enumerate the set B'.
11. Compute $P(B')$
12. Check that $P(B) + P(B') = 1$.
13. Compute $P(A|B)$ and $P(B|A)$. Are they equal?



An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=685#h5p-211

Your Turn!

A fair, six-sided die is rolled. Describe the sample space *S*, identify each of the following events with a subset of *S* and compute its probability (an outcome is the number of dots that show up).

1. Event *T* = the outcome is two.
2. Event *A* = the outcome is an even number.
3. Event *B* = the outcome is less than four.
4. The complement of *A*.

5. *A* GIVEN *B*
6. *B* GIVEN *A*
7. *A* AND *B*
8. *A* OR *B*
9. *A* OR *B*′
10. Event *N* = the outcome is a prime number.
11. Event *I* = the outcome is seven.

Solution

1. $T = \{2\}$, $P(T) = \frac{1}{6}$
2. $A = \{2, 4, 6\}$, $P(A) = \frac{1}{2}$
3. $B = \{1, 2, 3\}$, $P(B) = \frac{1}{2}$
4. $A' = \{1, 3, 5\}$, $P(A') = \frac{1}{2}$
5. *Reduced sample space:* $A|B = \{2\}$, $P(A|B) = \frac{1}{3}$
6. *Reduced sample space:* $B|A = \{2\}$, $P(B|A) = \frac{1}{3}$
7. *A* AND *B* = {2}, $P(A \text{ AND } B) = \frac{1}{6}$
8. *A* OR *B* = {1, 2, 3, 4, 6}, $P(A \text{ OR } B) = \frac{5}{6}$
9. *A* OR *B*′ = {2, 4, 5, 6}, $P(A \text{ OR } B') = \frac{2}{3}$
10. $N = \{2, 3, 5\}$, $P(N) = \frac{1}{2}$
11. A six-sided die does not have seven dots. $P(7) = 0$.

Your turn!

The table below shows the enrollment of 100 students in two freshman courses, English and Speech.

	Taking English	Not Taking English
Taking Speech	43	9
Not Taking Speech	44	4

Denote the events $S =$ taking Speech, $NS =$ Not taking Speech, E = taking English, and $NE =$ Not taking English. Compute the following probabilities.

1. $P(E)$
2. $P(S)$
3. The probability that a student is not taking English.
4. The probability that a student is not taking Speech.
5. $P(NE \cap NS)$
6. The probability that a student is taking both English and Speech.
7. The probability that a student is taking English and not Speech.
8. The probability that a student is taking English or Speech.
9. $P(E')$
10. The probability that a student is taking English or not taking Speech.
11. The probability that a given Speech student is taking English.
12. The probability that a given English student is taking Speech.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=685#h5p-212

References

"Countries List by Continent." Worldatlas, 2013. Available online at http://www.worldatlas.com/cntycont.htm (accessed May 2, 2013).

3.2 INDEPENDENT AND MUTUALLY EXCLUSIVE EVENTS

Learning Objectives

By the end of this section, you should be able to:

- Identify two events as mutually exclusive, or not
- Identify two events as independent, or not
- Use the relationships between events to compute probabilities

Often we are not just interested in a single event, but multiple events happening at the same time. In order to find probabilities relating to multiple events, we first have to know about the relationship (or lack thereof) between them. The two main relationship terms we will look for are independence and mutually exclusive. Remember, these two terms do not mean the same thing, neither are they opposites.

Independent Events

Two events A and B are **independent** if the knowledge that one occurred does not affect the chance the other occurs. For example, the outcomes of two rolls of a fair die are independent events, because the outcome of the first roll does not change the probability for the outcome of the second roll. If two events are NOT independent, then we say that they are **dependent**.

Two events are independent if and only if all of the following are true:

- $P(A|B) = P(A)$
- $P(B|A) = P(B)$
- $P(A \text{ AND } B) = P(A \cap B) = P(A)P(B)$.

So to show two events are independent, you must show just one of the above conditions holds.

How you sample can have implications on independence. Sampling may be done **with replacement** or **without replacement**.

- **With replacement**: If each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once. When sampling is done with replacement, then events are independent because the result of the first pick will not change the probabilities for the second pick.
- **Without replacement**: When sampling is done without replacement, each member of a population may be chosen only once. In this case, the probabilities for the second pick are affected by the result of the first pick. The events are dependent.

Example

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, *J* (jack), *Q* (queen), *K* (king) of that suit.

a. Sampling with replacement:

Suppose you pick three cards with replacement. The first card you pick out of the 52 cards is the *Q* of spades. You put this card back, reshuffle the cards and pick a second card from the 52-card deck. It is the ten of clubs. You put this card back, reshuffle the cards and pick a third card from the 52-card deck. This time, the card is the *Q* of spades again. Your picks are {*Q* of spades, ten of clubs, *Q* of spades}. You have picked the *Q* of spades twice. On each of the three draws, you pick from the full 52-card deck.

b. Sampling without replacement:

Suppose you pick three cards without replacement. The first card you pick out of the 52 cards is the *K* of hearts. You put this card aside and pick the second card from the 51 cards remaining in the deck. It is the three of diamonds. You put this card aside and pick the third card from the remaining 50 cards in the deck. The third card is the *J* of spades. Your picks are {*K* of hearts, three of diamonds, *J* of spades}. Because you have picked the cards without replacement, you cannot pick the same card twice.

Your Turn!

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, *J* (jack), *Q* (queen), *K* (king) of that suit.

1. Suppose you know that the picked cards are *Q* of spades, *K* of hearts and *Q* of spades. Can you decide if the sampling was with or without replacement?
2. Suppose you are dealt three cards for the start of a game, the *Q* of spades, *K* of hearts, and *J* of spades. Can you decide if the sampling was with or without replacement?

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=687#h5p-92

Your Turn!

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts, and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, *J* (jack), *Q* (queen), and *K* (king) of that suit. *S* = spades, *H* = Hearts, *D* = Diamonds, *C* = Clubs. Suppose that you sample four cards without replacement. Which of the following outcomes are possible? Answer the same question for sampling with replacement.

1. *QS*, 1*D*, 1*C*, *QD*
2. *KH*, 7*D*, 6*D*, *KH*
3. *QS*, 7*D*, 6*D*, *KS*



An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=687#h5p-213

Mutually Exclusive Events

Events A and B are **mutually exclusive events** if they cannot occur at the same time. This means that A and B do not share any outcomes, $A \cap B = \emptyset$ where $\emptyset = \{\}$ is the empty set, and $P(A \text{ AND } B) = P(A \cap B) = 0$.

Example

Suppose the sample space contains the integers one through ten, $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. Let $A = \{1, 2, 3, 4, 5\}, B = \{4, 5, 6, 7, 8\}$, and C = \{7, 9\}.

Then $A \cap B = \{4, 5\}$. Furthermore, $P(A \cap B) = \frac{2}{10}$, which is not zero. Therefore, A and B are not mutually exclusive.

However, A and C have no numbers in common. Thus $A \cap C = \emptyset$ and $P(A \cap C) = 0$, so A and C are mutually exclusive.

Example

Flip two fair coins. (This is an experiment.)

The sample space is {*HH*, *HT*, *TH*, *TT*} where *T* = tails and *H* = heads. The outcomes are *HH*, *HT*, *TH*, and *TT*. The outcomes HT and TH are different. The *HT* means that the first coin showed heads and the second coin showed tails. The *TH* means that the first coin showed tails and the second coin showed heads.

- Let *A* = the event of getting **at most one tail**. (At most one tail means zero or one tail.) Then *A* can be written as {*HH*, *HT*, *TH*}. The outcome *HH* shows zero tails. *HT* and *TH* each show one tail.
- Let *B* = the event of getting all tails. *B* can be written as {*TT*}. *B* is the **complement** of *A*, so *B* = *A*′. Also, *P*(*A*) + *P*(*B*) = *P*(*A*) + *P*(*A*′) = 1.
- The probabilities for *A* and for *B* are *P*(*A*) = $\frac{3}{4}$ and *P*(*B*) = $\frac{1}{4}$.
- Let *C* = the event of getting all heads. *C* = {*HH*}. Since *B* = {*TT*}, *P*(*B* AND *C*) = 0. *B* and *C* are mutually exclusive. (*B* and *C* have no members in common because you cannot have all tails and all heads at the same time.)
- Let *D* = event of getting **more than one** tail. *D* = {*TT*}. *P*(*D*) = $\frac{1}{4}$
- Let *E* = event of getting a head on the first roll. (This implies you can get either a head or tail on the second roll.) *E* = {*HT*, *HH*}. *P*(*E*) = $\frac{2}{4}$
- Find the probability of getting **at least one** (one or two) tail in two flips. Let *F* = event of getting at least one tail in two flips. *F* = {*HT*, *TH*, *TT*}. *P*(*F*) = $\frac{3}{4}$

Your Turn!

Draw two cards from a standard 52-card deck with replacement. Find the probability of getting at least one black card. (Half of the cards are red, half are black.)

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=687#h5p-214

Your Turn!

Flip two fair coins. Find the probabilities of the events.

1. Let F = the event of getting at most one tail (zero or one tail).
2. Let G = the event of getting two faces that are the same.
3. Let H = the event of getting a head on the first flip followed by a head or tail on the second flip.
4. Are F and G mutually exclusive?
5. Let J = the event of getting all tails. Are J and H mutually exclusive?

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=687#h5p-215

The Axioms of Probability

Finding probabilities in more complicated situations starts with the three Axioms of Probability. Two were mentioned in the previous section, and the third we can now understand.

1. $P(S) = 1$, that is, the probability of the sample space is one.

2. $0 \leq P(E) \leq 1$, that is, for any event E, the probability is between 0 and 1 inclusive.
3. For any two mutually exclusive events E_1 and E_2, that is, $E_1 \cap E_2 = \emptyset$, $P(E_1 \cup E_2) = P(E_1) + P(E_2)$.

The first two Axioms should be fairly intuitive. Axiom 1 says that the probabilities of all outcomes in a sample space will always add up to 1. Axiom 2 says the probability of any event must be between 0 and 1. The Third Axiom is called the disjoint addition rule; if we know two events are mutually exclusive, then we can find the probability of their OR event by adding their probabilities.

Assumptions about Events

Consider two events. If it is not known whether they are mutually exclusive, assume they are not until you can show otherwise. Likewise, if it is not known whether the events are independent, assume they are dependent until you can show otherwise. This "default" starting point is illustrated by the 4th position in the following table:

Figure 3.2: Relationships Between Events

		Independent?	
		Yes	**No**
Mutually Exclusive?	**Yes**	1*	2
	No	3	4

Depending on the information given in the problem and assumptions you are able to make, as you move around the above grid, you will apply slightly different versions of each important probability rule.

**Note:* You will rarely, if ever, find yourself in this case

Your Turn!

Let event C = taking an English class. Let event D = taking a speech class.

Suppose $P(C) = 0.75$, $P(D) = 0.3$, $P(C|D) = 0.75$ and $P(C \text{ AND } D) = 0.225$.

Justify your answers to the following questions numerically.

1. Are C and D independent?
2. Are C and D mutually exclusive?
3. What is $P(D|C)$?

Solution

1. Yes, because $P(C|D) = P(C)$.
2. No, because $P(C \text{ AND } D)$ is not equal to zero.
3. $P(D|C) = \dfrac{P(C \text{ AND } D)}{P(C)} = \dfrac{0.225}{0.75} = 0.3$

Your Turn!

A box has two balls, one white and one red. We select one ball, put it back in the box, and select a second ball (sampling with replacement). Find the probability of the following events:

1. Let F = the event of getting the white ball twice.
2. Let G = the event of getting two balls of different colors.
3. Let H = the event of getting white on the first pick.
4. Are F and G mutually exclusive?
5. Are G and H mutually exclusive?

Solution

1. $P(F) = \dfrac{1}{4}$
2. $P(G) = \dfrac{1}{2}$
3. $P(H) = \dfrac{1}{2}$
4. Yes

5. No

Your Turn!

Roll one fair, six-sided die. The sample space is {1, 2, 3, 4, 5, 6}. Let event A = a face is odd. Then A = {1, 3, 5}. Let event B = a face is even. Then B = {2, 4, 6}.

- Find the complement of A, A'. The complement of A, A', is B because A and B together make up the sample space. $P(A) + P(B) = P(A) + P(A') = 1$. Also, $P(A) = \frac{3}{6}$ and $P(B) = \frac{3}{6}$.
- Let event C = odd faces larger than two. Then C = {3, 5}. Let event D = all even faces smaller than five. Then D = {2, 4}. $P(C \text{ AND } D) = 0$ because you cannot have an odd and even face at the same time. Therefore, C and D are mutually exclusive events.
- Let event E = all faces less than five. E = {1, 2, 3, 4}.

Are C and E mutually exclusive events? (Answer yes or no.) Why or why not?

Solution

No. C = {3, 5} and E = {1, 2, 3, 4}. $P(C \text{ AND } E) = \frac{1}{6}$. To be mutually exclusive, $P(C \text{ AND } E)$ must be zero.

We can also find $P(C|A)$, a conditional probability. Recall that the event C is {3, 5} and event A is {1, 3, 5}. To find $P(C|A)$, find the probability of C using the sample space A. You have reduced the sample space from the original sample space {1, 2, 3, 4, 5, 6} to {1, 3, 5}. So, $P(C|A) = \frac{2}{3}$.

Your Turn!

Let event G = taking a math class. Let event H = taking a science class. Then, G AND H = taking a math class and a science class. Suppose $P(G) = 0.6$, $P(H) = 0.5$, and $P(G \text{ AND } H) = 0.3$. Are G and H independent?

If G and H are independent, then you must show **ONE** of the following:

- $P(G|H) = P(G)$
- $P(H|G) = P(H)$
- $P(G \text{ AND } H) = P(G)P(H)$

NOTE

The choice you make depends on the information you have. You could choose any of the methods here because you have the necessary information.

a. Show that $P(G|H) = P(G)$.

Solution

$$P(G|H) = \frac{P(G \text{ AND } H)}{P(H)} = \frac{0.3}{0.5} = 0.6 = P(G)$$

b. Show $P(G \text{ AND } H) = P(G)P(H)$.

Solution

$P(G)P(H) = (0.6)(0.5) = 0.3 = P(G \text{ AND } H)$

Since G and H are independent, knowing that a person is taking a science class does not change the chance that he or she is taking a math class. If the two events had not been independent (that is, they are dependent) then knowing that a person is taking a science class would change the chance he or she is taking math. For practice, show that $P(H|G) = P(H)$ to show that G and H are independent events.

Your Turn!

Mark is deciding which route to take to work. His choices are I = the Interstate and F = Fifth Street.

- $P(I) = 0.44$ and $P(F) = 0.56$
- $P(I \text{ AND } F) = 0$ because Mark will take only one route to work.

What is the probability of $P(I \text{ OR } F)$?

Solution

Because $P(I \text{ AND } F) = 0$,

$P(I \text{ OR } F) = P(I) + P(F) – P(I \text{ AND } F) = 0.44 + 0.56 – 0 = 1$

Your Turn!

In a particular college class, 60% of the students are female. Fifty percent of all students in the class have long hair. Forty-five percent of the students are female and have long hair. Of the female students, 75% have long hair. Let F be the event that a student is female. Let L be the event that a student has long hair. One student is picked randomly. Are the events of being female and having long hair independent?

- The following probabilities are given in this example:
- $P(F) = 0.60$; $P(L) = 0.50$
- $P(F \text{ AND } L) = 0.45$
- $P(L|F) = 0.75$

NOTE: THE CHOICE YOU MAKE FOR HOW TO SHOW INDEPENDENCE DEPENDS ON THE

INFORMATION YOU HAVE. YOU COULD USE THE FIRST OR LAST CONDITION ON THE LIST FOR THIS EXAMPLE. YOU DO NOT KNOW $P(F|L)$ YET, SO YOU CANNOT USE THE SECOND CONDITION.

Condition 1: Check whether $P(F \text{ AND } L) = P(F)P(L)$. We are given that $P(F \text{ AND } L) = 0.45$, but $P(F)P(L) = (0.60)(0.50) = 0.30$. The events of being female and having long hair are not independent because $P(F \text{ AND } L)$ does not equal $P(F)P(L)$.

Condition 3: Check whether $P(L|F)$ equals $P(L)$. We are given that $P(L|F) = 0.75$, but $P(L) = 0.50$; they are not equal. The events of being female and having long hair are not independent.

Interpretation of Results: The events of being female and having long hair are not independent; knowing that a student is female changes the probability that a student has long hair.

Your Turn!

In a basketball arena,

- 70% of the fans are rooting for the home team.
- 25% of the fans are wearing blue.
- 20% of the fans are wearing blue and are rooting for the away team.
- Of the fans rooting for the away team, 67% are wearing blue.

Let A be the event that a fan is rooting for the away team.

Let B be the event that a fan is wearing blue.

Are the events of rooting for the away team and wearing blue independent? Are they mutually exclusive?

Solution
$P(B|A) = 0.67$

$P(B) = 0.25$

So $P(B)$ does not equal $P(B|A)$ which means that B and A are not independent (wearing blue and rooting for the away team are not independent). They are also not mutually exclusive, because P(B AND A) = 0.20, not 0.

References

Lopez, Shane, Preety Sidhu. "U.S. Teachers Love Their Lives, but Struggle in the Workplace." Gallup Wellbeing, 2013. http://www.gallup.com/poll/161516/teachers-love-lives-struggle-workplace.aspx (accessed May 2, 2013).

Data from Gallup. Available online at www.gallup.com/ (accessed May 2, 2013).

Chapter Review

Two events A and B are independent if the knowledge that one occurred does not affect the chance the other occurs. If two events are not independent, then we say that they are dependent.

In sampling with replacement, each member of a population is replaced after it is picked, so that member has the possibility of being chosen more than once, and the events are considered to be independent. In sampling without replacement, each member of a population may be chosen only once, and the events are considered not to be independent. When events do not share outcomes, they are mutually exclusive of each other.

Formula Review

If A and B are independent, $P(A \text{ AND } B) = P(A)P(B)$, $P(A|B) = P(A)$ and $P(B|A) = P(B)$.

If A and B are mutually exclusive, $P(A \text{ OR } B) = P(A) + P(B)$ and $P(A \text{ AND } B) = 0$.

3.3 TWO BASIC RULES OF PROBABILITY

Learning Objectives

By the end of this section, you should be able to:

- state and apply the multiplication rule to any two events
- state and apply the multiplication rule for independent events
- state and apply the addition rule for any two events
- state and apply the addition rule for mutually exclusive events

Recall the different combinations of relationships between two events:

Figure 3.10: Relationships Between Events

		Independent?	
		Yes	**No**
Disjoint?	**Yes**	1*	2
	No	3	4

We must always go into a problem assuming two events are not **mutually exclusive** or **independent**. This "default" starting point is illustrated by the 4th position in the table above. Depending on the information you are given and assumptions you are able to make, you may move potions on this grid. Where you fall on the grid will dictate how we apply the rules we will discuss in this section to find probabilities of compound events.

There are two types of compound events we may be interested in, Unions (OR) and Intersections (AND), each with their own set of rules and assumptions. When calculating probability, there are two rules to consider when determining if the two events are independent or dependent and if they are mutually exclusive or not.

*__Note:__ You will rarely, if ever, find yourself in this case

The Multiplication Rule

If A and B are two events defined on a sample space, then: $P(A \text{ AND } B) = P(B)P(A|B)$, or equivalently, $P(A \cap B) = P(B)P(A|B)$.

This rule may also be written as: $P(A|B) = \frac{P(A \text{ AND } B)}{P(B)}$, that is, the probability of A given B equals the probability of both A and B divided by the probability of B.

If A and B are **independent**, then the probability of A does not depend on the probability of B and $P(A|B) = P(A)$. As a result, $P(A \text{ AND } B) = P(A|B)P(B)$ becomes $P(A \text{ AND } B) = P(A)P(B)$.

Example

A school has 200 seniors of whom 140 will be going to college next year. Forty will be going directly to work. The remainder are taking a gap year. Of the seniors going to college, fifty play sports. Of the seniors going directly to work, thirty play sports. Five of the seniors taking a gap year play sports.

1. What is the probability that a senior is taking a gap year?
2. Are "taking a gap year" and "playing sports" independent events?
3. Are "going to college" and "taking a gap year" independent events?

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=689#h5p-231

Your Turn!

Suppose you are going to roll a dice twice. Show that A = "Roll a 6 on the first roll" and B = "Roll a 4 on the second roll" are independent events using the Multiplication Rule.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=689#h5p-233

The Addition Rule

If A and B are defined on a sample space, then: $P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B)$.

If A and B are **mutually exclusive**, then $P(A \text{ AND } B) = 0$. Then $P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B)$ becomes $P(A \text{ OR } B) = P(A) + P(B)$.

Example

Klaus is trying to choose where to go on vacation. His two choices are: A = New Zealand and B = Alaska. Klaus can only afford one vacation; he is leaning towards New Zealand so the probability that he chooses A is $P(A) = 0.6$ and the probability that he chooses B is $P(B) = 0.35$.

- $P(A \text{ AND } B) = 0$ because Klaus can only afford to take one vacation.
- Therefore, the probability that he chooses either New Zealand or Alaska is $P(A \text{ OR } B) = P(A) + P(B) = 0.6 + 0.35 = 0.95$. Note that the probability that he does not choose to go anywhere on vacation must be 0.05.

Example

A student goes to the library. Let events B = the student checks out a book and M = the student checks out a movie. Suppose that $P(B) = 0.40$, $P(M) = 0.30$ and $P(M|B) = 0.5$.

1. Find $P(B \text{ AND } M)$.
2. Find P(B OR M).

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=689#h5p-232

Your Turn!

Carlos plays college soccer. He makes a goal 65% of the time he shoots. Carlos is going to attempt two goals in a row in the next game. A = the event Carlos is successful on his first attempt. $P(A) = 0.65$. B = the event Carlos is successful on his second attempt. $P(B) = 0.65$. Carlos tends to shoot in streaks. The probability that he makes the second goal **GIVEN** that he made the first goal is 0.90.

a. What is the probability that he makes both goals?

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=689#h5p-106

b. Are A and B independent?

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=689#h5p-104

c. What is the probability that Carlos makes either the first goal or the second goal?

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=689#h5p-107

d. Are *A* and *B* mutually exclusive?

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=689#h5p-105

Your Turn!

A community swim team has **150** members. **Seventy-five** of the members are advanced swimmers. **Forty-seven** of the members are intermediate swimmers. The remainder are novice swimmers. **Forty** of the advanced swimmers practice four times a week. **Thirty** of the intermediate swimmers practice four times a week. **Ten** of the novice swimmers practice four times a week. Suppose one member of the swim team is chosen randomly.

1. What is the probability that the member is a novice swimmer?
2. What is the probability that the member practices four times a week?
3. What is the probability that the member is an advanced swimmer and practices four times a week?
4. What is the probability that a member is an advanced swimmer and an intermediate swimmer? Are being an advanced swimmer and an intermediate swimmer mutually exclusive? Why or why not?
5. Are being a novice swimmer and practicing four times a week independent events? Why or why not?



An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=689#h5p-235

Let us now look at some examples where we include the complement of events. As a reminder, the complement of an event A is denoted A', and $P(A') = 1 - P(A)$.

Example

Studies show that about one woman in seven (approximately 14.3%) who live to be 90 will develop breast cancer. Suppose that of those women who develop breast cancer, a test is negative 2% of the time. Also suppose that in the general population of women, the test for breast cancer is negative about 85% of the time. Let *B* = woman develops breast cancer and let *N* = tests negative. Suppose one woman is selected at random.

1. What is the probability that the woman develops breast cancer? What is the probability that the woman tests negative?
2. Given that the woman has breast cancer, what is the probability that she tests negative?

3. What is the probability that the woman has breast cancer and tests negative?
4. What is the probability that the woman has breast cancer or tests negative?
5. Are having breast cancer and testing negative mutually exclusive?
6. Given that a woman develops breast cancer, what is the probability that she tests positive?
7. What is the probability that a woman develops breast cancer and tests positive?
8. What is the probability that a woman does not develop breast cancer?
9. What is the probability that a woman tests positive for breast cancer?



An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=689#h5p-236

Your Turn!

Returning to the example, a student goes to the library, where events B = the student checks out a book and M = the student checks out a movie. Suppose that $P(B) = 0.40$, $P(\text{M}) = 0.30$ and $P(\text{M}|B) = 0.5$.

1. Find $P(B')$.
2. Find $P(\text{M AND } B)$.
3. Find $P(B \mid \text{M})$.
4. Find $P(\text{M AND } B')$.
5. Find $P(\text{M} \mid B')$.



An interactive H5P element has been excluded from this version of the text. You can view it online

here:
https://louis.pressbooks.pub/introductorystatistics/?p=689#h5p-237

References

DiCamillo, Mark, Mervin Field. "The File Poll." Field Research Corporation. Available online at http://www.field.com/fieldpollonline/subscribers/Rls2443.pdf (accessed May 2, 2013).

Rider, David, "Ford support plummeting, poll suggests," The Star, September 14, 2011. Available online at http://www.thestar.com/news/gta/2011/09/14/ford_support_plummeting_poll_suggests.html (accessed May 2, 2013).

"Mayor's Approval Down." News Release by Forum Research Inc. Available online at http://www.forumresearch.com/forms/News Archives/News Releases/74209_TO_Issues_-_Mayoral_Approval_%28Forum_Research%29%2820130320%29.pdf (accessed May 2, 2013).

"Roulette." Wikipedia. Available online at http://en.wikipedia.org/wiki/Roulette (accessed May 2, 2013).

Shin, Hyon B., Robert A. Kominski. "Language Use in the United States: 2007." United States Census Bureau. Available online at http://www.census.gov/hhes/socdemo/language/data/acs/ACS-12.pdf (accessed May 2, 2013).

Data from the Baseball-Almanac, 2013. Available online at www.baseball-almanac.com (accessed May 2, 2013).

Data from U.S. Census Bureau.

Data from the Wall Street Journal.

Data from The Roper Center: Public Opinion Archives at the University of Connecticut. Available online at http://www.ropercenter.uconn.edu/ (accessed May 2, 2013).

Data from Field Research Corporation. Available online at www.field.com/fieldpollonline (accessed May 2,2 013).

3.4 CONTINGENCY TABLES

Learning Objectives

By the end of this section, the student should be able to:

- Use a given contingency table to compute probabilities
- Complete a contingency table given starting information about a sample

Sometimes, when the probability problems are complex, it can be helpful to visualize the situation. Contingency tables, tree diagrams and Venn diagrams are some tools that can help us visualize and solve conditional probabilities. Tree and Venn diagrams will be covered in the next section.

Contingency Tables

A contingency table provides a way of portraying data that can facilitate calculating probabilities, particularly conditional probabilities. The table displays sample values in relation to two different variables that may be dependent or contingent on one another. Later on, we will use **contingency tables** again, but in another manner.

Example

Suppose a study of speeding violations and drivers who use cell phones produced the following fictional data:

	Speeding violation in the last year	No speeding violation in the last year	Total
Cell phone user	25	280	305
Not a cell phone user	45	405	450
Total	70	685	755

The total number of people in the sample is 755. The row totals are 305 and 450. The column totals are 70 and 685. Notice that 305 + 450 = 755 and 70 + 685 = 755.

Calculate the following probabilities using the table.

a. Find *P*(Person is a car phone user).

Solution

a. $\dfrac{\text{number of car phone users}}{\text{total number in study}} = \dfrac{305}{755}$

b. Find *P*(person had no violation in the last year).

Solution

b. $\dfrac{\text{number that had no violation}}{\text{total number in study}} = \dfrac{685}{755}$

c. Find *P*(Person had no violation in the last year AND was a car phone user).

Solution

c. $\dfrac{280}{755}$

d. Find *P*(Person is a car phone user OR person had no violation in the last year).

Solution

d. $\left(\dfrac{305}{755} + \dfrac{685}{755}\right) - \dfrac{280}{755} = \dfrac{710}{755}$

e. Find *P*(Person is a car phone user GIVEN person had a violation in the last year).

Solution

e. $\frac{25}{70}$ (The sample space is reduced to the number of persons who had a violation.)

f. Find *P*(Person had no violation last year GIVEN person was not a car phone user)

Solution

f. $\frac{405}{450}$ (The sample space is reduced to the number of persons who were not car phone users.)

Your Turn!

[link] shows the number of athletes who stretch before exercising and how many had injuries within the past year.

	Injury in last year	No injury in last year	Total
Stretches	55	295	350
Does not stretch	231	219	450
Total	286	514	800

1. What is *P*(athlete stretches before exercising)?
2. What is *P*(athlete stretches before exercising | no injury in the last year)?

Solution

1. *P*(athlete stretches before exercising) = $\frac{350}{800}$ = 0.4375
2. *P*(athlete stretches before exercising | no injury in the last year) = $\frac{295}{514}$ = 0.5739

Key Takeaways

[link] shows a random sample of 200 cyclists broken down by type of bicycle and type of route they prefer. Let *M* = mountain bike and *H* = hilly path.

	Lake Path	Hilly Path	Wooded Path	Total
Road	45	38	27	110
Mountain	26	52	12	90
Total	71	90	39	200

1. Out of the mountain bike cyclists, what is the probability that the cyclist prefers a hilly path?
2. Are the events "prefers a mountain bike" and "preferring the hilly path" independent events?

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=690#h5p-238

Your Turn!

[link] relates the weights and heights of a group of individuals participating in an observational study.

Weight/Height	Tall	Medium	Short	Totals
Obese	18	28	14	
Normal	20	51	28	
Underweight	12	25	9	
Totals				

Complete the contingency table by finding the total for each row and column. Then use the table to find the probability that a randomly chosen individual from this group is:

1. Tall.
2. Obese and Tall.
3. Tall given that the individual is Obese.
4. Obese given that the individual is Tall.
5. Tall and Underweight.
6. Are the events Obese and Tall independent?

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=690#h5p-241

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=690#h5p-242

Your Turn!

Muddy Mouse lives in a cage with three doors. If Muddy goes out the first door, the probability that he gets caught by Alissa the Cat is $\frac{1}{5}$. If he goes out the second door, the probability he gets caught by Alissa is $\frac{1}{4}$. The probability that Alissa catches Muddy coming out of the third door is $\frac{1}{2}$. It is equally likely that Muddy will choose any of the three doors so the probability of choosing each door is $\frac{1}{3}$.

Door Choice

Caught or Not	Door One	Door Two	Door Three	Total
Caught	$\frac{1}{15}$	_____	_____	_____
Not Caught	$\frac{4}{15}$	_____	_____	_____
Total	____	____	____	1

The first column of the contingency table has been filled in for you. The first entry is computed as $P(\text{caught} \mid \text{Door 1})P(\text{Door 1}) = \frac{1}{5} \cdot \frac{1}{3} = \frac{1}{15}$. Similarly, the first entry in the second row is computed as

$$P(\text{not caught} \mid \text{Door 1})P(\text{Door 1}) = (1 - P(\text{caught} \mid \text{Door 1}))P(\text{Door 1}) = \frac{4}{5} \cdot \frac{1}{3} = \frac{4}{15}.$$

Complete the rest of the table, verifying that the totals sum to 1 as probabilities should. Then use the table to answer the following questions.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=690#h5p-240

1. What is the probability that Alissa does not catch Muddy?
2. What is the probability that Muddy chooses Door One OR Door Two given that Muddy is caught by Alissa?

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=690#h5p-239

References

"Blood Types." American Red Cross, 2013. Available online at http://www.redcrossblood.org/learn-about-blood/blood-types (accessed May 3, 2013).

Data from the National Center for Health Statistics, part of the United States Department of Health and Human Services.

Data from the United States Senate. Available online at www.senate.gov (accessed May 2, 2013).

Haiman, Christopher A., Daniel O. Stram, Lynn R. Wilkens, Malcolm C. Pike, Laurence N. Kolonel, Brien E. Henderson, and Loïc Le Marchand. "Ethnic and Racial Differences in the Smoking-Related Risk of Lung Cancer." *New England Journal of Medicine*, 2013. Available online at http://www.nejm.org/doi/full/10.1056/NEJMoa033250 (accessed May 2, 2013).

"Human Blood Types." United Blood Services, 2011. Available online at http://www.unitedbloodservices.org/learnMore.aspx (accessed May 2, 2013).

Samuel, T. M. "Strange Facts about RH Negative Blood." eHow Health, 2013. Available online at http://www.ehow.com/facts_5552003_strange-rh-negative-blood.html (accessed May 2, 2013).

3.5 TREE AND VENN DIAGRAMS

Learning Objectives

By the end of this section, you should be able to:

- Use tree diagrams to find probabilities for experiments with and without replacement
- Use Venn diagrams to visualize relationships between events

Sometimes, when the probability problems are complex, it can be helpful to graph the situation. Tree diagrams and Venn diagrams are two tools that can be used to visualize and solve conditional probabilities.

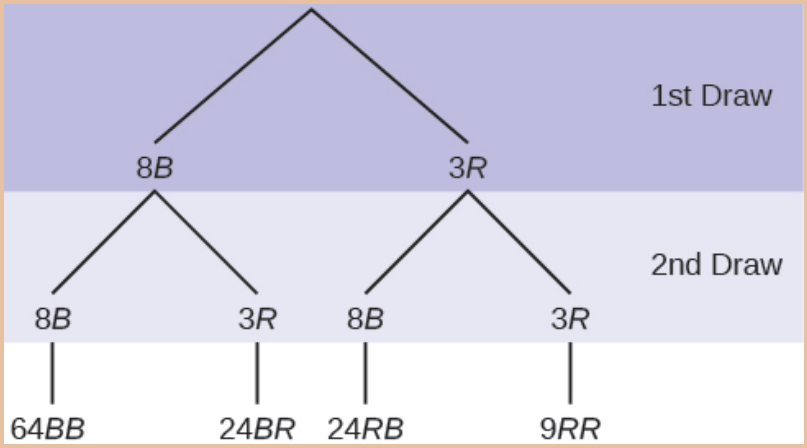
Tree Diagrams

A tree diagram is a special type of graph used to determine the outcomes of an experiment. It consists of "branches" that are labeled with either frequencies or probabilities. Tree diagrams can make some probability problems easier to visualize and solve. The following example illustrates how to use a **tree diagram**.

Example

In an urn, there are 11 balls. Three balls are red (*R*) and eight balls are blue (*B*). Draw two balls, one at a time, **with replacement,** which means that you put the first ball back in the urn before you

select the second ball. The tree diagram using frequencies that show all the possible outcomes follows.



The first set of branches represents the first draw, where there are 8 ways to first draw *B* and three ways to first draw *R*. The second set of branches represents the second draw, with the same potential outcomes for each of the original branches due to the replacement of the first draw. Each of the outcomes is distinct. In fact, we can label each red ball as *R*1, *R*2, and *R*3 and each blue ball as *B*1, *B*2, *B*3, *B*4, *B*5, *B*6, *B*7, and *B*8. Then the nine *RR* outcomes can be written as:

{ *R*1*R*1, *R*1*R*2, *R*1*R*3, *R*2*R*1, *R*2*R*2, *R*2*R*3, *R*3*R*1, *R*3*R*2, *R*3*R*3}. Counting these shows 9 ways we can draw two red balls. We can also get this total from our tree diagram by multiplying down the branches: 3 times 3 = 9 *RR* outcomes. The other branches are similar; there are 64 ways to draw two blue balls, 24 ways to draw blue then red, and 24 ways to draw red then blue, for a total of 64+24+24+9 = 121 outcomes of this experiment.

We can also find this total by multiplying: there are a total of 11 balls in the urn. Then the experiment to draw two balls, one at a time, with replacement has (11)(11) = 121 outcomes, the size of the sample space.

We can now use our tree diagram to compute some probabilities.

1. Calculate $P(RR)$.
2. Calculate $P(RB \text{ OR } BR)$.
3. Calculate $P(RB)$.
4. Calculate $P(R \text{ 2nd} \mid B \text{ 1st})$.
5. Calculate $P(BB)$.
6. Calculate $P(B \text{ 2nd} \mid R \text{ 1st})$.

Solution

b. $P(RR) = \left(\frac{3}{11}\right)\left(\frac{3}{11}\right) = \frac{9}{121}$

c. Using the tree diagram, calculate $P(RB \text{ OR } BR)$.

Solution

c. $P(RB \text{ OR } BR) = \left(\frac{3}{11}\right)\left(\frac{8}{11}\right) + \left(\frac{8}{11}\right)\left(\frac{3}{11}\right) = \frac{48}{121}$

d. Using the tree diagram, calculate $P(R \text{ on 1st draw AND } B \text{ on 2nd draw})$.

Solution

d. $P(R \text{ on 1st draw AND } B \text{ on 2nd draw}) = P(RB) = \left(\frac{3}{11}\right)\left(\frac{8}{11}\right) = \frac{24}{121}$

e. Using the tree diagram, calculate $P(R \text{ on 2nd draw GIVEN } B \text{ on 1st draw})$.

Solution

e. $P(R \text{ on 2nd draw GIVEN } B \text{ on 1st draw}) = P(R \text{ on 2nd} | B \text{ on 1st}) = \frac{24}{88}$

$= \frac{3}{11}$

This problem is a conditional one. The sample space has been reduced to those outcomes that already have a blue on the first draw. There are 24 + 64 = 88 possible outcomes (24 *BR* and 64 *BB*). Twenty-four of the 88 possible outcomes are *BR*. $\frac{24}{88} = \frac{3}{11}$.

f. Using the tree diagram, calculate $P(BB)$.

Solution

f. $P(BB) = \frac{64}{121}$

g. Using the tree diagram, calculate $P(B \text{ on the 2nd draw given } R \text{ on the first draw})$.

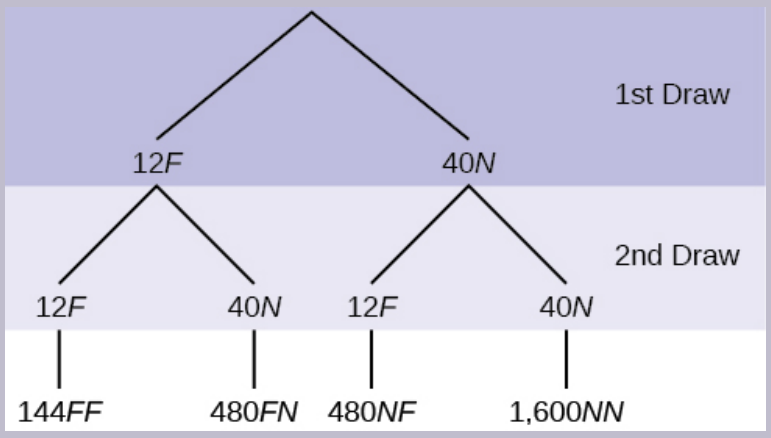
Solution

g. $P(B \text{ on 2nd draw} | R \text{ on 1st draw}) = \frac{8}{11}$

There are 9 + 24 outcomes that have *R* on the first draw (9 *RR* and 24 *RB*). The sample space is then 9 + 24 = 33. 24 of the 33 outcomes have *B* on the second draw. The probability is then $\frac{24}{33}$.

Your Turn!

In a standard deck, there are 52 cards. 12 cards are face cards (event *F*) and 40 cards are not face cards (event *N*). Draw two cards, one at a time, with replacement. All possible outcomes are shown in the tree diagram as frequencies. Using the tree diagram, calculate *P*(*FF*).



Solution

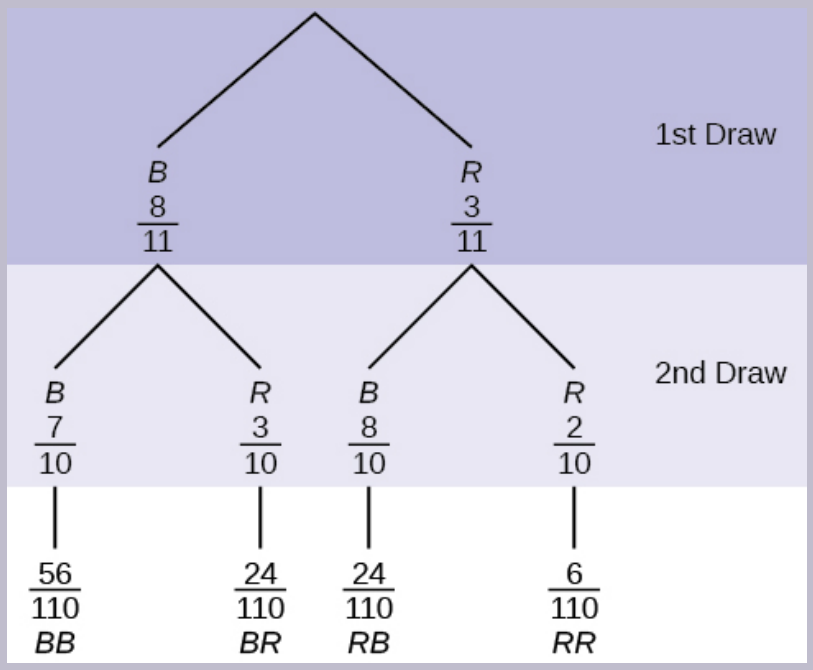
Total number of outcomes is 144 + 480 + 480 + 1600 = 2,704.

$$P(FF) = \frac{144}{144 + 480 + 480 + 1{,}600} = \frac{144}{2{,}704} = \frac{9}{169}$$

Your Turn!

An urn has three red marbles and eight blue marbles in it. Draw two marbles, one at a time, this time without replacement, from the urn. **"Without replacement"** means that you do not put the first ball back before you select the second marble. Following is a tree diagram for this situation. The branches are labeled with probabilities instead of frequencies. The numbers at the ends of the branches are calculated by multiplying the numbers on the two corresponding branches, for example, $\left(\frac{3}{11}\right)\left(\frac{2}{10}\right) = \frac{6}{110}$.

Total = $\frac{56 + 24 + 24 + 6}{110} = \frac{110}{110} = 1$



NOTE

Because we are drawing **without replacement**, on the second draw there are ten marbles left in the urn because we do not put back (replace) the first marble. This means that on the second draw, we have different numbers of red and blue marbles remaining in the two branches. If you draw a red on the first draw from the three red possibilities, there are two red marbles and eight blue marbles left to draw on the second draw.

Calculate the following probabilities using the tree diagram.

a. $P(RR)$ = _____

Solution

a. $P(RR) = \left(\frac{3}{11}\right)\left(\frac{2}{10}\right) = \frac{6}{110}$

b. Fill in the blanks:

$P(RB \text{ OR } BR) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) + (___)(___) = \frac{48}{110}$

Solution

b. $P(RB \text{ OR } BR) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) + \left(\frac{8}{11}\right)\left(\frac{3}{10}\right) = \frac{48}{110}$

c. $P(R \text{ on 2nd} | B \text{ on 1st}) =$

Solution

c. $P(R \text{ on 2nd} | B \text{ on 1st}) = \frac{3}{10}$

d. Fill in the blanks.

$P(R \text{ on 1st AND } B \text{ on 2nd}) = P(RB) = (___)(___) = \frac{24}{100}$

Solution

d. $P(R \text{ on 1st AND } B \text{ on 2nd}) = P(RB) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) = \frac{24}{100}$

e. Find $P(BB)$.

Solution

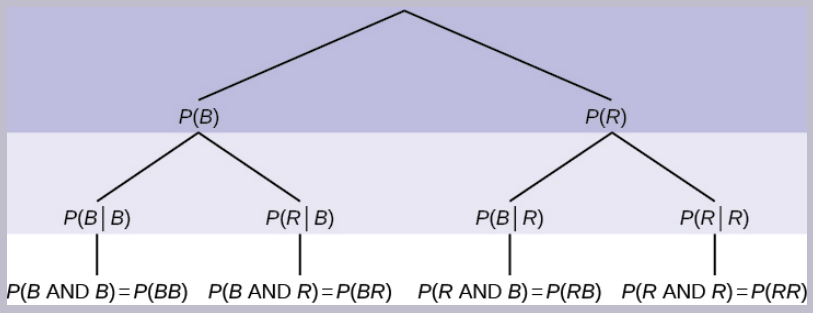
e. $P(BB) = \left(\frac{8}{11}\right)\left(\frac{7}{10}\right)$

f. Find $P(B \text{ on 2nd} | R \text{ on 1st})$.

Solution

f. Using the tree diagram, $P(B \text{ on 2nd} | R \text{ on 1st}) = P(R|B) = \frac{8}{10}$.

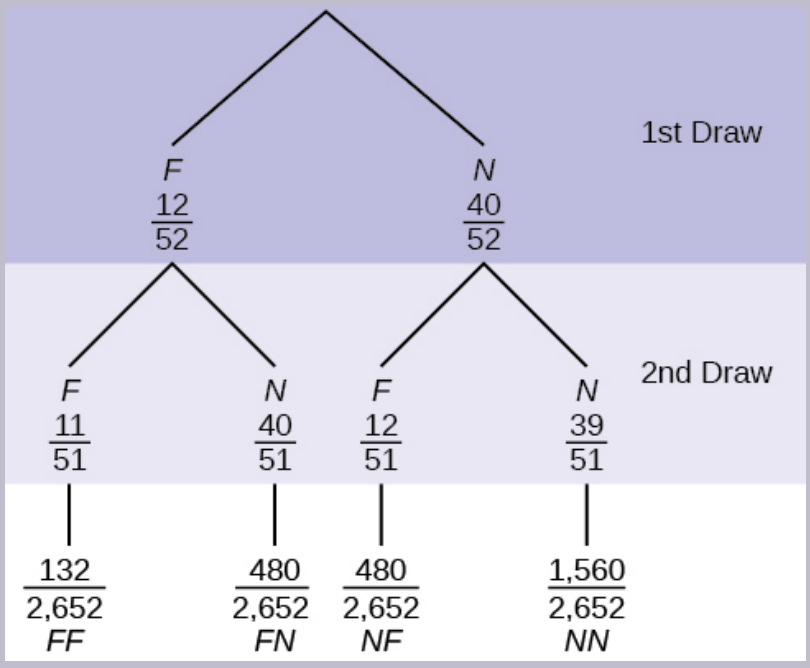
If we are using probabilities, we can label the tree in the following general way.



- $P(R|R)$ here means $P(R$ on 2nd$|R$ on 1st$)$
- $P(B|R)$ here means $P(B$ on 2nd$|R$ on 1st$)$
- $P(R|B)$ here means $P(R$ on 2nd$|B$ on 1st$)$
- $P(B|B)$ here means $P(B$ on 2nd$|B$ on 1st$)$

Your Turn!

In a standard deck, there are 52 cards. Twelve cards are face cards (F) and 40 cards are not face cards (N). Draw two cards, one at a time, without replacement. The tree diagram is labeled with all possible probabilities.



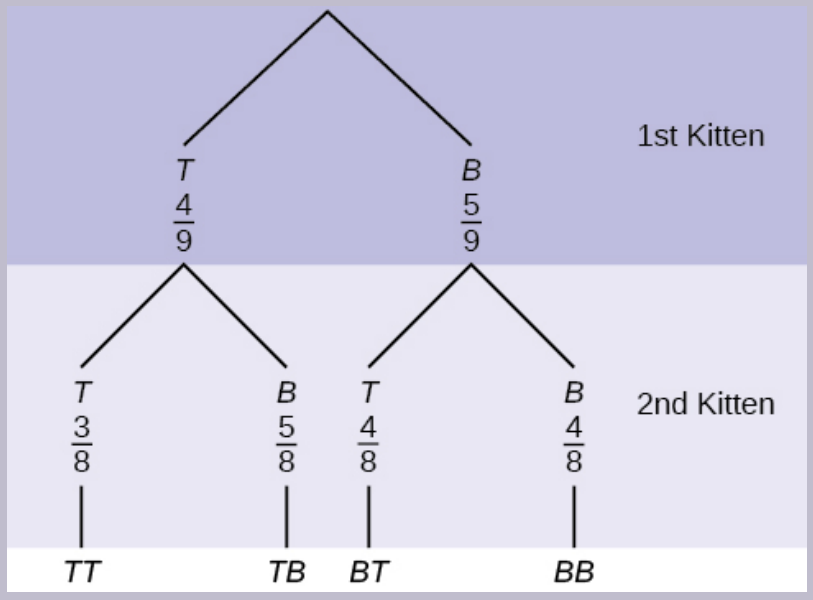
1. Find $P(FN \text{ OR } NF)$.
2. Find $P(N|F)$.
3. Find P(at most one face card).
   Hint: "At most one face card" means zero or one face card.
4. Find P(at least on face card).
   Hint: "At least one face card" means one or two face cards.

Solution

1. $P(FN \text{ OR } NF) = \frac{480}{2{,}652} + \frac{480}{2{,}652} = \frac{960}{2{,}652} = \frac{80}{221}$
2. $P(N|F) = \frac{40}{51}$
3. $P(\text{at most one face card}) = \frac{(480 + 480 + 1{,}560)}{2{,}652} = \frac{2{,}520}{2{,}652}$
4. $P(\text{at least one face card}) = \frac{(132 + 480 + 480)}{2{,}652} = \frac{1{,}092}{2{,}652}$

Your Turn!

A litter of kittens available for adoption at the Humane Society has four tabby kittens and five black kittens. A family comes in and randomly selects two kittens (without replacement) for adoption.



1. What is the probability that both kittens are tabby?
   a. $\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)$ b. $\left(\frac{4}{9}\right)\left(\frac{4}{9}\right)$ c. $\left(\frac{4}{9}\right)\left(\frac{3}{8}\right)$ d. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)$
2. What is the probability that one kitten of each coloring is selected?
   a. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)$ b. $\left(\frac{4}{9}\right)\left(\frac{5}{8}\right)$ c. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)+\left(\frac{5}{9}\right)\left(\frac{4}{9}\right)$ d. $\left(\frac{4}{9}\right)\left(\frac{5}{8}\right)+\left(\frac{5}{9}\right)\left(\frac{4}{8}\right)$
3. What is the probability that a tabby is chosen as the second kitten when a black kitten was chosen as the first?
4. What is the probability of choosing two kittens of the same color?

Solution

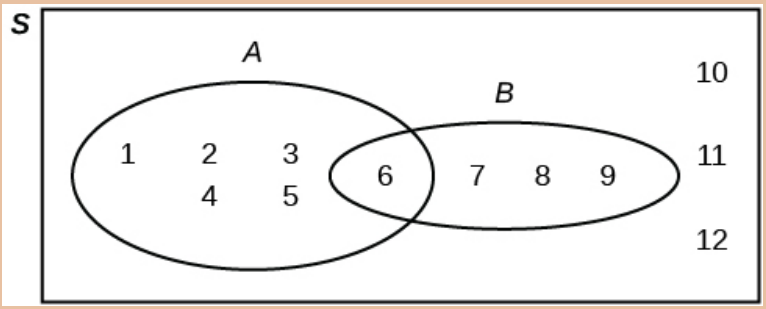
a. c, b. d, c. $\frac{4}{8}$, d. $\frac{32}{72}$

Venn Diagrams

A **Venn diagram** is a picture that represents the outcomes of an experiment. It generally consists of a box that represents the sample space S together with circles or ovals. The circles or ovals represent events.

Example

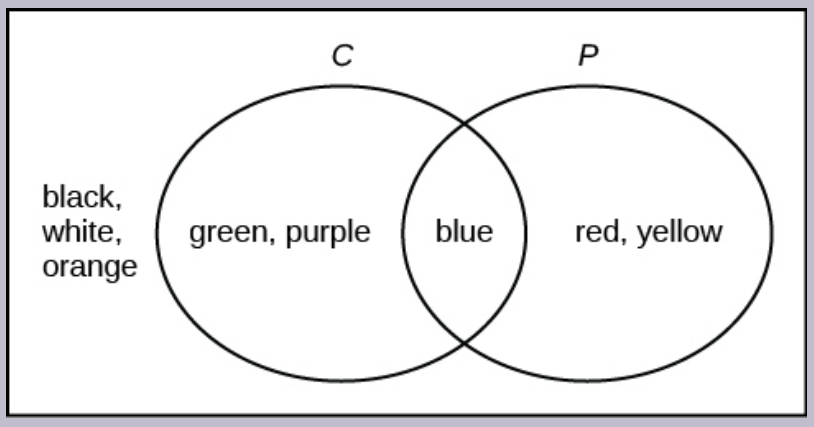
Suppose an experiment has the outcomes 1, 2, 3, ... , 12 where each outcome has an equal chance of occurring. Let event A = {1, 2, 3, 4, 5, 6} and event B = {6, 7, 8, 9}. Then A AND B = {6} and A OR B = {1, 2, 3, 4, 5, 6, 7, 8, 9}. The Venn diagram is as follows:



Your Turn!

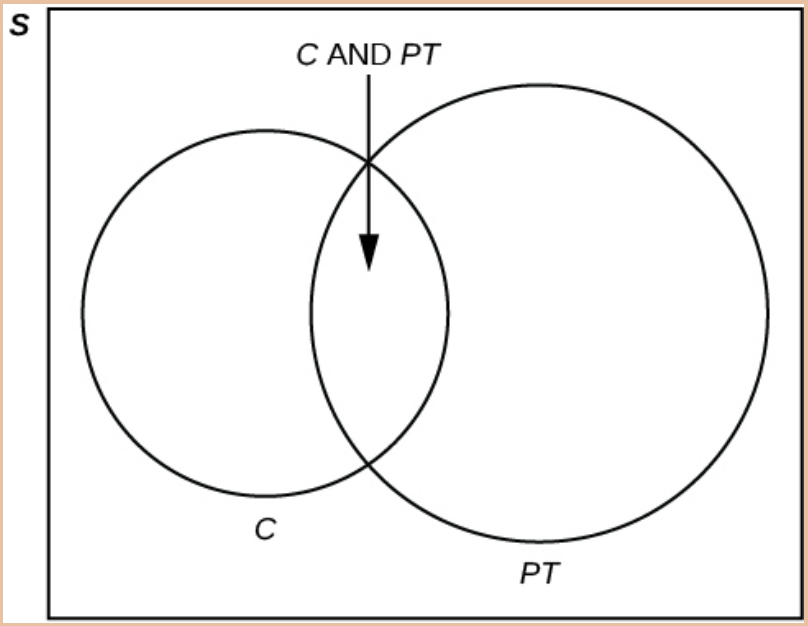
Suppose an experiment has outcomes black, white, red, orange, yellow, green, blue, and purple, where each outcome has an equal chance of occurring. Let event C = {green, blue, purple} and event P = {red, yellow, blue}. Then C AND P = {blue} and C OR P = {green, blue, purple, red, yellow}. Draw a Venn diagram representing this situation.

Solution



Example

Forty percent of the students at a local college belong to a club and **50%** work part time. **Five percent** of the students work part time and belong to a club. Draw a Venn diagram showing the relationships. Let *C* = student belongs to a club and *PT* = student works part time.



If a student is selected at random, find

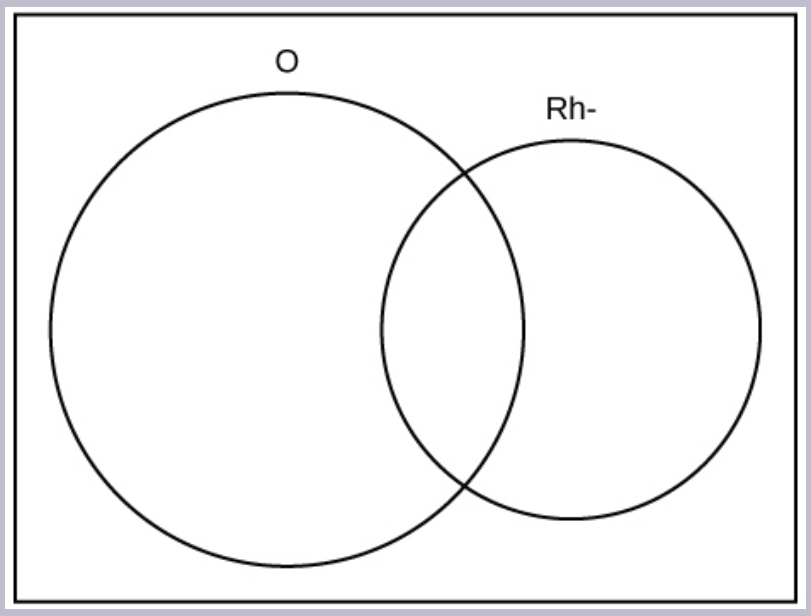
- the probability that the student belongs to a club. $P(C) = 0.40$
- the probability that the student works part time. $P(PT) = 0.50$
- the probability that the student belongs to a club AND works part time. $P(C \text{ AND } PT) = 0.05$
- the probability that the student belongs to a club **given** that the student works part time.

$$P(C|PT) = \frac{P(C \text{ AND } PT)}{P(PT)} = \frac{0.05}{0.50} = 0.1$$

- the probability that the student belongs to a club **OR** works part time. $P(C \text{ OR } PT) = P(C) + P(PT) - P(C \text{ AND } PT) = 0.40 + 0.50 - 0.05 = 0.85$

Your Turn!

A person with type O blood and a negative Rh factor (Rh-) can donate blood to any person with any blood type. Four percent of African Americans have type O blood and a negative RH factor, 5–10% of African Americans have the Rh- factor, and 51% have type O blood.



The "O" circle represents the African Americans with type O blood. The "Rh-" oval represents the African Americans with the Rh- factor.

We will take the average of 5% and 10% and use 7.5% as the percent of African Americans who have the Rh- factor. Let O = African American with Type O blood and R = African American with Rh- factor.

1. $P(O)$ = _____
2. $P(R)$ = _____
3. $P(O \text{ AND } R)$ = _____
4. $P(O \text{ OR } R)$ = _____
5. In the Venn diagram, describe the overlapping area using a complete sentence.
6. In the Venn diagram, describe the area in the rectangle but outside both the circle and the oval using a complete sentence.

Solution

a. 0.51; b. 0.075; c. 0.04; d. 0.545; e. The area represents the African Americans that have type O blood and the Rh- factor. f. The area represents the African Americans that have neither type O blood nor the Rh- factor.

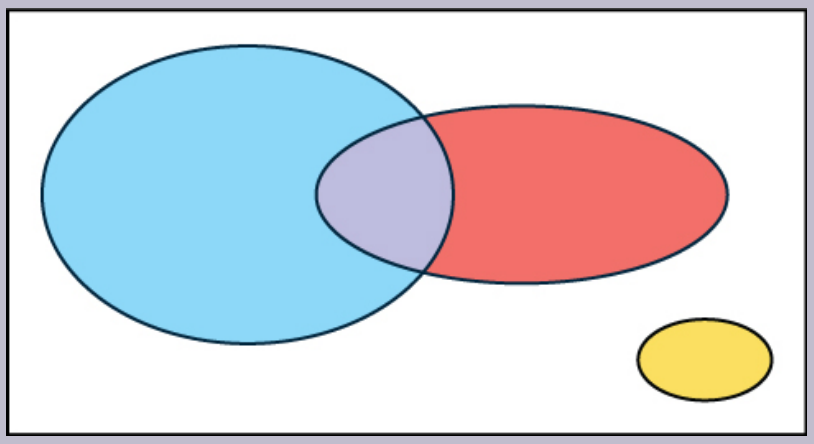
Your Turn!

In a bookstore, the probability that the customer buys a novel is 0.6, and the probability that the customer buys a nonfiction book is 0.4. Suppose that the probability that the customer buys both is 0.2.

1. Draw a Venn diagram representing the situation.
2. Find the probability that the customer buys either a novel or nonfiction book.
3. In the Venn diagram, describe the overlapping area using a complete sentence.
4. Suppose that some customers buy only compact disks. Draw an oval in your Venn diagram representing this event.

Solution

a. and d. In the following Venn diagram below, the blue oval represents customers buying a novel, the red oval represents customers buying nonfiction, and the yellow oval represents customers who buy compact disks.



b. P(novel or nonfiction) = P(Blue OR Red) = P(Blue) + P(Red) – P(Blue AND Red) = 0.6 + 0.4 – 0.2 = 0.8.

c. The overlapping area of the blue oval and red oval represents the customers buying both a novel and a nonfiction book.

References

Data from Clara County Public H.D.

Data from the American Cancer Society.

Data from The Data and Story Library, 1996. Available online at http://lib.stat.cmu.edu/DASL/ (accessed May 2, 2013).

Data from the Federal Highway Administration, part of the United States Department of Transportation.

Data from the United States Census Bureau, part of the United States Department of Commerce.

Data from USA Today.

"Environment." The World Bank, 2013. Available online at http://data.worldbank.org/topic/environment (accessed May 2, 2013).

"Search for Datasets." Roper Center: Public Opinion Archives, University of Connecticut., 2013. Available online at http://www.ropercenter.uconn.edu/data_access/data/search_for_datasets.html (accessed May 2, 2013).

CHAPTER 3 REVIEW

Section 3.1 Review

In this module we learned the basic terminology of probability. The set of all possible outcomes of an experiment is called the sample space. Events are subsets of the sample space, and they are assigned a probability that is a number between zero and one, inclusive.

Formula Review

A and B represent events

$0 \leq P(A) \leq 1$

If S is the sample space of an experiment, $P(S) = 1$

$$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)}$$

or equivalently, $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Section 3.2 Review

Two events A and B are independent if the knowledge that one occurred does not affect the chance the other occurs. If two events are not independent, then we say that they are dependent.

In sampling with replacement, each member of a population is replaced after it is picked, so that member has the possibility of being chosen more than once, and the events are considered to be independent. In sampling without replacement, each member of a population may be chosen only once, and the events are considered not to be independent. When events do not share outcomes, they are mutually exclusive of each other.

Formula Review

If A and B are independent, $P(A \text{ AND } B) = P(A)P(B)$, $P(A|B) = P(A)$ and $P(B|A) = P(B)$. Equivalently, $P(A \cap B) = P(A)P(B)$.

If A and B are mutually exclusive, $P(A \text{ OR } B) = P(A) + P(B)$ and $P(A \text{ AND } B) = 0$. Equivalently, $P(A \cup B) = P(A) + P(B)$ and $P(A \cap B) = 0$ because $A \cap B = \emptyset$.

Section 3.3 Review

The multiplication rule and the addition rule are used for computing the probability of A and B, as well as the probability of A and B for two given events A and B defined on the sample space. In sampling with replacement each member of a population is replaced after it is picked, so that member has the possibility of being chosen more than once, and the events are considered to be independent. In sampling without replacement, each member of a population may be chosen only once, and the events are considered to be not independent. The events A and B are mutually exclusive events when they do not have any outcomes in common.

Formula Review

The multiplication rule: $P(A \text{ AND } B) = P(A|B)P(B)$
or equivalently, $P(A \cap B) = P(A|B)P(B)$
The addition rule: $P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B)$

or equivalently, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Section 3.4 Review

There are several tools you can use to help organize and sort data when calculating probabilities. Contingency tables help display data and are particularly useful when calculating probabilities that have multiple dependent variables.

Section 3.5 Review

A tree diagram uses branches to show the different outcomes of experiments and makes complex probability questions easy to visualize.

A Venn diagram is a picture that represents the outcomes of an experiment. It generally consists of a box that represents the sample space *S* together with circles or ovals. The circles or ovals represent events. A Venn

diagram is especially helpful for visualizing the OR event, the AND event, and the complement of an event and for understanding conditional probabilities.

CHAPTER 3 PRACTICE

Section 3.1

Exercises

In a particular college class, there are male and female students. Some students have long hair and some students have short hair. Write the **symbols** for the probabilities of the events for parts a through j. (Note that you cannot find numerical answers here. You were not given enough information to find any probability values yet; concentrate on understanding the symbols.)

- Let F be the event that a student is female.
- Let M be the event that a student is male.
- Let S be the event that a student has short hair.
- Let L be the event that a student has long hair.

1. The probability that a student does not have long hair.
2. The probability that a student is male or has short hair.
3. The probability that a student is a female and has long hair.
4. The probability that a student is male, given that the student has long hair.
5. The probability that a student has long hair, given that the student is male.
6. Of all the female students, the probability that a student has short hair.
7. Of all students with long hair, the probability that a student is female.
8. The probability that a student is female or has long hair.
9. The probability that a randomly selected student is a male student with short hair.
10. The probability that a student is female.

Solution

1. $P(L') = P(S)$
2. $P(M \text{ OR } S)$

3. $P(F \text{ AND } L)$
4. $P(M|L)$
5. $P(L|M)$
6. $P(S|F)$
7. $P(F|L)$
8. $P(F \text{ OR } L)$
9. $P(M \text{ AND } S)$
10. $P(F)$

Exercises

Use the following information to answer the next four exercises. A box is filled with several party favors. It contains 12 hats, 15 noisemakers, ten finger traps, and five bags of confetti.

- Let H = the event of getting a hat.
- Let N = the event of getting a noisemaker.
- Let F = the event of getting a finger trap.
- Let C = the event of getting a bag of confetti.

Find $P(H)$.

Find $P(N)$.

Solution

$P(N) = \frac{15}{42} = \frac{5}{14} = 0.36$

Find $P(F)$.

Find $P(C)$.

Solution

$P(C) = \frac{5}{42} = 0.12$

Exercises

Use the following information to answer the next six exercises. A jar of 150 jelly beans contains 22 red jelly beans, 38 yellow, 20 green, 28 purple, 26 blue, and the rest are orange.

- Let B = the event of getting a blue jelly bean.
- Let G = the event of getting a green jelly bean.
- Let O = the event of getting an orange jelly bean.
- Let P = the event of getting a purple jelly bean.
- Let R = the event of getting a red jelly bean.
- Let Y = the event of getting a yellow jelly bean.

Find $P(B)$.

Find $P(G)$.

Solution

$P(G) = \frac{20}{150} = \frac{2}{15} = 0.13$

Find $P(P)$.

Find $P(R)$.

Solution

$P(R) = \frac{22}{150} = \frac{11}{75} = 0.15$

Find $P(Y)$.

Find $P(O)$.

Solution

$P(O) = \frac{150 - 22 - 38 - 20 - 28 - 26}{150} = \frac{16}{150} = \frac{8}{75} = 0.11$

Exercises

Use the following information to answer the next six exercises. There are 23 countries in North America, 12 countries in South America, 47 countries in Europe, 44 countries in Asia, 54 countries in Africa, and 14 in Oceania (Pacific Ocean region).

- Let A = the event that a country is in Asia.
- Let E = the event that a country is in Europe.
- Let F = the event that a country is in Africa.
- Let N = the event that a country is in North America.
- Let O = the event that a country is in Oceania.
- Let S = the event that a country is in South America.

Find $P(A)$.

Find $P(E)$.

Solution

$P(E) = \frac{47}{194} = 0.24$

Find $P(F)$.

Find $P(N)$.

Solution

$P(N) = \frac{23}{194} = 0.12$

Find $P(O)$.

Find $P(S)$.

Solution

$P(S) = \frac{12}{194} = \frac{6}{97} = 0.06$

Exercises

What is the probability of drawing a red card in a standard deck of 52 cards?

What is the probability of drawing a club in a standard deck of 52 cards?

Solution

$\frac{13}{52} = \frac{1}{4} = 0.25$

Exercises

What is the probability of rolling an even number of dots with a fair, six-sided die numbered one through six?

What is the probability of rolling a prime number of dots with a fair, six-sided die numbered one through six?

Solution

$\frac{3}{6} = \frac{1}{2} = 0.5$

Exercises

Use the following information to answer the next two exercises. You see a game at a local fair. You have to throw a dart at a color wheel. Each section on the color wheel is equal in area.



- Let B = the event of landing on blue.
- Let R = the event of landing on red.
- Let G = the event of landing on green.
- Let Y = the event of landing on yellow.

If you land on Y, you get the biggest prize. Find $P(Y)$.

If you land on red, you don't get a prize. What is $P(R)$?

Solution

$$P(R) = \frac{4}{8} = 0.5$$

Exercises

Use the following information to answer the next ten exercises. On a baseball team, there are infielders and outfielders. Some players are great hitters, and some players are not great hitters.

- Let I = the event that a player is an infielder.
- Let O = the event that a player is an outfielder.
- Let H = the event that a player is a great hitter.
- Let N = the event that a player is not a great hitter.

Write the symbols for the probability that a player is not an outfielder.

Write the symbols for the probability that a player is an outfielder or is a great hitter.

Solution
$P(O \text{ OR } H)$

Write the symbols for the probability that a player is an infielder and is not a great hitter.

Write the symbols for the probability that a player is a great hitter, given that the player is an infielder.

Solution
$P(H|I)$

Write the symbols for the probability that a player is an infielder, given that the player is a great hitter.

Write the symbols for the probability that of all the outfielders, a player is not a great hitter.

Solution
$P(N|O)$

Write the symbols for the probability that of all the great hitters, a player is an outfielder.

Write the symbols for the probability that a player is an infielder or is not a great hitter.

Solution
$P(I \text{ OR } N)$

Write the symbols for the probability that a player is an outfielder and is a great hitter.

Write the symbols for the probability that a player is an infielder.

Solution
$P(I)$

Exercises

What is the word for the set of all possible outcomes?

What is conditional probability?

Solution
The likelihood that an event will occur given that another event has already occurred.

Exercises

A shelf holds 12 books. Eight are fiction and the rest are nonfiction. Each is a different book with a unique title. The fiction books are numbered one to eight. The nonfiction books are numbered one to four. Randomly select one book

- Let F = event that book is fiction
- Let N = event that book is nonfiction

What is the sample space?

What is the sum of the probabilities of an event and its complement?

Solution
1

Exercises

Use the following information to answer the next two exercises. You are rolling a fair, six-sided number cube. Let E = the event that it lands on an even number. Let M = the event that it lands on a multiple of three.

What does $P(E|M)$ mean in words?

What does $P(E \text{ OR } M)$ mean in words?

Solution
the probability of landing on an even number or a multiple of three

Section 3.2

Exercises

E and F are mutually exclusive events. $P(E) = 0.4$; $P(F) = 0.5$. Find $P(E \mid F)$.

Exercises

J and K are independent events. $P(J|K) = 0.3$. Find $P(J)$.

Solution
$P(J) = 0.3$

Exercises

U and V are mutually exclusive events. $P(U) = 0.26$; $P(V) = 0.37$. Find:

1. $P(U \text{ AND } V) =$
2. $P(U|V) =$
3. $P(U \text{ OR } V) =$

Exercises

Q and R are independent events. $P(Q) = 0.4$ and $P(Q \text{ AND } R) = 0.1$. Find $P(R)$.

Solution

$P(Q \text{ AND } R) = P(Q)P(R)$

$0.1 = (0.4)P(R)$

$P(R) = 0.25$

Section 3.3

Exercises

Use the following information to answer the next ten exercises. Forty-eight percent of all Californians registered voters prefer life in prison without parole over the death penalty for a person convicted of first degree murder. Among Latino California registered voters, 55% prefer life in prison without parole over the death penalty for a person convicted of first degree murder. 37.6% of all Californians are Latino.

In this problem, let:

- C = Californians (registered voters) preferring life in prison without parole over the death penalty for a person convicted of first degree murder.
- L = Latino Californians

Suppose that one Californian is randomly selected.

Find $P(C)$.

Find $P(L)$.

Solution
0.376

Find $P(C|L)$.

In words, what is $C|L$?

Solution
$C|L$ means, given the person chosen is a Latino Californian, the person is a registered voter who prefers life in prison without parole for a person convicted of first degree murder.

Find $P(L \text{ AND } C)$.

In words, what is L AND C?

Solution
L AND C is the event that the person chosen is a Latino California registered voter who prefers life without parole over the death penalty for a person convicted of first degree murder.

Are L and C independent events? Show why or why not.

Find $P(L \text{ OR } C)$.

Solution

0.6492

In words, what is *L* OR *C*?

Are *L* and *C* mutually exclusive events? Show why or why not.

Solution
No, because *P*(*L* AND *C*) does not equal 0.

Section 3.4

Exercises

Use the following information to answer the next four exercises. The following table shows a random sample of musicians and how they learned to play their instruments. Find *P*(musician is a female).

Gender	Self-taught	Studied in School	Private Instruction	Total
Female	12	38	21	71
Male	19	24	15	58
Other	0	0	1	1
Total	31	62	37	130

Find *P*(musician is a male AND had private instruction).

Solution

P(musician is a male AND had private instruction) = $\frac{15}{130} = \frac{3}{26} = 0.12$

Find *P*(musician is a female OR is self taught).

Are the events “being a female musician” and “learning music in school” independent events?

Solution

P(being a female musician AND learning music in school) = $\frac{38}{130} = \frac{19}{65}$ = 0.29

P(being a female musician)P(learning music in school) = $\left(\frac{71}{130}\right)\left(\frac{62}{130}\right) = \frac{4,402}{16,900} = \frac{2,201}{8,450} \approx 0.26$

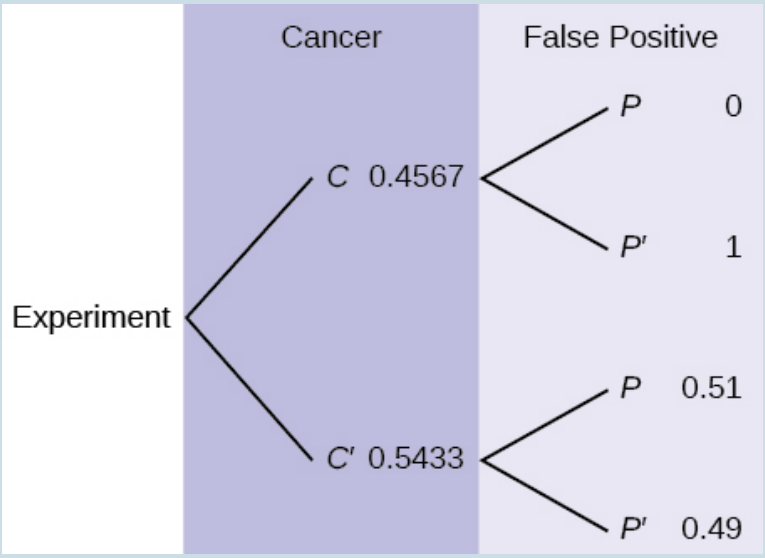
No, they are not independent because P(being a female musician AND learning music in school) is not equal to P(being a female musician)P(learning music in school).

Section 3.5

Exercises

The probability that a man develops some form of cancer in his lifetime is 0.4567. The probability that a man has at least one false positive test result (meaning the test comes back for cancer when the man does not have it) is 0.51. Let: C = a man develops cancer in his lifetime; P = man has at least one false positive. Construct a tree diagram of the situation.

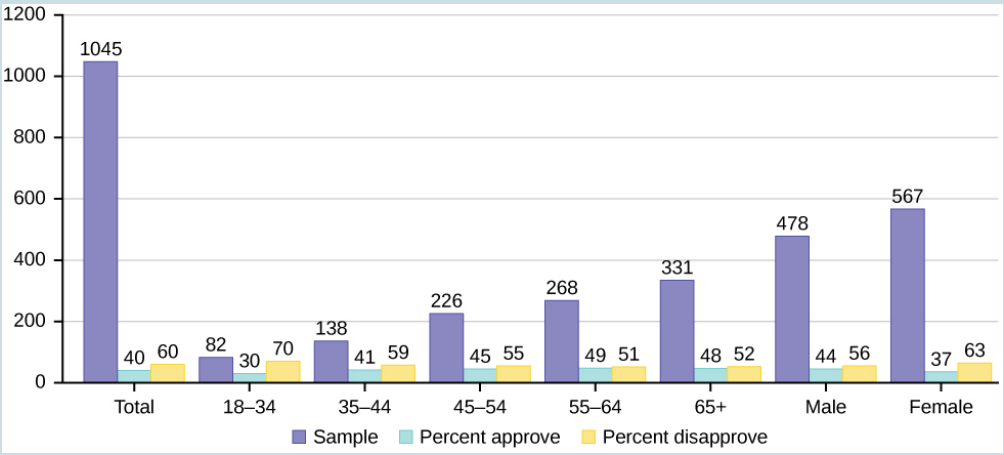
Solution



CHAPTER 3 HOMEWORK

Section 3.1

Exercises



The graph in [link] displays the sample sizes and percentages of people in different age and gender groups who were polled concerning their approval of Mayor Ford's actions in office. The total number in the sample of all the age groups is 1,045.

1. Define three events in the graph.
2. Describe in words what the entry 40 means.
3. Describe in words the complement of the entry in question 2.
4. Describe in words what the entry 30 means.
5. Out of the males and females, what percent are males?
6. Out of the females, what percent disapprove of Mayor Ford?
7. Out of all the age groups, what percent approve of Mayor Ford?
8. Find $P(\text{Approve}|\text{Male})$.
9. Out of the age groups, what percent are more than 44 years old?
10. Find $P(\text{Approve}|\text{Age} < 35)$.

Exercises

Explain what is wrong with the following statements. Use complete sentences.

1. If there is a 60% chance of rain on Saturday and a 70% chance of rain on Sunday, then there is a 130% chance of rain over the weekend.
2. The probability that a baseball player hits a home run is greater than the probability that he gets a successful hit.

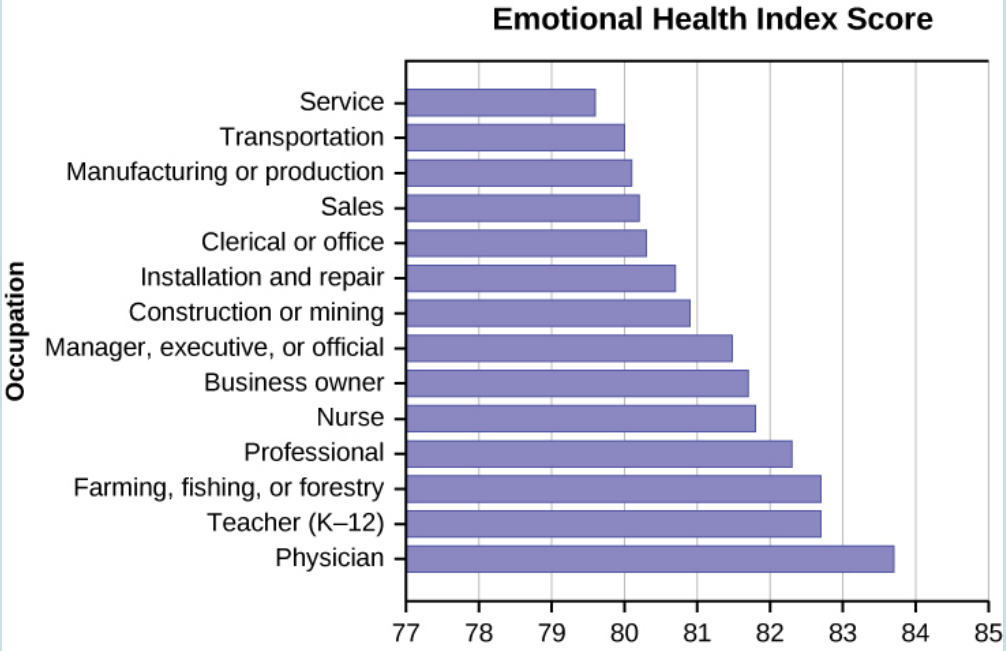
Solution

1. You can't calculate the joint probability knowing the probability of both events occurring, which is not in the information given; the probabilities should be multiplied, not added; and probability is never greater than 100%
2. A home run by definition is a successful hit, so he has to have at least as many successful hits as home runs.

Section 3.2

Exercises

Use the following information to answer the next 12 exercises. The graph shown is based on more than 170,000 interviews done by Gallup that took place from January through December 2012. The sample consists of employed Americans 18 years of age or older. The Emotional Health Index Scores are the sample space. We randomly sample one Emotional Health Index Score.



1. Find the probability that an Emotional Health Index Score is 82.7.

2. Find the probability that an Emotional Health Index Score is 81.0.

Solution
0

3. Find the probability that an Emotional Health Index Score is more than 81?

4. Find the probability that an Emotional Health Index Score is between 80.5 and 82?

Solution
0.3571

5. If we know an Emotional Health Index Score is 81.5 or more, what is the probability that it is 82.7?

6. What is the probability that an Emotional Health Index Score is 80.7 or 82.7?

Solution
0.2142

7. What is the probability that an Emotional Health Index Score is less than 80.2 given that it is already less than 81.

8. What occupation has the highest emotional index score?

Solution
Physician (83.7)

9. What occupation has the lowest emotional index score?

10. What is the range of the data?

Solution
83.7 – 79.6 = 4.1

11. Compute the average EHIS.

12. If all occupations are equally likely for a certain individual, what is the probability that he or she will have an occupation with lower than average EHIS?

Solution
P(Occupation < 81.3) = 0.5

Bringing It Together

Exercises

A previous year, the weights of the members of the **San Francisco 49ers** and the **Dallas Cowboys** were published in the SAN JOSE MERCURY NEWS. The factual data are compiled into [link].

Shirt#	≤ 210	211–250	251–290	290≤
1–33	21	5	0	0
34–66	6	18	7	4
66–99	6	12	22	5

Suppose that you randomly select one player from the 49ers or Cowboys. If having a shirt number

from one to 33 and weighing at most 210 pounds were independent events, then what should be true about P(Shirt# 1–33|≤ 210 pounds)?

Exercises

The probability that a male develops some form of cancer in his lifetime is 0.4567. The probability that a male has at least one false positive test result (meaning the test comes back for cancer when the man does not have it) is 0.51. Some of the following questions do not have enough information for you to answer them. Write "not enough information" for those answers. Let C = a man develops cancer in his lifetime and P = man has at least one false positive.

a. $P(C)$ = _____
b. $P(P|C)$ = _____
c. $P(P|C')$ = _____
d. If a test comes up positive, based upon numerical values, can you assume that man has cancer? Justify numerically and explain why or why not.
Solution
a. $P(C)$ = 0.4567

b. not enough information

c. not enough information

c. No, because over half (0.51) of men have at least one false positive text

Exercises

Given events G and H: $P(G)$ = 0.43; $P(H)$ = 0.26; $P(H \text{ AND } G)$ = 0.14

a. Find $P(H \text{ OR } G)$.

b. Find the probability of the complement of event (H AND G).

c. Find the probability of the complement of event (H OR G).

Exercises

Given events J and K: $P(J) = 0.18$; $P(K) = 0.37$; $P(J \text{ OR } K) = 0.45$

a. Find $P(J \text{ OR } K)$.

b. Find the probability of the complement of event (J AND K).

c. Find the probability of the complement of event (J AND K).

Solution

a. $P(J \text{ OR } K) = P(J) + P(K) - P(J \text{ AND } K)$; $0.45 = 0.18 + 0.37 - P(J \text{ AND } K)$; solve to find $P(J \text{ AND } K) = 0.10$

b. $P(\text{NOT } (J \text{ AND } K)) = 1 - P(J \text{ AND } K) = 1 - 0.10 = 0.90$

c. $P(\text{NOT } (J \text{ OR } K)) = 1 - P(J \text{ OR } K) = 1 - 0.45 = 0.55$

Section 3.3

Exercises

On February 28, 2013, a Field Poll Survey reported that 61% of California registered voters approved of allowing two people of the same gender to marry and have regular marriage laws apply to them. Among 18 to 39 year olds (California registered voters), the approval rating was 78%. Six in ten

California registered voters said that the upcoming Supreme Court's ruling about the constitutionality of California's Proposition 8 was either very or somewhat important to them. Out of those CA registered voters who support same-sex marriage, 75% say the ruling is important to them.

In this problem, let:

- C = California registered voters who support same-sex marriage.
- B = California registered voters who say the Supreme Court's ruling about the constitutionality of California's Proposition 8 is very or somewhat important to them
- A = California registered voters who are 18 to 39 years old.

1. Find $P(C)$.
2. Find $P(B)$.
3. Find $P(C|A)$.
4. Find $P(B|C)$.
5. In words, what is $C|A$?
6. In words, what is $B|C$?
7. Find $P(C \text{ AND } B)$.
8. In words, what is C AND B?
9. Find $P(C \text{ OR } B)$.
10. Are C and B mutually exclusive events? Show why or why not.

Exercises

After Rob Ford, the mayor of Toronto, announced his plans to cut budget costs in late 2011, the Forum Research polled 1,046 people to measure the mayor's popularity. Everyone polled expressed either approval or disapproval. These are the results their poll produced:

- In early 2011, 60 percent of the population approved of Mayor Ford's actions in office.
- In mid-2011, 57 percent of the population approved of his actions.
- In late 2011, the percentage of popular approval was measured at 42 percent.

1. What is the sample size for this study?
2. What proportion in the poll disapproved of Mayor Ford, according to the results from late 2011?
3. How many people polled responded that they approved of Mayor Ford in late 2011?
4. What is the probability that a person supported Mayor Ford, based on the data collected in mid-2011?
5. What is the probability that a person supported Mayor Ford, based on the data collected in early 2011?

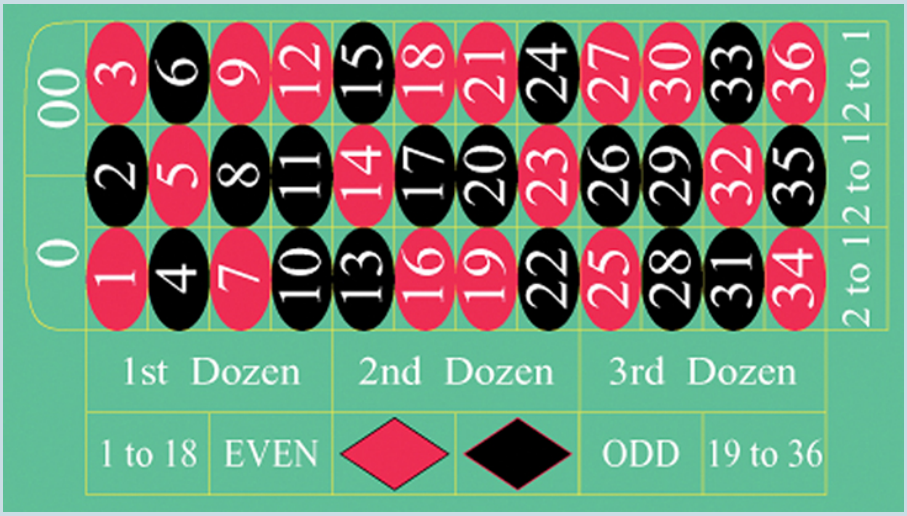
Solution

1. The Forum Research surveyed 1,046 Torontonians.
2. 58%
3. 42% of 1,046 = 439 (rounding to the nearest integer)
4. 0.57
5. 0.60.

Exercises

Use the following information to answer the following exercises. The casino game, roulette, allows the gambler to bet on the probability of a ball, which spins in the roulette wheel, landing on a particular color, number, or range of numbers. The table used to place bets contains 38 numbers, and each number is assigned to a color and a range.

(credit: film8ker/wikibooks)



1. Find the following probabilities.
    1. List the sample space of the 38 possible outcomes in roulette.
    2. You bet on red. Find P(red).
    3. You bet on -1st 12- (1st Dozen). Find P(-1st 12-).
    4. You bet on an even number. Find P(even number). Is getting an odd number the complement of getting an even number? Why?
2. Find two mutually exclusive events.
3. Are the events Even and 1st Dozen independent?
4. Compute the probability of winning the following types of bets:
    1. Betting on two lines that touch each other on the table as in 1-2-3-4-5-6
    2. Betting on three numbers in a line, as in 1-2-3
    3. Betting on one number
    4. Betting on four numbers that touch each other to form a square, as in 10-11-13-14
    5. Betting on two numbers that touch each other on the table, as in 10-11 or 10-13
    6. Betting on 0-00-1-2-3
    7. Betting on 0-1-2; or 0-00-2; or 00-2-3
5. Compute the probability of winning the following types of bets:
    1. Betting on a color
    2. Betting on one of the dozen groups
    3. Betting on the range of numbers from 1 to 18
    4. Betting on the range of numbers 19–36

5. Betting on one of the columns
6. Betting on an even or odd number (excluding zero)

Solution to number 4

1. *P*(Betting on two line that touch each other on the table) = $\frac{6}{38}$
2. *P*(Betting on three numbers in a line) = $\frac{3}{38}$
3. *P*(Betting on one number) = $\frac{1}{38}$
4. *P*(Betting on four number that touch each other to form a square) = $\frac{4}{38}$
5. *P*(Betting on two number that touch each other on the table) = $\frac{2}{38}$
6. *P*(Betting on 0-00-1-2-3) = $\frac{5}{38}$
7. *P*(Betting on 0-1-2; or 0-00-2; or 00-2-3) = $\frac{3}{38}$

Exercises

Suppose that you have eight cards. Five are green and three are yellow. The five green cards are numbered 1, 2, 3, 4, and 5. The three yellow cards are numbered 1, 2, and 3. The cards are well shuffled. You randomly draw one card.

- *G* = card drawn is green
- *E* = card drawn is even-numbered

1. List the sample space.
2. *P*(*G*) = _____

3. $P(G|E)$ = _____
4. $P(G \text{ AND } E)$ = _____
5. $P(G \text{ OR } E)$ = _____
6. Are G and E mutually exclusive? Justify your answer numerically.

Solution

1. {$G1$, $G2$, $G3$, $G4$, $G5$, $Y1$, $Y2$, $Y3$}
2. $\frac{5}{8}$
3. $\frac{2}{3}$
4. $\frac{2}{8}$
5. $\frac{6}{8}$
6. No, because $P(G \text{ AND } E)$ does not equal 0.

Exercises

Roll two fair dice. Each die has six faces.

1. List the sample space.
2. Let A be the event that either a three or four is rolled first, followed by an even number. Find $P(A)$.
3. Let B be the event that the sum of the two rolls is at most seven. Find $P(B)$.
4. In words, explain what "$P(A|B)$" represents. Find $P(A|B)$.
5. Are A and B mutually exclusive events? Explain your answer in one to three complete sentences, including numerical justification.
6. Are A and B independent events? Explain your answer in one to three complete sentences, including numerical justification.

Exercises A special deck of cards has ten cards. Four are green, three are blue, and three are red. When a card is picked, its color is recorded. An experiment consists of first picking a card and then tossing a coin.

1. List the sample space.
2. Let *A* be the event that a blue card is picked first, followed by landing a head on the coin toss. Find *P*(*A*).
3. Let *B* be the event that a red or green is picked, followed by landing a head on the coin toss. Are the events *A* and *B* mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.
4. Let *C* be the event that a red or blue is picked, followed by landing a head on the coin toss. Are the events *A* and *C* mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.

Solution

Note: The coin toss is independent of the card picked first.

1. {(*G*,*H*) (*G*,*T*) (*B*,*H*) (*B*,*T*) (*R*,*H*) (*R*,*T*)}
2. *P*(*A*) = *P*(blue)*P*(head) = $\left(\frac{3}{10}\right)\left(\frac{1}{2}\right) =< /span > \frac{3}{20}$
3. Yes, *A* and *B* are mutually exclusive because they cannot happen at the same time; you cannot pick a card that is both blue and also (red or green). *P*(*A* AND *B*) = 0
4. No, *A* and *C* are not mutually exclusive because they can occur at the same time. In fact, *C* includes all of the outcomes of *A*; if the card chosen is blue it is also (red or blue). *P*(*A* AND *C*) = *P*(*A*) = $\frac{3}{20}$

Exercises

An experiment consists of first rolling a die and then tossing a coin.

52. List the sample space.

53. Let A be the event that either a three or a four is rolled first, followed by landing a head on the coin toss. Find $P(A)$.

54. Let B be the event that the first and second tosses land on heads. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.

Exercises

An experiment consists of tossing a nickel, a dime, and a quarter. Of interest is the side the coin lands on.

1. List the sample space.
2. Let A be the event that there are at least two tails. Find $P(A)$.
3. Let B be the event that the first and second tosses land on heads. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including justification.

Solution

1. $S = \{(HHH), (HHT), (HTH), (HTT), (THH), (THT), (TTH), (TTT)\}$
2. $\frac{4}{8}$
3. Yes, because if A has occurred, it is impossible to obtain two tails. In other words, $P(A \text{ AND } B) = 0$.

Exercises

Consider the following scenario:

- Let $P(C) = 0.4$.
- Let $P(D) = 0.5$.
- Let $P(C|D) = 0.6$.

1. Find $P(C \text{ AND } D)$.
2. Are C and D mutually exclusive? Why or why not?
3. Are C and D independent events? Why or why not?
4. Find $P(C \text{ OR } D)$.
5. Find $P(D|C)$.

Exercises

Y and Z are independent events.

1. Rewrite the basic Addition Rule $P(Y \text{ OR } Z) = P(Y) + P(Z) - P(Y \text{ AND } Z)$ using the information that Y and Z are independent events.
2. Use the rewritten rule to find $P(Z)$ if $P(Y \text{ OR } Z) = 0.71$ and $P(Y) = 0.42$.

Solution

1. If Y and Z are independent, then $P(Y \text{ AND } Z) = P(Y)P(Z)$, so $P(Y \text{ OR } Z) = P(Y) + P(Z) - P(Y)P(Z)$.
2. 0.5

Exercises

G and H are mutually exclusive events. $P(G) = 0.5$ $P(H) = 0.3$

65. Explain why the following statement MUST be false: $P(H|G) = 0.4$.

66. Find $P(H \text{ OR } G)$.

67. Are G and H independent or dependent events? Explain in a complete sentence.

Exercises

Approximately 281,000,000 people over age five live in the United States. Of these people, 55,000,000 speak a language other than English at home. Of those who speak another language at home, 62.3% speak Spanish.

Let: E = speaks English at home; E' = speaks another language at home; S = speaks Spanish;

Finish each probability statement by matching the correct answer.

Probability Statements	Answers
a. $P(E') =$	i. 0.8043
b. $P(E) =$	ii. 0.623
c. $P(S \text{ and } E') =$	iii. 0.1957
d. $P(S\|E') =$	iv. 0.1219

Exercises

In 1994, the U.S. government held a lottery to issue 55,000 Green Cards (permits for non-citizens to work legally in the U.S.). Renate Deutsch, from Germany, was one of approximately 6.5 million people who entered this lottery. Let G = won green card.

1. What was Renate's chance of winning a Green Card? Write your answer as a probability statement.
2. In the summer of 1994, Renate received a letter stating she was one of 110,000 finalists chosen. Once the finalists were chosen, assuming that each finalist had an equal chance to win, what was Renate's chance of winning a Green Card? Write your answer as a conditional probability statement. Let F = was a finalist.
3. Are G and F independent or dependent events? Justify your answer numerically and also explain why.
4. Are G and F mutually exclusive events? Justify your answer numerically and explain why.

Exercises

Three professors at George Washington University did an experiment to determine if economists are more selfish than other people. They dropped 64 stamped, addressed envelopes with $10 cash in different classrooms on the George Washington campus. 44% were returned overall. From the economics classes 56% of the envelopes were returned. From the business, psychology, and history classes 31% were returned.

Let: R = money returned; E = economics classes; O = other classes

1. Write a probability statement for the overall percent of money returned.
2. Write a probability statement for the percent of money returned out of the economics classes.

3. Write a probability statement for the percent of money returned out of the other classes.
4. Is money being returned independent of the class? Justify your answer numerically and explain it.
5. Based upon this study, do you think that economists are more selfish than other people? Explain why or why not. Include numbers to justify your answer.

Solution

1. $P(R) = 0.44$
2. $P(R|E) = 0.56$
3. $P(R|O) = 0.31$
4. No, whether the money is returned is not independent of which class the money was placed in. There are several ways to justify this mathematically, but one is that the money placed in economics classes is not returned at the same overall rate; $P(R|E) \neq P(R)$.
5. No, this study definitely does not support that notion; *in fact*, it suggests the opposite. The money placed in the economics classrooms was returned at a higher rate than the money place in all classes collectively; $P(R|E) > P(R)$.

Exercises

The following table of data obtained from www.baseball-almanac.com shows hit information for four players. Suppose that one hit from the table is randomly selected.

Name	Single	Double	Triple	Home Run	Total Hits
Babe Ruth	1,517	506	136	714	2,873
Jackie Robinson	1,054	273	54	137	1,518
Ty Cobb	3,603	174	295	114	4,189
Hank Aaron	2,294	624	98	755	3,771
Total	8,471	1,577	583	1,720	12,351

Are "the hit being made by Hank Aaron" and "the hit being a double" independent events? Choose the best answer.

a. Yes, because P(hit by Hank Aaron|hit is a double) = P(hit by Hank Aaron)
b. No, because P(hit by Hank Aaron|hit is a double) ≠ P(hit is a double)
c. No, because P(hit is by Hank Aaron|hit is a double) ≠ P(hit by Hank Aaron)
d. Yes, because P(hit is by Hank Aaron|hit is a double) = P(hit is a double)

Exercises

United Blood Services is a blood bank that serves more than 500 hospitals in 18 states. According to their website, a person with type O blood and a negative Rh factor (Rh-) can donate blood to any person with any blood type. Their data show that 43% of people have type O blood and 15% of people have Rh- factor; 52% of people have type O or Rh- factor.

1. Find the probability that a person has both type O blood and the Rh- factor.
2. Find the probability that a person does NOT have both type O blood and the Rh- factor.

Solution

1. P(type O OR Rh-) = P(type O) + P(Rh-) – P(type O AND Rh-)

   0.52 = 0.43 + 0.15 – P(type O AND Rh-); solve to find P(type O AND Rh-) = 0.06

   6% of people have type O, Rh- blood

2. P(NOT(type O AND Rh-)) = 1 – P(type O AND Rh-) = 1 – 0.06 = 0.94

   94% of people do not have type O, Rh- blood

Exercises

At a college, 72% of courses have final exams and 46% of courses require research papers. Suppose that 32% of courses have a research paper and a final exam. Let *F* be the event that a course has a final exam. Let *R* be the event that a course requires a research paper.

1. Find the probability that a course has a final exam or a research project.
2. Find the probability that a course has NEITHER of these two requirements.

Exercises

In a box of assorted cookies, 36% contain chocolate and 12% contain nuts. Of those, 8% contain both chocolate and nuts. Sean is allergic to both chocolate and nuts.

1. Find the probability that a cookie contains chocolate or nuts (he can't eat it).
2. Find the probability that a cookie does not contain chocolate or nuts (he can eat it).

Solution
Let *C* = be the event that the cookie contains chocolate. Let *N* = the event that the cookie contains nuts.

1. $P(C \text{ OR } N) = P(C) + P(N) - P(C \text{ AND } N) = 0.36 + 0.12 - 0.08 = 0.40$
2. $P(\text{NEITHER chocolate NOR nuts}) = 1 - P(C \text{ OR } N) = 1 - 0.40 = 0.60$

Exercises

A college finds that 10% of students have taken a distance learning class and that 40% of students are part time students. Of the part time students, 20% have taken a distance learning class. Let *D* =

event that a student takes a distance learning class and E = event that a student is a part time student

1. Find $P(D \text{ AND } E)$.
2. Find $P(E|D)$.
3. Find $P(D \text{ OR } E)$.
4. Using an appropriate test, show whether D and E are independent.
5. Using an appropriate test, show whether D and E are mutually exclusive.

Section 3.4

Exercises

Use the information in the [link] to answer the next eight exercises. The table shows the political party affiliation of each of 67 members of the US Senate in June 2012, and when they are up for reelection.

Up for reelection:	Democratic Party	Republican Party	Other	Total
November 2014	20	13	0	
November 2016	10	24	0	
Total				

What is the probability that a randomly selected senator has an "Other" affiliation?

Solution
0

What is the probability that a randomly selected senator is up for reelection in November 2016?

What is the probability that a randomly selected senator is a Democrat and up for reelection in November 2016?

Solution

$$\frac{10}{67}$$

What is the probability that a randomly selected senator is a Republican or is up for reelection in November 2014?

Suppose that a member of the US Senate is randomly selected. Given that the randomly selected senator is up for reelection in November 2016, what is the probability that this senator is a Democrat?

Solution

$$\frac{10}{34}$$

Suppose that a member of the US Senate is randomly selected. What is the probability that the senator is up for reelection in November 2014, knowing that this senator is a Republican?

The events “Republican” and “Up for reelection in 2016” are _____

1. mutually exclusive.
2. independent.
3. both mutually exclusive and independent.
4. neither mutually exclusive nor independent.

Solution

d

The events “Other” and “Up for reelection in November 2016” are _____

1. mutually exclusive.
2. independent.
3. both mutually exclusive and independent.
4. neither mutually exclusive nor independent.

Exercises

[link] gives the number of suicides estimated in the U.S. for a recent year by age, race (black or white), and sex. We are interested in possible relationships between age, race, and sex. We will let suicide victims be our population.

Race and Sex	1–14	15–24	25–64	over 64	TOTALS
white, male	210	3,360	13,610		22,050
white, female	80	580	3,380		4,930
black, male	10	460	1,060		1,670
black, female	0	40	270		330
all others					
TOTALS	310	4,650	18,780		29,760

Do not include "all others" for parts f and g.

Fill in the column for the suicides for individuals over age 64.
Fill in the row for all other races.
Find the probability that a randomly selected individual was a white male.
Find the probability that a randomly selected individual was a black female.
Find the probability that a randomly selected individual was black
Find the probability that a randomly selected individual was male.
Out of the individuals over age 64, find the probability that a randomly selected individual was a black or white male.
Solution

1.

Race and Sex	1–14	15–24	25–64	over 64	TOTALS
white, male	210	3,360	13,610	4,870	22,050
white, female	80	580	3,380	890	4,930
black, male	10	460	1,060	140	1,670
black, female	0	40	270	20	330
all others				100	
TOTALS	310	4,650	18,780	6,020	29,760

2.

Race and Sex	1–14	15–24	25–64	over 64	TOTALS
white, male	210	3,360	13,610	4,870	22,050
white, female	80	580	3,380	890	4,930
black, male	10	460	1,060	140	1,670
black, female	0	40	270	20	330
all others	10	210	460	100	780
TOTALS	310	4,650	18,780	6,020	29,760

3. $\frac{22,050}{29,760}$

4. $\frac{330}{29,760}$

5. $\frac{2,000}{29,760}$

6. $\frac{23,720}{29,760}$

7. $\frac{5,010}{6,020}$

Exercises

Use the following information to answer the next two exercises. The table of data obtained from WWW.BASEBALL-ALMANAC.COM shows hit information for four well known baseball players. Suppose that one hit from the table is randomly selected.

NAME	Single	Double	Triple	Home Run	TOTAL HITS
Babe Ruth	1,517	506	136	714	2,873
Jackie Robinson	1,054	273	54	137	1,518
Ty Cobb	3,603	174	295	114	4,189
Hank Aaron	2,294	624	98	755	3,771
TOTAL	8,471	1,577	583	1,720	12,351

Find *P*(hit was made by Babe Ruth). Choose the correct letter.

a. $\frac{1518}{2873}$
b. $\frac{2873}{12351}$
c. $\frac{583}{12351}$
d. $\frac{4189}{12351}$

Find *P*(hit was made by Ty Cobb|The hit was a Home Run). Choose the correct letter.

a. $\frac{4189}{12351}$
b. $\frac{114}{1720}$
c. $\frac{1720}{4189}$
d. $\frac{114}{12351}$

Solution

b

Exercises

[link] identifies a group of children by one of four hair colors, and by type of hair.

Hair Type	Brown	Blond	Black	Red	Totals
Wavy	20		15	3	43
Straight	80	15		12	
Totals		20			215

1. Complete the table.
2. What is the probability that a randomly selected child will have wavy hair?
3. What is the probability that a randomly selected child will have either brown or blond hair?
4. What is the probability that a randomly selected child will have wavy brown hair?
5. What is the probability that a randomly selected child will have red hair, given that he or she has straight hair?
6. If B is the event of a child having brown hair, find the probability of the complement of B.
7. In words, what does the complement of B represent?

Exercises

In a previous year, the weights of the members of the **San Francisco 49ers** and the **Dallas Cowboys** were published in the SAN JOSE MERCURY NEWS. The factual data were compiled into the following table.

Shirt#	≤ 210	211–250	251–290	> 290
1–33	21	5	0	0
34–66	6	18	7	4
66–99	6	12	22	5

For the following, suppose that you randomly select one player from the 49ers or Cowboys.

Find the probability that his shirt number is from 1 to 33.
Find the probability that he weighs at most 210 pounds.
Find the probability that his shirt number is from 1 to 33 AND he weighs at most 210 pounds.
Find the probability that his shirt number is from 1 to 33 OR he weighs at most 210 pounds.
Find the probability that his shirt number is from 1 to 33 GIVEN that he weighs at most 210 pounds.
Solution

1. $\frac{26}{106}$
2. $\frac{33}{106}$
3. $\frac{21}{106}$
4. $\left(\frac{26}{106}\right) + \left(\frac{33}{106}\right) - \left(\frac{21}{106}\right) = \left(\frac{38}{106}\right)$
5. $\frac{21}{33}$

Exercises

Use the following information to answer the next seven exercises. An article in the NEW ENGLAND JOURNAL OF MEDICINE, reported about a study of smokers in California and Hawaii. In one part of the report, the self-reported ethnicity and smoking levels per day were given. Of the people smoking at most ten cigarettes per day, there were 9,886 African Americans, 2,745 Native

Hawaiians, 12,831 Latinos, 8,378 Japanese Americans, and 7,650 Whites. Of the people smoking 11 to 20 cigarettes per day, there were 6,514 African Americans, 3,062 Native Hawaiians, 4,932 Latinos, 10,680 Japanese Americans, and 9,877 Whites. Of the people smoking 21 to 30 cigarettes per day, there were 1,671 African Americans, 1,419 Native Hawaiians, 1,406 Latinos, 4,715 Japanese Americans, and 6,062 Whites. Of the people smoking at least 31 cigarettes per day, there were 759 African Americans, 788 Native Hawaiians, 800 Latinos, 2,305 Japanese Americans, and 3,970 Whites.

Complete the table using the data provided. Suppose that one person from the study is randomly selected. Find the probability that person smoked 11 to 20 cigarettes per day.

Smoking Levels by Ethnicity

Smoking Level	African American	Native Hawaiian	Latino	Japanese Americans	White	TOTALS
1–10						
11–20						
21–30						
31+						
TOTALS						

Suppose that one person from the study is randomly selected. Find the probability that person smoked 11 to 20 cigarettes per day.

Solution

$$\frac{35,065}{100,450}$$

Find the probability that the person was Latino.

In words, explain what it means to pick one person from the study who is "Japanese American **AND** smokes 21 to 30 cigarettes per day." Also, find the probability.

Solution

To pick one person from the study who is Japanese American AND smokes 21 to 30 cigarettes per day means that the person has to meet both criteria: both Japanese American and smokes 21 to 30 cigarettes. The sample space should include everyone in the study. The probability is $\frac{4,715}{100,450}$.

In words, explain what it means to pick one person from the study who is "Japanese American **OR** smokes 21 to 30 cigarettes per day." Also, find the probability.

In words, explain what it means to pick one person from the study who is "Japanese American **GIVEN** that person smokes 21 to 30 cigarettes per day." Also, find the probability.

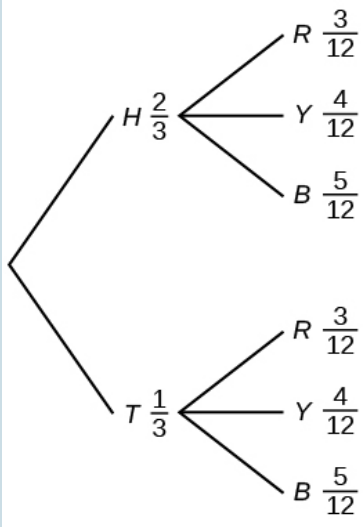
Solution
To pick one person from the study who is Japanese American given that person smokes 21-30 cigarettes per day, means that the person must fulfill both criteria and the sample space is reduced to those who smoke 21-30 cigarettes per day. The probability is $\frac{4715}{15,273}$.

Prove that smoking level/day and ethnicity are dependent events.

Section 3.5

Exercises

Use the following information to answer the next two exercises. This tree diagram shows the tossing of an unfair coin followed by drawing one bead from a cup containing three red (*R*), four yellow (*Y*) and five blue (*B*) beads. For the coin, $P(H) = \frac{2}{3}$ and $P(T) = \frac{1}{3}$ where *H* is heads and *T* is tails.



Find *P*(tossing a Head on the coin AND a Red bead)

1. $\frac{2}{3}$
2. $\frac{5}{15}$
3. $\frac{6}{36}$
4. $\frac{5}{36}$

Find *P*(Blue bead).

1. $\frac{15}{36}$
2. $\frac{10}{36}$
3. $\frac{10}{12}$
4. $\frac{6}{36}$

Solution

a

Exercises

A box of cookies contains three chocolate and seven butter cookies. Miguel randomly selects a cookie and eats it. Then he randomly selects another cookie and eats it. (How many cookies did he take?)

1. Draw the tree that represents the possibilities for the cookie selections. Write the probabilities along each branch of the tree.
2. Are the probabilities for the flavor of the SECOND cookie that Miguel selects independent of his first selection? Explain.
3. For each complete path through the tree, write the event it represents and find the probabilities.
4. Let S be the event that both cookies selected were the same flavor. Find $P(S)$.
5. Let T be the event that the cookies selected were different flavors. Find $P(T)$ by two different methods: by using the complement rule and by using the branches of the tree. Your answers should be the same with both methods.
6. Let U be the event that the second cookie selected is a butter cookie. Find $P(U)$.

Exercises

Use the following information to answer the next two exercises. Suppose that you have eight cards. Five are green and three are yellow. The cards are well shuffled.

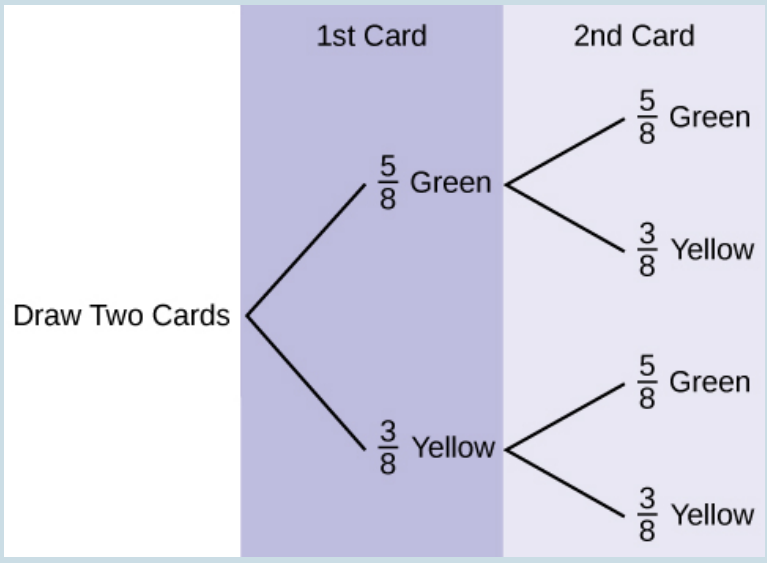
Suppose that you randomly draw two cards, one at a time, **with replacement**.

- Let G_1 = first card is green
- Let G_2 = second card is green

1. Draw a tree diagram of the situation.

2. Find $P(G_1 \text{ AND } G_2)$.
3. Find $P(\text{at least one green})$.
4. Find $P(G_2|G_1)$.
5. Are G_2 and G_1 independent events? Explain why or why not.

Solution

1. 

2. $P(GG) = \left(\frac{5}{8}\right)\left(\frac{5}{8}\right) = \frac{25}{64}$
3. $P(\text{at least one green}) = P(GG) + P(GY) + P(YG) = \frac{25}{64} + \frac{15}{64} + \frac{15}{64} = \frac{55}{64}$
4. $P(G|G) = \frac{5}{8}$
5. Yes, they are independent because the first card is placed back in the bag before the second card is drawn; the composition of cards in the bag remains the same from draw one to draw two.

Exercises

Suppose that you randomly draw two cards, one at a time, **without replacement**.

G_1 = first card is green

G_2 = second card is green

1. Draw a tree diagram of the situation.
2. Find $P(G_1 \text{ AND } G_2)$.
3. Find P(at least one green).
4. Find $P(G_2|G_1)$.
5. Are G_2 and G_1 independent events? Explain why or why not.

Exercises

Use the following information to answer the next two exercises. The percent of licensed U.S. drivers (from a recent year) that are female is 48.60. Of the females, 5.03% are age 19 and under; 81.36% are age 20–64; 13.61% are age 65 or over. Of the licensed U.S. male drivers, 5.04% are age 19 and under; 81.43% are age 20–64; 13.53% are age 65 or over.

Complete the following.

1. Construct a table or a tree diagram of the situation.
2. Find P(driver is female).
3. Find P(driver is age 65 or over|driver is female).
4. Find P(driver is age 65 or over AND female).
5. In words, explain the difference between the probabilities in part c and part d.
6. Find P(driver is age 65 or over).
7. Are being age 65 or over and being female mutually exclusive events? How do you know?

Solution

1.

	<20	20–64	>64	Totals
Female	0.0244	0.3954	0.0661	0.486
Male	0.0259	0.4186	0.0695	0.514
Totals	0.0503	0.8140	0.1356	1

2. $P(F) = 0.486$
3. $P(>64|F) = 0.1361$
4. $P(>64 \text{ and } F) = P(F)\,P(>64|F) = (0.486)(0.1361) = 0.0661$
5. $P(>64|F)$ is the percentage of female drivers who are 65 or older and $P(>64 \text{ and } F)$ is the percentage of drivers who are female and 65 or older.
6. $P(>64) = P(>64 \text{ and } F) + P(>64 \text{ and } M) = 0.1356$
7. No, being female and 65 or older are not mutually exclusive because they can occur at the same time $P(>64 \text{ and } F) = 0.0661$.

Exercises

Suppose that 10,000 U.S. licensed drivers are randomly selected.

1. How many would you expect to be male?
2. Using the table or tree diagram, construct a contingency table of gender versus age group.
3. Using the contingency table, find the probability that out of the age 20–64 group, a randomly selected driver is female.

Exercises

Approximately 86.5% of Americans commute to work by car, truck, or van. Out of that group, 84.6% drive alone and 15.4% drive in a carpool. Approximately 3.9% walk to work and approximately 5.3% take public transportation.

1. Construct a table or a tree diagram of the situation. Include a branch for all other modes of transportation to work.
2. Assuming that the walkers walk alone, what percent of all commuters travel alone to work?
3. Suppose that 1,000 workers are randomly selected. How many would you expect to travel alone to work?
4. Suppose that 1,000 workers are randomly selected. How many would you expect to drive in a carpool?

Solution

1.

	Car, Truck or Van	Walk	Public Transportation	Other	Totals
Alone	0.7318				
Not Alone	0.1332				
Totals	0.8650	0.0390	0.0530	0.0430	1

2. If we assume that all walkers are alone and that none from the other two groups travel alone (which is a big assumption) we have: $P(\text{Alone}) = 0.7318 + 0.0390 = 0.7708$.
3. Make the same assumptions as in (b) we have: $(0.7708)(1{,}000) = 771$
4. $(0.1332)(1{,}000) = 133$

Exercises

When the Euro coin was introduced in 2002, two math professors had their statistics students test whether the Belgian one Euro coin was a fair coin. They spun the coin rather than tossing it and found that out of 250 spins, 140 showed a head (event *H*) while 110 showed a tail (event *T*). On that basis, they claimed that it is not a fair coin.

1. Based on the given data, find $P(H)$ and $P(T)$.
2. Use a tree to find the probabilities of each possible outcome for the experiment of tossing the coin twice.
3. Use the tree to find the probability of obtaining exactly one head in two tosses of the coin.
4. Use the tree to find the probability of obtaining at least one head.

Exercises

Use the following information to answer the next two exercises. The following are real data from Santa Clara County, CA. As of a certain time, there had been a total of 3,059 documented cases of AIDS in the county. They were grouped into the following categories:

*** includes homosexual/bisexual IV drug users**

	Homosexual/Bisexual	IV Drug User*	Heterosexual Contact	Other	Totals
Female	0	70	136	49	____
Male	2,146	463	60	135	____
Totals	____	____	____	____	____

Suppose a person with AIDS in Santa Clara County is randomly selected.

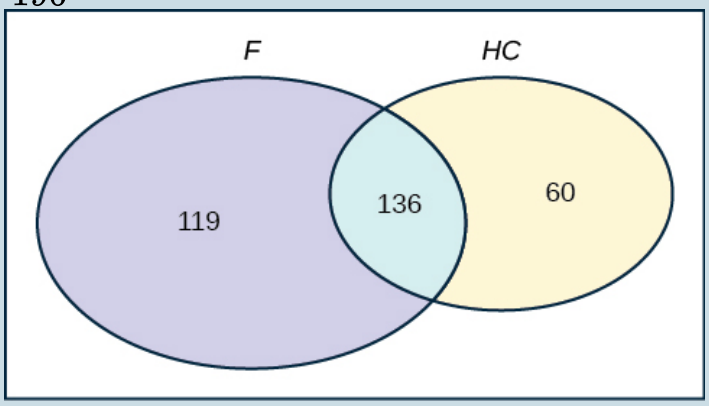
1. Find P(Person is female).
2. Find P(Person has a risk factor heterosexual contact).
3. Find P(Person is female OR has a risk factor of IV drug user).
4. Find P(Person is female AND has a risk factor of homosexual/bisexual).
5. Find P(Person is male AND has a risk factor of IV drug user).
6. Find P(Person is female GIVEN person got the disease from heterosexual contact).
7. Construct a Venn diagram. Make one group females and the other group heterosexual contact.

Solution

The completed contingency table is as follows:

*** includes homosexual/bisexual IV drug users**

	Homosexual/Bisexual	IV Drug User*	Heterosexual Contact	Other	Totals
Female	0	70	136	49	**255**
Male	2,146	463	60	135	**2,804**
Totals	**2,146**	**533**	**196**	**184**	**3,059**

1. $\frac{255}{3059}$
2. $\frac{196}{3059}$
3. $\frac{718}{3059}$
4. 0
5. $\frac{463}{3059}$
6. $\frac{136}{196}$
7. 

Exercises

Answer these questions using probability rules. Do NOT use the contingency table. Three thousand fifty-nine cases of AIDS had been reported in Santa Clara County, CA, through a certain date. Those

cases will be our population. Of those cases, 6.4% obtained the disease through heterosexual contact and 7.4% are female. Out of the females with the disease, 53.3% got the disease from heterosexual contact.

1. Find *P*(Person is female).
2. Find *P*(Person obtained the disease through heterosexual contact).
3. Find *P*(Person is female GIVEN person got the disease from heterosexual contact)
4. Construct a Venn diagram representing this situation. Make one group females and the other group heterosexual contact. Fill in all values as probabilities.

CHAPTER 4: DISCRETE RANDOM VARIABLES

INTRODUCTION TO CHAPTER 4: DISCRETE RANDOM VARIABLES



Figure 4.1. You can use probability and discrete random variables to calculate the likelihood of lightning striking the ground five times during a half-hour thunderstorm. (Credit: Leszek Leszczynski)

Let us begin with two different examples of the types of probability problems we'll be able to model using the techniques of this chapter.

Example 1: Suppose a student takes a ten-question, true-false quiz. Because the student had such a busy schedule, they could not study and guess randomly at each answer. What is the probability of the student passing the quiz with at least a 70%?

Example 2: Small companies might be interested in the number of unread emails sitting in their employee's mailboxes. If the average number of unread emails is 100 at the end of the workday, what is the probability that on a given day, the employees have more than 120 unread emails at the end of the workday?

Each of these probabilities can be found if we know the *distribution* of the corresponding *discrete random variable.* Recall that discrete data are data that you can count. A **random variable** describes the outcomes of a statistical experiment in words. The values of a random variable can vary with each repetition of an experiment.

Collaborative Exercises

Section 4.1

Collaborative Exercise

Toss a coin ten times and record the number of heads. After all members of the class have completed the experiment (tossed a coin ten times and counted the number of heads), fill in Table 1. Let X = the number of heads in ten tosses of the coin.

x	Frequency of x	Relative Frequency of x

Table 1

1. Which value(s) of x occurred most frequently?
2. If you tossed the coin 1,000 times, what values could x take on? Which value(s) of x do you think would occur most frequently?
3. What does the relative frequency column sum to?

Media Attributions

- Chapter 4 Introduction Figure © OpenStax Introductory Statistics is licensed under a CC BY (Attribution) license

4.1 PROBABILITY DISTRIBUTION FUNCTION (PDF) FOR A DISCRETE RANDOM VARIABLE

Learning Objectives

By the end of this section, you should be able to:

- Validate the properties of a discrete probability distribution function
- Write a probability distribution function for given discrete data

Random Variables

Random variables are **probability models** quantifying situations. Upper case letters such as X or Y denote a random variable. Lower case letters like x or y denote the value of a random variable. **If X is a random variable, then X is written in words, and x is given as a number.**

For example, let X = the number of heads you get when you toss three fair coins. The sample space for the toss of three fair coins is *TTT*; *THH*; *HTH*; *HHT*; *HTT*; *THT*; *TTH*; *HHH*. Then, in these outcomes, we can have no heads (when *TTT* occurs) or up to three heads (when *HHH* occurs). Therefore $x = 0, 1, 2,$ and 3.

Observe again that X is in words and x is a number. For this example, the x values are a countable number of outcomes, having four possibilities. Because you can count the possible values that X can take on and the outcomes are random, X is a *discrete random variable*.

We have seen the word discrete before associated with types of data: **discrete** means we have a countable number of outcomes. So a **discrete random variable** is a **random variable** that models a process or experiment that produces discrete data.

There are both continuous and discrete random variables. We will begin with **discrete random variables** and examine **continuous random variables** in the next chapter.

Characteristics and Notation

The **probability mass function (PMF)** of a discrete random variable tells you the probability of each outcome in the sample space. The probability that random variable X takes on value x is represented by $P(X = x)$ or just $P(x)$. The PMF is also sometimes called the probability distribution function (PDF).

The distribution of a discrete random variable is often pictured in a table, but may also be represented by a graph or formula. The two main characteristics the PMF should exhibit are:

1. Each probability is between zero and one, inclusive.
2. The sum of the probabilities is one.

Example

A child psychologist is interested in the number of times a newborn baby's crying wakes one of their parents after midnight. For a random sample of 50 families, the following information was obtained. Let X = the number of times per week a newborn baby's crying wakes one of their parents after midnight. In the sample, the largest number of times observed during a week was 5, so $x = 0, 1, 2, 3, 4, 5$.

$P(x)$ = probability that X takes on a value x.

Figure 4.2: Newborn Baby Crying

x	$P(x)$
0	$P(0) = \frac{2}{50}$
1	$P(1) = \frac{11}{50}$
2	$P(2) = \frac{23}{50}$
3	$P(3) = \frac{9}{50}$
4	$P(4) = \frac{4}{50}$
5	$P(5) = \frac{1}{50}$

Is this a valid discrete probability distribution?



An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=713#h5p-4

YOUR TURN!

A hospital researcher is interested in the number of times the average post-op patient will ring the nurse during a 12-hour shift. For a random sample of 50 patients, the following information was

obtained. Let X = the number of times a patient rings the nurse during a 12-hour shift. The largest number of patient calls was 5, so x = 0, 1, 2, 3, 4, 5.

Figure 4.3: Post-Op Patients

X	$P(x)$
0	$P(0) = \frac{4}{50}$
1	$P(1) = \frac{8}{50}$
2	$P(2) = \frac{16}{50}$
3	$P(3) = \frac{14}{50}$
4	$P(4) = \frac{6}{50}$
5	$P(5) = \frac{2}{50}$

Why is this a discrete probability distribution function (two reasons)?

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=713#h5p-197

The **cumulative distribution function (CDF)** of a discrete random variable tells you the probability of being less than or equal to a value, denoted $P(X \leq x)$.

A probability distribution function is a pattern. You try to fit a probability problem into a pattern or distribution in order to perform the necessary calculations. These distributions are tools to make solving probability problems easier. Each distribution has its own special characteristics. Learning the characteristics enables you to distinguish among the different distributions.

Example

Suppose Nancy has classes three days a week. She attends classes three days a week 80% of the time, two days 15% of the time, one day 4% of the time, and no days 1% of the time. Suppose one week is randomly selected.

a. Let X = the number of days Nancy _____ .

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=713#h5p-5

b. X takes on what values?

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=713#h5p-6

c. Suppose one week is randomly chosen. Construct a probability distribution table (called a PDF table) like the one below. The table should have two columns labeled x and $P(x)$. What does the $P(x)$ column sum to?

Figure 4.4: Blank PDF

x	$P(x)$
0	
1	
2	
3	

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=713#h5p-7

d. Construct the cumulative probability distribution function

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=713#h5p-8

Your turn!

Jeremiah has basketball practice two days a week. Ninety percent of the time, he attends both

practices. Eight percent of the time, he attends one practice. Two percent of the time, he does not attend either practice. What is X and what values does it take on?

4.2 MEASURES OF GENERAL DISCRETE RANDOM VARIABLES

Learning Objectives

By the end of this section, the student should be able to:

- Calculate and interpret expected values of general random variables
- Calculate and interpret the variance and standard deviation of general random variables

Once we know how to work with Discrete Random Variables we may be interested in some other measures such as the mean, variance, and standard deviation. The ideas here are slightly different than we have seen before within our new context of Random Variables.

The Expected Value (Mean) of a Discrete Random Variable

The Law of Large Numbers states: as the number of trials in a probability experiment increases, our results become closer to what we "expect." For instance, the student who was guessing on the true-false quiz in the chapter introduction would expect to get about half of the questions correct, since there are two options.

When evaluating the long-term results of statistical experiments, we often want to know the "average" outcome. This long-term average is known as the **mean** or **expected value** of the random variable and is denoted by the Greek letter μ, or in the context of random variables, $E[X]$. In other words, after conducting many trials of an experiment, you would expect this average value.

To find the expected value, we multiply each value of the random variable by its probability, then add the products.

Mean or Expected Value: $\mu = \sum_{x \in X} xP(x)$.

Example

A university soccer team plays soccer zero, one, or two days a week. The probability that they play zero days is 0.2, the probability that they play one day is 0.5, and the probability that they play two days is 0.3. Find the long-term average or expected value, μ, of the number of days per week the team plays soccer.

We first let the random variable X = the number of days the team plays soccer per week. Then x takes on the values 0, 1, 2. Construct a PDF table adding a column $x \cdot P(x)$. In this column, you multiply each value by its probability.

Figure 4.5: Expected Value Table. This table helps you calculate the expected value by giving you a column of products to sum.

x	$P(x)$	$x \cdot P(x)$
0	0.2	(0)(0.2) = 0
1	0.5	(1)(0.5) = 0.5
2	0.3	(2)(0.3) = 0.6

What is the expected value?

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=1441#h5p-119

Your turn!

A hospital researcher is interested in the number of times the average post-op patient will ring the nurse during a 12-hour shift. For a random sample of 50 patients, the following information was obtained. What is the expected value?

Figure 4.3 (repeat): Post-Op Patients

x	$P(x)$
0	$P(0) = \frac{4}{50}$
1	$P(1) = \frac{8}{50}$
2	$P(2) = \frac{16}{50}$
3	$P(3) = \frac{14}{50}$
4	$P(4) = \frac{6}{50}$
5	$P(5) = \frac{2}{50}$

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=1441#h5p-198

The Variance and Standard Deviation of a Discrete Random Variable

Like data, probability distributions have standard deviations. The standard deviation (σ) of a probability distribution is the sum of the squares of the difference between its outcomes and expected value weighted by the probability of the outcome.

Computing the variance, σ^2 or $V[X]$, and standard deviation, σ or $SD[X]$. of a random variable starts similar to what we have seen before for data but differs at step 4:

1. Find the mean μ.
2. Subtract the mean from each value of x to get your deviations.
3. Square each deviation.
4. Multiply each squared deviation by its probability, P(x).
5. Sum each of the products.

At this point you now have the variance then can of course take the square root of the variance to get your standard deviation. The formula looks like this:

$$\sigma = \sqrt{\sum_{x \in X} (x - \mu)^2 P(x)}$$

Example

Find the expected value of the number of times a newborn baby's crying wakes one of their parents after midnight. The expected value is the expected number of times per week a newborn baby's crying wakes one of their parents after midnight. Calculate the standard deviation of the variable as well.

Figure 4.6: Newborn Baby Crying. You expect a newborn to wake a parent after midnight 2.1 times per week, on the average.

x	$P(x)$	$x \cdot P(x)$	$(x-\mu)^2 \cdot P(x)$
0	$P(0) = \frac{2}{50}$	$(0)\left(\frac{2}{50}\right) = 0$	$(0 - 2.1)^2 \cdot 0.04 = 0.1764$
1	$P(1) = \frac{11}{50}$	$(1)\left(\frac{11}{50}\right) = \frac{11}{50}$	$(1 - 2.1)^2 \cdot 0.22 = 0.2662$
2	$P(2) = \frac{23}{50}$	$(2)\left(\frac{23}{50}\right) = \frac{46}{50}$	$(2 - 2.1)^2 \cdot 0.46 = 0.0046$
3	$P(3) = \frac{9}{50}$	$(3)\left(\frac{9}{50}\right) = \frac{27}{50}$	$(3 - 2.1)^2 \cdot 0.18 = 0.1458$
4	$P(4) = \frac{4}{50}$	$(4)\left(\frac{4}{50}\right) = \frac{16}{50}$	$(4 - 2.1)^2 \cdot 0.08 = 0.2888$
5	$P(5) = \frac{1}{50}$	$(5)\left(\frac{1}{50}\right) = \frac{5}{50}$	$(5 - 2.1)^2 \cdot 0.02 = 0.1682$

a. Add the values in the third column of the table to find the expected value of X.

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://louis.pressbooks.pub/introductorystatistics/?p=1441#h5p-120

b. Use μ to complete the table. The fourth column of this table will provide the values you need to calculate the standard deviation. For each value x, multiply the square of its deviation by its probability. (Each deviation has the format $x - \mu$.)

c. Add the values in the fourth column of the table:

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://louis.pressbooks.pub/introductorystatistics/?p=1441#h5p-121

d. The standard deviation of X is the square root of this sum.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=1441#h5p-122

e. The mean, μ, of a discrete probability function is the expected value.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=1441#h5p-123

f. The standard deviation, $\sum$, of the PDF is the square root of the variance.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=1441#h5p-124

When all outcomes in the probability distribution are equally likely, these formulas coincide with the mean and standard deviation of the set of possible outcomes.

Your turn!

Suppose you are booking a trip two days away that costs $1,000 and choose to purchase travel insurance for an additional $30. If you cannot go on your trip due to serious illness, you will receive back your $1,000 and have spent $30. Suppose the probability of getting ill in the next 48 hours is about 1.08%. Let X = the cost of booking the trip. Find the mean and standard deviation of X.

If you travel at this cost frequently and always purchase travel insurance, will you save or lose money vs. if you did not purchase travel insurance?

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=1441#h5p-202

Note on Calculations

Generally for probability distributions, we use a calculator or a computer to calculate μ and σ to reduce roundoff error. For many special cases of probability distributions, there are short-cut formulas for calculating μ, σ, and associated probabilities. We will see some of these in the future.

Your turn!

Toss a fair, six-sided die twice. Let X = the number of faces that show an even number. Construct a table like Figure 4.6 and calculate the mean μ and standard deviation σ of X.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=1441#h5p-203

4.3 BINOMIAL DISTRIBUTION

Learning Objectives

By the end of this section, you should be able to:

- Identify the components of a binomial experiment
- Use the formulas for a binomial random variable to compute mean, variance, and standard deviation

Binomial Experiments

There are three characteristics of a binomial experiment.

1. There are a fixed number of trials. Think of trials as repetitions of an experiment. The letter n denotes the number of trials.
2. There are only two possible outcomes, called "success" and "failure," for each trial. The letter p denotes the probability of a success on one trial, and q denotes the probability of a failure on one trial $p + q = 1$.
3. The n trials are independent and are repeated using identical conditions. Because the n trials are independent, the outcome of one trial does not help in predicting the outcome of another trial. Another way of saying this is that for each individual trial, the probability, p, of a success and probability, q, of a failure remain the same.

For example, randomly guessing at a true-false statistics question has only two outcomes. If a success is guessing correctly, then a failure is guessing incorrectly. Suppose Taylor consistently guesses correctly on any statistics true-false question with probability $p = 0.6$. Then, $q = 0.4$. Consistency in guessing means that for every true-false statistics question Taylor answers, the probability of success ($p = 0.6$) and the probability of failure ($q = 0.4$) remain the same, a necessary criteria for a situation to be binomial.

Any experiment that has characteristics two and three and where $n = 1$ is called a **Bernoulli Trial** (named after Jacob Bernoulli who, in the late 1600s, studied them extensively). A binomial experiment takes place when the number of successes is counted in one or more Bernoulli Trials.

The outcomes of a **binomial experiment** fit a **binomial probability distribution**. The random variable $X =$ the number of successes obtained in the n independent trials.

Notation for the Binomial

We use B to represent the binomial probability distribution, and when X fits the binomial distribution, we write $X \sim B(n, p)$. Read this as "X is a random variable with a binomial distribution with parameters n and p," which again represent the number of trials and the probability of "success."

Example

At ABC College, the withdrawal rate from an elementary physics course is 30% for any given term. This implies that, for any given term, 70% of the students stay in the class for the entire term. A "success" could be defined as an individual who withdrew. The random variable $X =$ the number of students who withdraw from the randomly selected elementary physics class.

If at the start of a particular term, 300 students are enrolled in the elementary physics course, $n = 300$ and $p = 0.3$ and $X \sim B(300, 0.3)$. The possible outcomes are $x = 0, \ldots, 300$ and the probability $P(X = x)$ is the probability that x students will withdraw during the term.

Your turn!

Approximately 70% of statistics students do their homework in time for it to be collected and graded. Each student does homework independently. In a statistics class of 50 students, what is the probability that at least 40 will do their homework on time?

a. This is a binomial problem because there is only a success or a ———, there are a fixed number of trials, and the probability of a success is 0.70 for each trial.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=717#h5p-125

b. If we are interested in the number of students who do their homework on time, then how do we define X?

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=717#h5p-126

c. What values does x take on?

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=717#h5p-127

d. What is a "failure," in words?

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=717#h5p-128

e. If $p + q = 1$, then what is q?

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=717#h5p-129

f. The words "at least" translate as what kind of inequality for the probability question $P(X____40)$.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=717#h5p-130

Your turn!

Suppose you play a game that you can only either win or lose. The probability that you win any game is 55%, and the probability that you lose is 45%. Each game you play is independent. If you play the game 20 times, write the function that describes the probability that you win 15 of the 20 times.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=717#h5p-204

Experiments That Are Not Binomial

Here are some common experiments that are not binomial:

1. Flipping a coin until you get one head. While there are only two outcomes (heads and tails) to each coin flip, and while the probability of getting a head on each flip is consistent, the number of trials will vary.
2. Most characteristics about people, such as weight, height, ethnicity, and gender. When we survey a group of people and ask for their weight, we get far more than two responses. When surveying ethnicity, as discussed in Chapter 1 an Other or Unknown category is needed to include all people in our results, specifically people who did not feel they fit into any of the ethnicity categories or declined to respond. While gender has historically been treated as a binomial outcome (male and female), not all people will feel they fit those two categories, so we may lose information and reliability by modeling gender as binomial.
3. Drawing cards without replacement. While you can specify a set number of n trials, and you can set it up as two outcomes (success = red, failure = black), the draws are not independent. If you draw a red on the first draw, it is now slightly more likely to draw black than red on the second draw.

Example

ABC College has a student advisory committee made up of ten staff members and six students. The committee wishes to choose a chairperson and a recorder by putting the names of all committee members into a box, and two names are drawn without replacement. The first name drawn determines the chairperson and the second name the recorder. Suppose the college wishes to find the probability that the chairperson and recorder are both students. Is this binomial?

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=717#h5p-206

Your turn!

Suppose 55% of people pass the state driver's exam on the first try. You survey fifty randomly chosen adults in the state.

Which of these are binomial problems?

- The number of the adults who have a driver's license.
- The number of the adults who got their driver's license on the first try.
- The number of times each adult has taken the driver's exam.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=717#h5p-205

Binomial Probability Function

Once we have decided we can use the binomial for a given situation, we can use the binomial probability function to find the probability of a specific number of successes, $P(X = x)$. The binomial **PMF** is made up of two parts:

First, we need to find out how many different ways we can get x successes in n trials. To do this we can use the "Choose" function, also called the binomial coefficient, written as:

$$\text{nCx} = = \binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Note: The ! mark is the factorial operator.

The next part gives us the probability of a single one of those ways to get x successes in n trials. We can do this by using our independent multiplication rule. We multiply the probability of success (p) raised to the number of successes (x) by the probability of failure ($q = 1 - p$) raised to the number of failures ($n - x$).

$$p^x q^{(n-x)}$$

Since we know each of these ways are equally likely and how many ways are possible we can now put the two

pieces together. We multiply the probability of one way by how many we have to give us our overall probability of x successes in n trials.

$$P(X = x) = \frac{n!}{x!(n-x)!} p^x q^{(n-x)}$$

Unfortunately the binomial does not have a nice form of **CDF**, but it is simply the sum of PDFs up until that point. Consider the following example to demonstrate this point.

Example

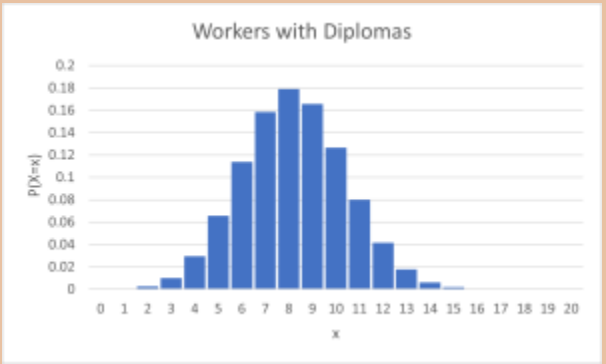
It has been stated that about 41% of adult workers have a high school diploma but do not pursue any further education. Twenty adult workers are randomly selected.

Let $X =$ the number of workers who have a high school diploma but do not pursue any further education.

Then X takes on the values $0, 1, 2, \ldots, 20$ where $n = 20, p = 0.41$, and $q = 1 - 0.41 = 0.59$. Finally, $X \sim B(20, 0.41)$.

The *y*-axis contains the probability of x, where $X =$ the number of workers who have only a high school diploma.

The graph of $X \sim B(20, 0.41)$ is as follows:



Find the probability that:

(a) Exactly 12 of them have a high school diploma

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=717#h5p-131

(b) At most 12 of them have a high school diploma but do not pursue any further education. How many adult workers do you expect to have a high school diploma but do not pursue any further education?

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=717#h5p-132

Note

To compute binomial probabilities on a graphing calculator, go into 2nd DISTR. The syntax for the instructions are as follows:

To calculate $P(X = x)$: binompdf(*n*, *p*, x).
If x is left out, the result is the binomial probability table.

To calculate $P(X \leq x)$: binomcdf(*n*, *p*, x).
If x is left out, the result is the cumulative binomial probability table.

For the example problem: After you are in 2nd DISTR, arrow down to binomcdf(. Press ENTER. Enter 20,0.41,12). The result is $P(X \leq 12) = 0.9738$.

If you wanted to instead find $P(X > 12)$, use 1 – binomcdf(20,0.41,12).

In Excel, both binomial probabilities are computed using BINOM.DIST(x, n, p, True/False), where False computes $P(X = x)$ and is equivalent to binompdf, and True computes $P(X \leq x)$ and is equivalent to binomcdf.

Your turn!

About 32% of students participate in a community volunteer program outside of school. If 30 students are selected at random, find:

(a) The probability that exactly 14 of them participate in a community volunteer program outside of school. First try plugging in to the binomial formula by hand, then check yourself with technology.

(b) The probability that exactly 14 of them participate in a community volunteer program outside of school. Rely on technology for this probability.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=717#h5p-216

Expected Value and Standard Deviation

The mean, μ, and variance, σ^2, for the binomial probability distribution are $\mu = np$ and $\sigma^2 = npq$. The standard deviation is then $\sigma = \sqrt{npq}$.

Example

In the 2013 *Jerry's Artarama* art supplies catalog, there are 560 pages. Eight of the pages feature signature artists. Suppose we randomly sample 100 pages. Let X = the number of pages that feature signature artists.

1. What values does x take on?
2. What is the probability distribution? Find the following probabilities:
    1. the probability that two pages feature signature artists
    2. the probability that at most six pages feature signature artists
    3. the probability that more than three pages feature signature artists.
3. Using the formulas, calculate the (i) mean and (ii) standard deviation.

Solution

1. $x = 0, 1, 2, 3, \ldots, n$, where n=100
2. $X \sim B\left(100, \frac{8}{560}\right)$
    1. $P(x = 2) = \text{binompdf}\left(100, \frac{8}{560}, 2\right) = 0.2466$
    2. $P(x \leq 6) = \text{binomcdf}\left(100, \frac{8}{560}, 6\right) = 0.9994$
    3. $P(x > 3) = 1 - P(x \leq 3) = 1 - \text{binomcdf}\left(100, \frac{8}{560}, 3\right) = 1 - 0.9443 = 0.0557$
    1. Mean = $np = (100)\left(\frac{8}{560}\right) = \frac{800}{560} \approx 1.4286$
    2. Standard Deviation = $\sqrt{npq} = \sqrt{(100)\left(\frac{8}{560}\right)\left(\frac{552}{560}\right)} \approx 1.1867$

Your turn!

According to a Gallup poll, 60% of American adults prefer saving over spending. Let X = the number of American adults out of a random sample of 50 who prefer saving to spending.

1. What is the probability distribution for X?
2. Use your calculator to find the following probabilities:
    1. the probability that 25 adults in the sample prefer saving over spending
    2. the probability that at most 20 adults prefer saving
    3. the probability that more than 30 adults prefer saving
3. Using the formulas, calculate the (i) mean and (ii) standard deviation of X.



An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=717#h5p-217

Your Turn!

During the 2013 regular NBA season, DeAndre Jordan of the Los Angeles Clippers had the highest field goal completion rate in the league. DeAndre scored with 61.3% of his shots. Suppose you choose a random sample of 80 shots made by DeAndre during the 2013 season. Let X = the number of shots that scored points.

1. What is the probability distribution for X?
2. Using the formulas, calculate the (i) mean and (ii) standard deviation of X.

3. Use your calculator to find the probability that DeAndre scored with 60 of these shots.
4. Find the probability that DeAndre scored with more than 50 of these shots.



An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=717#h5p-218

References

"Access to electricity (% of population)," The World Bank, 2013. Available online at http://data.worldbank.org/indicator/EG.ELC.ACCS.ZS?order=wbapi_data_value_2009%20wbapi_data_value%20wbapi_data_value-first&sort=asc (accessed May 15, 2015).

"Distance Education." Wikipedia. Available online at http://en.wikipedia.org/wiki/Distance_education (accessed May 15, 2013).

"NBA Statistics – 2013," ESPN NBA, 2013. Available online at http://espn.go.com/nba/statistics/_/seasontype/2 (accessed May 15, 2013).

Newport, Frank. "Americans Still Enjoy Saving Rather than Spending: Few demographic differences seen in these views other than by income," GALLUP® Economy, 2013. Available online at http://www.gallup.com/poll/162368/americans-enjoy-saving-rather-spending.aspx (accessed May 15, 2013).

Pryor, John H., Linda DeAngelo, Laura Palucki Blake, Sylvia Hurtado, Serge Tran. *The American Freshman: National Norms Fall 2011*. Los Angeles: Cooperative Institutional Research Program at the Higher Education Research Institute at UCLA, 2011. Also available online at http://heri.ucla.edu/PDFs/pubs/TFS/Norms/Monographs/TheAmericanFreshman2011.pdf (accessed May 15, 2013).

"The World FactBook," Central Intelligence Agency. Available online at https://www.cia.gov/library/publications/the-world-factbook/geos/af.html (accessed May 15, 2013).

"What are the key statistics about pancreatic cancer?" American Cancer Society, 2013. Available online at http://www.cancer.org/cancer/pancreaticcancer/detailedguide/pancreatic-cancer-key-statistics (accessed May 15, 2013).

Media Attributions

- Picture1

4.4 GEOMETRIC DISTRIBUTION

Learning Objectives

By the end of this section, you should be able to:

- Identify the components of a geometric experiment
- Use the formulas for a geometric random variable to compute mean, variance, and standard deviation

There are three main characteristics of a **geometric experiment**.

1. There are one or more Bernoulli trials with all failures except the last one, which is a success. In other words, you keep repeating what you are doing until the first success. Then you stop. For example, you throw a dart at a bullseye until you hit the bullseye. The first time you hit the bullseye is a "success" so you stop throwing the dart. It might take six tries until you hit the bullseye. You can think of the trials as failure, failure, failure, failure, failure, success, STOP.
2. In theory, the number of trials could go on forever. There must be at least one trial.
3. The probability p of a success and the probability q of a failure is the same for each trial with $p + q = 1$, so $q = 1 - p$.

The geometric random variable $X =$ the number of independent trials until the first success. The probability that X equals a particular value x, that is, the probability we first succeed on a particular trial, is

$$P(X = x) = q^{(x-1)}p.$$

Example

Suppose that when rolling a fair die, you want to know the probability of getting the first three on the fifth roll.

The experiment here is to roll a fair die until you get a three, then stop. This is a geometric experiment because the probability of rolling a three when you throw one fair die is $\frac{1}{6}$ no matter how many times you roll the die, and you either get a 3 (success) or get one of the other 5 numbers (failure).

So for the event to occur, on rolls one through four, you do not get a face with a three. The probability for each of the rolls is $q = \frac{5}{6}$, the probability of a failure. The probability of getting a three on the fifth roll is $\left(\frac{5}{6}\right)\left(\frac{5}{6}\right)\left(\frac{5}{6}\right)\left(\frac{5}{6}\right)\left(\frac{1}{6}\right)$ = 0.0804

Example

A safety engineer feels that 35% of all industrial accidents in her plant are caused by failure of employees to follow instructions. She decides to look at the accident reports (selected randomly and replaced in the pile after reading) **until** she finds one that shows an accident caused by failure of employees to follow instructions. On average, how many reports would the safety engineer **expect** to look at until she finds a report showing an accident caused by employee failure to follow instructions? What is the probability that the safety engineer will have to examine at least three reports until she finds a report showing an accident caused by employee failure to follow instructions?

Let X = the number of accidents the safety engineer must examine **until** she finds a report showing an accident caused by employee failure to follow instructions. X takes on the values 1, 2, 3, ... The first question asks you to find the **expected value** or the mean. The second question asks you to find $P(x \geq 3)$. ("At least" translates to a "greater than or equal to" symbol.)

Your Turn!

You play a game of chance that you can either win or lose (there are no other possibilities) **until** you lose. Your probability of losing is $p = 0.57$. What is the probability that it takes five games until you lose?

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=719#h5p-219

Your Turn!

Suppose that you are looking for a student at your college who lives within five miles of you. You know that 55% of the 25,000 students do live within five miles of you. You randomly contact students from the college **until** one says he or she lives within five miles of you. What is the probability that you need to contact four people?

This is a geometric problem because you may have a number of failures before you have the one success you desire. Also, the probability of success stays the same each time you ask a student if he or she lives within five miles of you. There is no definite number of trials (number of times you ask a student).

a. Let X = the number of _____ you must ask _____ one says yes.

Solution
a. Let X = the number of **students** you must ask **until** one says yes.

b. What values does X take on?

Solution
b. 1, 2, 3, ..., (total number of students)

c. What are *p* and *q*?

Solution
c. $p = 0.55$; $q = 0.45$

d. The probability question is $P(_____)$.

Solution
d. $P(x = 4)$

Expected Value and Standard Deviation

The notation for the Geometric Probability Distribution function is G, and we denote "X is a random variable with a **geometric distribution**" as $X \sim G(p)$.

The parameter is p, the probability of a success for each trial.

The formula for the mean is $\mu = \frac{1}{p}$.

The formula for the variance is $\sigma^2 = \left(\frac{1}{p}\right)\left(\frac{1}{p} - 1\right)$.

The standard deviation is $\sigma = \sqrt{\left(\frac{1}{p}\right)\left(\frac{1}{p} - 1\right)}$.

Note

To compute geometric probabilities on a graphing calculator, go into 2^{nd} DISTR. The syntax for the instructions are as follows:

To calculate $P(X = x)$: geometpdf(*p*, x).

To calculate $P(X \leq x)$: geometcdf(*p*, x).

If you wanted to instead find $P(X > x)$, use 1 – geometcdf(p, x).

In Excel, both geometric probabilities are computed using NEGBINOM.DIST(x-1, 1, p, True/False), where False computes $P(X = x)$ and is equivalent to geometpdf, and True computes $P(X \leq x)$ and is equivalent to geometcdf.

Example

Suppose that at a particular computer component manufacturing facility that the probability of a defective computer component is 0.02. For quality assurance, components are randomly selected. Find the probability that the first defect is found in the seventh component tested. How many components do you expect to test until one is found to be defective?

Let *X* = the number of computer components tested until the first defect is found; *X* takes on the values 1, 2, 3, ... where *p* = 0.02. We write $X \sim G(0.02$.

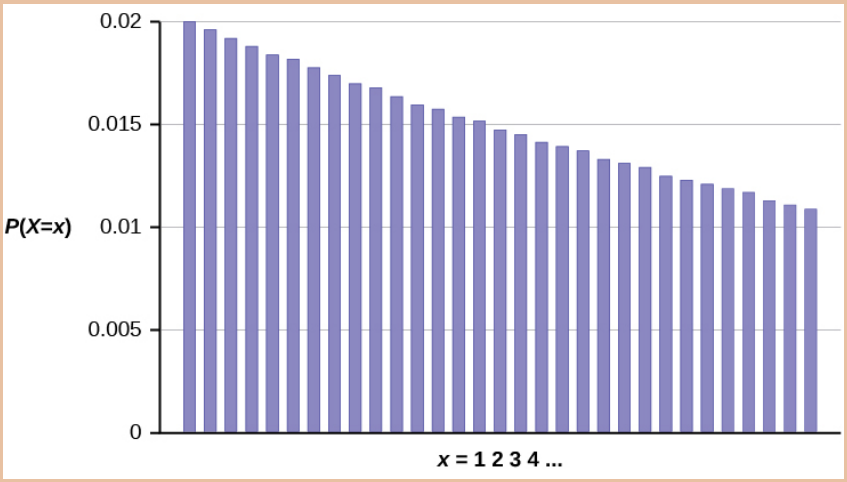
We are looking for $P(X = 7)$. To find the probability that *x* = 7 using a calculator,

- Enter 2nd, DISTR
- Scroll down and select geometpdf(
- Press ENTER
- Enter 0.02, 7); press ENTER to see the result: $P(X = 7) = 0.0177$.

To find the probability $P(X \leq 7)$, follow the same instructions EXCEPT select E:geometcdf(as the distribution function.

The probability that the seventh component is the first defect is 0.0177.

The graph of $X \sim G(0.02)$ is below.



The *y*-axis contains the probability of x where $X =$ the number of computer components tested.

The number of components that you would expect to test until you find the first defective one is the mean, $\mu = \frac{1}{p} = \frac{1}{0.02} = 50$.

The variance is $\sigma^2 = \left(\frac{1}{p}\right)\left(\frac{1}{p} - 1\right) = \left(\frac{1}{0.02}\right)\left(\frac{1}{0.02} - 1\right) = 2{,}450$.

The standard deviation is $\sigma = \sqrt{\left(\frac{1}{p}\right)\left(\frac{1}{p} - 1\right)} = \sqrt{\left(\frac{1}{0.02}\right)\left(\frac{1}{0.02} - 1\right)} = 49.5$.

Your Turn!

The probability of a defective steel rod is 0.01. Steel rods are selected at random. Find the probability that the first defect occurs on the ninth steel rod. Use a calculator or computer to find the answer.

An interactive H5P element has been excluded from this version of the text. You can view it online

here:
https://louis.pressbooks.pub/introductorystatistics/?p=719#h5p-220

Your Turn!

The literacy rate for a nation measures the proportion of people age 15 and over who can read and write. The literacy rate for women in Afghanistan is 12%. Let $X =$ the number of Afghani women you ask until one says that she is literate.

1. What is the probability distribution of *X*?
2. What is the probability that you ask five women before one says she is literate?
3. What is the probability that you must ask ten women?
4. Find the (i) mean and (ii) standard deviation of *X*.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=719#h5p-221

References

"Millennials: A Portrait of Generation Next," PewResearchCenter. Available online at http://www.pewsocialtrends.org/files/2010/10/millennials-confident-connected-open-to-change.pdf (accessed May 15, 2013).

"Millennials: Confident. Connected. Open to Change." Executive Summary by PewResearch Social &

Demographic Trends, 2013. Available online at http://www.pewsocialtrends.org/2010/02/24/millennials-confident-connected-open-to-change/ (accessed May 15, 2013).

"Prevalence of HIV, total (% of populations ages 15-49)," The World Bank, 2013. Available online at http://data.worldbank.org/indicator/SH.DYN.AIDS.ZS?order=wbapi_data_value_2011+wbapi_data_value+wbapi_data_value-last&sort=desc (accessed May 15, 2013).

Pryor, John H., Linda DeAngelo, Laura Palucki Blake, Sylvia Hurtado, Serge Tran. *The American Freshman: National Norms Fall 2011.* Los Angeles: Cooperative Institutional Research Program at the Higher Education Research Institute at UCLA, 2011. Also available online at http://heri.ucla.edu/PDFs/pubs/TFS/Norms/Monographs/TheAmericanFreshman2011.pdf (accessed May 15, 2013).

"Summary of the National Risk and Vulnerability Assessment 2007/8: A profile of Afghanistan," The European Union and ICON-Institute. Available online at http://ec.europa.eu/europeaid/where/asia/documents/afgh_brochure_summary_en.pdf (accessed May 15, 2013).

"The World FactBook," Central Intelligence Agency. Available online at https://www.cia.gov/library/publications/the-world-factbook/geos/af.html (accessed May 15, 2013).

"UNICEF reports on Female Literacy Centers in Afghanistan established to teach women and girls basic resading [*sic*] and writing skills," UNICEF Television. Video available online at http://www.unicefusa.org/assets/video/afghan-female-literacy-centers.html (accessed May 15, 2013).

4.5 HYPERGEOMETRIC DISTRIBUTION

Learning Objectives

By the end of this section, you should be able to:

- Identify the components of a hypergeometric experiment
- Use the formulas for a hypergeometric random variable to compute the mean

There are five characteristics of a hypergeometric experiment.

1. You take samples from **two** groups.
2. You are concerned with a group of interest, called the first group.
3. You sample **without replacement** from the combined groups. For example, you want to choose a softball team from a combined group of 11 men and 13 women. The team consists of ten players.
4. Each pick is **not** independent, since sampling is without replacement. In the softball example, the probability of picking a woman first is $\frac{13}{24}$. The probability of picking a man second is $\frac{11}{23}$ if a woman was picked first. It is $\frac{10}{23}$ if a man was picked first. The probability of the second pick depends on what happened in the first pick.
5. You are **not** dealing with Bernoulli Trials.

The outcomes of a **hypergeometric experiment** fit a **hypergeometric probability distribution**. The random variable X = the number of items from the group of interest.

Example

A candy dish contains 100 jelly beans and 80 gumdrops. Fifty candies are picked at random. What is the probability that 35 of the 50 are gumdrops?

This experiment is hypergeometric: the two groups are jelly beans and gumdrops. Since the probability question asks for the probability of picking gumdrops, the group of interest (first group) is gumdrops and size 80. The size of the second group, jelly beans, is 100. The size of the sample is 50 (jelly beans or gumdrops).

Let X = the number of gumdrops in the sample of 50. X takes on the values $x = 0, 1, 2, ..., 50$. What is the probability statement written mathematically?

Solution
$P(x = 35)$

Your Turn!

A bag contains letter tiles. Forty-four of the tiles are vowels, and 56 are consonants. Seven tiles are picked at random. You want to know the probability that four of the seven tiles are vowels. What is the group of interest, the size of the group of interest, and the size of the sample?



An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=721#h5p-222

Your Turn!

You are president of an on-campus special events organization. You need a committee of seven students to plan a special birthday party for the president of the college. Your organization consists of 18 women and 15 men. If the members of the committee are randomly selected, what is the probability that your committee has more than four men?

a. Are you choosing with or without replacement?

Solution
a. without

b. What is the group of interest?

Solution
b. the men

c. How many are in the group of interest?

Solution
c. 15 men

d. How many are in the other group?

Solution
d. 18 women

e. Let X = _____ on the committee. What values does X take on?

Solution
e. Let X = **the number of men** on the committee. $x = 0, 1, 2, ..., 7$.

f. The probability question is P(_____).

Solution
f. $P(x > 4)$

Notation and Mean for the Hypergeometric

The notation for the Geometric Probability Distribution function is H, and we denote "X is a random variable with a **hypergeometric probability distribution** as $X \sim H(r, b, n)$.

The parameters are r, b, and n; r = the size of the group of interest (first group), b = the size of the second group, n = the size of the chosen sample. Most of the formulas for the Hypergeometric distribution are beyond the scope of this book; the exception is the mean.

$$\mu = \frac{nr}{r+b}$$

Note

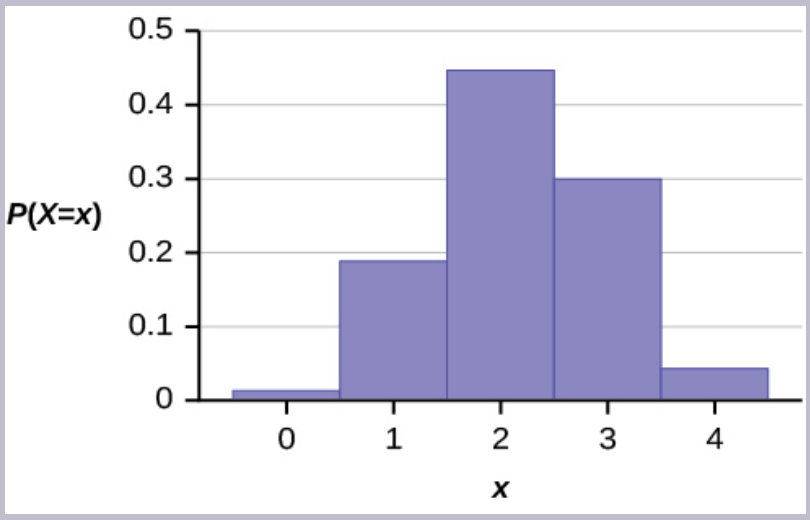
Currently, the TI-83/84 calculators do not have a built-in hypergeometric distribution button. You can calculate probabilities in Excel using HYPGEO.DIST(x, n, r, r+b, True/False), where False computes $P(X = x)$, and True computes $P(X \leq x)$.

Your Turn!

A school site committee is to be chosen randomly from five men and six women. (1) If the committee consists of four members chosen randomly, what is the probability that two of them are women? (2) How many women do you expect to be on the committee?

Let X = the number of women on the committee of four. The women are the group of interest (first group). The random variable X takes on the values 0, 1, 2, 3, 4, where $r = 6$, $b = 5$, and $n = 4$, and $X \sim H(6, 5, 4)$.

The graph of $X \sim H(6, 5, 4)$ is below.



The *y*-axis contains the probability of *X*, where *X* = the number of men on the committee.

Solution

(1) Using Excel's HypGeom.Dist,

$P(X = 2) = \text{HYPGEOM.DIST}(2, 4, 6, 11, \text{FALSE}) \approx 0.4545$. The probability that there are exactly two women on the committee is about 0.45.

(2) The formula for the mean is $\mu = \dfrac{nr}{r + b} = \dfrac{(4)\,(6)}{6 + 5} = 2.18$. You would expect $\mu = 2.18$ (about two) women on the committee.

Your Turn!

An intramural basketball team is to be chosen randomly from 15 upperclassmen and 12 underclassmen. The team has ten slots. You want to know the probability that eight of the players will be upperclassmen. What is the group of interest and the sample? Find the probability using a computer.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

https://louis.pressbooks.pub/introductorystatistics/?p=721#h5p-224

4.6 POISSON DISTRIBUTION

Learning Objectives

By the end of this section, the student should be able to:

- Identify the components of a Poisson experiment
- Use the formulas for a Poisson random variable to compute the mean, variance, and standard deviation

There are two main characteristics of a Poisson experiment.

1. The **Poisson probability distribution** gives the probability of a number of events occurring in a **fixed interval** of time or space if these events happen with a known average rate and independently of the time since the last event. For example, a book editor might be interested in the number of words spelled incorrectly in a particular book. It might be that, on the average, there are five words spelled incorrectly in 100 pages. The interval is the 100 pages.
2. The number of events occurring has a constant average and the events are independent of the "time" since the last event. Continuing the book example, the average number of misspelled words is 5, and finding a word that is misspelled does not influence the page on which the next misspelled word lies. Additionally, two events shouldn't occur simultaneously.

The Poisson distribution is frequently used for experiments where there are a large number of possible but rare events. For instance, the Poisson distribution may be used to approximate the binomial if the probability of success is "small" (such as 0.01) and the number of trials is "large" (such as 1,000). You will verify the relationship in the homework exercises.

The random variable X = the number of occurrences in the interval of interest. Because the average per interval is constant, we can find the average for any size interval by scaling to the size of the interval.

Example

The average number of loaves of bread put on a shelf in a bakery in a half-hour period is 12. Of interest is the number of loaves of bread put on the shelf in five minutes, selected randomly. The time interval of interest is five minutes. What is the probability that the number of loaves put on the shelf in five minutes is three?

Let $X =$ the number of loaves of bread put on the shelf in five minutes. If the average number of loaves put on the shelf in 30 minutes (half-hour) is 12, then the average number of loaves put on the shelf in five minutes is $\left(\frac{5}{30}\right)(12) = 2$ loaves of bread.

The probability question asks you to find $P(X = 3)$.

Your Turn!

You notice that a news reporter says "uh," on average, two times per broadcast. What is the probability that the news reporter says "uh" more than two times per broadcast.

This is a Poisson problem because you are interested in knowing the number of times the news reporter says "uh" during a broadcast.

a. What is the interval of interest?

Solution
a. one broadcast

b. What is the average number of times the news reporter says "uh" during one broadcast?

Solution
b. 2

c. Let X = _____. What values does X take on?

Solution

c. Let X = the number of times the news reporter says "uh" during one broadcast.

x = 0, 1, 2, 3, ...

d. The probability question is P(_____).

Solution

d. $P(x > 2)$

Your Turn!

An emergency room at a particular hospital gets an average of five patients per hour. A doctor wants to know the probability that the ER gets more than five patients per hour. Give the reasons why this would be a Poisson distribution.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=723#h5p-225

Your Turn!

The average number of fish caught in an hour is eight. Of interest is the number of fish caught in 15 minutes. What is the average number of fish caught in 15 minutes?

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=723#h5p-226

Notation and Mean

The notation for the Poisson Probability Distribution function is P, and we denote "X is a random variable with a **Poisson probability distribution** as $X \sim P(\lambda)$. The parameter is λ, which is also the mean μ. So, we may also write $X \sim P(\mu)$.

The mean is given as part of a Poisson experiment, $\mu = \lambda$.

The formula for the variance is $\sigma^2 = \lambda = \mu$, that is, the average is also the variance of a Poisson random variable.

The formula for the standard deviation is $\sigma = \sqrt{\lambda}$.

Note

To compute Poisson probabilities on a graphing calculator, go into 2nd DISTR. The syntax for the instructions are as follows. The TI calculators use λ (lambda) for the mean.

To calculate $P(X = x)$: poissonpdf(λ, x).

To calculate $P(X \leq x)$: poissoncdf(λ, x).

If you wanted to instead find $P(X > x)$, use 1 – poissoncdf(λ, x).

In Excel, both Poisson probabilities are computed using POISSON.DIST(x, λ, True/False), where False computes $P(X = x)$ and is equivalent to poissonpdf, and True computes $P(X \leq x)$ and is equivalent to poissoncdf.

Example

Leah's school email receives an average of six emails between 8 a.m. and 10 a.m. each day. What is the probability that Leah receives more than one email in the next 15 minutes?

Let $X =$ the number of emails Leah receives in 15 minutes. (The interval of interest is 15 minutes or $\frac{1}{4}$ hour.)

x = 0, 1, 2, 3, ...

If Leah receives, on the average, six emails in two hours, and there are eight 15 minute intervals in two hours, then Leah receives

$\left(\frac{1}{8}\right)(6) = 0.75$ emails in 15 minutes, on average. So, $\lambda = \mu = 0.75$ for this problem, and $X \sim P(0.75)$.

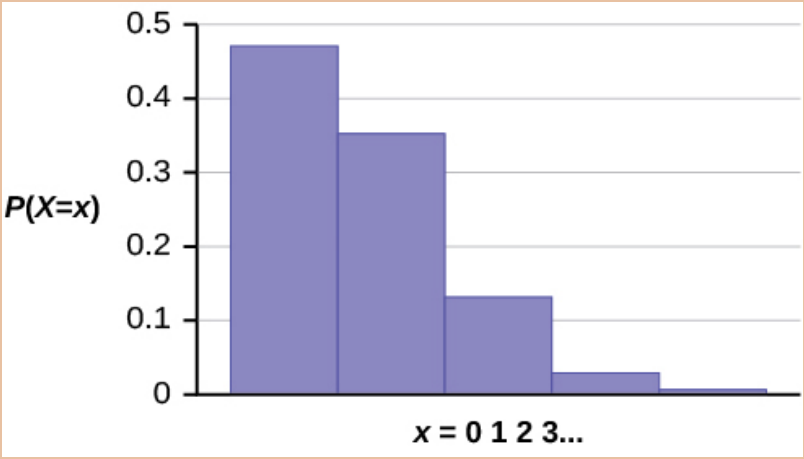
Find $P(X > 1)$: (calculator or computer)

- Press 1 – and then press 2nd DISTR.
- Arrow down to poissoncdf(. Press ENTER.
- Enter .75,1).
- The result is $P(X > 1) = 0.1734.$

The probability that Leah receives more than one email in the next 15 minutes is about 0.1734:

$P(X > 1) = 1 - poissoncdf(0.75, 1) = 0.1734.$

The graph of $X \sim P(0.75)$ is:



The *y*-axis contains the probability of *x* where *X* = the number of emails in 15 minutes.

Your Turn!

Atlanta's Hartsfield-Jackson International Airport is the busiest airport in the world. On average there are 2,500 arrivals and departures each day.

1. How many airplanes arrive and depart the airport per hour?
2. What is the probability that there are exactly 100 arrivals and departures in one hour?
3. What is the probability that there are at most 100 arrivals and departures in one hour?

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=723#h5p-227

Example

According to Baydin, an email management company, an email user gets, on average, 147 emails per day. Let $X =$ the number of emails an email user receives per day. The discrete random variable X takes on the values $x = 0, 1, 2, \ldots$. The random variable X has a Poisson distribution: $X \sim P(147)$. The mean is 147 emails.

1. What is the probability that an email user receives exactly 160 emails per day?
2. What is the probability that an email user receives at most 160 emails per day?
3. What is the standard deviation?

Solution

1. $P(x = 160)$ = poissonpdf(147, 160) ≈ 0.0180
2. $P(x \leq 160)$ = poissoncdf(147, 160) ≈ 0.8666
3. Standard Deviation = $\sigma = \sqrt{\mu} = \sqrt{147} \approx 12.1244$

Your Turn!

According to a recent poll by the Pew Internet Project, girls between the ages of 14 and 17 send an average of 187 text messages each day. Let *X* = the number of texts that a girl aged 14 to 17 sends per day. The discrete random variable *X* takes on the values *x* = 0, 1, 2 .... The random variable *X* has a Poisson distribution: *X* ~ *P*(187). The mean is 187 text messages.

1. What is the probability that a teen girl sends exactly 175 texts per day?
2. What is the probability that a teen girl sends at most 150 texts per day?
3. What is the standard deviation?

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=723#h5p-228

The Poisson distribution can be used to approximate probabilities for a binomial distribution. The Poisson approximation to a binomial distribution was commonly used in the days before technology made both values very easy to calculate.

Let n represent the number of binomial trials and let p represent the probability of a success for each trial. If n is large enough and p is small enough then the Poisson approximates the binomial very well. In general, n is considered "large enough" if it is greater than or equal to 20. The probability p from the binomial distribution should be less than or equal to 0.05. When the Poisson is used to approximate the binomial, we use the binomial mean $\mu = np$.

Example

On May 13, 2013, starting at 4:30 PM, the probability of low seismic activity for the next 48 hours in Alaska was reported as about 1.02%. Use this information for the next 200 days to find the probability that there will be low seismic activity in ten of the next 200 days. Use both the binomial and Poisson distributions to calculate the probabilities. Are they similar?



An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=723#h5p-229

Your Turn!

On May 13, 2013, starting at 4:30 PM, the probability of moderate seismic activity for the next 48 hours in the Kuril Islands off the coast of Japan was reported at about 1.43%. Use this information for the next 100 days to find the probability that there will be low seismic activity in five of the next

100 days. Use both the binomial and Poisson distributions to calculate the probabilities. Are they close?

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=723#h5p-230

References

"ATL Fact Sheet," Department of Aviation at the Hartsfield-Jackson Atlanta International Airport, 2013. Available online at http://www.atlanta-airport.com/Airport/ATL/ATL_FactSheet.aspx (accessed May 15, 2013).

Center for Disease Control and Prevention. "Teen Drivers: Fact Sheet," Injury Prevention & Control: Motor Vehicle Safety, October 2, 2012. Available online at http://www.cdc.gov/Motorvehiclesafety/Teen_Drivers/teendrivers_factsheet.html (accessed May 15, 2013).

"Children and Childrearing," Ministry of Health, Labour, and Welfare. Available online at http://www.mhlw.go.jp/english/policy/children/children-childrearing/index.html (accessed May 15, 2013).

"Eating Disorder Statistics," South Carolina Department of Mental Health, 2006. Available online at http://www.state.sc.us/dmh/anorexia/statistics.htm (accessed May 15, 2013).

"Giving Birth in Manila: The maternity ward at the Dr Jose Fabella Memorial Hospital in Manila, the busiest in the Philippines, where there is an average of 60 births a day," theguardian, 2013. Available online at http://www.theguardian.com/world/gallery/2011/jun/08/philippines-health#/?picture=375471900&index=2 (accessed May 15, 2013).

"How Americans Use Text Messaging," Pew Internet, 2013. Available online at http://pewinternet.org/Reports/2011/Cell-Phone-Texting-2011/Main-Report.aspx (accessed May 15, 2013).

Lenhart, Amanda. "Teens, Smartphones & Testing: Texting volume is up while the frequency of voice calling is down. About one in four teens say they own smartphones," Pew Internet, 2012. Available online at http://www.pewinternet.org/~/media/Files/Reports/2012/PIP_Teens_Smartphones_and_Texting.pdf (accessed May 15, 2013).

"One born every minute: the maternity unit where mothers are THREE to a bed," MailOnline. Available online at http://www.dailymail.co.uk/news/article-2001422/Busiest-maternity-ward-planet-averages-60-babies-day-mothers-bed.html (accessed May 15, 2013).

Vanderkam, Laura. "Stop Checking Your Email, Now." CNNMoney, 2013. Available online at http://management.fortune.cnn.com/2012/10/08/stop-checking-your-email-now/ (accessed May 15, 2013).

"World Earthquakes: Live Earthquake News and Highlights," World Earthquakes, 2012. http://www.world-earthquakes.com/index.php?option=ethq_prediction (accessed May 15, 2013).

CHAPTER 4 REVIEW

Section 4.1 Review

The characteristics of a probability distribution function (PDF) for a discrete random variable are as follows:

1. Each probability is between zero and one, inclusive (*inclusive* means to include zero and one).
2. The sum of the probabilities is one.

Section 4.2 Review

The expected value, or mean, of a discrete random variable predicts the long-term results of a statistical experiment that has been repeated many times. The **standard deviation of a probability distribution** is used to measure the variability of possible outcomes.

Formula Review

Mean or Expected Value: $\mu = \sum_{x \in X} xP(x)$

Standard Deviation: $\sigma = \sqrt{\sum_{x \in X} (x - \mu)^2 P(x)}$

Section 4.3 Review

A statistical experiment can be classified as a binomial experiment if the following conditions are met:

1. There are a fixed number of trials n.
2. There are only two possible outcomes, called "success" and "failure" for each trial. The letter p denotes the probability of a success on one trial and q denotes the probability of a failure on one trial.
3. The n trials are independent and are repeated using identical conditions.

The outcomes of a binomial experiment fit a binomial probability distribution. The random variable

$X =$ the number of successes obtained in the n independent trials. The mean of X can be calculated using the formula $\mu = np$, and the standard deviation is given by the formula $\sigma = \sqrt{npq}$.

Formula Review

$X \sim B(n, p)$ means that the discrete random variable X has a binomial probability distribution with n trials and probability of success p.

$X =$ the number of successes in n independent trials

n = \text{the number of independent trials}

X takes on the values $x = 0, 1, 2, 3, \ldots, n$

$p =$ the probability of a success for any trial

q = \text{the probability of a failure for any trial}

$p + q = 1$

$q = 1 - p$

The mean of X is $\mu = np$. The standard deviation of X is $\sigma = \sqrt{npq}$.

Section 4.4 Review

There are three characteristics of a geometric experiment:

1. There are one or more Bernoulli trials with all failures except the last one, which is a success.
2. In theory, the number of trials could go on forever. There must be at least one trial.
3. The probability, p, of a success and the probability, q, of a failure are the same for each trial.

In a geometric experiment, define the discrete random variable X as the number of independent trials until the first success. We say that X has a geometric distribution and write $X \sim G(p)$.

Formula Review

$X \sim G(p)$ means that the discrete random variable X has a geometric probability distribution with probability of success in a single trial p.

X = \text{the number of independent trials until the first success}

X takes on the values $x = 1, 2, 3, \ldots$

$p =$ the probability of a success for any trial

$q =$ the probability of a failure for any trial

$p + q = 1$

$q = 1 - p$

The mean is $\mu = \frac{1}{p}$.

The standard deviation is $\sigma = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1}{p}\left(\frac{1}{p} - 1\right)}$.

Section 4.5 Review

A hypergeometric experiment is a statistical experiment with the following properties:

1. You take samples from two groups.
2. You are concerned with a group of interest, called the first group.
3. You sample without replacement from the combined groups.
4. Each pick is not independent, since sampling is without replacement.
5. You are not dealing with Bernoulli Trials.

The outcomes of a hypergeometric experiment fit a hypergeometric probability distribution. The random variable $X =$ the number of items from the group of interest. The distribution of X is denoted $X \sim H(r, b, n)$, where $r =$ the size of the group of interest (first group), $b =$ the size of the second group, and $n =$ the size of the chosen sample. It follows that $n \leq r + b$. The mean of X is $\mu = \frac{nr}{r+b}$ and the standard deviation is $\sigma = \sqrt{\frac{rbn(r+b-n)}{(r+b)^2(r+b-1)}}$.

.

Formula Review

$X \sim H(r, b, n)$ means that the discrete random variable X has a hypergeometric probability distribution with $r =$ the size of the group of interest (first group), $b =$ the size of the second group, and $n =$ the size of the chosen sample.

$X =$ the number of items from the group of interest that are in the chosen sample

X may take on the values $x = 0, 1, 2, \ldots$ up to the size of the group of interest. (The minimum value for X may be larger than zero in some instances.)

$n \leq r + b$

$$\mu = \frac{nr}{r+b}$$

$$\sigma = \sqrt{\frac{rbn(r+b-n)}{(r+b)^2(r+b-1)}}$$

Section 4.6 Review

A Poisson probability distribution of a discrete random variable gives the probability of a number of events occurring in a fixed interval of time or space, if these events happen at a known average rate and independently of the time since the last event. The Poisson distribution may be used to approximate the binomial, if the probability of success is "small" (less than or equal to 0.05) and the number of trials is "large" (greater than or equal to 20).

Formula Review

$X \sim P(\mu)$ means that X has a Poisson probability distribution where $X =$ the number of occurrences in the interval of interest.

X takes on the values $x = 0, 1, 2, 3, \ldots$

The mean μ is typically given.

The variance is $\sigma^2 = \mu$, and the standard deviation is $\sigma = \sqrt{\mu}$.

When $P(\mu)$ is used to approximate a binomial distribution, $\mu = np$ where n represents the number of independent trials and p represents the probability of success in a single trial.

CHAPTER 4 PRACTICE

Section 4.1

Exercises

Use the following information to answer the next four exercises. A baker is deciding how many batches of muffins to make to sell in his bakery. He wants to make enough to sell every one and no fewer. Through observation, the baker has established a probability distribution.

x	$P(x)$
1	0.15
2	0.35
3	0.40
4	0.10

Define the random variable *X*.

Solution
Let X = the number of batches that the baker will sell.

What is the probability the baker will sell more than one batch? $P(x > 1)$ = _____

Solution
$$0.35 + 0.40 + 0.10 = 0.85$$

What is the probability the baker will sell exactly one batch? $P(x = 1)$ = _____

Solution
0.15

On average, how many batches should the baker make?

Solution
$$1(0.15) + 2(0.35) + 3(0.40) + 4(0.10) = 0.15 + 0.70 + 1.20 + 0.40 = 2.45$$

Exercises

Use the following information to answer the next three exercises. Ellen has music practice three days a week. She practices for all of the three days 85% of the time, two days 8% of the time, one day 4% of the time, and no days 3% of the time. One week is selected at random.

Define the random variable *X*.

Solution
Let X = the number of days Ellen attends practice per week.

Construct a probability distribution table for the data.

Solution

x	$P(x)$
0	0.03
1	0.04
2	0.08
3	0.85

We know that for a probability distribution function to be discrete, it must have two characteristics. One is that the sum of the probabilities is one. What is the other characteristic?

Solution

Each probability is between zero and one, inclusive.

Exercises

Use the following information to answer the next five exercises. Javier volunteers in community events each month. He does not do more than five events in a month. He attends exactly five events 35% of the time, four events 25% of the time, three events 20% of the time, two events 10% of the time, one event 5% of the time, and no events 5% of the time.

Define the random variable X.

Solution

Let X = the number of events Javier volunteers for each month.

What values does x take on?

Solution

0, 1, 2, 3, 4, 5

Construct a PDF table.

Solution

x	$P(x)$
0	0.05
1	0.05
2	0.10
3	0.20
4	0.25
5	0.35

Find the probability that Javier volunteers for less than three events each month.

Solution

$$P(x < 3) = 0.05 + 0.05 + 0.10 = 0.20$$

Find the probability that Javier volunteers for at least one event each month.

Solution

$$P(x > 0)1 - 0.05 = 0.95$$

Section 4.2

Exercises

Complete the expected value table.

x	$P(x)$	$x*P(x)$
0	0.2	
1	0.2	
2	0.4	
3	0.2	

Solution

x	$P(x)$	$x*P(x)$
2	0.1	$2(0.1) = 0.2$
4	0.3	$4(0.3) = 1.2$
6	0.4	$6(0.4) = 2.4$
8	0.2	$8(0.2) = 1.6$

Find the expected value from the expected value table.

Solution

$$0.2 + 1.2 + 2.4 + 1.6 = 5.4$$

Find the standard deviation.

x	$P(x)$	$x*P(x)$	$(x - \mu)^2 P(x)$
2	0.1	$2(0.1) = 0.2$	$(2–5.4)^2(0.1) = 1.156$
4	0.3	$4(0.3) = 1.2$	$(4–5.4)^2(0.3) = 0.588$
6	0.4	$6(0.4) = 2.4$	$(6–5.4)^2(0.4) = 0.144$
8	0.2	$8(0.2) = 1.6$	$(8–5.4)^2(0.2) = 1.352$

Solution

$$\sigma = 1.156 + 0.588 + 0.144 + 1.352 = 3.24$$

$$\sqrt{3.24} = 1.8$$

Exercises

Identify the mistake in the probability distribution table.

x	$P(x)$	$x*P(x)$
1	0.15	0.15
2	0.25	0.50
3	0.30	0.90
4	0.20	0.80
5	0.15	0.75

Solution

The values of $P(x)$ do not sum to one.

Exercises

Identify the mistake in the probability distribution table.

x	$P(x)$	$x*P(x)$
1	0.15	0.15
2	0.25	0.40
3	0.25	0.65
4	0.20	0.85
5	0.15	1

Solution

The values of $x * P(x)$ are not correct.

Exercises

Use the following information to answer the next five exercises: A physics professor wants to know what percent of physics majors will spend the next several years doing post-graduate research. He has the following probability distribution.

x	$P(x)$	$x*P(x)$
1	0.35	
2	0.20	
3	0.15	
4		
5	0.10	
6	0.05	

Define the random variable X.

Solution

Let X = the number of years a physics major will spend doing post-graduate research.

Define $P(x)$, or the probability of x.

Solution
Let P(x) = the probability that a physics major will do post-graduate research for x years.

Find the probability that a physics major will do post-graduate research for four years. $P(x = 4)$ = _____

Solution
$1 - 0.35 - 0.20 - 0.15 - 0.10 - 0.05 = 0.15$

Find the probability that a physics major will do post-graduate research for at most three years. $P(x \leq 3)$ = _____

Solution
$0.35 + 0.20 + 0.15 = 0.70$

On average, how many years would you expect a physics major to spend doing post-graduate research?

Solution
$1(0.35) + 2(0.20) + 3(0.15) + 4(0.15) + 5(0.10) + 6(0.05) = 0.35 + 0.40 + 0.45 + 0.60 + 0.50 + 0.30 = 2.6$ years

Exercises

Use the following information to answer the next seven exercises: A ballet instructor is interested in knowing what percent of each year's class will continue on to the next, so that she can plan what classes to offer. Over the years, she has established the following probability distribution.

- Let X = the number of years a student will study ballet with the teacher.

- Let $P(x)$ = the probability that a student will study ballet x years.

Complete [link] using the data provided.

x	$P(x)$	$x*P(x)$
1	0.10	
2	0.05	
3	0.10	
4		
5	0.30	
6	0.20	
7	0.10	

In words, define the random variable X.

Solution
X is the number of years a student studies ballet with the teacher.

$P(x = 4)$ = _____

Solution
$$1-0.10-0.05-0.10-0.30-0.20-0.10 = 0.15$$

$P(x < 4)$ = _____

Solution
$$0.10 + 0.05 + 0.10 = 0.25$$

On average, how many years would you expect a child to study ballet with this teacher?

Solution
$$1(0.10) + 2(0.05) + 3(0.10) + 4(0.15) + 5(0.30) + 6(0.20) + 7(0.10) = 4.5$$
years

What does the column "$P(x)$" sum to and why?

Solution
The sum of the probabilities sum to one because it is a probability distribution.

What does the column "$x*P(x)$" sum to and why?

Solution
The sum of x*P(x) = 4.5; it is the mean of the distribution.

Exercises

You are playing a game by drawing a card from a standard deck and replacing it. If the card is a face card, you win $30. If it is not a face card, you pay $2. There are 12 face cards in a deck of 52 cards. What is the expected value of playing the game?

Solution

$$-2\left(\frac{40}{52}\right)+30\left(\frac{12}{52}\right)=-1.54+6.92=5.38$$

You are playing a game by drawing a card from a standard deck and replacing it. If the card is a face card, you win $30. If it is not a face card, you pay $2. There are 12 face cards in a deck of 52 cards. Should you play the game?

Solution
Yes, because there is a positive expected value, and the more you play, the more likely you are to get closer to the expected value.

Section 4.3

Exercises

Use the following information to answer the next eight exercises: The Higher Education Research Institute at UCLA collected data from 203,967 incoming first-time, full-time freshmen from 270 four-year colleges and universities in the U.S. 71.3% of those students replied that, yes, they believe that same-sex couples should have the right to legal marital status. Suppose that you randomly pick eight first-time, full-time freshmen from the survey. You are interested in the number that believes that same sex-couples should have the right to legal marital status.

In words, define the random variable X.

Solution
X = the number that reply "yes"

$X \sim$ _____(_____,_____)

Solution
B(8,0.713)

What values does the random variable X take on?

Solution
0, 1, 2, 3, 4, 5, 6, 7, 8

Construct the probability distribution function (PDF).

On average (μ), how many would you expect to answer yes?

Solution
5.7

What is the standard deviation (σ)?

Solution
1.2795

What is the probability that at most five of the freshmen reply "yes"?

Solution
0.4151

What is the probability that at least two of the freshmen reply "yes"?

Solution
0.9990

Section 4.4

Exercises

Use the following information to answer the next six exercises. The Higher Education Research Institute at UCLA collected data from 203,967 incoming first-time, full-time freshmen from 270 four-year colleges and universities in the U.S. 71.3% of those students replied that, yes, they believe that same-sex couples should have the right to legal marital status. Suppose that you randomly select a freshman from the study until you find one who replies "yes." You are interested in the number of freshmen you must ask.

In words, define the random variable X.

Solution
X = the number of freshmen selected from the study until one replied "yes" that same-sex couples should have the right to legal marital status.

X ~ _____(_____,_____)

Solution
G(0.713)

What values does the random variable X take on?

Solution
1,2,...

Construct the probability distribution function (PDF). Stop at $x = 6$.

x	$P(x)$
1	
2	
3	
4	
5	
6	

On average (μ), how many freshmen would you expect to have to ask until you found one who replies "yes?"

Solution
1.4

What is the probability that you will need to ask fewer than three freshmen?

Solution
0.9176

Section 4.5

Exercises

Use the following information to answer the next five exercises: Suppose that a group of statistics students is divided into two groups: business majors and non-business majors. There are 16 business majors in the group and seven non-business majors in the group. A random sample of nine students is taken. We are interested in the number of business majors in the sample.

In words, define the random variable X.

Solution
X = the number of business majors in the sample.

X ~ _____(_____,_____)

What values does X take on?

Solution
2, 3, 4, 5, 6, 7, 8, 9

Find the standard deviation.

On average (μ), how many would you expect to be business majors?

Solution
6.26

Section 4.6

Exercises

Use the following information to answer the next six exercises: On average, a clothing store gets 120 customers per day.

Assume the event occurs independently on any given day. Define the random variable *X*.

Solution
X ~ P(120)

What values does *X* take on?

Solution
0, 1, 2, 3, 4, ...

What is the probability of getting 150 customers in one day?

Solution
0.0010

What is the probability of getting 35 customers in the first four hours? Assume the store is open 12 hours each day.

Solution
0.0485

What is the probability that the store will have more than 12 customers in the first hour?

Solution
0.2084

What is the probability that the store will have fewer than 12 customers in the first two hours?

Solution
0.0214

Exercises

Which type of distribution can the Poisson model be used to approximate? When would you do this?

Solution
The Poisson distribution can approximate a binomial distribution, which you would do if the probability of success is small and the number of trials is large.

Exercises

Use the following information to answer the next six exercises: On average, eight teens in the U.S. die from motor vehicle injuries per day. As a result, states across the country are debating raising the driving age.

Assume the event occurs independently on any given day. In words, define the random variable X.

Solution
X = the number of U.S. teens who die from motor vehicle injuries per day.

$X \sim$ _____(_____,_____)

Solution
P(8)

What values does X take on?

Solution
0, 1, 2, 3, 4, ...

For the given values of the random variable X, fill in the corresponding probabilities.

Is it likely that there will be no teens killed from motor vehicle injuries on any given day in the U.S.? Justify your answer numerically.

Solution

No

Is it likely that there will be more than 20 teens killed from motor vehicle injuries on any given day in the U.S.? Justify your answer numerically.

Solution

No

CHAPTER 4 HOMEWORK

Section 4.1 Homework

Exercises

Suppose that the PDF for the number of years it takes to earn a Bachelor of Science (B.S.) degree is given in [link].

x	$P(x)$
3	0.05
4	0.40
5	0.30
6	0.15
7	0.10

1. In words, define the random variable *X*.
2. What does it mean that the values zero, one, and two are not included for *x* in the PDF?

Section 4.2 Homework

Exercises

A theater group holds a fund-raiser. It sells 100 raffle tickets for $5 a piece. Suppose you purchase four tickets. The prize is two passes to a Broadway show, worth a total of $150.

1. What are you interested in here?
2. In words, define the random variable *X*.
3. List the values that *X* may take on.
4. Construct a PDF.
5. If this fund-raiser is repeated often and you always purchase four tickets, what would be your expected average winnings per raffle?

Solution

1. I am interested in the average profit or loss.
2. Let X = the return from the raffle
3. Win($150) or Lose ($0)
4.

Net Gain	Probability
$150	$\frac{4}{100}$
$0	$\frac{0}{100}$

5. $150(\frac{4}{100}) + 0(\frac{99}{100}) - 20 = -\$14$

Exercises

A game involves selecting a card from a regular 52-card deck and tossing a coin. The coin is a fair coin and is equally likely to land on heads or tails.

- If the card is a face card, and the coin lands on Heads, you win $6
- If the card is a face card, and the coin lands on Tails, you win $2
- If the card is not a face card, you lose $2, no matter what the coin shows.

1. Find the expected value for this game (expected net gain or loss).
2. Explain what your calculations indicate about your long-term average profits and losses on this game.
3. Should you play this game to win money?

Solution
The variable of interest is *X*, or the gain or loss, in dollars.

The face cards jack, queen, and king. There are $(3)(4) = 12$ face cards and $52–12 = 40$ cards that are not face cards.

We first need to construct the probability distribution for *X*. We use the card and coin events to determine the probability for each outcome, but we use the monetary value of *X* to determine the expected value.

Card Event	*X* net gain/loss	*P*(*X*)
Face Card and Heads	6	$\left(\frac{12}{52}\right)\left(\frac{1}{2}\right)=\left(\frac{6}{52}\right)$
Face Card and Tails	2	$\left(\frac{12}{52}\right)\left(\frac{1}{2}\right)=\left(\frac{6}{52}\right)$
(Not Face Card) and (H or T)	–2	$\left(\frac{40}{52}\right)(1)=\left(\frac{40}{52}\right)$

1. $\text{Expected value}=(6)\left(\frac{6}{52}\right)+(2)\left(\frac{6}{52}\right)+(-2)\left(\frac{40}{52}\right)=-\frac{32}{52}=-\$0.62$
2. If you play this game repeatedly, over a long string of games, you would expect to lose 62 cents per game, on average.
3. You should not play this game to win money because the expected value indicates an expected average loss.

Exercises

You buy a lottery ticket to a lottery that costs $10 per ticket. There are only 100 tickets available to be sold in this lottery. In this lottery there are one $500 prize, two $100 prizes, and four $25 prizes. Find your expected gain or loss.

Solution

Start by writing the probability distribution. X is net gain or loss = prize (if any) less $10 cost of ticket

Expected Value

=(490)(\frac{1}{100})+(90)(\frac{2}{100})+(15)(\frac{4}{100})+(−10)(\frac{93}{100})=−$2.

There is an expected loss of $2 per ticket, on average.

Complete the PDF and answer the questions.

x	$P(x)$	$xP(x)$
0	0.3	
1	0.2	
2		
3	0.4	

1. Find the probability that $x = 2$.
2. Find the expected value.

Solution

1. 0.1
2. 1.6

Exercises

Suppose that you are offered the following "deal." You roll a die. If you roll a six, you win $10. If you roll a four or five, you win $5. If you roll a one, two, or three, you pay $6.

1. What are you ultimately interested in here (the value of the roll or the money you win)?
2. In words, define the Random Variable X.
3. List the values that X may take on.
4. Construct a PDF.
5. Over the long run of playing this game, what are your expected average winnings per game?
6. Based on numerical values, should you take the deal? Explain your decision in complete

sentences.

Solution

1. the money won
2. X = the amount of money won or lost
3. $5, –$6, $10
4.

x	P(x)
5	$\frac{2}{6}$
-6	$\frac{3}{6}$
10	$\frac{1}{6}$

5. Expected Value = (10)(\frac{1}{6})+ (5)(\frac{2}{6})– (6)(\frac{3}{6})= 0.33
6. Yes, the expected value is 33 cents

Exercises

A venture capitalist, willing to invest $1,000,000, has three investments to choose from. The first investment, a software company, has a 10% chance of returning $5,000,000 profit, a 30% chance of returning $1,000,000 profit, and a 60% chance of losing the million dollars. The second company, a hardware company, has a 20% chance of returning $3,000,000 profit, a 40% chance of returning $1,000,000 profit, and a 40% chance of losing the million dollars. The third company, a biotech firm, has a 10% chance of returning $6,000,000 profit, a 70% of no profit or loss, and a 20% chance of losing the million dollars.

1. Construct a PDF for each investment.
2. Find the expected value for each investment.
3. Which is the safest investment? Why do you think so?
4. Which is the riskiest investment? Why do you think so?
5. Which investment has the highest expected return, on average?

Solution

1.

Software Company	
x	$P(x)$
5,000,000	0.10
1,000,000	0.30
−1,000,000	0.60

Hardware Company	
x	$P(x)$
3,000,000	0.20
1,000,000	0.40
−1,000,00	0.40

Biotech Firm	
x	$P(x)$
6,00,000	0.10
0	0.70
−1,000,000	0.20

2. $200,000; $600,000; $400,000
3. third investment because it has the lowest probability of loss
4. first investment because it has the highest probability of loss
5. second investment

S

Exercises

Suppose that 20,000 married adults in the United States were randomly surveyed as to the number of children they have. The results are compiled and are used as theoretical probabilities. Let X = the number of children married people have.

x	$P(x)$	$xP(x)$
0	0.10	
1	0.20	
2	0.30	
3		
4	0.10	
5	0.05	
6 (or more)	0.05	

1. Find the probability that a married adult has three children.
2. In words, what does the expected value in this example represent?
3. Find the expected value.
4. Is it more likely that a married adult will have two to three children or four to six children? How do you know?

Solution

1. 0.2
2. The average number of children married adults have.
3. 2.35
4. two of three children

Exercises

Suppose that the PDF for the number of years it takes to earn a Bachelor of Science (B.S.) degree is given as in [link].

x	$P(x)$
3	0.05
4	0.40
5	0.30
6	0.15
7	0.10

On average, how many years do you expect it to take for an individual to earn a B.S.?

Solution
4.85 years

Exercises

People visiting video rental stores often rent more than one DVD at a time. The probability distribution for DVD rentals per customer at Video To Go is given in the following table. There is a five-video limit per customer at this store, so nobody ever rents more than five DVDs.

x	$P(x)$
0	0.03
1	0.50
2	0.24
3	
4	0.07
5	0.04

1. Describe the random variable X in words.

2. Find the probability that a customer rents three DVDs.
3. Find the probability that a customer rents at least four DVDs.
4. Find the probability that a customer rents at most two DVDs.
   Another shop, Entertainment Headquarters, rents DVDs and video games. The probability distribution for DVD rentals per customer at this shop is given as follows. They also have a five-DVD limit per customer.

x	$P(x)$
0	0.35
1	0.25
2	0.20
3	0.10
4	0.05
5	0.05

5. At which store is the expected number of DVDs rented per customer higher?
6. If Video to Go estimates that they will have 300 customers next week, how many DVDs do they expect to rent next week? Answer in sentence form.
7. If Video to Go expects 300 customers next week, and Entertainment HQ projects that they will have 420 customers, for which store is the expected number of DVD rentals for next week higher? Explain.
8. Which of the two video stores experiences more variation in the number of DVD rentals per customer? How do you know that?

Solution

1. X = the number of video rentals per customer
2. 0.12
3. 0.11
4. 0.77
5. Video To Go (1.82 expected value vs. 1.4 for Entertainment Headquarters)
6. The expected number of videos rented to 300 Video To Go customers is 546.
7. The expected number of videos rented to 420 Entertainment Headquarters customers is 588. Entertainment Headquarters will rent more videos.
8. The standard deviation for the number of videos rented at Video To Go is 1.1609. The standard deviation for the number of videos rented at Entertainment Headquarters is 1.4293.

Entertainment Headquarters has more variation.

Exercises

A "friend" offers you the following "deal." For a $10 fee, you may pick an envelope from a box containing 100 seemingly identical envelopes. However, each envelope contains a coupon for a free gift.

- Ten of the coupons are for a free gift worth $6.
- Eighty of the coupons are for a free gift worth $8.
- Six of the coupons are for a free gift worth $12.
- Four of the coupons are for a free gift worth $40.

Choose the best answer below:

a. Based upon the financial gain or loss over the long run, should you play the game?
b. Yes, I expect to come out ahead in money.
c. No, I expect to come out behind in money.
d. It doesn't matter. I expect to break even.

Solution
c

Exercises

Florida State University has 14 statistics classes scheduled for its Summer 2013 term. One class has

space available for 30 students, eight classes have space for 60 students, one class has space for 70 students, and four classes have space for 100 students.

1. What is the average class size assuming each class is filled to capacity?
2. Space is available for 980 students. Suppose that each class is filled to capacity and select a statistics student at random. Let the random variable *X* equal the size of the student's class. Define the PDF for *X*.
3. Find the mean of *X*.
4. Find the standard deviation of *X*.

Solution

1. The average class size is $\frac{30+8(60)+70+4(100)}{14} = 70$.
2. $P(x=30)=\frac{1}{14}; P(x=60)=\frac{8}{14}; P(x=70)=\frac{1}{14}; P(x=100)=\frac{4}{14}$
3. $\text{Mean of } X = \frac{30}{14}+\frac{480}{14}+\frac{70}{14}+\frac{400}{14}=\frac{980}{14}=70$
4. $\text{Standard Deviation of } X = 114.2857+57.1429+0+257.1429=20.702$

Exercises

In a lottery, there are 250 prizes of $5, 50 prizes of $25, and ten prizes of $100. Assuming that 10,000 tickets are to be issued and sold, what is a fair price to charge to break even?

Solution

Let *X* = the amount of money to be won on a ticket. The following table shows the PDF for *X*.

x	$P(x)$
0	0.969
5	$\frac{250}{10{,}000} = 0.025$
25	$\frac{50}{10{,}000} = 0.005$
100	$\frac{10}{10{,}000} = 0.001$

Calculate the expected value of X.

$$0(0.969) + 5(0.025) + 25(0.005) + 100(0.001) = 0.35$$

A fair price for a ticket is \$0.35. Any price over \$0.35 will enable the lottery to raise money.

Section 4.3 Homework

Exercises

According to a recent article the average number of babies born with significant hearing loss (deafness) is approximately two per 1,000 babies in a healthy baby nursery. The number climbs to an average of 30 per 1,000 babies in an intensive care nursery.

Suppose that 1,000 babies from healthy baby nurseries were randomly surveyed. Find the probability that exactly two babies were born deaf.

Solution
0.2709

Exercises

Use the following information to answer the next four exercises. Recently, a nurse commented that when a patient calls the medical advice line claiming to have the flu, the chance that he or she truly has the flu (and not just a nasty cold) is only about 4%. Of the next 25 patients calling in claiming to have the flu, we are interested in how many actually have the flu.

Define the random variable and list its possible values.

Solution
X = the number of patients calling in claiming to have the flu, who actually have the flu.

X = 0, 1, 2, ...25

State the distribution of X.

Solution
B(25,0.04)

Find the probability that at least four of the 25 patients actually have the flu.

Solution
0.0165

On average, for every 25 patients calling in, how many do you expect to have the flu?

Solution
one

Exercises

People visiting video rental stores often rent more than one DVD at a time. The probability

distribution for DVD rentals per customer at Video To Go is given [link]. There is a five-video limit per customer at this store, so nobody ever rents more than five DVDs.

x	$P(x)$
0	0.03
1	0.50
2	0.24
3	
4	0.07
5	0.04

1. Describe the random variable X in words.
2. Find the probability that a customer rents three DVDs.
3. Find the probability that a customer rents at least four DVDs.
4. Find the probability that a customer rents at most two DVDs.

Solution

1. X = the number of DVDs a Video to Go customer rents
2. 0.12
3. 0.11
4. 0.77

Exercises

A school newspaper reporter decides to randomly survey 12 students to see if they will attend Tet (Vietnamese New Year) festivities this year. Based on past years, she knows that 18% of students attend Tet festivities. We are interested in the number of students who will attend the festivities.

1. In words, define the random variable X.
2. List the values that X may take on.

3. Give the distribution of *X*. *X* ~ _____(_____,_____)
4. How many of the 12 students do we expect to attend the festivities?
5. Find the probability that at most four students will attend.
6. Find the probability that more than two students will attend.

Solution

1. X = the number of students who will attend Tet.
2. 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
3. X ~ B(12,0.18)
4. 2.16
5. 0.9511
6. 0.3702

Exercises

Use the following information to answer the next two exercises. The probability that the San Jose Sharks will win any given game is 0.3694 based on a 13-year win history of 382 wins out of 1,034 games played (as of a certain date). An upcoming monthly schedule contains 12 games.

The expected number of wins for that upcoming month is:

a. 1.67
b. 12
c. $\frac{382}{1043}$
d. 4.43

Solution
d. 4.43

Let X = the number of games won in that upcoming month. What is the probability that the San Jose Sharks win six games in that upcoming month?

a. 0.1476
b. 0.2336
c. 0.7664
d. 0.8903

Solution
a

What is the probability that the San Jose Sharks win at least five games in that upcoming month

a. 0.3694
b. 0.5266
c. 0.4734
d. 0.2305

Solution
c

Exercises

A student takes a ten-question true-false quiz but did not study and randomly guesses each answer. Find the probability that the student passes the quiz with a grade of at least 70% of the questions correct.

Solution
X = number of questions answered correctly
X ~ B(10, 0.5)
We are interested in AT LEAST 70% of ten questions correct. 70% of ten is seven. We want to find the probability that X is greater than or equal to seven. The event “at least seven” is the complement of “less than or equal to six”.
Using your calculator’s distribution menu: 1 – binomcdf(10, .5, 6) gives 0.171875

The probability of getting at least 70% of the ten questions correct when randomly guessing is approximately 0.172.

Exercises

A student takes a 32-question multiple-choice exam, but did not study and randomly guesses each answer. Each question has three possible choices for the answer. Find the probability that the student guesses **more than** 75% of the questions correctly.

Solution

- X = number of questions answered correctly
- $X \sim B\left(32, \frac{1}{3}\right)$
- We are interested in MORE THAN 75% of 32 questions correct. 75% of 32 is 24. We want to find $P(x > 24)$. The event "more than 24" is the complement of "less than or equal to 24."
- Using your calculator's distribution menu: 1 – binomcdf $\left(32, \frac{1}{3}, 24\right)$
- $P(x > 24) = 0$
- The probability of getting more than 75% of the 32 questions correct when randomly guessing is very small and practically zero.

Exercises

Six different colored dice are rolled. Of interest is the number of dice that show a one.

1. In words, define the random variable *X*.
2. List the values that *X* may take on.
3. Give the distribution of *X*. *X* ~ _____(_____,_____)
4. On average, how many dice would you expect to show a one?
5. Find the probability that all six dice show a one.
6. Is it more likely that three or that four dice will show a one? Use numbers to justify your answer numerically.

Solution

1. X = the number of dice that show a one
2. 0, 1, 2, 3, 4, 5, 6
3. X ~ B(6,16)
4. 1
5. 0.00002
6. three dice

Exercises

More than 96 percent of the very largest colleges and universities (more than 15,000 total enrollments) have some online offerings. Suppose you randomly pick 13 such institutions. We are interested in the number that offer distance learning courses.

1. In words, define the random variable *X*.
2. List the values that *X* may take on.
3. Give the distribution of *X*. *X* ~ _____(_____,_____)
4. On average, how many schools would you expect to offer such courses?
5. Find the probability that at most ten offer such courses.
6. Is it more likely that 12 or that 13 will offer such courses? Use numbers to justify your answer numerically and answer in a complete sentence.

Solution

1. X = the number of colleges and universities that offer online offerings.
2. 0, 1, 2, ..., 13
3. $X \sim B(13, 0.96)$
4. 12.48
5. 0.0135
6. $P(x = 12) = 0.3186$ $P(x = 13) = 0.5882$ More likely to get 13.

Exercises

Suppose that about 85% of graduating students attend their graduation. A group of 22 graduating students is randomly chosen.

1. In words, define the random variable X.
2. List the values that X may take on.
3. Give the distribution of X. $X \sim$ _____(_____,_____)
4. How many are expected to attend their graduation?
5. Find the probability that 17 or 18 attend.
6. Based on numerical values, would you be surprised if all 22 attended graduation? Justify your answer numerically.

Solution

1. X = the number of students who attend their graduation
2. 0, 1, 2, ..., 22
3. X ~ B(22, 0.85)
4. 18.7
5. 0.3249
6. P(x = 22) = 0.0280 (less than 3%) which is unusual

Exercises

At The Fencing Center, 60% of the fencers use the foil as their main weapon. We randomly survey 25 fencers at The Fencing Center. We are interested in the number of fencers who do **not** use the foil as their main weapon.

1. In words, define the random variable X.
2. List the values that X may take on.
3. Give the distribution of X. $X \sim$ _____(_____,_____)
4. How many are expected to **not** to use the foil as their main weapon?
5. Find the probability that six do **not** use the foil as their main weapon.
6. Based on numerical values, would you be surprised if all 25 did **not** use foil as their main weapon? Justify your answer numerically.

Solution

1. X = the number of fencers who do **not** use the foil as their main weapon
2. 0, 1, 2, 3,... 25
3. $X \sim B(25,0.40)$
4. 10
5. 0.0442
6. The probability that all 25 not use the foil is almost zero. Therefore, it would be very surprising.

Exercises

Approximately 8% of students at a local high school participate in after-school sports all four years of high school. A group of 60 seniors is randomly chosen. Of interest is the number who participated in after-school sports all four years of high school.

1. In words, define the random variable *X*.
2. List the values that *X* may take on.
3. Give the distribution of *X*. *X* ~ _____(_____,_____)
4. How many seniors are expected to have participated in after-school sports all four years of high school?
5. Based on numerical values, would you be surprised if none of the seniors participated in after-school sports all four years of high school? Justify your answer numerically.
6. Based upon numerical values, is it more likely that four or that five of the seniors participated in after-school sports all four years of high school? Justify your answer numerically.

Solution

1. X = the number of high school students who participate in after school sports all four years of high school.
2. 0, 1, 2, ..., 60
3. ~ B(60, 0.08)
4. 4.8
5. Yes, P(x = 0) = 0.0067, which is a small probability
6. P(x = 4) = 0.1873, P(x = 5) = 0.1824. More likely to get four.

Exercises

The chance of an IRS audit for a tax return with over $25,000 in income is about 2% per year. We are interested in the expected number of audits a person with that income has in a 20-year period. Assume each year is independent.

1. In words, define the random variable *X*.
2. List the values that *X* may take on.
3. Give the distribution of *X*. *X* ~ _____(_____,_____)
4. How many audits are expected in a 20-year period?
5. Find the probability that a person is not audited at all.

6. Find the probability that a person is audited more than twice.

Solution

1. X = the number of audits in a 20-year period
2. 0, 1, 2, ..., 20
3. $X \sim B(20, 0.02)$
4. 0.4
5. 0.6676
6. 0.0071

Exercises

It has been estimated that only about 30% of California residents have adequate earthquake supplies. Suppose you randomly survey 11 California residents. We are interested in the number who have adequate earthquake supplies.

1. In words, define the random variable X.
2. List the values that X may take on.
3. Give the distribution of X. $X \sim$ _____(_____,_____)
4. What is the probability that at least eight have adequate earthquake supplies?
5. Is it more likely that none or that all of the residents surveyed will have adequate earthquake supplies? Why?
6. How many residents do you expect will have adequate earthquake supplies?

Solution

1. X = the number of California residents who do have adequate earthquake supplies.
2. 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11
3. B(11, 0.30)
4. 0.0043
5. P(x = 0) = 0.0198; P(x = 11) = 0 or none

6. 3.3

Exercises

There are two similar games played for Chinese New Year and Vietnamese New Year. In the Chinese version, fair dice with numbers 1, 2, 3, 4, 5, and 6 are used, along with a board with those numbers. In the Vietnamese version, fair dice with pictures of a gourd, fish, rooster, crab, crayfish, and deer are used. The board has those six objects on it, also. We will play with bets being $1. The player places a bet on a number or object. The "house" rolls three dice. If none of the dice show the number or object that was bet, the house keeps the $1 bet. If one of the dice shows the number or object bet (and the other two do not show it), the player gets back his or her $1 bet, plus $1 profit. If two of the dice show the number or object bet (and the third die does not show it), the player gets back his or her $1 bet, plus $2 profit. If all three dice show the number or object bet, the player gets back his or her $1 bet, plus $3 profit. Let X = number of matches and Y = profit per game.

1. In words, define the random variable X.
2. List the values that X may take on.
3. Give the distribution of X. X ~ _____(_____,_____)
4. List the values that Y may take on. Then, construct one PDF table that includes both X and Y and their probabilities.
5. Calculate the average expected matches over the long run of playing this game for the player.
6. Calculate the average expected earnings over the long run of playing this game for the player.
7. Determine who has the advantage, the player or the house.

Solution

1. X = the number of matches
2. 0, 1, 2, 3

3. $X \sim B\left(3, \frac{1}{6}\right)$
4. In dollars: −1, 1, 2, 3
5. $\frac{1}{2}$
6. Multiply each Y value by the corresponding X probability from the PDF table. The answer is −0.0787. You lose about eight cents, on average, per game.
7. The house has the advantage.

Exercises

According to The World Bank, only 9% of the population of Uganda had access to electricity as of 2009. Suppose we randomly sample 150 people in Uganda. Let X = the number of people who have access to electricity.

1. What is the probability distribution for X?
2. Using the formulas, calculate the mean and standard deviation of X.
3. Use your calculator to find the probability that 15 people in the sample have access to electricity.
4. Find the probability that at most ten people in the sample have access to electricity.
5. Find the probability that more than 25 people in the sample have access to electricity.

Solution

1. X ~ B(150,0.09)
2. $\text{Mean} = np = 150(0.09) = 13.5$;
   $\text{Standard Deviation} = npq = 150(0.09)(0.91) \approx 3.5050$
3. P(x = 15) = binompdf(150, 0.09, 15) = 0.0988
4. P(x ≤ 10) = binomcdf(150, 0.09, 10) = 0.1987
5. P(x > 25) = 1 – P(x ≤ 25) = 1 – binomcdf(150, 0.09, 25) = 1 – 0.9991 = 0.0009

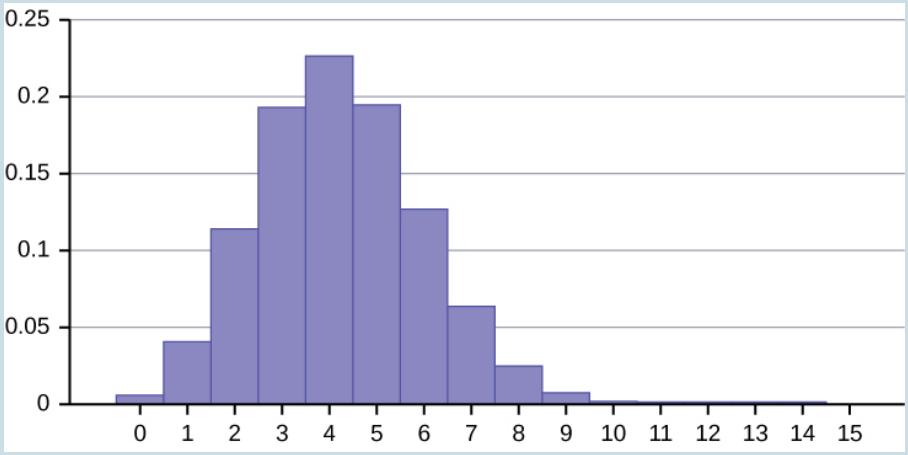
Exercises

The literacy rate for a nation measures the proportion of people age 15 and over that can read and write. The literacy rate in Afghanistan is 28.1%. Suppose you choose 15 people in Afghanistan at random. Let X = the number of people who are literate.

1. Sketch a graph of the probability distribution of X.
2. Using the formulas, calculate the (i) mean and (ii) standard deviation of X.
3. Find the probability that more than five people in the sample are literate. Is it more likely that three people or four people are literate?

Solution

1. $X \sim B(15, 0.281)$



   1. Mean = $\mu = np = 15(0.281) = 4.215$
   2. Standard Deviation = $\sigma = \sqrt{npq} = \sqrt{15\,(0.281)\,(0.719)} = 1.7409$

2. $P(x > 5) = 1 - P(x \leq 5) = 1 - \text{binomcdf}(15, 0.281, 5) = 1 - 0.7754 = 0.2246$
   $P(x = 3) = \text{binompdf}(15, 0.281, 3) = 0.1927$
   $P(x = 4) = \text{binompdf}(15, 0.281, 4) = 0.2259$
3. It is more likely that four people are literate than three people are.

Section 4.4 Homework

Exercises

A consumer looking to buy a used red Miata car will call dealerships until she finds a dealership that carries the car. She estimates the probability that any independent dealership will have the car will be 28%. We are interested in the number of dealerships she must call.

1. In words, define the random variable X.
2. List the values that X may take on.
3. Give the distribution of X. X ~ _____(_____,_____)
4. On average, how many dealerships would we expect her to have to call until she finds one that has the car?
5. Find the probability that she must call at most four dealerships.
6. Find the probability that she must call three or four dealerships.

Solution

1. X = the number of dealers she calls until she finds one with a used red Miata.
2. 1, 2, 3,...
3. X ~ G(0.28)
4. 3.57
5. 0.7313
6. 0.2497

Exercises

Suppose that the probability that an adult in America will watch the Super Bowl is 40%. Each

person is considered independent. We are interested in the number of adults in America we must survey until we find one who will watch the Super Bowl.

1. In words, define the random variable X.
2. List the values that X may take on.
3. Give the distribution of X. $X \sim$ _____(_____,_____)
4. How many adults in America do you expect to survey until you find one who will watch the Super Bowl?
5. Find the probability that you must ask seven people.
6. Find the probability that you must ask three or four people.

Solution

1. X = the number of adults in America who are surveyed until one says he or she will watch the Super Bowl.
2. 0, 1, 2, ...
3. $X \sim G(0.40)$
4. 2.5
5. 0.0187
6. 0.2304

Exercises

It has been estimated that only about 30% of California residents have adequate earthquake supplies. Suppose we are interested in the number of California residents we must survey until we find a resident who does **not** have adequate earthquake supplies.

1. In words, define the random variable X.
2. List the values that X may take on.
3. Give the distribution of X. $X \sim$ _____(_____,_____)
4. What is the probability that we must survey just one or two residents until we find a California resident who does not have adequate earthquake supplies?

5. What is the probability that we must survey at least three California residents until we find a California resident who does not have adequate earthquake supplies?
6. How many California residents do you expect to need to survey until you find a California resident who **does not** have adequate earthquake supplies?
7. How many California residents do you expect to need to survey until you find a California resident who **does** have adequate earthquake supplies?

Exercises

In one of its Spring catalogs, L.L. Bean® advertised footwear on 29 of its 192 catalog pages. Suppose we randomly survey 20 pages. We are interested in the number of pages that advertise footwear. Each page may be picked more than once.

1. In words, define the random variable X.
2. List the values that X may take on.
3. Give the distribution of X. $X \sim$ _____(_____,_____)
4. How many pages do you expect to advertise footwear on them?
5. Is it probable that all twenty will advertise footwear on them? Why or why not?
6. What is the probability that fewer than ten will advertise footwear on them?
7. Reminder: A page may be picked more than once. We are interested in the number of pages that we must randomly survey until we find one that has footwear advertised on it. Define the random variable X and give its distribution.
8. What is the probability that you only need to survey at most three pages in order to find one that advertises footwear on it?
9. How many pages do you expect to need to survey in order to find one that advertises footwear?

Solution

1. X = the number of pages that advertise footwear
2. X takes on the values 0, 1, 2, ..., 20

3. $X \sim B(20, \frac{29}{192})$
4. 3.02
5. No
6. 0.9997
7. X = the number of pages we must survey until we find one that advertises footwear.
8. X \sim G(\frac{29}{192}=0.3881)
9. 6.6207 pages

Exercises

Suppose that you are performing the probability experiment of rolling one fair six-sided die. Let F be the event of rolling a four or a five. You are interested in how many times you need to roll the die in order to obtain the first four or five as the outcome.

- p = probability of success (event F occurs)
- q = probability of failure (event F does not occur)

1. Write the description of the random variable X.
2. What are the values that X can take on?
3. Find the values of p and q.
4. Find the probability that the first occurrence of event F (rolling a four or five) is on the second trial.

Solution

1. X = the number of times you need to roll the die in order to get the face four or five.
2. X can take on the values 1, 2, 3, and so on.
3. $p = \frac{2}{6} \text{ and } q = \frac{4}{6}$
4. 0.2222

Exercises

Ellen has music practice three days a week. She practices for all of the three days 85% of the time, two days 8% of the time, one day 4% of the time, and no days 3% of the time. One week is selected at random. What values does *X* take on?

Solution
0, 1, 2, and 3

Exercises

The World Bank records the prevalence of HIV in countries around the world. According to their data, "Prevalence of HIV refers to the percentage of people ages 15 to 49 who are infected with HIV."[1] In South Africa, the prevalence of HIV is 17.3%. Let *X* = the number of people you test until you find a person infected with HIV.

1. Sketch a graph of the distribution of the discrete random variable *X*.
2. What is the probability that you must test 30 people to find one with HIV?
3. What is the probability that you must ask ten people?
4. Find the (i) mean and (ii) standard deviation of the distribution of *X*.

Solution

1. X ~ G(0.173)
2. P(x = 30) = geometpdf(0.173, 30) = 0.0007
3. P(x = 10) = geometpdf(0.173, 10) = 0.0313
4. $\text{Mean} = \mu = \frac{1}{p} = \frac{1}{0.173} \approx 5.7804$

   $\text{Standard Deviation} = \sigma = \frac{\sqrt{(1-p)}}{p} = \frac{\sqrt{1-0.173}}{0.173} = 5.2566$

Exercises

According to a recent Pew Research poll, 75% of millennials (people born between 1981 and 1995) have a profile on a social networking site. Let X = the number of millennials you ask until you find a person without a profile on a social networking site.

1. Describe the distribution of X.
2. Find the (i) mean and (ii) standard deviation of X.
3. What is the probability that you must ask ten people to find one person without a social networking site?
4. What is the probability that you must ask 20 people to find one person without a social networking site?
5. What is the probability that you must ask *at most* five people?

Solution

1. $X \sim G(0.25)$
2. (i) Mean = $\mu = \frac{1}{p} = \frac{1}{0.25} = 4$

   (ii) Standard Deviation = $\sigma = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1-0.25}{0.25^2}} \approx 3.4641$
3. $P(x = 10)$ = geometpdf(0.25, 10) = 0.0188
4. $P(x = 20)$ = geometpdf(0.25, 20) = 0.0011
5. $P(x \leq 5)$ = geometcdf(0.25, 5) = 0.7627

Footnotes

1. [1] "Prevalence of HIV, total (% of populations ages 15-49)," The World Bank, 2013. Available online at http://data.worldbank.org/indicator/SH.DYN.AIDS.ZS?order=wbapi_data_value_2011+wbapi_data_value+wbapi_data_value-last&sort=desc (accessed May 15, 2013).

Section 4.5 Homework

Exercises

A group of Martial Arts students is planning on participating in an upcoming demonstration. Six are students of Tae Kwon Do; seven are students of Shotokan Karate. Suppose that eight students are randomly picked to be in the first demonstration. We are interested in the number of Shotokan Karate students in that first demonstration.

1. In words, define the random variable *X*.
2. List the values that *X* may take on.
3. Give the distribution of *X*. *X* ~ _____(_____,_____)
4. How many Shotokan Karate students do we expect to be in that first demonstration?

Solution

1. X = the number of Shotokan Karate students in the first demonstration
2. 1, 2, 3, 4, 5, 6, 7
3. X ~ H(7, 6, 8)
4. 4.31

Exercises

In one of its Spring catalogs, L.L. Bean® advertised footwear on 29 of its 192 catalog pages. Suppose we randomly survey 20 pages. We are interested in the number of pages that advertise footwear. Each page may be picked at most once.

1. In words, define the random variable *X*.
2. List the values that *X* may take on.

3. Give the distribution of *X*. *X* ~ _____(_____,_____)
4. How many pages do you expect to advertise footwear on them?
5. Calculate the standard deviation.

Solution

1. *X* = the number of pages that advertise footwear
2. 0, 1, 2, 3, ..., 20
3. *X* ~ *H*(29, 163, 20); *r* = 29, *b* = 163, *n* = 20
4. 3.03
5. 1.5197

Exercises

Suppose that a technology task force is being formed to study technology awareness among instructors. Assume that ten people will be randomly chosen to be on the committee from a group of 28 volunteers, 20 who are technically proficient and eight who are not. We are interested in the number on the committee who are **not** technically proficient.

1. In words, define the random variable *X*.
2. List the values that *X* may take on.
3. Give the distribution of *X*. *X* ~ _____(_____,_____)
4. How many instructors do you expect on the committee who are **not** technically proficient?
5. Find the probability that at least five on the committee are not technically proficient.
6. Find the probability that at most three on the committee are not technically proficient.

Solution

1. X = the number of people on the committee who are not technically proficient
2. 0, 1, 2, 3, 4, 5, 6, 7, 8
3. X ~ B(10, 8, 28)
4. 2.8571

5. 0.0772
6. 0.7160

Exercises

Suppose that nine Massachusetts athletes are scheduled to appear at a charity benefit. The nine are randomly chosen from eight volunteers from the Boston Celtics and four volunteers from the New England Patriots. We are interested in the number of Patriots picked.

1. In words, define the random variable X.
2. List the values that X may take on.
3. Give the distribution of X. $X \sim$ _____(_____,_____)
4. Are you choosing the nine athletes with or without replacement?

Solution

1. X = the number of Patriots picked
2. 0, 1, 2, 3, 4
3. $X \sim H(4, 8, 9)$
4. Without replacement

Exercises

A bridge hand is defined as 13 cards selected at random and without replacement from a deck of 52 cards. In a standard deck of cards, there are 13 cards from each suit: hearts, spades, clubs, and diamonds. What is the probability of being dealt a hand that does not contain a heart?

1. What is the group of interest?
2. How many are in the group of interest?
3. How many are in the other group?
4. Let X = _____. What values does X take on?
5. The probability question is P(_____).
6. Find the probability in question.
7. Find the (i) mean and (ii) standard deviation of X.

Section 4.6 Homework

Exercises

The switchboard in a Minneapolis law office gets an average of 5.5 incoming phone calls during the noon hour on Mondays. Experience shows that the existing staff can handle up to six calls in an hour. Let X = the number of calls received at noon.

1. Find the mean and standard deviation of X.
2. What is the probability that the office receives at most six calls at noon on Monday?
3. Find the probability that the law office receives six calls at noon. What does this mean to the law office staff who get, on average, 5.5 incoming phone calls at noon?
4. What is the probability that the office receives more than eight calls at noon?

Solution

1. $X \sim P(5.5)$; $\mu = 5.5$; $\sigma = \sqrt{5.5} \approx 2.3452$
2. $P(x \leq 6)$ = poissoncdf(5.5, 6) ≈ 0.6860
3. There is a 15.7% probability that the law staff will receive more calls than they can handle.
4. $P(x > 8) = 1 - P(x \leq 8)$ = 1 – poissoncdf(5.5, 8) ≈ 1 – 0.8944 = 0.1056

Exercises

The maternity ward at Dr. Jose Fabella Memorial Hospital in Manila in the Philippines is one of the busiest in the world with an average of 60 births per day. Let X = the number of births in an hour.

1. Find the mean and standard deviation of X.
2. Sketch a graph of the probability distribution of X.
3. What is the probability that the maternity ward will deliver three babies in one hour?
4. What is the probability that the maternity ward will deliver at most three babies in one hour?
5. What is the probability that the maternity ward will deliver more than five babies in one hour?

Exercises

A manufacturer of Christmas tree light bulbs knows that 3% of its bulbs are defective. Find the probability that a string of 100 lights contains at most four defective bulbs using both the binomial and Poisson distributions.

Solution
Let X = the number of defective bulbs in a string.

Using the Poisson distribution:

$\mu = np = 100(0.03) = 3$
$X \sim P(3)$
$P(x \leq 4)$ = poissoncdf(3, 4) ≈ 0.8153
Using the binomial distribution:

$X \sim B(100, 0.03)$
$P(x \leq 4)$ = binomcdf(100, 0.03, 4) ≈ 0.8179
The Poisson approximation is very good—the difference between the probabilities is only 0.0026.

Exercises

The average number of children a Japanese woman has in her lifetime is 1.37. Suppose that one Japanese woman is randomly chosen.

1. In words, define the random variable *X*.
2. List the values that *X* may take on.
3. Give the distribution of *X*. *X* ~ _____(_____,_____)
4. Find the probability that she has no children.
5. Find the probability that she has fewer children than the Japanese average.
6. Find the probability that she has more children than the Japanese average.

Exercises

The average number of children a Spanish woman has in her lifetime is 1.47. Suppose that one Spanish woman is randomly chosen.

1. In words, define the Random Variable *X*.
2. List the values that *X* may take on.
3. Give the distribution of *X*. *X* ~ _____(_____,_____)
4. Find the probability that she has no children.
5. Find the probability that she has fewer children than the Spanish average.
6. Find the probability that she has more children than the Spanish average.

Solution

1. *X* = the number of children for a Spanish woman
2. 0, 1, 2, 3,...

3. $X \sim P(1.47)$
4. 0.2299
5. 0.5679
6. 0.4321

Exercises

Fertile, female cats produce an average of three litters per year. Suppose that one fertile, female cat is randomly chosen. In one year, find the probability she produces:

1. In words, define the random variable X.
2. List the values that X may take on.
3. Give the distribution of X. $X \sim$ _____
4. Find the probability that she has no litters in one year.
5. Find the probability that she has at least two litters in one year.
6. Find the probability that she has exactly three litters in one year.

Solution

1. X = the number of litters a fertile, female cat produces per year
2. 0, 1, 2, 3, ...
3. X ~ P(3)
4. 0.0498
5. 0.8009
6. 0.2240

Exercises

The chance of having an extra fortune in a fortune cookie is about 3%. Given a bag of 144 fortune cookies, we are interested in the number of cookies with an extra fortune. Two distributions may be used to solve this problem, but only use one distribution to solve the problem.

1. In words, define the random variable X.
2. List the values that X may take on.
3. Give the distribution of X. $X \sim$ _____(_____,_____)
4. How many cookies do we expect to have an extra fortune?
5. Find the probability that none of the cookies have an extra fortune.
6. Find the probability that more than three have an extra fortune.
7. As n increases, what happens involving the probabilities using the two distributions? Explain in complete sentences.

Solution

1. X = the number of fortune cookies that have an extra fortune
2. 0, 1, 2, 3,... 144
3. $X \sim B(144, 0.03)$ or $P(4.32)$
4. 4.32
5. 0.0124 or 0.0133
6. 0.6300 or 0.6264
7. As n gets larger, the probabilities get closer together.

Exercises

According to the South Carolina Department of Mental Health website, for every 200 U.S. women,

the average number who suffer from anorexia is one. Out of a randomly chosen group of 600 U.S. women determine the following.

1. In words, define the random variable *X*.
2. List the values that *X* may take on.
3. Give the distribution of *X*. *X* ~ _____(_____,_____)
4. How many are expected to suffer from anorexia?
5. Find the probability that no one suffers from anorexia.
6. Find the probability that more than four suffer from anorexia.

Solution

1. X = the number of women that suffer from anorexia
2. 0, 1, 2, 3, ..., 600
3. X ~ P(3)
4. 3
5. 0.0498
6. 0.1847

Exercises

The chance of an IRS audit for a tax return with over $25,000 in income is about 2% per year. Suppose that 100 people with tax returns over $25,000 are randomly picked. We are interested in the number of people audited in one year. Use a Poisson distribution to answer the following questions.

1. In words, define the random variable *X*.
2. List the values that *X* may take on.
3. Give the distribution of *X*. *X* ~ _____(_____,_____)
4. How many are expected to be audited?

5. Find the probability that no one was audited.
6. Find the probability that at least three were audited.

Solution

1. X = the number of people audited in one year
2. 0, 1, 2, ..., 100
3. $X \sim P(2)$
4. 2
5. 0.1353
6. 0.3233

Exercises

Approximately 8% of students at a local high school participate in after-school sports all four years of high school. A group of 60 seniors is randomly chosen. Of interest is the number that participated in after-school sports all four years of high school.

1. In words, define the random variable X.
2. List the values that X may take on.
3. Give the distribution of X. $X \sim$ _____(_____,_____)
4. How many seniors are expected to have participated in after-school sports all four years of high school?
5. Based on numerical values, would you be surprised if none of the seniors participated in after-school sports all four years of high school? Justify your answer numerically.
6. Based on numerical values, is it more likely that four or that five of the seniors participated in after-school sports all four years of high school? Justify your answer numerically.

Solution

1. X = the number of seniors who participated in after-school sports all four years of high school

2. 0, 1, 2, 3, ... 60
3. X ~ P(4.8)
4. 4.8
5. Yes
6. 4

Exercises

On average, Pierre, an amateur chef, drops three pieces of egg shell into every two cake batters he makes. Suppose that you buy one of his cakes.

1. In words, define the random variable *X*.
2. List the values that *X* may take on.
3. Give the distribution of *X*. *X* ~ _____(_____,_____)
4. On average, how many pieces of egg shell do you expect to be in the cake?
5. What is the probability that there will not be any pieces of egg shell in the cake?
6. Let's say that you buy one of Pierre's cakes each week for six weeks. What is the probability that there will not be any egg shell in any of the cakes?
7. Based upon the average given for Pierre, is it possible for there to be seven pieces of shell in the cake? Why?

Solution

1. *X* = the number of shell pieces in one cake
2. 0, 1, 2, 3,...
3. *X* ~ *P*(1.5)
4. 1.5
5. 0.2231
6. 0.0001
7. Yes

Exercises

Use the following information to answer the next two exercises: The average number of times per week that Mrs. Plum's cats wake her up at night because they want to play is ten. We are interested in the number of times her cats wake her up each week.

In words, the random variable X = _____

a. the number of times Mrs. Plum's cats wake her up each week.
b. the number of times Mrs. Plum's cats wake her up each hour.
c. the number of times Mrs. Plum's cats wake her up each night.
d. the number of times Mrs. Plum's cats wake her up.

Solution
a

Exercises

Find the probability that her cats will wake her up no more than five times next week.

a. 0.5000
b. 0.9329
c. 0.0378
d. 0.0671

Solution
d

CHAPTER 5: CONTINUOUS RANDOM VARIABLES

INTRODUCTION TO CHAPTER 5: CONTINUOUS RANDOM VARIABLES



Figure 5.1. The heights of these radish plants are continuous random variables. (Credit: Rev Stan)

Continuous random variables have many applications. Baseball batting averages, IQ scores, the length of time a long distance telephone call lasts, the amount of money a person carries, the length of time a computer chip lasts, and SAT scores are just a few. The field of reliability depends on a variety of continuous random variables.

Note

The values of discrete and continuous random variables can be ambiguous. For example, if X is equal to the number of miles (to the nearest mile) you drive to work, then X is a **discrete** random variable. You count the miles. If X is the distance you drive to work, then you measure values of X and X is a **continuous** random variable. For a second example, if X is equal to the number of books in a backpack, then X is a **discrete** random variable. If X is the weight of a book, then X is a **continuous** random variable because weights are measured.
How the random variable is defined is very important!

Collaborative Exercises

Section 5.3

Collaborative Exercise

Have each class member count the change he or she has in his or her pocket or purse. Your instructor will record the amounts in dollars and cents. Construct a histogram of the data taken by the class. Use five intervals. Draw a smooth curve through the bars. The graph should look approximately exponential. Then calculate the mean.

Let X = the amount of money a student in your class has in his or her pocket or purse.

The distribution for X is approximately exponential with mean, μ = _____ and m = _____. The standard deviation, σ = _____.

Draw the appropriate exponential graph. You should label the x– and y–axes, the decay rate, and

the mean. Shade the area that represents the probability that one student has less than $.40 in his or her pocket or purse. (Shade $P(x < 0.40)$).

Media Attributions

- Chapter 5 Introduction Figure © OpenStax Introductory Statistics is licensed under a CC BY (Attribution) license

5.1 CONTINUOUS PROBABILITY FUNCTIONS

Learning Objectives

By the end of this section, students will be able to:

- describe the properties of continuous probability distributions.

Properties of Continuous Probability Distributions

The graph of a continuous probability distribution is a curve. Probability is represented by area under the curve.

The curve is called the **probability density function** (abbreviated as **pdf**). We use the symbol $f(x)$ to represent the curve. $f(x)$ is the function that corresponds to the graph; we use the density function $f(x)$ to draw the graph of the probability distribution.

Area under the curve is given by a different function called the **cumulative distribution function** (abbreviated as **cdf**). The cumulative distribution function is used to evaluate probability as area.

- The outcomes are measured, not counted.
- The entire area under the curve and above the x-axis is equal to one.
- Probability is found for intervals of x values rather than for individual x values.
- $P(c < x < d)$ is the probability that the random variable X is in the interval between the values c and d. $P(c < x < d)$ is the area under the curve, above the x-axis, to the right of c and the left of d.
- $P(x = c) = 0$ The probability that x takes on any single individual value is zero. The area below the curve, above the x-axis, and between $x = c$ and $x = c$ has no width, and therefore no area (area = 0). Since the probability is equal to the area, the probability is also zero.

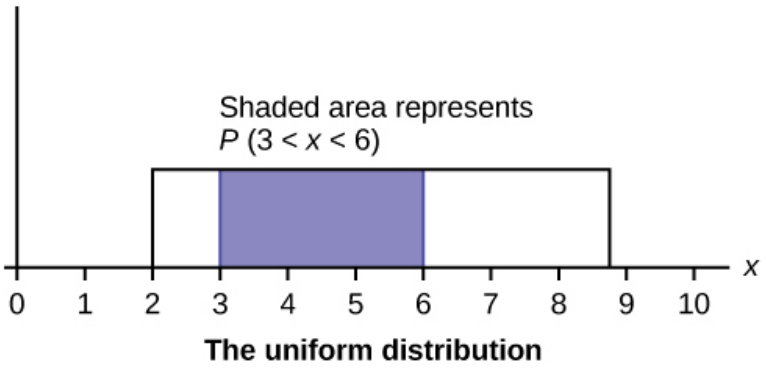
- $P(c < x < d)$ is the same as $P(c \leq x \leq d)$ because probability is equal to area.

We will find the area that represents probability by using geometry, formulas, technology, or probability tables. In general, calculus is needed to find the area under the curve for many probability density functions. When we use formulas to find the area in this textbook, the formulas were found by using the techniques of integral calculus. However, because most students taking this course have not studied calculus, we will not be using calculus in this textbook.

There are many continuous probability distributions. When using a continuous probability distribution to model probability, the distribution used is selected to model and fit the particular situation in the best way.

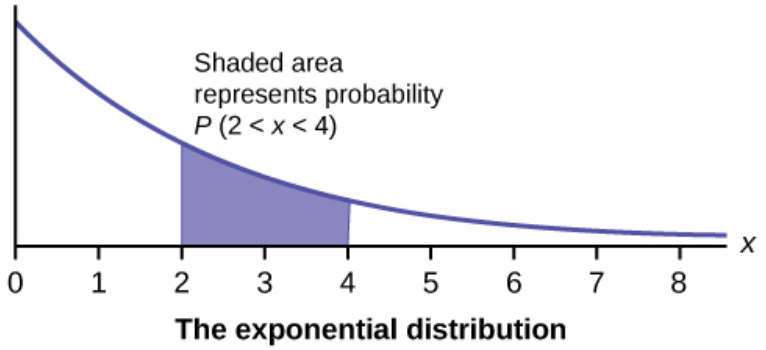
In this chapter and the next, we will study the **uniform distribution**, the exponential distribution, and the normal distribution. The following graphs illustrate these distributions.

The graph shows a **Uniform Distribution** with the area between $x = 3$ and $x = 6$ shaded to represent the probability that the value of the random variable X is in the interval between three and six.



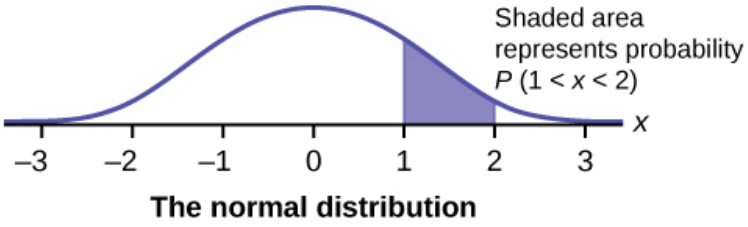
The uniform distribution

The graph shows an **Exponential Distribution** with the area between $x = 2$ and $x = 4$ shaded to represent the probability that the value of the random variable X is in the interval between two and four.



The exponential distribution

The graph shows the Standard Normal Distribution with the area between $x = 1$ and $x = 2$ shaded to represent the probability that the value of the random variable X is in the interval between one and two.



The normal distribution

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=750#oembed-1

Continuous Probability Functions

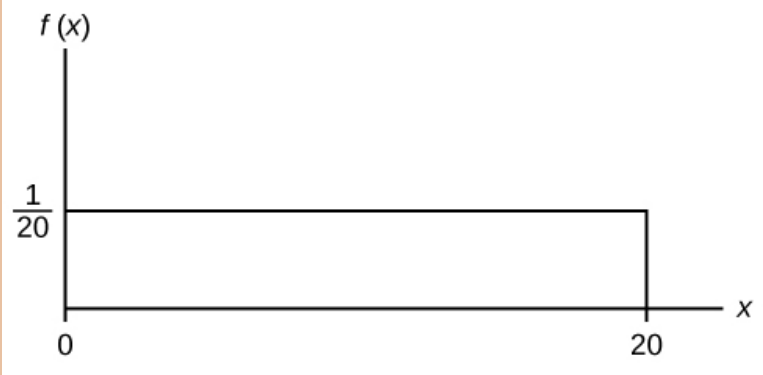
We begin by defining a continuous probability density function. We use the function notation $f(x)$. Intermediate algebra may have been your first formal introduction to functions. In the study of probability, the functions we study are special. We define the function $f(x)$ so that the area between it and the x-axis is equal to a probability. Since the maximum probability is one, the maximum area is also one. **For continuous probability distributions, PROBABILITY = AREA.**

Example

Consider the function

f(x) = $\frac{1}{20}$ for 0 ≤ x ≤ 20. x = a real number. The graph of

f(x) = $\frac{1}{20}$ is a horizontal line. However, since 0 ≤ x ≤ 20, f(x) is restricted to the portion between x = 0 and x = 20, inclusive.



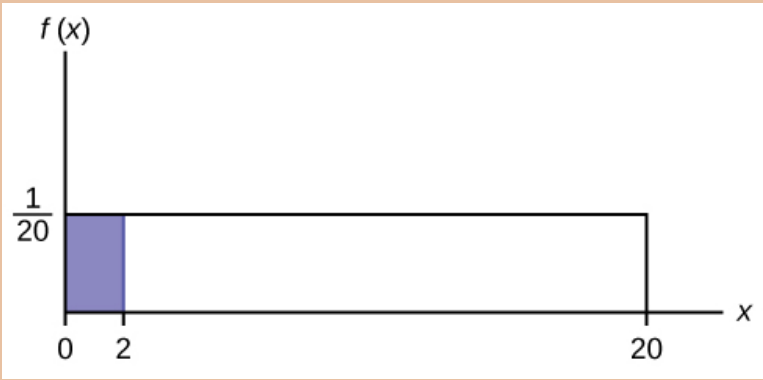
f(x) = $\frac{1}{20}$ for $0 \le x \le 20$.

The graph of f(x) = $\frac{1}{20}$ is a horizontal line segment when $0 \le x \le 20$.

The area between f(x) = $\frac{1}{20}$ where $0 \le x \le 20$ and the x-axis is the area of a rectangle with base = 20 and height = $\frac{1}{20}$.

$$\text{AREA} = 20\left(\frac{1}{20}\right) = 1$$

Suppose we want to find the area between f(x) = $\frac{1}{20}$ and the x-axis where $0 < x < 2$.



$$\text{AREA} = (2-0)\left(\frac{1}{20}\right) = 0.1$$

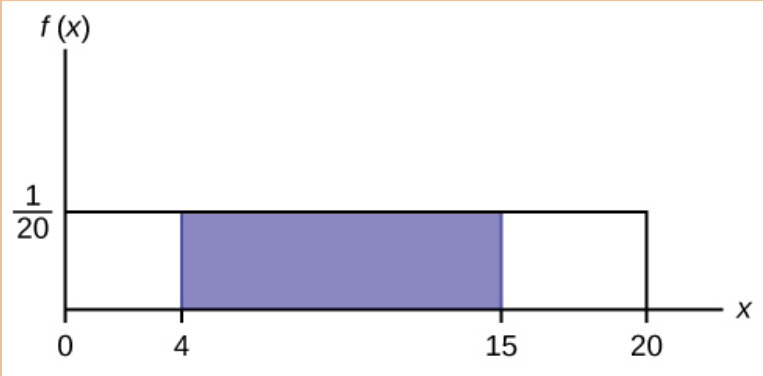
$$(2-0) = 2 = \text{base of a rectangle}$$

Reminder

area of a rectangle = (base)(height).

The area corresponds to a probability. The probability that x is between zero and two is 0.1, which can be written mathematically as P(0 < x < 2) = P(x < 2) = 0.1.

Suppose we want to find the area between f(x) = $\frac{1}{20}$ and the x-axis where 4 < x < 15.



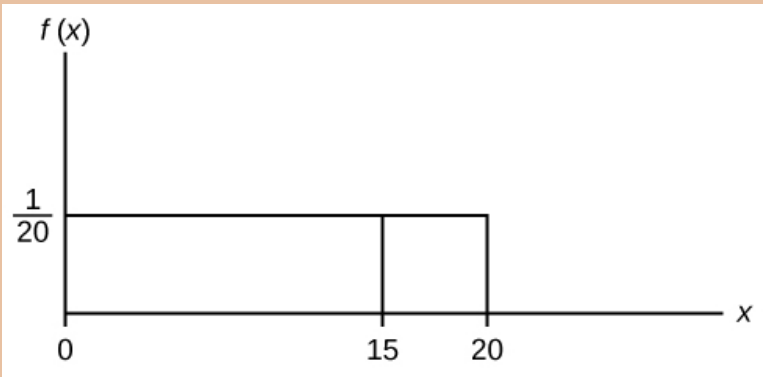
$$\text{AREA} = (15 - 4)\left(\frac{1}{20}\right) = 0.55$$

$$\text{AREA} = (15 - 4)\left(\frac{1}{20}\right) = 0.55$$

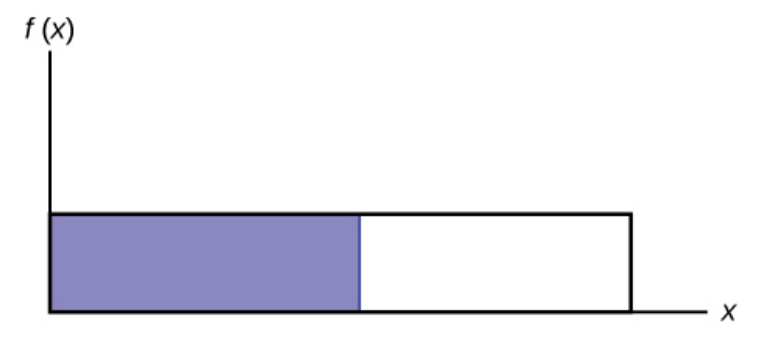
$$(15 - 4) = 11 = \text{the base of a rectangle}$$

The area corresponds to the probability P(4 < x < 15) = 0.55.

Suppose we want to find P(x = 15). On an x-y graph, x = 15 is a vertical line. A vertical line has no width (or zero width). Therefore, P(x = 15) = (base)(height) = (0)$\left(\frac{1}{20}\right)$ = 0



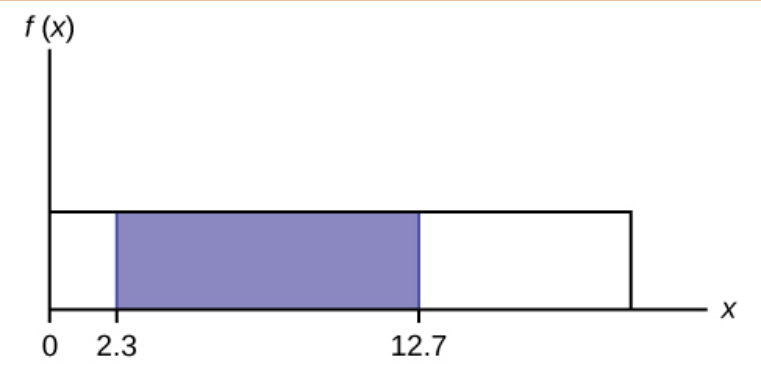
P(X ≤ x) (can be written as P(X < x) for continuous distributions) is called the cumulative distribution function or CDF. Notice the "less than or equal to" symbol. We can use the CDF to calculate P(X > x). The CDF gives "area to the left" and P(X > x) gives "area to the right." We calculate P(X > x) for continuous distributions as follows: P(X > x) = 1 – P (X < x).



Label the graph with

f(x) and x. Scale the x and y axes with the maximum x and y values. f(x) = $\frac{1}{20}$, 0 ≤ x ≤ 20.

To calculate the probability that x is between two values, look at the following graph. Shade the region between x = 2.3 and x = 12.7. Then calculate the shaded area of a rectangle.



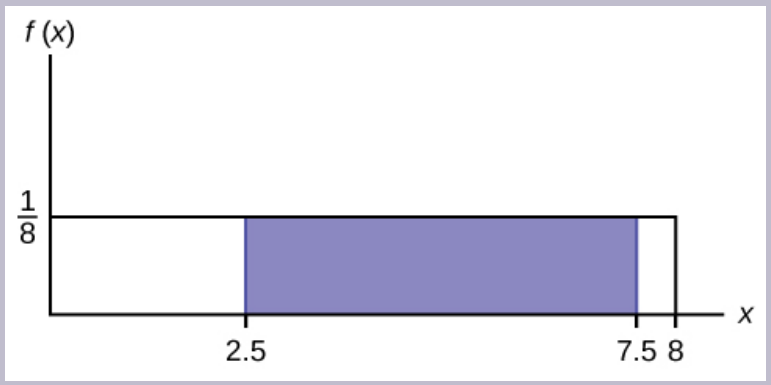
$$P(2.3 < x < 12.7) = (\text{base})(\text{height}) = (12.7 - 2.3)(\frac{1}{20}) = 0.52$$

Your Turn!

Consider the function $f(x) = \frac{1}{8}$

for 0 ≤ x ≤ 8. Draw the graph of $f(x)$ and find $P(2.5 < x < 7.5)$.

Solution



$P(2.5 < x < 7.5) = 0.625$

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=750#oembed-2

5.2 THE UNIFORM DISTRIBUTION

Learning Objectives

At the end of this section, students will be able to:

- describe uniform probability distribution and apply it appropriately.

The **uniform distribution** is a continuous probability distribution and is concerned with events that are equally likely to occur. When working out problems that have a **uniform distribution**, be careful to note if the data is inclusive or exclusive.

Example

The data in the table below are 55 smiling times, in seconds, of an eight-week-old baby.

10.4	19.6	18.8	13.9	17.8	16.8	21.6	17.9	12.5	11.1	4.9
12.8	14.8	22.8	20.0	15.9	16.3	13.4	17.1	14.5	19.0	22.8
1.3	0.7	8.9	11.9	10.9	7.3	5.9	3.7	17.9	19.2	9.8
5.8	6.9	2.6	5.8	21.7	11.8	3.4	2.1	4.5	6.3	10.7
8.9	9.4	9.4	7.6	10.0	3.3	6.7	7.8	11.6	13.8	18.6

The sample mean = 11.49 and the sample standard deviation = 6.23.

We will assume that the smiling times, in seconds, follow a uniform distribution between zero and

23 seconds, inclusive. This means that any smiling time from zero to and including 23 seconds is equally likely. The histogram that could be constructed from the sample is an empirical distribution that closely matches the theoretical uniform distribution.

Let X = length, in seconds, of an eight-week-old baby's smile.

The notation for the uniform distribution is

X ~ U(a, b) where a = the lowest value of x and b = the highest value of x.

The probability density function is

f(x) = $\frac{1}{b-a}$ for a ≤ x ≤ b.

For this example, X ~ U(0, 23) and f(x) = $\frac{1}{23-0}$ for 0 ≤ X ≤ 23.

Formulas for the theoretical mean and standard deviation are

$$\mu = \frac{a+b}{2} \text{ and } \sigma = \sqrt{\frac{(b-a)^2}{12}}$$

For this problem, the theoretical mean and standard deviation are

μ = $\frac{0 + 23}{2}$

= 11.50 seconds and σ = $\sqrt{\frac{(23 - 0)^2}{12}}$

= 6.64 seconds.

Notice that the theoretical mean and standard deviation are close to the sample mean and standard deviation in this example.

Your Turn!

The data that follow are the number of passengers on 35 different charter fishing boats. The sample mean = 7.9 and the sample standard deviation = 4.33. The data follow a **uniform**

distribution where all values between and including zero and 14 are equally likely. State the values of *a* and *b*. Write the distribution in proper notation, and calculate the theoretical mean and standard deviation.

1	12	4	10	4	14	11
7	11	4	13	2	4	6
3	10	0	12	6	9	10
5	13	4	10	14	12	11
6	10	11	0	11	13	2

Solution

a is zero; *b* is 14; $X \sim U(0, 14)$; $\mu = 7$ passengers; $\sigma = 4.04$ passengers

Example

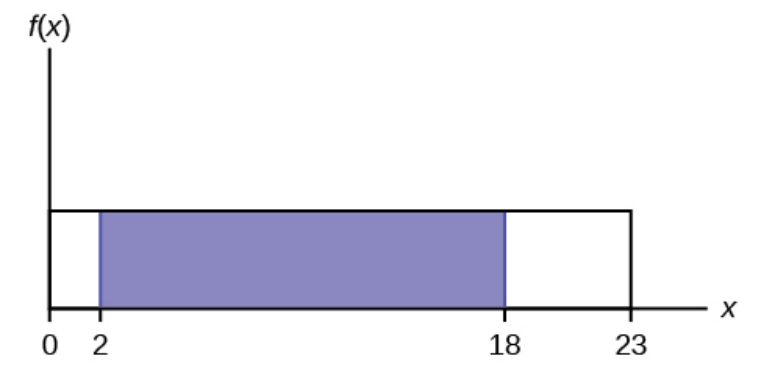
a. Refer to the data from the first example (copied below). What is the probability that a randomly chosen eight-week-old baby smiles between two and 18 seconds?

10.4	19.6	18.8	13.9	17.8	16.8	21.6	17.9	12.5	11.1	4.9
12.8	14.8	22.8	20.0	15.9	16.3	13.4	17.1	14.5	19.0	22.8
1.3	0.7	8.9	11.9	10.9	7.3	5.9	3.7	17.9	19.2	9.8
5.8	6.9	2.6	5.8	21.7	11.8	3.4	2.1	4.5	6.3	10.7
8.9	9.4	9.4	7.6	10.0	3.3	6.7	7.8	11.6	13.8	18.6

Solution

a. Find $P(2 < x < 18)$.

$P(2 < x < 18) = (\text{base})(\text{height}) = (18 - 2)\left(\frac{1}{23}\right) = \left(\frac{16}{23}\right)$.



b. Find the 90th percentile for an eight-week-old baby's smiling time.

Solution

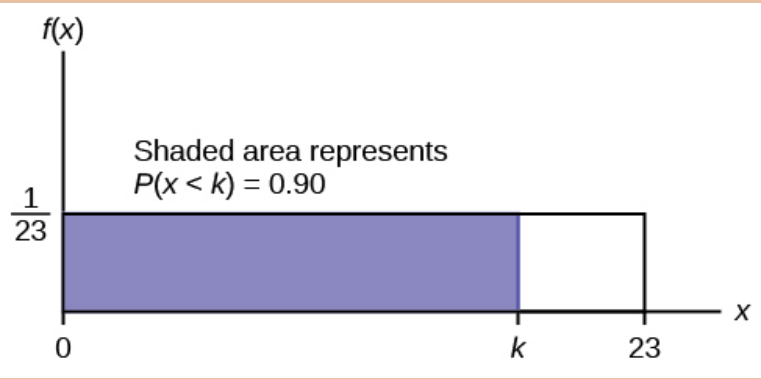
b. Ninety percent of the smiling times fall below the 90th percentile, k, so $P(x < k) = 0.90$

$$P(x < k) = 0.90$$

$$(\text{base})(\text{height}) = 0.90$$

$$(k - 0)\left(\frac{1}{23}\right) = 0.90$$

$$k = (23)(0.90) = 20.7$$



c. Find the probability that a random eight-week-old baby smiles more than 12 seconds **KNOWING** that the baby smiles **MORE THAN EIGHT SECONDS**.

Solution

c. This probability question is a **conditional**. You are asked to find the probability that an eight-week-old baby smiles more than 12 seconds when you **already know** the baby has smiled for more than eight seconds.

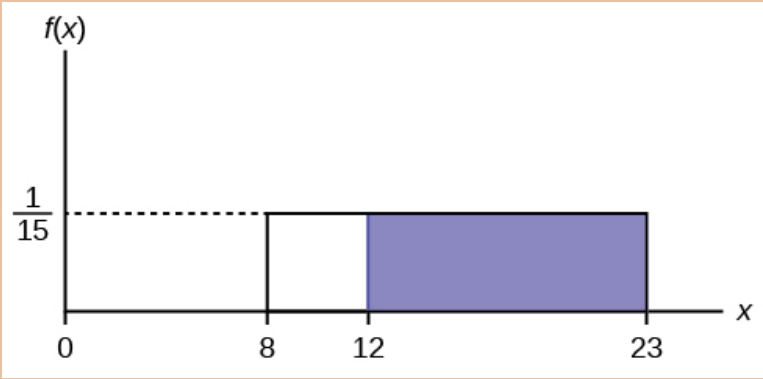
Find $P(x > 12|x > 8)$ There are two ways to do the problem. **For the first way**, use the fact that this

is a **conditional** and changes the sample space. The graph illustrates the new sample space. You already know the baby smiled for more than eight seconds.

Write a new $f(x)$: $f(x) = \frac{1}{23 - 8} = \frac{1}{15}$

for $8 < x < 23$

$P(x > 12|x > 8) = (23 - 12)\left(\frac{1}{15}\right) = \left(\frac{11}{15}\right)$

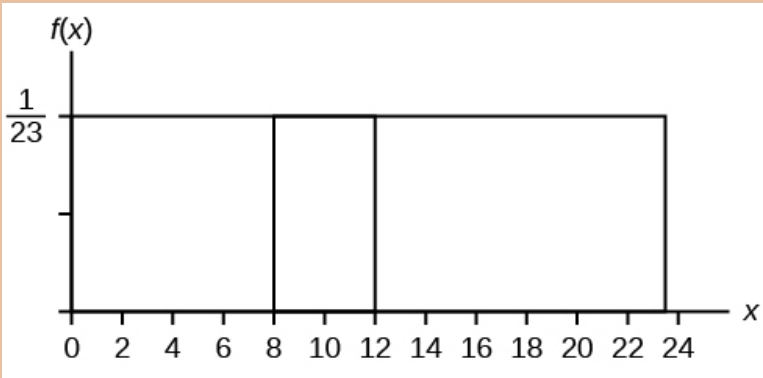


For the second way, use the conditional formula from Probability Topics with the original distribution $X \sim U(0, 23)$:

$$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)}$$

For this problem, A is $(x > 12)$ and B is $(x > 8)$.

So, $P(x > 12|x > 8) = \frac{(x > 12 \text{ AND } x > 8)}{P(x > 8)} = \frac{P(x > 12)}{P(x > 8)} = \frac{\frac{11}{23}}{\frac{15}{23}} = \frac{11}{15}$



Your Turn!

A distribution is given as $X \sim U(0, 20)$. What is $P(2 < x < 18)$? Find the 90^{th} percentile.

Solution

$P(2 < x < 18) = 0.8$; 90^{th} percentile = 18

Example

The amount of time, in minutes, that a person must wait for a bus is uniformly distributed between zero and 15 minutes, inclusive.

a. What is the probability that a person waits fewer than 12.5 minutes?

Solution

a. Let X = the number of minutes a person must wait for a bus. a = 0 and b = 15.

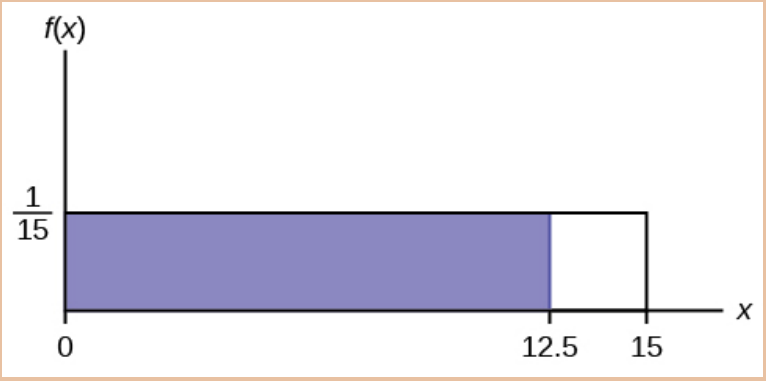
X ~ U(0, 15). Write the probability density function. f (x) = $\frac{1}{15 - 0} = \frac{1}{15}$

for $0 \le x \le 15$.

Find P (x < 12.5). Draw a graph.

$$P(x < k) = (\text{base})(\text{height}) = (12.5 - 0)(\frac{1}{15}) = 0.8333$$

The probability a person waits less than 12.5 minutes is 0.8333.



b. On the average, how long must a person wait? Find the mean, μ, and the standard deviation, σ.

Solution

$\mu = \frac{a+b}{2} = \frac{15+0}{2} = 7.5$ On the average, a person must wait 7.5 minutes.

$$\sigma = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(15-0)^2}{12}} = 4.3$$

The Standard deviation is 4.3 minutes.

c. Ninety percent of the time, the time a person must wait falls below what value?

This asks for the 90th percentile.

Solution

c. Find the 90th percentile. Draw a graph. Let k = the 90th percentile.

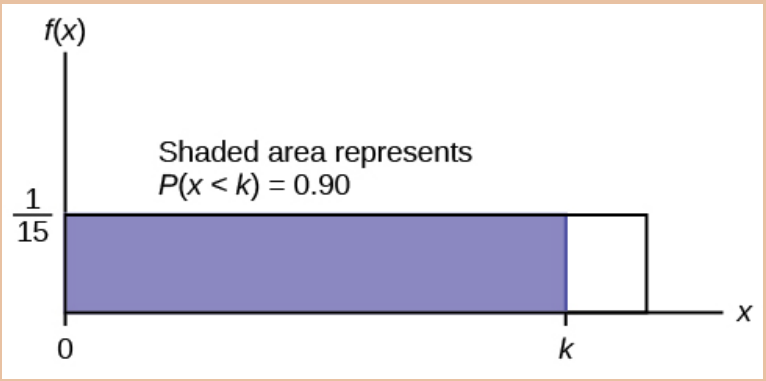
$$P(x < k) = (\text{base})(\text{height}) = (k-0)\left(\frac{1}{15}\right)$$

$$0.90 = (k)\left(\frac{1}{15}\right)$$

$$k = (0.90)(15) = 13.5$$

k is sometimes called a critical value.

The 90th percentile is 13.5 minutes. Ninety percent of the time, a person must wait at most 13.5 minutes.



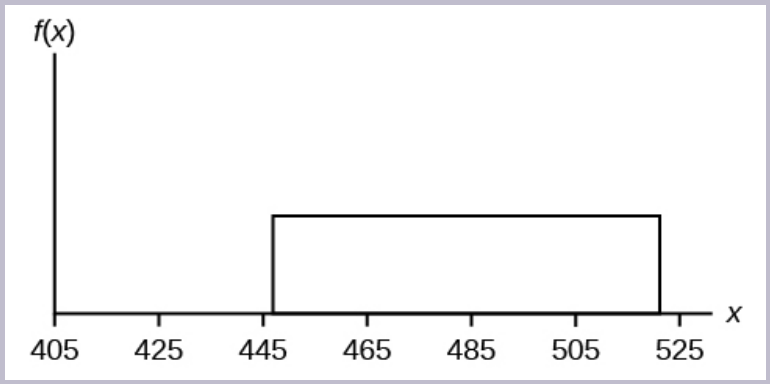
Your Turn!

The total duration of baseball games in the major league in the 2011 season is uniformly distributed between 447 hours and 521 hours inclusive.

1. Find *a* and *b* and describe what they represent.
2. Write the distribution.
3. Find the mean and the standard deviation.
4. What is the probability that the duration of games for a team for the 2011 season is between 480 and 500 hours?
5. What is the 65^{th} percentile for the duration of games for a team for the 2011 season?

Solution

1. *a* is 447, and *b* is 521. *a* is the minimum duration of games for a team for the 2011 season, and *b* is the maximum duration of games for a team for the 2011 season.
2. $X \sim U(447, 521)$.
3. $\mu = 484$, and $\sigma = 21.36$



4. $P(480 < x < 500) = 0.2703$
5. 65^{th} percentile is 495.1 hours.

Example

Suppose the time it takes a nine-year old to eat a donut is between 0.5 and 4 minutes, inclusive. Let X = the time, in minutes, it takes a nine-year old child to eat a donut. Then X ~ U (0.5, 4).

a. The probability that a randomly selected nine-year old child eats a donut in at least two minutes is _____.

Solution
a. 0.5714

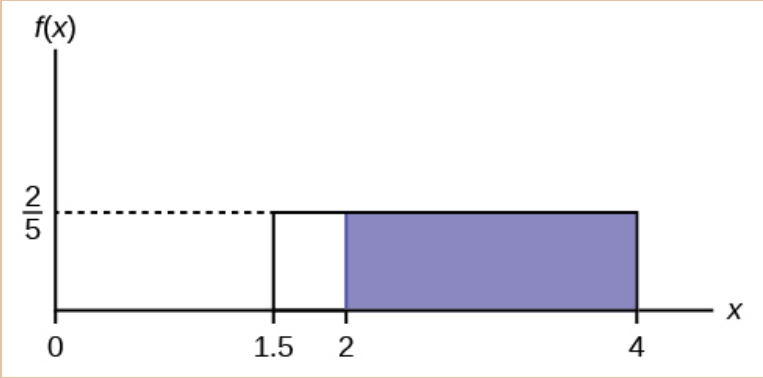
b. Find the probability that a different nine-year old child eats a donut in more than two minutes given that the child has already been eating the donut for more than 1.5 minutes.

The second question has a **conditional probability**. You are asked to find the probability that a nine-year old child eats a donut in more than two minutes given that the child has already been eating the donut for more than 1.5 minutes. Solve the problem two different ways. You must reduce the sample space. **First way**: Since you know the child has already been eating the donut for more than 1.5 minutes, you are no longer starting at $a = 0.5$ minutes. Your starting point is 1.5 minutes.

Write a new *f*(*x*):

$f(x) = \frac{1}{4 - 1.5} = \frac{2}{5}$ for $1.5 \leq x \leq 4$.

Find $P(x > 2|x > 1.5)$. Draw a graph.



$P(x > 2|x > 1.5)$ = (base)(new height) = $(4 - 2)\left(\frac{2}{5}\right)$ = ?

Solution

b. $\frac{4}{5}$

The probability that a nine-year old child eats a donut in more than two minutes given that the child has already been eating the donut for more than 1.5 minutes is $\frac{4}{5}$.

Second way: Draw the original graph for $X \sim U(0.5, 4)$. Use the conditional formula

$$P(x > 2|x > 1.5) = \frac{P(x > 2 \text{ AND } x > 1.5)}{P(x > 1.5)} = \frac{P(x > 2)}{P(x > 1.5)} = \frac{\frac{2}{3.5}}{\frac{2.5}{3.5}} = 0.8 = \frac{4}{5}$$

Your Turn!

Suppose the time it takes a student to finish a quiz is uniformly distributed between six and 15 minutes, inclusive. Let X = the time, in minutes, it takes a student to finish a quiz. Then $X \sim U(6, 15)$.

Find the probability that a randomly selected student needs at least eight minutes to complete the

quiz. Then find the probability that a different student needs at least eight minutes to finish the quiz given that she has already taken more than seven minutes.

Solution

$P(x > 8) = 0.7778$

$P(x > 8 \mid x > 7) = 0.875$

Example

Ace Heating and Air Conditioning Service finds that the amount of time a repairman needs to fix a furnace is uniformly distributed between 1.5 and four hours. Let x = the time needed to fix a furnace. Then x ~ U (1.5, 4).

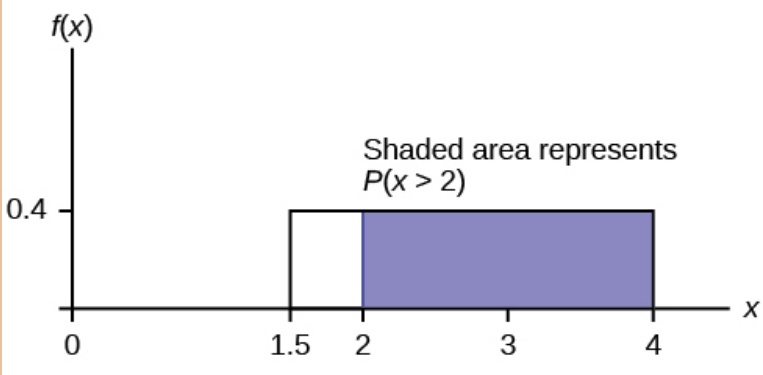
1. Find the probability that a randomly selected furnace repair requires more than two hours.
2. Find the probability that a randomly selected furnace repair requires less than three hours.
3. Find the 30th percentile of furnace repair times.
4. The longest 25% of furnace repair times take at least how long? (In other words: find the minimum time for the longest 25% of repair times.) What percentile does this represent?
5. Find the mean and standard deviation

Solution

a. To find f(x): f (x) = $\frac{1}{4 - 1.5} = \frac{1}{2.5}$ so f(x) = 0.4

P(x > 2) = (base)(height) = (4 – 2)(0.4) = 0.8

Uniform Distribution between 1.5 and four with shaded area between two and four representing the probability that the repair time x is greater than two.

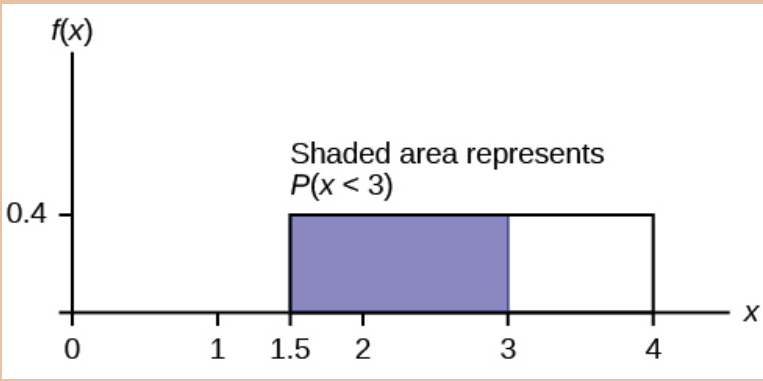


Solution

b. P(x < 3) = (base)(height) = (3 – 1.5)(0.4) = 0.6

The graph of the rectangle showing the entire distribution would remain the same. However the graph should be shaded between x = 1.5 and x = 3. Note that the shaded area starts at x = 1.5 rather than at x = 0; since X ~ U (1.5, 4), x can not be less than 1.5.

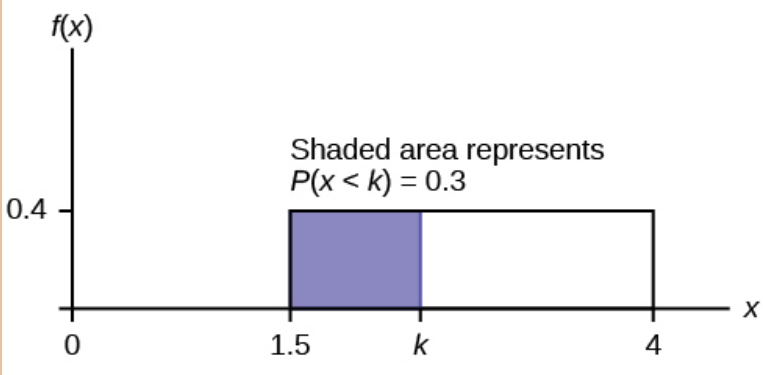
Uniform Distribution between 1.5 and four with shaded area between 1.5 and three representing the probability that the repair time x is less than three.



Solution

c.

Uniform Distribution between 1.5 and 4 with an area of 0.30 shaded to the left, representing the shortest 30% of repair times.



$P(x < k) = 0.30$

$P(x < k)$ = (base)(height) = $(k - 1.5)(0.4)$

$0.3 = (k - 1.5)(0.4)$; Solve to find k.

$0.75 = k - 1.5$, obtained by dividing both sides by 0.4

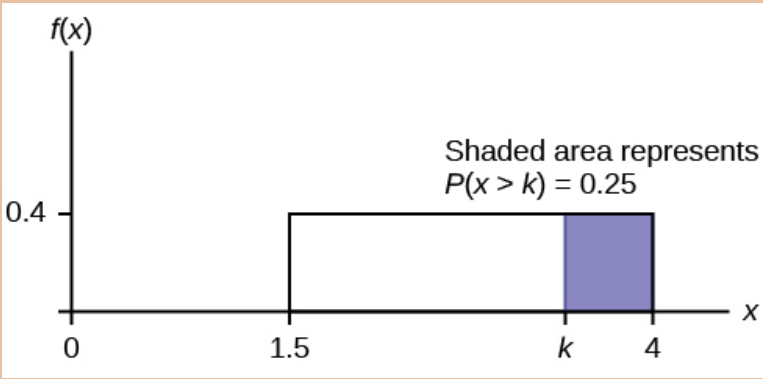
$k = 2.25$, obtained by adding 1.5 to both sides

The 30th percentile of repair times is 2.25 hours. 30% of repair times are 2.5 hours or less.

Solution

d.

Uniform Distribution between 1.5 and 4 with an area of 0.25 shaded to the right representing the longest 25% of repair times.



$P(x > k) = 0.25$

$P(x > k)$ = (base)(height) = $(4 - k)(0.4)$

$0.25 = (4 - k)(0.4)$; Solve for k.

$0.625 = 4 - k$,

obtained by dividing both sides by 0.4

$-3.375 = -k$,

obtained by subtracting four from both sides: ***k* = 3.375**

The longest 25% of furnace repairs take at least 3.375 hours (3.375 hours or longer).

Note: Since 25% of repair times are 3.375 hours or longer, that means that 75% of repair times are 3.375 hours or less. 3.375 hours is the **75th percentile** of furnace repair times.

Solution

e. $\mu = \frac{a+b}{2}$ and $\sigma = \sqrt{\frac{(b-a)^2}{12}}$

$$\mu = \frac{1.5+4}{2} = 2.75$$

hours and $\sigma = \sqrt{\frac{(4-1.5)^2}{12}} = 0.7217$ hours

Your Turn!

The amount of time a service technician needs to change the oil in a car is uniformly distributed between 11 and 21 minutes. Let X = the time needed to change the oil on a car.

1. Write the random variable X in words. X = _____.
2. Write the distribution.
3. Graph the distribution.
4. Find $P(x > 19)$.
5. Find the 50th percentile.

Solution

1. Let X = the time needed to change the oil in a car.
2. $X \sim U(11, 21)$.
3. Draw a graph that is shaded from 11 to 21.

4. $P(x > 19) = 0.2$
5. the 50th percentile is 16 minutes.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=769#oembed-1

HTTPS://WWW.YOUTUBE.COM/WATCH?V=44G51TLNCUY

5.3 THE EXPONENTIAL DISTRIBUTION

Learning Objectives

By the end of this section, the student should be able to:

- calculate exponential distribution and the probability density function.

The **exponential distribution** is often concerned with the amount of time until some specific event occurs. For example, the amount of time (beginning now) until an earthquake occurs has an **exponential distribution**. Other examples include the length, in minutes, of long distance business telephone calls, and the amount of time, in months, a car battery lasts. It can be shown, too, that the value of the change that you have in your pocket or purse approximately follows an **exponential distribution**.

Values for an exponential random variable occur in the following way. There are fewer large values and more small values. For example, the amount of money customers spend in one trip to the supermarket follows an **exponential distribution**. There are more people who spend small amounts of money and fewer people who spend large amounts of money.

The **exponential distribution** is widely used in the field of reliability. Reliability deals with the amount of time a product lasts.

Example

Let X = amount of time (in minutes) a postal clerk spends with his or her customer. The time is known to have an exponential distribution with the average amount of time equal to four minutes.

X is a continuous random variable since time is measured. It is given that $\mu = 4$ minutes. To do any calculations, you must know m, the **decay parameter**.

$m = \frac{1}{\mu}$. Therefore, $m = \frac{1}{4} = 0.25$.

The standard deviation, σ, is the same as the mean. $\mu = \sigma$

The distribution notation is $X \sim Exp(m)$. Therefore, $X \sim Exp(0.25)$.

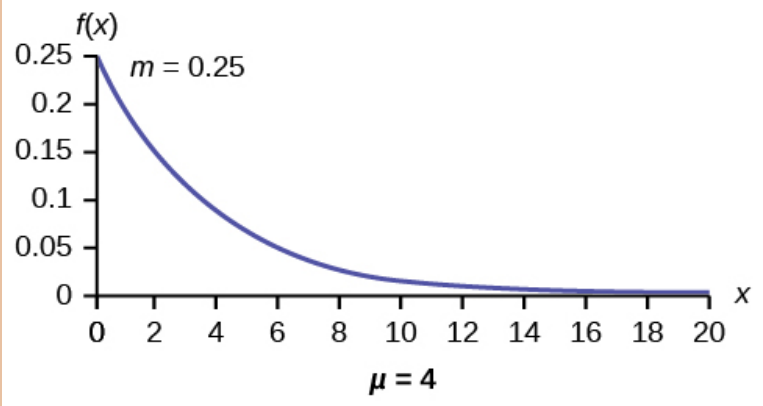
The probability density function is $f(x) = me - mx$. The number $e = 2.71828182846\ldots$ It is a number that is used often in mathematics. Scientific calculators have the key "e^x." If you enter one for x, the calculator will display the value e.

The curve is:

$f(x) = 0.25e{-}0.25x$ where x is at least zero and $m = 0.25$.

For example, $f(5) = 0.25e - (0.25)(5) = 0.072$. The postal clerk spends five minutes with the customers.

The graph is as follows:



Notice the graph is a declining curve. When $x = 0$,

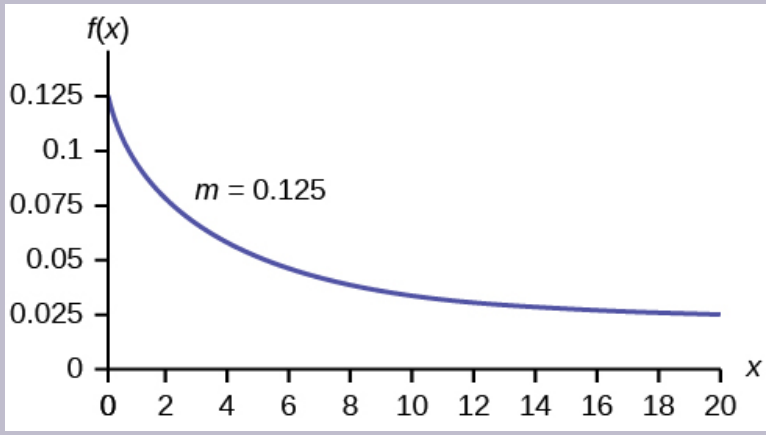
$f(x) = 0.25e(-0.25)(0) = (0.25)(1) = 0.25 = m$. The maximum value on the y-axis is m.

Your Turn!

The amount of time spouses shop for anniversary cards can be modeled by an exponential distribution with the average amount of time equal to eight minutes. Write the distribution, state the probability density function, and graph the distribution.

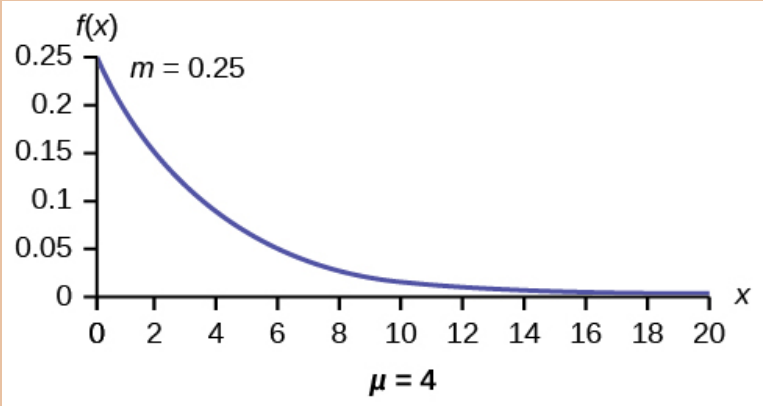
Solution

$X \sim Exp(0.125)$; $f(x) = 0.125e^{-0.125x}$;



Example

a. Using the information in the graph from the first example (shown below), find the probability that a clerk spends four to five minutes with a randomly selected customer.



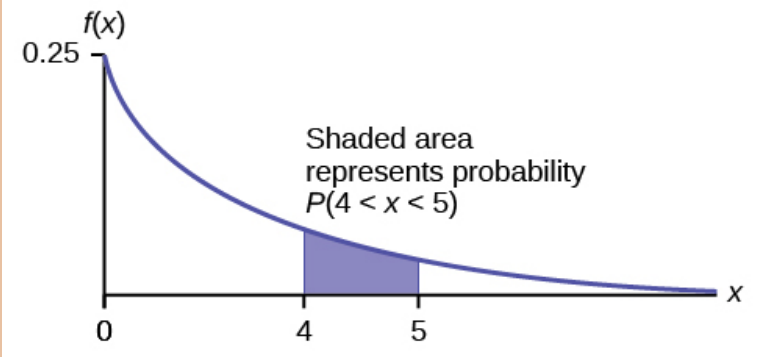
Solution

a. Find $P(4 < x < 5)$.

The cumulative distribution function (CDF) gives the area to the left.

$P(x < x) = 1 - e - mx$

$P(x < 5) = 1 - e(-0.25)(5) = 0.7135$ and $P(x < 4) = 1 - e(-0.25)(4) = 0.6321$



NOTE

You can do these calculations easily on a calculator.

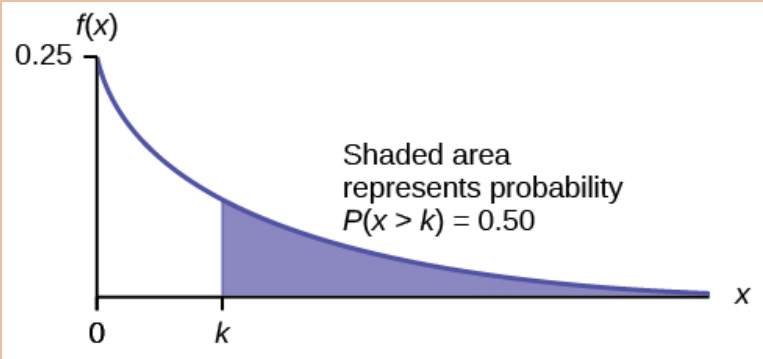
The probability that a postal clerk spends four to five minutes with a randomly selected customer is $P(4 < x < 5) = P(x < 5) - P(x < 4) = 0.7135 - 0.6321 = 0.0814$.

On the home screen, enter (1 – e^(–0.25*5))–(1–e^(–0.25*4)) or enter e^(–0.25*4) – e^(–0.25*5).

b. Half of all customers are finished within how long? (Find the 50th percentile)

Solution

b. Find the 50th percentile.



$P(x < k) = 0.50, k = 2.8$ minutes (calculator or computer)

Half of all customers are finished within 2.8 minutes.

You can also do the calculation as follows:

$P(x < k) = 0.50$ and $P(x < k) = 1 - e - 0.25k$

Therefore, $0.50 = 1 - e - 0.25k$ and $e - 0.25k = 1 - 0.50 = 0.5$

Take natural logs: $ln(e - 0.25k) = ln(0.50)$. So, $-0.25k = ln(0.50)$

Solve for k: $k = \frac{ln(0.50)}{-0.25} = 2.8$ minutes. The calculator simplifies the calculation for percentile k. See the following two notes.

Note

A formula for the percentile k is $k = \frac{ln(1 - AreaToTheLeft)}{-m}$ where ln is the natural log.

On the home screen, enter ln(1 – 0.50)/–0.25. Press the (-) for the negative.

c. Which is larger, the mean or the median?

Solution

c. From part b, the median or 50th percentile is 2.8 minutes. The theoretical mean is four minutes. The mean is larger.

Your Turn!

The number of days ahead travelers purchase their airline tickets can be modeled by an exponential distribution with the average amount of time equal to 15 days. Find the probability that a traveler will purchase a ticket fewer than ten days in advance. How many days do half of all travelers wait?

Solution

$P(x < 10) = 0.4866$

50^{th} percentile = 10.40

Example

On the average, a certain computer part lasts ten years. The length of time the computer part lasts is exponentially distributed.

a. What is the probability that a computer part lasts more than 7 years?

Solution

a. Let x = the amount of time (in years) a computer part lasts.

$\mu = 10$ so $m = \frac{1}{\mu} = \frac{1}{10} = 0.1$

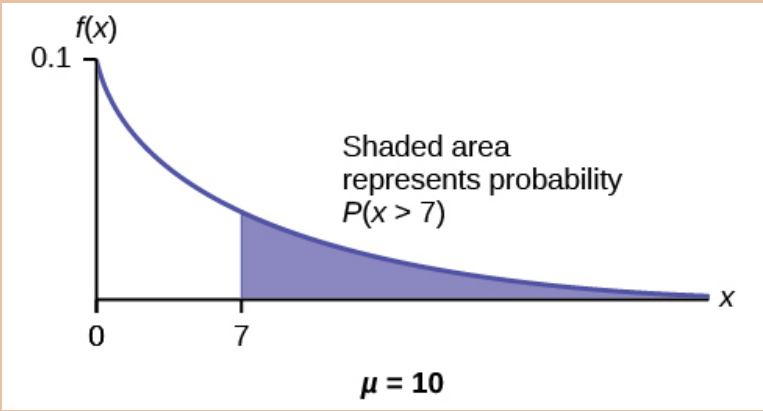
Find $P(x > 7)$. Draw the graph.

$P(x > 7) = 1 - P(x < 7)$.

Since $P(X < x) = 1 - e - mx$ then $P(X > x) = 1 - (1 - e - mx) = e - mx$

$P(x > 7) = e(-0.1)(7) = 0.4966$. The probability that a computer part lasts more than seven years is 0.4966.

On the home screen, enter e^(-.1*7).



b. On the average, how long would five computer parts last if they are used one after another?

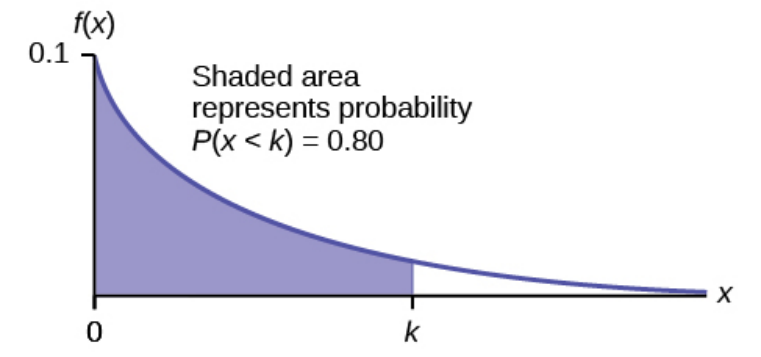
Solution

b. On the average, one computer part lasts ten years. Therefore, five computer parts, if they are used one right after the other, would last, on the average, (5)(10) = 50 years.

c. Eighty percent of computer parts last at most how long?

Solution

c. Find the 80th percentile. Draw the graph. Let k = the 80th percentile.



Solve for k: $k = \frac{ln\,(1-0.80)}{-0.1} = 16.1$ years

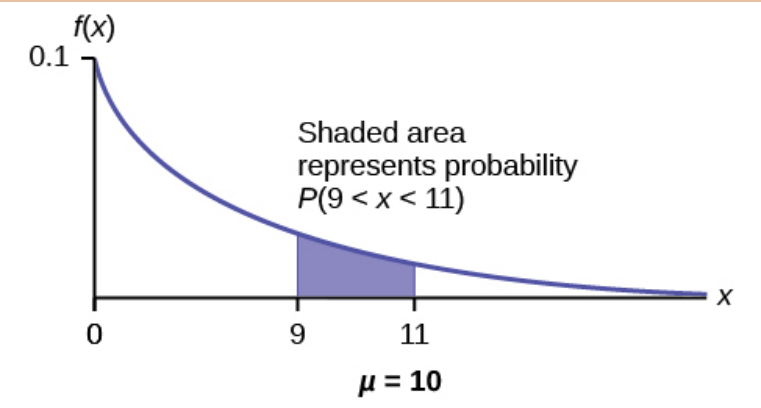
Eighty percent of the computer parts last at most 16.1 years.

On the home screen, enter $\frac{\ln\,(1-0.80)}{-0.1}$

d. What is the probability that a computer part lasts between nine and 11 years?

Solution

d. Find $P(9 < x < 11)$. Draw the graph.



$P(9 < x < 11) = P(x < 11) - P(x < 9) = (1 - e(-0.1)(11)) - (1 - e(-0.1)(9)) = 0.6671 - 0.5934 = 0.0737$. The probability that a computer part lasts between nine and 11 years is 0.0737.

On the home screen, enter e^(–0.1*9) – e^(–0.1*11).

Your Turn!

On average, a pair of running shoes can last 18 months if used every day. The length of time running shoes last is exponentially distributed. What is the probability that a pair of running shoes last more than 15 months? On average, how long would six pairs of running shoes last if they are used one after the other? Eighty percent of running shoes last at most how long if used every day?

Solution

$P(x > 15) = 0.4346$

Six pairs of running shoes would last 108 months on average.

80^{th} percentile = 28.97 months

Example

Suppose that the length of a phone call, in minutes, is an exponential random variable with decay parameter = $\frac{1}{12}$. If another person arrives at a public telephone just before you, find the probability that you will have to wait more than five minutes. Let X = the length of a phone call, in minutes.

What is m, μ, and σ? The probability that you must wait more than five minutes is _____ .

Solution

- $m = \frac{1}{12}$
- $\mu = 12$
- $\sigma = 12$

$P(x > 5) = 0.6592$

Your Turn!

Suppose that the distance, in miles, that people are willing to commute to work is an exponential random variable with a decay parameter $\frac{1}{20}$. Let X = the distance people are willing to commute in miles. What is m, μ, and σ? What is the probability that a person is willing to commute more than 25 miles?

Solution

$m = \frac{1}{20}; \mu = 20; \sigma = 20; P(x > 25) = 0.2865$

Example

The time spent waiting between events is often modeled using the exponential distribution. For example, suppose that an average of 30 customers per hour arrive at a store and the time between arrivals is exponentially distributed.

1. On average, how many minutes elapse between two successive arrivals?

2. When the store first opens, how long on average does it take for three customers to arrive?
3. After a customer arrives, find the probability that it takes less than one minute for the next customer to arrive.
4. After a customer arrives, find the probability that it takes more than five minutes for the next customer to arrive.
5. Seventy percent of the customers arrive within how many minutes of the previous customer?
6. Is an exponential distribution reasonable for this situation?

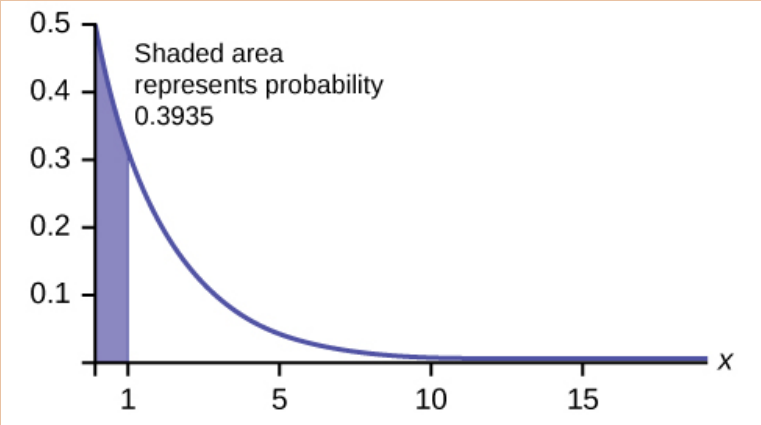
Solution

1. Since we expect 30 customers to arrive per hour (60 minutes), we expect on average one customer to arrive every two minutes on average.
2. Since one customer arrives every two minutes on average, it will take six minutes on average for three customers to arrive.
3. Let X = the time between arrivals, in minutes. By part a, $/mu = 2$, so $m = \frac{1}{2}$ = 0.5.

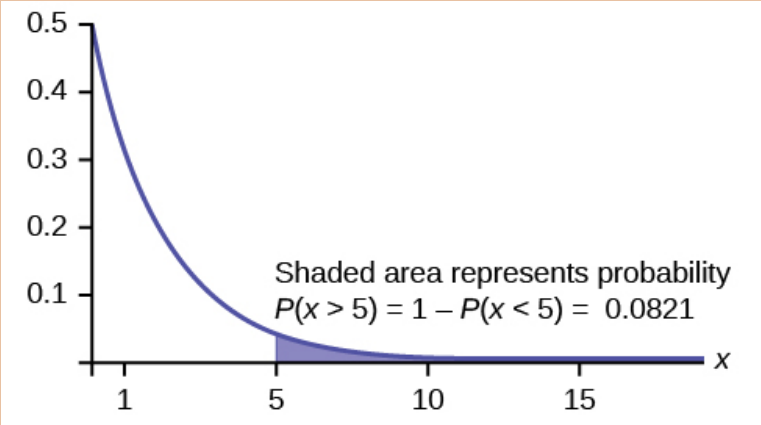
   Therefore, $X \sim Exp(0.5)$.

   The cumulative distribution function is $P(X < x) = 1 - e(-0.5x)e$.

   Therefore $P(X < 1) = 1 - e(-0.5)(1) \approx 0.3935$.



4. $P(X > 5) = 1 - P(X < 5) = 1 - (1 - e(-5)(0.5)) = e - 2.5 \approx 0.0821$.



5. We want to solve $0.70 = P(X < x)$ for x.

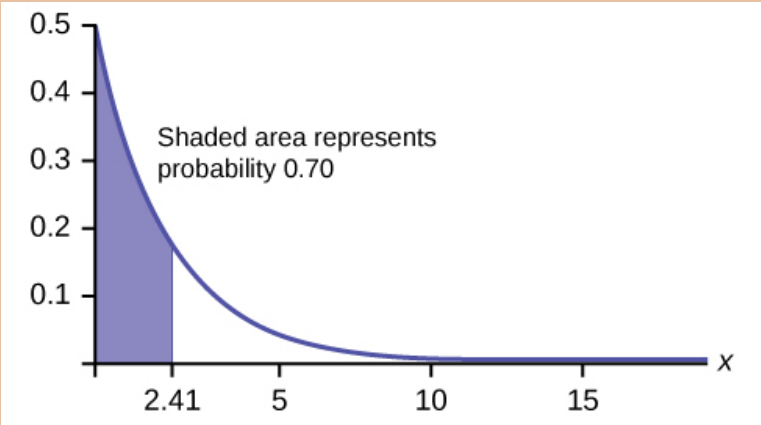
   Substituting in the cumulative distribution function gives $0.70 = 1 - e - 0.5x$, so that $e - 0.5x = 0.30$. Converting this to logarithmic form gives $-0.5x = ln(0.30)$, or $x = \frac{ln\,(0.30)}{-0.5} \approx 2.41$ minutes.

   Thus, seventy percent of customers arrive within 2.41 minutes of the previous customer.

   You are finding the 70th percentile k so you can use the formula

   $k = \frac{ln\,(1 - Area_T o_T he_L eft_O f_k)}{(-m)}$

   $k = \frac{ln\,(1 - 0.70)}{(-0.5)} \approx 2.41$ minutes



6. This model assumes that a single customer arrives at a time, which may not be reasonable since people might shop in groups, leading to several customers arriving at the same time. It also assumes that the flow of customers does not change throughout the day, which is not valid if some times of the day are busier than others.

Your Turn!

Suppose that on a certain stretch of highway, cars pass at an average rate of five cars per minute. Assume that the duration of time between successive cars follows the exponential distribution.

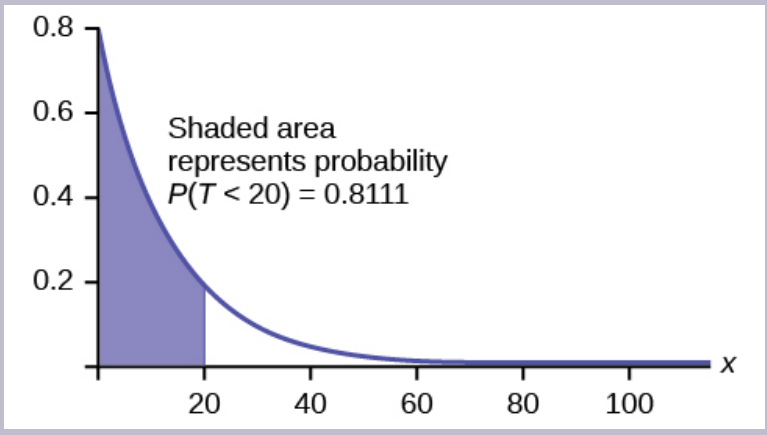
1. On average, how many seconds elapse between two successive cars?
2. After a car passes by, how long on average will it take for another seven cars to pass by?
3. Find the probability that after a car passes by, the next car will pass within the next 20 seconds.
4. Find the probability that after a car passes by, the next car will not pass for at least another 15 seconds.

Solution

1. At a rate of five cars per minute, we expect $\frac{60}{5}$ = 12 seconds to pass between successive cars on average.
2. Using the answer from part a, we see that it takes (12)(7) = 84 seconds for the next seven cars to pass by.
3. Let T = the time (in seconds) between successive cars.

   The mean of T is 12 seconds, so the decay parameter is $\frac{1}{12}$ and $T \sim Exp(\frac{1}{12})$. The cumulative distribution function of T is $P(T < t) = 1 - e^{-\frac{t}{12}}$. Then

   $$P(T < 20) = 1 - e^{-\frac{20}{12}} \approx 0.8111.$$



$$P(T > 15) = 1 - P(T < 15) = 1 - (1 - e^{-\frac{15}{12}}) = e^{-\frac{15}{12}} \approx 0.2865.$$

Memorylessness of the Exponential Distribution

In the first example, recall that the amount of time between customers is exponentially distributed with a mean of two minutes ($X \sim Exp(0.5)$). Suppose that five minutes have elapsed since the last customer arrived. Since an unusually long amount of time has now elapsed, it would seem to be more likely for a customer to arrive within the next minute. With the exponential distribution, this is not the case–the additional time spent waiting for the next customer does not depend on how much time has already elapsed since the last customer. This is referred to as the **memoryless property**. Specifically, the **memoryless property** says that

$$P(X > r + t | X > r) = P(X > t) \text{ for all } r \geq 0 and t \geq 0$$

For example, if five minutes has elapsed since the last customer arrived, then the probability that more than one minute will elapse before the next customer arrives is computed by using $r = 5$ and $t = 1$ in the foregoing equation.

P(X > 5 + 1 | X > 5) = P(X > 1) = {e}^{(–0.5)(1)} \approx{0.6065}

This is the same probability as that of waiting more than one minute for a customer to arrive after the previous arrival.

The **exponential distribution** is often used to model the longevity of an electrical or mechanical device. In [link], the lifetime of a certain computer part has the exponential distribution with a mean of ten years ($X \sim Exp(0.1)$). The **memoryless property** says that knowledge of what has occurred in the past has no effect on future probabilities. In this case it means that an old part is not any more likely to break down at any particular time than a brand new part. In other words, the part stays as good as new until it suddenly breaks. For example, if the part has already lasted ten years, then the probability that it lasts another seven years is $P(X > 17 | X > 10) = P(X > 7) = 0.4966$.

Example

Refer to [link] where the time a postal clerk spends with his or her customer has an exponential distribution with a mean of four minutes. Suppose a customer has spent four minutes with a postal

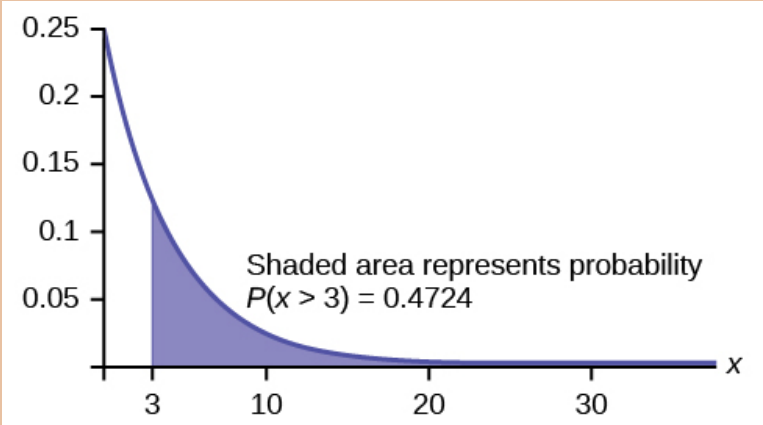
clerk. What is the probability that he or she will spend at least an additional three minutes with the postal clerk?

The decay parameter of X is m = \frac{1}{4} = 0.25, so $X \sim Exp(0.25)$.

The cumulative distribution function is $P(X < x) = 1 - e - 0.25x$.

We want to find $P(X > 7|X > 4)$. The memoryless property says that $P(X > 7|X > 4) = P(X > 3)$, so we just need to find the probability that a customer spends more than three minutes with a postal clerk.

This is $P(X > 3) = 1 - P(X < 3) = 1 - (1 - e - 0.25 \cdot 3) = e - 0.75 \approx\approx 0.4724$.



$1 - (1 - e^{-0.25*2}) = e^{-0.25*2}$

Your Turn!

Suppose that the longevity of a light bulb is exponential with a mean lifetime of eight years. If a bulb has already lasted 12 years, find the probability that it will last a total of over 19 years.

Solution

Let *T* = the lifetime of the light bulb. Then $T \sim Exp(\frac{1}{8})$.

The cumulative distribution function is $P(T < t) = 1 - e^{-\frac{t}{8}}$.

We need to find $P(T > 19|T = 12)$. By the **memoryless property**,

$P(T > 19|T = 12) = P(T > 7) = 1 - P(T < 7) = 1 - (1 - e^{-\frac{7}{8}}) = e^{-\frac{7}{8}} \approx 0.4169$

.

Relationship between the Poisson and the Exponential Distribution

There is an interesting relationship between the **exponential distribution** and the **Poisson distribution**. Suppose that the time that elapses between two successive events follows the **exponential distribution** with a mean of μ units of time. Also assume that these times are **independent**, meaning that the time between events is not affected by the times between previous events. If these assumptions hold, then the number of events per unit time follows a **Poisson distribution** with mean $\lambda = \frac{1}{\mu}$. Recall from the chapter on Discrete Random Variables that if X has the **Poisson distribution** with mean λ, then $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$. $(k! = k \cdot (k-1) \cdot (k-2) \cdot (k-3) \ldots 3 \cdot 2 \cdot 1)$ Conversely, if the number of events per unit time follows a **Poisson distribution**, then the amount of time between events follows the **exponential distribution**.

Suppose X has the **Poisson distribution** with mean λ. Compute $P(X = k)$ by entering 2nd, VARS(DISTR), C: poissonpdf(λ, k). To compute $P(X \leq k)$, enter 2nd, VARS (DISTR), D:poissoncdf(λ, k).

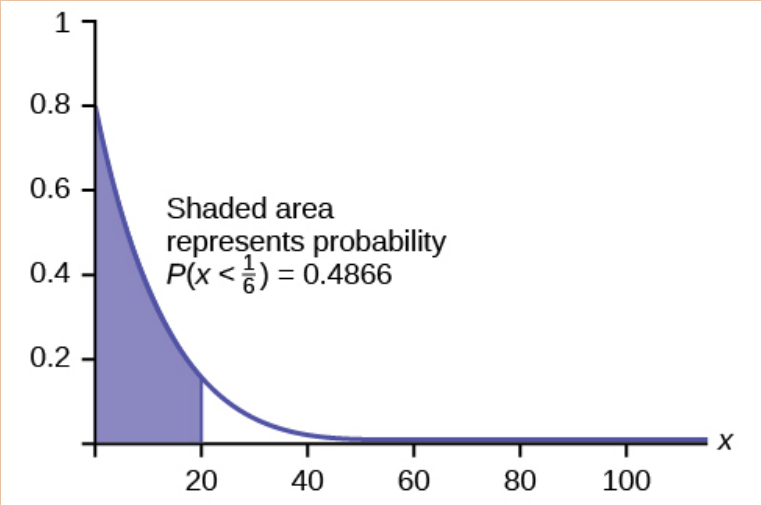
Example

At a police station in a large city, calls come in at an average rate of four calls per minute. Assume that the time that elapses from one call to the next has the exponential distribution. Take note that we are concerned only with the rate at which calls come in, and we are ignoring the time spent on the phone. We must also assume that the times spent between calls are independent. This means that a particularly long delay between two calls does not mean that there will be a shorter waiting period for the next call. We may then deduce that the total number of calls received during a time period has the Poisson distribution.

1. Find the average time between two successive calls.
2. Find the probability that after a call is received, the next call occurs in less than ten seconds.
3. Find the probability that exactly five calls occur within a minute.
4. Find the probability that less than five calls occur within a minute.
5. Find the probability that more than 40 calls occur in an eight-minute period.

Solution

1. On average there are four calls that occur per minute, so 15 seconds, or $\frac{15}{60}$ = 0.25 minutes, occur between successive calls on average.
2. Let T = time elapsed between calls. From part a, μ = 0.25, so m = $\frac{1}{0.25}$ = 4. Thus, T ~ Exp(4).
   The cumulative distribution function is P(T < t) = 1 – e–4t.
   The probability that the next call occurs in less than ten seconds (ten seconds = 1/6 minute) is
   $$P\left(T < \frac{1}{6}\right) = 1 - e^{-4\frac{1}{6}} \approx 0.4866.$$



3. Let X = the number of calls per minute. As previously stated, the number of calls per minute has a Poisson distribution, with a mean of four calls per minute.

   Therefore, X ~ Poisson(4), and so P(X = 5) = $\frac{4^5 e^{-4}}{5!}$ ≈ 0.1563. (5! = (5)(4)(3)(2)(1))

   poissonpdf(4, 5) = 0.1563.

4. Keep in mind that X must be a whole number, so P(X < 5) = P(X ≤ 4).
   To compute this, we could take P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4).

   Using technology, we see that P(X ≤ 4) = 0.6288.

   poisssoncdf(4, 4) = 0.6288

5. Let Y = the number of calls that occur during an eight minute period.
   Since there is an average of four calls per minute, there is an average of (8)(4) = 32 calls during each eight minute period.

   Hence, Y ~ Poisson(32). Therefore, P(Y > 40) = 1 – P (Y ≤ 40) = 1 – 0.9294 = 0.0707.

   1 – poissoncdf(32, 40). = 0.0707

Your Turn!

In a small city, the number of automobile accidents occur with a Poisson distribution at an average of three per week.

1. Calculate the probability that there are at most 2 accidents that occur in any given week.
2. What is the probability that there is at least two weeks between any 2 accidents?

Solution

1. Let X = the number of accidents per week, so that $X \sim Poisson(3)$. We need to find $P(X \leq 2) \approx 0.4232$

   poissoncdf(3, 2)

2. Let T = the time (in weeks) between successive accidents.
   Since the number of accidents occurs with a Poisson distribution, the time between accidents follows the exponential distribution.

   If there are an average of three per week, then on average there is $\mu = \frac{1}{3}$ of a week between accidents, and the decay parameter is $m = \frac{1}{\left(\frac{1}{3}\right)}$ = 3.

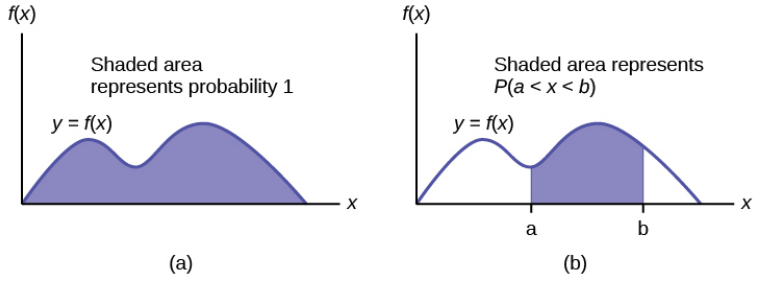
   To find the probability that there are at least two weeks between two accidents, $P(T > 2) = 1 - P(T < 2) = 1 - (1 - e(-3)(2)) = e^{-6} \approx 0.0025$.

   e^(-3*2).

CHAPTER 5 REVIEW

5.1 Review

The probability density function (pdf) is used to describe probabilities for continuous random variables. The area under the density curve between two points corresponds to the probability that the variable falls between those two values. In other words, the area under the density curve between points a and b is equal to $P(a < x < b)$. The cumulative distribution function (cdf) gives the probability as an area. If X is a continuous random variable, the probability density function (pdf), $f(x)$, is used to draw the graph of the probability distribution. The total area under the graph of $f(x)$ is one. The area under the graph of $f(x)$ and between values a and b gives the probability $P(a < x < b)$.



The cumulative distribution function (cdf) of X is defined by $P(X \leq x)$. It is a function of x that gives the probability that the random variable is less than or equal to x.

Formula Review

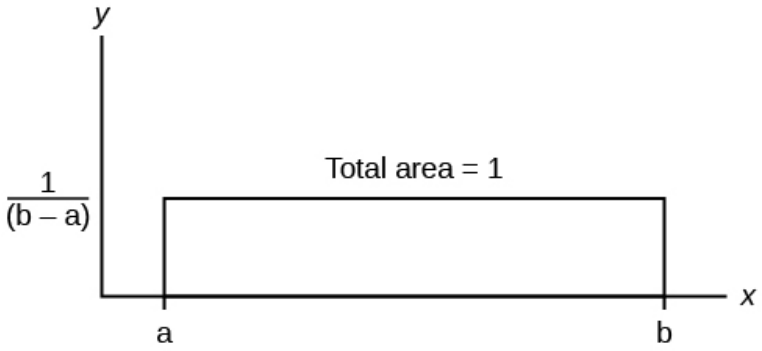
Probability density function (pdf) $f(x)$:

- $f(x) \geq 0$
- The total area under the curve $f(x)$ is one.

Cumulative distribution function (cdf): $P(X \leq x)$

5.2 Review

If X has a uniform distribution where $a < x < b$ or $a \leq x \leq b$, then X takes on values between a and b (may include a and b). All values x are equally likely. We write $X \sim U(a, b)$. The mean of X is $\mu = \frac{a+b}{2}$. The standard deviation of X is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$. The probability density function of X is $f(x) = \frac{1}{b-a}$ for $a \leq x \leq b$. The cumulative distribution function of X is P(X \le x) = \frac{x-a}{b-a}. X is continuous.



The probability $P(c < X < d)$ may be found by computing the area under $f(x)$, between c and d. Since the corresponding area is a rectangle, the area may be found simply by multiplying the width and the height.

Formula Review

X = a real number between a and b (in some instances, X can take on the values a and b). a = smallest X; b = largest X

$X \sim U(a, b)$

The mean is $\mu = \frac{a+b}{2}$

The standard deviation is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$

Probability density function: $f(x) = \frac{1}{b-a}$ for $a \leq X \leq b$

Area to the Left of x: P(X \lt x) = (x – a)(\frac{1}{b-a})

Area to the Right of x: P(X > x) = (b – x)(\frac{1}{b-a})

Area Between *c* and *d*: P(c \lt x \lt d) = (\text{base})(\text{height}) = (d – c) (\frac{1}{b-a})

Uniform: $X \sim U(a, b)$ where $a < x < b$

- pdf: $f(x) = \frac{1}{b-a}$ for $a \leq x \leq b$
- cdf: P(X \le x) =\frac{x-a}{b-a}
- mean: $\mu = \frac{a+b}{2}$
- standard deviation: $\sigma = \sqrt{\frac{(b-a)^2}{12}}$
- P(c \lt X \lt d) = (d – c)\left(\frac{1}{b–a}\right)

5.3 Review

If X has an **exponential distribution** with mean μ, then the **decay parameter** is $m = \frac{1}{\mu}$, and we write $X \sim Exp(m)$ where $x \geq 0$ and $m > 0$. The probability density function of X is $f(x) = me^{-mx}$ (or equivalently $f(x) = \frac{1}{\mu}e^{-x/\mu}$). The cumulative distribution function of X is $P(X \leq x) = 1 - e^{-mx}$.

The exponential distribution has the **memoryless property**, which says that future probabilities do not depend on any past information. Mathematically, it says that $P(X > x + k|X > x) = P(X > k)$.

If T represents the waiting time between events, and if $T \sim Exp(\lambda)$, then the number of events X per unit time follows the Poisson distribution with mean λ. The probability density function of P(X) is $(X = k) = \frac{\lambda^k e^{-k}}{k!}$. This may be computed using a TI-83, 83+, 84, 84+ calculator with the command poissonpdf(λ, k). The cumulative distribution function $P(X \leq k)$ may be computed using the TI-83, 83+,84, 84+ calculator with the command poissoncdf(λ, k).

Formula Review

Exponential: $X \sim Exp(m)$ where m = the decay parameter

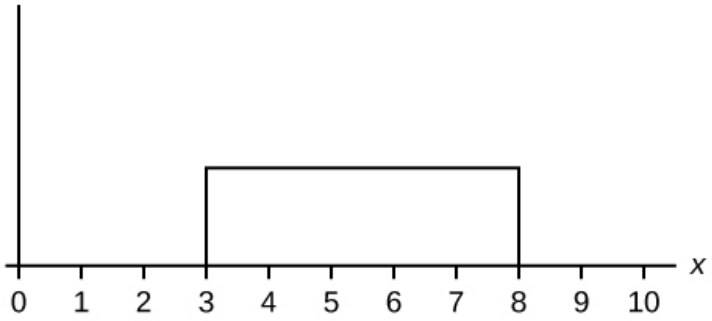
- pdf: $f(x) = me^{(-mx)}$ where $x \geq 0$ and $m > 0$
- cdf: $P(X \leq x) = 1 - e^{-mx}$

- mean: $\mu = \dfrac{1}{m}$
- standard deviation $\sigma = \mu$
- percentile k: $k = \dfrac{ln\left(1 - AreaToTheLeftOfk\right)}{(-m)}$
- Additionally
  - $P(X > x) = e^{(-mx)}$
  - $P(a < X < b) = e^{(-ma)} - e^{(-mb)}$
- Memoryless Property: $P(X > x + k | X > x) = P(X > k)$
- Poisson probability: $P(X = k) = \dfrac{\lambda^k e^{-k}}{k!}$ with mean λ
- $k! = k(k-1)(k-2)(k-3)\ldots 3 \cdot 2 \cdot 1$

CHAPTER 5 PRACTICE

5.1 Practice

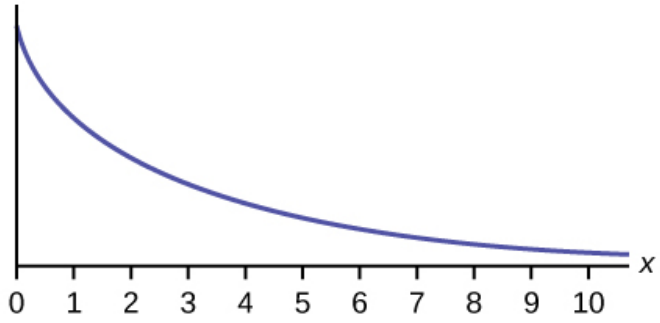
Which type of distribution does the graph illustrate?



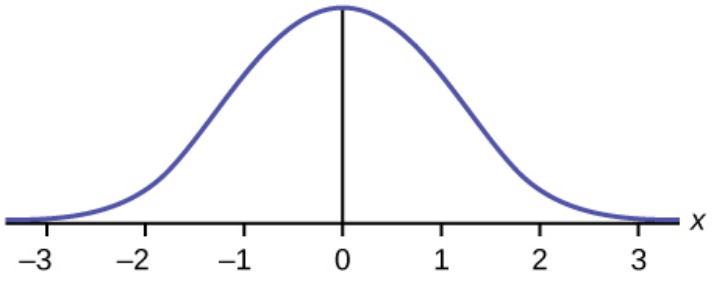
Solution

Uniform Distribution

Which type of distribution does the graph illustrate?



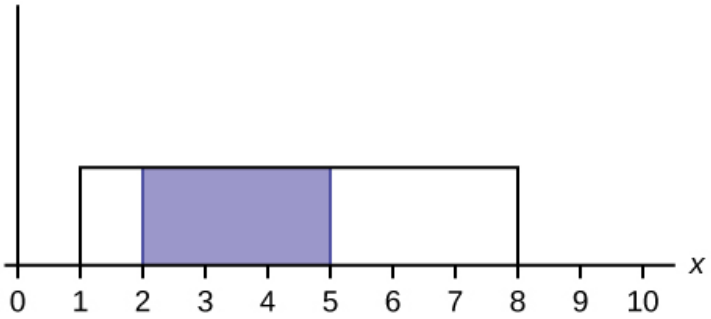
Which type of distribution does the graph illustrate?



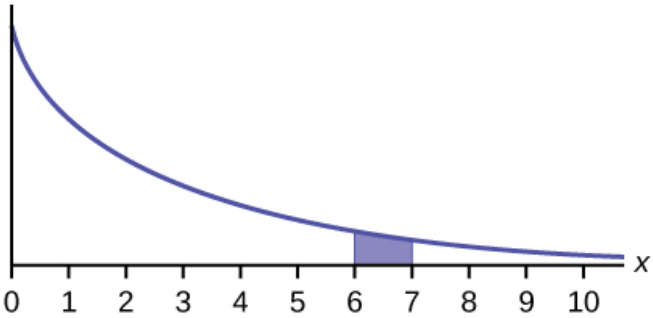
Solution

Normal Distribution

What does the shaded area represent? $P(___ < x < ___)$



What does the shaded area represent? $P(___ < x < ___)$



Solution

$P(6 < x < 7)$

For a continuous probability distribution, $0 \le x \le 15$. What is $P(x > 15)$?

What is the area under $f(x)$ if the function is a continuous probability density function?
Solution

one

For a continuous probability distribution, $0 \le x \le 10$. What is $P(x = 7)$?

A **continuous** probability function is restricted to the portion between $x = 0$ and 7. What is $P(x = 10)$?
Solution

zero

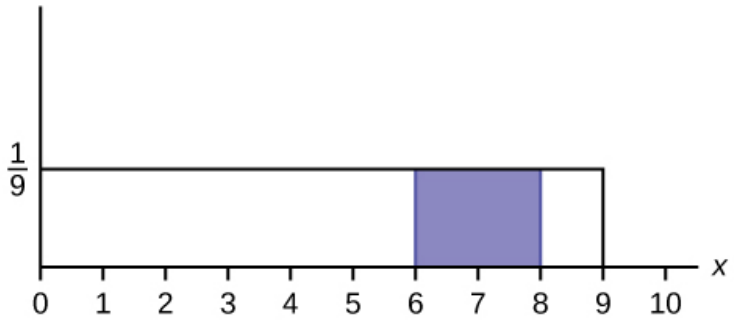
$f(x)$ for a continuous probability function is $\frac{1}{5}$, and the function is restricted to $0 \le x \le 5$. What is $P(x < 0)$?

$f(x)$, a continuous probability function, is equal to $\frac{1}{12}$, and the function is restricted to $0 \leq x \leq 12$. What is $P(0 < x < 12)$?

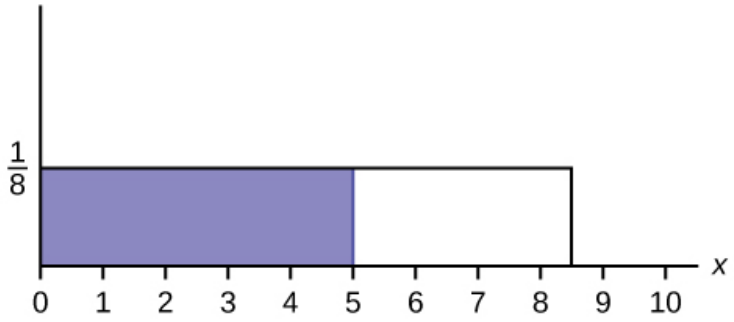
Solution

one

Find the probability that x falls in the shaded area.



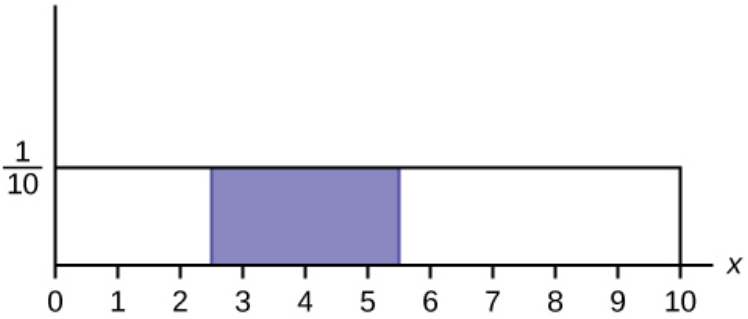
Find the probability that x falls in the shaded area.



Solution

0.625

Find the probability that x falls in the shaded area.



$f(x)$, a continuous probability function, is equal to $\frac{1}{3}$ and the function is restricted to $1 \leq x \leq 4$. Describe $P\left(x > \frac{3}{2}\right)$.

Solution

The probability is equal to the area from $x = \frac{3}{2}$ to $x = 4$ above the x-axis and up to $f(x) = \frac{1}{3}$.

5.2 Practice

McDougall, John A. *The McDougall Program for Maximum Weight Loss*. Plume, 1995.

Use the following information to answer the next ten questions. The data that follow are the square footage (in 1,000 feet squared) of 28 homes.

1.5	2.4	3.6	2.6	1.6	2.4	2.0
3.5	2.5	1.8	2.4	2.5	3.5	4.0
2.6	1.6	2.2	1.8	3.8	2.5	1.5
2.8	1.8	4.5	1.9	1.9	3.1	1.6

The sample mean = 2.50 and the sample standard deviation = 0.8302.

The distribution can be written as $X \sim U(1.5, 4.5)$.

What type of distribution is this?

In this distribution, outcomes are equally likely. What does this mean?

Solution

It means that the value of x is just as likely to be any number between 1.5 and 4.5.

References

What is the height of $f(x)$ for the continuous probability distribution?

What are the constraints for the values of x?

Solution

$1.5 \leq x \leq 4.5$

Graph $P(2 < x < 3)$.

What is $P(2 < x < 3)$?
Solution

0.3333

What is $P(x < 3.5 | x < 4)$?

What is $P(x = 1.5)$?
Solution

zero

What is the 90^{th} percentile of square footage for homes?

Find the probability that a randomly selected home has more than 3,000 square feet given that you already know the house has more than 2,000 square feet.
Solution

0.6

Use the following information to answer the next eight exercises. A distribution is given as $X \sim U(0, 12)$.

What is a? What does it represent?

What is b? What does it represent?
Solution

b is 12, and it represents the highest value of x.

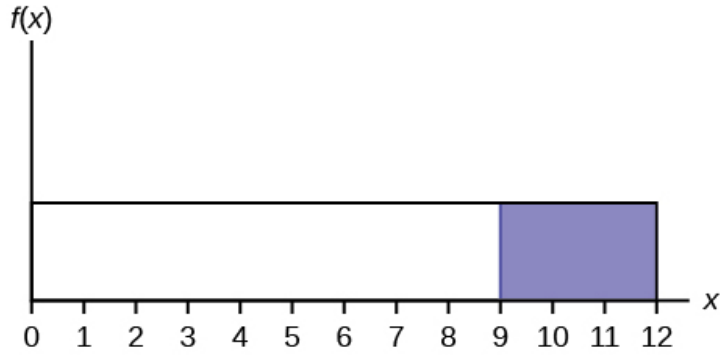
What is the probability density function?

What is the theoretical mean?
Solution

six

What is the theoretical standard deviation?

Draw the graph of the distribution for $P(x > 9)$.
Solution



Find $P(x > 9)$.

Find the 40^{th} percentile.
Solution

4.8

Use the following information to answer the next eleven exercises. The age of cars in the staff parking lot of a suburban college is uniformly distributed from six months (0.5 years) to 9.5 years.

What is being measured here?

In words, define the random variable X.
Solution

X = The age (in years) of cars in the staff parking lot

Are the data discrete or continuous?

The interval of values for x is _____.
Solution

0.5 to 9.5

The distribution for X is _____.

Write the probability density function.
Solution

$f(x) = \frac{1}{9}$ where x is between 0.5 and 9.5, inclusive.

Graph the probability distribution.

1. Sketch the graph of the probability distribution.



2. Identify the following values:
    1. Lowest value for $\overline{x}$: _____
    2. Highest value for $\overline{x}$: _____
    3. Height of the rectangle: _____
    4. Label for x-axis (words): _____
    5. Label for y-axis (words): _____

Find the average age of the cars in the lot.
Solution

$\mu = 5$

Find the probability that a randomly chosen car in the lot was less than four years old.

1. Sketch the graph, and shade the area of interest.



2. Find the probability. $P(x < 4)$ = _____

Considering only the cars less than 7.5 years old, find the probability that a randomly chosen car in the lot was less than four years old.

1. Sketch the graph, shade the area of interest.



2. Find the probability. $P(x < 4|x < 7.5)$ = _____

Solution

1. Check student's solution.
2. $\frac{3.5}{7}$

What has changed in the previous two problems that made the solutions different?

Find the third quartile of ages of cars in the lot. This means you will have to find the value such that $\frac{3}{4}$, or 75%, of the cars are at most (less than or equal to) that age.

1. Sketch the graph, and shade the area of interest.



2. Find the value k such that $P(x < k) = 0.75$.
3. The third quartile is _____

Solution

1. Check student's solution.
2. $k = 7.25$
3. 7.25

5.2 Practice

References

Data from the United States Census Bureau.

Data from World Earthquakes, 2013. Available online at http://www.world-earthquakes.com/ (accessed June 11, 2013).

"No-hitter." Baseball-Reference.com, 2013. Available online at http://www.baseball-reference.com/bullpen/No-hitter (accessed June 11, 2013).

Zhou, Rick. "Exponential Distribution lecture slides." Available online at www.public.iastate.edu/~riczw/stat330s11/lecture/lec13.pdf (accessed June 11, 2013).

Use the following information to answer the next ten exercises. A customer service representative must spend different amounts of time with each customer to resolve various concerns. The amount of time spent with each customer can be modeled by the following distribution: $X \sim Exp(0.2)$

What type of distribution is this?

Are outcomes equally likely in this distribution? Why or why not?
Solution

No, outcomes are not equally likely. In this distribution, more people require a little bit of time, and fewer people require a lot of time, so it is more likely that someone will require less time.

What is m? What does it represent?

What is the mean?
Solution

five

What is the standard deviation?

State the probability density function.
Solution

$f(x) = 0.2e^{-0.2x}$

Graph the distribution.

Find $P(2 < x < 10)$.
Solution

0.5350

Find $P(x > 6)$.

Find the 70^{th} percentile.
Solution

6.02

Use the following information to answer the next seven exercises. A distribution is given as $X \sim Exp(0.75)$.

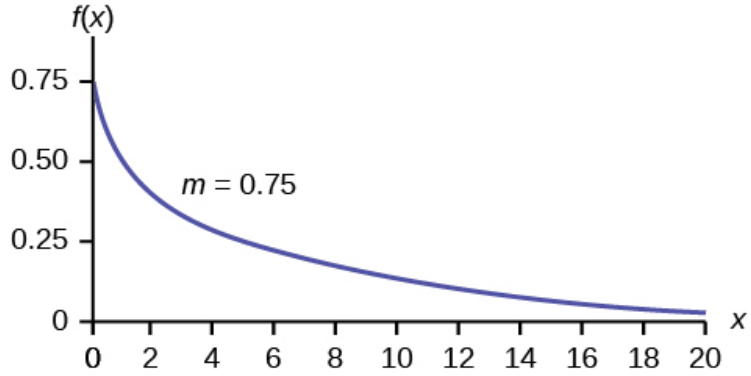
What is m?

What is the probability density function?
Solution

$f(x) = 0.75e^{-0.75x}$

What is the cumulative distribution function?

Draw the distribution.
Solution



Find $P(x < 4)$.

Find the 30^{th} percentile.
Solution

0.4756

Find the median.

Which is larger, the mean or the median?
Solution

The mean is larger. The mean is $\frac{1}{m} = \frac{1}{0.75} \approx 1.33$, which is greater than 0.9242.

Use the following information to answer the next 16 exercises. Carbon-14 is a radioactive element with a half-life of about 5,730 years. Carbon-14 is said to decay exponentially. The decay rate is 0.000121. We start with one gram of carbon-14. We are interested in the time (years) it takes to decay carbon-14.

What is being measured here?

Are the data discrete or continuous?
Solution

continuous

In words, define the random variable X.

What is the decay rate (m)?
Solution

m = 0.000121

The distribution for X is _____.

Find the amount (percent of one gram) of carbon-14 lasting less than 5,730 years. This means, find $P(x <$ 5,730).

1. Sketch the graph, and shade the area of interest.



2. Find the probability. $P(x < 5{,}730) =$ _____

Solution

1. Check student's solution
2. $P(x < 5{,}730) = 0.5001$

Find the percentage of carbon-14 lasting longer than 10,000 years.

1. Sketch the graph, and shade the area of interest.



2. Find the probability. $P(x > 10{,}000) =$ _____

Thirty percent (30%) of carbon-14 will decay within how many years?

1. Sketch the graph, and shade the area of interest.



2. Find the value k such that $P(x < k) = 0.30$.

Solution

1. Check student's solution.
2. $k = 2947.73$

CHAPTER 5 HOMEWORK

5.1 Homework

For each probability and percentile problem, draw the picture.

Consider the following experiment. You are one of 100 people enlisted to take part in a study to determine the percent of nurses in America with an R.N. (registered nurse) degree. You ask nurses if they have an R.N. degree. The nurses answer "yes" or "no." You then calculate the percentage of nurses with an R.N. degree. You give that percentage to your supervisor.

1. What part of the experiment will yield discrete data?
2. What part of the experiment will yield continuous data?

When age is rounded to the nearest year, do the data stay continuous, or do they become discrete? Why?
Solution

Age is a measurement, regardless of the accuracy used.

5.2 Homework

For each probability and percentile problem, draw the picture.

Births are approximately uniformly distributed between the 52 weeks of the year. They can be said to follow a uniform distribution from one to 53 (spread of 52 weeks).

1. $X \sim$ _____
2. Graph the probability distribution.
3. $f(x) =$ _____
4. $\mu =$ _____
5. $\sigma =$ _____
6. Find the probability that a person is born at the exact moment week 19 starts. That is, find $P(x = 19) =$ _____
7. $P(2 < x < 31) =$ _____

8. Find the probability that a person is born after week 40.
9. $P(12 < x | x < 28)$ = _____
10. Find the 70^{th} percentile.
11. Find the minimum for the upper quarter.

A random number generator picks a number from one to nine in a uniform manner.

1. $X \sim$ _____
2. Graph the probability distribution.
3. $f(x)$ = _____
4. μ = _____
5. σ = _____
6. $P(3.5 < x < 7.25)$ = _____
7. $P(x > 5.67)$
8. $P(x > 5 | x > 3)$ = _____
9. Find the 90^{th} percentile.

Solution

1. $X \sim U(1, 9)$
2. Check student's solution.
3. $f(x) = \frac{1}{8}$ where $1 \leq x \leq 9$
4. five
5. 2.3
6. $\frac{15}{32}$
7. $\frac{333}{800}$
8. $\frac{2}{3}$
9. 8.2

According to a study by Dr. John McDougall of his live-in weight loss program at St. Helena Hospital, the people who follow his program lose between six and 15 pounds a month until they approach trim body weight. Let's suppose that the weight loss is uniformly distributed. We are interested in the weight loss of a randomly selected individual following the program for one month.

1. Define the random variable. X = _____

2. $X \sim$ _____
3. Graph the probability distribution.
4. $f(x) =$ _____
5. $\mu =$ _____
6. $\sigma =$ _____
7. Find the probability that the individual lost more than ten pounds in a month.
8. Suppose it is known that the individual lost more than ten pounds in a month. Find the probability that he lost less than 12 pounds in the month.
9. $P(7 < x < 13 | x > 9) =$ _____. State this in a probability question, similarly to parts g and h, draw the picture, and find the probability.

A subway train on the Red Line arrives every eight minutes during rush hour. We are interested in the length of time a commuter must wait for a train to arrive. The time follows a uniform distribution.

1. Define the random variable. $X =$ _____
2. $X \sim$ _____
3. Graph the probability distribution.
4. $f(x) =$ _____
5. $\mu =$ _____
6. $\sigma =$ _____
7. Find the probability that the commuter waits less than one minute.
8. Find the probability that the commuter waits between three and four minutes.
9. Sixty percent of commuters wait more than how long for the train? State this in a probability question, similarly to parts g and h, draw the picture, and find the probability.

Solution

1. X represents the length of time a commuter must wait for a train to arrive on the Red Line.
2. $X \sim U(0, 8)$
3. $f(x) = \frac{1}{8}$ where $\leq x \leq 8$
4. four
5. 2.31
6. $\frac{1}{8}$
7. $\frac{1}{8}$
8. 3.2

The age of a first grader on September 1 at Garden Elementary School is uniformly distributed from 5.8 to 6.8 years. We randomly select one first grader from the class.

1. Define the random variable. X = _____
2. X ~ _____
3. Graph the probability distribution.
4. $f(x)$ = _____
5. μ = _____
6. σ = _____
7. Find the probability that she is over 6.5 years old.
8. Find the probability that she is between four and six years old.
9. Find the 70th percentile for the age of first graders on September 1 at Garden Elementary School.

Use the following information to answer the next three exercises. The Sky Train from the terminal to the rental–car and long–term parking center is supposed to arrive every eight minutes. The waiting times for the train are known to follow a uniform distribution.

What is the average waiting time (in minutes)?

1. zero
2. two
3. three
4. four

Solution

d

Find the 30th percentile for the waiting times (in minutes).

1. two
2. 2.4
3. 2.75
4. three

The probability of waiting more than seven minutes given a person has waited more than four minutes is?

1. 0.125

2. 0.25
3. 0.5
4. 0.75

Solution

b

The time (in minutes) until the next bus departs a major bus depot follows a distribution with $f(x) = \frac{1}{20}$ where x goes from 25 to 45 minutes.

1. Define the random variable. $X =$ _____
2. $X \sim$ _____
3. Graph the probability distribution.
4. The distribution is _____ (name of distribution). It is _____ (discrete or continuous).
5. $\mu =$ _____
6. $\sigma =$ _____
7. Find the probability that the time is at most 30 minutes. Sketch and label a graph of the distribution. Shade the area of interest. Write the answer in a probability statement.
8. Find the probability that the time is between 30 and 40 minutes. Sketch and label a graph of the distribution. Shade the area of interest. Write the answer in a probability statement.
9. $P(25 < x < 55) =$ _____. State this in a probability statement, similarly to parts g and h, draw the picture, and find the probability.
10. Find the 90^{th} percentile. This means that 90% of the time, the time is less than _____ minutes.
11. Find the 75^{th} percentile. In a complete sentence, state what this means. (See part j.)
12. Find the probability that the time is more than 40 minutes given (or knowing that) it is at least 30 minutes.

Suppose that the value of a stock varies each day from $16 to $25 with a uniform distribution.

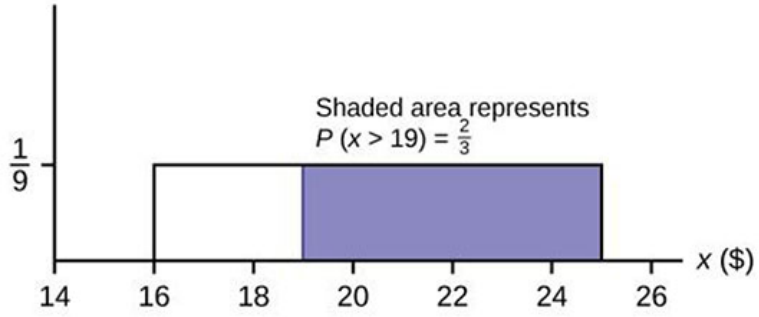
1. Find the probability that the value of the stock is more than $19.
2. Find the probability that the value of the stock is between $19 and $22.
3. Find the upper quartile – 25% of all days the stock is above what value? Draw the graph.
4. Given that the stock is greater than $18, find the probability that the stock is more than $21.

Solution

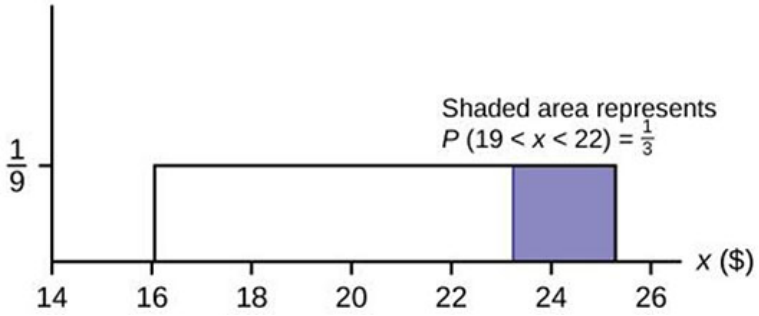
1. The probability density function of X is $\frac{1}{25-16} = \frac{1}{9}$.

   $P(X > 19) = (25 - 19)\left(\frac{1}{9}\right)$

   $= \frac{6}{9} = \frac{2}{3}$.



2. $P(19 < X < 22) = (22 - 19)\left(\frac{1}{9}\right) = \frac{3}{9} = \frac{1}{3}$.



3. The area must be 0.25, and 0.25 = (width)$\left(\frac{1}{9}\right)$, so width = (0.25)(9) = 2.25. Thus, the value is 25 – 2.25 = 22.75.
4. This is a conditional probability question. P(x > 21| x > 18). You can do this two ways:

   - Draw the graph where a is now 18 and b is still 25. The height is $\frac{1}{(25-18)} = \frac{1}{7}$

     So, $P(x > 21|x > 18) = (25 - 21)\left(\frac{1}{7}\right) = 4/7$.

   - Use the formula: $P(x > 21|x > 18) = \frac{P(x > 21 \text{ AND } x > 18)}{P(x > 18)}$

     $= \frac{P(x > 21)}{P(x > 18)} = \frac{(25-21)}{(25-18)} = \frac{4}{7}$.

A fireworks show is designed so that the time between fireworks is between one and five seconds, and follows a uniform distribution.

1. Find the average time between fireworks.
2. Find the probability that the time between fireworks is greater than four seconds.

The number of miles driven by a truck driver falls between 300 and 700, and follows a uniform distribution.

1. Find the probability that the truck driver goes more than 650 miles in a day.
2. Find the probability that the truck driver goes between 400 and 650 miles in a day.
3. At least how many miles does the truck driver travel on the farthest 10% of days?

Solution

1. $P(X > 650) = \frac{700 - 650}{700 - 300} = \frac{500}{400} = \frac{1}{8} = 0.125$.
2. $P(400 < X < 650) = \frac{700 - 650}{700 - 300} = \frac{250}{400} = 0.625$
3. $0.10 = \frac{\text{width}}{700 - 300}$, so width = 400(0.10) = 40. Since 700 – 40 = 660, the drivers travel at least 660 miles on the farthest 10% of days.

5.3 Homework

Suppose that the length of long distance phone calls, measured in minutes, is known to have an exponential distribution with the average length of a call equal to eight minutes.

1. Define the random variable. X = _____.
2. Is X continuous or discrete?
3. $X \sim$ _____
4. μ = _____
5. σ = _____
6. Draw a graph of the probability distribution. Label the axes.
7. Find the probability that a phone call lasts less than nine minutes.
8. Find the probability that a phone call lasts more than nine minutes.
9. Find the probability that a phone call lasts between seven and nine minutes.
10. If 25 phone calls are made one after another, on average, what would you expect the total to be? Why?

Suppose that the useful life of a particular car battery, measured in months, decays with parameter 0.025. We are interested in the life of the battery.

1. Define the random variable. X = _____.
2. Is X continuous or discrete?
3. X ~ _____
4. On average, how long would you expect one car battery to last?
5. On average, how long would you expect nine car batteries to last, if they are used one after another?
6. Find the probability that a car battery lasts more than 36 months.
7. Seventy percent of the batteries last at least how long?

Solution

1. X = the useful life of a particular car battery, measured in months.
2. X is continuous.
3. $X \sim Exp(0.025)$
4. 40 months
5. 360 months
6. 0.4066
7. 14.27

The percent of persons (ages five and older) in each state who speak a language at home other than English is approximately exponentially distributed with a mean of 9.848. Suppose we randomly pick a state.

1. Define the random variable. X = _____.
2. Is X continuous or discrete?
3. X ~ _____
4. μ = _____
5. σ = _____
6. Draw a graph of the probability distribution. Label the axes.
7. Find the probability that the percent is less than 12.
8. Find the probability that the percent is between eight and 14.
9. The percent of all individuals living in the United States who speak a language at home other than English is 13.8.
    1. Why is this number different from 9.848%?
    2. What would make this number higher than 9.848%?

The time (in years) **after** reaching age 60 that it takes an individual to retire is approximately exponentially distributed with a mean of about five years. Suppose we randomly pick one retired individual. We are interested in the time after age 60 to retirement.

1. Define the random variable. X = _____.
2. Is X continuous or discrete?
3. $X \sim$ = _____
4. μ = _____
5. σ = _____
6. Draw a graph of the probability distribution. Label the axes.
7. Find the probability that the person retired after age 70.
8. Do more people retire before age 65 or after age 65?
9. In a room of 1,000 people over age 80, how many do you expect will NOT have retired yet?

Solution

1. X = the time (in years) after reaching age 60 that it takes an individual to retire
2. X is continuous.
3. $X \sim Exp\left(\frac{1}{5}\right)$
4. five
5. five
6. Check student's solution.
7. 0.1353
8. before
9. 18.3

The cost of all maintenance for a car during its first year is approximately exponentially distributed with a mean of $ 150.

1. Define the random variable. X = _____.
2. $X \sim$ = _____
3. μ = _____
4. σ = _____
5. Draw a graph of the probability distribution. Label the axes.
6. Find the probability that a car required over $ 300 for maintenance during its first year.

Use the following information to answer the next three exercises. The average lifetime of a certain new cell phone is three years. The manufacturer will replace any cell phone failing within two years of the date of purchase. The lifetime of these cell phones is known to follow an exponential distribution.

The decay rate is:

1. 0.3333
2. 0.5000
3. 2
4. 3

Solution

a

What is the probability that a phone will fail within two years of the date of purchase?

1. 0.8647
2. 0.4866
3. 0.2212
4. 0.9997

What is the median lifetime of these phones (in years)?

1. 0.1941
2. 1.3863
3. 2.0794
4. 5.5452

Solution

c

Let $X \sim Exp(0.1)$.

1. decay rate = _____
2. μ = _____
3. Graph the probability distribution function.
4. On the graph, shade the area corresponding to $P(x < 6)$ and find the probability.

5. Sketch a new graph, shade the area corresponding to $P(3 < x < 6)$ and find the probability.
6. Sketch a new graph, shade the area corresponding to $P(x < 7)$ and find the probability.
7. Sketch a new graph, shade the area corresponding to the 40^{th} percentile and find the value.
8. Find the average value of x.

Suppose that the longevity of a light bulb is exponential with a mean lifetime of eight years.

1. Find the probability that a light bulb lasts less than one year.
2. Find the probability that a light bulb lasts between six and ten years.
3. Seventy percent of all light bulbs last at least how long?
4. A company decides to offer a warranty to give refunds to light bulbs whose lifetime is among the lowest two percent of all bulbs. To the nearest month, what should be the cutoff lifetime for the warranty to take place?
5. If a light bulb has lasted seven years, what is the probability that it fails within the 8^{th} year?

Solution

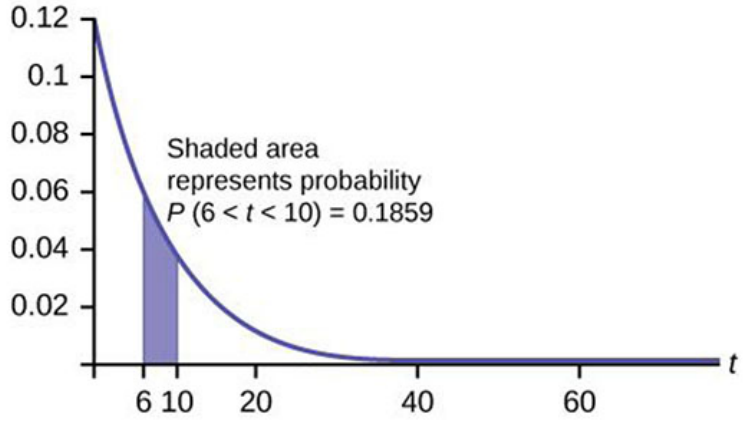
Let T = the lifetime of a light bulb.

The decay parameter is $m = 1/8$, and $T \sim \text{Exp}(1/8)$. The cumulative distribution function is $P(T < t) = 1 - e^{-\frac{t}{8}}$

1. Therefore, $P(T < 1) = 1 - e^{-\frac{1}{8}} \approx 0.1175$.
2. We want to find $P(6 < t < 10)$.

   To do this, $P(6 < t < 10) - P(t < 6)$

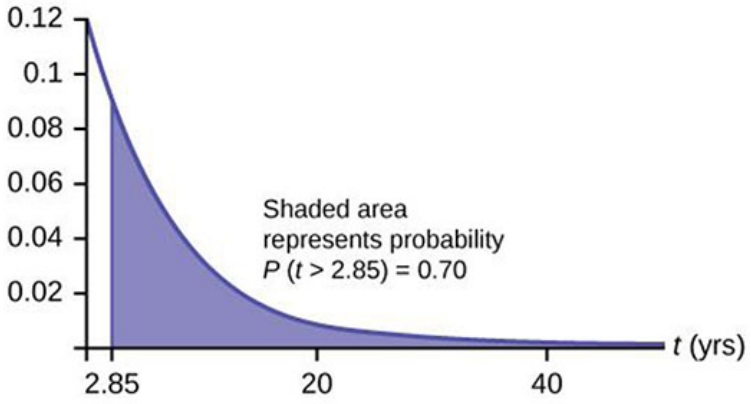
   $= = \left(1 - e^{-\frac{1}{8} * 10}\right) - \left(1 - e^{-\frac{1}{8} * 6}\right) \approx 0.7135 - 0.5276 = 0.1859$



3. We want to find $0.70 = P(T > t) = 1 - \left(1 - e^{-\frac{t}{8}}\right) = e^{-\frac{t}{8}}$.

   Solving for t, $e^{-\frac{t}{8}} = 0.70$, so $-\frac{t}{8} = ln(0.70)$, and $t = -8ln(0.70) \approx 2.85$ years.

   Or use $t = \frac{ln(\text{area_to_the_right})}{(-m)} = \frac{ln(0.70)}{-\frac{1}{8}} \approx 2.85 \text{ years}$.



4. We want to find $0.02 = P(T < t) = 1 - e^{-\frac{t}{8}}$.

   Solving for t, $e^{-\frac{t}{8}} = 0.98$, so $-\frac{t}{8} = ln(0.98)$, and t = $-8ln(0.98) \approx 0.1616$ years, or roughly two months.

   The warranty should cover light bulbs that last less than 2 months.

   Or use $\frac{\ln(\text{area_to_the_right})}{(-m)} = \frac{\ln(1 - 0.2)}{(-\frac{1}{8})} = 0.1616$.

5. We must find $P(T < 8 | T > 7)$.

Notice that by the rule of complement events, $P(T < 8|T > 7) = 1 - P(T > 8|T > 7)$.

By the memoryless property ($P(X > r + t|X > r) = P(X > t)$).

So $P(T > 8|T > 7) = P(T > 1) = 1 - \left(1 - e^{-\frac{1}{8}}\right) = e^{-\frac{1}{8}} \approx 0.8825$

Therefore, $P(T < 8|T > 7) = 1 - 0.8825 = 0.1175$.

At a 911 call center, calls come in at an average rate of one call every two minutes. Assume that the time that elapses from one call to the next has the exponential distribution.

1. On average, how much time occurs between five consecutive calls?
2. Find the probability that after a call is received, it takes more than three minutes for the next call to occur.
3. Ninety percent of all calls occur within how many minutes of the previous call?
4. Suppose that two minutes have elapsed since the last call. Find the probability that the next call will occur within the next minute.
5. Find the probability that less than 20 calls occur within an hour.

In major league baseball, a no-hitter is a game in which a pitcher, or pitchers, doesn't give up any hits throughout the game. No-hitters occur at a rate of about three per season. Assume that the duration of time between no-hitters is exponential.

1. What is the probability that an entire season elapses with a single no-hitter?
2. If an entire season elapses without any no-hitters, what is the probability that there are no no-hitters in the following season?
3. What is the probability that there are more than 3 no-hitters in a single season?

Solution

Let X = the number of no-hitters throughout a season. Since the duration of time between no-hitters is exponential, the number of no-hitters per season is Poisson with mean $\lambda = 3$.

Therefore, $(X = 0) = \frac{3^0 e^{-3}}{0!} = e^{-3} \approx 0.0498$

You could let T = duration of time between no-hitters. Since the time is exponential and there are 3 no-hitters per season, then the time between no-hitters is $\frac{1}{3}$ season. For the exponential, $\mu = \frac{1}{3}$.

Therefore, $m = \frac{1}{\mu} = 3$ and $T \sim Exp(3)$.

1. The desired probability is $P(T > 1) = 1 - P(T < 1) = 1 - (1 - e^{-3}) = e^{-3} \approx 0.0498$.
2. Let T = duration of time between no-hitters. We find $P(T > 2 | T > 1)$, and by the **memoryless property** this is simply $P(T > 1)$, which we found to be 0.0498 in part a.
3. Let X = the number of no-hitters is a season. Assume that X is Poisson with mean $\lambda = 3$. Then $P(X > 3) = 1 - P(X \leq 3) = 0.3528$.

During the years 1998–2012, a total of 29 earthquakes of magnitude greater than 6.5 have occurred in Papua New Guinea. Assume that the time spent waiting between earthquakes is exponential.

1. What is the probability that the next earthquake occurs within the next three months?
2. Given that six months has passed without an earthquake in Papua New Guinea, what is the probability that the next three months will be free of earthquakes?
3. What is the probability of zero earthquakes occurring in 2014?
4. What is the probability that at least two earthquakes will occur in 2014?

According to the American Red Cross, about one out of nine people in the U.S. have Type B blood. Suppose the blood types of people arriving at a blood drive are independent. In this case, the number of Type B blood types that arrive roughly follows the Poisson distribution.

1. If 100 people arrive, how many on average would be expected to have Type B blood?
2. What is the probability that over 10 people out of these 100 have type B blood?
3. What is the probability that more than 20 people arrive before a person with type B blood is found?

Solution

1. $\frac{100}{9} = 11.11$
2. $P(X > 10) = 1 - P(X \leq 10) = 1 - \text{Poissoncdf}(11.11, 10) \approx 0.5532$.
3. The number of people with Type B blood encountered roughly follows the Poisson distribution, so the number of people X who arrive between successive Type B arrivals is roughly exponential with mean $\mu = 9$ and $m = \frac{1}{9}$. The cumulative distribution function of X is $P(X < x) = 1 - e^{-\frac{x}{9}}$. Thus, $P(X >$

$$20) = 1 - P(X \leq 20) = 1 - \left(1 - e^{-\frac{20}{9}}\right) \approx 0.1084.$$

NOTE

We could also deduce that each person arriving has a 8/9 chance of not having Type B blood. So the probability that none of the first 20 people arrive have Type B blood is $\left(\frac{8}{9}\right)^{20} \approx 0.0948$. (The geometric distribution is more appropriate than the exponential because the number of people between Type B people is discrete instead of continuous.)

A web site experiences traffic during normal working hours at a rate of 12 visits per hour. Assume that the duration between visits has the exponential distribution.

1. Find the probability that the duration between two successive visits to the website is more than ten minutes.
2. The top 25% of durations between visits are at least how long?
3. Suppose that 20 minutes have passed since the last visit to the web site. What is the probability that the next visit will occur within the next 5 minutes?
4. Find the probability that less than 7 visits occur within a one-hour period.

At an urgent care facility, patients arrive at an average rate of one patient every seven minutes. Assume that the duration between arrivals is exponentially distributed.

1. Find the probability that the time between two successive visits to the urgent care facility is less than 2 minutes.
2. Find the probability that the time between two successive visits to the urgent care facility is more than 15 minutes.
3. If 10 minutes have passed since the last arrival, what is the probability that the next person will arrive within the next five minutes?
4. Find the probability that more than eight patients arrive during a half-hour period.

Solution

Let T = duration (in minutes) between successive visits. Since patients arrive at a rate of one patient every seven minutes, $\mu = 7$ and the decay constant is $m = \frac{1}{7}$. The cdf is $P(T < t) = 1 - e^{\frac{t}{7}}$

1. $P(T < 2) = 1 - 1 - e^{-\frac{2}{7}} \approx 0.2485$.
2. $P(T > 15) = 1 - P(T < 15) = 1 - \left(1 - e^{-\frac{15}{7}}\right) \approx e^{-\frac{15}{7}} \approx 0.1173$.
3. $P(T > 15 | T > 10) = P(T > 5) = 1 - \left(1 - e^{-\frac{5}{7}}\right) = e^{-\frac{5}{7}} \approx 0.4895$.
4. Let X = # of patients arriving during a half-hour period. Then X has the Poisson distribution with a mean of $\frac{30}{7}$, $X \sim \text{Poisson}\left(\frac{30}{7}\right)$. Find $P(X > 8) = 1 - P(X \leq 8) \approx 0.0311$.

CHAPTER 6: THE NORMAL DISTRIBUTION AND THE CENTRAL LIMIT THEOREM

INTRODUCTION TO CHAPTER 6A: THE NORMAL DISTRIBUTION



Figure 6.1. If you ask enough people about their shoe size, you will find that your graphed data is shaped like a bell curve and can be described as normally distributed. (credit: Ömer Ünlü)

The **normal distribution**, a continuous distribution, is the most important of all the distributions. It is widely used and even more widely abused. Its graph is bell-shaped. You see the bell curve in almost all disciplines. Some of these include psychology, business, economics, the sciences, nursing, and, of course, mathematics. Some of your instructors may use the normal distribution to help determine your grade. Most IQ scores are normally distributed. Often real-estate prices fit a normal distribution. The normal distribution is extremely important, but it cannot be applied to everything in the real world.

In this chapter, you will study the normal distribution, the standard normal distribution, and applications associated with them.

Collaborative Exercises

Sections 6.1 & 6.2

Collaborative Exercise

Your instructor will record the heights of both men and women in your class, separately. Draw histograms of your data. Then draw a smooth curve through each histogram. Is each curve somewhat bell-shaped? Do you think that if you had recorded 200 data values for men and 200 for women that the curves would look bell-shaped? Calculate the mean for each data set. Write the means on the *x*-axis of the appropriate graph below the peak. Shade the approximate area that represents the probability that one randomly chosen male is taller than 72 inches. Shade the approximate area that represents the probability that one randomly chosen female is shorter than 60 inches. If the total area under each curve is one, does either probability appear to be more than 0.5?

Media Attributions

- Chapter 6 Introduction Figure © OpenStax Introductory Statistics is licensed under a CC BY (Attribution) license

6.1 THE STANDARD NORMAL DISTRIBUTION

Learning Objectives

By the end of this section, the student should be able to:

- Calculate and interpret z-scores.
- Recognize characteristics of the standard normal distribution.
- Apply the empirical rule to solve real-world applications.

The normal distribution has two parameters (two numerical descriptive measures): the mean (μ) and the standard deviation (σ). If X is a quantity to be measured that has a normal distribution with mean (μ) and standard deviation (σ), we designate this by writing $X \sim N(\mu, \sigma)$.

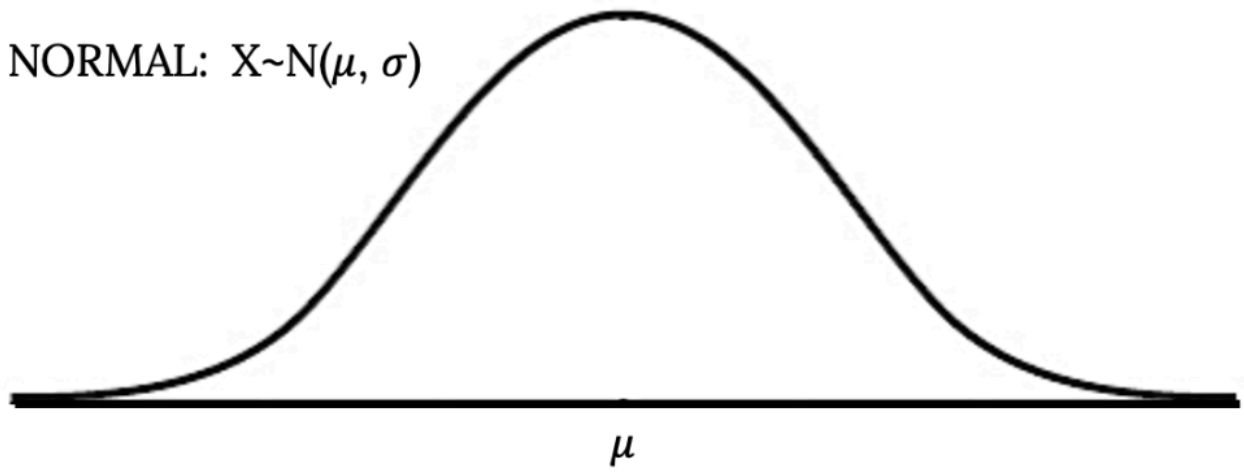


Figure 6.2.

The **probability density function** of this curve is as follows:

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{-\frac{1}{2} \cdot \left(\frac{x-\mu}{\sigma}\right)^2}$$

where:

- $-\infty < X < \infty$
- $-\infty < \mu < \infty$
- $\sigma > 0$

As you can see, the normal pdf is a rather complicated function. **Do not memorize it**. It is not necessary. But this could be a problem since the normal distribution is so widely used. However we will see some ways we can work around this.

The cumulative distribution function is $P(X < x)$. It is calculated either by a calculator or a computer, or it is looked up in a table. Technology has made the tables virtually obsolete.

The curve is symmetrical about a vertical line drawn through the mean, μ. In theory, the mean is the same as the median, because the graph is symmetric about μ. As the notation indicates, the normal distribution depends only on the mean and the standard deviation. Since the area under the curve must equal one, a change in the standard deviation, σ, causes a change in the shape of the curve; the curve becomes fatter or skinnier depending on σ. A change in μ causes the graph to shift to the left or right. This means there are an infinite number of normal probability distributions. One of special interest is called the **standard normal distribution**.

The **standard normal distribution** is a normal distribution of **standardized values called *z*-scores**. **A *z*-score is measured in units of the standard deviation.** For example, if the mean of a normal distribution is five and the standard deviation is two, the value 11 is three standard deviations above (or to the right of) the mean. The calculation is as follows:

$x = \mu + (z)(\sigma) = 5 + (3)(2) = 11$

The z-score is three.

The mean for the standard normal distribution is zero, and the standard deviation is one. The transformation $z = \frac{x - \mu}{\sigma}$ produces the distribution $Z \sim N(0, 1)$. The value x comes from a normal distribution with mean μ and standard deviation σ.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=792#oembed-1

Z-Scores

If X is a normally distributed random variable and $X \sim N(\mu, \sigma)$, then the z-score is:

$$z = \frac{x - \mu}{\sigma}$$

The z-score tells you how many standard deviations the value x is above (to the right of) or below (to the left of) the mean, μ. Values of x that are larger than the mean have positive z-scores, and values of x that are smaller than the mean have negative z-scores. If x equals the mean, then x has a z-score of zero.

Note

Suppose $X \sim N(5, 6)$. This says that x is a normally distributed random variable with mean $\mu = 5$ and standard deviation $\sigma = 6$. Suppose $x = 17$.

Then: $z = \frac{x-\mu}{\sigma} = \frac{17-5}{6} = 2$

This means that $x = 17$ is **two standard deviations** (2σ) above or to the right of the mean $\mu = 5$. The standard deviation is $\sigma = 6$.

Notice that: $5 + (2)(6) = 17$ (The pattern is $\mu + z\sigma = x$)

Now suppose $x = 1$. Then: $z = \frac{x-\mu}{\sigma} = \frac{1-5}{6} = -0.67$ (rounded to two decimal places)

This means that $x = 1$ is 0.67 standard deviations (-0.67σ) below or to the left of the mean $\mu = 5$. Notice that: $5 + (-0.67)(6)$ is approximately equal to one (This has the pattern $\mu + (-0.67)\sigma = 1$)

Summarizing, when z is positive, x is above or to the right of μ and when z is negative, x is to the left of or below μ. Or, when z is positive, x is greater than μ, and when z is negative x is less than μ.



One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=792#oembed-2

Example

Some doctors believe that a person can lose five pounds, on the average, in a month by reducing his or her fat intake and by exercising consistently. Suppose weight loss has a normal distribution. Let X = the amount of weight lost(in pounds) by a person in a month. Use a standard deviation of two pounds. $X \sim N(5, 2)$. Fill in the blanks.

a. Suppose a person **lost** ten pounds in a month. The z-score when $x = 10$ pounds is $z = 2.5$ (verify). This z-score tells you that $x = 10$ is _____ standard deviations to the _____ (right or left) of the mean _____ (What is the mean?).

Solution

This z-score tells you that $x = 10$ is **2.5** standard deviations to the **right** of the mean **five**.

b. Suppose a person **gained** three pounds (a negative weight loss). Then z = _____. This

z-score tells you that $x = -3$ is _____ standard deviations to the _____ (right or left) of the mean.

Solution

$z = \mathbf{-4}$. This z-score tells you that $x = -3$ is **four** standard deviations to the **left** of the mean.

c. Suppose the random variables X and Y have the following normal distributions: $X \sim N(5, 6)$ and $Y \sim N(2, 1)$. If $x = 17$, then $z = 2$. (This was previously shown.) If $y = 4$, what is z?

Solution

$z = \frac{y - \mu}{\sigma} = \frac{4 - 2}{1} = 2$ where $\mu = 2$ and $\sigma = 1$.

The z-score for $y = 4$ is $z = 2$. This means that four is $z = 2$ standard deviations to the right of the mean. Therefore, $x = 17$ and $y = 4$ are both two (of **their own**) standard deviations to the right of **their** respective means.

The z-score allows us to compare data that are scaled differently. To understand the concept, suppose $X \sim N(5, 6)$ represents weight gains for one group of people who are trying to gain weight in a six week period and $Y \sim N(2, 1)$ measures the same weight gain for a second group of people. A negative weight gain would be a weight loss. Since $x = 17$ and $y = 4$ are each two standard deviations to the right of their means, they represent the same, standardized weight gain **relative to their means**.

Your Turn!

What is the *z*-score of *x*, when $x = 1$ and $X \sim N(12,3)$?

Solution

$$z = \frac{1-12}{3} \approx -3.67$$

Example

The mean height of 15 to 18-year-old males from Chile from 2009 to 2010 was 170 cm with a standard deviation of 6.28 cm. Male heights are known to follow a normal distribution. Let X = the height of a 15 to 18-year-old male from Chile in 2009 to 2010. Then $X \sim N(170, 6.28)$.

a. Suppose a 15 to 18-year-old male from Chile was 168 cm tall from 2009 to 2010. The *z*-score when $x = 168$ cm is $z =$ _____. This *z*-score tells you that $x = 168$ is _____ standard deviations to the _____ (right or left) of the mean _____ (What is the mean?).

Solution

z = –0.32, z = 0.32, left, $\mu = 170$

b. Suppose that the height of a 15 to 18-year-old male from Chile from 2009 to 2010 has a *z*-score of $z = 1.27$. What is the male's height? The *z*-score ($z = 1.27$) tells you that the male's height is _____ standard deviations to the _____ (right or left) of the mean.

Solution

x = 177.98, 1.27, right

Your Turn!

1. Suppose a 15- to 18-year-old male from Chile was 176 cm tall from 2009 to 2010. The *z*-score when $x = 176$ cm is $z =$ _____. This *z*-score tells you that $x = 176$ cm is _____ standard deviations to the _____ (right or left) of the mean _____ (What is the mean?).

2. Suppose that the height of a 15 to 18-year-old male from Chile from 2009 to 2010 has a *z*-score of $z = -2$. What is the male's height? The *z*-score ($z = -2$) tells you that the male's height is _____ standard deviations to the _____ (right or left) of the mean.

Solution

Solve the equation $z = \frac{x - \mu}{\sigma}$ for x. $x = \mu + (z)(\sigma)$

1. $z = \frac{176 - 170}{6.28} \approx 0.96$, This *z*-score tells you that $x = 176$ cm is 0.96 standard deviations to the right of the mean 170 cm.
2. $X = 157.44$ cm, The *z*-score($z = -2$) tells you that the male's height is two standard deviations to the left of the mean

Example

From 1984 to 1985, the mean height of 15- to 18-year-old males from Chile was 172.36 cm, and the standard deviation was 6.34 cm. Let Y = the height of 15 to 18-year-old males from 1984 to 1985. Then $Y \sim N(172.36, 6.34)$.

The mean height of 15- to 18-year-old males from Chile from 2009 to 2010 was 170 cm with a standard deviation of 6.28 cm. Male heights are known to follow a normal distribution. Let X = the height of a 15 to 18-year-old male from Chile in 2009 to 2010. Then $X \sim N(170, 6.28)$.

Find the z-scores for $x = 160.58$ cm and $y = 162.85$ cm. Interpret each z-score. What can you say about $x = 160.58$ cm and $y = 162.85$ cm?

Solution

The z-score for $x = 160.58$ is $z = -1.5$. The z-score for $y = 162.85$ is $z = -1.5$. Both $x = 160.58$ and $y = 162.85$ deviate the same number of standard deviations from their respective means and in the same direction.

Your Turn!

In 2012, 1,664,479 students took the SAT exam. The distribution of scores in the verbal section of the SAT had a mean $\mu = 496$ and a standard deviation $\sigma = 114$. Let X = a SAT exam verbal section score in 2012. Then $X \sim N(496, 114)$.

Find the z-scores for $x_1 = 325$ and $x_2 = 366.21$. Interpret each z-score. What can you say about $x_1 = 325$ and $x_2 = 366.21$?

Solution

The z-score for $x_1 = 325$ is $z_1 = -1.14$.

The z-score for $x_2 = 366.21$ is $z_2 = -1.14$.

Student 2 scored closer to the mean than Student 1, and since they both had negative z-scores, Student 2 had the better score.

Your Turn!

Fill in the blanks.

Jerome averages 16 points a game with a standard deviation of four points. $X \sim N(16,4)$. Suppose Jerome scores ten points in a game. The z–score when $x = 10$ is –1.5. This score tells you that $x = 10$ is _____ standard deviations to the _____(right or left) of the mean_____(What is the mean?).

Solution

1.5, left, 16

The Empirical Rule

The Empirical Rule If X is a random variable and has a normal distribution with mean μ and standard deviation σ, then the **Empirical Rule** says the following:

- About 68% of the x values lie between -1σ and $+1\sigma$ of the mean μ (within one standard deviation of the mean). The z-scores for $+1\sigma$ and -1σ are +1 and –1, respectively.
- About 95% of the x values lie between -2σ and $+2\sigma$ of the mean μ (within two standard deviations of the mean). The z-scores for $+2\sigma$ and -2σ are +2 and –2, respectively.
- About 99.7% of the x values lie between -3σ and $+3\sigma$ of the mean μ (within three standard deviations of the mean). The z-scores for $+3\sigma$ and -3σ are +3 and –3 respectively.

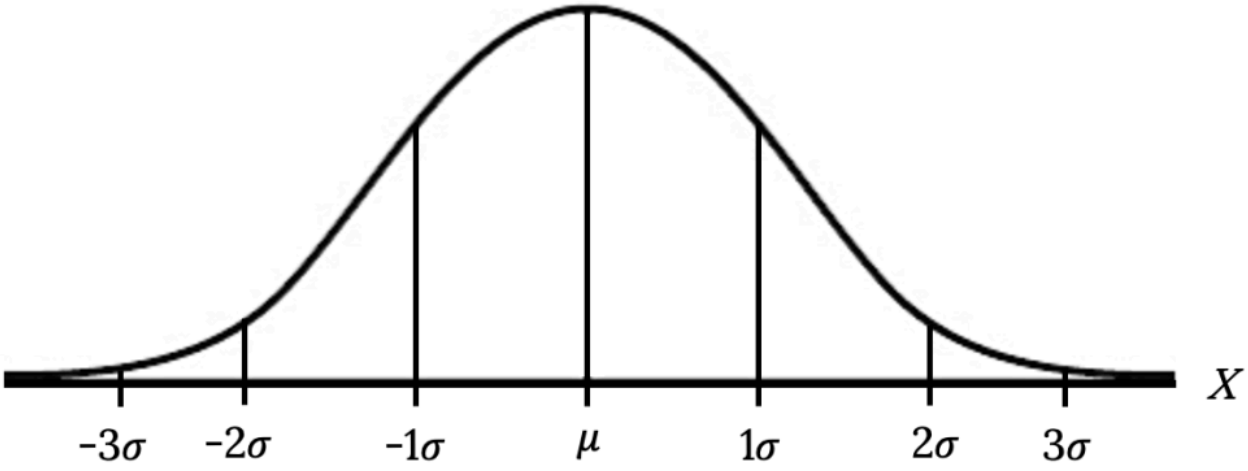


Figure 5.11

Notice that almost all the x values lie within three standard deviations of the mean. The empirical rule is also known as the 68-95-99.7 rule.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=792#oembed-3

Note

Suppose x has a normal distribution with mean 50 and standard deviation 6.

- About 68% of the x values lie between $-1\sigma = (-1)(6) = -6$ and $1\sigma = (1)(6) = 6$ of the mean 50. The values $50 - 6 = 44$ and $50 + 6 = 56$ are within one standard deviation of the mean 50.

The z-scores are –1 and +1 for 44 and 56, respectively.

- About 95% of the x values lie between $-2\sigma = (-2)(6) = -12$ and $2\sigma = (2)(6) = 12$. The values 50 – 12 = 38 and 50 + 12 = 62 are within two standard deviations of the mean 50. The z-scores are –2 and +2 for 38 and 62, respectively.
- About 99.7% of the x values lie between $-3\sigma = (-3)(6) = -18$ and $3\sigma = (3)(6) = 18$ of the mean 50. The values 50 – 18 = 32 and 50 + 18 = 68 are within three standard deviations of the mean 50. The z-scores are –3 and +3 for 32 and 68, respectively.

Example

From 1984 to 1985, the mean height of 15- to 18-year-old males from Chile was 172.36 cm, and the standard deviation was 6.34 cm. Let Y = the height of 15- to 18-year-old males in 1984 to 1985. Then $Y \sim N(172.36, 6.34)$.

About 68% of the y values lie between what two values? These values are _____.
The z-scores are _____, respectively.
About 95% of the y values lie between what two values? These values are _____.
The z-scores are _____ respectively.
About 99.7% of the y values lie between what two values? These values are _____.
The z-scores are _____, respectively.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=792#h5p-142

Your Turn!

The scores on a college entrance exam have an approximate normal distribution with mean, $\mu = 52$ points and a standard deviation, $\sigma = 11$ points.

1. About 68% of the y values lie between what two values? These values are _____. The z-scores are _____, respectively.
2. About 95% of the y values lie between what two values? These values are _____. The z-scores are _____, respectively.
3. About 99.7% of the y values lie between what two values? These values are _____. The z-scores are _____, respectively.

Solution

1. About 68% of the values lie between the values 41 and 63. The z-scores are –1 and 1, respectively.
2. About 95% of the values lie between the values 30 and 74. The z-scores are –2 and 2, respectively.
3. About 99.7% of the values lie between the values 19 and 85. The z-scores are –3 and 3, respectively.

Your Turn!

Suppose X has a normal distribution with mean 25 and standard deviation five. Between what values of x do 68% of the values lie?

Solution

between 20 and 30.

References

"Blood Pressure of Males and Females." StatCrunch, 2013. Available online at http://www.statcrunch.com/5.0/viewreport.php?reportid=11960 (accessed May 14, 2013).

"The Use of Epidemiological Tools in Conflict-affected populations: Open-access educational resources for policy-makers: Calculation of z-scores." London School of Hygiene and Tropical Medicine, 2009. Available online at http://conflict.lshtm.ac.uk/page_125.htm (accessed May 14, 2013).

"2012 College-Bound Seniors Total Group Profile Report." CollegeBoard, 2012. Available online at http://media.collegeboard.com/digitalServices/pdf/research/TotalGroup-2012.pdf (accessed May 14, 2013).

"Digest of Education Statistics: ACT score average and standard deviations by sex and race/ethnicity and percentage of ACT test takers, by selected composite score ranges and planned fields of study: Selected years, 1995 through 2009." National Center for Education Statistics. Available online at http://nces.ed.gov/programs/digest/d09/tables/dt09_147.asp (accessed May 14, 2013).

Data from the *San Jose Mercury News*.

Data from *The World Almanac and Book of Facts*.

"List of stadiums by capacity." Wikipedia. Available online at https://en.wikipedia.org/wiki/List_of_stadiums_by_capacity (accessed May 14, 2013).

Data from the National Basketball Association. Available online at www.nba.com (accessed May 14, 2013).

Media Attributions

- Private: Figure 5.10 © Significant Statistics by John Morgan Russell is licensed under a CC BY-SA (Attribution ShareAlike) license
- Private: Figure 5.11 © Significant Statistics by John Morgan Russell is licensed under a CC BY-SA (Attribution ShareAlike) license

6.2 USING THE NORMAL DISTRIBUTION

Learning Objectives

By the end of this section, the student should be able to:

- Find areas under the normal distribution curve.
- Use the inverse normal to identify values for a given area under the normal curve.
- Compute probabilities and percentiles for real-world applications using the normal distribution.

Finding Normal Probabilities

The shaded area in the following graph indicates the area to the left of x. This area is represented by the probability $P(X < x)$. Normal tables, computers, and calculators provide or calculate the probability $P(X < x)$.

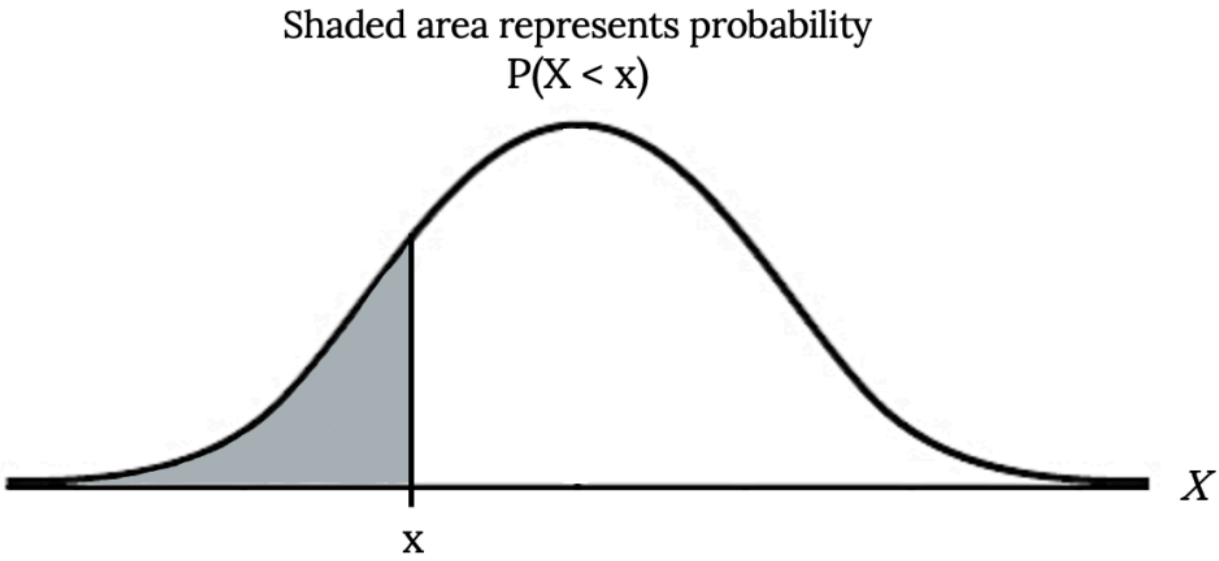


Figure 6.3.

The area to the right is then $P(X > x) = 1 - P(X < x)$. Remember, $P(X < x)$ = Area to the left of the vertical line through x. $P(X < x) = 1 - P(X < x)$ = Area to the right of the vertical line through x. $P(X < x)$ is the same as $P(X \leq x)$ and $P(X > x)$ is the same as $P(X \geq x)$ for continuous distributions.

Note

If the area to the left is 0.0228, then the area to the right is 1 – 0.0228 = 0.9772.

Your Turn!

If the area to the left of x is 0.012, then what is the area to the right?

Solution
1 − 0.012 = 0.988

There are 3 main ways we can find probabilities associated with the Normal Distribution. These include:

1. Math (via Calculus Integration)
2. The Standardizing Process
3. Technology

We would like to avoid complicated math if possible, so option 1 is out.

In order to avoid the math, a process called "Standardizing" can be used (option 2). This involves Z-scores, the Standard Normal Distribution and Tables. Although this tried and true process is now somewhat antiquated, it is a great place to start.

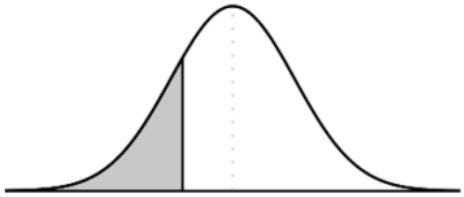
There are many technologies such as calculators and various statistical software that let us skip the entire standardizing process and instantaneously provide us with a probability. Although we typically have these at our disposal to use in practice, it is good to understand the process going on behind the scenes to make sure we apply our technology correctly.

The Standardizing Process

The **standard normal distribution** (SND) is the simplest form of the normal distribution you can think of. Recall some important facts from the previous section:

1. The mean for the standard normal distribution is zero, and the standard deviation is one.
2. If X is a normally distributed random variable and $X \sim N(\mu, \sigma)$, then the z-score is: $z = \frac{x - \mu}{\sigma}$. The transformation $z = \frac{x - \mu}{\sigma}$ produces the distribution $Z \sim N(0, 1)$. The value x in the given equation comes from a normal distribution with mean μ and standard deviation σ.
3. The ***z*-scores** are converted to units of the standard deviation.
4. A z-score tells you how many standard deviations the value x is above (to the right of) or below (to the left of) the mean, μ. Values of x that are larger than the mean have positive z-scores, and values of x that are smaller than the mean have negative z-scores. If x equals the mean, then x has a z-score of zero.

We have the Z table at our disposal with probabilities already calculated and organized. Note that some Z tables (like in Figure 6.4) give us the left tailed CDF, or "less than" probability. For example, using the Figure 6.4 below, the area to the left of a Z score of -3.37, P(Z ≤ -3.37) =0.0004.



Second decimal place of Z										
0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	Z
0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	−3.4
0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0005	−3.3
0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0007	−3.2
0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0010	−3.1
0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013	0.0013	−3.0

Figure 6.4.

We can then use these CDF values, P(Z ≤ z), and some probability rules to find greater than [P(Z ≥ z) = 1 – P(Z ≤ z)] or in between [P(a ≤ Z ≤ b) = P(Z ≤ b) – P(Z ≤ a)] probabilities.

Here is a link to another Z table example: Normal Table (***Source:*** NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, January 3, 2009). This table gives positive z-scores and the area in the table is the area under the normal curve from 0 (the mean) to that z-score. Using this table takes some practice and understanding symmetry and area under the normal curve.

A few examples are provided on the table link (***Source:*** NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, January 3, 2009):

1. The area to the left of a Z score of 1.53 can be found by going to 1.5 in the X column, go right to the 0.03 column to find the value 0.43699. Now add 0.5 (for the probability less than zero) to obtain the final result of 0.93699. P(Z ≤ 1.53) = 0.93699.
2. The probability that *z* is less than or equal to -1.53 (also called the area to the left of -1.53) can also be found using this table. Remember that the curve is symmetrical, so P(Z ≥ z) = P(Z ≤ -z). Since we found P(Z ≤ 1.53) = 0.93699, then P(Z ≥ 1.53) = 1 – 0.93699 = 0.06301 = P(Z ≤ -1.53).
3. Finding probabilities (area) between values is also possible. To find the probability that z is between -1 and 0.5, look up the values for 0.5 (0.5 + 0.19146 = 0.69146) and -1 (1 – (0.5 + 0.34134) = 0.15866).

Then subtract the results (0.69146 – 0.15866) to obtain the result 0.5328.

To use any of these Z tables **with a non-standard normal distribution** (either the location parameter is not 0 or the scale parameter is not 1), standardize your value by subtracting the mean and dividing the result by the standard deviation. Then look up the value for this standardized value. (***Source:*** NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, January 3, 2009)

Example

Use the Z table (Normal Table) to find the following probabilities:

a. $P(Z \leq -0.54)$

Solution

The area between 0.54 and 0 can be found by going to 0.5 in the X column, go right to the 0.04 column to find the value 0.20540. If you add 0.5 to that value, (0.5 + 0.20540 = 0.7054) you'll have the area all the way to the left of 0.54. $P(Z \leq 0.54) = 0.7054$. To find the equal negative z-score, subtract that from 1. So 1 – 0.7054 = **0.2946**.

b. $P(Z \geq 1.2)$

Solution

The area between 1.2 and 0 can be found by going to 1.2 in the X column, go right to the 0.00 column to find the value 0.38493. Since the curve has an area to the right of 0 that equals 0.5, then to the right of just 1.2 must be 0.5 – 0.38493 = **0.11507**.

c. $P(-1.5 \leq Z \leq 0.84)$

Solution

The area between 0 and 0.84 is found by going to 0.8 in the X column, go right to the 0.04 column to find the value 0.29955. The area between -1.5 and 0 is the same as the area between 1.5 and 0, which can be found on the table as a value of 0.43319. This area can be found by adding 0.29955 + 0.43319 = **0.73274**.

Your Turn

Use the Z table to find the following probabilities:

a. P(Z ≤ 1)

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=805#h5p-143

b. P(Z ≥ 1)

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=805#h5p-144

c. P(-1 ≤ Z ≤ 1)

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=805#h5p-145

So far we have seen the idea that we can convert any normal distribution with any mean and standard

deviation to the standard normal distribution in units of z-scores. We also have the associated probabilities in our Z table. Essentially, the work has been done for us if we know how to standardize and look up the associated probability in the table. The general process is: **$X \sim N(\mu,\sigma)$ -> $Z \sim N(0,1)$ -> Probability from Z table**

This process, while maybe outdated in our technology age, is good for beginners to understand and useful when we do not have access to technology.

Example

Height and weight are two measurements used to track a child's development. The World Health Organization measures child development by comparing the weights of children who are the same height and the same gender. In 2009, weights for all 80 cm girls in the reference population had a mean $\mu = 10.2$ kg and standard deviation $\sigma = 0.8$ kg. Weights are normally distributed. $X \sim N(10.2, 0.8)$. Calculate the z-scores that correspond to the following weights, then find the associated probabilities.

a. The probability that a child weighs **less than 11 kg**

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=805#h5p-146

b. The probability that a child weighs **more than 7.9 kg**

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=805#h5p-147

c. The probability that a child weighs **between 11.2 and 12.2 kg**

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=805#h5p-148

Your Turn!

The golf scores for a school team were normally distributed with a mean of 68 and a standard deviation of three.

a. Find the probability that a randomly selected golfer scored **less than 65**.

Solution
0.1587

b. Find the probability that a golfer scored **between 66 and 70**.

Solution
0.4950

The "Un-standardizing" Process

Sometimes we may be given a percentile or z-score and want to work backwards through the standardizing process to find a value on the original distribution. You could call this "un-standardizing" or finding a normal quantile. **quantile**.

The process looks like this: **Probability in Z table -> Z~N(0,1) -> X~N(μ,σ)**

For example, if the mean of a normal distribution is five and the standard deviation is two, what value is three standard deviations above (or to the right of) the mean (z-score = three). Rearranging the z-score formula, the calculation is as follows: $x = \mu + (z)(\sigma) = 5 + (3)(2) = 11$

Often we are given a percentile to find on the original distribution. For example, what if we want to know a value on the previous distribution that corresponds to the 90th percentile? We can look up a probability of 0.9 in the Z table and find a corresponding z-score of approximately 1.28.

$x = \mu + (z)(\sigma) = 5 + (1.28)(2) = 7.56$

Example

Two thousand students took an exam. The scores on the exam have an approximate normal distribution with a mean μ = 81 points and standard deviation σ = 15 points.

a. Calculate the **first** and **third quartile** scores for this exam.

Solution

Start by finding 0.25 (or the closest value possible) in the Z table. You can find 0.24537 in the table for a z-score of 0.66. Now using the rearranged z-score formula, $x = \mu + (z)(\sigma) = 81 + (0.66)(\,) = 7.56$

the $Q_1 = 25^{th}$ percentile = invNorm(0.25,81,15) = 70.9

$Q_3 = 75^{th}$ percentile = invNorm(0.75,81,15) = 91.9

b. The **middle 50%** of the exam scores are between what two values?

Solution

In part a, The middle 50% of the scores are between 70.9 and 91.1.

Your Turn

A citrus farmer who grows mandarin oranges finds that the diameters of mandarin oranges harvested on his farm follow a normal distribution with a mean diameter of 5.85 cm and a standard deviation of 0.24 cm.

a. Find the 90^{th} percentile for the diameters of mandarin oranges

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=805#h5p-149

b. The middle 20% of mandarin oranges from this farm have diameters between _____ and _____.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=805#h5p-150

Using Technology

Probabilities can be quickly calculated using **technology.** The technology that will be used for the following examples are the **TI-83+ and TI-84 calculators.** The best news about using the TI graphing calculator is that *there is no need to standardize*. You don't have to go from X~N(μ,σ) -> Z~N(0,1), you just use **X~N(μ,σ).** Although the calculator is a quick method, drawing the image of the shaded region can help you to see if your answer is valid.

Using the TI-83+ and TI-84 Calculators for Normal Probabilities

Go into **`2nd DISTR`**.

Press **`2:normalcdf`**.

The syntax for the instructions are as follows: **normalcdf(lower value, upper value, mean, standard deviation)**

In some instances:

- the upper number of the area might be 1E99 (= 10^{99}). You get 1E99 (= 10^{99}) by pressing `1`, the `EE` key (a 2nd key) and then `99`. Or, you can enter `10^99` instead. The number 10^{99} is way out in the right tail of the normal curve.
- the lower number of the area might be –1E99 (= -10^{99}). The number -10^{99} is way out in the left tail of the normal curve.



One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=805#oembed-1

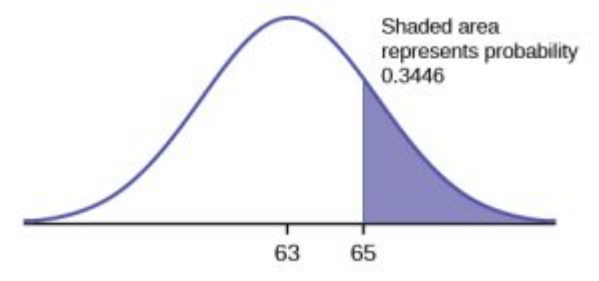
Example

The final exam scores in a statistics class were normally distributed with a mean of 63 and a standard deviation of five.

a. Find the probability that a randomly selected student scored more than 65 on the exam.

Solution

Draw the graph, then find $P(x > 65)$.



normalcdf(65, 1E99, 63, 5) = 0.3446.

b. Find the probability that a randomly selected student scored less than 85.

Solution

normalcdf(-1E99, 85, 63, 5) = 1 (rounds to one)

The probability that one student scores less than 85 is approximately one (or 100%).

Example

There are approximately one billion smartphone users in the world today. In the United States the ages 13 to 55+ of smartphone users approximately follow a normal distribution with approximate mean and standard deviation of 36.9 years and 13.9 years, respectively.

a. Determine the probability that a random smartphone user in the age range 13 to 55+ is between 23 and 64.7 years old.

Solution

normalcdf(23, 64.7, 36.9, 13.9) = 0.8186

b. Determine the probability that a randomly selected smartphone user in the age range 13 to 55+ is at most 50.8 years old.

Solution

normalcdf(-10^{99}, 50.8, 36.9, 13.9) = 0.8413

Your Turn!

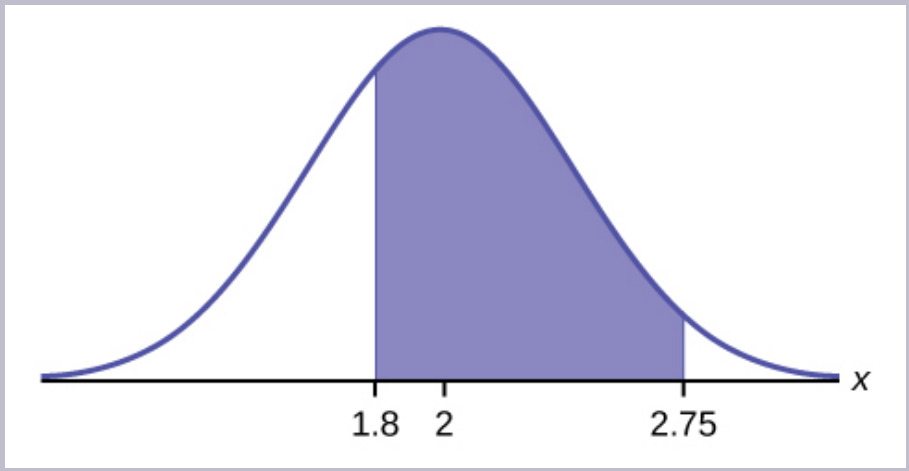
A personal computer is used for office work at home, research, communication, personal finances, education, entertainment, social networking, and a myriad of other things. Suppose that the average number of hours a household personal computer is used for entertainment is two hours per day. Assume the times for entertainment are normally distributed and the standard deviation

for the times is half an hour. Find the probability that a household personal computer is used for entertainment between 1.8 and 2.75 hours per day.

Solution

Let X = the amount of time (in hours) a household personal computer is used for entertainment. $X \sim N(2, 0.5)$ where $\mu = 2$ and $\sigma = 0.5$. Find $P(1.8 < x < 2.75)$.

The probability for which you are looking is the area **between** $x = 1.8$ and $x = 2.75$.



normalcdf(1.8, 2.75, 2, 0.5) = 0.5886

The probability that a household personal computer is used between 1.8 and 2.75 hours per day for entertainment is 0.5886.

Using the TI-83+ and TI-84 Calculators for Finding Percentiles (Z-scores)

Press 2nd **DISTR**

Press **3:invNorm**

The syntax for the instructions are as follows: **invNorm(area to the left, mean, standard deviation)**

**Note: Percentiles are values that are left area, but be careful to read if you are given a right area in the problem. Remember, Right Area = 1 – Left Area.



One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=805#oembed-2

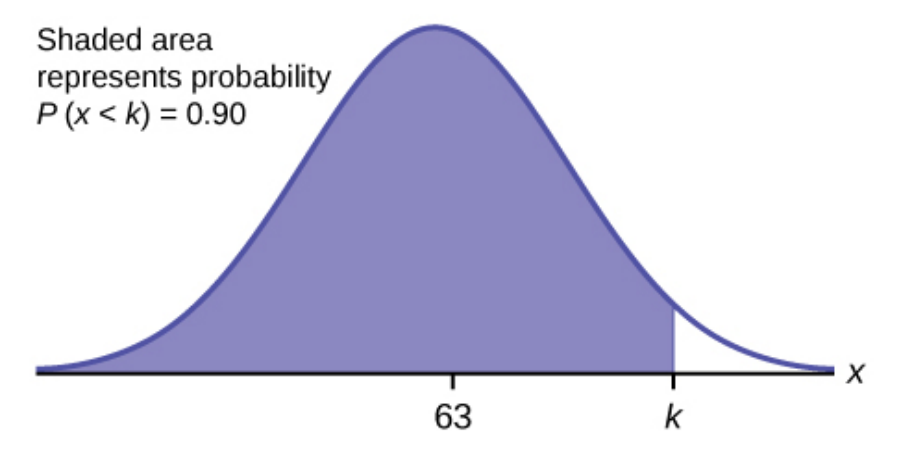
Example

The final exam scores in a statistics class were normally distributed with a mean of 63 and a standard deviation of five.

a. Find the 90th percentile (that is, find the score k that has 90% of the scores below k and 10% of the scores above k).

Solution

Draw a graph and shade the area that corresponds to the 90th percentile. **Let k = the 90th percentile.** The variable k is located on the x-axis. $P(x < k)$ is the area to the left of k. The 90th percentile k separates the exam scores into those that are the same or lower than k and those that are the same or higher. Ninety percent of the test scores are the same or lower than k, and ten percent are the same or higher.



For this problem, invNorm(0.90, 63, 5) = 69.4

The 90th percentile is 69.4. This means that 90% of the test scores fall at or below 69.4 and 10% fall at or above.

b. Find the 70th percentile (that is, find the score *k* such that 70% of scores are below *k* and 30% of the scores are above *k*).

Solution

invNorm(0.70,63,5) = 65.6

The 70th percentile is 65.6. This means that 70% of the test scores fall at or below 65.5 and 30% fall at or above.

Example

There are approximately one billion smartphone users in the world today. In the United States the ages 13 to 55+ of smartphone users approximately follow a normal distribution with approximate

mean and standard deviation of 36.9 years and 13.9 years respectively. Using this information, answer the following questions (round answers to one decimal place).

a. Calculate the interquartile range (*IQR*).

Solution

$IQR = Q_3 - Q_1$
Calculate $Q_3 = 75^{th}$ percentile and $Q_1 = 25^{th}$ percentile.
invNorm(0.75,36.9,13.9) = Q_3 = 46.2754
invNorm(0.25,36.9,13.9) = Q_1 = 27.5246
$IQR = Q_3 - Q_1$ = 18.7508

b. Forty percent of the ages that range from 13 to 55+ are at least what age?

Solution

Find k where $P(x > k) = 0.40$ ("At least" translates to "greater than or equal to.")
0.40 = the area to the right.
Area to the left = 1 – 0.40 = 0.60.
The area to the left of k = 0.60.
invNorm(0.60,36.9,13.9) = 40.4215.
k = 40.42.
Forty percent of the ages that range from 13 to 55+ are at least 40.42 years.

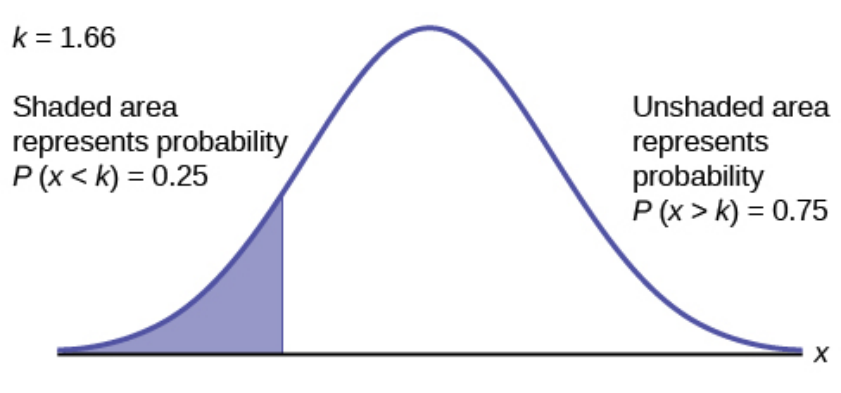
Your Turn!

A personal computer is used for office work at home, research, communication, personal finances, education, entertainment, social networking, and a myriad of other things. Suppose that the

average number of hours a household personal computer is used for entertainment is two hours per day. Assume the times for entertainment are normally distributed and the standard deviation for the times is half an hour. Find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment.

Solution

To find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment, **find the 25th percentile,** k, where $P(x < k) = 0.25$.



invNorm(0.25,2,0.5) = 1.66

The maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment is 1.66 hours.

Your Turn

A citrus farmer who grows mandarin oranges finds that the diameters of mandarin oranges harvested on his farm follow a normal distribution with a mean diameter of 5.85 cm and a standard

deviation of 0.24 cm. The middle 40% of mandarin oranges from this farm are between _____ and _____.

Solution

The middle area = 0.40, so each tail has an area of 0.30.

1 – 0.40 = 0.60

The tails of the graph of the normal distribution each have an area of 0.30.

Find *k1*, the 30^{th} percentile and *k2*, the 70^{th} percentile (0.40 + 0.30 = 0.70).

k1 = invNorm(0.30,5.85,0.24) = 5.72 cm

k2 = invNorm(0.70,5.85,0.24) = 5.98 cm

References

"Naegele's rule." Wikipedia. Available online at http://en.wikipedia.org/wiki/Naegele's_rule (accessed May 14, 2013).

"403: NUMMI." Chicago Public Media & Ira Glass, 2013. Available online at http://www.thisamericanlife.org/radio-archives/episode/403/nummi (accessed May 14, 2013).

"Scratch-Off Lottery Ticket Playing Tips." WinAtTheLottery.com, 2013. Available online at http://www.winatthelottery.com/public/department40.cfm (accessed May 14, 2013).

"Smart Phone Users, By The Numbers." Visual.ly, 2013. Available online at http://visual.ly/smart-phone-users-numbers (accessed May 14, 2013).

"Facebook Statistics." Statistics Brain. Available online at http://www.statisticbrain.com/facebook-statistics/(accessed May 14, 2013).

Media Attributions

- Private: Figure 5.12 © Significant Statistics by John Morgan Russell is licensed under a CC BY-SA (Attribution ShareAlike) license
- Private: Figure 5.13 © Significant Statistics by John Morgan Russell is licensed under a CC BY-SA

(Attribution ShareAlike) license

- fig-ch06_05_01-1

INTRODUCTION TO CHAPTER 6B: THE CENTRAL LIMIT THEOREM



Figure 7.1. If you want to figure out the distribution of the change people carry in their pockets, using the central limit theorem and assuming your sample is large enough, you will find that the distribution is normal and bell-shaped. (credit: John Lodder)

Why are we so concerned with means? Two reasons are: they give us a middle ground for comparison, and they are easy to calculate. In this chapter, you will study means and the **central limit theorem**.

The **central limit theorem** (clt for short) is one of the most powerful and useful ideas in all of statistics. There are two alternative forms of the theorem, and both alternatives are concerned with drawing finite samples size n from a population with a known mean, μ, and a known standard deviation, σ. The first alternative says that if we collect samples of size n with a "large enough n," calculate each sample's mean, and create a histogram of those means, then the resulting histogram will tend to have an approximate normal bell shape. The second alternative says that if we again collect samples of size n that are "large enough," calculate

the sum of each sample and create a histogram, then the resulting histogram will again tend to have a normal bell-shape.

In either case, it does not matter what the distribution of the original population is, or whether you even need to know it. The important fact is that the distribution of sample means and the sums tend to follow the normal distribution.

The size of the sample, n, that is required in order to be "large enough" depends on the original population from which the samples are drawn (the sample size should be at least 30 or the data should come from a normal distribution). If the original population is far from normal, then more observations are needed for the sample means or sums to be normal. **Sampling is done with replacement.** Given simple random samples of size n from a given population with a measured characteristic such as mean, proportion, or standard deviation for each sample, the probability distribution of all the measured characteristics is called a **sampling distribution**.

It would be difficult to overstate the importance of the central limit theorem in statistical theory. Knowing that data, even if its distribution is not normal, behaves in a predictable way is a powerful tool.



One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=811#oembed-1

Collaborative Exercises

Section 6.3

Collaborative Exercise

Suppose eight of you roll one fair die ten times, seven of you roll two fair dice ten times, nine of you roll five fair dice ten times, and 11 of you roll ten fair dice ten times.

Each time a person rolls more than one die, he or she calculates the sample mean of the faces showing. For example, one person might roll five fair dice and get 2, 2, 3, 4, 6 on one roll.

The mean is $\frac{2+2+3+4+6}{5}$ = 3.4. The 3.4 is one mean when five fair dice are rolled. This same person would roll the five dice nine more times and calculate nine more means for a total of ten means.

Your instructor will pass out the dice to several people. Roll your dice ten times. For each roll, record the faces, and find the mean. Round to the nearest 0.5.

Your instructor (and possibly you) will produce one graph (it might be a histogram) for one die, one graph for two dice, one graph for five dice, and one graph for ten dice. Since the "mean" when you roll one die is just the face on the die, what distribution do these **means** appear to be representing?

Draw the graph for the means using two dice. Do the sample means show any kind of pattern?

Draw the graph for the means using five dice. Do you see any pattern emerging?

Finally, draw the graph for the means using ten dice. Do you see any pattern to the graph? What can you conclude as you increase the number of dice?

As the number of dice rolled increases from one to two to five to ten, the following is happening:

1. The mean of the sample means remains approximately the same.
2. The spread of the sample means (the standard deviation of the sample means) gets smaller.
3. The graph appears steeper and thinner.

You have just demonstrated the central limit theorem (clt).

The central limit theorem tells you that as you increase the number of dice, **the sample means tend toward a normal distribution (the sampling distribution).**

Media Attributions

- Chapter 7 Introduction Figure © OpenStax Introductory Statistics is licensed under a CC BY (Attribution) license

6.3 THE CENTRAL LIMIT THEOREM FOR SAMPLE MEANS (AVERAGES)

Learning Objectives

By the end of this section, the student should be able to:

- Recognize characteristics of the Central Limit Theorem for the Sample Means
- Apply and interpret the Central Limit Theorem for the Sample Means and use it to solve real-world applications.

Applying the **law of large numbers** here, we could say that if you take larger and larger samples from a population, then the mean $\overline{x}$ of the sample tends to get closer and closer to μ. From the central limit theorem, we know that as n gets larger and larger, the sample means follow a normal distribution. The larger n gets, the smaller the standard deviation gets. (Remember that the standard deviation for $\overline{x}$ is $\frac{\sigma}{\sqrt{n}}$.) This means that the sample mean $\overline{x}$ must be close to the population mean μ. We can say that μ is the value that the sample means approach as n gets larger. The central limit theorem illustrates the law of large numbers.

The size of the sample, n, that is required in order to be "large enough" depends on the original population from which the samples are drawn (the sample size should be at least 30 or the data should come from a normal distribution). If the original population is far from normal, then more observations are needed for the sample means or sums to be normal. Sampling is done with replacement.

The following images look at sampling distributions of the sample mean built from taking 1000 samples of different sample sizes from a normal Population. What pattern do you notice?

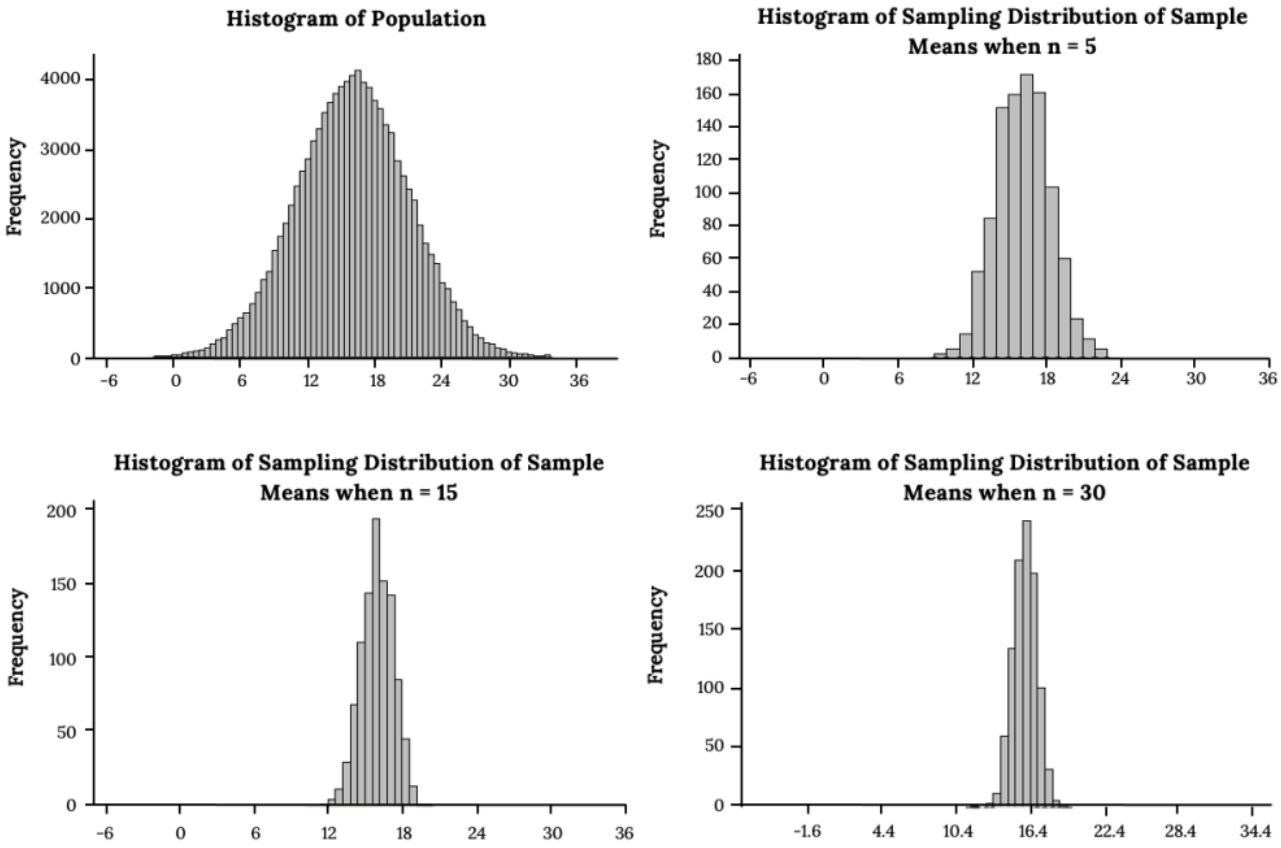


Figure 6.5. Sampling Distributions of the Sample Mean from a Non-Normal Population

What differences do you notice when sampling from a normal population vs. non-normal?

Example

Suppose:

- eight students roll one fair die ten times
- seven roll two fair dice ten times
- nine roll five fair dice ten times
- 11 roll ten fair dice ten times.

Each time a person rolls more than one die, he or she calculates the sample mean of the faces showing. For example, one person might roll five fair dice and get 2, 2, 3, 4, 6 on one roll.

The mean is $\frac{2+2+3+4+6}{5}$ = 3.4. The 3.4 is one mean when five fair dice are rolled. This same person would roll the five dice nine more times and calculate nine more means for a total of ten means.

As the number of dice rolled increases from one to two to five to ten, the following would happen:

1. The mean of the sample means remains approximately the same.
2. The spread of the sample means (the standard deviation of the sample means) gets smaller.
3. The graph appears steeper and thinner.

We have just demonstrated the idea of central limit theorem (clt) for means, that as you increase the sample size, the sampling distribution of the sample mean tends toward a normal distribution.

To summarize, the central limit theorem for sample means says that if you keep drawing larger and larger samples (such as rolling one, two, five, and finally, ten dice) and calculating their means, the sample means form their own normal distribution (the sampling distribution). The normal distribution has the same mean as the original distribution and a variance that equals the original variance divided by the sample size. Standard deviation is the square root of variance, so the standard deviation of the sampling distribution (a.k.a. standard error) is the standard deviation of the original distribution divided by the square root of n. The variable n is the number of values that are averaged together, not the number of times the experiment is done.

It would be difficult to overstate the importance of the central limit theorem in statistical theory. Knowing that data, even if its distribution is not normal, behaves in a predictable way is a powerful tool. We can simulate this idea using technology.

Suppose X is a random variable with a distribution that may be known or unknown (it can be any distribution). Using a subscript that matches the random variable, suppose:

1. μ_X = the mean of X
2. σ_X = the standard deviation of X

If you draw random samples of size n, then as n increases, the random variable $\overline{X}$ which consists of sample means, tends to be normally distributed and $\overline{X} \sim N\left(\mu_x, \frac{\sigma x}{\sqrt{n}}\right)$.

The **central limit theorem** for sample means says that if you keep drawing larger and larger samples (such

as rolling one, two, five, and finally, ten dice) and **calculating their means,** the sample means form their own **normal distribution** (the sampling distribution). The **normal distribution** has the same mean as the original distribution and a variance that equals the original variance divided by the sample size. The variable n is the number of values that are averaged together, not the number of times the experiment is done.

To put it more formally, if you draw random samples of size n, the distribution of the random variable $\overline{X}$, which consists of sample means, is called the **sampling distribution of the mean**. The sampling distribution of the mean approaches a normal distribution as n, the sample size, increases.

The random variable $\overline{X}$ has a different z-score associated with it from that of the random variable X. The mean $\overline{x}$ is the value of $\overline{X}$ in one sample.

$$z = \frac{\overline{x} - \mu_x}{\left(\frac{\sigma_x}{\sqrt{n}}\right)}$$

μ_X is the **average** of both X and $\overline{X}$.

$\sigma\overline{x} = \frac{\sigma x}{\sqrt{n}}$ = standard deviation of $\overline{X}$ and is called the **standard error of the mean**.

Using the CLT

The random variable $\overline{X}$ has a different z-score formula associated with it from that of a single observation. Remember, the mean $\overline{x}$ is the mean of one sample and μ_X is the average, or center, of both X (The original distribution) and $\overline{X}$.

$$z = \frac{\overline{x} - \mu_x}{\left(\frac{\sigma_x}{\sqrt{n}}\right)}$$

We can use our Z table and standardize just as we are already familiar with, or can use your technology of choice (shown after this example).

Example

An unknown distribution has a mean of 90 and a standard deviation of 15. Samples of size $n = 25$ are drawn randomly from the population.

a. Find the probability that the sample mean is between 85 and 92.

- Let X = one value from the original unknown population. The probability question asks you to find a probability for the **sample mean**.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=815#h5p-155

- Find $P(85 < \overline{x} < 92)$. Draw a graph.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=815#h5p-156

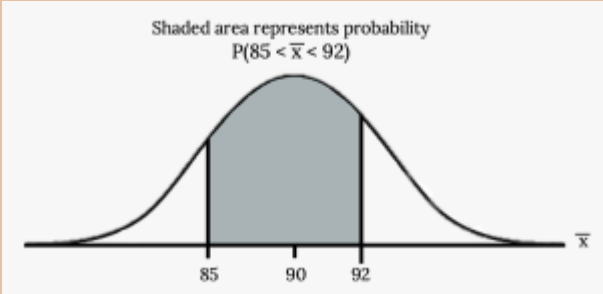


Figure 6.6: Area Under the Curve.

b. Find the value that is two standard deviations above the expected value, 90, of the sample mean.

An interactive H5P element has been excluded from this version of the text. You can view it online

here:
https://louis.pressbooks.pub/introductorystatistics/?p=815#h5p-157

Using the TI-83, 83+, 84, 84+ Calculator

To find probabilities for means on the calculator, follow these steps.

2nd DISTR

2:normalcdf( )

$$normalcdf\left(lower\ value\ of\ the\ area,\ upper\ value\ of\ the\ area,\ mean,\ \frac{standarddeviation}{\sqrt{samplesize}}\right)$$

where:

- *mean* is the mean of the original distribution
- *standard deviation* is the standard deviation of the original distribution
- *sample size* = *n*

Example

Using the graphing calculator for the previous problem:

An unknown distribution has a mean of 90 and a standard deviation of 15. Samples of size $n = 25$ are drawn randomly from the population. Find the probability that the sample mean is between 85 and 92.

Solution

Let X = one value from the original unknown population. The probability question asks you to find a probability for the **sample mean**.

Let $\overline{X}$ = the mean of a sample of size 25. Since $\mu_X = 90$, $\sigma_X = 15$, and $n = 25$,

$$\overline{X} \sim N\left(90, \frac{15}{\sqrt{25}}\right).$$

Find $P(85 < \overline{x} < 92)$. Draw a graph.

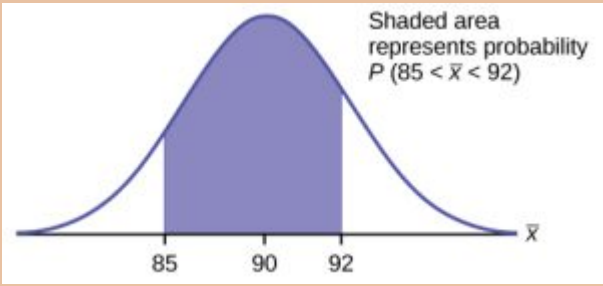


Figure 7.2. Shaded area represents P(85 < x-bar < 92).

`normalcdf`(lower value, upper value, mean, standard error of the mean)

The parameter list is abbreviated (lower value, upper value, μ, $\frac{\sigma}{\sqrt{n}}$)

`normalcdf`$(85,92,90,\frac{15}{\sqrt{25}})$ = 0.6997

Your Turn!

An unknown distribution has a mean of 45 and a standard deviation of eight. Samples of size n = 30 are drawn randomly from the population. Find the probability that the sample mean is between 42 and 50.

Solution

$P(42 < \overline{x} < 50) = \left(42,50,45,\frac{8}{\sqrt{30}}\right) = 0.9797$

Example

The length of time, in hours, it takes an "over 40" group of people to play one soccer match is normally distributed with a **mean of two hours** and a **standard deviation of 0.5 hours**. A **sample of size *n* = 50** is drawn randomly from the population. Find the probability that the **sample mean** is between 1.8 hours and 2.3 hours.

Solution

Let X = the time, in hours, it takes to play one soccer match.

The probability question asks you to find a probability for the **sample mean time, in hours**, it takes to play one soccer match.

Let $\overline{X}$ = the mean time, in hours, it takes to play one soccer match.

If μ_X = _____, σ_X = _____, and n = _____, then $X \sim N$(_____, _____) by the central limit theorem for means.

$\mu_X = 2$, $\sigma_X = 0.5$, $n = 50$, and $X \sim N\left(2, \frac{0.5}{\sqrt{50}}\right)$

Find $P(1.8 < \overline{x} < 2.3)$. Draw a graph.

$P(1.8 < \overline{x} < 2.3) = 0.9977$

$\texttt{normalcdf}\left(1.8,2.3,2,\frac{.5}{\sqrt{50}}\right) = 0.9977$

The probability that the mean time is between 1.8 hours and 2.3 hours is 0.9977.

Your Turn!

The length of time taken on the SAT for a group of students is normally distributed with a mean of 2.5 hours and a standard deviation of 0.25 hours. A sample size of $n = 60$ is drawn randomly from the population. Find the probability that the sample mean is between two hours and three hours.

Solution

$P(2 < \overline{x} < 3) = \text{normalcdf}\left(2, 3, 2.5, \frac{0.25}{\sqrt{60}}\right) = 1$

Using the TI-83, 83+, 84, 84+ Calculator

To find percentiles for means on the calculator, follow these steps.

2nd DIStR

3: invNorm( )

k = invNorm $\left(\text{area to the left of } k, \text{mean}, \frac{standarddeviation}{\sqrt{samplesize}}\right)$

where:

- k = the k^{th} percentile
- *mean* is the mean of the original distribution
- *standard deviation* is the standard deviation of the original distribution
- *sample size* = n

Example

In a recent study reported Oct. 29, 2012 on the Flurry Blog, the mean age of tablet users is 34 years. Suppose the standard deviation is 15 years. Take a sample of size n = 100.

1. What are the mean and standard deviation for the sample mean ages of tablet users?
2. What does the distribution look like?
3. Find the probability that the sample mean age is more than 30 years (the reported mean age of tablet users in this particular study).
4. Find the 95th percentile for the sample mean age (to one decimal place).

Solution

1. Since the sample mean tends to target the population mean, we have $\mu_{\bar{X}} = \mu = 34$. The sample standard deviation is given by $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{100}} = \frac{15}{10} = 1.5$
2. The central limit theorem states that for large sample sizes(n), the sampling distribution will be approximately normal.
3. The probability that the sample mean age is more than 30 is given by $P(\bar{X} > 30)$ = `normalcdf`(30,E99,34,1.5) = 0.9962
4. Let k = the 95th percentile. k = invNorm $\left(0.95, 34, \frac{15}{\sqrt{100}}\right)$ = 36.5

Your Turn!

In an article on Flurry Blog, a gaming marketing gap for men between the ages of 30 and 40 is identified. You are researching a startup game targeted at the 35-year-old demographic. Your idea is to develop a strategy game that can be played by men from their late 20s through their late 30s. Based on the article's data, industry research shows that the average strategy player is 28 years old with a standard deviation of 4.8 years. You take a sample of 100 randomly selected gamers. If your target market is 29- to 35-year-olds, should you continue with your development strategy?

Solution

You need to determine the probability for men whose mean age is between 29 and 35 years of age wanting to play a strategy game.

$P(29 < \overline{x} < 35)$ = `normalcdf` $\left(29, 35, 28, \frac{4.8}{\sqrt{100}}\right)$ = 0.0186

You can conclude there is approximately a 1.9% chance that your game will be played by men whose mean age is between 29 and 35.

Example

The mean number of minutes for app engagement by a tablet user is 8.2 minutes. Suppose the standard deviation is one minute. Take a sample of 60.

1. What are the mean and standard deviation for the sample mean number of app engagement by a tablet user?
2. What is the standard error of the mean?
3. Find the 90th percentile for the sample mean time for app engagement for a tablet user. Interpret this value in a complete sentence.
4. Find the probability that the sample mean is between eight minutes and 8.5 minutes.

Solution

1. $\mu_{\overline{x}} = \mu = 8.2\ \sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{60}} = 0.13$
2. This allows us to calculate the probability of sample means of a particular distance from the mean, in repeated samples of size 60.
3. Let k = the 90th percentile

   k = `invNorm` $\left(0.90, 8.2, \frac{1}{\sqrt{60}}\right)$ = 8.37. This value indicates that 90 percent of the average app engagement time for table users is less than 8.37 minutes.
4. $P(8 < \overline{x} < 8.5)$ = `normalcdf` $\left(8, 8.5, 8.2, \frac{1}{\sqrt{60}}\right)$ = 0.9293

Your Turn!

Cans of a cola beverage claim to contain 16 ounces. The amounts in a sample are measured and the statistics are $n = 34$, $\overline{x} = 16.01$ ounces. If the cans are filled so that $\mu = 16.00$ ounces (as labeled) and $\sigma = 0.143$ ounces, find the probability that a sample of 34 cans will have an average amount greater than 16.01 ounces. Do the results suggest that cans are filled with an amount greater than 16 ounces?

Solution

We have $P(\overline{x} > 16.01) = \texttt{normalcdf}\left(16.01, \text{E}99, 16, \frac{0.143}{\sqrt{34}}\right) = 0.3417$. Since there is a 34.17% probability that the average sample weight is greater than 16.01 ounces, we should be skeptical of the company's claimed volume. If I am a consumer, I should be glad that I am probably receiving free cola. If I am the manufacturer, I need to determine if my bottling processes are outside of acceptable limits.

Note

It is important for you to understand when to use the central limit theorem. If you are being asked to **find the probability of the mean**, **use the clt for the mean**. This also applies to percentiles for means. If you are being asked to find the probability of an **individual** value, do **not** use the clt. **Use the distribution of its random variable.**

Example

Suppose that a market research analyst for a cell phone company conducts a study of their customers who exceed the time allowance included on their basic cell phone contract; the analyst finds that for those people who exceed the time included in their basic contract, the **excess time used** follows an exponential distribution with a mean of 22 minutes.

Consider a random sample of 80 customers who exceed the time allowance included in their basic cell phone contract.

Let X = the excess time used by one INDIVIDUAL cell phone customer who exceeds his contracted time allowance.

$X \sim Exp\left(\frac{1}{22}\right)$. From previous chapters, we know that $\mu = 22$ and $\sigma = 22$.

Let $\overline{X}$ = the mean excess time used by a sample of $n = 80$ customers who exceed their contracted time allowance.

$\overline{X} \sim N\left(22, \frac{22}{\sqrt{80}}\right)$ by the central limit theorem for sample means

Using the clt to find probability

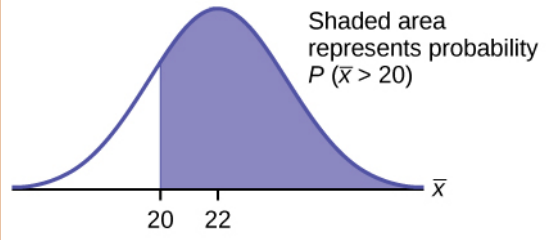
1. Find the probability that the mean excess time used by the 80 customers in the sample is longer than 20 minutes. This is asking us to find $P(\overline{x} > 20)$. Draw the graph.
2. Suppose that one customer who exceeds the time limit for his cell phone contract is randomly selected. Find the probability that this individual customer's excess time is longer than 20 minutes. This is asking us to find $P(x > 20)$.
3. Explain why the probabilities in parts a and b are different.

Solution

1. Find: $P(\overline{x} > 20)$

   $P(\overline{x} > 20) = 0.79199$ using `normalcdf` $\left(20,1\text{E}99,22,\frac{22}{\sqrt{80}}\right)$

The probability is 0.7919 that the mean excess time used is more than 20 minutes, for a sample of 80 customers who exceed their contracted time allowance.



REMINDER

1E99 = 10^{99} and –1E99 = -10^{99}. Press the EE key for E. Or just use 10^{99} instead of 1E99.

2. Find $P(x > 20)$. Remember to use the exponential distribution for an **individual:**

$X \sim Exp\left(\frac{1}{22}\right)$.

$P(x > 20) = e^{\left(-\left(\frac{1}{22}\right)(20)\right)}$ or $e^{(-0.04545(20))} = 0.4029$

   1. $P(x > 20) = 0.4029$ but $P(\bar{x} > 20) = 0.7919$
   2. The probabilities are not equal because we use different distributions to calculate the probability for individuals and for means.
   3. When asked to find the probability of an individual value, use the stated distribution of its random variable; do not use the clt. Use the clt with the normal distribution when you are being asked to find the probability for a mean.

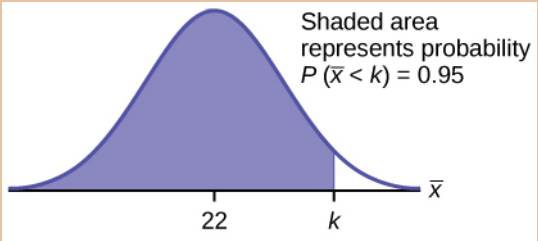
Using the clt to find percentiles

Find the 95th percentile for the **sample mean excess time** for samples of 80 customers who exceed their basic contract time allowances. Draw a graph.

Solution

Let k = the 95th percentile. Find k where $P(\bar{x} < k) = 0.95$

$k = 26.0$ using $\text{invNorm}\left(0.95, 22, \frac{22}{\sqrt{80}}\right) = 26.0$



The 95th percentile for the **sample mean excess time used** is about 26.0 minutes for random samples of 80 customers who exceed their contractual allowed time.

Ninety five percent of such samples would have means under 26 minutes; only five percent of such samples would have means above 26 minutes.

Example

Based on data from the National Health Survey, women between the ages of 18 and 24 have an average systolic blood pressures (in mmHg) of 114.8 with a standard deviation of 13.1. Systolic blood pressure for women between the ages of 18 to 24 follows a normal distribution.

1. If one woman from this population is randomly selected, find the probability that her systolic blood pressure is greater than 120.
2. If 40 women from this population are randomly selected, find the probability that their mean systolic blood pressure is greater than 120.
3. If the sample were four women between the ages of 18 to 24 and we did not know the original distribution, could the central limit theorem be used?

Solution

1. $P(x > 120)$ = `normalcdf`(120,99,114.8,13.1) = 0.0272. There is about a 3% that the randomly selected woman will have systolic blood pressure greater than 120.
2. $P(\overline{x} > 120)$ = `normalcdf` $\left(120, 114.8, \frac{13.1}{\sqrt{40}}\right)$ = 0.006. There is only a 0.6% chance that the average systolic blood pressure for the randomly selected group is greater than 120.
3. The central limit theorem could not be used if the sample size were four and we did not

know the original distribution was normal. The sample size would be too small.

Your Turn!

According to Boeing data, the 757 airliner carries 200 passengers and has doors with a mean height of 72 inches. Assume for a certain population of men we have a mean of 69.0 inches and a standard deviation of 2.8 inches.

1. What mean doorway height would allow 95% of men to enter the aircraft without bending?
2. Assume that half of the 200 passengers are men. What mean doorway height satisfies the condition that there is a 0.95 probability that this height is greater than the mean height of 100 men?
3. For engineers designing the 757, which result is more relevant: the height from part a or part b? Why?

Solution

1. We know that $\mu_{\bar{x}} = \mu = 69$ and we have $\sigma_{\bar{x}} = 2.8$. The height of the doorway is found to be `invNorm`(0.95,69,2.8) = 73.61
2. We know that $\mu_{\bar{x}} = \mu = 69$ and we have $\sigma_{\bar{x}} = 0.28$. So, `invNorm`(0.95,69,0.28) = 69.49
3. When designing the doorway heights, we need to incorporate as much variability as possible in order to accommodate as many passengers as possible. Therefore, we need to use the result based on part a.

References

Baran, Daya. "20 Percent of Americans Have Never Used Email." WebGuild, 2010. Available online at http://www.webguild.org/20080519/20-percent-of-americans-have-never-used-email (accessed May 17, 2013).

Data from The Flurry Blog, 2013. Available online at http://blog.flurry.com (accessed May 17, 2013).
Data from the United States Department of Agriculture.

Media Attributions

- Private: 6.2.1
- Private: 6.4
- Figure 7.2 © OpenStax Introductory Statistics is licensed under a CC BY (Attribution) license

6.4 THE CENTRAL LIMIT THEOREM FOR SUMS

Learning Objectives

By the end of this section, the student should be able to:

- Recognize characteristics of the Central Limit Theorem for the Sums
- Apply and interpret the Central Limit Theorem for the Sums and use it to solve real-world applications.

Suppose X is a random variable with a distribution that may be **known or unknown** (it can be any distribution) and suppose:

1. μ_X = the mean of X
2. σ_X = the standard deviation of X

If you draw random samples of size n, then as n increases, the random variable ΣX consisting of sums tends to be normally distributed and $\Sigma X \sim N((n)(\mu_X), (\sqrt{n})(\sigma_X))$.

The central limit theorem for sums says that if you keep drawing larger and larger samples and taking their sums, the sums form their own normal distribution (the sampling distribution), which approaches a normal distribution as the sample size increases. **The normal distribution has a mean equal to the original mean multiplied by the sample size and a standard deviation equal to the original standard deviation multiplied by the square root of the sample size.**

The random variable ΣX has the following z-score associated with it:

1. Σx is one sum.

2. $z = \dfrac{\Sigma x - (n)\,(\mu_X)}{(\sqrt{n})\,(\sigma_X)}$
    1. $(n)(\mu_X)$ = the mean of ΣX
    2. $(\sqrt{n})\,(\sigma_X)$ = standard deviation of ΣX

Using the TI-83, 83+, 84, 84+ Calculator

To find probabilities for sums on the calculator, follow these steps.

2^{nd} `DISTR`

2:`normalcdf()`

`normalcdf`(lower value of the area, upper value of the area, (n)(mean), ($\sqrt{n}$)(standard deviation))

where:

- *mean* is the mean of the original distribution
- *standard deviation* is the standard deviation of the original distribution
- *sample size* = n

Example

An unknown distribution has a mean of 90 and a standard deviation of 15. A sample of size 80 is drawn randomly from the population.

1. Find the probability that **the sum** of the 80 values (or the total of the 80 values) is **more than 7,500**.
2. Find **the sum** that is 1.5 standard deviations above the mean of the sums.

Solution

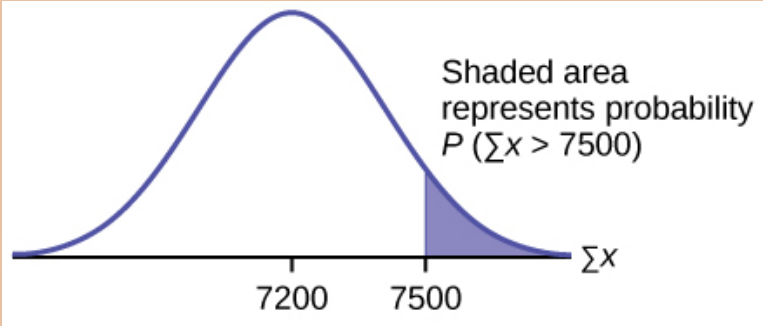
Let X = one value from the original unknown population. The probability question asks you to find a probability for **the sum (or total of) 80 values.**

ΣX = the sum or total of 80 values. Since $\mu_X = 90$, $\sigma_X = 15$, and $n = 80$, $\Sigma X \sim N((80)(90), (\sqrt{80})(15))$

- mean of the sums = $(n)(\mu_X) = (80)(90) = 7{,}200$
- standard deviation of the sums = $(\sqrt{n})(\sigma_X) = (\sqrt{80})(15)$
- sum of 80 values = $\Sigma x = 7{,}500$

1. Find $P(\Sigma x > 7{,}500)$

$P(\Sigma x > 7{,}500) = 0.0127$



`normalcdf`(lower value, upper value, mean of sums, `stdev` of sums)

The parameter list is abbreviated(lower, upper, $(n)(\mu_X, (\sqrt{n})(\sigma_X))$

`normalcdf` (7500,1E99,(80)(90), $(\sqrt{80})(15)) = 0.0127$

2. Find Σx where $z = 1.5$.

$\Sigma x = (n)(\mu_X) + (z)(\sqrt{n})(\sigma_X) = (80)(90) + (1.5)(\sqrt{80})(15) = 7{,}401.2$

Your Turn!

An unknown distribution has a mean of 45 and a standard deviation of eight. A sample size of 50 is drawn randomly from the population. Find the probability that **the sum** of the 50 values is **more than 2,400**.

Solution
0.0040

Using the TI-83, 83+, 84, 84+ Calculator

To find percentiles for sums on the calculator, follow these steps.

2^{nd} DIStR

3:invNorm( )

k = invNorm (area to the left of *k*, (*n*)(mean), Formula does not parse(standard deviation))

where:

- *k* is the k^{th} percentile
- *mean* is the mean of the original distribution
- *standard deviation* is the standard deviation of the original distribution
- *sample size* = *n*

Example

In an October 29, 2012, study reported on the Flurry Blog, the mean age of tablet users is 34 years. Suppose the standard deviation is 15 years. The sample size is 50.

1. What are the mean and standard deviation for the sum of the ages of tablet users? What is the distribution?
2. Find the probability that the sum of the ages is between 1,500 and 1,800 years.
3. Find the 80th percentile for the sum of the 50 ages.

Solution

1. $\mu_{\Sigma x} = n\mu_x = 50(34) = 1{,}700$ and $\sigma_{\Sigma x} = \sqrt{n}\sigma_x =$ Formula does not parse$(15) = 106.01$
   The distribution is normal for sums by the central limit theorem.
2. $P(1500 < \Sigma x < 1800) =$ `normalcdf`$(1{,}500, 1{,}800, (50)(34),$ Formula does not parse$(15)) = 0.7974$
3. Let $k =$ the 80th percentile.
   $k =$ `invNorm`$(0.80,(50)(34),$ Formula does not parse$(15)) = 1{,}789.3$

Your Turn!

In an October 29, 2012, study reported on the Flurry Blog, the mean age of tablet users is 35 years. Suppose the standard deviation is ten years. The sample size is 39.

1. What are the mean and standard deviation for the sum of the ages of tablet users? What is the distribution?

2. Find the probability that the sum of the ages is between 1,400 and 1,500 years.
3. Find the 90th percentile for the sum of the 39 ages.

Solution

1. $\mu_{\Sigma X} = n\mu_X$ = 1,365 and $\sigma_{\Sigma X} = \sqrt{n}\sigma_x$ = 62.4
   The distribution is normal for sums by the central limit theorem.
2. $P(1400 < \Sigma_X < 1500)$ = normalcdf(1400,1500,(39)(35),($\sqrt{39}$)(10)) = 0.2723
3. Let k = the 90th percentile. k = invNorm(0.90,(39)(35),($\sqrt{39}$)(10)) = 1445.0

Example

The mean number of minutes for app engagement by a tablet user is 8.2 minutes. Suppose the standard deviation is one minute. Take a sample of size 70.

1. What are the mean and standard deviation for the sums?
2. Find the 95th percentile for the sum of the sample. Interpret this value in a complete sentence.
3. Find the probability that the sum of the sample is at least ten hours.

Solution

1. $\mu_{\Sigma X} = n\mu_X$ = 70(8.2) = 574 minutes and $\sigma_{\Sigma X} = \left(\sqrt{n}\right)\left(\sigma_x\right)$ = Formula does not parse(1) = 8.37 minutes
2. Let k = the 95th percentile.
   k = invNorm(0.95,(70)(8.2), Formula does not parse(1)) = 587.76 minutes. Ninety five percent of the app engagement times are at most 587.76 minutes.
3. ten hours = 600 minutes

$P(\Sigma x \geq 600)$ = `normalcdf`(600,E99,(70)(8.2), **Formula does not parse** (1)) = 0.0009

Your Turn!

The mean number of minutes for app engagement by a tablet user is 8.2 minutes. Suppose the standard deviation is one minute. Take a sample size of 70.

1. What is the probability that the sum of the sample is between seven hours and ten hours? What does this mean in context of the problem?
2. Find the 84th and 16th percentiles for the sum of the sample. Interpret these values in context.

Solution

1. 7 hours = 420 minutes
   10 hours = 600 minutesnormalcdf
   $P\left(420 \leq \Sigma x \leq 600\right) = normalcdf\left(420, 600, (70)\left(8.2\right), \sqrt{70}\left(1\right)\right) = 0.9991$
   This means that for these sample sums, there is a 99.9% chance that the sums of usage minutes will be between 420 minutes and 600 minutes.
2. $invNorm\left(0.84, (70)\left(8.2\right), \sqrt{70}\left(1\right)\right) = 582.32$
   $invNorm\left(0.16, (70)\left(8.2\right), \sqrt{70}\left(1\right)\right) = 565.68$
   Since 84% of the app engagement times are at most 582.32 minutes and 16% of the app engagement times are at most 565.68 minutes, we may state that 68% of the app engagement times are between 565.68 minutes and 582.32 minutes.

Note

It is important for you to understand when to use the central limit theorem and when to differentiate between the **CLT for the Mean** *versus* the **CLT for the Sum (or Total)**. If you are being asked to **find the probability of the mean, use the clt for the mean, but if you are being asked to find the probability of a sum or total**, **use the clt for sums**. This also applies to percentiles for means and sums.

Example

A study involving stress is conducted among the students on a college campus. **The stress scores follow a uniform distribution** with the lowest stress score equal to one and the highest equal to five. Using a sample of 75 students. Find:

1. The probability that the **mean stress score** for the 75 students is less than two.
2. The 90^{th} percentile for the **mean stress score** for the 75 students.
3. The probability that the **total of the 75 stress scores** is less than 200.
4. The 90^{th} percentile for the **total stress score** for the 75 students.

Let X = one stress score. Problems 1 and 2 ask you to find a **probability** or a **percentile** for a **mean**. Problems 3 and 4 ask you to find a **probability** or a **percentile** for a **total or sum**.

The sample size, n, is equal to 75. Since the individual stress scores follow a uniform distribution, $X \sim U(1, 5)$ where $a = 1$ and $b = 5$.

$\mu_X = \frac{a+b}{2} = \frac{1+5}{2} = 3$

$\sigma_X = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(5-1)^2}{12}} = 1.15$

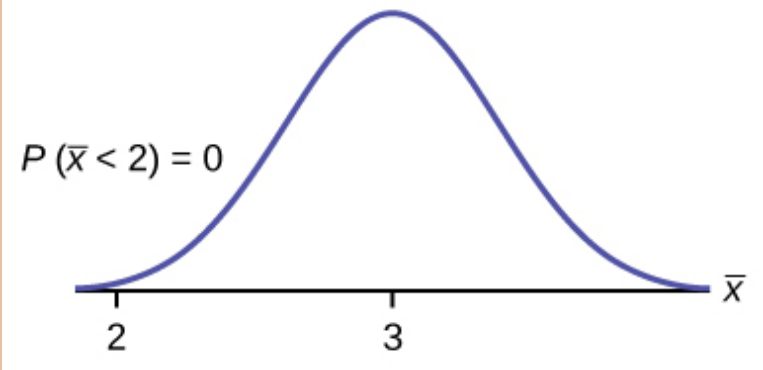
For problems 1 and 2., let $\overline{X}$ = the mean stress score for the 75 students. Then, $\overline{X} \sim N\left(3, \frac{1.15}{\sqrt{75}}\right)$ where $n = 75$.

a. Find $P(\overline{x} < 2)$.

Solution

$P(\overline{x} < 2) = 0$

The probability that the mean stress score is less than two is about zero.



normalcdf$\left(1,2,3,\frac{1.15}{\sqrt{75}}\right) = 0$

Reminder: The smallest stress score is one.

b. Find the 90th percentile for the mean of 75 stress scores. Draw a graph.

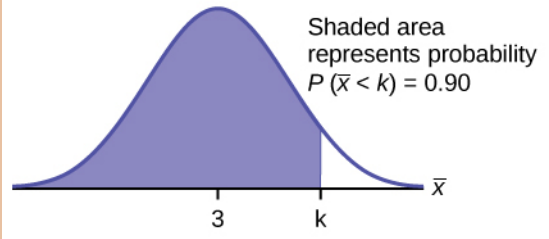
Solution

Let k = the 90th percentile.

Find k, where $P(\overline{x} < k) = 0.90$.

$k = 3.2$

This is a normal distribution curve. The peak of the curve coincides with the point 3 on the horizontal axis. A point, k, is labeled to the right of 3. A vertical line extends from k to the curve. The area under the curve to the left of k is shaded. The shaded area shows that P(x-bar < k) = 0.90



The 90th percentile for the mean of 75 scores is about 3.2. This tells us that 90% of all the means of 75 stress scores are at most 3.2, and that 10% are at least 3.2.

$\texttt{invNorm}\left(0.90, 3, \frac{1.15}{\sqrt{75}}\right)$ = 3.2

For problems 3 and 4, let ΣX = the sum of the 75 stress scores. Then, $\Sigma X \sim N[(75)(3), \left(\sqrt{75}\right)(1.15)]$

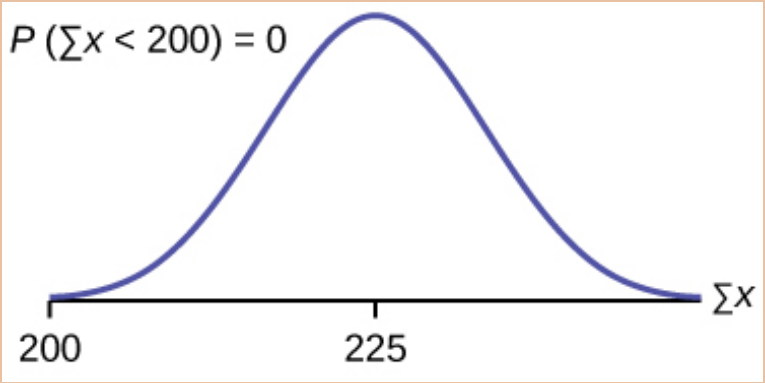
c. Find $P(\Sigma x < 200)$. Draw the graph.

Solution

The mean of the sum of 75 stress scores is (75)(3) = 225

The standard deviation of the sum of 75 stress scores is $\left(\sqrt{75}\right)(1.15)$ = 9.96

$P(\Sigma x < 200) = 0$



The probability that the total of 75 scores is less than 200 is about zero.

`normalcdf` (75,200,(75)(3),$\left(\sqrt{75}\right)$(1.15)).

Reminder: The smallest total of 75 stress scores is 75, because the smallest single score is one.

d. Find the 90th percentile for the total of 75 stress scores. Draw a graph.

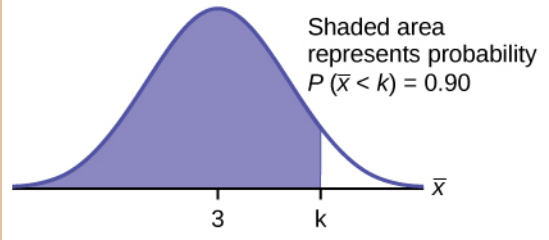
Solution

Let k = the 90th percentile.

Find k where $P(\Sigma x < k) = 0.90$.

k = 237.8

This is a normal distribution curve. The peak of the curve coincides with the point 225 on the horizontal axis. A point, k, is labeled to the right of 225. A vertical line extends from k to the curve. The area under the curve to the left of k is shaded. The shaded area shows that P(sum of x < k) = 0.90



The 90th percentile for the sum of 75 scores is about 237.8. This tells us that 90% of all the sums of 75 scores are no more than 237.8 and 10% are no less than 237.8.

`invNorm`(0.90,(75)(3),$\left(\sqrt{75}\right)$(1.15)) = 237.8

Example

In the United States, someone is sexually assaulted every two minutes, on average, according to a number of studies. Suppose the standard deviation is 0.5 minutes and the sample size is 100.

1. Find the median, the first quartile, and the third quartile for the **sample mean** time of sexual assaults in the United States.
2. Find the median, the first quartile, and the third quartile for the **sum** of sample times of sexual assaults in the United States.
3. Find the probability that a sexual assault occurs on **the average** between 1.75 and 1.85 minutes.
4. Find the value that is two standard deviations above the **sample mean**.
5. Find the *IQR* for **the sum** of the sample times.

Solution

1. We have $\mu_X = \mu = 2$ and $\sigma_X = \frac{\sigma}{\sqrt{n}} = \frac{0.5}{10} = 0.05$. Therefore:
    1. 50th percentile = $\mu_X = \mu = 2$
    2. 25th percentile = *invNorm*(0.25,2,0.05) = 1.97
    3. 75th percentile = *invNorm*(0.75,2,0.05) = 2.03
2. We have $\mu_{\Sigma X} = n(\mu_X) = 100(2) = 200$ and $\sigma_{\mu X} = \sqrt{n}(\sigma_X) = 10(0.5) = 5$. Therefore
    1. 50th percentile = $\mu_{\Sigma X} = n(\mu_X) = 100(2) = 200$
    2. 25th percentile = invNorm(0.25,200,5) = 196.63
    3. 75th percentile = invNorm(0.75,200,5) = 203.37
3. $P(1.75 < \overline{x} < 1.85)$ = `normalcdf`(1.75,1.85,2,0.05) = 0.0013
4. Using the *z*-score equation, $z = \frac{\overline{x} - \mu_{\overline{x}}}{\sigma_{\overline{x}}}$, and solving for *x*, we have *x* = 2(0.05) + 2 = 2.1
5. The *IQR* is 75th percentile – 25th percentile = 203.37 – 196.63 = 6.74

Your Turn!

A study was done about violence against prostitutes and the symptoms of the posttraumatic stress that they developed. The age range of the prostitutes was 14 to 61. The mean age was 30.9 years with a standard deviation of nine years.

1. In a sample of 25 prostitutes, what is the probability that **the mean** age of the prostitutes is less than 35?
2. Is it likely that **the mean** age of the sample group could be more than 50 years? Interpret the results.
3. In a sample of 49 prostitutes, what is the probability that **the sum** of the ages is no less than 1,600?
4. Is it likely that **the sum** of the ages of the 49 prostitutes is at most 1,595? Interpret the results.
5. Find the 95th percentile for **the sample mean** age of 65 prostitutes. Interpret the results.
6. Find the 90th percentile for **the sum** of the ages of 65 prostitutes. Interpret the results.

Solution

1. $P(\overline{x} < 35)$ = `normalcdf`(-1E99,35,30.9,1.8) = 0.9886
2. $P(\overline{x} > 50)$ = `normalcdf`(50, 1E99,30.9,1.8) ≈ 0. For this sample group, it is almost impossible for the group's average age to be more than 50. However, it is still possible for an individual in this group to have an age greater than 50.
3. $P(\Sigma x \geq 1{,}600)$ = `normalcdf`(1600, 1E99,1514.10,63) = 0.0864
4. $P(\Sigma x \leq 1{,}595)$ = `normalcdf`(-1E99, 1595, 1514.10, 63) = 0.9005. This means that there is a 90% chance that the sum of the ages for the sample group $n = 49$ is at most 1595.
5. The 95th percentile = `invNorm`(0.95,30.9,1.1) = 32.7. This indicates that 95% of the prostitutes in the sample of 65 are younger than 32.7 years, on average.
6. The 90th percentile = `invNorm`(0.90,2008.5,72.56) = 2101.5. This indicates that 90% of the prostitutes in the sample of 65 have a sum of ages less than 2,101.5 years.

References

Farago, Peter. "The Truth About Cats and Dogs: Smartphone vs Tablet Usage Differences." The Flurry Blog, October 29, 2012. Available online at https://www.flurry.com/blog (accessed May 17, 2013).

6.5 THE NORMAL APPROXIMATION TO THE BINOMIAL

Learning Objectives

By the end of this section, the student should be able to:

- Approximate the binomial distribution using the normal distribution

The **binomial formula** is cumbersome when the sample size (n) is large, particularly when we consider a range of observations. Consider the following example.

Example

Approximately 15% of the US population smokes cigarettes. A local government believed their community had a lower smoker rate and commissioned a survey of 400 randomly selected individuals. The survey found that only 42 of the 400 participants smoke cigarettes. If the true proportion of smokers in the community was really 15%, what is the probability of observing 42 or fewer smokers in a sample of 400 people?

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=824#h5p-151

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=824#h5p-152

The computations in the previous example are tedious, long, and near impossible **if you do not have access to technology.** Luckily, we have discovered the binomcdf() function previously, so this problem with the technology is not horrible to do (Try it for yourself: binomcdf(400, 0.15, 42) = 0.0054).

In some cases we may use the **normal distribution** as an easier and faster way to estimate binomial probabilities. In general, we should avoid such work if an alternative method exists that is faster, easier, and still accurate. Recall that calculating probabilities of a range of values is much easier in the normal model. We might wonder, is it reasonable to use the normal model in place of the binomial distribution? Surprisingly, yes, if certain conditions are met.

Historical Note

Historically, being able to compute binomial probabilities was one of the most important applications of the central limit theorem. Binomial probabilities with a small value for n (say, 20) were displayed in a table in a book. To calculate the probabilities with large values of n, you had to use the binomial formula, which could be very complicated. Using **the normal approximation to the binomial distribution** simplified the process.

To compute the normal approximation to the binomial distribution, take a simple random sample from a population. You must meet the conditions for a **binomial distribution**:

- there are a certain number n of independent trials
- the outcomes of any trial are success or failure
- each trial has the same probability of a success p

Recall that if X is the binomial random variable, then $X \sim B(n, p)$. The shape of the binomial distribution needs to be similar to the shape of the normal distribution. To ensure this, the quantities np and nq must both be greater than five ($np > 5$ and $nq > 5$; the approximation is better if they are both greater than or equal to 10). Then the binomial can be approximated by the normal distribution with mean $\mu = np$ and standard deviation $\sigma = \sqrt{npq}$. Remember that $q = 1 - p$.

Binomial Approximation Conditions

Consider the binomial model when the probability of a success is p = 0.10. The following figures show four hollow histograms for simulated samples from the binomial distribution using four different sample sizes: n = 10, 30, 100, 300. What happens to the shape of the distributions as the sample size increases? What distribution does the last histogram resemble?

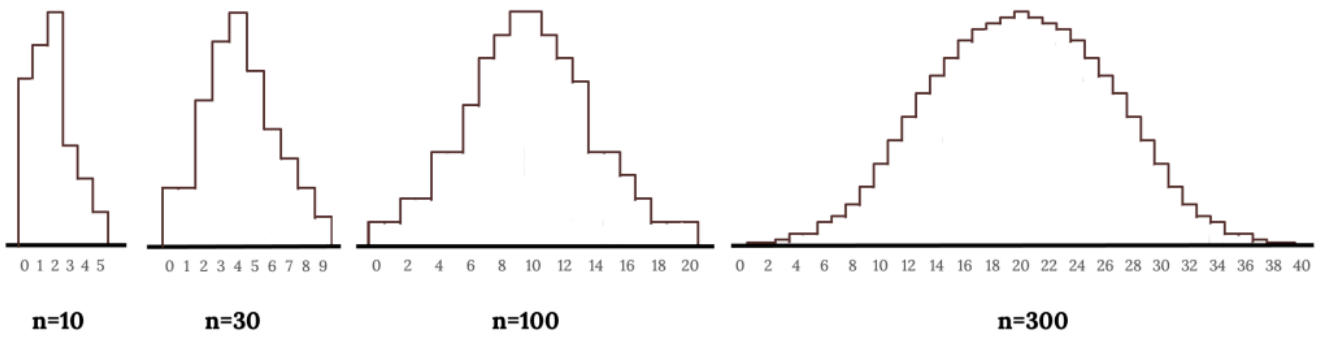


Figure 5.14. Hollow Histograms for Different Sample Sizes

It appears the distribution is transformed from a blocky and skewed distribution into one that rather resembles the normal distribution in the last hollow histogram.

The binomial distribution with probability of success p is nearly normal when the sample size n is

sufficiently large that np and n(1 − p) are both ***at least 10***. The approximate normal distribution has parameters corresponding to the mean and standard deviation of the binomial distribution: $\mu = np$ and $\sigma = np(1-p)$

The normal approximation may be used when computing the range of many possible successes. For instance, we may apply the normal distribution to the setting of the previous example:

Example

Use the normal approximation to estimate the probability of observing 42 or fewer smokers in a sample of 400, if the true proportion of smokers is p = 0.15.

Already knowing that the binomial model, we then verify that both np and n(1 − p) are at least 10:

- np = 400 × 0.15 = 60 n(1 − p) = 400 × 0.85 = 340

With these conditions met, we may use the normal approximation in place of the binomial distribution using the mean and standard deviation from the binomial model:

- $\mu = np = 60$ and $\sigma = np(1-p) = 7.14$

We want to find the probability of observing 42 or fewer smokers using this model. Use the normal model N($\mu = 60$, $\sigma = 7.14$) and standardize to estimate the probability of observing 42 or fewer smokers. Your answer should be approximately equal to the solution we found in the previous example, 0.0054.

Compute the Z-score first:

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=824#h5p-153

The Continuity Correction

The normal approximation to the binomial distribution tends to perform poorly when estimating the probability of a small range of counts, even when the conditions are met.

Suppose we wanted to compute the probability of observing 49, 50, or 51 smokers in 400 when p = 0.15. With such a large sample, we might be tempted to apply the normal approximation and use the range 49 to 51. However, we would find that the binomial solution and the normal approximation notably differ:

- Binomial: 0.0649
- Normal: 0.0421

We can identify the cause of this discrepancy in the next figure which shows the areas representing the binomial probability (outlined) and normal approximation (shaded). Notice that the width of the area under the normal distribution is 0.5 units too slim on both sides of the interval.

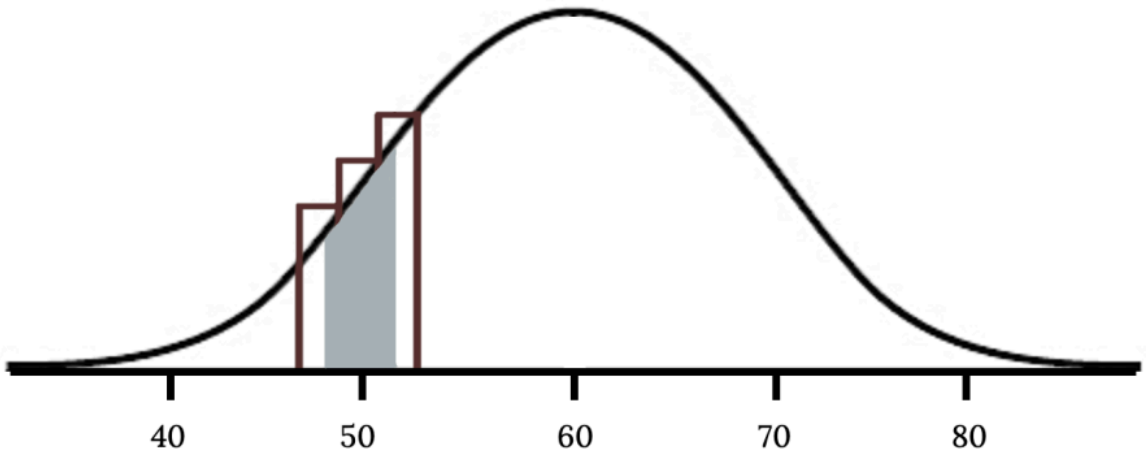


Figure 5.15. Continuity Correction

The normal approximation to the binomial distribution for intervals of values can usually be improved if cutoff values are modified slightly. The cutoff values for the lower end of a shaded region should be reduced by 0.5, and the cutoff value for the upper end should be increased by 0.5. This is called the **continuity correction**.

The tip to add extra area when applying the normal approximation is most often useful when examining a range of observations. In the example above, the revised normal distribution estimate is 0.0633, much closer to the exact value of 0.0649. While it is possible to also apply this correction when computing a tail area, the

benefit of the modification usually disappears since the total interval is typically quite wide. So, in order to get the best approximation, add 0.5 to x or subtract 0.5 from x (use $x + 0.5$ or $x - 0.5$). The number 0.5 is called the continuity correction factor and is used in the following example.

Example

Suppose in a local Kindergarten through 12th grade (K – 12) school district, 53 percent of the population favor a charter school for grades K through 5. A simple random sample of 300 is surveyed.

1. Find the probability that **at least 150** favor a charter school.
2. Find the probability that **at most 160** favor a charter school.
3. Find the probability that **more than 155** favor a charter school.
4. Find the probability that **fewer than 147** favor a charter school.
5. Find the probability that **exactly 175** favor a charter school.

Let X = the number that favors a charter school for grades K through 5. $X \sim B(n, p)$ where $n = 300$ and $p = 0.53$. Since $np > 5$ and $nq > 5$, **use the normal approximation to the binomial**.

Remember, the formulas for the mean and standard deviation are $\mu = np$ and $\sigma = \sqrt{npq}$. **The mean is 159 and the standard deviation is 8.6447.** The random variable for the normal distribution is Y: $Y \sim N(159, 8.6447)$.

Remember in Section 6.2, we learned how to use the **normalcdf()** function in the graphing calculator:

Using the TI-83+ and TI-84 Calculators for Normal Probabilities

Go into **2nd DISTR**.

Press **2 : normalcdf**.

The syntax for the instructions are as follows: **normalcdf(lower value, upper value, mean, standard deviation)**

In some instances:

- the upper number of the area might be 1E99 (= 10^{99}). You get 1E99 (= 10^{99}) by pressing 1, the EE key (a 2nd key) and then 99. Or, you can enter 10^99 instead. The number 10^{99} is way out in the right tail of the normal curve.
- the lower number of the area might be –1E99 (= -10^{99}). The number -10^{99} is way out in the left tail of the normal curve.

Use that function for this problem!

Solution

For part 1, you **include 150** so $P(X \geq 150)$ has normal approximation $P(Y \geq 149.5) = 0.8641$.

normalcdf(149.5,10^99,159,8.6447) = 0.8641.

Solution

For part 2, you **include 160** so $P(X \leq 160)$ has normal approximation $P(Y \leq 160.5) = 0.5689$.

normalcdf(0,160.5,159,8.6447) = 0.5689

Solution

For part 3, you **exclude 155** so $P(X > 155)$ has normal approximation $P(y > 155.5) = 0.6572$.

normalcdf(155.5,10^99,159,8.6447) = 0.6572.

Solution

For part 4, you **exclude 147** so $P(X < 147)$ has normal approximation $P(Y < 146.5) = 0.0741$.

`normalcdf`(0,146.5,159,8.6447) = 0.0741

Solution

For part 5, $P(X = 175)$ has normal approximation $P(174.5 < Y < 175.5) = 0.0083$.

`normalcdf`(174.5,175.5,159,8.6447) = 0.0083

Note

Because of calculators and computer software that let you calculate binomial probabilities for large values of n easily, it is not necessary to use the normal approximation to the binomial distribution, provided that you have access to these technology tools. Most school labs have Microsoft Excel, an example of computer software that calculates binomial probabilities. Many students have access to the TI-83 or 84 series calculators, and they easily calculate probabilities for the binomial distribution. If you type in "binomial probability distribution calculation" in an Internet browser, you can find at least one online calculator for the binomial.

For the previous example, the probabilities are calculated using the following binomial distribution: ($n = 300$ and $p = 0.53$). Compare the binomial and normal distribution answers.

$P(X \geq 150)$: `1 - binomcdf`(300,0.53,149) = 0.8641

$P(X \leq 160)$: `binomcdf`(300,0.53,160) = 0.5684

$P(X > 155)$: `1 - binomcdf`(300,0.53,155) = 0.6576

$P(X < 147)$: `binomcdf`(300,0.53,146) = 0.0742

$P(X = 175)$: `binompdf`(300,0.53,175) = 0.0083

Your Turn!

In a city, 46 percent of the population favor the incumbent, Dawn Morgan, for mayor. A simple random sample of 500 is taken. Using the **continuity correction factor**, find the probability that at least 250 favor Dawn Morgan for mayor.

Solution
0.0401

References

Data from the Wall Street Journal.

"National Health and Nutrition Examination Survey." Center for Disease Control and Prevention. Available online at http://www.cdc.gov/nchs/nhanes.htm (accessed May 17, 2013).

Media Attributions

- Private: Figure 5.14 © Significant Statistics by John Morgan Russell is licensed under a CC BY-SA (Attribution ShareAlike) license
- Private: Figure 5.15 © Significant Statistics by John Morgan Russell is licensed under a CC BY-SA (Attribution ShareAlike) license

CHAPTER 6 REVIEW

Chapter Review

Section 6.1

A z-score is a standardized value. Its distribution is the standard normal, $Z \sim N(0, 1)$. The mean of the z-scores is zero and the standard deviation is one. If z is the z-score for a value x from the normal distribution $N(\mu, \sigma)$ then z tells you how many standard deviations x is above (greater than) or below (less than) μ.

Section 6.2

The normal distribution, which is continuous, is the most important of all the probability distributions. Its graph is bell-shaped. This bell-shaped curve is used in almost all disciplines. Since it is a continuous distribution, the total area under the curve is one. The parameters of the normal are the mean μ and the standard deviation σ. A special normal distribution, called the standard normal distribution is the distribution of z-scores. Its mean is zero, and its standard deviation is one.

Section 6.3

In a population whose distribution may be known or unknown, if the size (n) of samples is sufficiently large, the distribution of the sample means will be approximately normal. The mean of the sample means will equal the population mean. The standard deviation of the distribution of the sample means, called the standard error of the mean, is equal to the population standard deviation divided by the square root of the sample size (n).

Section 6.4

The central limit theorem tells us that for a population with any distribution, the distribution of the sums for the sample means approaches a normal distribution as the sample size increases. In other words, if the sample size is large enough, the distribution of the sums can be approximated by a normal distribution even if the original population is not normally distributed. Additionally, if the original population has a mean of μ_X and a standard deviation of σ_X, the mean of the sums is $n\mu_X$ and the standard deviation is $\left(\sqrt{n}\right)(\sigma_X)$ where n is the sample size.

Section 6.5

The central limit theorem can be used to illustrate the law of large numbers. The law of large

numbers states that the larger the sample size you take from a population, the closer the sample mean $\overline{x}$ gets to μ.

Formula Review

Introduction – The Normal Distribution

$X \sim N(\mu, \sigma)$

μ = the mean; σ = the standard deviation

Section 6.1

z = a standardized value (z-score)

mean = 0; standard deviation = 1

To find the observed value, x, when the z-scores is known: $x = \mu + (z)\sigma$

z-score: $z = \dfrac{x - \mu}{\sigma}$

Z = the random variable for z-scores, $Z \sim N(0, 1)$

Section 6.2

Normal Distribution: $X \sim N(\mu, \sigma)$ where μ is the mean and σ is the standard deviation.

Standard Normal Distribution: $Z \sim N(0, 1)$.

Calculator function for probability: normalcdf (lower x value of the area, upper x value of the area, mean, standard deviation)

Calculator function for the k^{th} percentile: k = invNorm (area to the left of k, mean, standard deviation)

Section 6.3

The Central Limit Theorem for Sample Means: $\overline{X} \sim N\left(\mu_x, \frac{\sigma x}{\sqrt{n}}\right)$

The Mean $\overline{X}$: μ_X

Central Limit Theorem for Sample Means z-score and standard error of the mean:

$$z = \frac{\overline{x} - \mu_x}{\left(\frac{\sigma_x}{\sqrt{n}}\right)}$$

Standard Error of the Mean (Standard Deviation ($\overline{X}$)): $\frac{\sigma_x}{\sqrt{n}}$

Section 6.4

The Central Limit Theorem for Sums: $\sum X \sim N[(n)(\mu_X),(\sqrt{n})(\sigma_X)]$

Mean for Sums ($\sum X$): $(n)(\mu_X)$

The Central Limit Theorem for Sums z-score and standard deviation for sums:

$$z \text{ for the sample mean} = \frac{\Sigma x - (n)\,(\mu_X)}{(\sqrt{n})\,(\sigma_X)}$$

Standard deviation for Sums ($\sum X$): $(\sqrt{n})(\sigma_X)$

CHAPTER 6 PRACTICE

Section 6.1 Practice

A bottle of water contains 12.05 fluid ounces with a standard deviation of 0.01 ounces. Define the random variable X in words. X = _____.

Solution
ounces of water in a bottle

A normal distribution has a mean of 61 and a standard deviation of 15. What is the median?

Solution
61

If X ~ $N(1, 2)$.

σ = _____

Solution

2

A company manufactures rubber balls. The mean diameter of a ball is 12 cm with a standard deviation of 0.2 cm. Define the random variable X in words. X = _____.

Solution
diameter of a rubber ball

If $X \sim N(-4, 1)$.

What is the median?

Solution
–4

If $X \sim N(3, 5)$

σ = _____

Solution
5

If $X \sim N(-2, 1)$

$\mu =$ _____

Solution

–2

What does a z-score measure?

Solution

The number of standard deviations a value is from the mean.

What does standardizing a normal distribution do to the mean?

Solution

The mean becomes zero.

Is $X \sim N(0, 1)$ a standardized normal distribution? Why or why not?

Solution

Yes because the mean is zero, and the standard deviation is one.

What is the z-score of $x = 12$, if it is two standard deviations to the right of the mean?

Solution

$z = 2$

What is the z-score of $x = 9$, if it is 1.5 standard deviations to the left of the mean?

Solution

z = –1.5 –>

What is the z-score of $x = –2$, if it is 2.78 standard deviations to the right of the mean?

Solution

$z = 2.78$

What is the z-score of $x = 7$, if it is 0.133 standard deviations to the left of the mean?

Solution

z = –0.133

Suppose $X \sim N(2, 6)$. What value of x has a z-score of three?

Solution

$x = 20$

Suppose $X \sim N(8, 1)$. What value of x has a z-score of –2.25?

Solution

x = 5.75

Suppose $X \sim N(9, 5)$. What value of x has a z-score of –0.5?

Solution

$x = 6.5$

Suppose $X \sim N(2, 3)$. What value of x has a z-score of –0.67?

Solution

x = –0.01

Suppose $X \sim N(4, 2)$. What value of x is 1.5 standard deviations to the left of the mean?

Solution

$x = 1$

Suppose $X \sim N(4, 2)$. What value of x is two standard deviations to the right of the mean?

Solution

x = 8

Suppose $X \sim N(8, 9)$. What value of x is 0.67 standard deviations to the left of the mean?

Solution

$x = 1.97$

Suppose $X \sim N(-1, 2)$. What is the z-score of $x = 2$?

Solution

z = 1.5

Suppose $X \sim N(12, 6)$. What is the z-score of $x = 2$?

Solution

$z = -1.67$

Suppose $X \sim N(9, 3)$. What is the z-score of $x = 9$?

Solution

z = 0

Suppose a normal distribution has a mean of six and a standard deviation of 1.5. What is the z-score of $x = 5.5$?

Solution

$z \approx -0.33$

In a normal distribution, $x = 5$ and $z = -1.25$. This tells you that $x = 5$ is ____ standard deviations to the ____ (right or left) of the mean.

Solution

1.25, left

In a normal distribution, $x = 3$ and $z = 0.67$. This tells you that $x = 3$ is ____ standard deviations to the ____ (right or left) of the mean.

Solution
0.67, right

In a normal distribution, $x = –2$ and $z = 6$. This tells you that $x = –2$ is ____ standard deviations to the ____ (right or left) of the mean.

Solution
six, right

In a normal distribution, $x = –5$ and $z = –3.14$. This tells you that $x = –5$ is ____ standard deviations to the ____ (right or left) of the mean.

Solution
3.14, left

In a normal distribution, $x = 6$ and $z = -1.7$. This tells you that $x = 6$ is ____ standard deviations to the ____ (right or left) of the mean.

Solution
1.7, left

About what percent of x values from a normal distribution lie within one standard deviation (left and right) of the mean of that distribution?

Solution
about 68%

About what percent of the x values from a normal distribution lie within two standard deviations (left and right) of the mean of that distribution?

Solution
about 95.45%

About what percent of x values lie between the second and third standard deviations (both sides)?

Solution
about 4%

Suppose $X \sim N(15, 3)$. Between what x values does 68.27% of the data lie? The range of x values is centered at the mean of the distribution (i.e., 15).

Solution
between 12 and 18

Suppose $X \sim N(-3, 1)$. Between what x values does 95.45% of the data lie? The range of x values is centered at the mean of the distribution(i.e., –3).

Solution
between –5 and –1

Suppose $X \sim N(-3, 1)$. Between what x values does 34.14% of the data lie?

Solution
between –4 and –3 or between –3 and –2

About what percent of *x* values lie between the mean and three standard deviations?

Solution
about 50%

About what percent of *x* values lie between the mean and one standard deviation?

Solution
about 34.14%

About what percent of *x* values lie between the first and second standard deviations from the mean (both sides)?

Solution
about 27%

About what percent of *x* values lie between the first and third standard deviations(both sides)?

Solution

about 34.46%

Use the following information to answer the next two exercises:

The life of Sunshine CD players is normally distributed with a mean of 4.1 years and a standard deviation of 1.3 years. A CD player is guaranteed for three years. We are interested in the length of time a CD player lasts.

1. Define the random variable X in words. X = _____.

Solution

The lifetime of a Sunshine CD player measured in years.

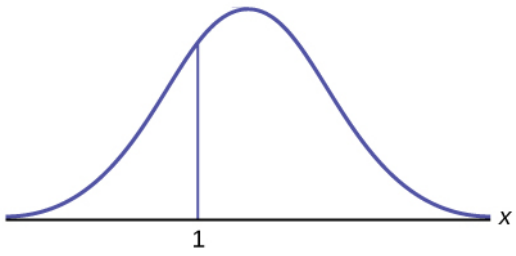
2. The distribution is: X ~ _____(_____,_____)

Solution

X ~ N(4.1, 1.3)

Section 6.2 Practice

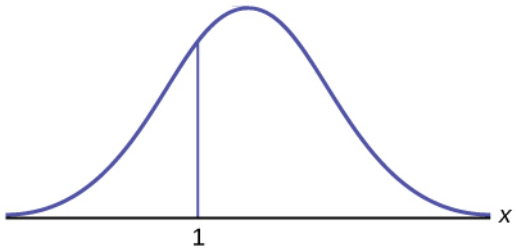
How would you represent the area to the left of one in a probability statement?



Solution

$P(x < 1)$

What is the area to the right of one?



Solution

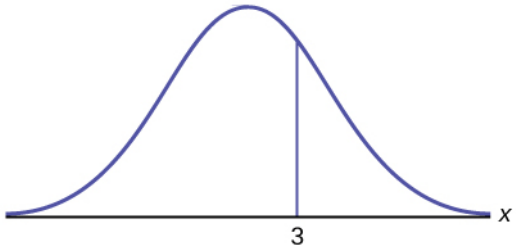
1 – P(x < 1) or P(x < 1)

Is $P(x < 1)$ equal to $P(x \leq 1)$? Why?

Solution

Yes, because they are the same in a continuous distribution: $P(x = 1) = 0$

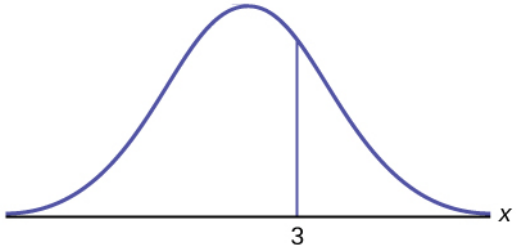
How would you represent the area to the left of three in a probability statement?



Solution

P(x < 3)

What is the area to the right of three?



Solution

$1 - P(x < 3)$ or $P(x > 3)$

If the area to the left of x in a normal distribution is 0.123, what is the area to the right of x?

Solution

1 – 0.123 = 0.877

If the area to the right of x in a normal distribution is 0.543, what is the area to the left of x?

Solution

1 – 0.543 = 0.457

Use the following information to answer the next four exercises:

$X \sim N(54, 8)$

1. Find the probability that $x > 56$.

Solution

0.4013

2. Find the probability that $x < 30$.

Solution

0.0013

3. Find the 80^{th} percentile.

Solution
60.73

4. Find the 60^{th} percentile.

Solution
56.03

$X \sim N(6, 2)$

Find the probability that x is between three and nine.

Solution
0.8664

$X \sim N(-3, 4)$

Find the probability that x is between one and four.

Solution
0.1186

$X \sim N(4, 5)$

Find the maximum of x in the bottom quartile.

Solution

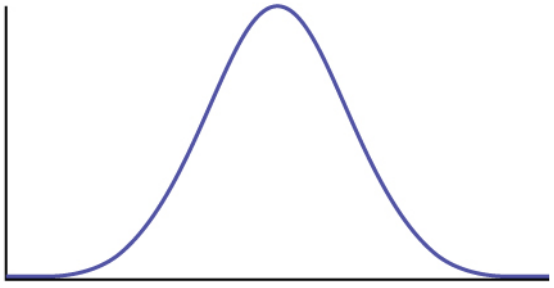
0.6276

Use the following information to answer the next three exercise:

The life of Sunshine CD players is normally distributed with a mean of 4.1 years and a standard deviation of 1.3 years. A CD player is guaranteed for three years. We are interested in the length of time a CD player lasts.

1. Find the probability that a CD player will break down during the guarantee period.

a. Sketch the situation. Label and scale the axes. Shade the region corresponding to the probability.



b. $P(0 < x <$ _____) = _____ (Use zero for the minimum value of x.)

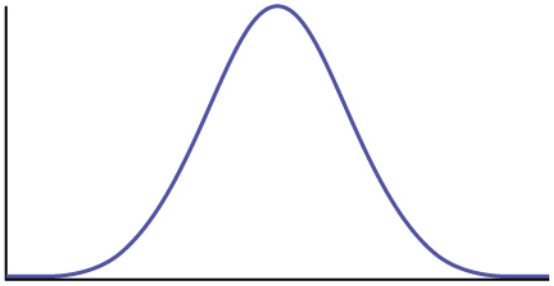
Solution

a. Check student's solution.

b. 3, 0.1979

2. Find the probability that a CD player will last between 2.8 and six years.

a. Sketch the situation. Label and scale the axes. Shade the region corresponding to the probability.



b. P(_____ < x < _____) = _____

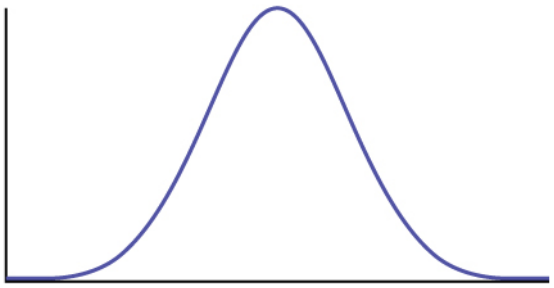
Solution

a. Check student's solution

b. 2.8, 6, 0.7694

3. Find the 70^{th} percentile of the distribution for the time a CD player lasts.

a. Sketch the situation. Label and scale the axes. Shade the region corresponding to the lower 70%.



b. P(x < k) = _____ Therefore, k = _____

Solution

a. Check student's solution.

b. 0.70, 4.78 years

Section 6.3 Practice

Use the following information to answer the next six exercises:

Yoonie is a personnel manager in a large corporation. Each month she must review 16 of the employees. From past experience, she has found that the reviews take her approximately four hours each to do with a population standard deviation of 1.2 hours. Let X be the random variable representing the time it takes her to complete one review. Assume X is normally distributed. Let $\overline{X}$ be the random variable representing the mean time to complete the 16 reviews. Assume that the 16 reviews represent a random set of reviews.

1. What is the mean, standard deviation, and sample size?

Solution

mean = 4 hours; standard deviation = 1.2 hours; sample size = 16

2. Complete the distributions.

X ~ _____(_____, _____)

$\overline{X}$ ~ _____(_____, _____)

Solution

X ~ N(4, 1.2).

$\overline{X}$ ~ N(4, 1.216)

3. Find the probability that **one** review will take Yoonie from 3.5 to 4.25 hours.

a. Sketch the graph, labeling and scaling the horizontal axis. Shade the region corresponding to the probability.



b. P(_____ < x < _____) = _____

Solution

a. Check student's solution.

b. 3.5, 4.25, 0.2441

4. Find the probability that the **mean** of a month's reviews will take Yoonie from 3.5 to 4.25 hrs.

a. Sketch the graph, labeling and scaling the horizontal axis. Shade the region corresponding to the probability.



b. P(_____) = _____

Solution

a. Check student's solution.

b. 0.7499

5. What causes the probabilities in [link] and [link] to be different?

Solution

The fact that the two distributions are different accounts for the different probabilities.

6. Find the 95th percentile for the mean time to complete one month's reviews.

a. Sketch the graph.



b. The 95th Percentile =_____

Solution

P(3.5 < $\overline{x}$ < 4.25) = `invNorm`(95,4,1.216) = 4.49

Section 6.4 Practice

Use the following information to answer the next four exercises:

An unknown distribution has a mean of 80 and a standard deviation of 12. A sample size of 95 is drawn randomly from the population.

1. Find the probability that the sum of the 95 values is greater than 7,650.

Solution

0.3345

2. Find the probability that the sum of the 95 values is less than 7,400.

Solution

0.0436

3. Find the sum that is two standard deviations above the mean of the sums.

Solution
7,833.92

4. Find the sum that is 1.5 standard deviations below the mean of the sums.

Solution
7,424.56

Use the following information to answer the next five exercises:

The distribution of results from a cholesterol test has a mean of 180 and a standard deviation of 20. A sample size of 40 is drawn randomly.

1. Find the probability that the sum of the 40 values is greater than 7,500.

Solution
0.0089

2. Find the probability that the sum of the 40 values is less than 7,000.

Solution

0.0569

3. Find the sum that is one standard deviation above the mean of the sums.

Solution
7,326.49

4. Find the sum that is 1.5 standard deviations below the mean of the sums.

Solution
7,010.26

5. Find the percentage of sums between 1.5 standard deviations below the mean of the sums and one standard deviation above the mean of the sums.

Solution
77.45%

Use the following information to answer the next six exercises:

A researcher measures the amount of sugar in several cans of the same soda. The mean is 39.01 with a standard deviation of 0.5. The researcher randomly selects a sample of 100.

1. Find the probability that the sum of the 100 values is greater than 3,910.

Solution

0.0359

2. Find the probability that the sum of the 100 values is less than 3,900.

Solution
0.4207

3. Find the probability that the sum of the 100 values falls between the numbers you found in [link] and [link].

Solution
0.5433

4. Find the sum with a *z*–score of –2.5.

Solution
3,888.5

5. Find the sum with a *z*–score of 0.5.

Solution
3,903.5

6. Find the probability that the sums will fall between the *z*-scores –2 and 1.

Solution
0.8186

Use the following information to answer the next four exercise:

An unknown distribution has a mean 12 and a standard deviation of one. A sample size of 25 is taken. Let X = the object of interest.

1. What is the mean of ΣX?

Solution
300

2. What is the standard deviation of ΣX?

Solution
5

3. What is $P(\Sigma x = 290)$?

Solution
0

4. What is $P(\Sigma x > 290)$?

Solution
0.9772

True or False: Only the sums of normal distributions are also normal distributions.

Solution
False, the sums of any distribution approach a normal distribution as the sample size increases.

In order for the sums of a distribution to approach a normal distribution, what must be true?

Solution
The sample size, n, gets larger.

What three things must you know about a distribution to find the probability of sums?

Solution
the mean of the distribution, the standard deviation of the distribution, and the sample size –>

An unknown distribution has a mean of 25 and a standard deviation of six. Let X = one object from this distribution. What is the sample size if the standard deviation of $\boldsymbol{\Sigma}X$ is 42?

Solution
49

An unknown distribution has a mean of 19 and a standard deviation of 20. Let X = the object of interest. What is the sample size if the mean of ΣX is 15,200?

Solution
800

Use the following information to answer the next three exercises:

A market researcher analyzes how many electronics devices customers buy in a single purchase. The distribution has a mean of three with a standard deviation of 0.7. She samples 400 customers.

1. What is the z-score for $\Sigma x = 840$?

Solution
26.00

2. What is the z-score for $\Sigma x = 1{,}186$?

Solution
–1

3. What is $P(\Sigma x < 1{,}186)$?

Solution
0.1587

Use the following information to answer the next three exercises:

An unknown distribution has a mean of 100, a standard deviation of 100, and a sample size of 100. Let X = one object of interest.

1. What is the mean of ΣX?

Solution
10,000

2. What is the standard deviation of ΣX?

Solution
1,000

3. What is $P(\Sigma x > 9{,}000)$?

Solution
0.8413

Section 6.5 Practice

Use the following information to answer the next ten exercises:

A manufacturer produces 25-pound lifting weights. The lowest actual weight is 24 pounds, and the highest is 26 pounds. Each weight is equally likely so the distribution of weights is uniform. A sample of 100 weights is taken.

1. Answer the following:

a. What is the distribution for the weights of one 25-pound lifting weight? What is the mean and standard deviation?

b. What is the distribution for the mean weight of 100 25-pound lifting weights?

c. Find the probability that the mean actual weight for the 100 weights is less than 24.9.

Solution

a. $U(24, 26)$, 25, 0.5774

b. $N(25, 0.0577)$

c. 0.0416

2. Draw the graph from [link]

Solution

3. Find the probability that the mean actual weight for the 100 weights is greater than 25.2.

Solution

0.0003

4. Draw the graph from [link]

Solution

5. Find the 90th percentile for the mean weight for the 100 weights.

Solution
25.07

6. Draw the graph from [link]

Solution

7. Answer the following:

a. What is the distribution for the sum of the weights of 100 25-pound lifting weights?

b. Find $P(\Sigma x < 2{,}450)$.

Solution
a. $N(2{,}500, 5.7735)$

b. 0

8. Draw the graph from [link]

Solution

9. Find the 90th percentile for the total weight of the 100 weights.

Solution
2,507.40

10. Draw the graph from [link]

Solution

Use the following information to answer the next eight exercises:

The length of time a particular smartphone's battery lasts follows an exponential distribution with a mean of ten months. A sample of 64 of these smartphones is taken.

1. Answer the following:

a. What is the standard deviation?

b. What is the parameter *m*?

Solution
a. 10

b. $\frac{1}{10}$

2. What is the distribution for the length of time one battery lasts?

Solution
Exp(1/10)

3. What is the distribution for the mean length of time 64 batteries last?

Solution

$N\left(10, \frac{10}{8}\right)$

4. What is the distribution for the total length of time 64 batteries last?

Solution

N(640, 80)

5. Find the probability that the sample mean is between seven and 11.

Solution

0.7799

6. Find the 80^{th} percentile for the total length of time 64 batteries last.

Solution

707.3 –>

7. Find the *IQR* for the mean amount of time 64 batteries last.

Solution

1.69

8. Find the middle 80% for the total amount of time 64 batteries last.

Solution
205.05

Use the following information to answer the next six exercises:

A uniform distribution has a minimum of six and a maximum of ten. A sample of 50 is taken.

1. Find $P(\Sigma x > 420)$.

Solution
0.0072

2. Find the 90^{th} percentile for the sums.

Solution
410.46

3. Find the 15^{th} percentile for the sums.

Solution
391.54

4. Find the first quartile for the sums.

Solution
394.49

5. Find the third quartile for the sums.

Solution
405.51

Find the 80^{th} percentile for the sums.

Solution
406.87

CHAPTER 6 HOMEWORK

Section 6.1 Homework

Use the following information to answer the next two exercises:

The patient recovery time from a particular surgical procedure is normally distributed with a mean of 5.3 days and a standard deviation of 2.1 days.

What is the median recovery time?

1. 2.7
2. 5.3
3. 7.4
4. 2.1

Solution

5.3

What is the *z*-score for a patient who takes ten days to recover?

1. 1.5
2. 0.2
3. 2.2
4. 7.3

Solution

2.2

The length of time to find it takes to find a parking space at 9 A.M. follows a normal distribution with a mean of five minutes and a standard deviation of two minutes. If the mean is significantly greater than the standard deviation, which of the following statements is *true?*

1. The data cannot follow the uniform distribution.
2. The data cannot follow the exponential distribution.
3. The data cannot follow the normal distribution.

1. I only
2. II only
3. III only
4. I, II, and III

Solution

II only

The heights of the 430 National Basketball Association players were listed on team rosters at the start of the 2005–2006 season. The heights of basketball players have an approximate normal distribution with mean, $\mu = 79$ inches and a standard deviation, $\sigma = 3.89$ inches. For each of the following heights, calculate the z-score and interpret it using complete sentences.

1. 77 inches
2. 85 inches
3. If an NBA player reported his height had a z-score of 3.5, would you believe him? Explain your answer.

Solution

1. Use the z-score formula. $z = -0.5141$. The height of 77 inches is 0.5141 standard deviations below the mean. An NBA player whose height is 77 inches is shorter than

average.

2. Use the *z*-score formula. $z = 1.5424$. The height 85 inches is 1.5424 standard deviations above the mean. An NBA player whose height is 85 inches is taller than average.
3. Height = 79 + 3.5(3.89) = 90.67 inches, which is over 7.7 feet tall. There are very few NBA players this tall so the answer is no, not likely.

The systolic blood pressure (given in millimeters) of males has an approximately normal distribution with mean $\mu = 125$ and standard deviation $\sigma = 14$. Systolic blood pressure for males follows a normal distribution.

1. Calculate the *z*-scores for the male systolic blood pressures 100 and 150 millimeters.
2. If a male friend of yours said he thought his systolic blood pressure was 2.5 standard deviations below the mean, but that he believed his blood pressure was between 100 and 150 millimeters, what would you say to him?

Solution

1. $z \approx -1.8$ and $z \approx 1.8$
2. I would tell him that 2.5 standard deviations below the mean would give him a blood pressure reading of 90, which is below the range of 100 to 150.

Kyle's doctor told him that the *z*-score for his systolic blood pressure is 1.75. Which of the following is the best interpretation of this standardized score? The systolic blood pressure (given in millimeters) of males has an approximately normal distribution with mean $\mu = 125$ and standard deviation $\sigma = 14$. If $X =$ a systolic blood pressure score then $X \sim N(125, 14)$.

1. Which answer(s) **is/are** correct?

1. Kyle's systolic blood pressure is 175.
2. Kyle's systolic blood pressure is 1.75 times the average blood pressure of men his age.
3. Kyle's systolic blood pressure is 1.75 above the average systolic blood pressure of men his age.
4. Kyle's systolic blood pressure is 1.75 standard deviations above the average systolic blood pressure for men.

2. Calculate Kyle's blood pressure.

Solution

1. Kyle's systolic blood pressure is 1.75 standard deviations above the average systolic blood pressure for men.
2. Kyle's blood pressure is equal to 125 + (1.75)(14) = 149.5.

Height and weight are two measurements used to track a child's development. The World Health Organization measures child development by comparing the weights of children who are the same height and the same gender. In 2009, weights for all 80 cm girls in the reference population had a mean μ = 10.2 kg and standard deviation $\boldsymbol{\sigma}$ = 0.8 kg. Weights are normally distributed. $X \sim N(10.2, 0.8)$.

Calculate the z-scores that correspond to the following weights and interpret them:

1. 11 kg
2. 7.9 kg
3. 12.2 kg

Solution

1. z = 1; A child who weighs 11 kg is one standard deviation above the mean of 10.2 kg.

2. z = –2.875; A child who weighs 7.9 kg is 2.875 standard deviations below the mean of 10.2 kg.
3. z = 2.5; A child who weighs 12.2 kg is 2.5 standard deviation above the mean of 10.2 kg.

In 2005, 1,475,623 students heading to college took the SAT. The distribution of scores in the math section of the SAT follows a normal distribution with mean $\mu = 520$ and standard deviation $\sigma = 115$.

1. Calculate the *z*-score for an SAT score of 720. Interpret it using a complete sentence.
2. What math SAT score is 1.5 standard deviations above the mean? What can you say about this SAT score?
3. For 2012, the SAT math test had a mean of 514 and standard deviation 117. The ACT math test is an alternate to the SAT and is approximately normally distributed with mean 21 and standard deviation 5.3. If one person took the SAT math test and scored 700 and a second person took the ACT math test and scored 30, who did better with respect to the test they took?

Solution

Let X = an SAT math score and Y = an ACT math score.

1. $X = 720$ $\frac{720 - 520}{15} = 1.74$. The exam score of 720 is 1.74 standard deviations above the mean of 520.
2. $z = 1.5$;The math SAT score is 520 + 1.5(115) ≈ 692.5. The exam score of 692.5 is 1.5 standard deviations above the mean of 520.
3. $\frac{X - \mu}{\sigma} = \frac{700 - 514}{117} \approx 1.59$, the *z*-score for the SAT. $\frac{Y - \mu}{\sigma} = \frac{30 - 21}{5.3} \approx 1.70$, the *z*-scores for the ACT. With respect to the test they took, the person who took the ACT did better (has the higher *z*-score).

Section 6.2 Homework

Use the following information to answer the next two exercises:

The patient recovery time from a particular surgical procedure is normally distributed with a mean of 5.3 days and a standard deviation of 2.1 days.

What is the probability of spending more than two days in recovery?

1. 0.0580
2. 0.8447
3. 0.0553
4. 0.9420

Solution
0.9420

The 90^{th} percentile for recovery times is?

1. 8.89
2. 7.07
3. 7.99
4. 4.32

Solution
7.99

Use the following information to answer the next three exercises:

The length of time it takes to find a parking space at 9 A.M. follows a normal distribution with a mean of five minutes and a standard deviation of two minutes.

Based upon the given information and numerically justified, would you be surprised if it took less than one minute to find a parking space?

1. Yes
2. No
3. Unable to determine

Solution
Yes

Find the probability that it takes at least eight minutes to find a parking space.

1. 0.0001
2. 0.9270
3. 0.1862
4. 0.0668

Solution
0.0668

Seventy percent of the time, it takes more than how many minutes to find a parking space?

1. 1.24
2. 2.41
3. 3.95
4. 6.05

Solution

3.95

According to a study done by De Anza students, the height for Asian adult males is normally distributed with an average of 66 inches and a standard deviation of 2.5 inches. Suppose one Asian adult male is randomly chosen. Let X = height of the individual.

1. X ~ _____(_____, _____)
2. Find the probability that the person is between 65 and 69 inches. Include a sketch of the graph, and write a probability statement.
3. Would you expect to meet many Asian adult males over 72 inches? Explain why or why not, and justify your answer numerically.
4. The middle 40% of heights fall between what two values? Sketch the graph, and write the probability statement.

Solution

1. $X \sim N(66, 2.5)$
2. 0.5404
3. No, the probability that an Asian male is over 72 inches tall is 0.0082
4.

IQ is normally distributed with a mean of 100 and a standard deviation of 15. Suppose one individual is randomly chosen. Let X = IQ of an individual.

1. X ~ _____(_____, _____)
2. Find the probability that the person has an IQ greater than 120. Include a sketch of the graph, and write a probability statement.

3. MENSA is an organization whose members have the top 2% of all IQs. Find the minimum IQ needed to qualify for the MENSA organization. Sketch the graph, and write the probability statement.
4. The middle 50% of IQs fall between what two values? Sketch the graph and write the probability statement.

Solution

1. $X \sim N(100, 15)$
2. The probability that a person has an IQ greater than 120 is 0.0918.
3. A person has to have an IQ over 130 to qualify for MENSA.
4. The middle 50% of IQ scores falls between 89.95 and 110.05.

The percent of fat calories that a person in America consumes each day is normally distributed with a mean of about 36 and a standard deviation of 10. Suppose that one individual is randomly chosen. Let X = percent of fat calories.

1. $X \sim$ _____(_____, _____)
2. Find the probability that the percent of fat calories a person consumes is more than 40. Graph the situation. Shade in the area to be determined.
3. Find the maximum number for the lower quarter of percent of fat calories. Sketch the graph and write the probability statement.

Solution

1. $X \sim N(36, 10)$
2. The probability that a person consumes more than 40% of their calories as fat is 0.3446.
3. Approximately 25% of people consume less than 29.26% of their calories as fat.

Suppose that the distance of fly balls hit to the outfield (in baseball) is normally distributed with a mean of 250 feet and a standard deviation of 50 feet.

1. If X = distance in feet for a fly ball, then X ~ _____(_____, _____)
2. If one fly ball is randomly chosen from this distribution, what is the probability that this ball traveled fewer than 220 feet? Sketch the graph. Scale the horizontal axis X. Shade the region corresponding to the probability. Find the probability.
3. Find the 80^{th} percentile of the distribution of fly balls. Sketch the graph, and write the probability statement.

Solution

1. X ~ N(250, 50)
2. The probability that a fly ball travels less than 220 feet is 0.2743.
3. Eighty percent of the fly balls will travel less than 292 feet.

In China, four-year-olds average three hours a day unsupervised. Most of the unsupervised children live in rural areas, considered safe. Suppose that the standard deviation is 1.5 hours and the amount of time spent alone is normally distributed. We randomly select one Chinese four-year-old living in a rural area. We are interested in the amount of time the child spends alone per day.

1. In words, define the random variable X.
2. X ~ _____(_____,_____)
3. Find the probability that the child spends less than one hour per day unsupervised. Sketch the graph, and write the probability statement.
4. What percent of the children spend over ten hours per day unsupervised?
5. Seventy percent of the children spend at least how long per day unsupervised?

Solution

1. X = number of hours that a Chinese four-year-old in a rural area is unsupervised during the day.
2. $X \sim N(3, 1.5)$
3. The probability that the child spends less than one hour a day unsupervised is 0.0918.
4. The probability that a child spends over ten hours a day unsupervised is less than 0.0001.
5. 2.21 hours

In the 1992 presidential election, Alaska's 40 election districts averaged 1,956.8 votes per district for President Clinton. The standard deviation was 572.3. (There are only 40 election districts in Alaska.) The distribution of the votes per district for President Clinton was bell-shaped. Let X = number of votes for President Clinton for an election district.

1. State the approximate distribution of X.
2. Is 1,956.8 a population mean or a sample mean? How do you know?
3. Find the probability that a randomly selected district had fewer than 1,600 votes for President Clinton. Sketch the graph and write the probability statement.
4. Find the probability that a randomly selected district had between 1,800 and 2,000 votes for President Clinton.
5. Find the third quartile for votes for President Clinton.

Solution

1. $X \sim N(1956.8, 572.3)$
2. This is a population mean, because all election districts are included.
3. The probability that a district had less than 1,600 votes for President Clinton is 0.2676.
4. 0.3798
5. Seventy-five percent of the districts had fewer than 2,340 votes for President Clinton.

Suppose that the duration of a particular type of criminal trial is known to be normally distributed with a mean of 21 days and a standard deviation of 7 days.

1. In words, define the random variable *X*.
2. *X* ~ _____(_____,_____)
3. If one of the trials is randomly chosen, find the probability that it lasted at least 24 days. Sketch the graph and write the probability statement.
4. Sixty percent of all trials of this type are completed within how many days?

Solution

1. *X* = the distribution of the number of days a particular type of criminal trial will take
2. *X* ~ *N*(21, 7)
3. The probability that a randomly selected trial will last more than 24 days is 0.3336.
4. 22.77

Terri Vogel, an amateur motorcycle racer, averages 129.71 seconds per 2.5 mile lap (in a seven-lap race) with a standard deviation of 2.28 seconds. The distribution of her race times is normally distributed. We are interested in one of her randomly selected laps.

1. In words, define the random variable *X*.
2. *X* ~ _____(_____,_____)
3. Find the percent of her laps that are completed in less than 130 seconds.
4. The fastest 3% of her laps are under _____.
5. The middle 80% of her laps are from _____ seconds to _____ seconds.

Solution

1. X = the distribution of race times that Terry Vogel produces
2. X ~ N(129.71, 2.28)

3. Terri completes 55.17% of her laps in less than 130 seconds.
4. Terri completes 55.17% of her laps in less than 130 seconds.
5. 124.4 and 135.02

Thuy Dau, Ngoc Bui, Sam Su, and Lan Voung conducted a survey as to how long customers at Lucky claimed to wait in the checkout line until their turn. Let X = time in line. [link] displays the ordered real data (in minutes):

0.50	4.25	5	6	7.25
1.75	4.25	5.25	6	7.25
2	4.25	5.25	6.25	7.25
2.25	4.25	5.5	6.25	7.75
2.25	4.5	5.5	6.5	8
2.5	4.75	5.5	6.5	8.25
2.75	4.75	5.75	6.5	9.5
3.25	4.75	5.75	6.75	9.5
3.75	5	6	6.75	9.75
3.75	5	6	6.75	10.75

1. Calculate the sample mean and the sample standard deviation.
2. Construct a histogram.
3. Draw a smooth curve through the midpoints of the tops of the bars.
4. In words, describe the shape of your histogram and smooth curve.
5. Let the sample mean approximate μ and the sample standard deviation approximate σ. The distribution of X can then be approximated by $X \sim$ _____(_____,_____)
6. Use the distribution in part e to calculate the probability that a person will wait fewer than 6.1 minutes.
7. Determine the cumulative relative frequency for waiting less than 6.1 minutes.
8. Why aren't the answers to part f and part g exactly the same?

9. Why are the answers to part f and part g as close as they are?
10. If only ten customers have been surveyed rather than 50, do you think the answers to part f and part g would have been closer together or farther apart? Explain your conclusion.

Solution

1. mean = 5.51, s = 2.15
2. Check student's solution.
3. Check student's solution.
4. Check student's solution.
5. $X \sim N(5.51, 2.15)$
6. 0.6029
7. The cumulative frequency for less than 6.1 minutes is 0.64.
8. The answers to part f and part g are not exactly the same, because the normal distribution is only an approximation to the real one.
9. The answers to part f and part g are close, because a normal distribution is an excellent approximation when the sample size is greater than 30.
10. The approximation would have been less accurate, because the smaller sample size means that the data does not fit the normal curve as well.

Suppose that Ricardo and Anita attend different colleges. Ricardo's GPA is the same as the average GPA at his school. Anita's GPA is 0.70 standard deviations above her school average. In complete sentences, explain why each of the following statements may be false.

1. Ricardo's actual GPA is lower than Anita's actual GPA.
2. Ricardo is not passing because his z-score is zero.
3. Anita is in the 70^{th} percentile of students at her college.

Solution

1. If the average GPA is less at Anita's school than it is at Ricardo's, then Ricardo's actual score could be higher.
2. Passing can be defined differently at different schools. Also, since Ricardo's z-score is 0, his GPA is actually the average for his school, which is typically a passing GPA.
3. Anita's percentile is higher than the 70th percentile.

[link] shows a sample of the maximum capacity (maximum number of spectators) of sports stadiums. The table does not include horse-racing or motor-racing stadiums.

40,000	40,000	45,050	45,500	46,249	48,134
49,133	50,071	50,096	50,466	50,832	51,100
51,500	51,900	52,000	52,132	52,200	52,530
52,692	53,864	54,000	55,000	55,000	55,000
55,000	55,000	55,000	55,082	57,000	58,008
59,680	60,000	60,000	60,492	60,580	62,380
62,872	64,035	65,000	65,050	65,647	66,000
66,161	67,428	68,349	68,976	69,372	70,107
70,585	71,594	72,000	72,922	73,379	74,500
75,025	76,212	78,000	80,000	80,000	82,300

1. Calculate the sample mean and the sample standard deviation for the maximum capacity of sports stadiums (the data).
2. Construct a histogram.
3. Draw a smooth curve through the midpoints of the tops of the bars of the histogram.
4. In words, describe the shape of your histogram and smooth curve.
5. Let the sample mean approximate μ and the sample standard deviation approximate σ. The distribution of X can then be approximated by $X \sim$ _____(_____,_____).
6. Use the distribution in part e to calculate the probability that the maximum capacity of sports stadiums is less than 67,000 spectators.

7. Determine the cumulative relative frequency that the maximum capacity of sports stadiums is less than 67,000 spectators. Hint: Order the data and count the sports stadiums that have a maximum capacity less than 67,000. Divide by the total number of sports stadiums in the sample.
8. Why aren't the answers to part f and part g exactly the same?

Solution

1. mean = 60,136; s = 10,468
2. Answers will vary.
3. Answers will vary.
4. Answers will vary.
5. $X \sim N(60136, 10468)$
6. 0.7440
7. The cumulative relative frequency is 43/60 = 0.717.
8. The answers for part f and part g are not the same, because the normal distribution is only an approximation.

An expert witness for a paternity lawsuit testifies that the length of a pregnancy is normally distributed with a mean of 280 days and a standard deviation of 13 days. An alleged father was out of the country from 240 to 306 days before the birth of the child, so the pregnancy would have been less than 240 days or more than 306 days long if he was the father. The birth was uncomplicated, and the child needed no medical intervention.

What is the probability that he was NOT the father? What is the probability that he could be the father? Calculate the z-scores first, and then use those to calculate the probability.

Solution

For x = 240, z =−3.0769

For x = 306, z = 2

P(240 < x < 306) = P(–3.0769 < z < 2) = normalcdf(–3.0769,2,0,1) = 0.9762.

According to the scenario given, this means that there is a 97.62% chance that he is not the father.

To answer the second part of the question, there is a 1 – 0.9762 = 0.0238 = 2.38% chance that he is the father.

A NUMMI assembly line, which has been operating since 1984, has built an average of 6,000 cars and trucks a week. Generally, 10% of the cars were defective coming off the assembly line. Suppose we draw a random sample of $n = 100$ cars. Let X represent the number of defective cars in the sample.

What can we say about X in regard to the 68-95-99.7 empirical rule (one standard deviation, two standard deviations and three standard deviations from the mean are being referred to)? Assume a normal distribution for the defective cars in the sample.

Solution

- $n = 100$; $p = 0.1$; $q = 0.9$
- $\mu = np = (100)(0.10) = 10$
- $\sigma = \sqrt{npq} = \sqrt{(100)(0.1)(0.9)} = 3$

1. $z = \pm 1$: $x_1 = \mu + z\sigma = 10 + 1(3) = 13$ and $x2 = \mu - z\sigma = 10 - 1(3) = 7$. 68% of the defective cars will fall between seven and 13.
2. $z = \pm 2$: $x_1 = \mu + z\sigma = 10 + 2(3) = 16$ and $x2 = \mu - z\sigma = 10 - 2(3) = 4$. 95 % of the defective cars will fall between four and 16
3. $z = \pm 3$: $x_1 = \mu + z\sigma = 10 + 3(3) = 19$ and $x2 = \mu - z\sigma = 10 - 3(3) = 1$. 99.7% of the defective cars will fall between one and 19.

We flip a coin 100 times ($n = 100$) and note that it only comes up heads 20% ($p = 0.20$) of the time. The mean and standard deviation for the number of times the coin lands on heads is $\mu = 20$ and $\sigma = 4$ (verify the mean and standard deviation). Solve the following:

1. There is about a 68% chance that the number of heads will be somewhere between ___ and ___.
2. There is about a ____chance that the number of heads will be somewhere between 12 and 28.
3. There is about a ____ chance that the number of heads will be somewhere between eight and 32.

Solution

1. There is about a 68% chance that the number of heads will be somewhere between 16 and 24. $z = \pm 1$: $x1 = \mu + z\sigma = 20 + 1(4) = 24$ and $x2 = \mu - z\sigma = 20 - 1(4) = 16$.
2. There is about a 95% chance that the number of heads will be somewhere between 12 and 28. For this problem: normalcdf(12,28,20,4) = 0.9545 = 95.45%
3. There is about a 99.73% chance that the number of heads will be somewhere between eight and 32. For this problem: normalcdf(8,32,20,4) = 0.9973 = 99.73%.

A $1 scratch-off lotto ticket will be a winner one out of five times. Out of a shipment of $n = 190$ lotto tickets, find the probability for the lotto tickets that there are

1. somewhere between 34 and 54 prizes.
2. somewhere between 54 and 64 prizes.
3. more than 64 prizes.

Solution

- $n = 190$; $p = \frac{1}{5}$ = 0.2; $q = 0.8$
- $\mu = np = (190)(0.2) = 38$
- $\sigma = \sqrt{npq} = \sqrt{(190)(0.2)(0.8)} = 5.5136$

1. For this problem: $P(34 < x < 54)$ = normalcdf(34,54,48,5.5136) = 0.7641
2. For this problem: $P(54 < x < 64)$ = normalcdf(54,64,48,5.5136) = 0.0018
3. For this problem: $P(x > 64)$ = normalcdf(64,10^{99},48,5.5136) = 0.0000012 (approximately 0)

Facebook provides a variety of statistics on its website that detail the growth and popularity of the site.

On average, 28 percent of 18- to 34-year-olds check their Facebook profiles before getting out of bed in the morning. Suppose this percentage follows a normal distribution with a standard deviation of 5 percent.

1. Find the probability that the percent of 18- to 34-year-olds who check Facebook before getting out of bed in the morning is at least 30.
2. Find the 95th percentile, and express it in a sentence.

Solution

X = the percent of 18- to 34-year-olds who check Facebook before getting out of bed in the morning.
X ~ N(28, 5)
P(x ≥ 30) = 0.3446; normalcdf(30,1EE99,28,5) = 0.3446

invNorm(0.95,0.28,0.05) = 0.3622.95% of the percentage of 18- to 34-year-olds who check Facebook before getting out of bed in the morning is at most 36.22%.
P(25 < x < 55) = normalcdf(25,55,28,5) = 0.7257(0.7257)(400) = 290.28

Section 6.3 Homework

Previously, De Anza statistics students estimated that the amount of change daytime statistics students carry is exponentially distributed with a mean of $ 0.88. Suppose that we randomly pick 25 daytime statistics students.

1. In words, X = _____
2. X ~ _____(_____,_____)
3. In words, $\overline{X}$ = _____
4. $\overline{X}$ ~ _____ (_____, _____)
5. Find the probability that an individual had between $ 0.80 and $ 1.00. Graph the situation, and shade in the area to be determined.
6. Find the probability that the average of the 25 students was between $ 0.80 and $ 1.00. Graph the situation, and shade in the area to be determined.
7. Explain why there is a difference in part e and part f.

Solution

1. X = amount of change students carry
2. X ~ $E(0.88, 0.88)$
3. $\overline{X}$ = average amount of change carried by a sample of 25 students.
4. $\overline{X}$ ~ $N(0.88, 0.176)$
5. 0.0819
6. 0.1882
7. The distributions are different. Part a is exponential and part b is normal.

Suppose that the distance of fly balls hit to the outfield (in baseball) is normally distributed with a mean of 250 feet and a standard deviation of 50 feet. We randomly sample 49 fly balls.

1. If $\overline{X}$ = average distance in feet for 49 fly balls, then $\overline{X}$ ~ _____(_____,_____)
2. What is the probability that the 49 balls traveled an average of less than 240 feet? Sketch the graph. Scale the horizontal axis for $\overline{X}$. Shade the region corresponding to the probability. Find the probability.
3. Find the 80th percentile of the distribution of the average of 49 fly balls.

Solution

N(250, 50/7)

0.0808

256.01 feet

According to the Internal Revenue Service, the average length of time for an individual to complete (keep records for, learn, prepare, copy, assemble, and send) IRS Form 1040 is 10.53 hours (without any attached schedules). The distribution is unknown. Let us assume that the standard deviation is two hours. Suppose we randomly sample 36 taxpayers.

1. In words, X = _____
2. In words, $\overline{X}$ = _____
3. $\overline{X}$ ~ _____(_____,_____)
4. Would you be surprised if the 36 taxpayers finished their Form 1040s in an average of more than 12 hours? Explain why or why not in complete sentences.
5. Would you be surprised if one taxpayer finished his or her Form 1040 in more than 12 hours? In a complete sentence, explain why.

Solution

1. length of time for an individual to complete IRS form 1040, in hours.
2. mean length of time for a sample of 36 taxpayers to complete IRS form 1040, in hours.
3. $N\left(10.53, \frac{1}{3}\right)$
4. Yes. I would be surprised, because the probability is almost 0.
5. No. I would not be totally surprised because the probability is 0.2312

Suppose that a category of world-class runners are known to run a marathon (26 miles) in an average of 145 minutes with a standard deviation of 14 minutes. Consider 49 of the races. Let $\overline{X}$ the average of the 49 races.

1. $\overline{X}$ ~ _____(_____,_____)
2. Find the probability that the runner will average between 142 and 146 minutes in these 49 marathons.
3. Find the 80th percentile for the average of these 49 marathons.
4. Find the median of the average running times.

Solution

N(145, 2)

0.6247

146.68

145 minutes

The length of songs in a collector's iTunes album collection is uniformly distributed from two

to 3.5 minutes. Suppose we randomly pick five albums from the collection. There are a total of 43 songs on the five albums.

1. In words, X = _____
2. X ~ _____
3. In words, $\overline{X}$ = _____
4. $\overline{X}$ ~ _____(_____,_____)
5. Find the first quartile for the average song length.
6. The IQR(interquartile range) for the average song length is from _____–_____.

Solution

1. the length of a song, in minutes, in the collection
2. $U(2, 3.5)$
3. the average length, in minutes, of the songs from a sample of five albums from the collection
4. $N(2.75, 0.0220)$
5. 2.74 minutes
6. 0.03 minutes

In 1940 the average size of a U.S. farm was 174 acres. Let's say that the standard deviation was 55 acres. Suppose we randomly survey 38 farmers from 1940.

1. In words, X = _____
2. In words, $\overline{X}$ = _____
3. $\overline{X}$ ~ _____(_____,_____)
4. The IQR for $\overline{X}$ is from _____ acres to _____ acres.

Solution

1. the size of a U.S. farm in 1940
2. the average size of a U.S. farm, in acres
3. N(174,8.9222)
4. 168.0, 180.0

Determine which of the following are true and which are false. Then, in complete sentences, justify your answers.

1. When the sample size is large, the mean of $\overline{X}$ is approximately equal to the mean of X.
2. When the sample size is large, $\overline{X}$ is approximately normally distributed.
3. When the sample size is large, the standard deviation of $\overline{X}$ is approximately the same as the standard deviation of X.

Solution

1. True. The mean of a sampling distribution of the means is approximately the mean of the data distribution.
2. True. According to the Central Limit Theorem, the larger the sample, the closer the sampling distribution of the means becomes normal.
3. The standard deviation of the sampling distribution of the means will decrease making it approximately the same as the standard deviation of X as the sample size increases.

The percentage of fat calories that a person in America consumes each day is normally distributed with a mean of about 36 and a standard deviation of about ten. Suppose that 16 individuals are randomly chosen. Let $\overline{X}$ = average percent of fat calories.

1. $\overline{X}$ ~ _____(_____, _____)
2. For the group of 16, find the probability that the average percent of fat calories consumed is more than five. Graph the situation and shade in the area to be determined.
3. Find the first quartile for the average percent of fat calories.

Solution

1. N(36, 2.5)
2. 1
3. 34.31

The distribution of income in some Third World countries is considered wedge shaped (many very poor people, very few middle income people, and even fewer wealthy people). Suppose we pick a country with a wedge shaped distribution. Let the average salary be $ 2,000 per year with a standard deviation of $ 8,000. We randomly survey 1,000 residents of that country.

1. In words, X = _____
2. In words, $\overline{X}$ = _____
3. $\overline{X}$ ~ _____(_____,_____)
4. How is it possible for the standard deviation to be greater than the average?
5. Why is it more likely that the average of the 1,000 residents will be from $ 2,000 to $ 2,100 than from $ 2,100 to $ 2,200?

Solution

1. X = the yearly income of someone in a third world country
2. the average salary from samples of 1,000 residents of a third world country

3. $\overline{X} \sim N\left(2000, \frac{8000}{\sqrt{1000}}\right)$
4. Very wide differences in data values can have averages smaller than standard deviations.
5. The distribution of the sample mean will have higher probabilities closer to the population mean.
   $P(2000 < \overline{X} < 2100) = 0.1537 P(2100 < \overline{X} < 2200) = 0.1317$

Which of the following is NOT TRUE about the distribution for averages?

1. The mean, median, and mode are equal.
2. The area under the curve is one.
3. The curve never touches the *x*-axis.
4. The curve is skewed to the right.

Solution

The curve is skewed to the right.

The cost of unleaded gasoline in the Bay Area once followed an unknown distribution with a mean of $4.59 and a standard deviation of $0.10. Sixteen gas stations from the Bay Area are randomly chosen. We are interested in the average cost of gasoline for the 16 gas stations.

The distribution to use for the average cost of gasoline for the 16 gas stations is:

1. $\overline{X} \sim N(4.59, 0.10)$

2. $\overline{X} \sim N\left(4.59, \frac{0.10}{\sqrt{16}}\right)$
3. $\overline{X} \sim N\left(4.59, \frac{16}{0.10}\right)$
4. $\overline{X} \sim N$ Formula does not parse

Solution

$\overline{X} \sim N\left(4.59, \frac{0.10}{\sqrt{16}}\right)$

Section 6.4 Homework

Which of the following is NOT TRUE about the theoretical distribution of sums?

1. The mean, median and mode are equal.
2. The area under the curve is one.
3. The curve never touches the *x*-axis.
4. The curve is skewed to the right.

Solution

The curve is skewed to the right.

Suppose that the duration of a particular type of criminal trial is known to have a mean of 21 days and a standard deviation of seven days. We randomly sample nine trials.

1. In words, ΣX = _____
2. ΣX ~ _____(_____,_____)
3. Find the probability that the total length of the nine trials is at least 225 days.
4. Ninety percent of the total of nine of these types of trials will last at least how long?

Solution

1. the total length of time for nine criminal trials
2. $N(189, 21)$
3. 0.0432
4. 162.09; ninety percent of the total nine trials of this type will last 162 days or more.

Suppose that the weight of open boxes of cereal in a home with children is uniformly distributed from two to six pounds with a mean of four pounds and standard deviation of 1.1547. We randomly survey 64 homes with children.

1. In words, X = _____
2. The distribution is _____.
3. In words, ΣX = _____
4. ΣX ~ _____(_____,_____)
5. Find the probability that the total weight of open boxes is less than 250 pounds.
6. Find the 35^{th} percentile for the total weight of open boxes of cereal.

Solution

1. X = weight of open cereal boxes
2. The distribution is uniform.

3. ΣX = the sum of the weights of 64 randomly chosen cereal boxes.
4. N(256, 9.24)
5. 0.2581
6. 252.44 pounds

Salaries for teachers in a particular elementary school district are normally distributed with a mean of $ 44,000 and a standard deviation of $ 6,500. We randomly survey ten teachers from that district.

1. In words, X = _____
2. X ~ _____(_____,_____)
3. In words, ΣX = _____
4. ΣX ~ _____(_____,_____)
5. Find the probability that the teachers earn a total of over $ 400,000.
6. Find the 90th percentile for an individual teacher's salary.
7. Find the 90th percentile for the sum of ten teachers' salary.
8. If we surveyed 70 teachers instead of ten, graphically, how would that change the distribution in part d?
9. If each of the 70 teachers received a $ 3,000 raise, graphically, how would that change the distribution in part b?

Solution

1. X = the salary of one elementary school teacher in the district
2. $X \sim N(44{,}000, 6{,}500)$
3. ΣX ~ sum of the salaries of ten elementary school teachers in the sample
4. $\Sigma X \sim N(44000, 20554.80)$
5. 0.9742
6. $ 52,330.09
7. 466,342.04

8. Sampling 70 teachers instead of ten would cause the distribution to be more spread out. It would be a more symmetrical normal curve.
9. If every teacher received a $3,000 raise, the distribution of X would shift to the right by $3,000. In other words, it would have a mean of $47,000.

Section 6.5 Homework

The attention span of a two-year-old is exponentially distributed with a mean of about eight minutes. Suppose we randomly survey 60 two-year-olds.

1. In words, X = _____
2. X ~ _____(_____,_____)
3. In words, $\overline{X}$ = _____
4. $\overline{X}$ ~ _____(_____,_____)
5. Before doing any calculations, which do you think will be higher? Explain why.
    1. The probability that an individual attention span is less than ten minutes.
    2. The probability that the average attention span for the 60 children is less than ten minutes?
6. Calculate the probabilities in part e.
7. Explain why the distribution for $\overline{X}$ is not exponential.

Solution

1. X = the attention span of a two-year-old
2. **X** ~ Exp(18)
3. $\bar{X}$ = the mean (average) attention span of a two-year-old
4. $\bar{X}$ ~ N(8,860)
5. The standard deviation is smaller, so there is more area under the normal curve.
6. Exponential: 0.7135
7. Normal: 0.7579
8. By the central limit theorem, as n gets larger, the means tend to follow a normal distribution.

The closing stock prices of 35 U.S. semiconductor manufacturers are given as follows.

8.62530.2527.62546.7532.87518.2550.1252.93756.87528.2524.25211.530.257143.549.252.56253116.59.518.518910.516.6251.251812.87712.8752.87560.2529.25

1. In words, ***X*** = _____
    1. $\overline{x}$ = _____
    2. s_X = _____
    3. n = _____
2. Construct a histogram of the distribution of the averages. Start at $x = -0.0005$. Use bar widths of ten.
3. In words, describe the distribution of stock prices.
4. Randomly average five stock prices together. (Use a random number generator.) Continue averaging five pieces together until you have ten averages. List those ten averages.
5. Use the ten averages from part e to calculate the following.
    1. $\overline{x}$ = _____
    2. s_X = _____

6. Construct a histogram of the distribution of the averages. Start at $x = -0.0005$. Use bar widths of ten.
7. Does this histogram look like the graph in part c?
8. In one or two complete sentences, explain why the graphs either look the same or look different?
9. Based upon the theory of the **central limit theorem**, $\overline{X}$ ~ _____(_____,_____)

Solution

1. X = the closing stock prices for U.S. semiconductor manufacturers
2. i. $ 20.71; ii. $ 17.31; iii. 35
3. Exponential distribution, $X \sim Exp\left(\frac{1}{20.71}\right)$
4. Answers will vary.
5. i. $ 20.71; ii. $ 11.14
6. Answers will vary.
7. Answers will vary.
8. Answers will vary.
9. $N\left(20.71, \frac{17.31}{\sqrt{5}}\right)$

Use the following information to answer the next three exercises:

Richard's Furniture Company delivers furniture from 10 A.M. to 2 P.M. continuously and uniformly. We are interested in how long (in hours) past the 10 A.M. start time that individuals wait for their delivery.

X ~ _____(_____,_____)

1. $U(0,4)$

2. $U(10,2)$
3. $Exp(2)$
4. $N(2,1)$

Solution

$U(0,4)$

The average wait time is:

1. one hour.
2. two hours.
3. two and a half hours.
4. four hours.

Solution

two hours

Suppose that it is now past noon on a delivery day. The probability that a person must wait at least one and a half **more** hours is:

1. $\frac{1}{4}$
2. $\frac{1}{2}$
3. $\frac{3}{4}$
4. $\frac{3}{8}$

Solution

$\frac{1}{4}$

Use the following information to answer the next two exercises:

The time to wait for a particular rural bus is distributed uniformly from zero to 75 minutes. One hundred riders are randomly sampled to learn how long they waited.

The 90^{th} percentile sample average wait time (in minutes) for a sample of 100 riders is:

1. 315.0
2. 40.3
3. 38.5
4. 65.2

Solution

40.3

Would you be surprised, based upon numerical calculations, if the sample average wait time (in minutes) for 100 riders was less than 30 minutes?

1. yes
2. no
3. There is not enough information.

Solution

yes

Use the following to answer the next two exercises:

The cost of unleaded gasoline in the Bay Area once followed an unknown distribution with a mean of $ 4.59 and a standard deviation of $ 0.10. Sixteen gas stations from the Bay Area are randomly chosen. We are interested in the average cost of gasoline for the 16 gas stations.

What's the approximate probability that the average price for 16 gas stations is over $ 4.69?

1. almost zero
2. 0.1587
3. 0.0943
4. unknown

Solution
almost zero

Find the probability that the average price for 30 gas stations is less than $ 4.55.

1. 0.6554
2. 0.3446
3. 0.0142
4. 0.9858
5. 0

Solution
0.0142

Suppose in a local Kindergarten through 12^{th} grade (K – 12) school district, 53 percent of the

population favor a charter school for grades K through five. A simple random sample of 300 is surveyed. Calculate following using the normal approximation to the binomial distribution.

1. Find the probability that less than 100 favor a charter school for grades K through 5.
2. Find the probability that 170 or more favor a charter school for grades K through 5.
3. Find the probability that no more than 140 favor a charter school for grades K through 5.
4. Find the probability that there are fewer than 130 that favor a charter school for grades K through 5.
5. Find the probability that exactly 150 favor a charter school for grades K through 5.

If you have access to an appropriate calculator or computer software, try calculating these probabilities using the technology.

Solution

1. 0
2. 0.1123
3. 0.0162
4. 0.0003
5. 0.0268

Four friends, Janice, Barbara, Kathy and Roberta, decided to carpool together to get to school. Each day the driver would be chosen by randomly selecting one of the four names. They carpool to school for 96 days. Use the normal approximation to the binomial to calculate the following probabilities. Round the standard deviation to four decimal places.

1. Find the probability that Janice is the driver for at most 20 days.
2. Find the probability that Roberta is the driver for more than 16 days.
3. Find the probability that Barbara drives exactly 24 of those 96 days.

Solution

1. 0.2047
2. 0.9615
3. 0.0938

$X \sim N(60, 9)$. Suppose that you form random samples of 25 from this distribution. Let $\overline{X}$ be the random variable of averages. Let ΣX be the random variable of sums. For parts c through f, sketch the graph, shade the region, label and scale the horizontal axis for $\overline{X}$, and find the probability.

1. Sketch the distributions of X and $\overline{X}$ on the same graph.
2. $\overline{X} \sim$ _____(_____,_____)
3. $P(\overline{x} < 60) =$ _____
4. Find the 30th percentile for the mean.
5. $P(56 < \overline{x} < 62) =$ _____
6. $P(18 < \overline{x} < 58) =$ _____
7. $\Sigma x \sim$ _____(_____,_____)
8. Find the minimum value for the upper quartile for the sum.
9. $P(1{,}400 < \Sigma x < 1{,}550) =$ _____

Solution

1. Check student's solution.
2. $\overline{X} \sim N\left(60, \frac{9}{\sqrt{25}}\right)$
3. 0.5000
4. 59.06
5. 0.8536
6. 0.1333

7. $N(1500, 45)$
8. 1530.35
9. 0.6877

Suppose that the length of research papers is uniformly distributed from ten to 25 pages. We survey a class in which 55 research papers were turned in to a professor. The 55 research papers are considered a random collection of all papers. We are interested in the average length of the research papers.

1. In words, X = _____
2. X ~ _____(_____,_____)
3. μ_X = _____
4. σ_X = _____
5. In words, $\overline{X}$ = _____
6. $\overline{X}$ ~ _____(_____,_____)
7. In words, ΣX = _____
8. ΣX ~ _____(_____,_____)
9. Without doing any calculations, do you think that it's likely that the professor will need to read a total of more than 1,050 pages? Why?
10. Calculate the probability that the professor will need to read a total of more than 1,050 pages.
11. Why is it so unlikely that the average length of the papers will be less than 12 pages?

Solution

U(10, 25)

17.5

225

12

= 4.3301

N(17.5, 0.5839)
N(962.5, 32.11)
0.0032

Salaries for teachers in a particular elementary school district are normally distributed with a mean of $ 44,000 and a standard deviation of $ 6,500. We randomly survey ten teachers from that district.

1. Find the 90th percentile for an individual teacher's salary.
2. Find the 90th percentile for the average teacher's salary.

Solution

1. $ 52,330
2. $ 46,634

The average length of a maternity stay in a U.S. hospital is said to be 2.4 days with a standard deviation of 0.9 days. We randomly survey 80 women who recently bore children in a U.S. hospital.

1. In words, X = _____
2. In words, $\overline{X}$ = _____
3. $\overline{X}$ ~ _____(_____,_____)
4. In words, ΣX = _____
5. ΣX ~ _____(_____,_____)
6. Is it likely that an individual stayed more than five days in the hospital? Why or why not?

7. Is it likely that the average stay for the 80 women was more than five days? Why or why not?
8. Which is more likely:
    1. An individual stayed more than five days.
    2. The average stay of 80 women was more than five days.
9. If we were to sum up the women's stays, is it likely that, collectively, they spent more than a year in the hospital? Why or why not?

Solution

1. the length of a maternity stay in a U.S. hospital, in days
2. the average length of a maternity stay in a U.S. hospital, in days
3. N(2.4, 0.980)
4. N(192, 8.05)
5. Not likely, but possible. The mean stay is 2.4 days.
6. No, the probability is 0.
7. Individual
8. No, the probability is 0.

For each problem, wherever possible, provide graphs and use the calculator.

NeverReady batteries has engineered a newer, longer lasting AAA battery. The company claims this battery has an average life span of 17 hours with a standard deviation of 0.8 hours. Your statistics class questions this claim. As a class, you randomly select 30 batteries and find that the sample mean life span is 16.7 hours. If the process is working properly, what is the probability of getting a random sample of 30 batteries in which the sample mean lifetime is 16.7 hours or less? Is the company's claim reasonable?

Solution

- We have $\mu = 17$, $\sigma = 0.8$, $\overline{x} = 16.7$, and $n = 30$. To calculate the probability, we use normalcdf(lower, upper, μ, $\frac{\sigma}{\sqrt{n}}$) = normalcdf $\left(E-99,16.7,17,\frac{0.8}{\sqrt{30}}\right)$ = 0.0200.
- If the process is working properly, then the probability that a sample of 30 batteries would have at most 16.7 lifetime hours is only 2%. Therefore, the class was justified to question the claim.

Men have an average weight of 172 pounds with a standard deviation of 29 pounds.

1. Find the probability that 20 randomly selected men will have a sum weight greater than 3,600 lbs.
2. If 20 men have a sum weight greater than 3,500 lbs., then their total weight exceeds the safety limits for water taxis. Based on (a), is this a safety concern? Explain.

Solution

To calculate the probability, we use normalcdf(3600, E99, 3440, 129.69) = 0.1087

While the probability is not exceptionally large, $P(\bar{X} > 3600) = 0.1087$, it could be wise to be concerned. After all, we do have a 1 in 10 chance that a random sample of men could have an average sum weight that exceeds the safety limit.

M&M candies large candy bags have a claimed net weight of 396.9 g. The standard deviation for the weight of the individual candies is 0.017 g. The following table is from a stats experiment conducted by a statistics class.

Red	Orange	Yellow	Brown	Blue	Green
0.751	0.735	0.883	0.696	0.881	0.925
0.841	0.895	0.769	0.876	0.863	0.914
0.856	0.865	0.859	0.855	0.775	0.881
0.799	0.864	0.784	0.806	0.854	0.865
0.966	0.852	0.824	0.840	0.810	0.865
0.859	0.866	0.858	0.868	0.858	1.015
0.857	0.859	0.848	0.859	0.818	0.876
0.942	0.838	0.851	0.982	0.868	0.809
0.873	0.863			0.803	0.865
0.809	0.888			0.932	0.848
0.890	0.925			0.842	0.940
0.878	0.793			0.832	0.833
0.905	0.977			0.807	0.845
	0.850			0.841	0.852
	0.830			0.932	0.778
	0.856			0.833	0.814
	0.842			0.881	0.791
	0.778			0.818	0.810
	0.786			0.864	0.881
	0.853			0.825	
	0.864			0.855	
	0.873			0.942	
	0.880			0.825	
	0.882			0.869	
	0.931			0.912	
				0.887	

The bag contained 465 candies, and the listed weights in the table came from randomly selected candies. Count the weights.

1. Find the mean sample weight and the standard deviation of the sample weights of candies in the table.
2. Find the sum of the sample weights in the table and the standard deviation of the sum of the weights.
3. If 465 M&Ms are randomly selected, find the probability that their weights sum to at least 396.9.
4. Is the Mars Company's M&M labeling accurate?

Solution

1. For the sample, we have $n = 100$, $\overline{x} = 0.862$, $s = 0.05$
2. $\sum\overline{x} = 85.65$, $\Sigma s = 5.18$
3. `normalcdf`(396.9,*E*99,(465)(0.8565),(0.05)($\sqrt{465}$)) ≈ 1
4. Since the probability of a sample of size 465 having at least a mean sum of 396.9 is approximately 1, we can conclude that Mars is correctly labeling their M&M packages.

The Screw Right Company claims their $\frac{3}{4}$-inch screws are within ±0.23 of the claimed mean diameter of 0.750 inches with a standard deviation of 0.115 inches. The following data were recorded.

0.757	0.723	0.754	0.737	0.757	0.741	0.722	0.741	0.743	0.742
0.740	0.758	0.724	0.739	0.736	0.735	0.760	0.750	0.759	0.754
0.744	0.758	0.765	0.756	0.738	0.742	0.758	0.757	0.724	0.757
0.744	0.738	0.763	0.756	0.760	0.768	0.761	0.742	0.734	0.754
0.758	0.735	0.740	0.743	0.737	0.737	0.725	0.761	0.758	0.756

The screws were randomly selected from the local home repair store.

1. Find the mean diameter and standard deviation for the sample
2. Find the probability that 50 randomly selected screws will be within the stated tolerance levels. Is the company's diameter claim plausible?

Solution

$\bar{x}$ = 0.75 and s 0.01

We have `normalcdf`(0.52,0.98,0.75,0.11550) ≈ 1 and we can conclude that the company's diameter claim is justified.

Your company has a contract to perform preventive maintenance on thousands of air-conditioners in a large city. Based on service records from previous years, the time that a technician spends servicing a unit averages one hour with a standard deviation of one hour. In the coming week, your company will service a simple random sample of 70 units in the city. You plan to budget an average of 1.1 hours per technician to complete the work. Will this be enough time?

Solution

Use `normalcdf` $\left(E-99,1.1,1,\frac{1}{\sqrt{70}}\right)$ = 0.7986. This means that there is an 80% chance that the service time will be less than 1.1 hours. It could be wise to schedule more time since there is an associated 20% chance that the maintenance time will be greater than 1.1 hours.

A typical adult has an average IQ score of 105 with a standard deviation of 20. If 20 randomly selected adults are given an IQ test, what is the probability that the sample mean scores will be between 85 and 125 points?

Solution

The probability that the sample score is between 85 and 125 points is given by `normalcdf`(85,125,105,2020)= 0.9999. Therefore, it is almost a guarantee that a well selected sample of 20 adults will have an average score between 85 and 125.

Certain coins have an average weight of 5.201 grams with a standard deviation of 0.065 g. If a

vending machine is designed to accept coins whose weights range from 5.111 g to 5.291 g, what is the expected number of rejected coins when 280 randomly selected coins are inserted into the machine?

Solution

Since we have `normalcdf` $\left(5.111,5.291,5.201,\frac{0.065}{\sqrt{280}}\right)$ ≈ 1, we can conclude that practically all the coins are within the limits, therefore, there should be no rejected coins out of a well selected sample of size 280.

CHAPTER 7: CONFIDENCE INTERVALS

INTRODUCTION TO CHAPTER 7: CONFIDENCE INTERVALS



Figure 8.1. Have you ever wondered what the average number of M&Ms in a bag at the grocery store is? You can use confidence intervals to answer this question. (credit: comedy_nose/flickr)

Suppose you were trying to determine the mean rent of a two-bedroom apartment in your town. You might look in the classified section of the newspaper, write down several rents listed, and average them together. You would have obtained a point estimate of the true mean. If you are trying to determine the percentage of times you make a basket when shooting a basketball, you might count the number of shots you make and divide that by the number of shots you attempted. In this case, you would have obtained a point estimate for the true proportion.

We use sample data to make generalizations about an unknown population. This part of statistics is called **inferential statistics**. **The sample data help us to make an estimate of a population parameter.** We

realize that the point estimate is most likely not the exact value of the population parameter, but close to it. After calculating **point estimates**, we construct interval estimates, called confidence intervals.

In this chapter, you will learn to construct and interpret confidence intervals. You will also learn a new distribution, the Student's-t, and how it is used with these intervals. Throughout the chapter, it is important to keep in mind that the confidence interval is a random variable. It is the population parameter that is fixed.

If you worked in the marketing department of an entertainment company, you might be interested in the mean number of songs a consumer downloads a month from iTunes. If so, you could conduct a survey and calculate the sample mean, $\overline{x}$, and the sample standard deviation, s. You would use $\overline{x}$ to estimate the population mean and s to estimate the population standard deviation. The sample mean, $\overline{x}$, is the point estimate for the population mean, μ. The sample standard deviation, s, is the point estimate for the population standard deviation, σ.

Each of $\overline{x}$ and s is called a statistic.

A confidence interval is another type of estimate, but instead of being just one number, it is an interval of numbers. The interval of numbers is a range of values calculated from a given set of sample data. The confidence interval is likely to include an unknown population parameter.

Suppose for the iTunes example, we do not know the population mean μ, but we do know that the population standard deviation is $\sigma = 1$ and our sample size is 100. Then, by the central limit theorem, the standard deviation for the sample mean is $\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.1$.

The empirical rule, which applies to bell-shaped distributions, says that in approximately 95% of the samples, the sample mean, $\overline{x}$, will be within two standard deviations of the population mean μ. For our iTunes example, two standard deviations are (2)(0.1) = 0.2. The sample mean $\overline{x}$ is likely to be within 0.2 units of μ.

Because $\overline{x}$ is within 0.2 units of μ, which is unknown, then μ is likely to be within 0.2 units of $\overline{x}$ in 95% of the samples. The population mean μ is contained in an interval whose lower number is calculated by taking the sample mean and subtracting two standard deviations (2)(0.1) and whose upper number is calculated by taking the sample mean and adding two standard deviations. In other words, μ is between $\overline{x} - 0.2$ and $\overline{x} + 0.2$ in 95% of all the samples.

For the iTunes example, suppose that a sample produced a sample mean $\overline{x} = 2$. Then the unknown population mean μ is between $\overline{x} - 0.2 = 2 - 0.2 = 1.8$ and $\overline{x} + 0.2 = 2 + 0.2 = 2.2$

We say that we are **95% confident** that the unknown population mean number of songs downloaded from iTunes per month is between 1.8 and 2.2. **The 95% confidence interval is (1.8, 2.2).**

The 95% confidence interval implies two possibilities. Either the interval (1.8, 2.2) contains the true mean μ or our sample produced an $\overline{x}$ that is not within 0.2 units of the true mean μ. The second possibility happens for only 5% of all the samples (95–100%).

Remember that a **confidence interval** is created for an unknown population **parameter** like the population mean, μ. Confidence intervals for some parameters have the form:

(point estimate – margin of error, point estimate + margin of error)

The margin of error depends on the confidence level or percentage of confidence and the standard error of the mean.

When you read newspapers and journals, some reports will use the phrase "margin of error." Other reports will not use that phrase but include a confidence interval as the point estimate plus or minus the margin of error. These are two ways of expressing the same concept.

Note

Although the text only covers symmetrical confidence intervals, there are non-symmetrical confidence intervals (for example, a confidence interval for the standard deviation).

Collaborative Exercises

Section 7.1

Collaborative Exercise

Have your instructor record the number of meals each student in your class eats out in a week. Assume that the standard deviation is known to be three meals. Construct an approximate 95% confidence interval for the true mean number of meals students eat out each week.

1. Calculate the sample mean.

2. Let σ.= 3 and n = the number of students surveyed.
3. Construct the interval $\left(\overline{x} - 2 \cdot \frac{\sigma}{\sqrt{\mathrm{n}}}, \overline{x} + 2 \cdot \frac{\sigma}{\sqrt{\mathrm{n}}}\right)$.

We say we are approximately 95% confident that the true mean number of meals that students eat out in a week is between _____ and _____.

Media Attributions

- Chapter 8 Introduction Figure © OpenStax Introductory Statistics is licensed under a CC BY (Attribution) license

7.1 A SINGLE POPULATION MEAN USING THE NORMAL DISTRIBUTION

Learning Objectives

By the end of this chapter, the student should be able to:

- Calculate the confidence interval
- Calculate the error bound (EBM)
- Determine the **confidence level** and sample size

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of $\overline{x} = 10$ and we have constructed the 90% confidence interval (5, 15) where $EBM = 5$.

The Confidence Interval (Known Standard Deviation)

To construct a confidence interval for a single unknown population, mean μ, **where the population standard deviation is known**, we need $\overline{x}$ as an estimate for μ and we need the margin of error. Here, the margin of error (EBM) is called the error bound for a population mean (abbreviated ***EBM***). The sample mean $\overline{x}$ is the **point estimate** of the unknown population mean μ.

The confidence interval estimate will have the form:
(point estimate – error bound, point estimate + error bound) or, in symbols,
$(\overline{x} - EBM, \overline{x} + EBM)$
The margin of error (EBM) depends on the confidence level (abbreviated ***CL***). The confidence level is often

considered the probability that the calculated confidence interval estimate will contain the true population parameter. However, it is more accurate to state that the confidence level is the percent of confidence intervals that contain the true population parameter when repeated samples are taken. Most often, it is the choice of the person constructing the **confidence interval** to choose a confidence level of 90% or higher because that person wants to be reasonably certain of his or her conclusions.

There is another probability called alpha (α). α is related to the **confidence level**, CL. α is the probability that the interval does not contain the unknown population parameter.

Mathematically, $\alpha + CL = 1$.

Example

- Suppose we have collected data from a sample. We know the sample mean but we do not know the mean for the entire population.
- The sample mean is seven, and the error bound for the mean is 2.5.

$\overline{x} = 7$ and $EBM = 2.5$

The confidence interval is (7 – 2.5, 7 + 2.5), and calculating the values gives (4.5, 9.5).

If the confidence level (CL) is 95%, then we say, "We estimate with 95% confidence that the true value of the population mean is between 4.5 and 9.5."

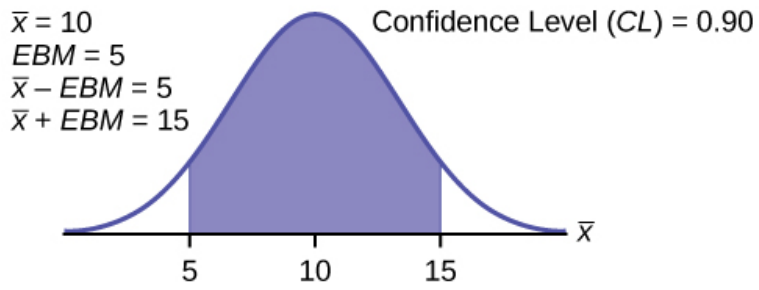
Your Turn!

An interactive H5P element has been excluded from this version of the text. You can view it online

here:
https://louis.pressbooks.pub/introductorystatistics/?p=845#h5p-10

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of $\overline{x}$ = 10, and we have constructed the 90% confidence interval (5, 15) where *EBM* = 5.

To get a 90% confidence interval, we must include the central 90% of the probability of the normal distribution. If we include the central 90%, we leave out a total of α = 10% in both tails, or 5% in each tail, of the normal distribution.



To capture the central 90%, we must go out 1.645 "standard deviations" on either side of the calculated sample mean. The value 1.645 is the *z*-score from a standard normal probability distribution that puts an area of 0.90 in the center, an area of 0.05 in the far-left tail, and an area of 0.05 in the far-right tail.

It is important that the "standard deviation" used must be appropriate for the parameter we are estimating, so in this section we need to use the standard deviation that applies to sample means, which is $\frac{\sigma}{\sqrt{n}}$. The fraction $\frac{\sigma}{\sqrt{n}}$ is commonly called the "standard error of the mean" in order to distinguish clearly the standard deviation for a mean from the population standard deviation σ.

In summary, as a result of the central limit theorem:

- $\overline{X}$ is normally distributed, that is, $\overline{X} \sim N\left(\mu_X, \frac{\sigma}{\sqrt{n}}\right)$.

- When the population standard deviation σ is known, we use a normal distribution to calculate the error bound.

Calculating the Confidence Interval

To construct a confidence interval, estimate for an unknown population mean, we need data from a random sample. The steps to construct and interpret the confidence interval are:

- Calculate the sample mean $\overline{x}$ from the sample data. Remember, in this section we already know the population standard deviation σ.
- Find the z-score that corresponds to the confidence level.
- Calculate the error bound EBM.
- Construct the confidence interval.
- Write a sentence that interprets the estimate in the context of the situation in the problem. (Explain what the confidence interval means, in the words of the problem.)

We will first examine each step in more detail, and then illustrate the process with some examples.



One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=845#oembed-1

Finding the z-score for the Stated Confidence Level

When we know the population standard deviation σ, we use a standard normal distribution to calculate the error bound EBM and construct the confidence interval. We need to find the value of z that puts an area equal to the confidence level (in decimal form) in the middle of the standard normal distribution $Z \sim N(0, 1)$.

The confidence level, CL, is the area in the middle of the standard normal distribution. $CL = 1 - \alpha$, so α is the area that is split equally between the two tails. Each of the tails contains an area equal to $\frac{\alpha}{2}$.

The z-score that has an area to the right of $\frac{\alpha}{2}$ is denoted by $z_{\frac{\alpha}{2}}$.

For example, when $CL = 0.95$, $\alpha = 0.05$ and $\frac{\alpha}{2} = 0.025$; we write $z_{\frac{\alpha}{2}} = z_{0.025}$.

The area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is $1 - 0.025 = 0.975$.

$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$, using a calculator, computer or a standard normal probability table.

`invNorm`(0.975, 0, 1) = 1.96

Note

Remember to use the area to the LEFT of $z_{\frac{\alpha}{2}}$; in this chapter the last two inputs in the invNorm command are 0, 1, because you are using a standard normal distribution $Z \sim N(0, 1)$.

Calculating the **Error Bound (*EBM*)**

The error bound formula for an unknown population mean μ when the population standard deviation σ is known is

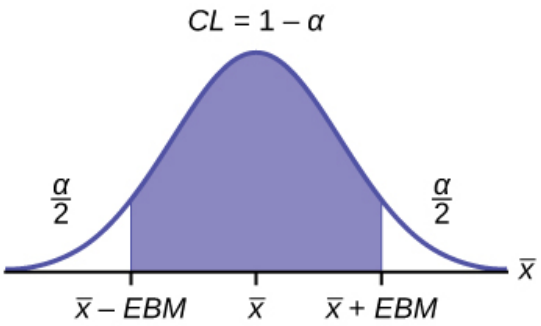
- $EBM = z_{\frac{\alpha}{2}} \left(\frac{\sigma}{\sqrt{n}} \right)$

Constructing the Confidence Interval

- The confidence interval estimate has the format $(\overline{x} - EBM, \overline{x} + EBM)$.

The graph gives a picture of the entire situation.

$CL + \frac{\alpha}{2} + \frac{\alpha}{2} = CL + \alpha = 1.$



Writing the Interpretation

The interpretation should clearly state the confidence level (*CL*), explain what population parameter is being estimated (here, a **population mean**), and state the confidence interval (both endpoints). "We estimate with ___% confidence that the true population mean (include the context of the problem) is between ___ and ___ (include appropriate units)."

Example

Suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of three points. A random sample of 36 scores is taken and gives a sample mean (sample mean score) of 68. Find a confidence interval estimate for the population mean exam score (the mean score on all exams).

Find a 90% confidence interval for the true (population) mean of statistics exam scores.

Solution

- You can use technology to calculate the confidence interval directly.
- The first solution is shown step-by-step (Solution A).
- The second solution uses the TI-83, 83+, and 84+ calculators (Solution B).

Solution

To find the confidence interval, you need the sample mean, $\overline{x}$, and the *EBM*.

$\overline{x}$ = 68

$EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$

σ = 3; n = 36; The confidence level is 90% (CL = 0.90)

CL = 0.90 so $\alpha = 1 - CL = 1 - 0.90 = 0.10$

$\frac{\alpha}{2} = 0.05$ $z_{\frac{\alpha}{2}} = z_{0.05}$

The area to the right of $z_{0.05}$ is 0.05 and the area to the left of $z_{0.05}$ is 1 – 0.05 = 0.95.

$z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$

using invNorm(0.95, 0, 1) on the TI-83,83+, and 84+ calculators. This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the standard normal distribution.

$EBM = (1.645)\left(\frac{3}{\sqrt{36}}\right) = 0.8225$

$\overline{x} - EBM = 68 - 0.8225 = 67.1775$

$\overline{x} + EBM = 68 + 0.8225 = 68.8225$

The 90% confidence interval is **(67.1775, 68.8225).**

Solution

Solution B

Press `STAT` and arrow over to `TESTS`.

Arrow down to `7: Z Interval`.

Press `ENTER`.

Arrow to `Stats` and press `ENTER`.

Arrow down and enter three for σ, 68 for $\overline{x}$, 36 for n, and .90 for `C-level`.

Arrow down to `Calculate` and press `ENTER`.

The confidence interval is (to three decimal places) (67.178, 68.822).

Interpretation We estimate with 90% confidence that the true population mean exam score for all statistics students is between 67.18 and 68.82.

Explanation of 90% Confidence Level

Ninety percent of all confidence intervals constructed in this way contain the true mean statistics exam score. For example, if we constructed 100 of these confidence intervals, we would expect 90 of them to contain the true population mean exam score.

Your Turn!

Suppose average pizza delivery times are normally distributed with an unknown population mean and a population standard deviation of six minutes. A random sample of 28 pizza delivery restaurants is taken and has a sample mean delivery time of 36 minutes.

Find a 90% confidence interval estimate for the population mean delivery time.

Solution
(34.1347, 37.8653)

Example

The Specific Absorption Rate (SAR) for a cell phone measures the amount of radio frequency (RF) energy absorbed by the user's body when using the handset. Every cell phone emits RF energy. Different phone models have different SAR measures. To receive certification from the Federal Communications Commission (FCC) for sale in the United States, the SAR level for a cell phone must

be no more than 1.6 watts per kilogram. [link] shows the highest SAR level for a random selection of cell phone models as measured by the FCC.

Phone Model	SAR	Phone Model	SAR	Phone Model	SAR
Apple iPhone 4S	1.11	LG Ally	1.36	Pantech Laser	0.74
BlackBerry Pearl 8120	1.48	LG AX275	1.34	Samsung Character	0.5
BlackBerry Tour 9630	1.43	LG Cosmos	1.18	Samsung Epic 4G Touch	0.4
Cricket TXTM8	1.3	LG CU515	1.3	Samsung M240	0.867
HP/Palm Centro	1.09	LG Trax CU575	1.26	Samsung Messager III SCH-R750	0.68
HTC One V	0.455	Motorola Q9h	1.29	Samsung Nexus S	0.51
HTC Touch Pro 2	1.41	Motorola Razr2 V8	0.36	Samsung SGH-A227	1.13
Huawei M835 Ideos	0.82	Motorola Razr2 V9	0.52	SGH-a107 GoPhone	0.3
Kyocera DuraPlus	0.78	Motorola V195s	1.6	Sony W350a	1.48
Kyocera K127 Marbl	1.25	Nokia 1680	1.39	T-Mobile Concord	1.38

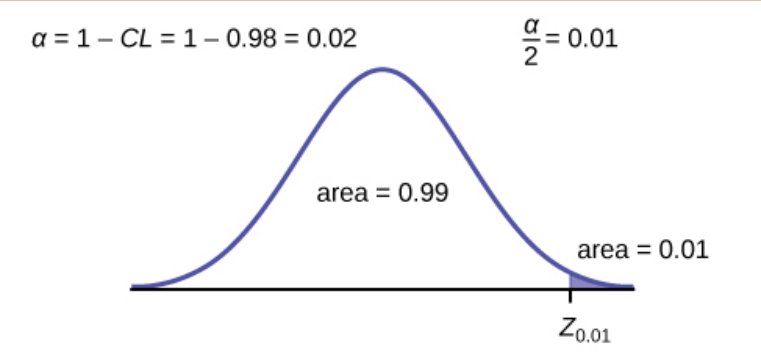
Find a 98% confidence interval for the true (population) mean of the Specific Absorption Rates (SARs) for cell phones. Assume that the population standard deviation is $\sigma = 0.337$.

Solution

To find the confidence interval, start by finding the point estimate: the sample mean.

$$\overline{x} = 1.024$$

Next, find the *EBM*. Because you are creating a 98% confidence interval, $CL = 0.98$.



You need to find $z_{0.01}$ having the property that the area under the normal density curve to the right of $z_{0.01}$ is 0.01 and the area to the left is 0.99. Use your calculator, a computer, or a probability table for the standard normal distribution to find $z_{0.01} = 2.326$.

$$EBM = (z_{0.01})\frac{\sigma}{\sqrt{n}} = (2.326)\frac{0.337}{\sqrt{30}} = 0.1431$$

To find the 98% confidence interval, find $\overline{x} \pm EBM$.

$\overline{x}$ – *EBM* = 1.024 – 0.1431 = 0.8809

$\overline{x}$ – *EBM* = 1.024 – 0.1431 = 1.1671

We estimate with 98% confidence that the true SAR mean for the population of cell phones in the United States is between 0.8809 and 1.1671 watts per kilogram.

Solution
Solution B

- Press STAT and arrow over to TESTS.
- Arrow down to 7:ZInterval.
- Press ENTER.
- Arrow to Stats and press ENTER.
- Arrow down and enter the following values:
    - σ: 0.337
    - $\overline{x}$: 1.024
    - *n*: 30
    - *C*-level: 0.98
- Arrow down to Calculate and press ENTER.
- The confidence interval is (to three decimal places) (0.881, 1.167).

Your Turn!

[link] shows a different random sampling of 20 cell phone models. Use this data to calculate a 93% confidence interval for the true mean SAR for cell phones certified for use in the United States. As previously, assume that the population standard deviation is σ = 0.337.

Phone Model	SAR	Phone Model	SAR
Blackberry Pearl 8120	1.48	Nokia E71x	1.53
HTC Evo Design 4G	0.8	Nokia N75	0.68
HTC Freestyle	1.15	Nokia N79	1.4
LG Ally	1.36	Sagem Puma	1.24
LG Fathom	0.77	Samsung Fascinate	0.57
LG Optimus Vu	0.462	Samsung Infuse 4G	0.2
Motorola Cliq XT	1.36	Samsung Nexus S	0.51
Motorola Droid Pro	1.39	Samsung Replenish	0.3
Motorola Droid Razr M	1.3	Sony W518a Walkman	0.73
Nokia 7705 Twist	0.7	ZTE C79	0.869

Solution

$\overline{x} = 0.940$

$$\frac{\alpha}{2} = \frac{1 - CL}{2} = \frac{1 - 0.93}{2} = 0.035$$

$Z_{0.035} = 1.812$

$$EBM = (z_{0.035})\left(\frac{\sigma}{\sqrt{n}}\right) = (1.812)\left(\frac{0.337}{\sqrt{20}}\right) = 0.1365$$

$\overline{x} - EBM = 0.940 - 0.1365 = 0.8035$

$\overline{x} + EBM = 0.940 + 0.1365 = 1.0765$

We estimate with 93% confidence that the true SAR mean for the population of cell phones in the United States is between 0.8035 and 1.0765 watts per kilogram.

Notice the difference in the confidence intervals calculated in [link] and the following Try It exercise. These intervals are different for several reasons: they were calculated from different samples, the samples were different sizes, and the intervals were calculated for different levels of confidence. Even though the intervals are different, they do not yield conflicting information. The effects of these kinds of changes are the subject of the next section in this chapter.

Changing the Confidence Level or Sample Size

Example

Suppose we change the original problem in [link] by using a 95% **confidence level**. Find a 95% confidence interval for the true (population) mean statistics exam score.

Solution

To find the **confidence interval**, you need the sample mean, $\overline{x}$, and the *EBM*.

- $\overline{x}$ = 68
- *EBM* = $\left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$
- σ = 3; n = 36; The confidence level is 95% (*CL* = 0.95).

CL = 0.95 so α = 1 – *CL* = 1 – 0.95 = 0.05

$$\frac{\alpha}{2} = 0.025 \text{ } \backslash\text{hspace}\{0.17\text{em}\}\backslash\text{hspace}\{0.17\text{em}\} z_{\frac{\alpha}{2}} = z_{0.025}$$

The area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is 1 – 0.025 = 0.975.

$$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$$

when using invnorm(0.975,0,1) on the TI-83, 83+, or 84+ calculators. (This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the standard normal distribution.)

EBM = (1.96)$\left(\frac{3}{\sqrt{36}}\right)$ = 0.98

$\overline{x}$ – *EBM* = 68 – 0.98 = 67.02

$\overline{x}$ + *EBM* = 68 + 0.98 = 68.98

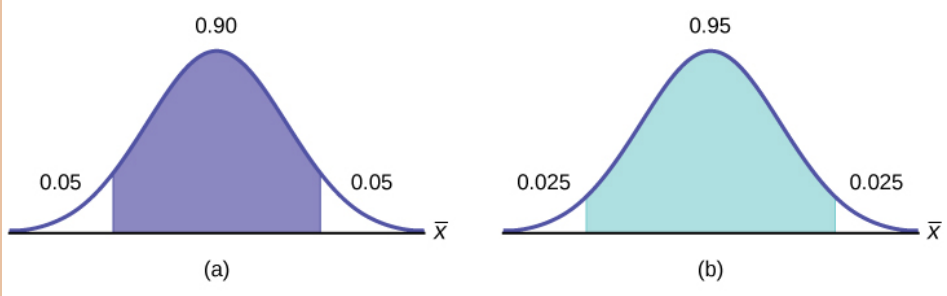
Notice that the *EBM* is larger for a 95% confidence level in the original problem.

Interpretation

We estimate with 95% confidence that the true population mean for all statistics exam scores is between 67.02 and 68.98.

Explanation of 95% Confidence Level
Ninety-five percent of all confidence intervals constructed in this way contain the true value of the population mean statistics exam score.

Comparing the results
The 90% confidence interval is (67.18, 68.82). The 95% confidence interval is (67.02, 68.98). The 95% confidence interval is wider. If you look at the graphs, because the area 0.95 is larger than the area 0.90, it makes sense that the 95% confidence interval is wider. To be more confident that the confidence interval actually does contain the true value of the population mean for all statistics exam scores, the confidence interval necessarily needs to be wider.



Summary: Effect of Changing the Confidence Level

- Increasing the confidence level increases the error bound, making the confidence interval wider.
- Decreasing the confidence level decreases the error bound, making the confidence interval narrower.

Your Turn!

Refer back to the pizza-delivery Try It exercise. The population standard deviation is six minutes and the sample mean delivery time is 36 minutes. Use a sample size of 20. Find a 95% confidence interval estimate for the true mean pizza delivery time.

Solution
(33.37, 38.63)

Example

Suppose we change the original problem in [link] to see what happens to the error bound if the sample size is changed.

Leave everything the same except the sample size. Use the original 90% **confidence level.** What happens to the error bound and the confidence interval if we increase the sample size and use n = 100 instead of n = 36? What happens if we decrease the sample size to n = 25 instead of n = 36?

- $\overline{x}$ = 68
- $EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$
- σ = 3; The confidence level is 90% (CL=0.90); $z_{\frac{\alpha}{2}} = z_{0.05}$ = 1.645.

Solution
Solution A: If we **increase** the sample size n to 100, we **decrease** the error bound.

When n = 100: $EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right) = (1.645)\left(\frac{3}{\sqrt{100}}\right) = 0.4935$.

Solution
Solution B: If we **decrease** the sample size n to 25, we **increase** the error bound.

When $n = 25$: $EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right) = (1.645)\left(\frac{3}{\sqrt{25}}\right) = 0.987$.

Summary: Effect of Changing the Sample Size

- Increasing the sample size causes the error bound to decrease, making the **confidence interval** narrower.
- Decreasing the sample size causes the error bound to increase, making the **confidence interval** wider.

Your Turn!

Refer back to the pizza-delivery Try It exercise. The mean delivery time is 36 minutes and the population standard deviation is six minutes. Assume the sample size is changed to 50 restaurants with the same sample mean. Find a 90% confidence interval estimate for the population mean delivery time.

Solution
(34.6041, 37.3958)

Working Backwards to Find the Error Bound or Sample Mean

When we calculate a confidence interval, we find the sample mean, calculate the error bound, and use them to calculate the confidence interval. However, sometimes when we read statistical studies, the study may state

the confidence interval only. If we know the confidence interval, we can work backwards to find both the error bound and the sample mean.

Finding the Error Bound

- From the upper value for the interval, subtract the sample mean,
- OR, from the upper value for the interval, subtract the lower value. Then divide the difference by two.

Finding the Sample Mean

- Subtract the error bound from the upper value of the confidence interval,
- OR, average the upper and lower endpoints of the confidence interval.

Notice that there are two methods to perform each calculation. You can choose the method that is easier to use with the information you know.

Example

Suppose we know that a confidence interval is **(67.18, 68.82)** and we want to find the error bound. We may know that the sample mean is 68, or perhaps our source only gave the confidence interval and did not tell us the value of the sample mean.

Calculate the Error Bound:

- If we know that the sample mean is 68: *EBM* = 68.82 – 68 = 0.82.
- If we don't know the sample mean: *EBM* = $\frac{(68.82 - 67.18)}{2}$ = 0.82.

Calculate the Sample Mean:

- If we know the error bound:
  $\overline{x}$ = 68.82 – 0.82 = 68
- If we don't know the error bound: $\overline{x}$ = $\frac{(67.18 + 68.82)}{2}$ = 68.

Your Turn!

Suppose we know that a confidence interval is (42.12, 47.88). Find the error bound and the sample mean.

Solution

Sample mean is 45, error bound is 2.88

The Sample Size *n*

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size.

The error bound formula for a population mean when the population standard deviation is known is $EBM = z\frac{\alpha}{2}\left(\frac{\sigma}{\sqrt{n}}\right)$.

The formula for sample size is $n = \frac{z^2\sigma^2}{EBM^2}$, found by solving the error bound formula for n.

In this formula, z is $z_{\frac{\alpha}{2}}$, corresponding to the desired confidence level. A researcher planning a study who wants a specified confidence level and error bound can use this formula to calculate the size of the sample needed for the study.

Example

The population standard deviation for the age of Foothill College students is 15 years. If we want to be 95% confident that the sample mean age is within two years of the true population mean age of Foothill College students, how many randomly selected Foothill College students must be surveyed?

- From the problem, we know that $\sigma = 15$ and $EBM = 2$.
- $z = z_{0.025} = 1.96$, because the confidence level is 95%.
- $n = \frac{z^2\sigma^2}{EBM^2} = \frac{(1.96)^2(15)^2}{2^2} = 216.09$ using the sample size equation.
- Use $n = 217$: Always round the answer UP to the next higher integer to ensure that the sample size is large enough.

Therefore, 217 Foothill College students should be surveyed in order to be 95% confident that we are within two years of the true population mean age of Foothill College students.

Your Turn!

The population standard deviation for the height of high school basketball players is three inches. If we want to be 95% confident that the sample mean height is within one inch of the true population mean height, how many randomly selected students must be surveyed?

Solution
35 students

References

"American Fact Finder." U.S. Census Bureau. Available online at http://factfinder2.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t (accessed July 2, 2013).

"Disclosure Data Catalog: Candidate Summary Report 2012." U.S. Federal Election Commission. Available online at http://www.fec.gov/data/index.jsp (accessed July 2, 2013).

"Headcount Enrollment Trends by Student Demographics Ten-Year Fall Trends to Most Recently Completed Fall." Foothill De Anza Community College District. Available online at http://research.fhda.edu/factbook/FH_Demo_Trends/FoothillDemographicTrends.htm (accessed September 30,2013).

Kuczmarski, Robert J., Cynthia L. Ogden, Shumei S. Guo, Laurence M. Grummer-Strawn, Katherine M. Flegal, Zuguo Mei, Rong Wei, Lester R. Curtin, Alex F. Roche, Clifford L. Johnson. "2000 CDC Growth Charts for the United States: Methods and Development." Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/growthcharts/2000growthchart-us.pdf (accessed July 2, 2013).

La, Lynn, Kent German. "Cell Phone Radiation Levels." c|net part of CBX Interactive Inc. Available online at http://reviews.cnet.com/cell-phone-radiation-levels/ (accessed July 2, 2013).

"Mean Income in the Past 12 Months (in 2011 Inflation-Adjusted Dollars): 2011 American Community Survey 1-Year Estimates." American Fact Finder, U.S. Census Bureau. Available online at http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_11_1YR_S1902&prodType=table (accessed July 2, 2013).

"Metadata Description of Candidate Summary File." U.S. Federal Election Commission. Available online at http://www.fec.gov/finance/disclosure/metadata/metadataforcandidatesummary.shtml (accessed July 2, 2013).

"National Health and Nutrition Examination Survey." Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/nchs/nhanes.htm (accessed July 2, 2013).

7.2 A SINGLE POPULATION MEAN USING THE STUDENT T DISTRIBUTION

Learning Objectives

By the end of this section, the student should be able to:

- Calculate Student's t-probabilities.
- Calculate the confidence interval

Student's t-probabilities.

In practice, we rarely know the population standard deviation. In the past, when the sample size was large, this did not present a problem to statisticians. They used the sample **standard deviation** s as an estimate for σ and proceeded as before to calculate a confidence interval with close enough results. However, statisticians ran into problems when the sample size was small. A small sample size caused inaccuracies in the confidence interval.

William S. Goset (1876–1937) of the Guinness brewery in Dublin, Ireland ran into this problem. His experiments with hops and barley produced very few samples. Just replacing σ with s did not produce accurate results when he tried to calculate a confidence interval. He realized that he could not use a **normal distribution** for the calculation; he found that the actual distribution depends on the sample size. This problem led him to "discover" what is called the Student's t-distribution. The name comes from the fact that Gosset wrote under the pen name "Student."

Up until the mid-1970s, some statisticians used the normal distribution approximation for large sample sizes and only used the Student's t-distribution only for sample sizes of at most 30. With graphing calculators and computers, the practice now is to use the Student's t-distribution whenever s is used as an estimate for σ.

If you draw a simple random sample of size n from a population that has an approximately a normal distribution with mean μ and unknown population **standard deviation** σ and calculate the t-score t =

$\frac{\overline{x}-\mu}{\left(\frac{s}{\sqrt{n}}\right)}$, then the *t*-scores follow a **Student's t-distribution with $n-1$ degrees of freedom**. The *t*-score has the same interpretation as the *z*-score. It measures how far $\overline{x}$ is from its mean μ. For each sample size *n*, there is a different **Student's t-distribution**.

The degrees of freedom, $n-1$, come from the calculation of the sample **standard deviation** s. In [link], we used *n* deviations $(x-\overline{x}\text{values})$ to calculate s. Because the sum of the deviations is zero, we can find the last deviation once we know the other $n-1$ deviations. The other $n-1$ deviations can change or vary freely. **We call the number $n-1$ the degrees of freedom (df).**

Properties of the Student's t-Distribution

- The graph for the **Student's t-distribution** is similar to the standard normal curve.
- The mean for the **Student's t-distribution** is zero and the distribution is symmetric about zero.
- The **Student's t-distribution** has more probability in its tails than the standard normal distribution because the spread of the t-distribution is greater than the spread of the standard normal. So the graph of the **Student's t-distribution** will be thicker in the tails and shorter in the center than the graph of the standard normal distribution.
- The exact shape of the **Student's t-distribution** depends on the degrees of freedom. As the degrees of freedom increases, the graph of Student's t-distribution becomes more like the graph of the standard normal distribution.
- The underlying population of individual observations is assumed to be normally distributed with unknown population mean μ and unknown population standard deviation σ. The size of the underlying population is generally not relevant unless it is very small. If it is bell shaped (normal) then the assumption is met and doesn't need discussion. Random sampling is assumed, but that is a completely separate assumption from normality.

Calculators and computers can easily calculate any Student's t-probabilities. The TI-83,83+, and 84+ have a tcdf function to find the probability for given values of *t*. The grammar for the tcdf command is tcdf(lower bound, upper bound, degrees of freedom). However for confidence intervals, we need to use **inverse** probability to find the value of *t* when we know the probability.

For the TI-84+ you can use the invT command on the DISTRibution menu. The invT command works similarly to the invnorm. The invT command requires two inputs: **invT(area to the left, degrees of freedom)** The output is the t-score that corresponds to the area we specified.

The TI-83 and 83+ do not have the invT command. (The TI-89 has an inverse T command.)

A probability table for the Student's t-distribution can also be used. The table gives t-scores that correspond to the confidence level (column) and degrees of freedom (row). (The TI-86 does not have an invT program or command, so if you are using that calculator, you need to use a probability table for the **Student's t-**

Distribution.) When using a *t*-table, note that some tables are formatted to show the confidence level in the column headings, while the column headings in some tables may show only the corresponding area in one or both tails.

A Student's t table (See [link]) gives *t*-scores given the degrees of freedom and the right-tailed probability. The table is very limited. **Calculators and computers can easily calculate any Student's t-probabilities.**
The notation for the Student's t-distribution (using *T* as the random variable) is:

- $T \sim t_{df}$ where $df = n - 1$.
- For example, if we have a sample of size $n = 20$ items, then we calculate the degrees of freedom as $df = n - 1 = 20 - 1 = 19$ and we write the distribution as $T \sim t_{19}$.

The Confidence Interval

If the population standard deviation is not known, the error bound for a population mean is:

- $EBM = \left(t_{\frac{\alpha}{2}}\right)\left(\frac{s}{\sqrt{n}}\right)$,
- $t_{\frac{\sigma}{2}}$ is the *t*-score with area to the right equal to $\frac{\alpha}{2}$,
- use $df = n - 1$ degrees of freedom, and
- s = sample standard deviation.

The format for the confidence interval is:

$$(\overline{x} - EBM, \overline{x} + EBM).$$

To calculate the confidence interval directly:

Press STAT.
Arrow over to TESTS.
Arrow down to 8:TInterval and press ENTER (or just press 8).

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=851#oembed-1

Example

Suppose you do a study of acupuncture to determine how effective it is in relieving pain. You measure sensory rates for 15 subjects with the results given. Use the sample data to construct a 95% confidence interval for the mean sensory rate for the population (assumed normal) from which you took the data.

The solution is shown step-by-step and by using the TI-83, 83+, or 84+ calculators.

8.69.47.96.88.37.39.29.68.711.410.35.48.15.56.9
Solution

- The first solution is step-by-step (Solution A).
- The second solution uses the TI-83+ and TI-84 calculators (Solution B).

Solution A To find the confidence interval, you need the sample mean, $\overline{x}$, and the *EBM*.

$\overline{x}$ = 8.2267 s = 1.6722 n = 15

df = 15 – 1 = 14 CL so α = 1 – CL = 1 – 0.95 = 0.05

$\frac{\alpha}{2}$ = 0.025 $t_{\frac{\alpha}{2}} = t_{0.025}$

The area to the right of $t_{0.025}$ is 0.025, and the area to the left of $t_{0.025}$ is 1 – 0.025 = 0.975

$t_{\frac{\alpha}{2}} = t_{0.025} = 2.14$ using invT(.975,14) on the TI-84+ calculator.

$$EBM = \left(t_{\frac{\alpha}{2}}\right)\left(\frac{s}{\sqrt{n}}\right)$$

$$EBM = (2.14)\left(\frac{1.6722}{\sqrt{15}}\right) = 0.924$$

$\overline{x}$ – *EBM* = 8.2267 – 0.9240 = 7.3

$\overline{x}$ + *EBM* = 8.2267 + 0.9240 = 9.15

The 95% confidence interval is (7.30, 9.15).

We estimate with 95% confidence that the true population mean sensory rate is between 7.30 and 9.15.

Solution

Press `STAT` and arrow over to `TESTS`.

Arrow down to 8`:TInterval` and press `ENTER` (or you can just press 8).

Arrow to `Data` and press `ENTER`.

Arrow down to `List` and enter the list name where you put the data.

There should be a 1 after `Freq`.

Arrow down to `C-level` and enter 0.95

Arrow down to `Calculate` and press `ENTER`.

The 95% confidence interval is (7.3006, 9.1527)

Note

When calculating the error bound, a probability table for the **Student's t-distribution** can also be used to find the value of *t*. The table gives *t*-scores that correspond to the confidence level

(column) and degrees of freedom (row); the *t*-score is found where the row and column intersect in the table.

Your Turn!

You do a study of hypnotherapy to determine how effective it is in increasing the number of hours of sleep subjects get each night. You measure hours of sleep for 12 subjects with the following results. Construct a 95% confidence interval for the mean number of hours slept for the population (assumed normal) from which you took the data.

8.2; 9.1; 7.7; 8.6; 6.9; 11.2; 10.1; 9.9; 8.9; 9.2; 7.5; 10.5

Solution
(8.1634, 9.8032)

Exercises

Suppose you do a study of acupuncture to determine how effective it is in relieving pain. You measure sensory rates for 15 subjects with the results given. Plots of the data show no skewness or outliers. Which distribution is appropriate to use here?

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=851#h5p-166

Example

The Human Toxome Project (HTP) is working to understand the scope of industrial pollution in the human body. Industrial chemicals may enter the body through pollution or as ingredients in consumer products. In October 2008, the scientists at HTP tested cord blood samples for 20 newborn infants in the United States. The cord blood of the "In utero/newborn" group was tested for 430 industrial compounds, pollutants, and other chemicals, including chemicals linked to brain and nervous system toxicity, immune system toxicity, and reproductive toxicity, and fertility problems. There are health concerns about the effects of some chemicals on the brain and nervous system. [link] shows how many of the targeted chemicals were found in each infant's cord blood.

79	145	147	160	116	100	159	151	156	126
137	83	156	94	121	144	123	114	139	99

Use this sample data to construct a 90% confidence interval for the mean number of targeted industrial chemicals to be found in an infant's blood.

Solution

Solution A: From the sample, you can calculate $\overline{x}$ = 127.45 and s = 25.965. There are 20 infants in the sample, so n = 20, and df = 20 – 1 = 19.

You are asked to calculate a 90% confidence interval: *CL* = 0.90, so $\alpha = 1 - CL = 1 - 0.90 = 0.10$

$$\frac{\alpha}{2} = 0.05,\ t_{\frac{\alpha}{2}} = t_{0.05}$$

By definition, the area to the right of $t_{0.05}$ is 0.05 and so the area to the left of $t_{0.05}$ is 1 – 0.05 = 0.95.

Use a table, calculator, or computer to find that $t_{0.05} = 1.729$.

$$EBM = t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right) = 1.729 \left(\frac{25.965}{\sqrt{20}} \right) \approx 10.038$$

$\overline{x}$ – *EBM* = 127.45 – 10.038 = 117.412

$\overline{x}$ + *EBM* = 127.45 + 10.038 = 137.488

We estimate with 90% confidence that the mean number of all targeted industrial chemicals found in cord blood in the United States is between 117.412 and 137.488.

Solution

Solution B

Enter the data as a list.

Press `STAT` and arrow over to `TESTS`.

Press `STAT` and arrow over to `TESTS`.

Arrow down to `8:TInterval` and press `ENTER` (or you can just press 8).

Arrow to `Data` and press `ENTER`.

Arrow down to `List` and enter the list name where you put the data.

There should be a 1 after `Freq`.

Arrow down to `C-level` and enter 0.90

Arrow down to `Calculate` and press `ENTER`.

The 90% confidence interval is (117.41, 137.49)

Your Turn!

A random sample of statistics students were asked to estimate the total number of hours they spend watching television in an average week. The responses are recorded in [link]. Use this sample data to construct a 98% confidence interval for the mean number of hours statistics students will spend watching television in one week.

0	3	1	20	9
5	10	1	10	4
14	2	4	4	5

Solution

Solution A

$\overline{x}$ = 6.133, s = 5.514, n = 15, and df = 15 – 1 = 14

CL = 0.98, so α = 1 – CL = 1 – 0.98 = 0.02

$$\frac{\alpha}{2} = 0.01 t_{\frac{\alpha}{2}} = t_{0.01} = 2.624$$

$$EBM = t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right) = 2.624 \left(\frac{5.514}{\sqrt{15}} \right) \ 3.736$$

$\overline{x}$ – EBM = 6.133 – 3.736 = 2.397

$\overline{x}$ + EBM = 6.133 + 3.736 = 9.869

We estimate with 98% confidence that the mean number of all hours that statistics students spend watching television in one week is between 2.397 and 9.869.

Solution

Solution B

Enter the data as a list.

Press `STAT` and arrow over to `TESTS`.

Arrow down to `8:TInterval`.

Press ENTER.

Arrow to Data and press ENTER.

Arrow down and enter the name of the list where the data is stored.

Enter Freq: 1

Enter C-Level: 0.98

Arrow down to Calculate and press Enter.

The 98% confidence interval is (2.3965, 9.8702).



One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=851#oembed-2

References

"America's Best Small Companies." Forbes, 2013. Available online at http://www.forbes.com/best-small-companies/list/ (accessed July 2, 2013).

Data from *Microsoft Bookshelf*.

Data from http://www.businessweek.com/.

Data from http://www.forbes.com/.

"Disclosure Data Catalog: Leadership PAC and Sponsors Report, 2012." Federal Election Commission. Available online at http://www.fec.gov/data/index.jsp (accessed July 2,2013).

"Human Toxome Project: Mapping the Pollution in People." Environmental Working Group. Available online at http://www.ewg.org/sites/humantoxome/participants/participant-group.php?group=in+utero%2Fnewborn (accessed July 2, 2013).

"Metadata Description of Leadership PAC List." Federal Election Commission. Available online at http://www.fec.gov/finance/disclosure/metadata/metadataLeadershipPacList.shtml (accessed July 2, 2013).

7.3 A POPULATION PROPORTION

Learning Objectives

By the end of this section, the student should be able to:

- Calculate the error bound for a proportion.
- Calculate the confidence interval of p.
- Calculate the sample size n.

The error bound for a proportion.

During an election year, we see articles in the newspaper that state confidence intervals in terms of proportions or percentages. For example, a poll for a particular candidate running for president might show that the candidate has 40% of the vote within three percentage points (if the sample is large enough). Often, election polls are calculated with 95% confidence, so, the pollsters would be 95% confident that the true proportion of voters who favored the candidate would be between 0.37 and 0.43: (0.40 – 0.03,0.40 + 0.03).

Investors in the stock market are interested in the true proportion of stocks that go up and down each week. Businesses that sell personal computers are interested in the proportion of households in the United States that own personal computers. Confidence intervals can be calculated for the true proportion of stocks that go up or down each week and for the true proportion of households in the United States that own personal computers.

The procedure to find the confidence interval, the sample size, the error bound, and the confidence level for a proportion is similar to that for the population mean, but the
formulas are different.

How do you know you are dealing with a proportion problem? First, the underlying **distribution is a** binomial distribution. (There is no mention of a mean or average.) If X is a binomial random variable, then $X \sim B(n, p)$ where n is the number of trials and p is the probability of a success. To form a proportion, take X,

the random variable for the number of successes and divide it by n, the number of trials (or the sample size). The random variable P' (read "P prime") is that proportion,

$$P' = \frac{X}{n}$$

(Sometimes the random variable is denoted as $\hat{P}$, read "P hat".)

When n is large and p is not close to zero or one, we can use the normal distribution to approximate the binomial.

$$X \sim N\left(np, \sqrt{npq}\right)$$

If we divide the random variable, the mean, and the standard deviation by n, we get a normal distribution of proportions with P', called the estimated proportion, as the random variable. (Recall that a proportion as the number of successes divided by n.)

$$\frac{X}{n} = P' \sim N\left(\frac{np}{n}, \frac{\sqrt{npq}}{n}\right)$$

Using algebra to simplify : $\frac{\sqrt{npq}}{n} = \sqrt{\frac{pq}{n}}$

***P'* follows a normal distribution for proportions**:

$$\frac{X}{n} = P' \sim N\left(\frac{np}{n}, \frac{\sqrt{npq}}{n}\right)$$

The confidence interval has the form ($p' - EBP$, $p' + EBP$). ***EBP*** is error bound for the proportion.

$p' = \frac{x}{n}$

p' = the **estimated proportion** of successes (p' is a **point estimate** for p, the true proportion.)

x = the **number** of successes

n = the size of the sample

The error bound for a proportion is

$$EBP = \left(z_{\frac{\alpha}{2}}\right)\left(\sqrt{\frac{p'q'}{n}}\right) \text{ where } q' = 1 - p'$$

This formula is similar to the error bound formula for a mean, except that the "appropriate standard deviation" is different. For a mean, when the population standard deviation is known, the appropriate standard deviation that we use is $\frac{\sigma}{\sqrt{n}}$. For a proportion, the appropriate standard deviation is $\sqrt{\frac{pq}{n}}$.

However, in the error bound formula, we use

$\sqrt{\frac{p'q'}{n}}$ as the standard deviation, instead of $\sqrt{\frac{pq}{n}}$.

In the error bound formula, the **sample proportions p' and q' are estimates of the unknown**

population proportions p and q. The estimated proportions p' and q' are used because p and q are not known. The sample proportions p' and q' are calculated from the data: p' is the estimated proportion of successes, and q' is the estimated proportion of failures.

The confidence interval can be used only if the number of successes np' and the number of failures nq' are both greater than five.

Example

A random sample of 25 statistics students was asked: "Have you smoked a cigarette in the past week?" Six students reported smoking within the past week. Use the "plus-four" method to find a 95% confidence interval for the true proportion of statistics students who smoke.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=855#h5p-172

Note

For the normal distribution of proportions, the z-score formula is as follows.

If $P' \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$ then the z-score formula is $z = \frac{p' - p}{\sqrt{\frac{pq}{n}}}$

Example

Suppose that a market research firm is hired to estimate the percent of adults living in a large city who have cell phones. Five hundred randomly selected adult residents in this city are surveyed to determine whether they have cell phones. Of the 500 people surveyed, 421 responded yes – they own cell phones. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of adult residents of this city who have cell phones.

Solution

Solution A

- The first solution is step-by-step (Solution A).
- The second solution uses a function of the TI-83, 83+ or 84 calculators (Solution B).

Let X = the number of people in the sample who have cell phones. X is binomial. $X\ B\left(500, \frac{421}{500}\right)$.

To calculate the confidence interval, you must find p', q', and ***EBP***.

n = 500

x = the number of successes = 421

$$p' = \frac{x}{n} = \frac{421}{500} = 0.842$$

p' = 0.842 is the sample proportion; this is the point estimate of the population proportion.

$q' = 1 - p' = 1 - 0.842 = 0.158$

Since CL = 0.95, then $\alpha = 1 - CL = 1 - 0.95 = 0.05$ $\left(\frac{\alpha}{2}\right)$ = 0.025.

Then $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$

Use the TI-83, 83+, or 84+ calculator command invNorm(0.975,0,1) to find $z_{0.025}$. Remember that the area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is 0.975. This can also be found using appropriate commands on other calculators, using a computer, or using a Standard Normal probability table.

$$EBP = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{p'q'}{n}} = (1.96)\sqrt{\frac{(0.842)\,(0.158)}{500}} = 0.032$$

$$p\text{'} - EBP = 0.842 - 0.032 = 0.81$$

$$p' + EBP = 0.842 + 0.032 = 0.874$$

The confidence interval for the true binomial population proportion is (p' – *EBP*, p' + *EBP*) = (0.810, 0.874).

Interpretation we estimate with 95% confidence that between 81% and 87.4% of all adult residents of this city have cell phones.

Explanation of 95% Confidence Level

Ninety-five percent of the confidence intervals constructed in this way would contain the true value for the population proportion of all adult residents of this city who have cell phones. The confidence interval is (0.81003, 0.87397).
The confidence interval is (0.81003, 0.87397).

Solution
Solution B

Press `STAT` and arrow over to `TESTS`.

Arrow down to `A:1-PropZint`. Press `ENTER`.

Arrow down to x and enter 421.

Arrow down to n and enter 500.

Arrow down to `C-Level` and enter .95.

Arrow down to `Calculate` and press `ENTER`.

The confidence interval is (0.81003, 0.87397).

Your Turn!

Suppose 250 randomly selected people are surveyed to determine if they own a tablet. Of the 250 surveyed, 98 reported owning a tablet. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of people who own tablets.

Solution
(0.3315, 0.4525)

Example

For a class project, a political science student at a large university wants to estimate the percent of students who are registered voters. He surveys 500 students and finds that 300 are registered voters. Compute a 90% confidence interval for the true percent of students who are registered voters, and interpret the confidence interval.

Solution

- The first solution is step-by-step (Solution A).
- The second solution uses a function of the TI-83, 83+, or 84 calculators (Solution B).

Solution A$x = 300$ and $n = 500$

$$p' = \frac{x}{n} = \frac{300}{500} = 0.600$$

$$q' = 1 - p' = 1 - 0.600 = 0.400$$

Since $CL = 0.90$, then $\alpha = 1 - CL = 1 - 0.90 = 0.10$ $\left(\frac{\alpha}{2}\right) = 0.05$

$z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$

Use the TI-83, 83+, or 84+ calculator command invNorm(0.95,0,1) to find $z_{0.05}$. Remember that the area to the right of $z_{0.05}$ is 0.05 and the area to the left of $z_{0.05}$ is 0.95. This can also be found using appropriate commands on other calculators, using a computer, or using a standard normal probability table.

$$EBP = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{p'q'}{n}} = (1.645)\sqrt{\frac{(0.60)\,(0.40)}{500}} = 0.036$$

$$p' - EBP = 0.60 - 0.036 = 0.564$$

$$p' + EBP = 0.60 + 0.036 = 0.636$$

The confidence interval for the true binomial population proportion is (p' – *EBP*, p' + *EBP*) = (0.564,0.636).

Interpretation

- We estimate with 90% confidence that the true percent of all students that are registered voters is between 56.4% and 63.6%.
- Alternate Wording: We estimate with 90% confidence that between 56.4% and 63.6% of ALL students are registered voters.

Explanation of 90% Confidence Level Ninety percent of all confidence intervals constructed in this way contain the true value for the population percent of students that are registered voters.

Solution
Solution B

Press `STAT` and arrow over to `TESTS`.

Arrow down to `A:1-PropZint`. Press `ENTER`.

Arrow down to x and enter 300.

Arrow down to n and enter 500.

Arrow down to `C-Level` and enter 0.90.

Arrow down to `Calculate` and press `ENTER`.

The confidence interval is (0.564, 0.636).

Your Turn!

A student polls his school to see if students in the school district are for or against the new legislation regarding school uniforms. She surveys 600 students and finds that 480 are against the new legislation.

a. Compute a 90% confidence interval for the true percent of students who are against the new legislation, and interpret the confidence interval.

Solution
(0.7731, 0.8269); We estimate with 90% confidence that the true percent of all students in the district who are against the new legislation is between 77.31% and 82.69%.

b. In a sample of 300 students, 68% said they own an iPod and a smartphone. Compute a 97% confidence interval for the true percent of students who own an iPod and a smartphone.

Solution
Solution A
Sixty-eight percent (68%) of students own an iPod and a smartphone.

$$p' = 0.68$$

$$q' = 1 - p' = 1 - 0.68 = 0.32$$

Since $CL = 0.97$, we know $\alpha = 1 - 0.97 = 0.03$ and $\frac{\alpha}{2} = 0.015$.

The area to the left of $z_{0.015}$ is 0.015, and the area to the right of $z_{0.015}$ is 1 – 0.015 = 0.985.

Using the TI 83, 83+, or 84+ calculator function InvNorm(.985,0,1),

$$z_{0.015} = 2.17$$

$$EPB = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{p'q'}{n}} = 2.17\sqrt{\frac{0.68\,(0.32)}{300}} \approx 0.0269$$

p' – EPB = 0.68 – 0.0269 = 0.6531

p' + EPB = 0.68 + 0.0269 = 0.7069

We are 97% confident that the true proportion of all students who own an iPod and a smartphone is between 0.6531 and 0.7069.

Solution
Solution B

Press STAT and arrow over to TESTS.

Arrow down to A:1-PropZint. Press ENTER.

Arrow down to x and enter 300*0.68.

Arrow down to n and enter 300.

Arrow down to C-Level and enter 0.97.

Arrow down to Calculate and press ENTER.

The confidence interval is (0.6531, 0.7069).

Exercises

Suppose that a market research firm is hired to estimate the percent of adults living in a large city who have cell phones. Five hundred randomly selected adult residents in this city are surveyed to determine whether they have cell phones. Of the 500 people surveyed, 421 responded yes – they own cell phones. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of adult residents of this city who have cell phones.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=855#h5p-170

"Plus Four" Confidence Interval for *p*

There is a certain amount of error introduced into the process of calculating a confidence interval for a

proportion. Because we do not know the true proportion for the population, we are forced to use point estimates to calculate the appropriate standard deviation of the sampling distribution. Studies have shown that the resulting estimation of the standard deviation can be flawed.

Fortunately, there is a simple adjustment that allows us to produce more accurate confidence intervals. We simply pretend that we have four additional observations. Two of these observations are successes and two are failures. The new sample size, then, is $n + 4$, and the new count of successes is $x + 2$.

Computer studies have demonstrated the effectiveness of this method. It should be used when the confidence level desired is at least 90% and the sample size is at least ten.

Example

A random sample of 25 statistics students was asked: "Have you smoked a cigarette in the past week?" Six students reported smoking within the past week. Use the "plus-four" method to find a 95% confidence interval for the true proportion of statistics students who smoke.

Solution

Solution A

Six students out of 25 reported smoking within the past week, so $x = 6$ and $n = 25$. Because we are using the "plus-four" method, we will use $x = 6 + 2 = 8$ and $n = 25 + 4 = 29$.

$$p' = \frac{x}{n} = \frac{8}{29} \approx 0.276$$

$$q' = 1 - p' = 1 - 0.276 = 0.724$$

Since $CL = 0.95$, we know $\alpha = 1 - 0.95 = 0.05$ and $\frac{\alpha}{2} = 0.025$.

$$z_{0.025} = 1.96$$

$$EPB = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{p'q'}{n}} = (1.96)\sqrt{\frac{0.276\,(0.724)}{29}} \approx 0.163$$

$p' - EPB = 0.276 - 0.163 = 0.113$

$p' + EPB = 0.276 + 0.163 = 0.439$

We are 95% confident that the true proportion of all statistics students who smoke cigarettes is between 0.113 and 0.439.

Solution
Solution B

Press STAT and arrow over to TESTS.

Arrow down to A:1-PropZint. Press ENTER.

Reminder

Remember that the plus-four method assumes an additional four trials: two successes and two failures. You do not need to change the process for calculating the confidence interval; simply update the values of x and n to reflect these additional trials.

Arrow down to *x* and enter eight.

Arrow down to *n* and enter 29.

Arrow down to C-Level and enter 0.95.

Arrow down to Calculate and press ENTER.

The confidence interval is (0.113, 0.439).

Your Turn!

Out of a random sample of 65 freshmen at State University, 31 students have declared a major. Use the "plus-four" method to find a 96% confidence interval for the true proportion of freshmen at State University who have declared a major.

Solution
Solution A
Using "plus four," we have $x = 31 + 2 = 33$ and $n = 65 + 4 = 69$.

$$p' = \frac{33}{69} \approx 0.478$$

$$q' = 1 - p' = 1 - 0.478 = 0.522$$

Since *CL* = 0.96, we know α = 1 – 0.96 = 0.04 and $\frac{\alpha}{2}$ = 0.02.

$$z_{0.02} = 2.054$$

$$EPB = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{p'q'}{n}} = (2.054)\left(\sqrt{\frac{(0.478)\,(0.522)}{69}}\right)\ 0.124$$

p′ – *EPB* = 0.478 – 0.124 = 0.354

p′ + *EPB* = 0.478 + 0.124 = 0.602

We are 96% confident that between 35.4% and 60.2% of all freshmen at State U have declared a major.

Solution
Solution B

Press STAT and arrow over to TESTS.

Arrow down to A:1-PropZint. Press ENTER.

Arrow down to *x* and enter 33.

Arrow down to *n* and enter 69.

Arrow down to C-Level and enter 0.96.

Arrow down to Calculate and press ENTER.

The confidence interval is (0.355, 0.602).

Example

The Berkman Center for Internet & Society at Harvard recently conducted a study analyzing the privacy management habits of teen internet users. In a group of 50 teens, 13 reported having more than 500 friends on Facebook. Use the "plus four" method to find a 90% confidence interval for the true proportion of teens who would report having more than 500 Facebook friends.

Solution

Solution A

Using "plus-four," we have x = 13 + 2 = 15 and n = 50 + 4 = 54.

$$p' = \frac{15}{54} \approx 0.278$$

$$q' = 1 - p' = 1 - 0.241 = 0.722$$

Since CL = 0.90, we know α = 1 – 0.90 = 0.10 and $\frac{\alpha}{2}$ = 0.05.

$$z_{0.05} = 1.645$$

$$EPB = \left(z_{\frac{\alpha}{2}}\right)\left(\sqrt{\frac{p'q'}{n}}\right) = (1.645)\left(\sqrt{\frac{(0.278)\,(0.722)}{54}}\right) \approx 0.100$$

p' – EPB = 0.278 – 0.100 = 0.178

p' + EPB = 0.278 + 0.100 = 0.378

We are 90% confident that between 17.8% and 37.8% of all teens would report having more than 500 friends on Facebook.

Solution

Solution B

Press STAT and arrow over to TESTS.

Arrow down to A:1-PropZint. Press ENTER.

Arrow down to x and enter 15.

Arrow down to n and enter 54.

Arrow down to C-Level and enter 0.90.

Arrow down to Calculate and press ENTER.

The confidence interval is (0.178, 0.378).

Your Turn!

The Berkman Center Study referenced in [link] talked to teens in smaller focus groups, but also interviewed additional teens over the phone. When the study was complete, 588 teens had answered the question about their Facebook friends with 159 saying that they have more than 500 friends. Use the "plus-four" method to find a 90% confidence interval for the true proportion of teens that would report having more than 500 Facebook friends based on this larger sample. Compare the results to those in [link].

Solution

Solution A

Using "plus-four," we have $x = 159 + 2 = 161$ and $n = 588 + 4 = 592$.

{p}^{\prime }= \frac{161}{592} \approx 0.272

$$q' = 1 - p' = 1 - 0.272 = 0.728$$

Since $CL = 0.90$, we know $\alpha = 1 - 0.90 = 0.10$ and $\frac{\alpha}{2} = 0.05$

$$EPB = \left(z_{\frac{\alpha}{2}}\right)\left(\sqrt{\frac{p'q'}{n}}\right) = (1.645)\left(\sqrt{\frac{(0.272)(0.728)}{592}}\right) \approx 0.030$$

$p' - EPB = 0.272 - 0.030 = 0.242$

$p' + EPB = 0.272 + 0.030 = 0.302$

We are 90% confident that between 24.2% and 30.2% of all teens would report having more than 500 friends on Facebook.

Solution

Solution B

Press STAT and arrow over to TESTS.

Arrow down to A:1-PropZint. Press ENTER.

Arrow down to x and enter 161.

Arrow down to n and enter 592.

Arrow down to C-Level and enter 0.90.

Arrow down to Calculate and press ENTER.

The confidence interval is (0.242, 0.302).

Conclusion: The confidence interval for the larger sample is narrower than the interval from [link].

Larger samples will always yield more precise confidence intervals than smaller samples. The "plus four" method has a greater impact on the smaller sample. It shifts the point estimate from 0.26 (13/50) to 0.278 (15/54). It has a smaller impact on the *EPB*, changing it from 0.102 to 0.100. In the larger sample, the point estimate undergoes a smaller shift: from 0.270 (159/588) to 0.272 (161/592). It is easy to see that the plus-four method has the greatest impact on smaller samples.



One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=855#oembed-1

Watch this video: Confidence Interval for a population proportion by Excel is Fun [8:34]



One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=855#oembed-2

Watch this video: Confidence Interval for a population proportion by Excel is Fun [4:51]

The Sample Size *n*

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size.

The error bound formula for a population proportion is

- $EBP = \left(z_{\frac{\alpha}{2}}\right)\left(\sqrt{\frac{p'q'}{n}}\right)$
- Solving for n gives you an equation for the sample size.
- $n = \frac{\left(z_{\frac{\alpha}{2}}\right)^2 (p'q')}{EBP^2}$

Example

Suppose a mobile phone company wants to determine the current percentage of customers aged 50+ who use text messaging on their cell phones. How many customers aged 50+ should the company survey in order to be 90% confident that the estimated (sample) proportion is within three percentage points of the true population proportion of customers aged 50+ who use text messaging on their cell phones.

Solution

From the problem, we know that ***EBP* = 0.03** (3%=0.03) and

$z_{\frac{\alpha}{2}}$ $z_{0.05}$ = 1.645 because the confidence level is 90%.

However, in order to find *n*, we need to know the estimated (sample) proportion *p*'. Remember that *q*' = 1 – *p*'. But, we do not know *p*' yet. Since we multiply *p*' and *q*' together, we make them both equal to 0.5 because *p*'*q*' = (0.5)(0.5) = 0.25 results in the largest possible product. (Try other products: (0.6)(0.4) = 0.24; (0.3)(0.7) = 0.21; (0.2)(0.8) = 0.16 and so on). The largest possible product gives us the largest *n*. This gives us a large enough sample so that we can be 90% confident that we are within three percentage points of the true population proportion. To calculate the sample size *n*, use the formula and make the substitutions.

$$n = \frac{z^2 p' q'}{EBP^2} \text{ gives } n = \frac{1.645^2\ (0.5)\ (0.5)}{0.03^2} = 751.7$$

Round the answer to the next higher value. The sample size should be 752 cell phone customers aged 50+ in order to be 90% confident that the estimated (sample) proportion is within three percentage points of the true population proportion of all customers aged 50+ who use text messaging on their cell phones.

Your Turn!

Suppose an internet marketing company wants to determine the current percentage of customers who click on ads on their smartphones. How many customers should the company survey in order to be 90% confident that the estimated proportion is within five percentage points of the true population proportion of customers who click on ads on their smartphones?

Solution

271 customers should be surveyed. Check the Real Estate section in your local

Exercises

Suppose a mobile phone company wants to determine the current percentage of customers aged 50+ who use text messaging on their cell phones. How many customers aged 50+ should the company survey in order to be 90% confident that the estimated (sample) proportion is within three percentage points of the true population proportion of customers aged 50+ who use text messaging on their cell phones?



An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=855#h5p-171

References

Jensen, Tom. "Democrats, Republicans Divided on Opinion of Music Icons." Public Policy Polling. Available online at http://www.publicpolicypolling.com/Day2MusicPoll.pdf (accessed July 2, 2013).

Madden, Mary, Amanda Lenhart, Sandra Coresi, Urs Gasser, Maeve Duggan, Aaron Smith, and Meredith Beaton. "Teens, Social Media, and Privacy." PewInternet, 2013. Available online at http://www.pewinternet.org/Reports/2013/Teens-Social-Media-And-Privacy.aspx (accessed July 2, 2013).

Prince Survey Research Associates International. "2013 Teen and Privacy Management Survey." Pew Research Center: Internet and American Life Project. Available online at http://www.pewinternet.org/~/media//Files/Questionnaire/2013/Methods%20and%20Questions_Teens%20and%20Social%20Media.pdf (accessed July 2, 2013).

Saad, Lydia. "Three in Four U.S. Workers Plan to Work Past Retirement Age: Slightly more say they will do this by choice rather than necessity." Gallup® Economy, 2013. Available online at http://www.gallup.com/poll/162758/three-four-workers-plan-work-past-retirement-age.aspx (accessed July 2, 2013).

The Field Poll. Available online at http://field.com/fieldpollonline/subscribers/ (accessed July 2, 2013).

Zogby. "New SUNYIT/Zogby Analytics Poll: Few Americans Worry about Emergency Situations Occurring in Their Community; Only one in three have an Emergency Plan; 70% Support Infrastructure 'Investment' for National Security." Zogby Analytics, 2013. Available online at http://www.zogbyanalytics.com/news/299-americans-neither-worried-nor-prepared-in-case-of-a-disaster-sunyit-zogby-analytics-poll (accessed July 2, 2013).

"52% Say Big-Time College Athletics Corrupt Education Process." Rasmussen Reports, 2013. Available online at http://www.rasmussenreports.com/public_content/lifestyle/sports/may_2013/52_say_big_time_college_athletics_corrupt_education_process (accessed July 2, 2013).

CHAPTER 7 REVIEW

7.1 Chapter Review

In this module, we learned how to calculate the confidence interval for a single population mean where the population standard deviation is known. When estimating a population mean, the margin of error is called the error bound for a population mean (*EBM*). A confidence interval has the general form:

(lower bound, upper bound) = (point estimate – *EBM*, point estimate + *EBM*)

The calculation of *EBM* depends on the size of the sample and the level of confidence desired. The confidence level is the percent of all possible samples that can be expected to include the true population parameter. As the confidence level increases, the corresponding *EBM* increases as well. As the sample size increases, the *EBM* decreases. By the central limit theorem,

$$EBM = z\frac{\sigma}{\sqrt{n}}$$

Given a confidence interval, you can work backwards to find the error bound (*EBM*) or the sample mean. To find the error bound, find the difference of the upper bound of the interval and the mean. If you do not know the sample mean, you can find the error bound by calculating half the difference of the upper and lower bounds. To find the sample mean given a confidence interval, find the difference of the upper bound and the error bound. If the error bound is unknown, then average the upper and lower bounds of the confidence interval to find the sample mean.

Sometimes researchers know in advance that they want to estimate a population mean within a specific margin of error for a given level of confidence. In that case, solve the *EBM* formula for n to discover the size of the sample that is needed to achieve this goal:

$$n = \frac{z^2\sigma^2}{EBM^2}$$

Formula Review

$\overline{X} \sim N\left(\mu_X, \frac{\sigma}{\sqrt{n}}\right)$ The distribution of sample means is normally distributed with mean equal to the population mean and standard deviation given by the population standard deviation divided by the square root of the sample size.

The general form for a confidence interval for a single population mean, known standard deviation, normal distribution is given by

(lower bound, upper bound) = (point estimate – *EBM*, point estimate + *EBM*)

$= (\overline{x} - EBM, \overline{x} + EBM)$

$= \left(\overline{x} - z\frac{\sigma}{\sqrt{n}}, \overline{x} + z\frac{\sigma}{\sqrt{n}}\right)$

EBM = $z\frac{\sigma}{\sqrt{n}}$ = the error bound for the mean, or the margin of error for a single population mean; this formula is used when the population standard deviation is known.

CL = confidence level, or the proportion of confidence intervals created that are expected to contain the true population parameter

$\alpha = 1 - CL$ = the proportion of confidence intervals that will not contain the population parameter

$z_{\frac{\alpha}{2}}$ = the *z*-score with the property that the area to the right of the *z*-score is $\frac{\alpha}{2}$ this is the *z*-score used in the calculation of "*EBM* where $\alpha = 1 - CL$.

$n = \frac{z^2\sigma^2}{EBM^2}$ = the formula used to determine the sample size (n) needed to achieve a desired margin of error at a given level of confidence

General form of a confidence interval

(lower value, upper value) = (point estimate–error bound, point estimate + error bound)

To find the error bound when you know the confidence interval

error bound = upper value–point estimate OR error bound = $\frac{\text{upper value} - \text{lower value}}{2}$

Single Population Mean, Known Standard Deviation, Normal Distribution

Use the Normal Distribution for Means, Population Standard Deviation is Known $EBM = z\frac{\alpha}{2} \cdot \frac{\sigma}{\sqrt{n}}$

The confidence interval has the format $(\overline{x} - EBM, \overline{x} + EBM)$.

Use the following information to answer the next five practice exercises:

7.2 Chapter Review

In many cases, the researcher does not know the population standard deviation, σ, of the measure being studied. In these cases, it is common to use the sample standard deviation, s, as an estimate of σ. The **normal distribution** creates accurate confidence intervals when σ is known, but it is not as accurate when s is used

as an estimate. In this case, the Student's t-distribution is much better. Define a t-score using the following formula:

$$t = \frac{\overline{x} - \mu}{\left(\frac{s}{\sqrt{n}}\right)},$$

The *t*-score follows the Student's t-distribution with $n - 1$ degrees of freedom. The confidence interval under this distribution is calculated with $EBM = \left(t_{\frac{\alpha}{2}}\right)\frac{s}{\sqrt{n}}$ where $t_{\frac{\alpha}{2}}$ is the *t*-score with area to the right equal to $\frac{\alpha}{2}$, s is the sample standard deviation, and n is the sample size. Use a table, calculator, or computer to find $t_{\frac{\alpha}{2}}$ for a given α.

Formula Review

s = the standard deviation of sample values.

$t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}$ is the formula for the *t*-score which measures how far away a measure is from the population mean in the Student's t-distribution

$df = n - 1$; the degrees of freedom for a **Student's t-distribution** where n represents the size of the sample

$T \sim t_{df}$ the random variable, T, has a **Student's t-distribution** with df degrees of freedom

$EBM = t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}$ = the error bound for the population mean when the population standard deviation is unknown

$t_{\frac{\alpha}{2}}$ is the *t*-score in the Student's t-distribution with area to the right equal to $\frac{\alpha}{2}$

The general form for a confidence interval for a single mean, population standard deviation unknown, Student's t is given by (lower bound, upper bound)

= (point estimate – *EBM*, point estimate + *EBM*)

$$= \left(\overline{x} - \frac{ts}{\sqrt{n}}, \overline{x} + \frac{ts}{\sqrt{n}}\right)$$

Use the following information to answer the next five practice exercises.

7.3 Chapter Review

Some statistical measures, like many survey questions, measure qualitative rather than quantitative data. In this case, the population parameter being estimated is a proportion. It is possible to create a confidence interval for the true population proportion following procedures similar to those used in creating confidence intervals for population means. The formulas are slightly different, but they follow the same reasoning.

Let p' represent the sample proportion, x/n, where x represents the number of successes and n represents the sample size. Let $q' = 1 - p'$. Then the confidence interval for a population proportion is given by the following formula:

(lowerbound,upperbound)
=\left({p}^{\prime}−EBP,{p}^{\prime}+EBP\right)= \left({p}^{\prime}−z\
sqrt{\frac{{p}^{\prime}{q}^{\prime}}{n}},{p}^{\prime}+z\
sqrt{\frac{{p}^{\prime}{q}^{\prime}}{n}}\right)

The "plus four" method for calculating confidence intervals is an attempt to balance the error introduced by using estimates of the population proportion when calculating the standard deviation of the sampling distribution. Simply imagine four additional trials in the study; two are successes and two are failures. Calculate $p' = \frac{x+2}{n+4}$, and proceed to find the confidence interval. When sample sizes are small, this method has been demonstrated to provide more accurate confidence intervals than the standard formula used for larger samples.

Formula Review

$p' = x / n$ where x represents the number of successes and n represents the sample size. The variable p' is the sample proportion and serves as the point estimate for the true population proportion.

$q' = 1 - p'$

$p' \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$ The variable p' has a **binomial distribution** that can be approximated with the normal distribution shown here.

EBP = the error bound for a proportion = $z_{\frac{\alpha}{2}}\sqrt{\frac{p'q'}{n}}$

Confidence interval for a proportion:

(lower bound, upper bound) $= (p' - EBP, p' + EBP) = \left(p' - z\sqrt{\frac{p'q'}{n}}, p' + z\sqrt{\frac{p'q'}{n}}\right)$

$n = \frac{z_{\frac{\alpha}{2}}{}^2 p' q'}{EBP^2}$ provides the number of participants needed to estimate the population proportion with confidence $1 - \alpha$ and margin of error *EBP*.

Use the normal distribution for a single population proportion $p' = \frac{x}{n}$

$$EBP = \left(z_{\frac{\alpha}{2}}\right) \sqrt{\frac{p'q'}{n}} p' + q' = 1$$

The confidence interval has the format $(p' - EBP, p' + EBP)$.

$\overline{x}$ is a point estimate for μ

p' is a point estimate for ρ

s is a point estimate for σ

Use the following information to answer the next two practice exercises:

CHAPTER 7 PRACTICE

7.1 Practice

The standard deviation of the weights of elephants is known to be approximately 15 pounds. We wish to construct a 95% confidence interval for the mean weight of newborn elephant calves. Fifty newborn elephants are weighed. The sample mean is 244 pounds. The sample standard deviation is 11 pounds.

Identify the following:

1. $\overline{x}$ = _____
2. σ = _____
3. n = _____

Solution

1. 244
2. 15
3. 50

In words, define the random variables X and $\overline{X}$.

Which distribution should you use for this problem?
Solution

$$N\left(244, \frac{15}{\sqrt{50}}\right)$$

Construct a 95% confidence interval for the population mean weight of newborn elephants. State the confidence interval, sketch the graph, and calculate the error bound.

What will happen to the confidence interval obtained, if 500 newborn elephants are weighed instead of 50? Why?
Solution

As the sample size increases, there will be less variability in the mean, so the interval size decreases.

Use the following information to answer the next seven exercises: The U.S. Census Bureau conducts a study to determine the time needed to complete the short form. The Bureau surveys 200 people. The sample mean is 8.2 minutes. There is a known standard deviation of 2.2 minutes. The population distribution is assumed to be normal.

Identify the following:

1. $\overline{x}$ = _____
2. σ = _____
3. n = _____

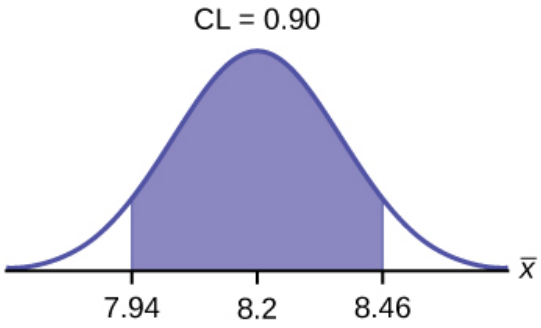
In words, define the random variables X and $\overline{X}$.
Solution

X is the time in minutes it takes to complete the U.S. Census short form. $\overline{X}$ is the mean time it took a sample of 200 people to complete the U.S. Census short form.

Which distribution should you use for this problem?

Construct a 90% confidence interval for the population mean time to complete the forms. State the confidence interval, sketch the graph, and calculate the error bound.
Solution

CI: (7.9441, 8.4559)



EBM = 0.26

If the Census wants to increase its level of confidence and keep the error bound the same by taking another survey, what changes should it make?

If the Census did another survey, kept the error bound the same, and surveyed only 50 people instead of 200, what would happen to the level of confidence? Why?
Solution

The level of confidence would decrease because decreasing n makes the confidence interval wider, so at the same error bound, the confidence level decreases.

Suppose the Census needed to be 98% confident of the population mean length of time. Would the Census have to survey more people? Why or why not?

Use the following information to answer the next ten exercises: A sample of 20 heads of lettuce was selected. Assume that the population distribution of head weight is normal. The weight of each head of lettuce was then recorded. The mean weight was 2.2 pounds with a standard deviation of 0.1 pounds. The population standard deviation is known to be 0.2 pounds.

Identify the following:

1. $\overline{x}$ = _____
2. σ = _____
3. n = _____

Solution

1. $\overline{x} = 2.2$
2. $\sigma = 0.2$
3. $n = 20$

In words, define the random variable X.

In words, define the random variable $\overline{X}$.
Solution

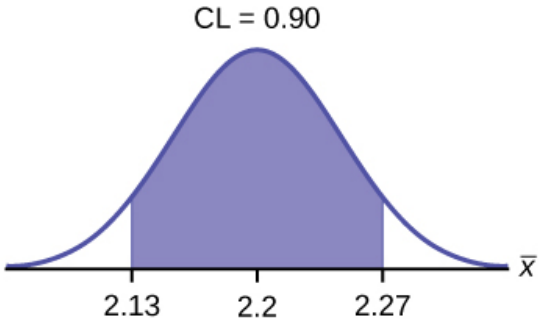
$\overline{X}$ is the mean weight of a sample of 20 heads of lettuce.

Which distribution should you use for this problem?

Construct a 90% confidence interval for the population mean weight of the heads of lettuce. State the confidence interval, sketch the graph, and calculate the error bound.
Solution

$EBM = 0.07$

CI: (2.1264, 2.2736)



Construct a 95% confidence interval for the population mean weight of the heads of lettuce. State the confidence interval, sketch the graph, and calculate the error bound.

In complete sentences, explain why the confidence interval in [link] is larger than in [link].
Solution

The interval is greater because the level of confidence increased. If the only change made in the analysis is a change in confidence level, then all we are doing is changing how much area is being calculated for the normal distribution. Therefore, a larger confidence level results in larger areas and larger intervals.

In complete sentences, give an interpretation of what the interval in [link] means.

What would happen if 40 heads of lettuce were sampled instead of 20, and the error bound remained the same?
Solution

The confidence level would increase.

What would happen if 40 heads of lettuce were sampled instead of 20, and the confidence level remained the same?

Use the following information to answer the next 14 exercises: The mean age for all Foothill College students for a recent Fall term was 33.2. The population standard deviation has been pretty consistent at 15. Suppose that twenty-five Winter students were randomly selected. The mean age for the sample was 30.4. We are interested in the true mean age for Winter Foothill College students. Let X = the age of a Winter Foothill College student.

$\overline{x}$ = _____
Solution

30.4

n = _____

_____ = 15
Solution

σ

In words, define the random variable $\overline{X}$.

What is $\overline{x}$ estimating?
Solution

μ

Is σ_x known?

As a result of your answer to [link], state the exact distribution to use when calculating the confidence interval.
Solution

normal

Construct a 95% Confidence Interval for the true mean age of Winter Foothill College students by working out then answering the next seven exercises.

How much area is in both tails (combined)? α =_____

How much area is in each tail? $\frac{\alpha}{2}$ =_____
Solution

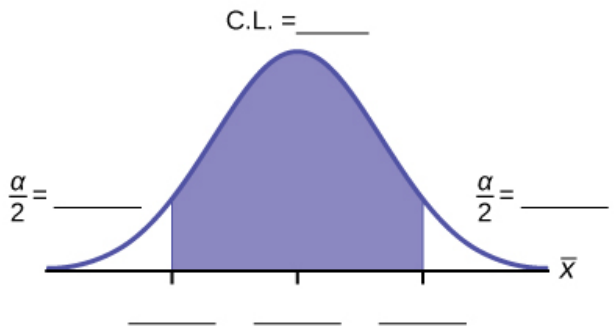
0.025

Identify the following specifications:
lower limit
upper limit
error bound

The 95% confidence interval is:_____.
Solution

(24.52,36.28)

Fill in the blanks on the graph with the areas, upper and lower limits of the confidence interval, and the sample mean.



In one complete sentence, explain what the interval means.
Solution

We are 95% confident that the true mean age for Winger Foothill College students is between 24.52 and 36.28.

Using the same mean, standard deviation, and level of confidence, suppose that n were 69 instead of 25. Would the error bound become larger or smaller? How do you know?

Using the same mean, standard deviation, and sample size, how would the error bound change if the confidence level were reduced to 90%? Why?
Solution

The error bound for the mean would decrease because as the CL decreases, you need less area under the normal curve (which translates into a smaller interval) to capture the true population mean.

7.2 Practice

A hospital is trying to cut down on emergency room wait times. It is interested in the amount of time patients must wait before being called back to be examined. An investigation committee randomly surveyed 70 patients. The sample mean was 1.5 hours with a sample standard deviation of 0.5 hours.

Identify the following:

1. $\overline{x}$ =_____
2. s_x =_____
3. n =_____
4. $n - 1$ =_____

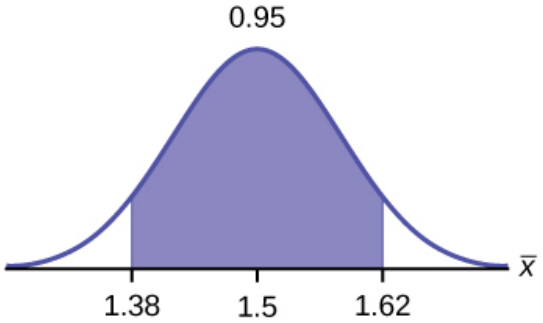
Define the random variables X and $\overline{X}$ in words.
Solution

X is the number of hours a patient waits in the emergency room before being called back to be examined. $\overline{X}$ is the mean wait time of 70 patients in the emergency room.

Which distribution should you use for this problem?

Construct a 95% confidence interval for the population mean time spent waiting. State the confidence interval, sketch the graph, and calculate the error bound.
Solution

CI: (1.3808, 1.6192)



$EBM = 0.12$

Explain in complete sentences what the confidence interval means.

Use the following information to answer the next six exercises: One hundred eight Americans were surveyed to determine the number of hours they spend watching television each month. It was revealed that they watched an average of 151 hours each month with a standard deviation of 32 hours. Assume that the underlying population distribution is normal.

Identify the following:

1. $\overline{x}$ =_____
2. s_x =_____
3. n =_____
4. $n - 1$ =_____

Solution

1. $\overline{x} = 151$
2. $s_x = 32$
3. $n = 108$
4. $n - 1 = 107$

Define the random variable X in words.

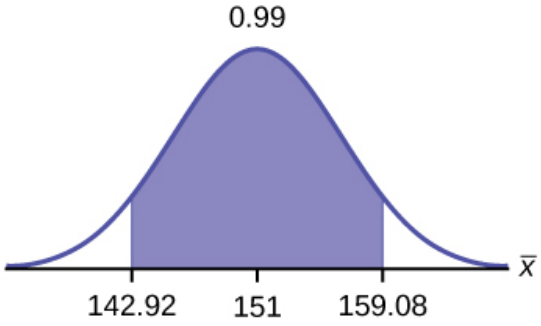
Define the random variable $\overline{X}$ in words.
Solution

$\overline{X}$ is the mean number of hours spent watching television per month from a sample of 108 Americans.

Which distribution should you use for this problem?

Construct a 99% confidence interval for the population mean hours spent watching television per month. (a) State the confidence interval, (b) sketch the graph, and (c) calculate the error bound.
Solution

CI: (142.92, 159.08)



$EBM = 8.08$

Why would the error bound change if the confidence level were lowered to 95%?

Use the following information to answer the next 13 exercises: The data in [link] are the result of a random survey of 39 national flags (with replacement between picks) from various countries. We are interested in finding a confidence interval for the true mean number of colors on a national flag. Let X = the number of colors on a national flag.

X	Freq.
1	1
2	7
3	18
4	7
5	6

Calculate the following:

1. $\overline{x}$ =_____
2. s_x =_____
3. n =_____

Solution

1. 3.26
2. 1.02
3. 39

Define the random variable $\overline{X}$ in words.

What is $\overline{x}$ estimating?
Solution

μ

Is σ_x known?

As a result of your answer to [link], state the exact distribution to use when calculating the confidence interval.
Solution

t_{38}

Construct a 95% confidence interval for the true mean number of colors on national flags.

How much area is in both tails (combined)?

How much area is in each tail?

Solution

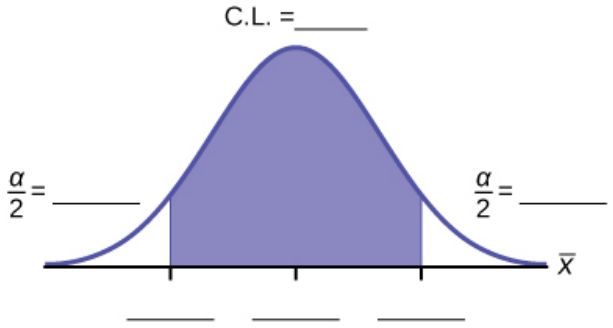
0.025

Calculate the following:
lower limit
upper limit
error bound

The 95% confidence interval is_____.
Solution

(2.93, 3.59)

Fill in the blanks on the graph with the areas, the upper and lower limits of the Confidence Interval and the sample mean.



In one complete sentence, explain what the interval means.
Solution

We are 95% confident that the true mean number of colors for national flags is between 2.93 colors and 3.59 colors.

Using the same $\overline{x}$, s_x, and level of confidence, suppose that n were 69 instead of 39. Would the error bound become larger or smaller? How do you know?
Solution

The error bound would become EBM = 0.245. This error bound decreases because as sample sizes increase, variability decreases and we need less interval length to capture the true mean.

Using the same $\overline{x}$, s_x, and n = 39, how would the error bound change if the confidence level were reduced to 90%? Why?

7.3 Practice

Marketing companies are interested in knowing the population percent of women who make the majority of household purchasing decisions.

When designing a study to determine this population proportion, what is the minimum number you would need to survey to be 90% confident that the population proportion is estimated to within 0.05?

If it were later determined that it was important to be more than 90% confident and a new survey were commissioned, how would it affect the minimum number you need to survey? Why?
Solution

It would decrease, because the z-score would decrease, which would reduce the numerator and lower the number.

Use the following information to answer the next five exercises: Suppose the marketing company did do a survey. They randomly surveyed 200 households and found that in 120 of them, the woman made the majority of the purchasing decisions. We are interested in the population proportion of households where women make the majority of the purchasing decisions.

Identify the following:

1. $x =$ _____
2. $n =$ _____
3. $p' =$ _____

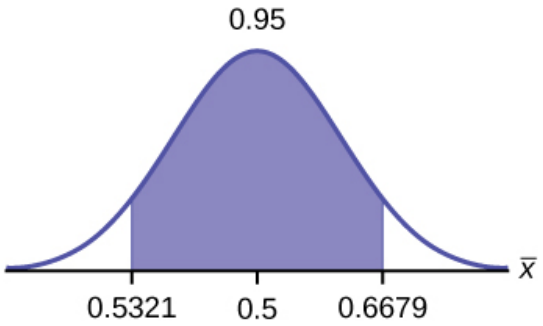
Define the random variables X and P' in words.
Solution

X is the number of "successes" where the woman makes the majority of the purchasing decisions for the household. P' is the percentage of households sampled where the woman makes the majority of the purchasing decisions for the household.

Which distribution should you use for this problem?

Construct a 95% confidence interval for the population proportion of households where the women make the majority of the purchasing decisions. State the confidence interval, sketch the graph, and calculate the error bound.
Solution

CI: (0.5321, 0.6679)



EBM: 0.0679

List two difficulties the company might have in obtaining random results, if this survey were done by email.

Use the following information to answer the next five exercises: Of 1,050 randomly selected adults, 360 identified themselves as manual laborers, 280 identified themselves as non-manual wage earners, 250 identified themselves as mid-level managers, and 160 identified themselves as executives. In the survey, 82% of manual laborers preferred trucks, 62% of non-manual wage earners preferred trucks, 54% of mid-level managers preferred trucks, and 26% of executives preferred trucks.

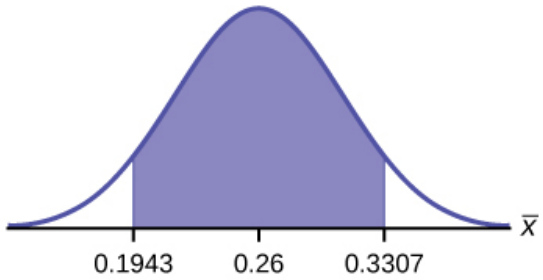
We are interested in finding the 95% confidence interval for the percent of executives who prefer trucks. Define random variables X and P' in words.
Solution

X is the number of "successes" where an executive prefers a truck. P' is the percentage of executives sampled who prefer a truck.

Which distribution should you use for this problem?

Construct a 95% confidence interval. State the confidence interval, sketch the graph, and calculate the error bound.
Solution

CI: (0.19432, 0.33068)



EBM: 0.0707

Suppose we want to lower the sampling error. What is one way to accomplish that?

The sampling error given in the survey is ±2%. Explain what the ±2% means.
Solution

The sampling error means that the true mean can be 2% above or below the sample mean.

Use the following information to answer the next five exercises: A poll of 1,200 voters asked what the most significant issue was in the upcoming election. Sixty-five percent answered the economy. We are interested in the population proportion of voters who feel the economy is the most important.

Define the random variable X in words.

Define the random variable P' in words.
Solution

P' is the proportion of voters sampled who said the economy is the most important issue in the upcoming election.

Which distribution should you use for this problem?

Construct a 90% confidence interval, and state the confidence interval and the error bound.
Solution

CI: (0.62735, 0.67265)
EBM: 0.02265

What would happen to the confidence interval if the level of confidence were 95%?

Use the following information to answer the next 16 exercises: The Ice Chalet offers dozens of different beginning ice-skating classes. All of the class names are put into a bucket. The 5 P.M., Monday night, ages 8 to 12, beginning ice-skating class was picked. In that class were 64 girls and 16 boys. Suppose that we are interested in the true proportion of girls, ages 8 to 12, in all beginning ice-skating classes at the Ice Chalet. Assume that the children in the selected class are a random sample of the population.

What is being counted?
Solution

The number of girls, ages 8 to 12, in the 5 P.M. Monday night beginning ice-skating class.

In words, define the random variable X.

Calculate the following:
x = _____
n = _____
p' = _____
Solution

1. $x = 64$
2. $n = 80$
3. $p' = 0.8$

State the estimated distribution of X. X~_____

Define a new random variable P'. What is p' estimating?
Solution

p

In words, define the random variable P'.

State the estimated distribution of P'. Construct a 92% Confidence Interval for the true proportion of girls in the ages 8 to 12 beginning ice-skating classes at the Ice Chalet.
Solution

$$P' \ N\left(0.8, \sqrt{\frac{(0.8)\,(0.2)}{80}}\right). (0.72171, 0.87829).$$

How much area is in both tails (combined)?

How much area is in each tail?
Solution

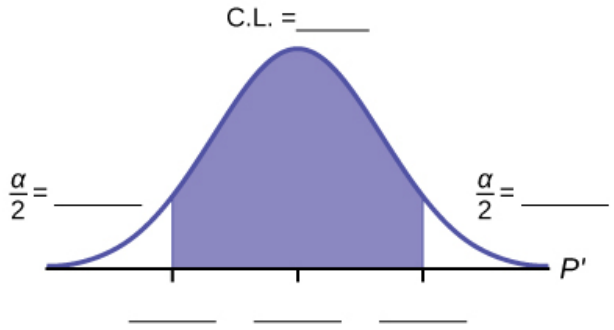
0.04

Calculate the following:
lower limit
upper limit
error bound

The 92% confidence interval is _____.
Solution

(0.72; 0.88)

Fill in the blanks on the graph with the areas, upper and lower limits of the confidence interval, and the sample proportion.



In one complete sentence, explain what the interval means.
Solution

With 92% confidence, we estimate the proportion of girls, ages 8 to 12, in a beginning ice-skating class at the Ice Chalet to be between 72% and 88%.

Using the same p' and level of confidence, suppose that n was increased to 100. Would the error bound become larger or smaller? How do you know?

Using the same p' and $n = 80$, how would the error bound change if the confidence level were increased to 98%? Why?
Solution

The error bound would increase. Assuming all other variables are kept constant, as the confidence level increases, the area under the curve corresponding to the confidence level becomes larger, which creates a wider interval and thus a larger error.

If you decrease the allowable error bound, why would the minimum sample size increase (keeping the same level of confidence)?

CHAPTER 7 HOMEWORK

7.1 Homework

Among various ethnic groups, the standard deviation of heights is known to be approximately three inches. We wish to construct a 95% confidence interval for the mean height of male Swedes. Forty-eight male Swedes are surveyed. The sample mean is 71 inches. The sample standard deviation is 2.8 inches.

$\overline{x}$ =_____

σ =_____

n =_____

In words, define the random variables X and $\overline{X}$.

Which distribution should you use for this problem? Explain your choice.

Construct a 95% confidence interval for the population mean height of male Swedes.

State the confidence interval.

Sketch the graph.

Calculate the error bound.

What will happen to the level of confidence obtained if 1,000 male Swedes are surveyed instead of 48? Why?

Solution

71

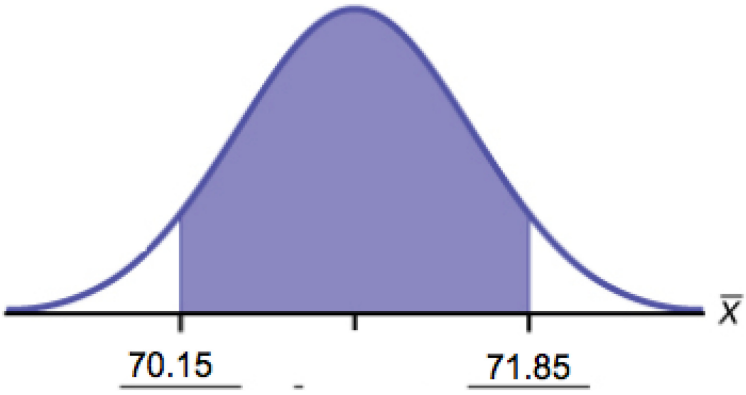
3

48

X is the height of a Swiss male and is the mean height from a sample of 48 Swiss males.

Normal. We know the standard deviation for the population, and the sample size is greater than 30.

CI: (70.151, 71.49)



EBM = 0.849

The confidence interval will decrease in size, because the sample size increased. Recall, when all factors remain unchanged, an increase in sample size decreases variability. Thus, we do not need as large an interval to capture the true population mean.

Announcements for 84 upcoming engineering conferences were randomly picked from a stack of IEEE Spectrum magazines. The mean length of the conferences was 3.94 days, with a standard deviation of 1.28 days. Assume the underlying population is normal.

1. In words, define the random variables X and $\overline{X}$.
2. Which distribution should you use for this problem? Explain your choice.
3. Construct a 95% confidence interval for the population mean length of engineering conferences.
    1. State the confidence interval.
    2. Sketch the graph.
    3. Calculate the error bound.

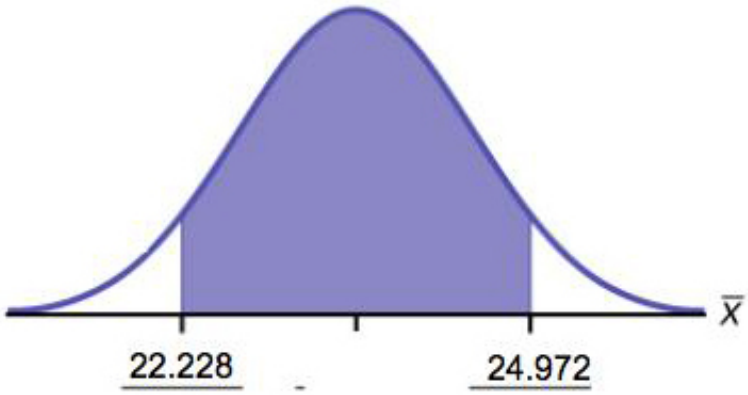
Suppose that an accounting firm does a study to determine the time needed to complete one person's tax forms. It randomly surveys 100 people. The sample mean is 23.6 hours. There is a known standard deviation of 7.0 hours. The population distribution is assumed to be normal.

    1. $\overline{x}$ =_____
    2. σ =_____
    3. n =_____
2. In words, define the random variables X and $\overline{X}$.
3. Which distribution should you use for this problem? Explain your choice.
4. Construct a 90% confidence interval for the population mean time to complete the tax forms.
    1. State the confidence interval.
    2. Sketch the graph.
    3. Calculate the error bound.
5. If the firm wished to increase its level of confidence and keep the error bound the same by taking another survey, what changes should it make?
6. If the firm did another survey, kept the error bound the same, and only surveyed 49 people, what would happen to the level of confidence? Why?
7. Suppose that the firm decided that it needed to be at least 96% confident of the population mean length of time to within one hour. How would the number of people the firm surveys change? Why?

Solution

1. $\overline{x} = 23.6$
2. $\sigma = 7$
3. n = 100

2. X is the time needed to complete an individual tax form. $\overline{X}$ is the mean time to complete tax forms from a sample of 100 customers.
3. $N\left(23.6, \frac{7}{\sqrt{100}}\right)$ because we know sigma.
    1. (22.228, 24.972)



    2.
    3. $EBM = 1.372$
4. It will need to change the sample size. The firm needs to determine what the confidence level should be, then apply the error bound formula to determine the necessary sample size.
5. The confidence level would increase as a result of a larger interval. Smaller sample sizes result in more variability. To capture the true population mean, we need to have a larger interval.
6. According to the error bound formula, the firm needs to survey 206 people. Since we increase the confidence level, we need to increase either our error bound or the sample size.

A sample of 16 small bags of the same brand of candies was selected. Assume that the population distribution of bag weights is normal. The weight of each bag was then recorded. The mean weight was two ounces with a standard deviation of 0.12 ounces. The population standard deviation is known to be 0.1 ounce.

1. $\overline{x}$ =_____
2. σ =_____
3. s_x =_____

2. In words, define the random variable X.
3. In words, define the random variable $\overline{X}$.
4. Which distribution should you use for this problem? Explain your choice.
5. Construct a 90% confidence interval for the population mean weight of the candies.
    1. State the confidence interval.

2. Sketch the graph.
3. Calculate the error bound.

6. Construct a 98% confidence interval for the population mean weight of the candies.
    1. State the confidence interval.
    2. Sketch the graph.
    3. Calculate the error bound.
7. In complete sentences, explain why the confidence interval in part f is larger than the confidence interval in part e.
8. In complete sentences, give an interpretation of what the interval in part f means.

A camp director is interested in the mean number of letters each child sends during his or her camp session. The population standard deviation is known to be 2.5. A survey of 20 campers is taken. The mean from the sample is 7.9 with a sample standard deviation of 2.8.

1. $\overline{x}$ =_____
2. σ =_____
3. n =_____

2. Define the random variables X and $\overline{X}$ in words.
3. Which distribution should you use for this problem? Explain your choice.
4. Construct a 90% confidence interval for the population mean number of letters campers send home.
    1. State the confidence interval.
    2. Sketch the graph.
    3. Calculate the error bound.
5. What will happen to the error bound and confidence interval if 500 campers are surveyed? Why?

Solution

7.9

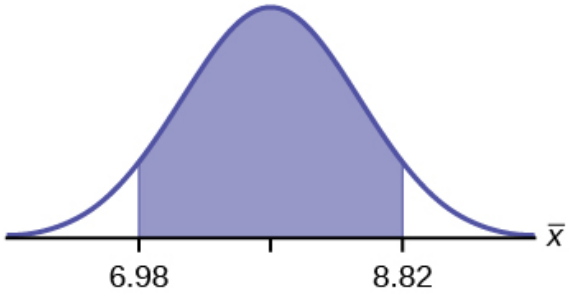
2.5

20

X is the number of letters a single camper will send home. $\overline{X}$ is the mean number of letters sent home from a sample of 20 campers.

$N7.9\left(\frac{2.5}{\sqrt{20}}\right)$

CI: (6.98, 8.82)



EBM: 0.92
The error bound and confidence interval will decrease.

What is meant by the term "90% confident" when constructing a confidence interval for a mean?

1. If we took repeated samples, approximately 90% of the samples would produce the same confidence interval.
2. If we took repeated samples, approximately 90% of the confidence intervals calculated from those samples would contain the sample mean.
3. If we took repeated samples, approximately 90% of the confidence intervals calculated from those samples would contain the true value of the population mean.
4. If we took repeated samples, the sample mean would equal the population mean in approximately 90% of the samples.

The Federal Election Commission collects information about campaign contributions and disbursements for candidates and political committees each election cycle. During the 2012 campaign season, there were 1,619 candidates for the House of Representatives across the United States who received contributions from individuals. [link] shows the total receipts from individuals for a random selection of 40 House candidates rounded to the nearest $ 100. The standard deviation for this data to the nearest hundred is $\sigma =$ $ 909,200.

$ 3,600	$ 1,243,900	$ 10,900	$ 385,200	$ 581,500
$ 7,400	$ 2,900	$ 400	$ 3,714,500	$ 632,500
$ 391,000	$ 467,400	$ 56,800	$ 5,800	$ 405,200
$ 733,200	$ 8,000	$ 468,700	$ 75,200	$ 41,000
$ 13,300	$ 9,500	$ 953,800	$ 1,113,500	$ 1,109,300
$ 353,900	$ 986,100	$ 88,600	$ 378,200	$ 13,200
$ 3,800	$ 745,100	$ 5,800	$ 3,072,100	$ 1,626,700
$ 512,900	$ 2,309,200	$ 6,600	$ 202,400	$ 15,800

1. Find the point estimate for the population mean.
2. Using 95% confidence, calculate the error bound.
3. Create a 95% confidence interval for the mean total individual contributions.
4. Interpret the confidence interval in the context of the problem.

Solution

1. $\overline{x} =$ $568,873
2. $CL = 0.95$ $\alpha = 1 - 0.95 = 0.05$ $z_{\frac{\alpha}{2}} = 1.96$

   $$EBM = z_{0.025}\frac{\sigma}{\sqrt{n}} = 1.96\frac{909200}{\sqrt{40}} =$$ $281,764
3. $\overline{x} - EBM = 568,873 - 281,764 = 287,109$

   $\overline{x} + EBM = 568,873 + 281,764 = 850,637$

   Alternate solution:
   1. Press `STAT` and arrow over to `TESTS`.
   2. Arrow down to `7:ZInterval`.
   3. Press `ENTER`.
   4. Arrow to Stats and press `ENTER`.
   5. Arrow down and enter the following values:
      - σ: 909,200
      - $\overline{x}$: 568,873
      - n: 40
      - CL: 0.95
   6. Arrow down to Calculate and press `ENTER`.
   7. The confidence interval is ($287,114, $850,632).
   8. Notice the small difference between the two solutions—these differences are simply due to rounding errors in the hand calculations.
4. We estimate with 95% confidence that the mean amount of contributions received from all individuals by House candidates is between $287,109 and $850,637.

The American Community Survey (ACS), part of the United States Census Bureau, conducts a yearly census similar to the one taken every ten years, but with a smaller percentage of participants. The most recent survey estimates with 90% confidence that the mean household income in the U.S. falls between $69,720 and

$ 69,922. Find the point estimate for mean U.S. household income and the error bound for mean U.S. household income.

The average height of young adult males has a normal distribution with standard deviation of 2.5 inches. You want to estimate the mean height of students at your college or university to within one inch with 93% confidence. How many male students must you measure?
Solution

Use the formula for *EBM*, solved for *n*:

$$n = \left(\frac{z\sigma}{EBM}\right)^2$$

$$n = \frac{z^2\sigma^2}{EBM^2}$$

From the statement of the problem, you know that $\sigma = 2.5$, and you need $EBM = 1$.

$z = z_{0.035} = 1.812$

(This is the value of z for which the area under the density curve to the ***right*** of z is 0.035.)

$$n = \frac{z^2\sigma^2}{EBM^2} = \frac{1.812^2 2.5^2}{1^2} \approx 20.52$$

You need to measure at least 21 male students to achieve your goal.

7.2 Homework

In six packages of "The Flintstones® Real Fruit Snacks" there were five Bam-Bam snack pieces. The total number of snack pieces in the six bags was 68. We wish to calculate a 96% confidence interval for the population proportion of Bam-Bam snack pieces.

1. Define the random variables X and P' in words.
2. Which distribution should you use for this problem? Explain your choice
3. Calculate p'.
4. Construct a 96% confidence interval for the population proportion of Bam-Bam snack pieces per bag.
    1. State the confidence interval.
    2. Sketch the graph.
    3. Calculate the error bound.
5. Do you think that six packages of fruit snacks yield enough data to give accurate results? Why or why not?

A random survey of enrollment at 35 community colleges across the United States yielded the following figures: 6,414; 1,550; 2,109; 9,350; 21,828; 4,300; 5,944; 5,722; 2,825; 2,044; 5,481; 5,200; 5,853; 2,750; 10,012; 6,357; 27,000; 9,414; 7,681; 3,200; 17,500; 9,200; 7,380; 18,314; 6,557; 13,713; 17,768; 7,493; 2,771; 2,861; 1,263; 7,285; 28,165; 5,080; 11,622. Assume the underlying population is normal.

$\overline{x}$ = _____

s_x = _____

n = _____

$n - 1$ = _____

Define the random variables X and $\overline{X}$ in words.

Which distribution should you use for this problem? Explain your choice.

Construct a 95% confidence interval for the population mean enrollment at community colleges in the United States.

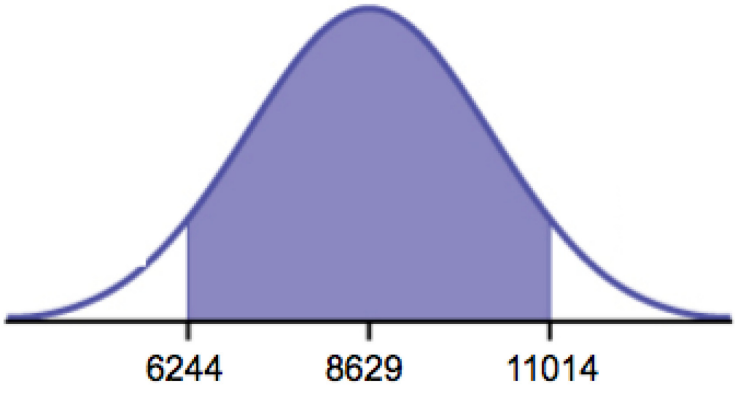
State the confidence interval.

Sketch the graph.

Calculate the error bound.

What will happen to the error bound and confidence interval if 500 community colleges were surveyed? Why?

Solution

1.
    1. 8629
    2. 6944
    3. 35
    4. 34
2. t_{34}
    1. CI: (6244, 11,014)
    2. 

    3. EB = 2385
3. It will become smaller

Suppose that a committee is studying whether or not there is waste of time in our judicial system. It is interested in the mean amount of time individuals waste at the courthouse waiting to be called for jury duty. The committee randomly surveyed 81 people who recently served as jurors. The sample mean wait time was eight hours with a sample standard deviation of four hours.

1. $\overline{x}$ = _____
2. s_x = _____
3. n = _____
4. $n - 1$ = _____

2. Define the random variables X and $\overline{X}$ in words.
3. Which distribution should you use for this problem? Explain your choice.
4. Construct a 95% confidence interval for the population mean time wasted.
    1. State the confidence interval.
    2. Sketch the graph.
    3. Calculate the error bound.
5. Explain in a complete sentence what the confidence interval means.

A pharmaceutical company makes tranquilizers. It is assumed that the distribution for the length of time they last is approximately normal. Researchers in a hospital used the drug on a random sample of nine patients. The effective period of the tranquilizer for each patient (in hours) was as follows: 2.7; 2.8; 3.0; 2.3; 2.3; 2.2; 2.8; 2.1; and 2.4.

1. $\overline{x}$ = _____
2. s_x = _____
3. n = _____
4. $n - 1$ = _____

2. Define the random variable X in words.
3. Define the random variable $\overline{X}$ in words.
4. Which distribution should you use for this problem? Explain your choice.
5. Construct a 95% confidence interval for the population mean length of time.
    1. State the confidence interval.
    2. Sketch the graph.
    3. Calculate the error bound.
6. What does it mean to be "95% confident" in this problem?

Solution

    1. $\overline{x} = 2.51$
    2. $s_x = 0.318$
    3. $n = 9$
    4. $n - 1 = 8$
2. the effective length of time for a tranquilizer
3. the mean effective length of time of tranquilizers from a sample of nine patients
4. We need to use a Student's-t distribution, because we do not know the population standard deviation.
    1. CI: (2.27, 2.76)
    2. Check student's solution.
    3. *EBM*: 0.25
5. If we were to sample many groups of nine patients, 95% of the samples would contain the true population mean length of time.

Suppose that 14 children, who were learning to ride two-wheel bikes, were surveyed to determine how long they had to use training wheels. It was revealed that they used them an average of six months with a sample standard deviation of three months. Assume that the underlying population distribution is normal.

    1. $\overline{x} =$ _____
    2. $s_x =$ _____
    3. $n =$ _____
    4. $n - 1 =$ _____
2. Define the random variable X in words.
3. Define the random variable $\overline{X}$ in words.
4. Which distribution should you use for this problem? Explain your choice.
5. Construct a 99% confidence interval for the population mean length of time using training wheels.
    1. State the confidence interval.
    2. Sketch the graph.
    3. Calculate the error bound.
6. Why would the error bound change if the confidence level were lowered to 90%?

The Federal Election Commission (FEC) collects information about campaign contributions and disbursements for candidates and political committees each election cycle. A political action committee (PAC) is a committee formed to raise money for candidates and campaigns. A Leadership PAC is a PAC formed by a federal politician (senator or representative) to raise money to help other candidates' campaigns.

The FEC has reported financial information for 556 Leadership PACs that operated during the 2011–2012 election cycle. The following table shows the total receipts during this cycle for a random selection of 20 Leadership PACs.

$ 46,500.00	$ 0	$ 40,966.50	$ 105,887.20	$ 5,175.00
$ 29,050.00	$ 19,500.00	$ 181,557.20	$ 31,500.00	$ 149,970.80
$ 2,555,363.20	$ 12,025.00	$ 409,000.00	$ 60,521.70	$ 18,000.00
$ 61,810.20	$ 76,530.80	$ 119,459.20	$ 0	$ 63,520.00
$ 6,500.00	$ 502,578.00	$ 705,061.10	$ 708,258.90	$ 135,810.00
$ 2,000.00	$ 2,000.00	$ 0	$ 1,287,933.80	$ 219,148.30

$\overline{x} = \$251,854.23$

$s = \$521,130.41$

Use this sample data to construct a 96% confidence interval for the mean amount of money raised by all Leadership PACs during the 2011–2012 election cycle. Use the Student's t-distribution.

Solution

$\overline{x} = \$251,854.23$

$s = \$521,130.41$

Note that we are not given the population standard deviation, only the standard deviation of the sample.

There are 30 measures in the sample, so $n = 30$, and $df = 30 - 1 = 29$

$CL = 0.96$, so $\alpha = 1 - CL = 1 - 0.96 = 0.04$

$\frac{\alpha}{2} = 0.02 t_{\frac{\alpha}{2}} = t_{0.02}$

= 2.150

$$EBM = t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right) = 2.150 \left(\frac{521,130.41}{\sqrt{30}} \right) \approx \$204,561.66$$

$\overline{x} - EBM$ = \$251,854.23 – \$204,561.66 = \$47,292.57

$\overline{x} + EBM$ = \$251,854.23+ \$204,561.66 = \$456,415.89

We estimate with 96% confidence that the mean amount of money raised by all Leadership PACs during the 2011–2012 election cycle lies between \$47,292.57 and \$456,415.89.

Alternate Solution

Enter the data as a list.

Press `STAT` and arrow over to `TESTS`.

Arrow down to `8:TInterval`.

Press `ENTER`.

Arrow to Data and press `ENTER`.

Arrow down and enter the name of the list where the data is stored.

Enter `Freq`: 1

Enter `C-Level`: 0.96

Arrow down to `Calculate` and press `Enter`.

The 96% confidence interval is ($47,262, $456,447).

The difference between solutions arises from rounding differences.

Forbes magazine published data on the best small firms in 2012. These were firms that had been publicly traded for at least a year, have a stock price of at least $5 per share, and have reported annual revenue between $5 million and $1 billion. The [link] shows the ages of the corporate CEOs for a random sample of these firms.

48	58	51	61	56
59	74	63	53	50
59	60	60	57	46
55	63	57	47	55
57	43	61	62	49
67	67	55	55	49

Use this sample data to construct a 90% confidence interval for the mean age of CEO's for these top small firms. Use the **Student's t-distribution**.

Unoccupied seats on flights cause airlines to lose revenue. Suppose a large airline wants to estimate its mean number of unoccupied seats per flight over the past year. To accomplish this, the records of 225 flights are randomly selected and the number of unoccupied seats is noted for each of the sampled flights. The sample mean is 11.6 seats and the sample standard deviation is 4.1 seats.

1. $\overline{x}$ = _____
2. s_x = _____
3. n = _____
4. n-1 = _____

2. Define the random variables X and $\overline{X}$ in words.
3. Which distribution should you use for this problem? Explain your choice.
4. Construct a 92% confidence interval for the population mean number of unoccupied seats per flight.

1. State the confidence interval.
2. Sketch the graph.
3. Calculate the error bound.

Solution

1. $\overline{x}$ =
2. s_x =
3. n =
4. $n - 1$ =

2. X is the number of unoccupied seats on a single flight. $\overline{X}$ is the mean number of unoccupied seats from a sample of 225 flights.
3. We will use a **Student's-t distribution**, because we do not know the population standard deviation.
    1. CI: (11.12 , 12.08)
    2. Check student's solution.
    3. *EBM*: 0.48

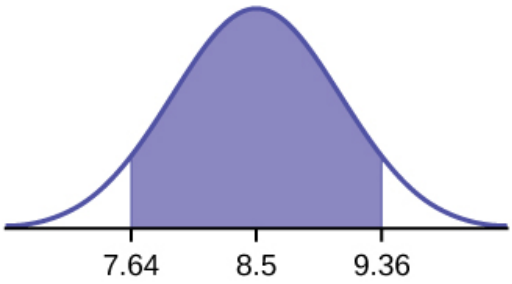
In a recent sample of 84 used car sales costs, the sample mean was $6,425 with a standard deviation of $3,156. Assume the underlying distribution is approximately normal.

1. Which distribution should you use for this problem? Explain your choice.
2. Define the random variable $\overline{X}$ in words.
3. Construct a 95% confidence interval for the population mean cost of a used car.
    1. State the confidence interval.
    2. Sketch the graph.
    3. Calculate the error bound.
4. Explain what a "95% confidence interval" means for this study.

Six different national brands of chocolate chip cookies were randomly selected at the supermarket. The grams of fat per serving are as follows: 8; 8; 10; 7; 9; 9. Assume the underlying distribution is approximately normal.

1. Construct a 90% confidence interval for the population mean grams of fat per serving of chocolate chip cookies sold in supermarkets.
    1. State the confidence interval.
    2. Sketch the graph.
    3. Calculate the error bound.
2. If you wanted a smaller error bound while keeping the same level of confidence, what should have been changed in the study before it was done?
3. Go to the store and record the grams of fat per serving of six brands of chocolate chip cookies.
4. Calculate the mean.
5. Is the mean within the interval you calculated in part a? Did you expect it to be? Why or why not?

Solution

1. CI: (7.64 , 9.36)
2. 

3. *EBM*: 0.86

2. The sample should have been increased.
3. Answers will vary.
4. Answers will vary.
5. Answers will vary.

A survey of the mean number of cents off that coupons give was conducted by randomly surveying one coupon per page from the coupon sections of a recent San Jose Mercury News. The following data were collected: 20¢; 75¢; 50¢; 65¢; 30¢; 55¢; 40¢; 40¢; 30¢; 55¢; $1.50; 40¢; 65¢; 40¢. Assume the underlying distribution is approximately normal.

1. $\overline{x}$ = _____
2. s_x = _____
3. n = _____
4. n-1 = _____

2. Define the random variables X and $\overline{X}$ in words.
3. Which distribution should you use for this problem? Explain your choice.
4. Construct a 95% confidence interval for the population mean worth of coupons.
    1. State the confidence interval.
    2. Sketch the graph.
    3. Calculate the error bound.
5. If many random samples were taken of size 14, what percent of the confidence intervals constructed should contain the population mean worth of coupons? Explain why.

Use the following information to answer the next two exercises: A quality control specialist for a restaurant chain takes a random sample of size 12 to check the amount of soda served in the 16 oz. serving size. The sample mean is 13.30 with a sample standard deviation of 1.55. Assume the underlying population is normally distributed.

Find the 95% Confidence Interval for the true population mean for the amount of soda served.

1. (12.42, 14.18)
2. (12.32, 14.29)
3. (12.50, 14.10)
4. Impossible to determine

Solution

b

What is the error bound?

1. 0.87
2. 1.98
3. 0.99
4. 1.74

7.3 Homework

Insurance companies are interested in knowing the population percentage of drivers who always buckle up before riding in a car.

1. When designing a study to determine this population proportion, what is the minimum number you would need to survey to be 95% confident that the population proportion is estimated to within 0.03?
2. If it were later determined that it was important to be more than 95% confident and a new survey was commissioned, how would that affect the minimum number you would need to survey? Why?

Solution

1. 1,068
2. The sample size would need to be increased since the critical value increases as the confidence level increases.

Suppose that the insurance companies did do a survey. They randomly surveyed 400 drivers and found that 320 claimed they always buckle up. We are interested in the population proportion of drivers who claim they always buckle up.

    1. x = _____
    2. n = _____
    3. p' = _____
2. Define the random variables X and P', in words.
3. Which distribution should you use for this problem? Explain your choice.
4. Construct a 95% confidence interval for the population proportion who claim they always buckle up.
    1. State the confidence interval.
    2. Sketch the graph.
    3. Calculate the error bound.
5. If this survey were done by telephone, list three difficulties the companies might have in obtaining random results.

According to a recent survey of 1,200 people, 61% feel that the president is doing an acceptable job. We are interested in the population proportion of people who feel the president is doing an acceptable job.

1. Define the random variables X and P' in words.
2. Which distribution should you use for this problem? Explain your choice.
3. Construct a 90% confidence interval for the population proportion of people who feel the president is doing an acceptable job.
    1. State the confidence interval.
    2. Sketch the graph.
    3. Calculate the error bound.

Solution

1. X = the number of people who feel that the president is doing an acceptable job;
   P' = the proportion of people in a sample who feel that the president is doing an acceptable job.
2. $N\left(0.61, \sqrt{\frac{(0.61)(0.39)}{1200}}\right)$
    1. CI: (0.59, 0.63)
    2. Check student's solution
    3. *EBM*: 0.02

An article regarding interracial dating and marriage recently appeared in the Washington Post. Of the 1,709 randomly selected adults, 315 identified themselves as Latinos, 323 identified themselves as blacks, 254 identified themselves as Asians, and 779 identified themselves as whites. In this survey, 86% of blacks said that

they would welcome a white person into their families. Among Asians, 77% would welcome a white person into their families, 71% would welcome a Latino, and 66% would welcome a black person.

1. We are interested in finding the 95% confidence interval for the percent of all black adults who would welcome a white person into their families. Define the random variables X and P, in words.
2. Which distribution should you use for this problem? Explain your choice.
3. Construct a 95% confidence interval.
    1. State the confidence interval.
    2. Sketch the graph.
    3. Calculate the error bound.

Refer to the information in [link].

1. Construct three 95% confidence intervals.
    1. percent of all Asians who would welcome a white person into their families.
    2. percent of all Asians who would welcome a Latino into their families.
    3. percent of all Asians who would welcome a black person into their families.
2. Even though the three-point estimates are different, do any of the confidence intervals overlap? Which?
3. For any intervals that do overlap, in words, what does this imply about the significance of the differences in the true proportions?
4. For any intervals that do not overlap, in words, what does this imply about the significance of the differences in the true proportions?

Solution

1. (0.72, 0.82)
    2. (0.65, 0.76)
    3. (0.60, 0.72)
2. Yes, the intervals (0.72, 0.82) and (0.65, 0.76) overlap, and the intervals (0.65, 0.76) and (0.60, 0.72) overlap.
3. We can say that there does not appear to be a significant difference between the proportion of Asian adults who say that their families would welcome a white person into their families and the proportion of Asian adults who say that their families would welcome a Latino person into their families.
4. We can say that there is a significant difference between the proportion of Asian adults who say that their families would welcome a white person into their families and the proportion of Asian adults who say that their families would welcome a black person into their families.

Stanford University conducted a study of whether running is healthy for men and women over age 50. During the first eight years of the study, 1.5% of the 451 members of the 50-Plus Fitness Association died. We are interested in the proportion of people over 50 who ran and died in the same eight-year period.

1. Define the random variables X and P' in words.
2. Which distribution should you use for this problem? Explain your choice.
3. Construct a 97% confidence interval for the population proportion of people over 50 who ran and died in the same eight–year period.
    1. State the confidence interval.
    2. Sketch the graph.
    3. Calculate the error bound.
4. Explain what a "97% confidence interval" means for this study.

A telephone poll of 1,000 adult Americans was reported in an issue of Time Magazine. One of the questions asked was "What is the main problem facing the country?" Twenty percent answered "crime." We are interested in the population proportion of adult Americans who feel that crime is the main problem.
Define the random variables X and P' in words.
Which distribution should you use for this problem? Explain your choice.

Construct a 95% confidence interval for the population proportion of adult Americans who feel that crime is the main problem.
State the confidence interval.
Sketch the graph.
Calculate the error bound.
Suppose we want to lower the sampling error. What is one way to accomplish that?
The sampling error given by Yankelovich Partners; Inc. (which conducted the poll) is ±3%. In one to three complete sentences, explain what the ±3% represents.
Solution
X = the number of adult Americans who feel that crime is the main problem; P' = the proportion of adult Americans who feel that crime is the main problem
Since we are estimating a proportion, given $P' = 0.2$ and $n = 1000$, the distribution we should use is

$$N\left(0.2, \sqrt{\frac{(0.2)(0.8)}{1000}}\right).$$

CI: (0.18, 0.22)
Check student's solution.
EBM: 0.02
One way to lower the sampling error is to increase the sample size.

The stated "± 3%" represents the maximum error bound. This means that those doing the study are reporting a maximum error of 3%. Thus, they estimate the percentage of adult Americans who feel that crime is the main problem to be between 18% and 22%.

Refer to [link]. Another question in the poll was "[How much are] you worried about the quality of education in our schools?" Sixty-three percent responded "a lot". We are interested in the population proportion of adult Americans who are worried a lot about the quality of education in our schools.
Define the random variables X and P in words.
Which distribution should you use for this problem? Explain your choice.

Construct a 95% confidence interval for the population proportion of adult Americans who are worried a lot about the quality of education in our schools.
State the confidence interval.
Sketch the graph.
Calculate the error bound.
The sampling error given by Yankelovich Partners, Inc. (which conducted the poll) is ±3%. In one to three complete sentences, explain what the ±3% represents.

Use the following information to answer the next three exercises: According to a Field Poll, 79% of California adults (actual results are 400 out of 506 surveyed) feel that "education and our schools" is one of the top issues facing California. We wish to construct a 90% confidence interval for the true proportion of California adults who feel that education and the schools is one of the top issues facing California.

A point estimate for the true population proportion is:

1. 0.90
2. 1.27
3. 0.79
4. 400

Solution

c

A 90% confidence interval for the population proportion is _____.

1. (0.761, 0.820)
2. (0.125, 0.188)
3. (0.755, 0.826)

4. (0.130, 0.183)

The error bound is approximately _____.

1. 1.581
2. 0.791
3. 0.059
4. 0.030

Solution

d

Use the following information to answer the next two exercises: Five hundred and eleven (511) homes in a certain southern California community are randomly surveyed to determine if they meet minimal earthquake preparedness recommendations. One hundred seventy-three (173) of the homes surveyed met the minimum recommendations for earthquake preparedness, and 338 did not.

Find the confidence interval at the 90% Confidence Level for the true population proportion of southern California community homes meeting at least the minimum recommendations for earthquake preparedness.

1. (0.2975, 0.3796)
2. (0.6270, 0.6959)
3. (0.3041, 0.3730)
4. (0.6204, 0.7025)

The point estimate for the population proportion of homes that do not meet the minimum recommendations for earthquake preparedness is _____.

1. 0.6614
2. 0.3386
3. 173
4. 338

Solution

a

On May 23, 2013, Gallup reported that of the 1,005 people surveyed, 76% of U.S. workers believe that they

will continue working past retirement age. The confidence level for this study was reported at 95% with a ±3% margin of error.

1. Determine the estimated proportion from the sample.
2. Determine the sample size.
3. Identify *CL* and α.
4. Calculate the error bound based on the information provided.
5. Compare the error bound in part d to the margin of error reported by Gallup. Explain any differences between the values.
6. Create a confidence interval for the results of this study.
7. A reporter is covering the release of this study for a local news station. How should she explain the confidence interval to her audience?

A national survey of 1,000 adults was conducted on May 13, 2013 by Rasmussen Reports. It concluded with 95% confidence that 49% to 55% of Americans believe that big-time college sports programs corrupt the process of higher education.

1. Find the point estimate and the error bound for this confidence interval.
2. Can we (with 95% confidence) conclude that more than half of all American adults believe this?
3. Use the point estimate from part a and n = 1,000 to calculate a 75% confidence interval for the proportion of American adults that believe that major college sports programs corrupt higher education.
4. Can we (with 75% confidence) conclude that at least half of all American adults believe this?

Solution

1. $p' = \frac{(0.55 + 0.49)}{2}$ = 0.52; EBP = 0.55 – 0.52 = 0.03
2. No, the confidence interval includes values less than or equal to 0.50. It is possible that less than half of the population believe this.
3. CL = 0.75, so α = 1 – 0.75 = 0.25 and $\frac{\alpha}{2} = 0.125 z_{\frac{\alpha}{2}} = 1.150$. (The area to the right of this z is 0.125, so the area to the left is 1 – 0.125 = 0.875.)

$$EBP = (1.150)\sqrt{\frac{0.52\,(0.48)}{1,000}} \approx 0.018$$

(p' – EBP, p' + EBP) = (0.52 – 0.018, 0.52 + 0.018) = (0.502, 0.538)

Alternate Solution

STAT TESTS A: 1-PropZinterval with $x = (0.52)(1{,}000)$, $n = 1{,}000$, CL = 0.75.

Answer is (0.502, 0.538)

4. Yes – this interval does not fall less than 0.50 so we can conclude that at least half of all American adults believe that major sports programs corrupt education – but we do so with only 75% confidence.

Public Policy Polling recently conducted a survey asking adults across the U.S. about music preferences. When asked, 80 of the 571 participants admitted that they have illegally downloaded music.

1. Create a 99% confidence interval for the true proportion of American adults who have illegally downloaded music.
2. This survey was conducted through automated telephone interviews on May 6 and 7, 2013. The error bound of the survey compensates for sampling error, or natural variability among samples. List some factors that could affect the survey's outcome that are not covered by the margin of error.
3. Without performing any calculations, describe how the confidence interval would change if the confidence level changed from 99% to 90%.

You plan to conduct a survey on your college campus to learn about the political awareness of students. You want to estimate the true proportion of college students on your campus who voted in the 2012 presidential election with 95% confidence and a margin of error no greater than five percent. How many students must you interview?

Solution

$CL = 0.95$ $\alpha = 1 - 0.95 = 0.05$ $\frac{\alpha}{2} = 0.025$ $z_{\frac{\alpha}{2}} = 1.96$. Use $p' = q' = 0.5$.

$$n = \frac{z_{\frac{\alpha}{2}}{}^2 p' q'}{EBP^2} = \frac{1.96^2 \, (0.5) \, (0.5)}{0.05^2} = 384.16$$

You need to interview at least 385 students to estimate the proportion to within 5% at 95% confidence.

In a recent Zogby International Poll, nine of 48 respondents rated the likelihood of a terrorist attack in their community as "likely" or "very likely." Use the "plus four" method to create a 97% confidence interval for the proportion of American adults who believe that a terrorist attack in their community is likely or very likely. Explain what this confidence interval means in the context of the problem.

CHAPTER 8: HYPOTHESIS TESTING WITH ONE SAMPLE

INTRODUCTION TO CHAPTER 8: HYPOTHESIS TESTING WITH ONE SAMPLE



Figure 9.1. You can use a hypothesis test to decide if a dog breeder's claim that every Dalmatian has 35 spots is statistically sound. (Credit: Robert Neff)

One job of a statistician is to make statistical inferences about populations based on samples taken from the population. Confidence intervals are one way to estimate a population parameter. Another way to make a statistical inference is to make a decision about a parameter. For instance, a car dealer advertises that its new small truck gets 35 miles per gallon, on average. A tutoring service claims that its method of tutoring helps 90% of its students get an A or a B. A company says that women managers in their company earn an average of $60,000 per year.

A statistician will make a decision about these claims. This process is called **"hypothesis testing."** A

hypothesis test involves collecting data from a sample and evaluating the data. Then, the statistician makes a decision as to whether or not there is sufficient evidence, based upon analyses of the data, to reject the null hypothesis.

In this chapter, you will conduct hypothesis tests on single means and single proportions. You will also learn about the errors associated with these tests.

Hypothesis testing consists of two contradictory hypotheses or statements, a decision based on the data, and a conclusion. To perform a hypothesis test, a statistician will:

1. Set up two contradictory hypotheses.
2. Collect sample data (in homework problems, the data or summary statistics will be given to you).
3. Determine the correct distribution to perform the hypothesis test.
4. Analyze sample data by performing the calculations that ultimately will allow you to reject or decline to reject the null hypothesis.
5. Make a decision and write a meaningful conclusion.

Note

To do the hypothesis test homework problems for this chapter and later chapters, make copies of the appropriate special solution sheets. See Appendix E.

Collaborative Exercises

Collaborative Exercise

Bring to class a newspaper, some news magazines, and some Internet articles. In groups, find articles from which your group can write null and alternative hypotheses. Discuss your hypotheses with the rest of the class.

Media Attributions

- Chapter 9 Introduction Figure © OpenStax Introductory Statistics is licensed under a CC BY (Attribution) license

8.1 NULL AND ALTERNATIVE HYPOTHESES

Learning Objectives

By the end of this section, the student should be able to:

- Describe hypothesis testing in general and in practice.

Hypothesis Testing

The actual test begins by considering two **hypotheses**. They are called the null hypothesis and the alternative hypothesis. These hypotheses contain opposing viewpoints.

H_0: **The null hypothesis:** It is a statement about the population that either is believed to be true or is used to put forth an argument unless it can be shown to be incorrect beyond a reasonable doubt.

H_a: **The alternative hypothesis:** It is a claim about the population that is contradictory to H_0 and what we conclude when we reject H_0.

Since the null and alternative hypotheses are contradictory, you must examine evidence to decide if you have enough evidence to reject the null hypothesis or not. The evidence is in the form of sample data.

After you have determined which hypothesis the sample supports, you make a **decision.** There are two options for a decision. They are "reject H_0" if the sample information favors the alternative hypothesis or "do not reject H_0" or "decline to reject H_0" if the sample information is insufficient to reject the null hypothesis.

Mathematical Symbols Used in H_0 and H_a:

H_0	H_a
equal (=)	not equal (≠) **or** greater than (>) **or** less than (<)
greater than or equal to (≥)	less than (<)
less than or equal to (≤)	more than (>)



One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=863#oembed-1

Note

H_0 always has a symbol with an equal in it. H_a never has a symbol with an equal in it. The choice of symbol depends on the wording of the hypothesis test. However, be aware that many researchers (including one of the co-authors in research work) use = in the null hypothesis, even with > or < as the symbol in the alternative hypothesis. This practice is acceptable because we only make the decision to reject or not reject the null hypothesis.

H_0: No more than 30% of the registered voters in Santa Clara County voted in the primary election. $p \leq 30$

H_a: More than 30% of the registered voters in Santa Clara County voted in the primary election. $p > 30$

Example

A medical trial is conducted to test whether or not a new medicine reduces cholesterol by 25%. State the null and alternative hypotheses.

Solution

H_0: The drug reduces cholesterol by 25%. $p = 0.25$

H_a: The drug does not reduce cholesterol by 25%. $p \neq 0.25$

Your turn!

We want to test whether the mean GPA of students in American colleges is different from 2.0 (out of 4.0). The null and alternative hypotheses are:

Solution

H_0: $\mu = 2.0$

H_a: $\mu \neq 2.0$

Exercises

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=863#h5p-200

Example

We want to test whether the mean height of eighth graders is 66 inches. State the null and alternative hypotheses. Fill in the correct symbol (=, ≠, ≥, <, ≤, >) for the null and alternative hypotheses.

H_0: μ __ 66
Ha: μ __ 66
Solution

1. $H0$: μ = 66
2. Ha : μ ≠ 66

Your Turn!

We want to test if college students take less than five years to graduate from college, on the average. The null and alternative hypotheses are:

Solution
H_0: $\mu \geq 5$

H_a: $\mu < 5$

Example

We want to test if it takes fewer than 45 minutes to teach a lesson plan. State the null and alternative hypotheses. Fill in the correct symbol (=, ≠, ≥, <, ≤, >) for the null and alternative hypotheses.

H0: μ __ 45
Ha: μ __ 45
Solution

1. H0: μ ≥ 45
2. Ha: μ < 45

Your Turn!

In an issue of *U.S. News and World Report*, an article on school standards stated that about half of all students in France, Germany, and Israel take advanced placement exams and a third pass. The same article stated that 6.6% of U.S. students take advanced placement exams and 4.4% pass. Test if the percentage of U.S. students who take advanced placement exams is more than 6.6%. State the null and alternative hypotheses.

Solution
H0: p ≤ 0.066

Ha: p > 0.066

Example

On a state driver's test, about 40% pass the test on the first try. We want to test if more than 40% pass on the first try. Fill in the correct symbol (=, ≠, ≥, <, ≤, >) for the null and alternative hypotheses.

H_0: p __ 0.40
H_a: p __ 0.40
Solution

1. H_0: $p = 0.40$
2. H_a: $p > 0.40$

References

Data from the National Institute of Mental Health. Available online at http://www.nimh.nih.gov/publicat/depression.cfm.

8.2 OUTCOMES AND THE TYPE I AND TYPE II ERRORS

Learning Objectives

By the end of this section, the student should be able to:

- Differentiate between Type I and Type II errors in a hypothesis test.

Type I and Type II Errors

When you perform a hypothesis test, there are four possible outcomes depending on the actual truth (or falseness) of the null hypothesis H_0 and the decision to reject or not. The outcomes are summarized in the following table:

ACTION	H_0 **IS ACTUALLY**	**...**
	True	False
Do not reject H_0	Correct Outcome	**Type II error**
Reject H_0	**Type I Error**	Correct Outcome

The four possible outcomes in the table are:
The decision is **not to reject H_0** when **H_0 is true (correct decision).**
The decision is to **reject H_0** when **H_0 is true** (incorrect decision known as a Type I error).
The decision is **not to reject H_0** when, in fact, **H_0 is false** (incorrect decision known as a Type II error).
The decision is to **reject H_0** when **H_0 is false** (**correct decision** whose probability is called the **Power of the Test**).

Each of the errors occurs with a particular probability. The Greek letters α and β represent the probabilities.

α = probability of a **Type I error** = P**(Type I error)** = probability of rejecting the null hypothesis when the null hypothesis is true.

β = probability of a **Type II error** = P**(Type II error)** = probability of not rejecting the null hypothesis when the null hypothesis is false.

α and β should be as small as possible because they are probabilities of errors. They are rarely zero.

The Power of the Test is $1 - \beta$. Ideally, we want a high power that is as close to one as possible. Increasing the sample size can increase the Power of the Test.

The following are examples of **Type I** and **Type II** errors.

Suppose the null hypothesis, H0, is: Frank's rock climbing equipment is safe.

Type I error: Frank thinks that his rock climbing equipment may not be safe when, in fact, it really is safe. **Type II error:** Frank thinks that his rock climbing equipment may be safe when, in fact, it is not safe.

α = probability that Frank thinks his rock climbing equipment may not be safe when, in fact, it really is safe. β = probability that Frank thinks his rock climbing equipment may be safe when, in fact, it is not safe.

Notice that, in this case, the error with the greater consequence is the Type II error. (If Frank thinks his rock climbing equipment is safe, he will go ahead and use it.)

Example

Suppose the null hypothesis, H_0, is: the blood cultures contain no traces of pathogen X. State the Type I and Type II errors.

Solution

Type I error: The researcher thinks the blood cultures do contain traces of pathogen X, when in fact, they do not.

Type II error: The researcher thinks the blood cultures do not contain traces of pathogen X, when in fact, they do.

Your Turn!

Suppose the null hypothesis, H0, is: The victim of an automobile accident is alive when he arrives at the emergency room of a hospital.

Type I error: The emergency crew thinks that the victim is dead when, in fact, the victim is alive.
Type II error: The emergency crew does not know if the victim is alive when, in fact, the victim is dead.

α = probability that the emergency crew thinks the victim is dead when, in fact, he is really alive = P(Type I error).
β = probability that the emergency crew does not know if the victim is alive when, in fact, the victim is dead = P(Type II error).

The error with the greater consequence is the Type I error. (If the emergency crew thinks the victim is dead, they will not treat him.)

Example

Suppose the null hypothesis, H_0, is: a patient is not sick. Which type of error has the greater consequence, Type I or Type II?

Solution
The error with the greater consequence is the Type II error: the patient will be thought well when, in fact, he is sick, so he will not get treatment.

Exercises

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=864#h5p-201

Example

It's a Boy Genetic Labs claim to be able to increase the likelihood that a pregnancy will result in a boy being born. Statisticians want to test the claim. Suppose that the null hypothesis, H0, is: It's a Boy Genetic Labs has no effect on gender outcome.

Type I error: This results when a true null hypothesis is rejected. In the context of this scenario, we would state that we believe that It's a Boy Genetic Labs influences the gender outcome, when in fact it has no effect. The probability of this error occurring is denoted by the Greek letter alpha, **α**.

Type II error: This results when we fail to reject a false null hypothesis. In context, we would state that It's a Boy Genetic Labs does not influence the gender outcome of a pregnancy when, in fact, it does. The probability of this error occurring is denoted by the Greek letter beta, **β**.

The error of greater consequence would be the Type I error since couples would use the It's a Boy Genetic Labs product in hopes of increasing the chances of having a boy.

Your Turn!

"Red tide" is a bloom of poison-producing algae—a few different species of a class of plankton called dinoflagellates. When the weather and water conditions cause these blooms, shellfish such as clams living in the area develop dangerous levels of a paralysis-inducing toxin. In Massachusetts, the Division of Marine Fisheries (DMF) monitors levels of the toxin in shellfish by regular sampling of shellfish along the coastline. If the mean level of toxin in clams exceeds 800 μg (micrograms) of toxin per kg of clam meat in any area, clam harvesting is banned there until the bloom is over and levels of toxin in clams subside. Describe both a Type I and a Type II error in this context, and state which error has the greater consequence.

Solution
In this scenario, an appropriate null hypothesis would beH0: the mean level of toxins is at most 800 μg, H0 : μ0 ≤ 800 μg.

Type I error: The DMF believes that toxin levels are still too high when, in fact, toxin levels are at most 800 μg. The DMF continues the harvesting ban.

Type II error: The DMF believes that toxin levels are within acceptable levels (are at least 800 μg) when, in fact, toxin levels are still too high (more than 800 μg). The DMF lifts the harvesting ban. This error could be the most serious. If the ban is lifted and clams are still toxic, consumers could possibly eat tainted food.

In summary, the more dangerous error would be to commit a Type II error, because this error involves the availability of tainted clams for consumption.

Example

A certain experimental drug claims a cure rate of at least 75% for males with prostate cancer. Describe both the Type I and Type II errors in context. Which error is the more serious?

Type I: A cancer patient believes the cure rate for the drug is less than 75% when it actually is at least 75%.

Type II: A cancer patient believes the experimental drug has at least a 75% cure rate when it has a cure rate that is less than 75%.

In this scenario, the Type II error contains the more severe consequence. If a patient believes the drug works at least 75% of the time, this most likely will influence the patient's (and doctor's) choice about whether to use the drug as a treatment option.

Your Turn!

Determine both Type I and Type II errors for the following scenario:

Assume a null hypothesis, H_0, that states the percentage of adults with jobs is at least 88%.

Identify the Type I and Type II errors from these four statements.

Not to reject the null hypothesis that the percentage of adults who have jobs is at least 88% when that percentage is actually less than 88%
Not to reject the null hypothesis that the percentage of adults who have jobs is at least 88% when the percentage is actually at least 88%.
Reject the null hypothesis that the percentage of adults who have jobs is at least 88% when the percentage is actually at least 88%.
Reject the null hypothesis that the percentage of adults who have jobs is at least 88% when that percentage is actually less than 88%.
Solution
Type I error: c

Type I error: b



One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=864#oembed-1

8.3 DISTRIBUTION NEEDED FOR HYPOTHESIS TESTING

Learning Objectives

By the end of this section, the student should be able to:

- Identify the distribution required to conduct a hypothesis test.

Distributions Required to Conduct a Hypothesis Test

Earlier in the course, we discussed sampling distributions. **Particular distributions are associated with hypothesis testing.** Perform tests of a population mean using a **normal distribution** or a **Student's *t*-distribution**. (Remember, use a **Student's *t*-distribution** when the population standard deviation is unknown and the distribution of the sample mean is approximately normal.) We perform tests of a population proportion using a **normal distribution** (usually n is large or the sample size is large).

If you are testing a single population mean, the distribution for the test is for **means**:

$$\overline{X} \sim N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right) \text{ or } t_{df}$$

The population parameter is μ. The estimated value (point estimate) for μ is $\overline{x}$, the sample mean.

If you are testing a single population proportion, the distribution for the test is for proportions or percentages:

$$P' \sim N\left(p, \sqrt{\frac{p \cdot q}{n}}\right)$$

The population parameter is p. The estimated value (point estimate) for p is p'. $p' = \frac{x}{n}$ where x is the number of successes and n is the sample size.

Assumptions

When you perform a hypothesis test **of a single population mean μ** using a **Student's t-distribution** (often called a t-test), there are fundamental assumptions that need to be met in order for the test to work properly. Your data should be a simple random sample that comes from a population that is approximately normally distributed. You use the sample **standard deviation** to approximate the population standard deviation. (Note that if the sample size is sufficiently large, a t-test will work even if the population is not approximately normally distributed.)

When you perform a **hypothesis test of a single population mean μ** using a normal distribution (often called a z-test), you take a simple random sample from the population. The population you are testing is normally distributed or your sample size is sufficiently large. You know the value of the population standard deviation which, in reality, is rarely known.

When you perform a **hypothesis test of a single population proportion p**, you take a simple random sample from the population. You must meet the conditions for a **binomial distribution** which are: there are a certain number n of independent trials, the outcomes of any trial are success or failure, and each trial has the same probability of a success p. The shape of the binomial distribution needs to be similar to the shape of the normal distribution. To ensure this, the quantities np and nq must both be greater than five ($np > 5$ and $nq > 5$). Then the binomial distribution of a sample (estimated) proportion can be approximated by the normal distribution with $\mu = p$ and $\sigma = \sqrt{\frac{pq}{n}}$.

Remember that $q = 1 - p$.

8.4 RARE EVENTS, THE SAMPLE, DECISION, AND CONCLUSION

Learning Objectives

By the end of this section, the student should be able to:

- Identify rare events
- Conduct and interpret hypothesis tests.
- Describe a decision and conclusion.

Establishing the type of distribution, sample size, and known or unknown standard deviation can help you figure out how to go about a hypothesis test. However, there are several other factors you should consider when working out a hypothesis test.

Rare Events

Suppose you make an assumption about a property of the population (this assumption is the null hypothesis). Then you gather sample data randomly. If the sample has properties that would be very **unlikely** to occur if the assumption is true, then you would conclude that your assumption about the population is probably incorrect. (Remember that your assumption is just an assumption—it is not a fact and it may or may not be true. But your sample data are real and the data are showing you a fact that seems to contradict your assumption.)

For example, Didi and Ali are at a birthday party of a very wealthy friend. They hurry to be first in line to grab a prize from a tall basket that they cannot see inside because they will be blindfolded. There are 200 plastic bubbles in the basket and Didi and Ali have been told that there is only one with a $100 bill. Didi is the first person to reach into the basket and pull out a bubble. Her bubble contains a $100 bill. The probability of

this happening is $\frac{1}{200}$ = 0.005. Because this is so unlikely, Ali is hoping that what the two of them were told is wrong and there are more $ 100 bills in the basket. A "rare event" has occurred (Didi getting the $ 100 bill) so Ali doubts the assumption about only one $ 100 bill being in the basket.

Using the Sample to Test the Null Hypothesis

Use the sample data to calculate the actual probability of getting the test result, called the *p*-value. The *p*-value is the **probability that, if the null hypothesis is true, the results from another randomly selected sample will be as extreme or more extreme as the results obtained from the given sample.**

A large ***p*-value** calculated from the data indicates that we should not reject the null hypothesis. The smaller the *p*-value, the more unlikely the outcome, and the stronger the evidence is against the null hypothesis. We would reject the null hypothesis if the evidence is strongly against it.

Draw a graph that shows the *p*-value. The hypothesis test is easier to perform if you use a graph because you see the problem more clearly.

Example

Suppose a baker claims that his bread height is more than 15 cm, on average. Several of his customers do not believe him. To persuade his customers that he is right, the baker decides to do a hypothesis test. He bakes 10 loaves of bread. The mean height of the sample loaves is 17 cm. The baker knows from baking hundreds of loaves of bread that the standard deviation for the height is 0.5 cm. and the distribution of heights is normal.

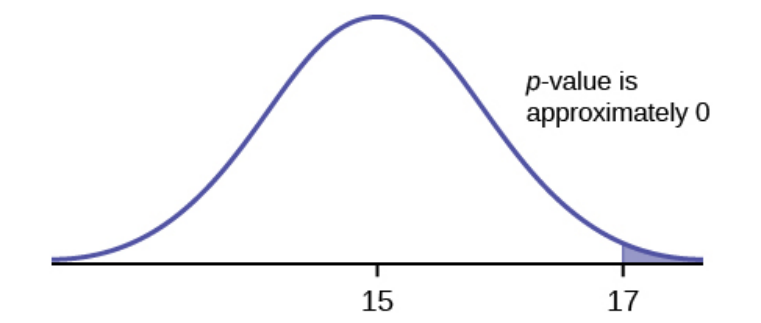
The null hypothesis could be H_0: $\mu \leq 15$ The alternate hypothesis is H_a: $\mu > 15$

The words **"is more than"** translates as a ">" so "$\mu > 15$" goes into the alternate hypothesis. The null hypothesis must contradict the alternate hypothesis.

Since **σ is known** (σ = 0.5 cm.), the distribution for the population is known to be normal with mean μ = 15 and standard deviation $\frac{\sigma}{\sqrt{n}} = \frac{0.5}{\sqrt{10}} = 0.16$.

Suppose the null hypothesis is true (the mean height of the loaves is no more than 15 cm). Then is the mean height (17 cm) calculated from the sample unexpectedly large? The hypothesis test works by asking the question how **unlikely** the sample mean would be if the null hypothesis were true. The graph shows how far out the sample mean is on the normal curve. The *p*-value is the probability that, if we were to take other samples, any other sample mean would fall at least as far out as 17 cm.

The *p*-value, then, is the probability that a sample mean is the same or greater than 17 cm. when the population mean is, in fact, 15 cm. We can calculate this probability using the normal distribution for means.



p-value= $P(\overline{x} > 17)$ which is approximately zero.

A *p*-value of approximately zero tells us that it is highly unlikely that a loaf of bread rises no more than 15 cm, on average. That is, almost 0% of all loaves of bread would be at least as high as 17 cm. **purely by CHANCE** had the population mean height really been 15 cm. Because the outcome of 17 cm. is so **unlikely (meaning it is happening NOT by chance alone)**, we conclude that the evidence is strongly against the null hypothesis (the mean height is at most 15 cm.). There is sufficient evidence that the true mean height for the population of the baker's loaves of bread is greater than 15 cm.

Your Turn!

A normal distribution has a standard deviation of 1. We want to verify a claim that the mean is greater than 12. A sample of 36 is taken with a sample mean of 12.5.

H_0: $\mu \leq 12$

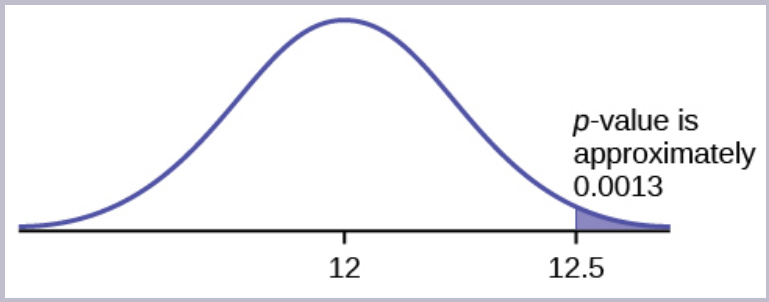
H_a: $\mu > 12$

The p-value is 0.0013

Draw a graph that shows the p-value.

Solution

p-value = 0.0013



Decision and Conclusion

A systematic way to make a decision of whether to reject or not reject the null hypothesis is to compare the p-value and a **preset or preconceived α (also called a "significance level")**. A preset α is the probability of a Type I error (rejecting the null hypothesis when the null hypothesis is true). It may or may not be given to you at the beginning of the problem.

When you make a **decision** to reject or not reject H_0, do as follows:

If $\alpha > p$-value, reject H_0. The results of the sample data are significant. There is sufficient evidence to conclude that H_0 is an incorrect belief and that the **alternative hypothesis**, H_a, may be correct.

If $\alpha \leq p$-value, do not reject H_0. The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis, H_a, may be correct.

When you "do not reject H_0", it does not mean that you should believe that H_0 is true. It simply means that the sample data have **failed** to provide sufficient evidence to cast serious doubt about the truthfulness of H_o.

Conclusion: After you make your decision, write a thoughtful **conclusion** about the hypotheses in terms of the given problem.

Example

When using the *p*-value to evaluate a hypothesis test, it is sometimes useful to use the following memory device

If the *p*-value is low, the null must go.

If the *p*-value is high, the null must fly.

This memory aid relates a *p*-value less than the established alpha (the *p* is low) as rejecting the null hypothesis and, likewise, relates a *p*-value higher than the established alpha (the *p* is high) as not rejecting the null hypothesis.

Fill in the blanks.

Reject the null hypothesis when _____.

The results of the sample data _____.

Do not reject the null when hypothesis when

_____.

The results of the sample data _____.

Solution

Reject the null hypothesis when **the *p*-value is less than the established alpha value**. The results of the sample data **support the alternative hypothesis**.

Do not reject the null hypothesis when **the *p*-value is greater than the established alpha value**. The results of the sample data **do not support the alternative hypothesis**.

Your Turn!

It's a Boy Genetics Labs claim their procedures improve the chances of a boy being born. The results for a test of a single population proportion are as follows:

H_0: $p = 0.50$, H_a: $p > 0.50$

$\alpha = 0.01$

p-value = 0.025

Interpret the results and state a conclusion in simple, non-technical terms.

Solution

Since the p-value is greater than the established alpha value (the p-value is high), we do not reject the null hypothesis. There is not enough evidence to support It's a Boy Genetics Labs' stated claim that their procedures improve the chances of a boy being born.

8.5 ADDITIONAL INFORMATION AND FULL HYPOTHESIS TEST EXAMPLES

Learning Objectives

By the end of this section, the student should be able to:

- Conduct and interpret hypothesis tests for a single population mean, population standard deviation known.
- Conduct and interpret hypothesis tests for a single population mean, population standard deviation unknown.
- Conduct and interpret hypothesis tests for a single population proportion.

Hypothesis Tests for a Single Population Mean, Population Standard Deviation Known

- In a hypothesis test problem, you may see words such as "the level of significance is 1%." The "1%" is the preconceived or preset α.
- The statistician setting up the hypothesis test selects the value of α to use **before** collecting the sample data.
- **If no level of significance is given, a common standard to use is $\alpha = 0.05$.**
- When you calculate the p-value and draw the picture, the p-value is the area in the left tail, the right tail, or split evenly between the two tails. For this reason, we call the hypothesis test left, right, or two tailed.
- The **alternative hypothesis**, H_a, tells you if the test is left, right, or two-tailed. It is the **key** to conducting the appropriate test.

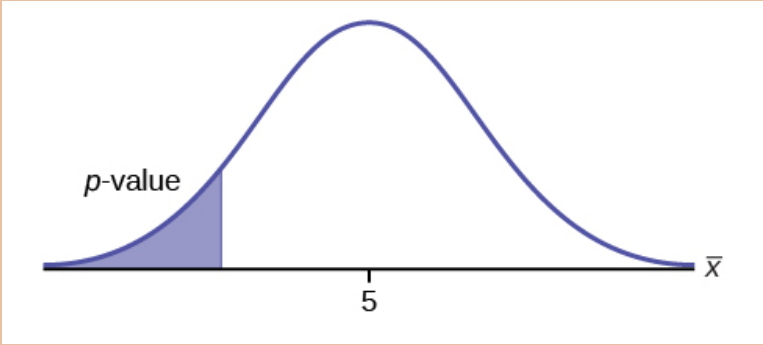
- H_a **never** has a symbol that contains an equal sign.
- **Thinking about the meaning of the** *p*-value: A data analyst (and anyone else) should have more confidence that he made the correct decision to reject the null hypothesis with a smaller *p*-value (for example, 0.001 as opposed to 0.04) even if using the 0.05 level for alpha. Similarly, for a large *p*-value such as 0.4, as opposed to a *p*-value of 0.056 (alpha = 0.05 is less than either number), a data analyst should have more confidence that she made the correct decision in not rejecting the null hypothesis. This makes the data analyst use judgment rather than mindlessly applying rules.

The following examples illustrate a left-, right-, and two-tailed test.

Example

H_0: μ = 5, H_a: μ < 5

Test of a single population mean. H_a tells you the test is left-tailed. The picture of the p-value is as follows:



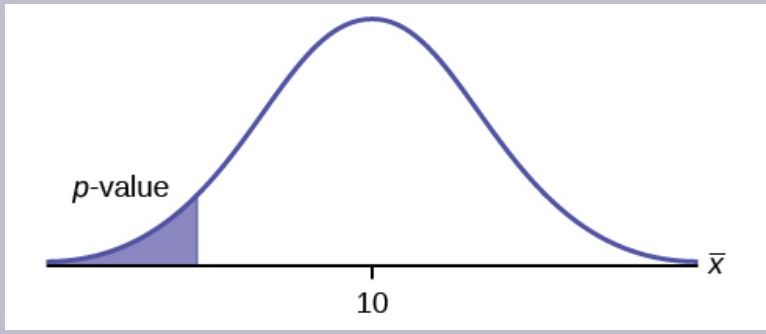
Your Turn!

H_0: $\mu = 10$, H_a: $\mu < 10$

Assume the *p*-value is 0.0935. What type of test is this? Draw the picture of the *p*-value.

Solution

left-tailed test



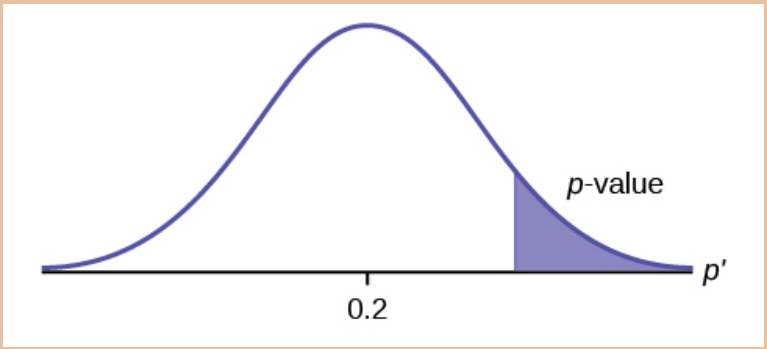
One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=893#oembed-1

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=893#oembed-2

Example

$H_0: p \leq 0.2 \quad H_a: p > 0.2$

This is a test of a single population proportion. H_a tells you the test is right-tailed. The picture of the p-value is as follows:



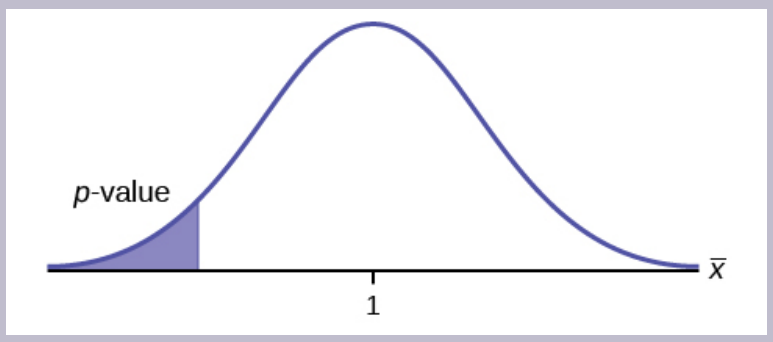
Your Turn!

H_0: $\mu \leq 1$, H_a: $\mu > 1$

Assume the p-value is 0.1243. What type of test is this? Draw the picture of the p-value.

Solution

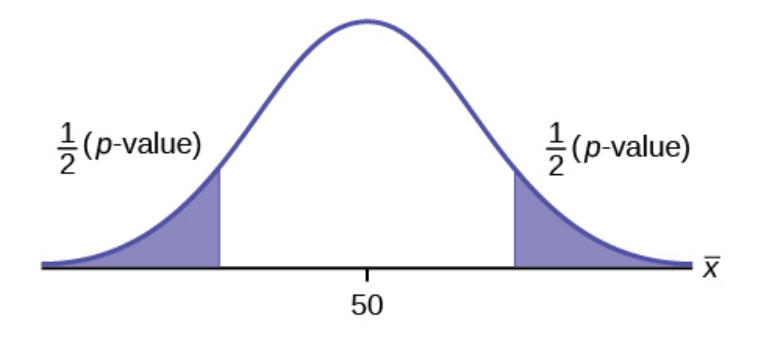
right-tailed test



Example

H_0: p = 50 H_a: p ≠ 50

This is a test of a single population mean. H_a tells you the test is two-tailed. The picture of the p-value is as follows.

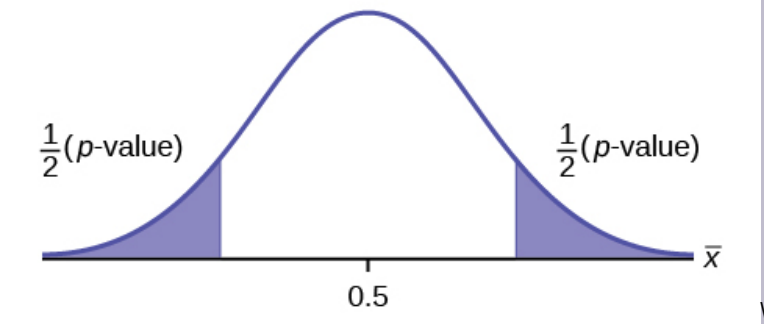


Your Turn!

H_0: p = 0.5, H_a: p ≠ 0.5

Assume the p-value is 0.2564. What type of test is this? Draw the picture of the p-value.

Solution
two-tailed test



\

Full Hypothesis Test Examples

Example

Jeffrey, as an eight-year old, established a mean time of 16.43 seconds for swimming the 25-yard freestyle, with a standard deviation of 0.8 seconds. His dad, Frank, thought that Jeffrey could swim the 25-yard freestyle faster using goggles. Frank bought Jeffrey a new pair of expensive goggles and timed Jeffrey for 15 25-yard freestyle swims. For the 15 swims, Jeffrey's mean time was 16 seconds. Frank thought that the goggles helped Jeffrey to swim faster than the 16.43 seconds. Conduct a hypothesis test using a preset $\alpha = 0.05$. Assume that the swim times for the 25-yard freestyle are normal.

Solution

Set up the Hypothesis Test:

Since the problem is about a mean, this is a test of a single population mean.

$H_0: \mu = 16.43 \quad H_a: \mu < 16.43$

For Jeffrey to swim faster, his time will be less than 16.43 seconds. The "<" tells you this is left-tailed.

Determine the distribution needed:

Random variable: $\overline{X}$ = the mean time to swim the 25-yard freestyle.

Distribution for the test: $\overline{X}$ is normal (population standard deviation is known: $\sigma = 0.8$)

$$\overline{X}\ N\left(\mu, \frac{\sigma_X}{\sqrt{n}}\right)$$

Therefore, $\overline{X}\ N\left(16.43, \frac{0.8}{\sqrt{15}}\right)$

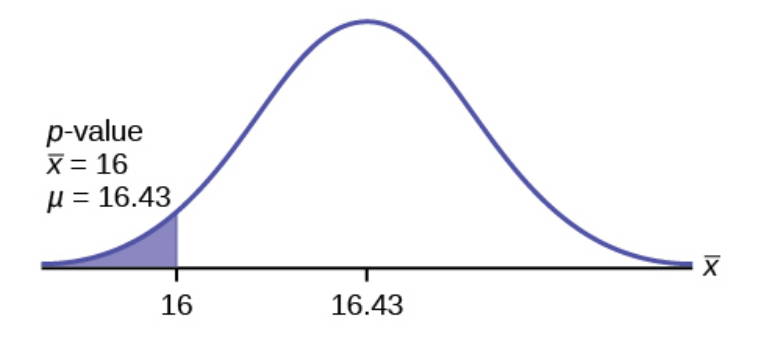
$\mu = 16.43$ comes from H_0 and not the data. $\sigma = 0.8$, and $n = 15$.

Calculate the p-value using the normal distribution for a mean:

p-value = $P(\overline{x} < 16) = 0.0187$ where the sample mean in the problem is given as 16.

p-value = 0.0187 (This is called the actual level of significance.) The p-value is the area to the left of the sample mean is given as 16.

Graph:



μ = 16.43 comes from H_0. Our assumption is μ = 16.43.

Interpretation of the p-value: If H_0 is true, there is a 0.0187 probability (1.87%)that Jeffrey's mean time to swim the 25-yard freestyle is 16 seconds or less. Because a 1.87% chance is small, the mean time of 16 seconds or less is unlikely to have happened randomly. It is a rare event.

Compare **α** and the p-value:

α = 0.05 p-value = 0.0187 **α** > p-value

Make a decision: Since **α** > p-value, reject H_0.

<!– CONTINUE HERE –>

This means that you reject μ = 16.43. In other words, you do not think Jeffrey swims the 25-yard freestyle in 16.43 seconds but faster with the new goggles.

Conclusion: At the 5% significance level, we conclude that Jeffrey swims faster using the new goggles. The sample data show there is sufficient evidence that Jeffrey's mean time to swim the 25-yard freestyle is less than 16.43 seconds.

The p-value can easily be calculated.

Press STAT and arrow over to TESTS. Press 1:Z-Test. Arrow over to Stats and press ENTER. Arrow down and enter 16.43 for μ0 (null hypothesis), .8 for **σ**, 16 for the sample mean, and 15 for n. Arrow down to μ : (alternate hypothesis) and arrow over to < μ0. Press ENTER. Arrow down to Calculate and press ENTER. The calculator not only calculates the p-value (p = 0.0187) but it also calculates the test statistic (z-score) for the sample mean. μ < 16.43 is the alternative hypothesis. Do this set of instructions again except arrow to Draw(instead of Calculate). Press ENTER. A shaded graph appears with z = -2.08 (test statistic) and p = 0.0187 (p-value). Make sure when you use Draw that no other equations are highlighted in Y = and the plots are turned off.

When the calculator does a Z-Test, the Z-Test function finds the p-value by doing a normal probability calculation using the central limit theorem:

$P(\overline{x} < 16)$ =2nd DISTR normcdf$\left(-10^9 9, 16, 16.43, 0.8/\sqrt{15}\right)$

.

The Type I and Type II errors for this problem are as follows:

The Type I error is to conclude that Jeffrey swims the 25-yard freestyle, on average, in less than 16.43 seconds when, in fact, he actually swims the 25-yard freestyle, on average, in 16.43 seconds. (Reject the null hypothesis when the null hypothesis is true.)

The Type II error is that there is no evidence to conclude that Jeffrey swims the 25-yard free-style, on average, in less than 16.43 seconds when, in fact, he actually does swim the 25-yard free-style, on average, in less than 16.43 seconds. (Do not reject the null hypothesis when the null hypothesis is false.)

Your Turn!

The mean throwing distance of a football for Marco, a high school freshman quarterback, is 40 yards, with a standard deviation of two yards. The team coach tells Marco to adjust his grip to get more distance. The coach records the distances for 20 throws. For the 20 throws, Marco's mean distance was 45 yards. The coach thought the different grip helped Marco throw farther than 40 yards. Conduct a hypothesis test using a preset $\alpha = 0.05$. Assume the throw distances for footballs are normal.

First, determine what type of test this is, set up the hypothesis test, find the *p*-value, sketch the graph, and state your conclusion.

Press STAT and arrow over to TESTS. Press 1:Z-Test. Arrow over to Stats and press ENTER. Arrow down and enter 40 for $\mu 0$ (null hypothesis), 2 for σ, 45 for the sample mean, and 20 for *n*. Arrow down to μ: (alternative hypothesis) and set it either as <, ≠, or >. Press ENTER. Arrow down to Calculate and press ENTER. The calculator not only calculates the *p*-value but it also calculates the test statistic (*z*-score) for the sample mean. Select <, ≠, or > for the alternative hypothesis. Do this set of instructions again except arrow to Draw (instead of Calculate). Press ENTER. A shaded graph appears with test statistics and *p*-value. Make sure when you use Draw that no other equations are highlighted in *Y* = and the plots are turned off.

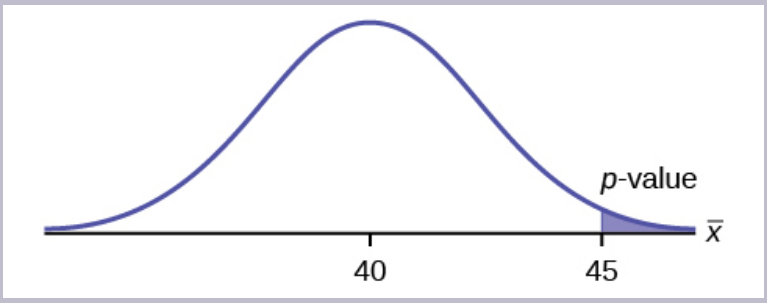
Solution

Since the problem is about a mean, this is a test of a single population mean.

$H_0 : \mu = 40$

$H_a : \mu > 40$

$p = 0.0062$

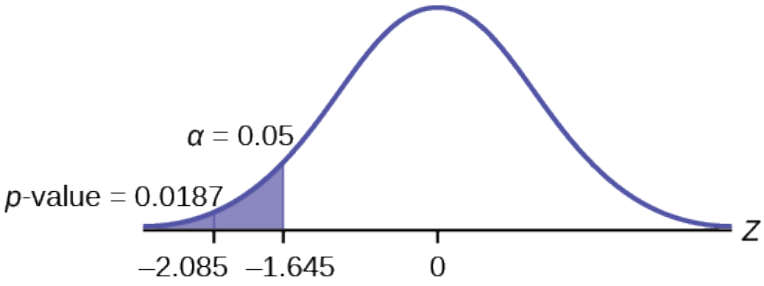


Because $p < \alpha$, we reject the null hypothesis. There is sufficient evidence to suggest that the change in grip improved Marco's throwing distance.

Historical Note ([link])

The traditional way to compare the two probabilities, α and the p-value, is to compare the critical value (z-score from α) to the test statistic (z-score from data). The calculated test statistic for the p-value is –2.08. (From the **Central Limit Theorem**, the test statistic formula is $z = \frac{\overline{x} - \mu_X}{\left(\frac{\sigma_X}{\sqrt{n}}\right)}$.)

For this problem, $\overline{x}$ = 16, μ_X = 16.43 from the null hypothesis is σ_X = 0.8, and n = 15.) You can find the critical value for α = 0.05 in the normal table (see **15.Tables** in the Table of Contents). The z-score for an area to the left equal to 0.05 is midway between –1.65 and –1.64 (0.05 is midway between 0.0505 and 0.0495). The z-score is –1.645. Since –1.645 > –2.08 (which demonstrates that $\alpha > p$-value), reject H_0. Traditionally, the decision to reject or not reject was done in this way. Today, comparing the two probabilities α and the p-value is very common. For this problem, the p-value, 0.0187 is considerably smaller than α, 0.05. You can be confident about your decision to reject. The graph shows α, the p-value, and the test statistics and the critical value.



Example

A college football coach thought that his players could bench press a mean weight of 275 pounds. It is known that the standard deviation is 55 pounds. Three of his players thought that the mean weight was more than that amount. They asked 30 of their teammates for their estimated maximum lift on the bench press exercise. The data ranged from 205 pounds to 385 pounds. The actual different weights were (frequencies are in parentheses) 205(3)215(3)225(1)241(2)252(2)265(2)275(2)313(2)316(5)338(2)341(1)345(2)368(2)385(1).

Conduct a hypothesis test using a 2.5% level of significance to determine if the bench press mean is more than 275 pounds.

Solution
Set up the Hypothesis Test:

Since the problem is about a mean weight, this is a test of a single population mean.

H_0: μ = 275

H_a: μ > 275

This is a right-tailed test.

Calculating the distribution needed:

Random variable: $\overline{X}$ = the mean weight, in pounds, lifted by the football players.

Distribution for the test: It is normal because σ is known.

$$\overline{X}\ N\left(275, \frac{55}{\sqrt{30}}\right)$$

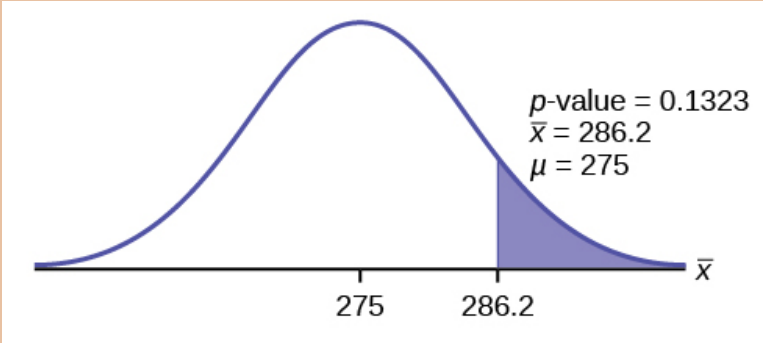
$\overline{x} = 286.2$ pounds (from the data).

σ = 55 pounds (Always use σ if you know it.) We assume μ = 275 pounds unless our data shows us otherwise.

Calculate the p-value using the normal distribution for a mean and using the sample mean as input (see [link] for using the data as input):

$$p\text{-value} = P(\overline{x} > 286.2) = 0.1323.$$

Interpretation of the p-value: If H_0 is true, then there is a 0.1331 probability (13.23%) that the football players can lift a mean weight of 286.2 pounds or more. Because a 13.23% chance is large enough, a mean weight lift of 286.2 pounds or more is not a rare event.



Compare α and the p-value:

α = 0.025 p-value = 0.1323

Make a decision: Since α <p-value, do not reject H_0.

Conclusion: At the 2.5% level of significance, from the sample data, there is not sufficient evidence to conclude that the true mean weight lifted is more than 275 pounds.

The p-value can easily be calculated.

Put the data and frequencies into lists. Press STAT and arrow over to TESTS. Press 1:Z-Test. Arrow over to Data and press ENTER. Arrow down and enter 275 for μ0, 55 for σ, the name of the list where you put the data, and the name of the list where you put the frequencies. Arrow down to μ: and arrow over to > μ0. Press ENTER. Arrow down to Calculate and press ENTER. The calculator not only calculates the p-value (p = 0.1331, a little different from the previous calculation—in it we used the sample mean rounded to one decimal place instead of the data) but it also calculates the test statistic (z-score) for the sample mean, the sample mean, and the sample standard deviation. μ > 275 is the alternative hypothesis. Do this set of instructions again except arrow to Draw (instead of Calculate). Press ENTER. A shaded graph appears with z = 1.112 (test statistic) and p = 0.1331 (p-value). Make sure when you use Draw that no other equations are highlighted in Y = and the plots are turned off.

Hypothesis Tests for a Single Population Mean, Population Standard Deviation Unknown

Example

Statistics students believe that the mean score on the first statistics test is 65. A statistics instructor thinks the mean score is higher than 65. He samples ten statistics students and obtains the scores 65657067666363687271. He performs a hypothesis test using a 5% level of significance. The data are assumed to be from a normal distribution.

Solution
Set up the hypothesis test:

A 5% level of significance means that $\alpha = 0.05$. This is a test of a single population mean.

$H_0: \mu = 65$ $H_a: \mu > 65$

Since the instructor thinks the average score is higher, use a “>”. The “>” means the test is right-tailed.

Determine the distribution needed:

Random variable: $\overline{X}$ = average score on the first statistics test.

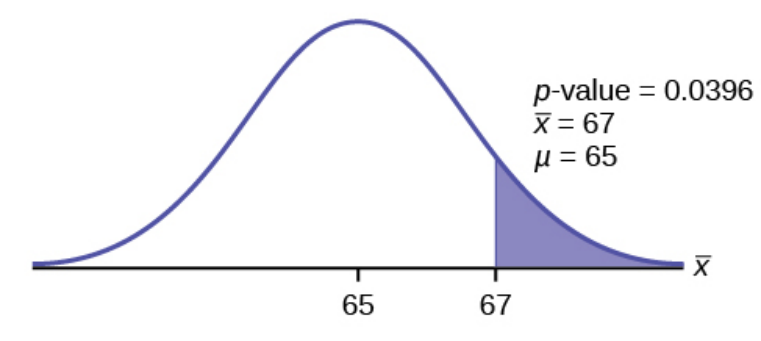
Distribution for the test: If you read the problem carefully, you will notice that there is no population standard deviation given. You are only given n = 10 sample data values. Notice also that the data come from a normal distribution. This means that the distribution for the test is a student’s t.

Use tdf. Therefore, the distribution for the test is t9 where n = 10 and df = 10 – 1 = 9.

Calculate the p-value using the Student’s t-distribution:

p-value = $P(\overline{x} > 67) = 0.0396$ where the sample mean and sample standard deviation are calculated as 67 and 3.1972 from the data.

Interpretation of the p-value: If the null hypothesis is true, then there is a 0.0396 probability (3.96%) that the sample mean is 65 or more.



Compare **α** and the p-value:

Since **α** = 0.05 and p-value = 0.0396. **α** > p-value.

Make a decision: Since **α** > p-value, reject H_0.

This means you reject μ = 65. In other words, you believe the average test score is more than 65.

Conclusion: At a 5% level of significance, the sample data show sufficient evidence that the mean (average) test score is more than 65, just as the math instructor thinks.

The p-value can easily be calculated.

Put the data into a list. Press STAT and arrow over to TESTS. Press 2:T-Test. Arrow over to Data and press ENTER. Arrow down and enter 65 for μ0, the name of the list where you put the data, and 1 for Freq:. Arrow down to μ: and arrow over to > μ0. Press ENTER. Arrow down to Calculate and press ENTER. The calculator not only calculates the p-value (p = 0.0396) but it also calculates the test statistic (t-score) for the sample mean, the sample mean, and the sample standard deviation. μ > 65 is the alternative hypothesis. Do this set of instructions again except arrow to Draw (instead of Calculate). Press ENTER. A shaded graph appears with t = 1.9781 (test statistic) and p = 0.0396 (p-value). Make sure when you use Draw that no other equations are highlighted in Y = and the plots are turned off.

Your Turn!

It is believed that a stock price for a particular company will grow at a rate of $ 5 per week with a standard deviation of $ 1. An investor believes the stock won't grow as quickly. The changes in

stock price is recorded for ten weeks and are as follows: $4, $3, $2, $3, $1, $7, $2, $1, $1, $2. Perform a hypothesis test using a 5% level of significance. State the null and alternative hypotheses, find the *p*-value, state your conclusion, and identify the Type I and Type II errors.

Solution

H_0: $\mu = 5$

H_a: $\mu < 5$

$p = 0.0082$

Because $p < \alpha$, we reject the null hypothesis. There is sufficient evidence to suggest that the stock price of the company grows at a rate less than $5 a week.

Type I Error: To conclude that the stock price is growing slower than $5 a week when, in fact, the stock price is growing at $5 a week (reject the null hypothesis when the null hypothesis is true).

Type II Error: To conclude that the stock price is growing at a rate of $5 a week when, in fact, the stock price is growing slower than $5 a week (do not reject the null hypothesis when the null hypothesis is false).

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=893#oembed-3

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=893#oembed-4

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=893#oembed-5

Hypothesis Tests for a Single Population Proportion

Example

Joon believes that 50% of first-time brides in the United States are younger than their grooms. She performs a hypothesis test to determine if the percentage is the same or different from 50%. Joon samples 100 first-time brides and 53 reply that they are younger than their grooms. For the hypothesis test, she uses a 1% level of
significance.

Solution
Set up the hypothesis test:

The 1% level of significance means that $\alpha = 0.01$. This is a test of a single population proportion.

$H_0: p = 0.50$ $H_a: p \neq 0.50$

The words "is the same or different from" tell you this is a two-tailed test.

Calculate the distribution needed:

Random variable: P' = the percentage of first-time brides who are younger than their grooms.

Distribution for the test: The problem contains no mention of a mean. The
information is given in terms of percentages. Use the distribution for P', the estimated proportion.

$P' \ N\left(p, \sqrt{\frac{p \cdot q}{n}}\right)$ Therefore, $P' \ N\left(0.5, \sqrt{\frac{0.5 \cdot 0.5}{100}}\right)$

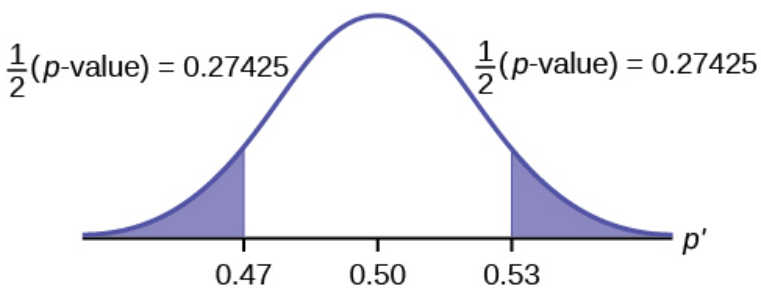
where $p = 0.50$, $q = 1-p = 0.50$, and $n = 100$

Calculate the p-value using the normal distribution for proportions:

p-value = P (p' < 0.47 or p' > 0.53) = 0.5485

where x = 53, p' = $\frac{x}{n} = \frac{53}{100}$ = 0.53.

Interpretation of the p-value: If the null hypothesis is true, there is 0.5485 probability (54.85%) that the sample (estimated) proportion p' is 0.53 or more OR 0.47 or less (see the graph in [link]).



μ = p = 0.50 comes from H_0, the null hypothesis.

p' = 0.53. Since the curve is symmetrical and the test is two-tailed, the p' for the left tail is equal to 0.50 – 0.03 = 0.47 where μ = p = 0.50. (0.03 is the difference between 0.53 and 0.50.)

Compare **α** and the p-value:

Since **α** = 0.01 and p-value = 0.5485. **α** < p-value.

Make a decision: Since **α** < p-value, you cannot reject H_0.

Conclusion: At the 1% level of significance, the sample data do not show sufficient evidence that the percentage of first-time brides who are younger than their grooms is different from 50%.

The p-value can easily be calculated.

Press STAT and arrow over to TESTS. Press 5:1-PropZTest. Enter .5 for p0, 53 for x and 100 for n. Arrow down to Prop and arrow to not equals p0. Press ENTER. Arrow down to Calculate and press ENTER. The calculator calculates the p-value (p = 0.5485) and the test statistic (z-score). Prop not equals .5 is the alternate hypothesis. Do this set of instructions again except arrow to Draw (instead of Calculate). Press ENTER. A shaded graph appears with z = 0.6 (test statistic) and p = 0.5485 (p-value). Make sure when you use Draw that no other equations are highlighted in Y = and the plots are turned off.

The Type I and Type II errors are as follows:

The Type I error is to conclude that the proportion of first-time brides who are younger than their

grooms is different from 50% when, in fact, the proportion is actually 50%. (Reject the null hypothesis when the null hypothesis is true).

The Type II error is there is not enough evidence to conclude that the proportion of first time brides who are younger than their grooms differs from 50% when, in fact, the proportion does differ from 50%. (Do not reject the null hypothesis when the null hypothesis is false.)

Your Turn!

A teacher believes that 85% of students in the class will want to go on a field trip to the local zoo. She performs a hypothesis test to determine if the percentage is the same or different from 85%. The teacher samples 50 students and 39 reply that they would want to go to the zoo. For the hypothesis test, use a 1% level of significance.

First, determine what type of test this is, set up the hypothesis test, find the *p*-value, sketch the graph, and state your conclusion.

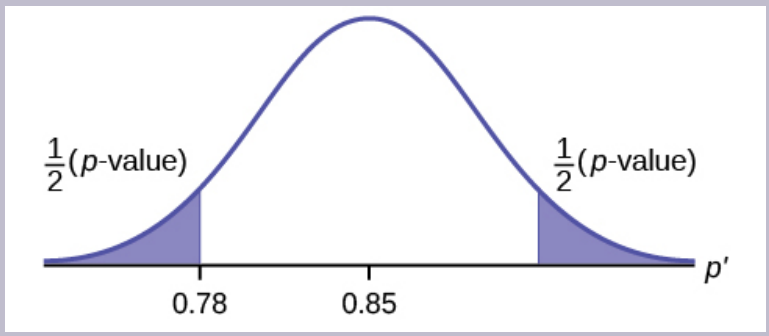
Solution

Since the problem is about percentages, this is a test of single population proportions.

$H_0 : p = 0.85$

$H_a: p \neq 0.85$

$p = 0.7554$



Because $p > \alpha$, we fail to reject the null hypothesis. There is not sufficient evidence to suggest that the proportion of students that want to go to the zoo is not 85%.

Example

Suppose a consumer group suspects that the proportion of households that have three cell phones is 30%. A cell phone company has reason to believe that the proportion is not 30%. Before they start a big advertising campaign, they conduct a hypothesis test. Their marketing people survey 150 households with the result that 43 of the households have three cell phones.

Solution
Set up the Hypothesis Test:

H_0: p = 0.30 H_a: p ≠ 0.30

Determine the distribution needed:

The random variable is P' = proportion of households that have three cell phones.

The distribution for the hypothesis test is $P' \ N\left(0.30, \sqrt{\frac{(0.30)\cdot(0.70)}{150}}\right)$

a. The value that helps determine the p-value is p'. Calculate p'.

Solution
a. p' = $\frac{x}{n}$ where x is the number of successes and n is the total number in the sample.

x = 43, n = 150

p' = $\frac{43}{150}$

b. What is a success for this problem?

Solution
b. A success is having three cell phones in a household.

c. What is the level of significance?

Solution
c. The level of significance is the preset α. Since α is not given, assume that α = 0.05.

d. Draw the graph for this problem. Draw the horizontal axis. Label and shade appropriately.

Calculate the *p*-value.

Solution
d. *p*-value = 0.7216

e. Make a decision. _____(Reject/Do not reject) H_0 because_____.

Solution
e. Assuming that $\alpha = 0.05$, $\alpha < p$-value. The decision is do not reject H_0 because there is not sufficient evidence to conclude that the proportion of households that have three cell phones is not 30%.

Your Turn!

Marketers believe that 92% of adults in the United States own a cell phone. A cell phone manufacturer believes that number is actually lower. 200 American adults are surveyed, of which, 174 report having cell phones. Use a 5% level of significance. State the null and alternative hypothesis, find the *p*-value, state your conclusion, and identify the Type I and Type II errors.

Solution
H_0: $p = 0.92$

H_a: $p < 0.92$

p-value = 0.0046

Because $p < 0.05$, we reject the null hypothesis. There is sufficient evidence to conclude that fewer than 92% of American adults own cell phones.

Type I Error: To conclude that fewer than 92% of American adults own cell phones when, in fact, 92% of American adults do own cell phones (reject the null hypothesis when the null hypothesis is true).

Type II Error: To conclude that 92% of American adults own cell phones when, in fact, fewer than 92% of American adults own cell phones (do not reject the null hypothesis when the null hypothesis is false).

The next example is a poem written by a statistics student named Nicole Hart. The solution to the problem follows the poem. Notice that the hypothesis test is for a single population proportion. This means that the null and alternate hypotheses use the parameter *p*. The distribution for the test is normal. The estimated proportion *p*′ is the proportion of fleas killed to the total fleas found on Fido. This is sample information. The problem gives a preconceived $\alpha = 0.01$, for comparison, and a 95% confidence interval computation. The poem is clever and humorous, so please enjoy it!

Example

My dog has so many fleas, they do not come off with ease. As for shampoo, I have tried many types, even one called Bubble Hype, which only killed 25% of the fleas, unfortunately I was not pleased. I've used all kinds of soap, until I had given up hope, one day I saw an ad that put me in awe. A shampoo used for dogs called GOOD ENOUGH to Clean a Hog guaranteed to kill more fleas. I gave Fido a bath and after doing the math his number of fleas started dropping by 3's! Before his shampoo I counted 42. At the end of his bath, I redid the math and the new shampoo had killed 17 fleas. So now I am pleased.

Now it is time for you to have some fun with the level of significance being .01, you must help me figure out use the new shampoo or go without?

Solution
Set up the hypothesis test:

H_0: $p \leq 0.25$ H_a: $p > 0.25$

Determine the distribution needed:

In words, CLEARLY state what your random variable $\overline{X}$ or P′ represents.

P′ = The proportion of fleas that are killed by the new shampoo

State the distribution to use for the test.

Normal: $N\left(0.25, \sqrt{\frac{(0.25)(1-0.25)}{42}}\right)$

Test Statistic: z = 2.3163

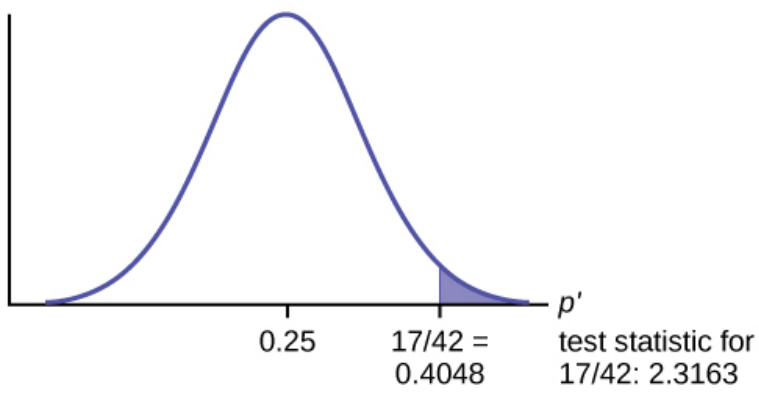
Calculate the p-value using the normal distribution for proportions:

p-value = 0.0103

In one to two complete sentences, explain what the p-value means for this problem.

If the null hypothesis is true (the proportion is 0.25), then there is a 0.0103 probability that the sample (estimated) proportion is 0.4048 $\left(\frac{17}{42}\right)$ or more.

Use the previous information to sketch a picture of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the p-value.



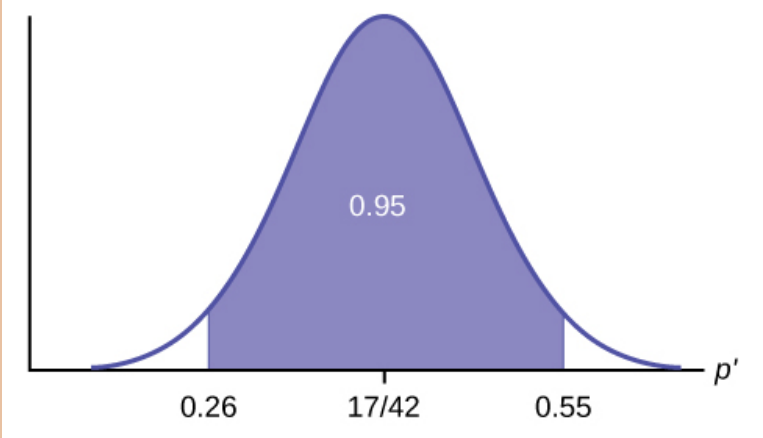
Compare α and the p-value:

Indicate the correct decision ("reject" or "do not reject" the null hypothesis), the reason for it, and write an appropriate conclusion, using complete sentences.

alpha	decision	reason for decision
0.01	Do not reject H_0	$\alpha <$ p-value

Conclusion: At the 1% level of significance, the sample data do not show sufficient evidence that the percentage of fleas that are killed by the new shampoo is more than 25%.

Construct a 95% confidence interval for the true mean or proportion. Include a sketch of the graph of the situation.
Label the point estimate and the lower and upper bounds of the confidence interval.



Confidence Interval: (0.26,0.55) We are 95% confident that the true population proportion p of fleas that are killed by the new shampoo is between 26% and 55%.

Note

This test result is not very definitive since the p-value is very close to alpha. In reality, one would probably do more tests by giving the dog another bath after the fleas have had a chance to return.

Example

The National Institute of Standards and Technology provides exact data on conductivity properties of materials. Following are conductivity measurements for 11 randomly selected pieces of a particular type of glass.

1.11; 1.07; 1.11; 1.07; 1.12; 1.08; .98; .98 1.02; .95; .95

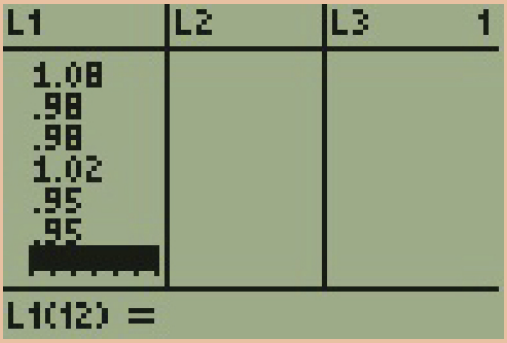
Is there convincing evidence that the average conductivity of this type of glass is greater than one? Use a significance level of 0.05. Assume the population is normal.

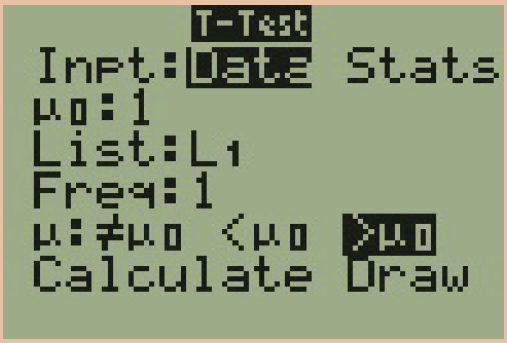
Solution
Let's follow a four-step process to answer this statistical question.

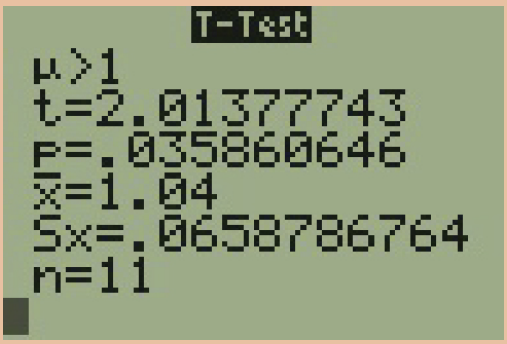
State the Question: We need to determine if, at a 0.05 significance level, the average conductivity of the selected glass is greater than one. Our hypotheses will be

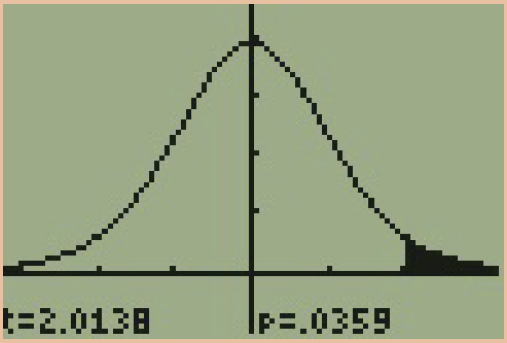
H_0: $\mu \leq 1$
H_a: $\mu > 1$
Plan: We are testing a sample mean without a known population standard deviation. Therefore, we need to use a Student's-t distribution. Assume the underlying population is normal.

Do the calculations: We will input the sample data into the TI-83 as follows.









State the Conclusions: Since the p-value* (p = 0.036) is less than our alpha value, we will reject the null hypothesis. It is reasonable to state that the data supports the claim that the average conductivity level is greater than one.

Example

In a study of 420,019 cell phone users, 172 of the subjects developed brain cancer. Test the claim that cell phone users developed brain cancer at a greater rate than that for non-cell phone users (the rate of brain cancer for non-cell phone users is 0.0340%). Since this is a critical issue, use a 0.005 significance level. Explain why the significance level should be so low in terms of a Type I error.

Solution
We will follow the four-step process.

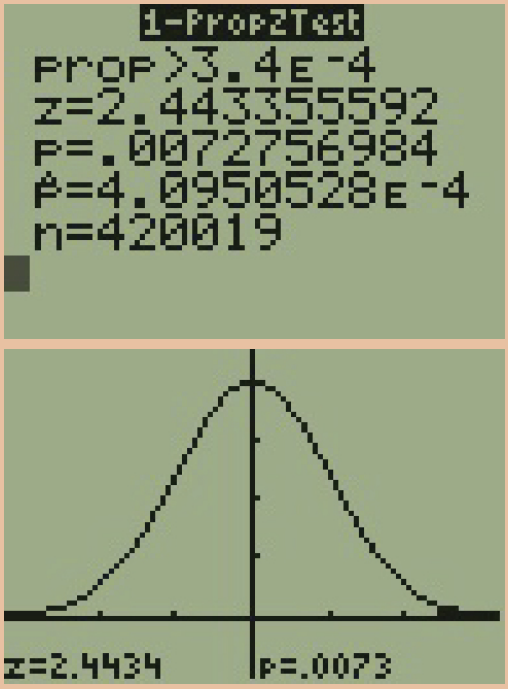
1. We need to conduct a hypothesis test on the claimed cancer rate. Our hypotheses will be

    1. H_0: $p \leq 0.00034$
    2. H_a: $p > 0.00034$

    If we commit a Type I error, we are essentially accepting a false claim. Since the claim describes cancer-causing environments, we want to minimize the chances of incorrectly identifying causes of cancer.
2. We will be testing a sample proportion with x = 172 and n = 420,019. The sample is sufficiently large because we have np = 420,019(0.00034) = 142.8, nq = 420,019(0.99966) =

419,876.2, two independent outcomes, and a fixed probability of success p = 0.00034. Thus we will be able to generalize our results to the population.

3. The associated TI results are



4. Since the p-value = 0.0073 is greater than our alpha value = 0.005, we cannot reject the null. Therefore, we conclude that there is not enough evidence to support the claim of higher brain cancer rates for the cell phone users.

Your Turn

According to the US Census there are approximately 268,608,618 residents aged 12 and older. Statistics from the Rape, Abuse, and Incest National Network indicate that, on average, 207,754 rapes occur each year (male and female) for persons aged 12 and older. This translates into a percentage of sexual assaults of 0.078%. In Daviess County, KY, there were reported 11 rapes for a population of 37,937. Conduct an appropriate hypothesis test to determine if there is a statistically

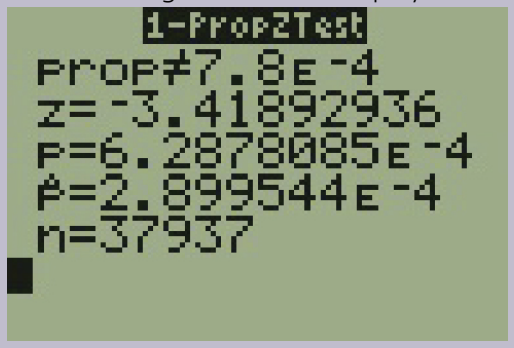
significant difference between the local sexual assault percentage and the national sexual assault percentage. Use a significance level of 0.01.

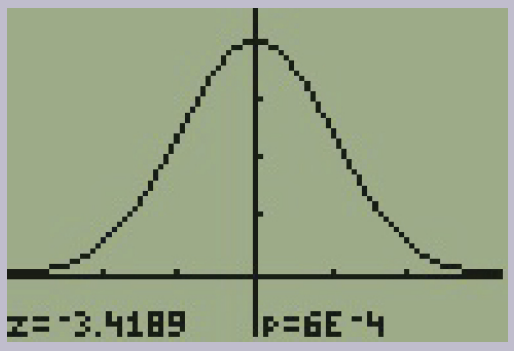
Solution
We will follow the four-step plan.

1. We need to test whether the proportion of sexual assaults in Daviess County, KY is significantly different from the national average.
2. Since we are presented with proportions, we will use a one-proportion z-test. The hypotheses for the test will be

    1. H_0: p = 0.00078
    2. H_a: p ≠ 0.00078

3. The following screenshots display the summary statistics from the hypothesis test.





4. Since the p-value, p = 0.00063, is less than the alpha level of 0.01, the sample data indicates that we should reject the null hypothesis. In conclusion, the sample data support the claim that the proportion of sexual assaults in Daviess County, Kentucky is different from the national average proportion.



One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=893#oembed-6

References

Data from Amit Schitai. Director of Instructional Technology and Distance Learning. LBCC.

Data from *Bloomberg Businessweek*. Available online at http://www.businessweek.com/news/2011-09-15/nyc-smoking-rate-falls-to-record-low-of-14-bloomberg-says.html.

Data from energy.gov. Available online at http://energy.gov (accessed June 27. 2013).

Data from Gallup®. Available online at www.gallup.com (accessed June 27, 2013).

Data from *Growing by Degrees* by Allen and Seaman.

Data from La Leche League International. Available online at http://www.lalecheleague.org/Law/BAFeb01.html.

Data from the American Automobile Association. Available online at www.aaa.com (accessed June 27, 2013).

Data from the American Library Association. Available online at www.ala.org (accessed June 27, 2013).

Data from the Bureau of Labor Statistics. Available online at http://www.bls.gov/oes/current/oes291111.htm.

Data from the Centers for Disease Control and Prevention. Available online at www.cdc.gov (accessed June 27, 2013)

Data from the U.S. Census Bureau, available online at http://quickfacts.census.gov/qfd/states/00000.html (accessed June 27, 2013).

Data from the United States Census Bureau. Available online at http://www.census.gov/hhes/socdemo/language/.

Data from Toastmasters International. Available online at http://toastmasters.org/artisan/detail.asp?CategoryID=1&SubCategoryID=10&ArticleID=429&Page=1.

Data from Weather Underground. Available online at www.wunderground.com (accessed June 27, 2013).

Federal Bureau of Investigations. "Uniform Crime Reports and Index of Crime in Daviess in the State of Kentucky enforced by Daviess County from 1985 to 2005." Available online at http://www.disastercenter.com/kentucky/crime/3868.htm (accessed June 27, 2013).

"Foothill-De Anza Community College District." De Anza College, Winter 2006. Available online at http://research.fhda.edu/factbook/DAdemofs/Fact_sheet_da_2006w.pdf.

Johansen, C., J. Boice, Jr., J. McLaughlin, J. Olsen. "Cellular Telephones and Cancer—a Nationwide Cohort Study in Denmark." Institute of Cancer Epidemiology and the Danish Cancer Society, 93(3):203-7. Available online at http://www.ncbi.nlm.nih.gov/pubmed/11158188 (accessed June 27, 2013).

Rape, Abuse & Incest National Network. "How often does sexual assault occur?" RAINN, 2009. Available online at http://www.rainn.org/get-information/statistics/frequency-of-sexual-assault (accessed June 27, 2013).

CHAPTER 8 REVIEW

8.1 Chapter Review

In a **hypothesis test**, sample data is evaluated in order to arrive at a decision about some type of claim. If certain conditions about the sample are satisfied, then the claim can be evaluated for a population. In a hypothesis test, we:

Evaluate the **null hypothesis**, typically denoted with H_0. The null is not rejected unless the hypothesis test shows otherwise. The null statement must always contain some form of equality (=, ≤ or ≥)

Always write the **alternative hypothesis**, typically denoted with H_a or H_1, using less than, greater than, or not equal symbols, i.e., (≠, >, or <).

If we reject the null hypothesis, then we can assume there is enough evidence to support the alternative hypothesis.

Never state that a claim is proven true or false. Keep in mind the underlying fact that hypothesis testing is based on probability laws; therefore, we can talk only in terms of non-absolute certainties.

Formula Review

H_0 and H_a are contradictory.

If H_0 has:	equal (=)	greater than or equal to (≥)	less than or equal to (≤)
then H_a has:	not equal (≠) **or** greater than (>) **or** less than (<)	less than (<)	greater than (>)

If $\alpha \leq p$-value, then do not reject H_0.

If $\alpha > p$-value, then reject H_0.

α is preconceived. Its value is set before the hypothesis test starts. The p-value is calculated from the data.

8.2 Chapter Review

In every hypothesis test, the outcomes are dependent on a correct interpretation of the data. Incorrect

calculations or misunderstood summary statistics can yield errors that affect the results. A **Type I** error occurs when a true null hypothesis is rejected. A **Type II error** occurs when a false null hypothesis is not rejected.

The probabilities of these errors are denoted by the Greek letters α and β, for a Type I and a Type II error respectively. The power of the test, $1 - \beta$, quantifies the likelihood that a test will yield the correct result of a true alternative hypothesis being accepted. A high power is desirable.

Formula Review

α = probability of a Type I error = P(Type I error) = probability of rejecting the null hypothesis when the null hypothesis is true.

β = probability of a Type II error = P(Type II error) = probability of not rejecting the null hypothesis when the null hypothesis is false.

8.3 Chapter Review

In order for a hypothesis test's results to be generalized to a population, certain requirements must be satisfied.

When testing for a single population mean:

1. A Student's t-test should be used if the data come from a simple, random sample and the population is approximately normally distributed, or the sample size is large, with an unknown standard deviation.
2. The normal test will work if the data come from a simple, random sample and the population is approximately normally distributed, or the sample size is large, with a known standard deviation.

When testing a single population proportion use a normal test for a single population proportion if the data comes from a simple, random sample, fill the requirements for a binomial distribution, and the mean number of success and the mean number of failures satisfy the conditions: $np > 5$ and $nq > n$ where n is the sample size, p is the probability of a success, and q is the probability of a failure.

Formula Review

If there is no given preconceived α, then use $\alpha = 0.05$.
Types of Hypothesis Tests

- Single population mean, **known** population variance (or standard deviation): **Normal test**.

- Single population mean, **unknown** population variance (or standard deviation): **Student's *t*-test**.
- Single population proportion: **Normal test**.
- For a **single population mean**, we may use a normal distribution with the following mean and standard deviation. Means: $\mu = \mu_{\overline{x}}$ and $\sigma_{\overline{x}} = \frac{\sigma_x}{\sqrt{n}}$
- A **single population proportion**, we may use a normal distribution with the following mean and standard deviation. Proportions: $\mu = p$ and $\sigma = \sqrt{\frac{pq}{n}}$.

8.4 Chapter Review

When the probability of an event occurring is low, and it happens, it is called a rare event. Rare events are important to consider in hypothesis testing because they can inform your willingness not to reject or to reject a null hypothesis. To test a null hypothesis, find the *p*-value for the sample data and graph the results. When deciding whether or not to reject the null the hypothesis, keep these two parameters in mind:
$\alpha > p$-value, reject the null hypothesis
$\alpha \leq p$-value, do not reject the null hypothesis

8.5 Chapter Review

The hypothesis test itself has an established process. This can be summarized as follows:
Determine H_0 and H_a. Remember, they are contradictory.
Determine the random variable.
Determine the distribution for the test.
Draw a graph, calculate the test statistic, and use the test statistic to calculate the *p*-value. (A *z*-score and a *t*-score are examples of test statistics.)
Compare the preconceived α with the *p*-value, make a decision (reject or do not reject H_0), and write a clear conclusion using English sentences.

Notice that in performing the hypothesis test, you use α and not β. β is needed to help determine the sample size of the data that is used in calculating the *p*-value. Remember that the quantity $1 - \beta$ is called the **Power of the Test**. A high power is desirable. If the power is too low, statisticians typically increase the sample size while keeping α the same. If the power is low, the null hypothesis might not be rejected when it should be.

CHAPTER 8 PRACTICE

8.1 Practice

You are testing that the mean speed of your cable Internet connection is more than three Megabits per second. What is the random variable? Describe in words.
Solution

The random variable is the mean Internet speed in Megabits per second.

You are testing that the mean speed of your cable Internet connection is more than three Megabits per second. State the null and alternative hypotheses.

The American family has an average of two children. What is the random variable? Describe in words.
Solution

The random variable is the mean number of children an American family has.

The mean entry level salary of an employee at a company is $58,000. You believe it is higher for IT professionals in the company. State the null and alternative hypotheses.

A sociologist claims the probability that a person picked at random in Times Square in New York City is visiting the area is 0.83. You want to test to see if the proportion is actually less. What is the random variable? Describe in words.
Solution

The random variable is the proportion of people picked at random in Times Square visiting the city.

A sociologist claims the probability that a person picked at random in Times Square in New York City is visiting the area is 0.83. You want to test to see if the claim is correct. State the null and alternative hypotheses.

In a population of fish, approximately 42% are female. A test is conducted to see if, in fact, the proportion is less. State the null and alternative hypotheses.
Solution
H_0: $p = 0.42$
H_a: $p < 0.42$

Suppose that a recent article stated that the mean time spent in jail by a first–time convicted burglar is 2.5 years. A study was then done to see if the mean time has increased in the new century. A random sample of 26 first-time convicted burglars in a recent year was picked. The mean length of time in jail from the survey was 3 years with a standard deviation of 1.8 years. Suppose that it is somehow known that the population standard deviation is 1.5. If you were conducting a hypothesis test to determine if the mean length of jail time has increased, what would the null and alternative hypotheses be? The distribution of the population is normal.
H_0: _____
H_a: _____

A random survey of 75 death row inmates revealed that the mean length of time on death row is 17.4 years with a standard deviation of 6.3 years. If you were conducting a hypothesis test to determine if the population mean time on death row could likely be 15 years, what would the null and alternative hypotheses be?

1. H_0: _____
2. H_a: _____

Solution

1. H_0: $\mu = 15$
2. H_a: $\mu \neq 15$

The National Institute of Mental Health published an article stating that in any one-year period, approximately 9.5 percent of American adults suffer from depression or a depressive illness. Suppose that in a survey of 100 people in a certain town, seven of them suffered from depression or a depressive illness. If you were conducting a hypothesis test to determine if the true proportion of people in that town suffering from depression or a depressive illness is lower than the percentage in the general adult American population, what would the null and alternative hypotheses be?
H_0: _____
H_a: _____

8.2 Practice

The mean price of mid-sized cars in a region is $32,000. A test is conducted to see if the claim is true. State the Type I and Type II errors in complete sentences.
Solution

Type I: The mean price of mid-sized cars is $32,000, but we conclude that it is not $32,000.

Type II: The mean price of mid-sized cars is not $32,000, but we conclude that it is $32,000.

A sleeping bag is tested to withstand temperatures of –15 °F. You think the bag cannot stand temperatures that low. State the Type I and Type II errors in complete sentences.

For Exercise 9.12, what are α and β in words?
Solution

α = the probability that you think the bag cannot withstand -15 degrees F, when in fact it can
β = the probability that you think the bag can withstand -15 degrees F, when in fact it cannot

In words, describe 1 – β For Exercise 9.12.

A group of doctors is deciding whether or not to perform an operation. Suppose the null hypothesis, H_0, is: the surgical procedure will go well. State the Type I and Type II errors in complete sentences.
Solution

Type I: The procedure will go well, but the doctors think it will not.
Type II: The procedure will not go well, but the doctors think it will.

A group of doctors is deciding whether or not to perform an operation. Suppose the null hypothesis, H_0, is: the surgical procedure will go well. Which is the error with the greater consequence?

The power of a test is 0.981. What is the probability of a Type II error?
Solution

0.019

A group of divers is exploring an old sunken ship. Suppose the null hypothesis, H_0, is: the sunken ship does not contain buried treasure. State the Type I and Type II errors in complete sentences.

A microbiologist is testing a water sample for E-coli. Suppose the null hypothesis, H_0, is: the sample does not contain E-coli. The probability that the sample does not contain E-coli, but the microbiologist thinks it does is 0.012. The probability that the sample does contain E-coli, but the microbiologist thinks it does not is 0.002. What is the power of this test?
Solution

0.998

A microbiologist is testing a water sample for E-coli. Suppose the null hypothesis, H_0, is: the sample contains E-coli. Which is the error with the greater consequence?

8.3 Practice

Which two distributions can you use for hypothesis testing for this chapter?
Solution

A normal distribution or a Student's *t*-distribution

Which distribution do you use when you are testing a population mean and the standard deviation is known? Assume sample size is large.

Which distribution do you use when the standard deviation is not known and you are testing one population mean? Assume sample size is large.
Solution

Use a Student's *t*-distribution

A population mean is 13. The sample mean is 12.8, and the sample standard deviation is two. The sample size is 20. What distribution should you use to perform a hypothesis test? Assume the underlying population is normal.

A population has a mean of 25 and a standard deviation of five. The sample mean is 24, and the sample size is 108. What distribution should you use to perform a hypothesis test?
Solution

a normal distribution for a single population mean

It is thought that 42% of respondents in a taste test would prefer Brand *A*. In a particular test of 100 people, 39% preferred Brand *A*. What distribution should you use to perform a hypothesis test?

You are performing a hypothesis test of a single population mean using a Student's *t*-distribution. What must you assume about the distribution of the data?
Solution

It must be approximately normally distributed.

You are performing a hypothesis test of a single population mean using a Student's *t*-distribution. The data are not from a simple random sample. Can you accurately perform the hypothesis test?

You are performing a hypothesis test of a single population proportion. What must be true about the quantities of np and nq?
Solution

They must both be greater than five.

You are performing a hypothesis test of a single population proportion. You find out that np is less than five. What must you do to be able to perform a valid hypothesis test?

You are performing a hypothesis test of a single population proportion. The data come from which distribution?
Solution

binomial distribution

8.4 Practice

When do you reject the null hypothesis?

The probability of winning the grand prize at a particular carnival game is 0.005. Is the outcome of winning very likely or very unlikely?
Solution

The outcome of winning is very unlikely.

The probability of winning the grand prize at a particular carnival game is 0.005. Michele wins the grand prize. Is this considered a rare or common event? Why?

It is believed that the mean height of high school students who play basketball on the school team is 73 inches with a standard deviation of 1.8 inches. A random sample of 40 players is chosen. The sample mean was 71 inches, and the sample standard deviation was 1.5 years. Do the data support the claim that the mean height is less than 73 inches? The *p*-value is almost zero. State the null and alternative hypotheses and interpret the *p*-value.
Solution

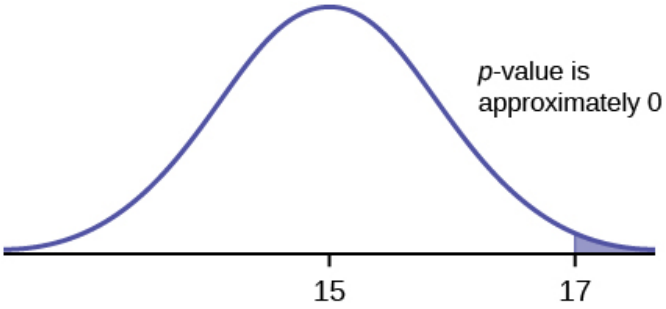
H_0: $\mu >= 73$

H_a: $\mu < 73$

The p-value is almost zero, which means there is sufficient data to conclude that the mean height of high school students who play basketball on the school team is less than 73 inches at the 5% level. The data do support the claim.

The mean age of graduate students at a University is at most 31 years with a standard deviation of two years. A random sample of 15 graduate students is taken. The sample mean is 32 years and the sample standard deviation is three years. Are the data significant at the 1% level? The p-value is 0.0264. State the null and alternative hypotheses and interpret the p-value.

Does the shaded region represent a low or a high p-value compared to a level of significance of 1%?



Solution

The shaded region shows a low p-value.

What should you do when $\alpha > p$-value?

What should you do if $\alpha = p$-value?
Solution

Do not reject H_0.

If you do not reject the null hypothesis, then it must be true. Is this statement correct? State why or why not in complete sentences.

Use the following information to answer the next seven exercises: Suppose that a recent article stated that the mean time spent in jail by a first-time convicted burglar is 2.5 years. A study was then done to see if the mean time has increased in the new century. A random sample of 26 first-time convicted burglars in a recent year was picked. The mean length of time in jail from the survey was three years with a standard deviation of 1.8 years. Suppose that it is somehow known that the population standard deviation is 1.5. Conduct a hypothesis test to determine if the mean length of jail time has increased. Assume the distribution of the jail times is approximately normal.

Is this a test of means or proportions?
Solution

means

<!–moved from 9.3, 32-41–>

What symbol represents the random variable for this test?

In words, define the random variable for this test.
Solution

the mean time spent in jail for 26 first time convicted burglars

Is the population standard deviation known and, if so, what is it?

Calculate the following:
$\overline{x}$ _____
σ _____
s_x _____
n _____
Solution

1. 3
2. 1.5
3. 1.8
4. 26

Since both σ and s_x are given, which should be used? In one to two complete sentences, explain why.

State the distribution to use for the hypothesis test.
Solution

$$\overline{X}\ N\left(2.5, \frac{1.5}{\sqrt{26}}\right)$$

A random survey of 75 death row inmates revealed that the mean length of time on death row is 17.4 years with a standard deviation of 6.3 years. Conduct a hypothesis test to determine if the population mean time on death row could likely be 15 years.

1. Is this a test of one mean or proportion?
2. State the null and alternative hypotheses.

   H_0: _____ H_a : _____
3. Is this a right-tailed, left-tailed, or two-tailed test?
4. What symbol represents the random variable for this test?
5. In words, define the random variable for this test.
6. Is the population standard deviation known and, if so, what is it?
7. Calculate the following:
   1. $\overline{x}$ = _____
   2. s = _____
   3. n = _____
8. Which test should be used?
9. State the distribution to use for the hypothesis test.
10. Find the p-value.
11. At a pre-conceived $\alpha = 0.05$, what is your:
    1. Decision:
    2. Reason for the decision:
    3. Conclusion (write out in a complete sentence):

8.5 Practice

Assume H_0: $\mu = 9$ and H_a: $\mu < 9$. Is this a left-tailed, right-tailed, or two-tailed test?
Solution

This is a left-tailed test.

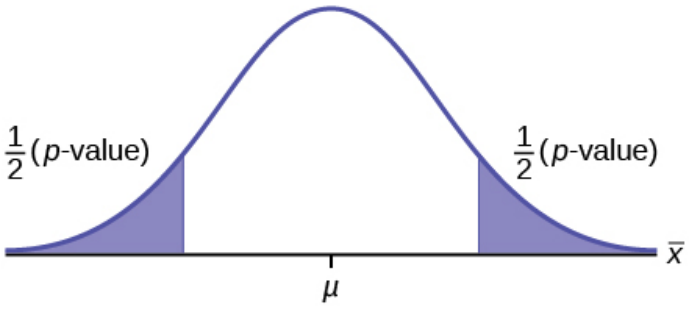
Assume H_0: $\mu \leq 6$ and H_a: $\mu > 6$. Is this a left-tailed, right-tailed, or two-tailed test?

Assume H_0: $p = 0.25$ and H_a: $p \neq 0.25$. Is this a left-tailed, right-tailed, or two-tailed test?
Solution

This is a two-tailed test.

Draw the general graph of a left-tailed test.

Draw the graph of a two-tailed test.
Solution



A bottle of water is labeled as containing 16 fluid ounces of water. You believe it is less than that. What type of test would you use?

Your friend claims that his mean golf score is 63. You want to show that it is higher than that. What type of test would you use?
Solution

a right-tailed test

A bathroom scale claims to be able to identify correctly any weight within a pound. You think that it cannot be that accurate. What type of test would you use?

You flip a coin and record whether it shows heads or tails. You know the probability of getting heads is 50%, but you think it is less for this particular coin. What type of test would you use?
Solution

a left-tailed test

If the alternative hypothesis has a not-equals ($\neq$) symbol, do you know which type of test to use?

Assume the null hypothesis states that the mean is at least 18. Is this a left-tailed, right-tailed, or two-tailed test?
Solution

This is a left-tailed test.

Assume the null hypothesis states that the mean is at most 12. Is this a left-tailed, right-tailed, or two-tailed test?

Assume the null hypothesis states that the mean is equal to 88. The alternative hypothesis states that the mean is not equal to 88. Is this a left-tailed, right-tailed, or two-tailed test?
Solution

This is a two-tailed test.

<!–moved from 9.3–>

CHAPTER 8 HOMEWORK

8.1 Homework

Some of the following statements refer to the null hypothesis, some to the alternate hypothesis.

State the null hypothesis, H_0, and the alternative hypothesis. H_a, in terms of the appropriate parameter (μ or p).

1. The mean number of years Americans work before retiring is 34.
2. At most 60% of Americans vote in presidential elections.
3. The mean starting salary for San Jose State University graduates is at least $ 100,000 per year.
4. Twenty-nine percent of high school seniors get drunk each month.
5. Fewer than 5% of adults ride the bus to work in Los Angeles.
6. The mean number of cars a person owns in her lifetime is not more than ten.
7. About half of Americans prefer to live away from cities, given the choice.
8. Europeans have a mean paid vacation each year of six weeks.
9. The chance of developing breast cancer is under 11% for women.
10. Private universities' mean tuition cost is more than $ 20,000 per year.

Solution

1. $H_0: \mu = 34$; $H_a: \mu \neq 34$
2. $H_0: p \leq 0.60$; $H_a: p > 0.60$
3. $H_0: \mu \geq 100{,}000$; $H_a: \mu < 100{,}000$
4. $H_0: p = 0.29$; $H_a: p \neq 0.29$
5. $H_0: p = 0.05$; $H_a: p < 0.05$
6. $H_0: \mu \leq 10$; $H_a: \mu > 10$
7. $H_0: p = 0.50$; $H_a: p \neq 0.50$
8. $H_0: \mu = 6$; $H_a: \mu \neq 6$
9. $H_0: p \geq 0.11$; $H_a: p < 0.11$
10. $H_0: \mu \leq 20{,}000$; $H_a: \mu > 20{,}000$

Over the past few decades, public health officials have examined the link between weight concerns and teen girls' smoking. Researchers surveyed a group of 273 randomly selected teen girls living in Massachusetts

(between 12 and 15 years old). After four years the girls were surveyed again. Sixty-three said they smoked to stay thin. Is there good evidence that more than thirty percent of the teen girls smoke to stay thin? The alternative hypothesis is:

1. $p < 0.30$
2. $p \leq 0.30$
3. $p \geq 0.30$
4. $p > 0.30$

A statistics instructor believes that fewer than 20% of Evergreen Valley College (EVC) students attended the opening night midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 attended the midnight showing. An appropriate alternative hypothesis is:

1. $p = 0.20$
2. $p > 0.20$
3. $p < 0.20$
4. $p \leq 0.20$

Solution

c

Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test. The null and alternative hypotheses are:

1. $H_o: \overline{x} = 4.5, H_a : \overline{x} > 4.5$
2. $H_o: \mu \geq 4.5, H_a: \mu < 4.5$
3. $H_o: \mu = 4.75, H_a: \mu > 4.75$
4. $H_o: \mu = 4.5, H_a: \mu > 4.5$

8.2 Homework

State the Type I and Type II errors in complete sentences given the following statements.

1. The mean number of years Americans work before retiring is 34.

2. At most 60% of Americans vote in presidential elections.
3. The mean starting salary for San Jose State University graduates is at least $100,000 per year.
4. Twenty-nine percent of high school seniors get drunk each month.
5. Fewer than 5% of adults ride the bus to work in Los Angeles.
6. The mean number of cars a person owns in his or her lifetime is not more than ten.
7. About half of Americans prefer to live away from cities, given the choice.
8. Europeans have a mean paid vacation each year of six weeks.
9. The chance of developing breast cancer is under 11% for women.
10. Private universities mean tuition cost is more than $20,000 per year.

Solution

1. Type I error: We conclude that the mean is not 34 years, when it really is 34 years. Type II error: We conclude that the mean is 34 years, when in fact it really is not 34 years.
2. Type I error: We conclude that more than 60% of Americans vote in presidential elections, when the actual percentage is at most 60%.Type II error: We conclude that at most 60% of Americans vote in presidential elections when, in fact, more than 60% do.
3. Type I error: We conclude that the mean starting salary is less than $100,000, when it really is at least $100,000. Type II error: We conclude that the mean starting salary is at least $100,000 when, in fact, it is less than $100,000.
4. Type I error: We conclude that the proportion of high school seniors who get drunk each month is not 29%, when it really is 29%. Type II error: We conclude that the proportion of high school seniors who get drunk each month is 29% when, in fact, it is not 29%.
5. Type I error: We conclude that fewer than 5% of adults ride the bus to work in Los Angeles, when the percentage that do is really 5% or more. Type II error: We conclude that 5% or more adults ride the bus to work in Los Angeles when, in fact, fewer than 5% do.
6. Type I error: We conclude that the mean number of cars a person owns in his or her lifetime is more than 10, when in reality it is not more than 10. Type II error: We conclude that the mean number of cars a person owns in his or her lifetime is not more than 10 when, in fact, it is more than 10.
7. Type I error: We conclude that the proportion of Americans who prefer to live away from cities is not about half, though the actual proportion is about half. Type II error: We conclude that the proportion of Americans who prefer to live away from cities is half when, in fact, it is not half.
8. Type I error: We conclude that the duration of paid vacations each year for Europeans is not six weeks, when in fact it is six weeks. Type II error: We conclude that the duration of paid vacations each year for Europeans is six weeks when, in fact, it is not.
9. Type I error: We conclude that the proportion is less than 11%, when it is really at least 11%. Type II error: We conclude that the proportion of women who develop breast cancer is at least 11%, when in

fact it is less than 11%.

10. Type I error: We conclude that the average tuition cost at private universities is more than $20,000, though in reality it is at most $20,000. Type II error: We conclude that the average tuition cost at private universities is at most $20,000 when, in fact, it is more than $20,000.

For statements a-j in Exercise 9.109, answer the following in complete sentences.

1. State a consequence of committing a Type I error.
2. State a consequence of committing a Type II error.

When a new drug is created, the pharmaceutical company must subject it to testing before receiving the necessary permission from the Food and Drug Administration (FDA) to market the drug. Suppose the null hypothesis is "the drug is unsafe." What is the Type II Error?

1. To conclude the drug is safe when, in fact, it is unsafe.
2. Not to conclude the drug is safe when, in fact, it is safe.
3. To conclude the drug is safe when, in fact, it is safe.
4. Not to conclude the drug is unsafe when, in fact, it is unsafe.

Solution

b

A statistics instructor believes that fewer than 20% of Evergreen Valley College (EVC) students attended the opening midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 of them attended the midnight showing. The Type I error is to conclude that the percent of EVC students who attended is _____.

1. at least 20%, when in fact, it is less than 20%.
2. 20%, when in fact, it is 20%.
3. less than 20%, when in fact, it is at least 20%.
4. less than 20%, when in fact, it is less than 20%.

It is believed that Lake Tahoe Community College (LTCC) Intermediate Algebra students get less than seven hours of sleep per night, on average. A survey of 22 LTCC Intermediate Algebra students generated a mean of 7.24 hours with a standard deviation of 1.93 hours. At a level of significance of 5%, do LTCC Intermediate Algebra students get less than seven hours of sleep per night, on average?

The Type II error is not to reject that the mean number of hours of sleep LTCC students get per night is at least seven when, in fact, the mean number of hours

1. is more than seven hours.
2. is at most seven hours.
3. is at least seven hours.
4. is less than seven hours.

Solution

d

Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test, the Type I error is:

1. to conclude that the current mean hours per week is higher than 4.5, when in fact, it is higher
2. to conclude that the current mean hours per week is higher than 4.5, when in fact, it is the same
3. to conclude that the mean hours per week currently is 4.5, when in fact, it is higher
4. to conclude that the mean hours per week currently is no higher than 4.5, when in fact, it is not higher

8.3 Homework

It is believed that Lake Tahoe Community College (LTCC) Intermediate Algebra students get less than seven hours of sleep per night, on average. A survey of 22 LTCC Intermediate Algebra students generated a mean of 7.24 hours with a standard deviation of 1.93 hours. At a level of significance of 5%, do LTCC Intermediate Algebra students get less than seven hours of sleep per night, on average? The distribution to be used for this test is $\overline{X}$ ~ _____

1. $N\left(7.24, \frac{1.93}{\sqrt{22}}\right)$
2. $N\left(7.24, 1.93\right)$
3. t_{22}
4. t_{21}

Solution

d

8.4 Homework

The National Institute of Mental Health published an article stating that in any one-year period, approximately 9.5 percent of American adults suffer from depression or a depressive illness. Suppose that in a survey of 100 people in a certain town, seven of them suffered from depression or a depressive illness. Conduct a hypothesis test to determine if the true proportion of people in that town suffering from depression or a depressive illness is lower than the percent in the general adult American population.

1. Is this a test of one mean or proportion?
2. State the null and alternative hypotheses.

   H_0: _____ H_a: _____
3. Is this a right-tailed, left-tailed, or two-tailed test?
4. What symbol represents the random variable for this test?
5. In words, define the random variable for this test.
6. Calculate the following:
   1. x = _____
   2. n = _____
   3. p' = _____
7. Calculate σ_x = _____. Show the formula set-up.
8. State the distribution to use for the hypothesis test.
9. Find the p-value.
10. At a pre-conceived $\alpha = 0.05$, what is your:
    1. Decision:
    2. Reason for the decision:
    3. Conclusion (write out in a complete sentence):

8.5 Homework

For each of the word problems, use a solution sheet to do the hypothesis test. The solution sheet is found in [link]. Please feel free to make copies of the solution sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.

Note

If you are using a Student's-*t* distribution for one of the following homework problems, you may assume that the underlying population is normally distributed. (In general, you must first prove that assumption, however.)

A particular brand of tires claims that its deluxe tire averages at least 50,000 miles before it needs to be replaced. From past studies of this tire, the standard deviation is known to be 8,000. A survey of owners of that tire design is conducted. From the 28 tires surveyed, the mean lifespan was 46,500 miles with a standard deviation of 9,800 miles. Using alpha = 0.05, is the data highly inconsistent with the claim?
Solution

1. H_0: $\mu \geq 50{,}000$
2. H_a: $\mu < 50{,}000$
3. Let $\overline{X}$ = the average lifespan of a brand of tires.
4. normal distribution
5. $z = -2.315$
6. p-value = 0.0103
7. Check student's solution.
    1. alpha: 0.05
    2. Decision: Reject the null hypothesis.
    3. Reason for decision: The p-value is less than 0.05.
    4. Conclusion: There is sufficient evidence to conclude that the mean lifespan of the tires is less than 50,000 miles.
8. (43,537, 49,463)

From generation to generation, the mean age when smokers first start to smoke varies. However, the standard deviation of that age remains constant of around 2.1 years. A survey of 40 smokers of this generation was done to see if the mean starting age is at least 19. The sample mean was 18.1 with a sample standard deviation of 1.3. Do the data support the claim at the 5% level?

The cost of a daily newspaper varies from city to city. However, the variation among prices remains steady with a standard deviation of 20¢. A study was done to test the claim that the mean cost of a daily newspaper is $1.00. Twelve costs yield a mean cost of 95¢ with a standard deviation of 18¢. Do the data support the claim at the 1% level?
Solution

1. H_0: $\mu =$ $1.00

2. $H_a: \mu \neq \$ 1.00$
3. Let $\overline{X}$ = the average cost of a daily newspaper.
4. normal distribution
5. $z = -0.866$
6. p-value = 0.3865
7. Check student's solution.
    1. Alpha: 0.01
    2. Decision: Do not reject the null hypothesis.
    3. Reason for decision: The p-value is greater than 0.01.
    4. Conclusion: There is sufficient evidence to support the claim that the mean cost of daily papers is $\$ 1$. The mean cost could be $\$ 1$.
8. ($\$ 0.84$, $\$ 1.06$)

An article in the *San Jose Mercury News* stated that students in the California state university system take 4.5 years, on average, to finish their undergraduate degrees. Suppose you believe that the mean time is longer. You conduct a survey of 49 students and obtain a sample mean of 5.1 with a sample standard deviation of 1.2. Do the data support your claim at the 1% level?

The mean number of sick days an employee takes per year is believed to be about ten. Members of a personnel department do not believe this figure. They randomly survey eight employees. The number of sick days they took for the past year are as follows: 12; 4; 15; 3; 11; 8; 6; 8. Let x = the number of sick days they took for the past year. Should the personnel team believe that the mean number is ten?
Solution

1. $H_0: \mu = 10$
2. $H_a: \mu \neq 10$
3. Let $\overline{X}$ the mean number of sick days an employee takes per year.
4. Student's t-distribution
5. $t = -1.12$
6. p-value = 0.300
7. Check student's solution.
    1. Alpha: 0.05
    2. Decision: Do not reject the null hypothesis.
    3. Reason for decision: The p-value is greater than 0.05.
    4. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that the mean number of sick days is not ten.
8. (4.9443, 11.806)

In 1955, *Life Magazine* reported that the 25 year-old mother of three worked, on average, an 80 hour week. Recently, many groups have been studying whether or not the women's movement has, in fact, resulted in an increase in the average work week for women (combining employment and at-home work). Suppose a study was done to determine if the mean work week has increased. 81 women were surveyed with the following results. The sample mean was 83; the sample standard deviation was ten. Does it appear that the mean work week has increased for women at the 5% level?

Your statistics instructor claims that 60 percent of the students who take her Elementary Statistics class go through life feeling more enriched. For some reason that she can't quite figure out, most people don't believe her. You decide to check this out on your own. You randomly survey 64 of her past Elementary Statistics students and find that 34 feel more enriched as a result of her class. Now, what do you think?
Solution

1. H_0: $p \geq 0.6$
2. H_a: $p < 0.6$
3. Let P' = the proportion of students who feel more enriched as a result of taking Elementary Statistics.
4. normal for a single proportion
5. 1.12
6. p-value = 0.1308
7. Check student's solution.
    1. Alpha: 0.05
    2. Decision: Do not reject the null hypothesis.
    3. Reason for decision: The p-value is greater than 0.05.
    4. Conclusion: There is insufficient evidence to conclude that less than 60 percent of her students feel more enriched.
8. Confidence Interval: (0.409, 0.654)

    The "plus-4s" confidence interval is (0.411, 0.648)

A Nissan Motor Corporation advertisement read, "The average man's I.Q. is 107. The average brown trout's I.Q. is 4. So why can't man catch brown trout?" Suppose you believe that the brown trout's mean I.Q. is greater than four. You catch 12 brown trout. A fish psychologist determines the I.Q.s as follows: 5; 4; 7; 3; 6; 4; 5; 3; 6; 3; 8; 5. Conduct a hypothesis test of your belief.

Refer to Exercise 9.119. Conduct a hypothesis test to see if your decision and conclusion would change if your belief were that the brown trout's mean I.Q. is **not** four.
Solution

1. H_0: $\mu = 4$
2. H_a: $\mu \neq 4$
3. Let $\overline{X}$ the average I.Q. of a set of brown trout.
4. two-tailed Student's t-test
5. $t = 1.95$
6. p-value = 0.076
7. Check student's solution.
    1. Alpha: 0.05
    2. Decision: Reject the null hypothesis.
    3. Reason for decision: The p-value is greater than 0.05
    4. Conclusion: There is insufficient evidence to conclude that the average IQ of brown trout is not four.
8. (3.8865,5.9468)

According to an article in *Newsweek*, the natural ratio of girls to boys is 100:105. In China, the birth ratio is 100: 114 (46.7% girls). Suppose you don't believe the reported figures of the percent of girls born in China. You conduct a study. In this study, you count the number of girls and boys born in 150 randomly chosen recent births. There are 60 girls and 90 boys born of the 150. Based on your study, do you believe that the percentage of girls born in China is 46.7?

A poll done for *Newsweek* found that 13% of Americans have seen or sensed the presence of an angel. A contingent doubts that the percentage is really that high. It conducts its own survey. Out of 76 Americans surveyed, only two had seen or sensed the presence of an angel. As a result of the contingent's survey, would you agree with the *Newsweek* poll? In complete sentences, also give three reasons why the two polls might give different results.
Solution

1. H_0: $p \geq 0.13$
2. H_a: $p < 0.13$
3. Let P' = the proportion of Americans who have seen or sensed angels
4. normal for a single proportion
5. −2.688
6. p-value = 0.0036
7. Check student's solution.
    1. alpha: 0.05
    2. Decision: Reject the null hypothesis.
    3. Reason for decision: The p-value is less than 0.05.
    4. Conclusion: There is sufficient evidence to conclude that the percentage of Americans who have

seen or sensed an angel is less than 13%.

8. (0, 0.0623).

The "plus-4s" confidence interval is (0.0022, 0.0978)

The mean work week for engineers in a start-up company is believed to be about 60 hours. A newly hired engineer hopes that it's shorter. She asks ten engineering friends in start-ups for the lengths of their mean work weeks. Based on the results that follow, should she count on the mean work week to be shorter than 60 hours?

Data (length of mean work week): 70; 45; 55; 60; 65; 55; 55; 60; 50; 55.

<!– LINK –>

Use the "Lap time" data for Lap 4 (see [link]) to test the claim that Terri finishes Lap 4, on average, in less than 129 seconds. Use all twenty races given.

Solution

1. H_0: $\mu \geq 129$
2. H_a: $\mu < 129$
3. Let $\overline{X}$ = the average time in seconds that Terri finishes Lap 4.
4. Student's t-distribution
5. $t = 1.209$
6. 0.8792
7. Check student's solution.
    1. Alpha: 0.05
    2. Decision: Do not reject the null hypothesis.
    3. Reason for decision: The p-value is greater than 0.05.
    4. Conclusion: There is insufficient evidence to conclude that Terri's mean lap time is less than 129 seconds.
8. (128.63, 130.37)

<!– LINK –>

Use the "Initial Public Offering" data (see [link]) to test the claim that the mean offer price was $ 18 per share. Do not use all the data. Use your random number generator to randomly survey 15 prices.

Note

The following questions were written by past students. They are excellent problems!

"Asian Family Reunion," by Chau Nguyen

Every two years it comes around.
We all get together from different towns.
In my honest opinion,
It's not a typical family reunion.
Not forty, or fifty, or sixty,
But how about seventy companions!
The kids would play, scream, and shout
One minute they're happy, another they'll pout.
The teenagers would look, stare, and compare
From how they look to what they wear.
The men would chat about their business
That they make more, but never less.
Money is always their subject
And there's always talk of more new projects.
The women get tired from all of the chats
They head to the kitchen to set out the mats.
Some would sit and some would stand
Eating and talking with plates in their hands.
Then come the games and the songs
And suddenly, everyone gets along!
With all that laughter, it's sad to say
That it always ends in the same old way.
They hug and kiss and say "good-bye"
And then they all begin to cry!
I say that 60 percent shed their tears
But my mom counted 35 people this year.
She said that boys and men will always have their pride,
So we won't ever see them cry.
I myself don't think she's correct,
So could you please try this problem to see if you object?

Solution

1. $H_0: p = 0.60$
2. $H_a: p < 0.60$
3. Let P' = the proportion of family members who shed tears at a reunion.
4. normal for a single proportion

5. −1.71
6. 0.0438
7. Check student's solution.
    1. alpha: 0.05
    2. Decision: Reject the null hypothesis.
    3. Reason for decision: *p*-value < alpha
    4. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the proportion of family members who shed tears at a reunion is less than 0.60. However, the test is weak because the *p*-value and alpha are quite close, so other tests should be done.
8. We are 95% confident that between 38.29% and 61.71% of family members will shed tears at a family reunion. (0.3829, 0.6171). The "plus-4s" confidence interval (see chapter 8) is (0.3861, 0.6139)

Note that here the "large-sample" 1 – PropZTest provides the approximate *p*-value of 0.0438. Whenever a *p*-value based on a normal approximation is close to the level of significance, the exact *p*-value based on binomial probabilities should be calculated whenever possible. This is beyond the scope of this course.

"The Problem with Angels," by Cyndy Dowling

Although this problem is wholly mine,
The catalyst came from the magazine, Time.
On the magazine cover I did find
The realm of angels tickling my mind.

Inside, 69% I found to be
In angels, Americans do believe.

Then, it was time to rise to the task,
Ninety-five high school and college students I did ask.
Viewing all as one group,
Random sampling to get the scoop.

So, I asked each to be true,
"Do you believe in angels?" Tell me, do!

Hypothesizing at the start,
Totally believing in my heart
That the proportion who said yes
Would be equal on this test.

Lo and behold, seventy-three did arrive,

Out of the sample of ninety-five.
Now your job has just begun,
Solve this problem and have some fun.

"Blowing Bubbles," by Sondra Prull

Studying stats just made me tense,
I had to find some sane defense.
Some light and lifting simple play
To float my math anxiety away.

Blowing bubbles lifts me high
Takes my troubles to the sky.
POIK! They're gone, with all my stress
Bubble therapy is the best.

The label said each time I blew
The average number of bubbles would be at least 22.
I blew and blew and this I found
From 64 blows, they all are round!

But the number of bubbles in 64 blows
Varied widely, this I know.
20 per blow became the mean
They deviated by 6, and not 16.

From counting bubbles, I sure did relax
But now I give to you your task.
Was 22 a reasonable guess?
Find the answer and pass this test!

Solution

1. H_0: $\mu \geq 22$
2. H_a: $\mu < 22$
3. Let $\overline{X}$ = the mean number of bubbles per blow.
4. Student's t-distribution
5. -2.667
6. p-value = 0.00486
7. Check student's solution.

1. Alpha: 0.05
2. Decision: Reject the null hypothesis.
3. Reason for decision: The *p*-value is less than 0.05.
4. Conclusion: There is sufficient evidence to conclude that the mean number of bubbles per blow is less than 22.

8. (18.501, 21.499)

"Dalmatian Darnation," by Kathy Sparling

A greedy dog breeder named Spreckles
Bred puppies with numerous freckles
The Dalmatians he sought
Possessed spot upon spot
The more spots, he thought, the more shekels.

His competitors did not agree
That freckles would increase the fee.
They said, "Spots are quite nice
But they don't affect price;
One should breed for improved pedigree."

The breeders decided to prove
This strategy was a wrong move.
Breeding only for spots
Would wreak havoc, they thought.
His theory they want to disprove.

They proposed a contest to Spreckles
Comparing dog prices to freckles.
In records they looked up
One hundred one pups:
Dalmatians that fetched the most shekels.

They asked Mr. Spreckles to name
An average spot count he'd claim
To bring in big bucks.
Said Spreckles, "Well, shucks,
It's for one hundred one that I aim."

Said an amateur statistician

Who wanted to help with this mission.
"Twenty-one for the sample
Standard deviation's ample:

They examined one hundred and one
Dalmatians that fetched a good sum.
They counted each spot,
Mark, freckle and dot
And tallied up every one.

Instead of one hundred one spots
They averaged ninety six dots
Can they muzzle Spreckles'
Obsession with freckles
Based on all the dog data they've got?

"Macaroni and Cheese, please!!" by Nedda Misherghi and Rachelle Hall

As a poor starving student I don't have much money to spend for even the bare necessities. So my favorite and main staple food is macaroni and cheese. It's high in taste and low in cost and nutritional value.

One day, as I sat down to determine the meaning of life, I got a serious craving for this, oh, so important, food of my life. So I went down the street to Greatway to get a box of macaroni and cheese, but it was SO expensive! $ 2.02!!! Can you believe it? It made me stop and think. The world is changing fast. I had thought that the mean cost of a box (the normal size, not some super-gigantic-family-value-pack) was at most $ 1, but now I wasn't so sure. However, I was determined to find out. I went to 53 of the closest grocery stores and surveyed the prices of macaroni and cheese. Here are the data I wrote in my notebook:
Price per box of Mac and Cheese:

- 5 stores @ $ 2.02
- 15 stores @ $ 0.25
- 3 stores @ $ 1.29
- 6 stores @ $ 0.35
- 4 stores @ $ 2.27
- 7 stores @ $ 1.50
- 5 stores @ $ 1.89
- 8 stores @ 0.75.

I could see that the cost varied but I had to sit down to figure out whether or not I was right. If it does turn out that this mouth-watering dish is at most $ 1, then I'll throw a big cheesy party in our next statistics lab, with

enough macaroni and cheese for just me. (After all, as a poor starving student I can't be expected to feed our class of animals!)

Solution

1. $H_0: \mu \leq 1$
2. $H_a: \mu > 1$
3. Let $\overline{X}$ = the mean cost in dollars of macaroni and cheese in a certain town.
4. Student's *t*-distribution
5. $t = 0.340$
6. *p*-value = 0.36756
7. Check student's solution.
    1. Alpha: 0.05
    2. Decision: Do not reject the null hypothesis.
    3. Reason for decision: The *p*-value is greater than 0.05
    4. Conclusion: The mean cost could be $ 1, or less. At the 5% significance level, there is insufficient evidence to conclude that the mean price of a box of macaroni and cheese is more than $ 1.
8. (0.8291, 1.241)

"William Shakespeare: The Tragedy of Hamlet, Prince of Denmark," by Jacqueline Ghodsi

THE CHARACTERS (in order of appearance):

- HAMLET, Prince of Denmark and student of Statistics
- POLONIUS, Hamlet's tutor
- HORATIO, friend to Hamlet and fellow student

Scene: The great library of the castle, in which Hamlet does his lessons

Act I

(The day is fair, but the face of Hamlet is clouded. He paces the large room. His tutor, Polonius, is reprimanding Hamlet regarding the latter's recent experience. Horatio is seated at the large table at right stage.)

POLONIUS: My Lord, how canst thou admit that thou hast seen a ghost! It is but a figment of your imagination!

HAMLET: I beg to differ; I know of a certainty that five-and-seventy in one hundred of us, condemned to the whips and scorns of time as we are, have gazed upon a spirit of health, or goblin damn'd, be their intents wicked or charitable.

POLONIUS If thou doest insist upon thy wretched vision then let me invest your time; be true to thy work and speak to me through the reason of the null and alternate hypotheses. (He turns to Horatio.) Did not Hamlet himself say, "What piece of work is man, how noble in reason, how infinite in faculties? Then let not

this foolishness persist. Go, Horatio, make a survey of three-and-sixty and discover what the true proportion be. For my part, I will never succumb to this fantasy, but deem man to be devoid of all reason should thy proposal of at least five-and-seventy in one hundred hold true.

HORATIO (to Hamlet): What should we do, my Lord?

HAMLET: Go to thy purpose, Horatio.

HORATIO: To what end, my Lord?

HAMLET: That you must teach me. But let me conjure you by the rights of our fellowship, by the consonance of our youth, but the obligation of our ever-preserved love, be even and direct with me, whether I am right or no.

(Horatio exits, followed by Polonius, leaving Hamlet to ponder alone.)

Act II

(The next day, Hamlet awaits anxiously the presence of his friend, Horatio. Polonius enters and places some books upon the table just a moment before Horatio enters.)

POLONIUS: So, Horatio, what is it thou didst reveal through thy deliberations?

HORATIO: In a random survey, for which purpose thou thyself sent me forth, I did discover that one-and-forty believe fervently that the spirits of the dead walk with us. Before my God, I might not this believe, without the sensible and true avouch of mine own eyes.

POLONIUS: Give thine own thoughts no tongue, Horatio. (Polonius turns to Hamlet.) But look to't I charge you, my Lord. Come Horatio, let us go together, for this is not our test. (Horatio and Polonius leave together.)

HAMLET: To reject, or not reject, that is the question: whether 'tis nobler in the mind to suffer the slings and arrows of outrageous statistics, or to take arms against a sea of data, and, by opposing, end them. (Hamlet resignedly attends to his task.)

(Curtain falls)

"Untitled," by Stephen Chen

I've often wondered how software is released and sold to the public. Ironically, I work for a company that sells products with known problems. Unfortunately, most of the problems are difficult to create, which makes them difficult to fix. I usually use the test program X, which tests the product, to try to create a specific problem. When the test program is run to make an error occur, the likelihood of generating an error is 1%.

So, armed with this knowledge, I wrote a new test program Y that will generate the same error that test program X creates, but more often. To find out if my test program is better than the original, so that I can convince the management that I'm right, I ran my test program to find out how often I can generate the same error. When I ran my test program 50 times, I generated the error twice. While this may not seem much better, I think that I can convince the management to use my test program instead of the original test program. Am I right?

Solution

1. H_0: $p = 0.01$
2. H_a: $p > 0.01$
3. Let P' = the proportion of errors generated
4. Normal for a single proportion
5. 2.13
6. 0.0165
7. Check student's solution.
    1. Alpha: 0.05
    2. Decision: Reject the null hypothesis
    3. Reason for decision: The p-value is less than 0.05.
    4. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the proportion of errors generated is more than 0.01.
8. Confidence interval: (0, 0.094).

    The "plus-4s" confidence interval is (0.004, 0.144).

"Japanese Girls' Names"
by Kumi Furuichi

It used to be very typical for Japanese girls' names to end with "ko." (The trend might have started around my grandmothers' generation and its peak might have been around my mother's generation.) "Ko" means "child" in Chinese characters. Parents would name their daughters with "ko" attached to other Chinese characters which have meanings that they want their daughters to become, such as Sachiko—happy child, Yoshiko—a good child, Yasuko—a healthy child, and so on.

However, I noticed recently that only two out of nine of my Japanese girlfriends at this school have names which end with "ko." More and more, parents seem to have become creative, modernized, and, sometimes, westernized in naming their children.

I have a feeling that, while 70 percent or more of my mother's generation would have names with "ko" at the end, the proportion has dropped among my peers. I wrote down all my Japanese friends', ex-classmates', co-workers, and acquaintances' names that I could remember. Following are the names. (Some are repeats.) Test to see if the proportion has dropped for this generation.

Ai, Akemi, Akiko, Ayumi, Chiaki, Chie, Eiko, Eri, Eriko, Fumiko, Harumi, Hitomi, Hiroko, Hiroko, Hidemi, Hisako, Hinako, Izumi, Izumi, Junko, Junko, Kana, Kanako, Kanayo, Kayo, Kayoko, Kazumi, Keiko, Keiko, Kei, Kumi, Kumiko, Kyoko, Kyoko, Madoka, Maho, Mai, Maiko, Maki, Miki, Miki, Mikiko, Mina, Minako, Miyako, Momoko, Nana, Naoko, Naoko, Naoko, Noriko, Rieko, Rika, Rika, Rumiko, Rei, Reiko, Reiko, Sachiko, Sachiko, Sachiyo, Saki, Sayaka, Sayoko, Sayuri, Seiko, Shiho, Shizuka, Sumiko, Takako, Takako, Tomoe, Tomoe, Tomoko, Touko, Yasuko, Yasuko, Yasuyo, Yoko, Yoko, Yoko, Yoshiko, Yoshiko, Yoshiko, Yuka, Yuki, Yuki, Yukiko, Yuko, Yuko.

"Phillip's Wish," by Suzanne Osorio

My nephew likes to play
Chasing the girls makes his day.
He asked his mother
If it is okay
To get his ear pierced.
She said, "No way!"
To poke a hole through your ear,
Is not what I want for you, dear.
He argued his point quite well,
Says even my macho pal, Mel,
Has gotten this done.
It's all just for fun.
C'mon please, mom, please, what the hell.
Again Phillip complained to his mother,
Saying half his friends (including their brothers)
Are piercing their ears
And they have no fears
He wants to be like the others.
She said, "I think it's much less.
We must do a hypothesis test.
And if you are right,
I won't put up a fight.
But, if not, then my case will rest."
We proceeded to call fifty guys
To see whose prediction would fly.
Nineteen of the fifty
Said piercing was nifty
And earrings they'd occasionally buy.
Then there's the other thirty-one,
Who said they'd never have this done.
So now this poem's finished.
Will his hopes be diminished,
Or will my nephew have his fun?

Solution

1. H_0: $p = 0.50$

2. H_a: $p < 0.50$
3. Let P' = the proportion of friends that has a pierced ear.
4. normal for a single proportion
5. –1.70
6. p-value = 0.0448
7. Check student's solution.
    1. Alpha: 0.05
    2. Decision: Reject the null hypothesis
    3. Reason for decision: The p-value is less than 0.05. (However, they are very close.)
    4. Conclusion: There is sufficient evidence to support the claim that less than 50% of his friends have pierced ears.
8. Confidence Interval: (0.245, 0.515): The "plus-4s" confidence interval is (0.259, 0.519).

"The Craven," by Mark Salangsang

Once upon a morning dreary
In stats class I was weak and weary.
Pondering over last night's homework
Whose answers were now on the board
This I did and nothing more.

While I nodded nearly napping
Suddenly, there came a tapping.
As someone gently rapping,
Rapping my head as I snore.
Quoth the teacher, "Sleep no more."

"In every class you fall asleep,"
The teacher said, his voice was deep.
"So a tally I've begun to keep
Of every class you nap and snore.
The percentage being forty-four."

"My dear teacher I must confess,
While sleeping is what I do best.
The percentage, I think, must be less,
A percentage less than forty-four."
This I said and nothing more.

"We'll see," he said and walked away,
And fifty classes from that day
He counted till the month of May
The classes in which I napped and snored.
The number he found was twenty-four.

At a significance level of 0.05,
Please tell me am I still alive?
Or did my grade just take a dive
Plunging down beneath the floor?
Upon thee I hereby implore.

Toastmasters International cites a report by Gallup Poll that 40% of Americans fear public speaking. A student believes that less than 40% of students at her school fear public speaking. She randomly surveys 361 schoolmates and finds that 135 report they fear public speaking. Conduct a hypothesis test to determine if the percent at her school is less than 40%.
Solution

1. H_0: $p = 0.40$
2. H_a: $p < 0.40$
3. Let P' = the proportion of schoolmates who fear public speaking.
4. normal for a single proportion
5. −1.01
6. p-value = 0.1563
7. Check student's solution.
    1. Alpha: 0.05
    2. Decision: Do not reject the null hypothesis.
    3. Reason for decision: The p-value is greater than 0.05.
    4. Conclusion: There is insufficient evidence to support the claim that less than 40% of students at the school fear public speaking.
8. Confidence Interval: (0.3241, 0.4240): The "plus-4s" confidence interval is (0.3257, 0.4250).

Sixty-eight percent of online courses taught at community colleges nationwide were taught by full-time faculty. To test if 68% also represents California's percent for full-time faculty teaching the online classes, Long Beach City College (LBCC) in California, was randomly selected for comparison. In the same year, 34 of the 44 online courses LBCC offered were taught by full-time faculty. Conduct a hypothesis test to determine if 68% represents California. NOTE: For more accurate results, use more California community colleges and this past year's data.

According to an article in *Bloomberg Businessweek*, New York City's most recent adult smoking rate is 14%. Suppose that a survey is conducted to determine this year's rate. Nine out of 70 randomly chosen N.Y. City residents reply that they smoke. Conduct a hypothesis test to determine if the rate is still 14% or if it has decreased.

Solution

1. H_0: $p = 0.14$
2. H_a: $p < 0.14$
3. Let P' = the proportion of NYC residents that smoke.
4. normal for a single proportion
5. –0.2756
6. p-value = 0.3914
7. Check student's solution.
    1. alpha: 0.05
    2. Decision: Do not reject the null hypothesis.
    3. Reason for decision: The p-value is greater than 0.05.
    4. At the 5% significance level, there is insufficient evidence to conclude that the proportion of NYC residents who smoke is less than 0.14.
8. Confidence Interval: (0.0502, 0.2070): The "plus-4s" confidence interval (see chapter 8) is (0.0676, 0.2297).

The mean age of De Anza College students in a previous term was 26.6 years old. An instructor thinks the mean age for online students is older than 26.6. She randomly surveys 56 online students and finds that the sample mean is 29.4 with a standard deviation of 2.1. Conduct a hypothesis test.

Registered nurses earned an average annual salary of $ 69,110. For that same year, a survey was conducted of 41 California registered nurses to determine if the annual salary is higher than $ 69,110 for California nurses. The sample average was $ 71,121 with a sample standard deviation of $ 7,489. Conduct a hypothesis test.

Solution

1. H_0: $\mu = 69{,}110$
2. H_a: $\mu > 69{,}110$
3. Let $\overline{X}$ = the mean salary in dollars for California registered nurses.
4. Student's t-distribution
5. $t = 1.719$
6. p-value: 0.0466
7. Check student's solution.
    1. Alpha: 0.05

2. Decision: Reject the null hypothesis.
3. Reason for decision: The p-value is less than 0.05.
4. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the mean salary of California registered nurses exceeds $ 69,110.

8. ($ 68,757, $ 73,485)

La Leche League International reports that the mean age of weaning a child from breastfeeding is age four to five worldwide. In America, most nursing mothers wean their children much earlier. Suppose a random survey is conducted of 21 U.S. mothers who recently weaned their children. The mean weaning age was nine months (3/4 year) with a standard deviation of 4 months. Conduct a hypothesis test to determine if the mean weaning age in the U.S. is less than four years old.

Over the past few decades, public health officials have examined the link between weight concerns and teen girls' smoking. Researchers surveyed a group of 273 randomly selected teen girls living in Massachusetts (between 12 and 15 years old). After four years the girls were surveyed again. Sixty-three said they smoked to stay thin. Is there good evidence that more than thirty percent of the teen girls smoke to stay thin?

After conducting the test, your decision and conclusion are

1. Reject H_0: There is sufficient evidence to conclude that more than 30% of teen girls smoke to stay thin.
2. Do not reject H_0: There is not sufficient evidence to conclude that less than 30% of teen girls smoke to stay thin.
3. Do not reject H_0: There is not sufficient evidence to conclude that more than 30% of teen girls smoke to stay thin.
4. Reject H_0: There is sufficient evidence to conclude that less than 30% of teen girls smoke to stay thin.

Solution

c

A statistics instructor believes that fewer than 20% of Evergreen Valley College (EVC) students attended the opening night midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 of them attended the midnight showing.

At a 1% level of significance, an appropriate conclusion is:

1. There is insufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is less than 20%.
2. There is sufficient evidence to conclude that the percent of EVC students who attended the midnight

showing of Harry Potter is more than 20%.
3. There is sufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is less than 20%.
4. There is insufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is at least 20%.

Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test.

At a significance level of
$a = 0.05$, what is the correct conclusion?

1. There is enough evidence to conclude that the mean number of hours is more than 4.75
2. There is enough evidence to conclude that the mean number of hours is more than 4.5
3. There is not enough evidence to conclude that the mean number of hours is more than 4.5
4. There is not enough evidence to conclude that the mean number of hours is more than 4.75

Solution

c

Instructions: For the following ten exercises,

Hypothesis testing: For the following ten exercises, answer each question.

1. State the null and alternate hypothesis.
2. State the *p*-value.
3. State alpha.
4. What is your decision?
5. Write a conclusion.
6. Answer any other questions asked in the problem.

According to the Center for Disease Control website, in 2011 at least 18% of high school students have smoked a cigarette. An Introduction to Statistics class in Davies County, KY conducted a hypothesis test at the local high school (a medium sized–approximately 1,200 students–small city demographic) to determine if the local high school's percentage was lower. One hundred fifty students were chosen at random and surveyed. Of

the 150 students surveyed, 82 have smoked. Use a significance level of 0.05 and using appropriate statistical evidence, conduct a hypothesis test and state the conclusions.

A recent survey in the *N.Y. Times Almanac* indicated that 48.8% of families own stock. A broker wanted to determine if this survey could be valid. He surveyed a random sample of 250 families and found that 142 owned some type of stock. At the 0.05 significance level, can the survey be considered to be accurate?
Solution

1. $H_0: p = 0.488$ $H_a: p \neq 0.488$
2. p-value = 0.0114
3. alpha = 0.05
4. Reject the null hypothesis.
5. At the 5% level of significance, there is enough evidence to conclude that 48.8% of families own stocks.
6. The survey does not appear to be accurate.

Driver error can be listed as the cause of approximately 54% of all fatal auto accidents, according to the American Automobile Association. Thirty randomly selected fatal accidents are examined, and it is determined that 14 were caused by driver error. Using $\alpha = 0.05$, is the AAA proportion accurate?

The US Department of Energy reported that 51.7% of homes were heated by natural gas. A random sample of 221 homes in Kentucky found that 115 were heated by natural gas. Does the evidence support the claim for Kentucky at the $\alpha = 0.05$ level in Kentucky? Are the results applicable across the country? Why?
Solution

1. $H_0: p = 0.517$ $H_a: p \neq 0.517$
2. p-value = 0.9203.
3. alpha = 0.05.
4. Do not reject the null hypothesis.
5. At the 5% significance level, there is not enough evidence to conclude that the proportion of homes in Kentucky that are heated by natural gas is 0.517.
6. However, we cannot generalize this result to the entire nation. First, the sample's population is only the state of Kentucky. Second, it is reasonable to assume that homes in the extreme north and south will have extreme high usage and low usage, respectively. We would need to expand our sample base to include these possibilities if we wanted to generalize this claim to the entire nation.

For Americans using library services, the American Library Association claims that at most 67% of patrons borrow books. The library director in Owensboro, Kentucky, feels this is not true, so she asked a local college statistics class to conduct a survey. The class randomly selected 100 patrons and found that 82 borrowed books.

Did the class demonstrate that the percentage was higher in Owensboro, KY? Use $\alpha = 0.01$ level of significance. What is the possible proportion of patrons that do borrow books from the Owensboro Library?

The Weather Underground reported that the mean amount of summer rainfall for the northeastern US is at least 11.52 inches. Ten cities in the northeast are randomly selected and the mean rainfall amount is calculated to be 7.42 inches with a standard deviation of 1.3 inches. At the $\alpha = 0.05$ level, can it be concluded that the mean rainfall was below the reported average? What if $\alpha = 0.01$? Assume the amount of summer rainfall follows a normal distribution.
Solution

1. $H_0: \mu \geq 11.52$ $H_a: \mu < 11.52$
2. p-value = 0.000002 which is almost 0.
3. alpha = 0.05.
4. Reject the null hypothesis.
5. At the 5% significance level, there is enough evidence to conclude that the mean amount of summer rain in the northeastern US is less than 11.52 inches, on average.
6. We would make the same conclusion if alpha was 1% because the p-value is almost 0.

A survey in the *N.Y. Times Almanac* finds the mean commute time (one way) is 25.4 minutes for the 15 largest US cities. The Austin, TX chamber of commerce feels that Austin's commute time is less and wants to publicize this fact. The mean for 25 randomly selected commuters is 22.1 minutes with a standard deviation of 5.3 minutes. At the $\alpha = 0.10$ level, is the Austin, TX commute significantly less than the mean commute time for the 15 largest US cities?

A report by the Gallup Poll found that a woman visits her doctor, on average, at most 5.8 times each year. A random sample of 20 women results in these yearly visit totals

3213729466805642 1341

At the $\alpha = 0.05$ level can it be concluded that the sample mean is higher than 5.8 visits per year?
Solution

1. $H_0: \mu \leq 5.8$ $H_a: \mu > 5.8$
2. p-value = 0.9987
3. alpha = 0.05
4. Do not reject the null hypothesis.
5. At the 5% level of significance, there is not enough evidence to conclude that a woman visits her doctor, on average, more than 5.8 times a year.

According to the *N.Y. Times Almanac* the mean family size in the U.S. is 3.18. A sample of a college math class resulted in the following family sizes:

5454436433556332745222 32

At $\alpha = 0.05$ level, is the class' mean family size greater than the national average? Does the Almanac result remain valid? Why?

The student academic group on a college campus claims that freshman students study at least 2.5 hours per day, on average. One Introduction to Statistics class was skeptical. The class took a random sample of 30 freshman students and found a mean study time of 137 minutes with a standard deviation of 45 minutes. At $\alpha = 0.01$ level, is the student academic group's claim correct?
Solution

1. H_0: $\mu \geq 150$ H_a: $\mu < 150$
2. p-value = 0.0622
3. alpha = 0.01
4. Do not reject the null hypothesis.
5. At the 1% significance level, there is not enough evidence to conclude that freshmen students study less than 2.5 hours per day, on average.
6. The student academic group's claim appears to be correct.

CHAPTER 9: HYPOTHESIS TESTING WITH TWO SAMPLES

INTRODUCTION TO CHAPTER 9: HYPOTHESIS TESTING WITH TWO SAMPLES



Figure 10.1. If you want to test a claim that involves two groups (the types of breakfasts eaten east and west of the Mississippi River) you can use a slightly different technique when conducting a hypothesis test. (credit: Chloe Lim)

Studies often compare two groups. For example, researchers are interested in the effect aspirin has in preventing heart attacks. Over the last few years, newspapers and magazines have reported various aspirin studies involving two groups. Typically, one group is given aspirin and the other group is given a placebo. Then, the heart attack rate is studied over several years.

There are other situations that deal with the comparison of two groups. For example, studies compare various diet and exercise programs. Politicians compare the proportion of individuals from different income

brackets who might vote for them. Students are interested in whether SAT or GRE preparatory courses really help raise their scores.

You have learned to conduct hypothesis tests on single means and single proportions. You will expand upon that in this chapter. You will compare two means or two proportions to each other. The general procedure is still the same, just expanded.

To compare two means or two proportions, you work with two groups. The groups are classified either as **independent** or matched pairs. Independent groups consist of two samples that are independent, that is, sample values selected from one population are not related in any way to sample values selected from the other population. **Matched pairs** consist of two samples that are dependent. The parameter tested using matched pairs is the population mean. The parameters tested using independent groups are either population means or population proportions.

Note

This chapter relies on either a calculator or a computer to calculate the degrees of freedom, the test statistics, and p-values. TI-83+ and TI-84 instructions are included as well as the test statistic formulas. When using a TI-83+ or TI-84 calculator, we do not need to separate two population means, independent groups, or population variances unknown into large and small sample sizes. However, most statistical computer software has the ability to differentiate these tests.

This chapter deals with the following hypothesis tests:

Independent groups (samples are independent)

- Test of two population means.
- Test of two population proportions.

Matched or paired samples (samples are dependent)

- Test of the two population proportions by testing one population mean of differences.

Media Attributions

- Chapter 10 Introduction Figure © OpenStax Introductory Statistics is licensed under a CC BY (Attribution) license

9.1 TWO POPULATION MEANS WITH UNKNOWN STANDARD DEVIATIONS

Learning Objectives

By the end of this section, the student should be able to:

- Conduct and interpret hypothesis tests for two population means, population standard deviations unknown.

Hypothesis Tests for Two Population Means, Population Standard Deviations Unknown

1. The two independent samples are simple random samples from two distinct populations.
2. For the two distinct populations:
    - if the sample sizes are small, the distributions are important (should be normal)
    - if the sample sizes are large, the distributions are not important (need not be normal)

The test comparing two independent population means with unknown and possibly unequal population standard deviations is called the Aspin-Welch t-test. The **degrees of freedom** formula was developed by Aspin-Welch.

The comparison of two population means is very common. A difference between the two samples depends on both the means and the standard deviations. Very different means can occur by chance if there is great variation among the individual samples. In order to account for the variation, we take the difference of the sample means,

$\overline{X}_1 - \overline{X}_2$, and divide by the standard error in order to standardize the difference. The result is a t-score test statistic.

Because we do not know the population **standard deviations**, we estimate them using the two sample standard deviations from our independent samples. For the hypothesis test, we calculate the estimated **standard deviation**, or standard error, of **the difference in sample means**, $\overline{X}_1 - \overline{X}_2$.

The standard error is: Formula does not parse

The test statistic (t-score) is calculated as follows:

$$\frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$$

where:

s_1 and s_2, the sample standard deviations, are estimates of σ_1 and σ_2, respectively.

σ_1 and σ_1 are the unknown population standard deviations.

$\overline{x}_1$

and

$\overline{x}_2$

are the sample means. μ_1 and μ_2 are the population means.

The number of **degrees of freedom** (df) requires a somewhat complicated calculation. However, a computer or calculator calculates it easily. The df are not always a whole number. The test statistic calculated previously is approximated by the Student's t-distribution with df as follows:

$$\text{Degrees of freedom } df = \frac{\left(\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}\right)^2}{\left(\frac{1}{n_1 - 1}\right)\left(\frac{(s_1)^2}{n_1}\right)^2 + \left(\frac{1}{n_2 - 1}\right)\left(\frac{(s_2)^2}{n_2}\right)^2}$$

When both sample sizes n_1 and n_2 are five or larger, the Student's t approximation is very good. Notice that the sample variances $(s_1)^2$ and $(s_2)^2$ are not pooled. (If the question comes up, do not pool the variances.)

It is not necessary to compute this by hand. A calculator or computer easily computes it.

Exercises

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=905#h5p-199

Example

Independent groups

The average amount of time boys and girls aged seven to 11 spend playing sports each day is believed to be the same. A study is done and data are collected, resulting in the data in [link]. Each population has a normal distribution.

	Sample Size	Average Number of Hours Playing Sports Per Day	Sample Standard Deviation
Girls	9	2	0.866
Boys	16	3.2	1.00

Is there a difference in the mean amount of time boys and girls aged seven to 11 play sports each day? Test at the 5% level of significance.

Solution

The population standard deviations are not known. Let g be the subscript for girls and b be the subscript for boys. Then, μg is the population mean for girls and μb is the population mean for boys. This is a test of two independent groups, two population means.

Random variable:

$\overline{X}_g - \overline{X}_b$

= difference in the sample mean amount of time girls and boys play sports each day.

H0: μg = μb H0: μg – μb = 0

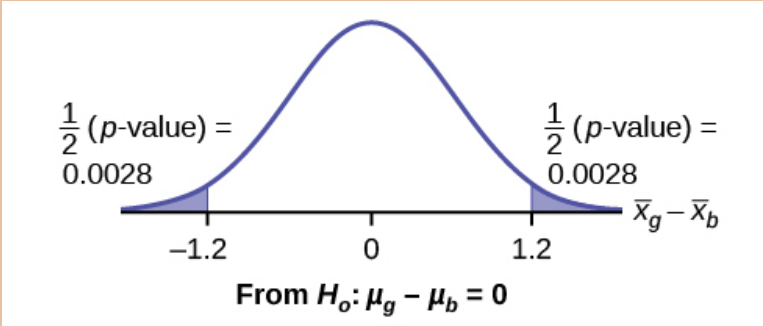
Ha: μg ≠ μb Ha: μg – μb ≠ 0

The words "the same" tell you H0 has an "=". Since there are no other words to indicate Ha, assume it says "is different."
This is a two-tailed test.

Distribution for the test:
Use tdf where df is calculated using the df formula for independent groups, two population means.
Using a calculator, df is approximately 18.8462. Do not pool the variances.

Calculate the p-value using a Student's t-distribution: p-value = 0.0054

Graph:



$s_g = 0.866$

$s_b = 1$

So, $\bar{x}_g - \bar{x}_b = 2 - 3.2 = -1.2$

Half the p-value is below –1.2 and half is above 1.2.

Make a decision: Since $\alpha > p$-value, reject H_0. This means you reject $\mu_g = \mu_b$. The means are different.

Press STAT. Arrow over to TESTS and press 4:2-SampTTest. Arrow over to Stats and press ENTER. Arrow down and enter 2 for the first sample mean, $\sqrt{0.866}$ for Sx1, 9 for n1, 3.2 for the second sample mean, 1 for Sx2, and 16 for n2. Arrow down to μ1: and arrow to does not equal μ2. Press ENTER. Arrow down to Pooled: and No. Press ENTER. Arrow down to Calculate and press ENTER. The p-value is p = 0.0054, the dfs are approximately 18.8462, and the test statistic is -3.14. Do the procedure again but instead of Calculate do Draw.

Conclusion: At the 5% level of significance, the sample data show there is sufficient evidence to conclude that the mean number of hours that girls and boys aged seven to 11 play sports per day is different (mean number of hours boys aged seven to 11 play sports per day is greater than the mean number of hours played by girls OR

the mean number of hours girls aged seven to 11 play sports per day is greater than the mean number of hours played by boys).

Your Turn!

Two samples are shown in [link]. Both have normal distributions. The means for the two populations are thought to be the same. Is there a difference in the means? Test at the 5% level of significance.

	Sample Size	Sample Mean	Sample Standard Deviation
Population A	25	5	1
Population B	16	4.7	1.2

Solution

The p-value is 0.4125, which is much higher than 0.05, so we decline to reject the null hypothesis. There is not sufficient evidence to conclude that the means of the two populations are not the same.

Note

When the sum of the sample sizes is larger than 30 ($n_1 + n_2 > 30$) you can use the normal distribution to approximate the Student's t.

Example

A study is done by a community group in two neighboring colleges to determine which one graduates students with more math classes. College A samples 11 graduates. Their average is four math classes with a standard deviation of 1.5 math classes. College B samples nine graduates. Their average is 3.5 math classes with a standard deviation of one math class. The community group believes that a student who graduates from college A has taken more math classes, on the average. Both populations have a normal distribution. Test at a 1% significance level. Answer the following questions.

a. Is this a test of two means or two proportions?

Solution
a. two means

b. Are the populations standard deviations known or unknown?

Solution
b. unknown

c. Which distribution do you use to perform the test?

Solution
c. Student's *t*

d. What is the random variable?

Solution
d. $\overline{X}_A - \overline{X}_B$

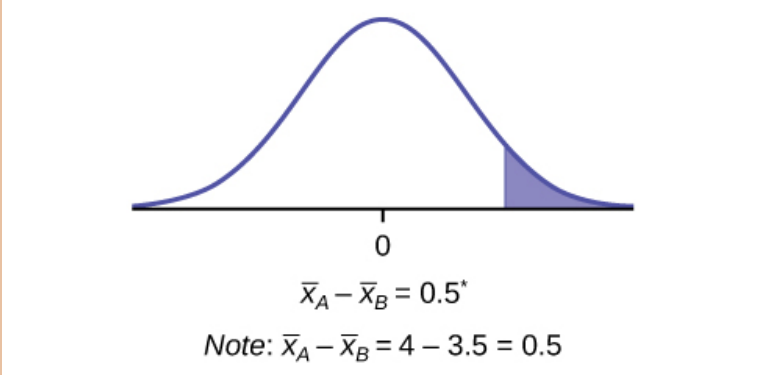
e. What are the null and alternate hypotheses? Write the null and alternate hypotheses in words and in symbols.

Solution
e.

$$H_o : \mu_A \leq \mu_B$$
$$H_a : \mu_A > \mu_B$$

f. Is this test right-, left-, or two-tailed?

Solution
f.



right

g. What is the *p*-value?

Solution
g. 0.1928

h. Do you reject or not reject the null hypothesis?

Solution
h. Do not reject.

i. **Conclusion:**

Solution
i. At the 1% level of significance, from the sample data, there is not sufficient evidence to conclude that a student who graduates from college A has taken more math classes, on the average, than a student who graduates from college B.

Your Turn!

A study is done to determine if Company A retains its workers longer than Company B. Company A samples 15 workers, and their average time with the company is five years with a standard deviation of 1.2. Company B samples 20 workers, and their average time with the company is 4.5 years with a standard deviation of 0.8. The populations are normally distributed.

1. Are the population standard deviations known?
2. Conduct an appropriate hypothesis test. At the 5% significance level, what is your conclusion?

Solution

1. They are unknown.
2. The *p*-value = 0.0878. At the 5% level of significance, there is insufficient evidence to conclude that the workers of Company A stay longer with the company.

Example

A professor at a large community college wanted to determine whether there is a difference in the means of final exam scores between students who took his statistics course online and the students who took his face-to-face statistics class. He believed that the mean of the final exam scores for the online class would be lower than that of the face-to-face class. Was the professor correct? The randomly selected 30 final exam scores from each group are listed in [link] and [link].

Online Class

67.6	41.2	85.3	55.9	82.4	91.2	73.5	94.1	64.7	64.7
70.6	38.2	61.8	88.2	70.6	58.8	91.2	73.5	82.4	35.5
94.1	88.2	64.7	55.9	88.2	97.1	85.3	61.8	79.4	79.4

Face-to-face Class

77.9	95.3	81.2	74.1	98.8	88.2	85.9	92.9	87.1	88.2
69.4	57.6	69.4	67.1	97.6	85.9	88.2	91.8	78.8	71.8
98.8	61.2	92.9	90.6	97.6	100	95.3	83.5	92.9	89.4

Is the mean of the Final Exam scores of the online class lower than the mean of the Final Exam scores of the face-to-face class? Test at a 5% significance level. Answer the following questions:

1. Is this a test of two means or two proportions?
2. Are the population standard deviations known or unknown?
3. Which distribution do you use to perform the test?
4. What is the random variable?
5. What are the null and alternative hypotheses? Write the null and alternative hypotheses in words and in symbols.
6. Is this test right, left, or two tailed?
7. What is the p-value?
8. Do you reject or not reject the null hypothesis?
9. At the ___ level of significance, from the sample data, there _____ (is/is not) sufficient evidence to conclude that _____.

(See the conclusion in [link], and write yours in a similar fashion.)

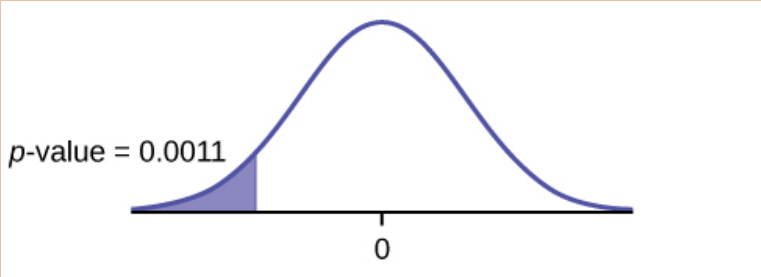
First put the data for each group into two lists (such as L1 and L2). Press STAT. Arrow over to TESTS and press 4:2SampTTest. Make sure Data is highlighted and press ENTER. Arrow down and enter L1 for the first list and L2 for the second list. Arrow down to μ1: and arrow to ≠ μ2 (does not equal). Press ENTER. Arrow down to Pooled: No. Press ENTER. Arrow down to Calculate and press ENTER.

Note

Be careful not to mix up the information for Group 1 and Group 2!

Solution

1. two means
2. unknown
3. Student's t
4. $\overline{X}_1 - \overline{X}_2$
    1. H0: μ1 = μ2 Null hypothesis: the means of the final exam scores are equal for the online and face-to-face statistics classes.
    2. Ha: μ1 < μ2 Alternative hypothesis: the mean of the final exam scores of the online class is less than the mean of the final exam scores of the face-to-face class.
5. left-tailed
6. p-value = 0.0011



7. Reject the null hypothesis
8. The professor was correct. The evidence shows that the mean of the final exam scores for the online class is lower than that of the face-to-face class.
   At the 5% level of significance, from the sample data, there is (is/is not) sufficient evidence to conclude that the mean of the final exam scores for the online class is less than the mean of final exam scores of the face-to-face class.

Cohen's Standards for Small, Medium, and Large Effect SizesCohen's d is a measure of effect size based on the differences between two means. Cohen's d, named for United States statistician Jacob Cohen, measures the relative strength of the differences between the means of two populations based on sample data. The calculated value of effect size is then compared to Cohen's standards of small, medium, and large effect sizes.

Cohen's Standard Effect Sizes

Size of effect	***d***
Small	0.2
medium	0.5
Large	0.8

Cohen's d is the measure of the difference between two means divided by the pooled standard deviation:

$$d = \frac{\overline{x}_1 - \overline{x}_2}{s_{pooled}} \text{ where } s_{pooled} = \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}}$$

Example

Calculate Cohen's d for [link]. Is the size of the effect small, medium, or large? Explain what the size of the effect means for this problem.

Solution
$\mu_1 = 4$ $s_1 = 1.5$ $n_1 = 11$

$\mu_2 = 3.5$ $s_2 = 1$ $n_2 = 9$

$d = 0.384$

The effect is small because 0.384 is between Cohen's value of 0.2 for small effect size and 0.5 for medium effect size. The size of the differences of the means for the two colleges is small indicating that there is not a significant difference between them.

Example

Calculate Cohen's d for [link]. Is the size of the effect small, medium or large? Explain what the size of the effect means for this problem.

Solution
$d = 0.834$; Large, because 0.834 is greater than Cohen's 0.8 for a large effect size. The size of the differences between the means of the Final Exam scores of online students and students in a face-to-face class is large indicating a significant difference.

Your Turn!

Weighted alpha is a measure of risk-adjusted performance of stocks over a period of a year. A high positive weighted alpha signifies a stock whose price has risen while a small positive weighted

alpha indicates an unchanged stock price during the time period. Weighted alpha is used to identify companies with strong upward or downward trends. The weighted alpha for the top 30 stocks of banks in the northeast and in the west as identified by Nasdaq on May 24, 2013, are listed in [link] and [link], respectively.

Northeast

94.2	75.2	69.6	52.0	48.0	41.9	36.4	33.4	31.5	27.6
77.3	71.9	67.5	50.6	46.2	38.4	35.2	33.0	28.7	26.5
76.3	71.7	56.3	48.7	43.2	37.6	33.7	31.8	28.5	26.0

West

126.0	70.6	65.2	51.4	45.5	37.0	33.0	29.6	23.7	22.6
116.1	70.6	58.2	51.2	43.2	36.0	31.4	28.7	23.5	21.6
78.2	68.2	55.6	50.3	39.0	34.1	31.0	25.3	23.4	21.5

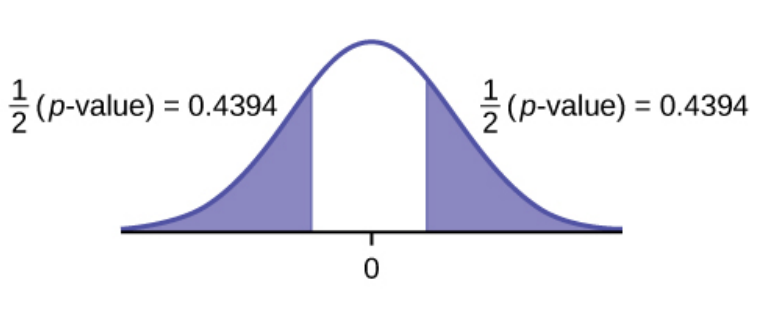
Is there a difference in the weighted alpha of the top 30 stocks of banks in the northeast and in the west? Test at a 5% significance level. Answer the following questions:

1. Is this a test of two means or two proportions?
2. Are the population standard deviations known or unknown?
3. Which distribution do you use to perform the test?
4. What is the random variable?
5. What are the null and alternative hypotheses? Write the null and alternative hypotheses in words and in symbols.
6. Is this test right, left, or two tailed?
7. What is the *p*-value?
8. Do you reject or not reject the null hypothesis?
9. At the ___ level of significance, from the sample data, there _____ (is/is not) sufficient evidence to conclude that _____.
10. Calculate Cohen's *d* and interpret it.

Solution

1. two means
2. unknown
3. Student's-t

4. $\overline{X}_1 - \overline{X}_2$
    1. $H_0 : \mu_1 = \mu_2$ Null hypothesis: the means of the weighted alphas are equal.
    2. $H_a : \mu_1 \neq \mu_2$ Alternative hypothesis : the means of the weighted alphas are not equal.
5. two-tailed
6. p-value = 0.8787
7. Do not reject the null hypothesis
8. This indicates that the trends in stocks are about the same in the top 30 banks in each region.



5% level of significance, from the sample data, there is not sufficient evidence to conclude that the mean weighted alphas for the banks in the northeast and the west are different

9. d = 0.040, Very small, because 0.040 is less than Cohen's value of 0.2 for small effect size. The size of the difference of the means of the weighted alphas for the two regions of banks is small indicating that there is not a significant difference between their trends in stocks.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=905#oembed-1

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=905#oembed-2

References

Data from Graduating Engineer + Computer Careers. Available online at http://www.graduatingengineer.com

Data from *Microsoft Bookshelf*.

Data from the United States Senate website, available online at www.Senate.gov (accessed June 17, 2013).

"List of current United States Senators by Age." Wikipedia. Available online at http://en.wikipedia.org/wiki/List_of_current_United_States_Senators_by_age (accessed June 17, 2013).

"Sectoring by Industry Groups." Nasdaq. Available online at http://www.nasdaq.com/markets/barchart-sectors.aspx?page=sectors&base=industry (accessed June 17, 2013).

"Strip Clubs: Where Prostitution and Trafficking Happen." Prostitution Research and Education, 2013. Available online at www.prostitutionresearch.com/ProsViolPosttrauStress.html (accessed June 17, 2013).

"World Series History." Baseball-Almanac, 2013. Available online at http://www.baseball-almanac.com/ws/wsmenu.shtml (accessed June 17, 2013).

9.2 TWO POPULATION MEANS WITH KNOWN STANDARD DEVIATIONS

Learning Objectives

By the end of this section, the student should be able to:

- Conduct and interpret hypothesis tests for two population means, population standard deviations known.

Hypothesis Tests for Two Population Means, Population Standard Deviations Known

Even though this situation is not likely (knowing the population standard deviations is not likely), the following example illustrates hypothesis testing for independent means, known population standard deviations. The sampling distribution for the difference between the means is normal and both populations must be normal. The random variable is $\overline{X}_1 - \overline{X}_2$. The normal distribution has the following format:

Normal distribution is: $\overline{X}_1 - \overline{X}_2 \sim N\left[\mu_1 - \mu_2, \sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}\right]$

The standard deviation is:

$$\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}$$

The test statistic (z-score) is:

$$z = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}}$$

Example

Independent groups, population standard deviations known: The mean lasting time of two competing floor waxes is to be compared. Twenty floors are randomly assigned to test each wax. Both populations have normal distributions. The data are recorded in [link].

Wax	Sample Mean Number of Months Floor Wax Lasts	Population Standard Deviation
1	3	0.33
2	2.9	0.36

Does the data indicate that wax 1 is more effective than wax 2? Test at a 5% level of significance.

Solution

This is a test of two independent groups, two population means, population standard deviations known.

Random Variable: $\overline{X}_1 - \overline{X}_2$ = difference in the mean number of months the competing floor waxes last.

H0: μ1 ≤ μ2

Ha: μ1 > μ2

The words “is more effective” says that wax 1 lasts longer than wax 2, on average. “Longer” is a “>” symbol and goes into Ha. Therefore, this is a right-tailed test.

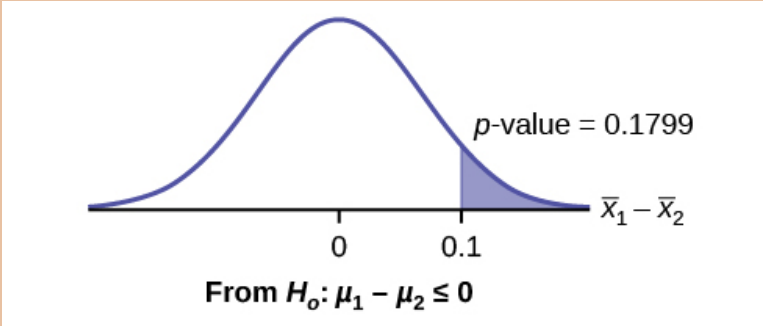
Distribution for the test: The population standard deviations are known so the distribution is normal. Using the formula, the distribution is:

$$\overline{X}_1 - \overline{X}_2 \; N\left(0, \sqrt{\frac{0.33^2}{20} + \frac{0.36^2}{20}}\right)$$

Since μ1 ≤ μ2 then μ1 – μ2 ≤ 0 and the mean for the normal distribution is zero.

Calculate the p-value using the normal distribution: p-value = 0.1799

Graph:



$\overline{X}_1 - \overline{X}_2$ = 3 – 2.9 = 0.1

Compare **α** and the p-value:**α** = 0.05 and p-value = 0.1799. Therefore, **α** < p-value.

Make a decision: Since **α** < p-value, do not reject H0.

Conclusion: At the 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the mean time wax 1 lasts is longer (wax 1 is more effective) than the mean time wax 2 lasts.

Press STAT. Arrow over to TESTS and press 3:2-SampZTest. Arrow over to Stats and press ENTER. Arrow down and enter .33 for sigma1, .36 for sigma2, 3 for the first sample mean, 20 for n1, 2.9 for the second sample mean, and 20 for n2. Arrow down to μ1: and arrow to > μ2. Press ENTER. Arrow down to Calculate and press ENTER. The p-value is p = 0.1799 and the test statistic is 0.9157. Do the procedure again, but instead of Calculate do Draw.

Your Turn!

The means of the number of revolutions per minute of two competing engines are to be compared. Thirty engines are randomly assigned to be tested. Both populations have normal distributions. [link] shows the result. Do the data indicate that Engine 2 has higher RPM than Engine 1? Test at a 5% level of significance.

Engine	Sample Mean Number of RPM	Population Standard Deviation
1	1,500	50
2	1,600	60

Solution

The *p*-value is almost zero, so we reject the null hypothesis. There is sufficient evidence to conclude that Engine 2 runs at a higher RPM than Engine 1.

Example

An interested citizen wanted to know if Democratic U. S. senators are older than Republican U.S. senators, on average. On May 26 2013, the mean age of 30 randomly selected Republican Senators was 61 years 247 days old (61.675 years) with a standard deviation of 10.17 years. The mean age of 30 randomly selected Democratic senators was 61 years 257 days old (61.704 years) with a standard deviation of 9.55 years.

Do the data indicate that Democratic senators are older than Republican senators, on average? Test at a 5% level of significance.

Solution

This is a test of two independent groups, two population means. The population standard deviations are unknown, but the sum of the sample sizes is 30 + 30 = 60, which is greater than 30, so we can use the normal approximation to the Student's-t distribution. Subscripts: 1: Democratic senators 2: Republican senators

Random variable: $\overline{X}_1 - \overline{X}_2$ = difference in the mean age of Democratic and Republican U.S. senators.

H0: μ1 ≤ μ2 H0: μ1 – μ2 ≤ 0

Ha: μ1 > μ2 Ha: μ1 – μ2 > 0

The words “older than” translates as a “>” symbol and goes into Ha. Therefore, this is a right-tailed test.

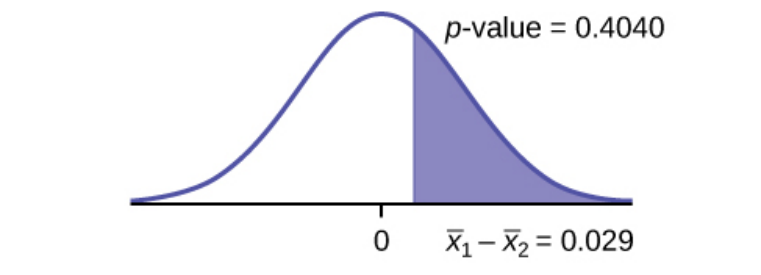
Distribution for the test: The distribution is the normal approximation to the Student’s t for means, independent groups. Using the formula, the distribution is:

$$\overline{X}_1 - \overline{X}_2 \sim N\left[0, \sqrt{\frac{(9.55)^2}{30} + \frac{(10.17)^2}{30}}\right]$$

Since μ1 ≤ μ2, μ1 – μ2 ≤ 0 and the mean for the normal distribution is zero.

(Calculating the p-value using the normal distribution gives p-value = 0.4040)

Graph:



Compare **α** and the p-value:**α** = 0.05 and p-value = 0.4040. Therefore, **α** < p-value.

Make a decision: Since **α** < p-value, do not reject H0.

Conclusion: At the 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the mean age of Democratic senators is greater than the mean age of the Republican senators.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=909#oembed-1

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=909#oembed-2

References

Data from the United States Census Bureau. Available online at http://www.census.gov/prod/cen2010/briefs/c2010br-02.pdf

Hinduja, Sameer. "Sexting Research and Gender Differences." Cyberbullying Research Center, 2013. Available online at http://cyberbullying.us/blog/sexting-research-and-gender-differences/ (accessed June 17, 2013).

"Smart Phone Users, By the Numbers." Visually, 2013. Available online at http://visual.ly/smart-phone-users-numbers (accessed June 17, 2013).

Smith, Aaron. "35% of American adults own a Smartphone." Pew Internet, 2013. Available online at http://www.pewinternet.org/~/media/Files/Reports/2011/PIP_Smartphones.pdf (accessed June 17, 2013).

"State-Specific Prevalence of Obesity Among Adults—United States, 2007." MMWR, CDC. Available online at http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5728a1.htm (accessed June 17, 2013).

"Texas Crime Rates 1960–1012." FBI, Uniform Crime Reports, 2013. Available online at: http://www.disastercenter.com/crime/txcrime.htm (accessed June 17, 2013).

9.3 COMPARING TWO INDEPENDENT POPULATION PROPORTIONS

Learning Objectives

By the end of this section, the student should be able to

- Conduct and interpret hypothesis tests for two population proportions

Hypothesis tests for two population proportions

When conducting a hypothesis test that compares two independent population proportions, the following characteristics should be present:

1. The two independent samples are simple random samples that are independent.
2. The number of successes is at least five, and the number of failures is at least five, for each of the samples.
3. Growing literature states that the population must be at least ten or 20 times the size of the sample. This keeps each population from being over-sampled and causing incorrect results.

Comparing two proportions, like comparing two means, is common. If two estimated proportions are different, it may be due to a difference in the populations or it may be due to chance. A hypothesis test can help determine if a difference in the estimated proportions reflects a difference in the population proportions.

The difference of two proportions follows an approximate normal distribution. Generally, the null hypothesis states that the two proportions are the same. That is, H_0: $p_A = p_B$. To conduct the test, we use a **pooled proportion**, p_c.

The pooled proportion is calculated as follows:

$$p_c = \frac{x_A + x_B}{n_A + n_B}$$

The distribution for the differences is:

$$P'_A - P'_B \sim N\left[0, \sqrt{p_c\left(1 - p_c\right)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}\right]$$

The test statistic (z-score) is:

$$z = \frac{(p'_A - p'_B) - (p_A - p_B)}{\sqrt{p_c\left(1 - p_c\right)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

Example

Two types of medication for hives are being tested to determine if there is a difference in the proportions of adult patient reactions. Twenty out of a random sample of 200 adults given medication A still had hives 30 minutes after taking the medication. Twelve out of another random sample of 200 adults given medication B still had hives 30 minutes after taking the medication. Test at a 1% level of significance.

Solution

The problem asks for a difference in proportions, making it a test of two proportions.

Let A and B be the subscripts for medication A and medication B, respectively. Then pA and pB are the desired population proportions.

Random Variable: P'A – P'B = difference in the proportions of adult patients who did not react after 30 minutes to medication A and to medication B.

H0: pA = pB

pA – pB = 0

Ha: pA ≠ pB

pA – pB ≠ 0

The words “is a difference” tell you the test is two-tailed.

Distribution for the test: Since this is a test of two binomial population proportions, the distribution is normal:

$$p_c = \frac{x_A + x_B}{n_A + n_B} = \frac{20 + 12}{200 + 200} = 0.08 \; 1 - p_c = 0.92$$

$$P'_A - P'_B \; N\left[0, \sqrt{(0.08)\,(0.92)\left(\frac{1}{200} + \frac{1}{200}\right)}\right]$$

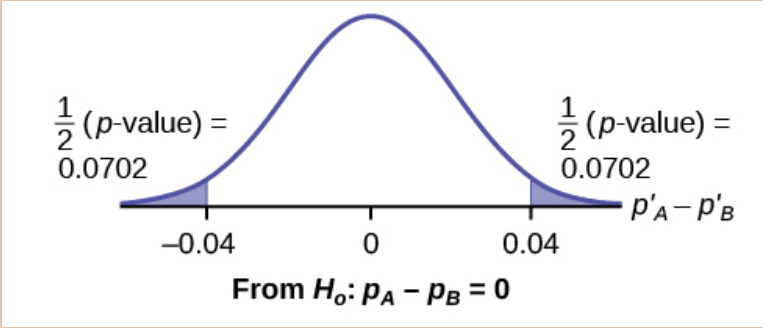
P'A – P'B follows an approximate normal distribution.

Calculate the p-value using the normal distribution: p-value = 0.1404.

Estimated proportion for group A: $p'_A = \frac{x_A}{n_A} = \frac{20}{200} = 0.1$

Estimated proportion for group B: $p'_B = \frac{x_B}{n_B} = \frac{12}{200} = 0.06$

Graph:



P'A – P'B = 0.1 – 0.06 = 0.04.

Half the p-value is below –0.04, and half is above 0.04.

Compare **α** and the p-value: **α** = 0.01 and the p-value = 0.1404. **α** < p-value.

Make a decision: Since **α** < p-value, do not reject H0.

Conclusion: At a 1% level of significance, from the sample data, there is not sufficient evidence to conclude that there is a difference in the proportions of adult patients who did not react after 30 minutes to medication A and medication B.

Press STAT. Arrow over to TESTS and press
6:2-PropZTest. Arrow down and enter 20 for x1, 200 for n1, 12 for x2,
and 200 for n2. Arrow down to p1: and arrow to not equal p2. Press ENTER. Arrow down to

Calculate and press ENTER. The p-value is p = 0.1404 and the test statistic is 1.47. Do the procedure again, but instead of Calculate do Draw.

Your Turn!

Two types of valves are being tested to determine if there is a difference in pressure tolerances. Fifteen out of a random sample of 100 of Valve *A* cracked under 4,500 psi. Six out of a random sample of 100 of Valve *B* cracked under 4,500 psi. Test at a 5% level of significance.

Solution

The *p*-value is 0.0379, so we can reject the null hypothesis. At the 5% significance level, the data support that there is a difference in the pressure tolerances between the two valves.

Example

A research study was conducted about gender differences in "sexting." The researcher believed that the proportion of girls involved in "sexting" is less than the proportion of boys involved. The data collected in the spring of 2010 among a random sample of middle and high school students in a large school district in the southern United States is summarized in [link]. Is the proportion of girls sending sexts less than the proportion of boys "sexting?" Test at a 1% level of significance.

	Males	**Females**
Sent "sexts"	183	156
Total number surveyed	2231	2169

Solution

This is a test of two population proportions. Let M and F be the subscripts for males and females. Then pM and pF are the desired population proportions.

Random variable:p'F – p'M = difference in the proportions of males and females who sent "sexts."

H0: pF = pM H0: pF – pM = 0

Ha: pF < pM Ha: pF – pM < 0

The words "less than" tell you the test is left-tailed.

Distribution for the test: Since this is a test of two population proportions, the distribution is normal:

$$p_c = \frac{x_F + x_M}{n_F + n_M} = \frac{156 + 183}{2169 + 2231} = 0.077$$

$$1 - p_c = 0.923$$

Therefore,

$$p'_F - p'_M \sim N\left(0, \sqrt{(0.077)(0.923)\left(\frac{1}{2169} + \frac{1}{2231}\right)}\right)$$

p'F – p'M follows an approximate normal distribution.

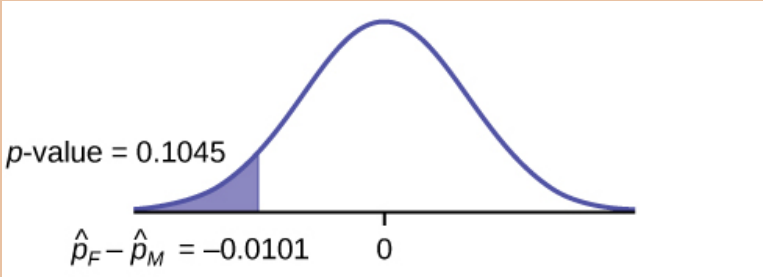
Calculate the p-value using the normal distribution:

p-value = 0.1045

Estimated proportion for females: 0.0719

Estimated proportion for males: 0.082

Graph:



Decision: Since α < p-value, Do not reject H0

Conclusion: At the 1% level of significance, from the sample data, there is not sufficient evidence to conclude that the proportion of girls sending "sexts" is less than the proportion of boys sending "sexts."

Press STAT. Arrow over to TESTS and press 6:2-PropZTest. Arrow down and enter 156 for x1, 2169 for n1, 183 for x2, and 2231 for n2. Arrow down to p1: and arrow to less than p2. Press ENTER. Arrow down to Calculate and press ENTER. The p-value is P = 0.1045 and the test statistic is z = -1.256.

Exercises

Two types of medication for hives are being tested to determine if there is a difference in the proportions of adult patient reactions. Twenty out of a random sample of 200 adults given medication A still had hives 30 minutes after taking the medication. Twelve out of another random sample of 200 adults given medication B still had hives 30 minutes after taking the medication. Test at a 1% level of significance.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://louis.pressbooks.pub/introductorystatistics/?p=917#h5p-181

Graph:

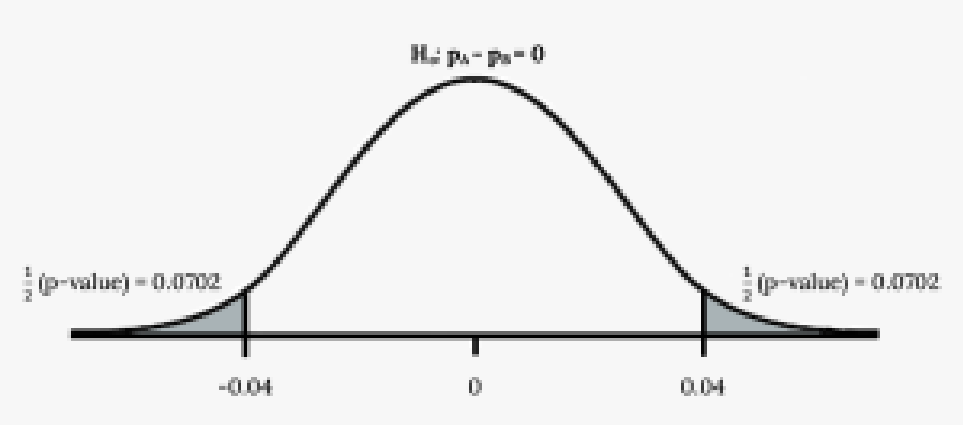


Figure 8.8: Medication A and B

Example

Researchers conducted a study of smartphone use among adults. A cell phone company claimed that iPhone smartphones are more popular with whites (non-Hispanic) than with African Americans. The results of the survey indicate that of the 232 African American cell phone owners randomly sampled, 5% have an iPhone. Of the 1,343 white cell phone owners randomly sampled, 10% own an iPhone. Test at the 5% level of significance. Is the proportion of white iPhone owners greater than the proportion of African American iPhone owners?

Solution

This is a test of two population proportions. Let W and A be the subscripts for the whites and African Americans. Then pW and pA are the desired population proportions.

Random variable: p'W – p'A = difference in the proportions of Android and iPhone users.

H0: pW = pA H0: pW – pA = 0

Ha: pW > pA Ha: pW – pA > 0

The words “more popular” indicate that the test is right-tailed.

Distribution for the test: The distribution is approximately normal:

$$p_c = \frac{x_W + x_A}{n_W + n_A} = \frac{134 + 12}{1343 + 232} = 0.0927$$

$$1 - p_c = 0.9073$$

Therefore,

$$p'_W - p'_A \backsim N\left(0, \sqrt{(0.0927)(0.9073)\left(\frac{1}{1343} + \frac{1}{232}\right)}\right)$$

$p'_W - p'_A$ follows an approximate normal distribution.

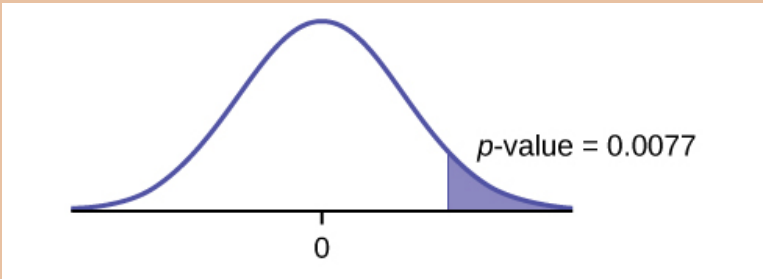
Calculate the p-value using the normal distribution:

p-value = 0.0077

Estimated proportion for group A: 0.10

Estimated proportion for group B: 0.05<!– LALALA ♪♫♪♫ CONTINUE INSERTING NEW EXAMPLE 3 HERE –>

Graph:



Decision: Since α > p-value, reject the H0.

Conclusion: At the 5% level of significance, from the sample data, there is sufficient evidence to conclude that a larger proportion of white cell phone owners use iPhones than African Americans.

TI-83+ and TI-84: Press STAT. Arrow over to TESTS and press 6:2-PropZTest. Arrow down and enter 135 for x1, 1343 for n1, 12 for x2, and 232 for n2. Arrow down to p1: and arrow to greater than p2. Press ENTER. Arrow down to Calculate and press ENTER. The P-value is P = 0.0092 and the test statistic is Z = 2.33.

Your Turn!

A concerned group of citizens wanted to know if the proportion of forcible rapes in Texas was different in 2011 than in 2010. Their research showed that of the 113,231 violent crimes in Texas in 2010, 7,622 of them were forcible rapes. In 2011, 7,439 of the 104,873 violent crimes were in the forcible rape category. Test at a 5% significance level. Answer the following questions:

a. Is this a test of two means or two proportions?

Solution
a. two proportions

b. Which distribution do you use to perform the test?

Solution
b. normal for two proportions

c. What is the random variable?

Solution
c. Subscripts: 1 = 2010, 2 = 2011

$P'_2 - P'_2$

d. What are the null and alternative hypotheses? Write the null and alternative hypothesis in symbols.

Solution
d. Subscripts: 1 = 2010, 2 = 2011

H_0: $p_1 = p_2$ H_0: $p_1 - p_2 = 0$

H_a: $p_1 \neq p_2$ H_a: $p_1 - p_2 \neq 0$

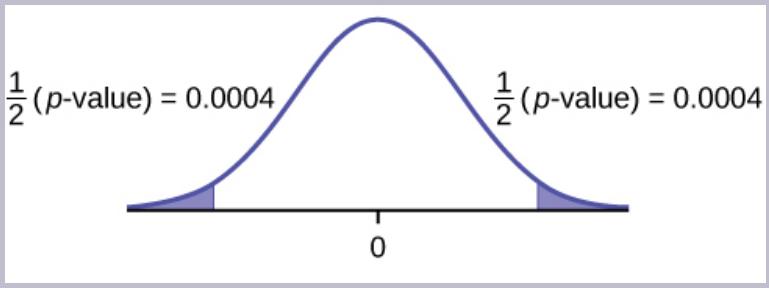
e. Is this test right-, left-, or two-tailed?

Solution
e. two-tailed

f. What is the *p*-value?

Solution

f. p-value = 0.00086



g. Do you reject or not reject the null hypothesis?

Solution

g. Reject the H_0.

h. At the ___ level of significance, from the sample data, there _____ (is/is not) sufficient evidence to conclude that _____.

Solution

h. At the 5% significance level, from the sample data, there is sufficient evidence to conclude that there is a difference between the proportion of forcible rapes in 2011 and 2010.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=917#oembed-1

References

Data from *Educational Resources*, December catalog.

Data from Hilton Hotels. Available online at http://www.hilton.com (accessed June 17, 2013).

Data from Hyatt Hotels. Available online at http://hyatt.com (accessed June 17, 2013).

Data from Statistics, United States Department of Health and Human Services.

Data from Whitney Exhibit on loan to San Jose Museum of Art.

Data from the American Cancer Society. Available online at http://www.cancer.org/index (accessed June 17, 2013).

Data from the Chancellor's Office, California Community Colleges, November 1994.

"State of the States." Gallup, 2013. Available online at http://www.gallup.com/poll/125066/State-States.aspx?ref=interactive (accessed June 17, 2013).

"West Nile Virus." Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/ncidod/dvbid/westnile/index.htm (accessed June 17, 2013).

Media Attributions

- Private: 8.8

9.4 MATCHED OR PAIRED SAMPLES

Learning Objectives

By the end of this section, the student should be able to

- Conduct and interpret hypothesis tests for matched or paired samples.

Hypothesis Tests for Matched or Paired Samples

When using a hypothesis test for matched or paired samples, the following characteristics should be present:

1. Simple random sampling is used.
2. Sample sizes are often small.
3. Two measurements (samples) are drawn from the same pair of individuals or objects.
4. Differences are calculated from the matched or paired samples.
5. The differences form the sample that is used for the hypothesis test.
6. Either the matched pairs have differences that come from a population that is normal or the number of differences is sufficiently large so that distribution of the sample mean of differences is approximately normal.

In a hypothesis test for matched or paired samples, subjects are matched in pairs and differences are calculated. The differences are the data. The population mean for the differences, μ_d, is then tested using a Student's t-test for a single population mean with $n - 1$ degrees of freedom, where n is the number of differences.
The test statistic (t-score) is:

$$t = \frac{\overline{x}_d - \mu_d}{\left(\frac{s_d}{\sqrt{n}}\right)}$$

Example

A study was conducted to investigate the effectiveness of hypnotism in reducing pain. Results for randomly selected subjects are shown in [link]. A lower score indicates less pain. The "before" value is matched to an "after" value and the differences are calculated. The differences have a normal distribution. Are the sensory measurements, on average, lower after hypnotism? Test at a 5% significance level.

Subject:	A	B	C	D	E	F	G	H
Before	6.6	6.5	9.0	10.3	11.3	8.1	6.3	11.6
After	6.8	2.4	7.4	8.5	8.1	6.1	3.4	2.0

Solution

Corresponding "before" and "after" values form matched pairs. (Calculate "after" – "before.")

After Data	Before Data	Difference
6.8	6.6	0.2
2.4	6.5	-4.1
7.4	9	-1.6
8.5	10.3	-1.8
8.1	11.3	-3.2
6.1	8.1	-2
3.4	6.3	-2.9
2	11.6	-9.6

The data for the test are the differences: {0.2, –4.1, –1.6, –1.8, –3.2, –2, –2.9, –9.6}

The sample mean and sample standard deviation of the differences are: $\overline{x_d} = -3.13$ and

$s_d = 2.91$

Verify these values.

Let μ_d be the population mean for the differences. We use the subscript d to denote "differences."

Random variable: $\overline{X}_d$ = the mean difference of the sensory measurements

H0: μd ≥ 0

The null hypothesis is zero or positive, meaning that there is the same or more pain felt after hypnotism. That means the subject shows no improvement. μd is the population mean of the differences.

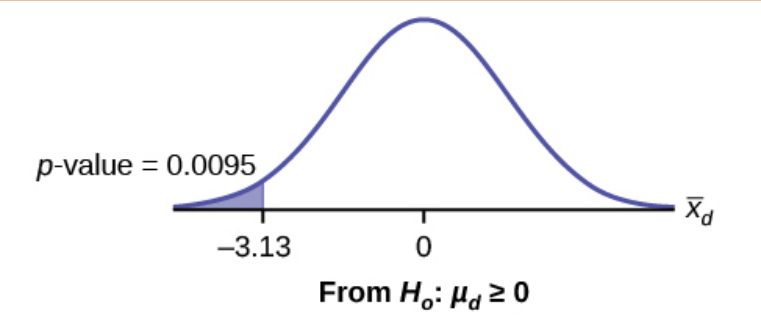
Ha: μd < 0

The alternative hypothesis is negative, meaning there is less pain felt after hypnotism. That means the subject shows improvement. The score should be lower after hypnotism, so the difference ought to be negative to indicate improvement.

Distribution for the test: The distribution is a Student's t with df = n – 1 = 8 – 1 = 7. Use t7. (Notice that the test is for a single population mean.)

Calculate the p-value using the Student's t-distribution: p-value = 0.0095

Graph:



$\overline{X}_d$ is the random variable for the differences.

The sample mean and sample standard deviation of the differences are:

$\overline{x}_d$ = –3.13

$\overline{s}_d$ = 2.91

Compare α and the p-value: α = 0.05 and p-value = 0.0095. α > p-value.

Make a decision: Since α > p-value, reject H0. This means that μd < 0 and there is improvement.

Conclusion: At a 5% level of significance, from the sample data, there is sufficient evidence to

conclude that the sensory measurements, on average, are lower after hypnotism. Hypnotism appears to be effective in reducing pain.

Note

For the TI-83+ and TI-84 calculators, you can either calculate the differences ahead of time (**after – before**) and put the differences into a list or you can put the **after** data into a first list and the **before** data into a second list. Then go to a third list and arrow up to the name. Enter 1st list name – 2nd list name. The calculator will do the subtraction, and you will have the differences in the third list.

Use your list of differences as the data. Press `STAT` and arrow over to `TESTS`. Press `2:T-Test`. Arrow over to `Data` and press `ENTER`. Arrow down and enter `0` for μ_0, the name of the list where you put the data, and `1` for Freq:. Arrow down to `μ:` and arrow over to $< \mu_0$. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The *p*-value is 0.0094, and the test statistic is -3.04. Do these instructions again except, arrow to `Draw` (instead of `Calculate`). Press `ENTER`.

Your Turn!

A study was conducted to investigate how effective a new diet was in lowering cholesterol. Results for the randomly selected subjects are shown in the table. The differences have a normal distribution. Are the subjects' cholesterol levels lower on average after the diet? Test at the 5% level.

Subject	A	B	C	D	E	F	G	H	I
Before	209	210	205	198	216	217	238	240	222
After	199	207	189	209	217	202	211	223	201

Solution

The *p*-value is 0.0130, so we can reject the null hypothesis. There is enough evidence to suggest that the diet lowers cholesterol.

Example

A college football coach was interested in whether the college's strength development class increased his players' maximum lift (in pounds) on the bench press exercise. He asked four of his players to participate in a study. The amount of weight they could each lift was recorded before they took the strength development class. After completing the class, the amount of weight they could each lift was again measured. The data are as follows:

Weight (in pounds)	**Player 1**	**Player 2**	**Player 3**	**Player 4**
Amount of weight lifted prior to the class	205	241	338	368
Amount of weight lifted after the class	295	252	330	360

The coach wants to know if the strength development class makes his players stronger, on average.

Record the differences data. Calculate the differences by subtracting the amount of weight lifted prior to the class from the weight lifted after completing the class. The data for the differences are: {90, 11, -8, -8}. Assume the differences have a normal distribution.

Using the differences data, calculate the sample mean and the sample standard deviation.

$\overline{x}_d$ = 21.3, sd = 46.7

Note:

The data given here would indicate that the distribution is actually right-skewed. The difference 90 may be an extreme outlier? It is pulling the sample mean to be 21.3 (positive). The means of the other three data values are actually negative.

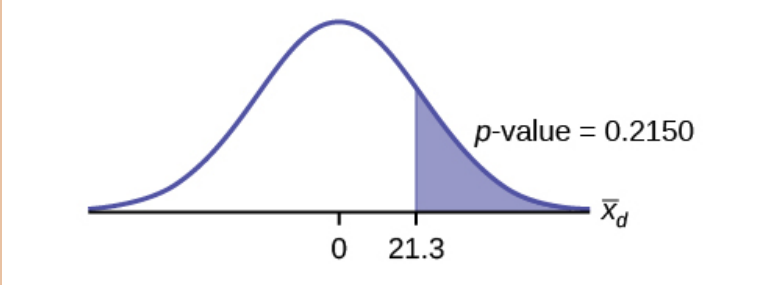
Using the difference data, this becomes a test of a single _____ (fill in the blank).

Define the random variable: $\overline{X}_d$ mean difference in the maximum lift per player.

The distribution for the hypothesis test is t3.

H0: μd ≤ 0, Ha: μd > 0

Graph:



Calculate the p-value: The p-value is 0.2150

Decision: If the level of significance is 5%, the decision is not to reject the null hypothesis, because **α** < p-value.

What is the conclusion?

At a 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the strength development class helped to make the players stronger, on average.

Your Turn!

A new prep class was designed to improve SAT test scores. Five students were selected at random. Their scores on two practice exams were recorded, one before the class and one after. The data recorded in [link]. Are the scores, on average, higher after the class? Test at a 5% level.

SAT Scores	Student 1	Student 2	Student 3	Student 4
Score before class	1840	1960	1920	2150
Score after class	1920	2160	2200	2100

Solution

The *p*-value is 0.0874, so we decline to reject the null hypothesis. The data do not support that the class improves SAT scores significantly.

Example

Seven eighth graders at Kennedy Middle School measured how far they could push the shot-put with their dominant (writing) hand and their weaker (non-writing) hand. They thought that they could push equal distances with either hand. The data were collected and recorded in [link].

Distance (in feet) using	Student 1	Student 2	Student 3	Student 4	Student 5	Student 6	Student 7
Dominant Hand	30	26	34	17	19	26	20
Weaker Hand	28	14	27	18	17	26	16

<!–MAGTULOY DITO...–>

Conduct a hypothesis test to determine whether the mean difference in distances between the children's dominant versus weaker hands is significant.

Record the differences data. Calculate the differences by subtracting the distances with the weaker hand from the distances with the dominant hand. The data for the differences are: {2, 12, 7, –1, 2, 0, 4}. The differences have a normal distribution.

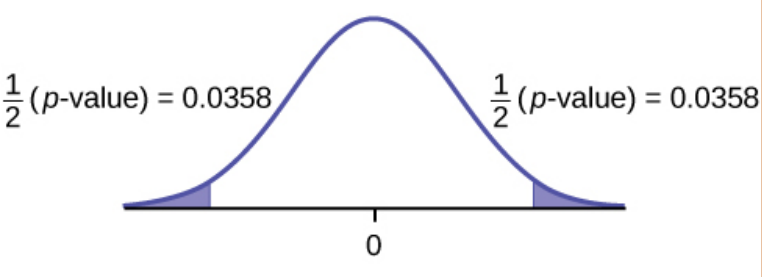
Using the differences data, calculate the sample mean and the sample standard deviation. $\overline{x}_d$ = 3.71, s_d = 4.5.

Random variable: $\overline{X}_d$ = mean difference in the distances between the hands.

Distribution for the hypothesis test:t6

H0: μd = 0 Ha: μd ≠ 0

Graph:



Calculate the p-value: The p-value is 0.0716 (using the data directly).

(test statistic = 2.18. p-value = 0.0719 using $(\overline{x}_d = 3.71,\ s_d = 4.5)$.)

Decision: Assume **α** = 0.05. Since **α** < p-value, Do not reject H0.

Conclusion: At the 5% level of significance, from the sample data, there is not sufficient evidence to conclude that there is a difference in the children's weaker and dominant hands to push the shot-put.

Your Turn!

Five ball players think they can throw the same distance with their dominant hand (throwing) and off-hand (catching hand). The data were collected and recorded in [link]. Conduct a hypothesis test to determine whether the mean difference in distances between the dominant and off-hand is significant. Test at the 5% level.

	Player 1	Player 2	Player 3	Player 4	Player 5
Dominant Hand	120	111	135	140	125
Off-hand	105	109	98	111	99

Solution

The *p*-level is 0.0230, so we can reject the null hypothesis. The data show that the players do not throw the same distance with their off-hands as they do with their dominant hands.



One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=924#oembed-1

CHAPTER 9 REVIEW

9.1 Chapter Review

Two population means from independent samples where the population standard deviations are not known

- Random Variable: $\overline{X}_1 - \overline{X}_2$ = the difference of the sampling means
- Distribution: Student's *t*-distribution with degrees of freedom (variances not pooled)

Formula Review

Standard error: $SE = \sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}$

Test statistic (*t*-score): $t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$

Degrees of freedom:

$$df = \frac{\left(\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}\right)^2}{\left(\frac{1}{n_1 - 1}\right)\left(\frac{(s_1)^2}{n_1}\right)^2 + \left(\frac{1}{n_2 - 1}\right)\left(\frac{(s_2)^2}{n_2}\right)^2}$$

where:

s_1 and s_2 are the sample standard deviations, and n_1 and n_2 are the sample sizes.

$\overline{x}_1$ and $\overline{x}_2$ are the sample means.

Cohen's *d* is the measure of effect size:

$$d = \frac{\overline{x}_1 - \overline{x}_2}{s_{pooled}}$$

where $s_{pooled} = \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}}$

9.2 Chapter Review

A hypothesis test of two population means from independent samples where the population standard deviations are known will have these characteristics:

- Random variable: $\overline{X_1} - \overline{X_2}$ = the difference of the means
- Distribution: normal distribution

Formula Review

Normal Distribution:

$$\overline{X}_1 - \overline{X}_2 \sim N\left[\mu_1 - \mu_2, \sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}\right].$$

Generally $\boldsymbol{\mu_1 - \mu_2 = 0.}$

Test Statistic (z-score):

$$z = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}}$$

Generally $\boldsymbol{\mu_1 - \mu_2 = 0.}$

where: σ_1 and σ_2 are the known population standard deviations. n_1 and n_2 are the sample sizes. x¯1x¯1″>x¯1 and x¯2x¯2″>x¯2 are the sample means. μ_1 and μ_2 are the population means.

9.3 Chapter Review

Test of two population proportions from independent samples.
Random variable: $p_A - p_B$ = difference between the two estimated proportions
Distribution: normal distribution

Formula Review

Pooled Proportion: $p_c = \frac{x_F + x_M}{n_F + n_M}$

Distribution for the differences:

$$p'_A - p'_B \sim N\left[0, \sqrt{p_c\,(1-p_c)\left(\frac{1}{n_A}+\frac{1}{n_B}\right)}\right]$$

where the null hypothesis is H_0: $p_A = p_B$ or H_0: $p_A - p_B = 0$.

Test Statistic (z-score): $$z = \frac{(p'_A - p'_B)}{\sqrt{p_c\,(1-p_c)\left(\frac{1}{n_A}+\frac{1}{n_B}\right)}}$$

where the null hypothesis is H_0: $p_A = p_B$ or H_0: $p_A - p_B = 0$.

where

p'_A and p'_B are the sample proportions, p_A and p_B are the population proportions,

P_c is the pooled proportion, and $\boldsymbol{n_A}$ and $\boldsymbol{n_B}$ are the sample sizes.

9.4 Chapter Review

A hypothesis test for matched or paired samples (t-test) has these characteristics:

- Test the differences by subtracting one measurement from the other measurement
- Random Variable: $\overline{x}_d$ = mean of the differences
- Distribution: Student's-t distribution with $n - 1$ degrees of freedom
- If the number of differences is small (less than 30), the differences must follow a normal distribution.
- Two samples are drawn from the same set of objects.
- Samples are dependent.

Formula Review

Test Statistic (t-score): $t = \dfrac{\overline{x}_d - \mu_d}{\left(\frac{s_d}{\sqrt{n}}\right)}$

where:

$\overline{x}_d$ is the mean of the sample differences. μ_d is the mean of the population differences. s_d is the sample standard deviation of the differences. n is the sample size.

CHAPTER 9 PRACTICE

9.1 Practice

Two Population Means with Unknown Standard Deviations

Use the following information to answer the next 15 exercises: Indicate if the hypothesis test is for

a. independent group means, population standard deviations, and/or variances known
b. independent group means, population standard deviations, and/or variances unknown
c. matched or paired samples
d. single mean
e. two proportions
f. single proportion

1.

It is believed that 70% of males pass their drivers test in the first attempt, while 65% of females pass the test in the first attempt. Of interest is whether the proportions are in fact equal.

2.

A new laundry detergent is tested on consumers. Of interest is the proportion of consumers who prefer the new brand over the leading competitor. A study is done to test this.

3.

A new windshield treatment claims to repel water more effectively. Ten windshields are tested by simulating rain without the new treatment. The same windshields are then treated, and the experiment is run again. A hypothesis test is conducted.

4.

The known standard deviation in salary for all mid-level professionals in the financial industry is $11,000. Company A and Company B are in the financial industry. Suppose samples are taken of mid-level professionals from Company A and from Company B. The sample mean salary for mid-level professionals in Company A

is $80,000. The sample mean salary for mid-level professionals in Company B is $96,000. Company A and Company B management want to know if their mid-level professionals are paid differently, on average.

5.

The average worker in Germany gets eight weeks of paid vacation.

6.

According to a television commercial, 80% of dentists agree that Ultra Fresh toothpaste is the best on the market.

7.

It is believed that the average grade on an English essay in a particular school system for females is higher than for males. A random sample of 31 females had a mean score of 82 with a standard deviation of three, and a random sample of 25 males had a mean score of 76 with a standard deviation of four.

8.

The league mean batting average is 0.280 with a known standard deviation of 0.06. The Rattlers and the Vikings belong to the league. The mean batting average for a sample of eight Rattlers is 0.210, and the mean batting average for a sample of eight Vikings is 0.260. There are 24 players on the Rattlers and 19 players on the Vikings. Are the batting averages of the Rattlers and Vikings statistically different?

9.

In a random sample of 100 forests in the United States, 56 were coniferous or contained conifers. In a random sample of 80 forests in Mexico, 40 were coniferous or contained conifers. Is the proportion of conifers in the United States statistically more than the proportion of conifers in Mexico?

10.

A new medicine is said to help improve sleep. Eight subjects are picked at random and given the medicine. The mean hours slept for each person were recorded before starting the medication and after.

11.

It is thought that teenagers sleep more than adults on average. A study is done to verify this. A sample of 16 teenagers has a mean of 8.9 hours slept and a standard deviation of 1.2. A sample of 12 adults has a mean of 6.9 hours slept and a standard deviation of 0.6.

12.

Varsity athletes practice five times a week, on average.

13.

A sample of 12 in-state graduate school programs at school A has a mean tuition of $64,000 with a standard deviation of $8,000. At school B, a sample of 16 in-state graduate programs has a mean of $80,000 with a standard deviation of $6,000. On average, are the mean tuitions different?

14.

A new WiFi range booster is being offered to consumers. A researcher tests the native range of 12 different routers under the same conditions. The ranges are recorded. Then the researcher uses the new WiFi range booster and records the new ranges. Does the new WiFi range booster do a better job?

15.

A high school principal claims that 30% of student athletes drive themselves to school, while 4% of non-athletes drive themselves to school. In a sample of 20 student athletes, 45% drive themselves to school. In a sample of 35 non-athlete students, 6% drive themselves to school. Is the percentage of student athletes who drive themselves to school more than the percentage of nonathletes?

Use the following information to answer the next three exercises: A study is done to determine which of two soft drinks has more sugar. There are 13 cans of Beverage A in a sample and six cans of Beverage B. The mean amount of sugar in Beverage A is 36 grams with a standard deviation of 0.6 grams. The mean amount of sugar in Beverage B is 38 grams with a standard deviation of 0.8 grams. The researchers believe that Beverage B has more sugar than Beverage A, on average. Both populations have normal distributions.

16.

Are standard deviations known or unknown?

17.

What is the random variable?

18.

Is this a one-tailed or two-tailed test?

Use the following information to answer the next 12 exercises: The U.S. Center for Disease Control reports that the mean life expectancy was 47.6 years for White people born in 1900 and 33.0 years for non-White people. Suppose that you randomly survey death records for people born in 1900 in a certain county. Of the 124 White people, the mean life span was 45.3 years with a standard deviation of 12.7 years. Of the 82 non-White people, the mean life span was 34.1 years with a standard deviation of 15.6 years. Conduct a hypothesis test to see if the mean life spans in the county were the same for White and non-White people.

19.

Is this a test of means or proportions?

20.

State the null and alternative hypotheses.

a. H_0: _____
b. H_a: _____

21.

Is this a right-tailed, left-tailed, or two-tailed test?

22.

In symbols, what is the random variable of interest for this test?

23.

In words, define the random variable of interest for this test.

24.

Which distribution (normal or Student's t) would you use for this hypothesis test?

25.

Explain why you chose the distribution you did for Exercise 10.24.

26.

Calculate the test statistic and p-value.

27.

Sketch a graph of the situation. Label the horizontal axis. Mark the hypothesized difference and the sample difference. Shade the area corresponding to the *p*-value.

28.

Find the *p*-value.

29.

At a pre-conceived $\alpha = 0.05$, what is your:

a. Decision:
b. Reason for the decision:
c. Conclusion (write out in a complete sentence):

30.

Does it appear that the means are the same? Why or why not?

9.2 Practice

Use the following information to answer the next five exercises. The mean speeds of fastball pitches from two different baseball pitchers are to be compared. A sample of 14 fastball pitches is measured from each pitcher. The populations have normal distributions. [link] shows the result. Scouters believe that Rodriguez pitches a speedier fastball.

Pitcher	Sample Mean Speed of Pitches (mph)	Population Standard Deviation
Wesley	86	3
Rodriguez	91	7

What is the random variable?
Solution

The difference in mean speeds of the fastball pitches of the two pitchers

State the null and alternative hypotheses.

What is the test statistic?
Solution

−2.46

What is the *p*-value?

At the 1% significance level, what is your conclusion?
Solution

At the 1% significance level, we can reject the null hypothesis. There is sufficient data to conclude that the mean speed of Rodriguez's fastball is faster than Wesley's.

Use the following information to answer the next five exercises. A researcher is testing the effects of plant food on plant growth. Nine plants have been given the plant food. Another nine plants have not been given the plant food. The heights of the plants are recorded after eight weeks. The populations have normal distributions. The following table is the result. The researcher thinks the food makes the plants grow taller.

Plant Group	Sample Mean Height of Plants (inches)	Population Standard Deviation
Food	16	2.5
No food	14	1.5

Is the population standard deviation known or unknown?

State the null and alternative hypotheses.
Solution

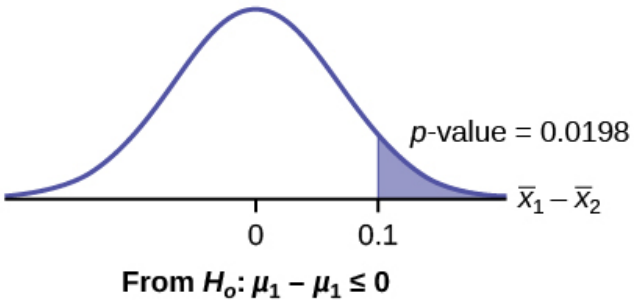
Subscripts: 1 = Food, 2 = No Food

H_0: $\mu_1 \leq \mu_2$

H_a: $\mu_1 > \mu_2$

What is the *p*-value?

Draw the graph of the *p*-value.
Solution



At the 1% significance level, what is your conclusion?

Use the following information to answer the next five exercises. Two metal alloys are being considered as material for ball bearings. The mean melting point of the two alloys is to be compared. 15 pieces of each metal are being tested. Both populations have normal distributions. The following table is the result. It is believed that Alloy Zeta has a different melting point.

	Sample Mean Melting Temperatures (°F)	**Population Standard Deviation**
Alloy Gamma	800	95
Alloy Zeta	900	105

State the null and alternative hypotheses.
Solution

Subscripts: 1 = Gamma, 2 = Zeta

H_0: $\mu_1 = \mu_2$

H_a: $\mu_1 \neq \mu_2$

Is this a right-, left-, or two-tailed test?

What is the *p*-value?
Solution

0.0062

Draw the graph of the *p*-value.

At the 1% significance level, what is your conclusion?
Solution

There is sufficient evidence to reject the null hypothesis. The data support that the melting point for Alloy Zeta is different from the melting point of Alloy Gamma.

9.3

Use the following information for the next five exercises. Two types of phone operating system are being tested to determine if there is a difference in the proportions of system failures (crashes). Fifteen out of a random sample of 150 phones with OS_1 had system failures within the first eight hours of operation. Nine out of another random sample of 150 phones with OS_2 had system failures within the first eight hours of operation. OS_2 is believed to be more stable (have fewer crashes) than OS_1.

Is this a test of means or proportions?

What is the random variable?
Solution

$P'_{OS1} - P'_{OS2}$ = difference in the proportions of phones that had system failures within the first eight hours of operation with OS_1 and OS_2.

State the null and alternative hypotheses.

What is the *p*-value?
Solution

0.1018

What can you conclude about the two operating systems?

9.3 Practice

Use the following information to answer the next twelve exercises. In the recent Census, three percent of the U.S. population reported being of two or more races. However, the percent varies tremendously from state to state. Suppose that two random surveys are conducted. In the first random survey, out of 1,000 North Dakotans, only nine people reported being of two or more races. In the second random survey, out of 500 Nevadans, 17 people reported being of two or more races. Conduct a hypothesis test to determine if the population percentages are the same for the two states or if the percent for Nevada is statistically higher than for North Dakota.

Is this a test of means or proportions?

Solution

proportions

State the null and alternative hypotheses.

1. H_0: _____
2. H_a: _____

Is this a right-tailed, left-tailed, or two-tailed test? How do you know?
Solution

right-tailed

What is the random variable of interest for this test?

In words, define the random variable for this test.
Solution

The random variable is the difference in proportions (percentages) of the populations that are of two or more races in Nevada and North Dakota.

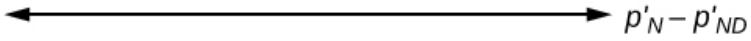
Which distribution (normal or Student's t) would you use for this hypothesis test?

Explain why you chose the distribution you did for the Exercise 10.56.
Solution

Our sample sizes are much greater than five each, so we use the normal for two proportions distribution for this hypothesis test.

Calculate the test statistic.

Sketch a graph of the situation. Mark the hypothesized difference and the sample difference. Shade the area corresponding to the p-value.



Solution

Check student's solution.

Find the p-value.

At a pre-conceived $\alpha = 0.05$, what is your:
Decision:
Reason for the decision:
Conclusion (write out in a complete sentence):
Solution

1. Reject the null hypothesis.
2. p-value < alpha
3. At the 5% significance level, there is sufficient evidence to conclude that the proportion (percent) of the population that is of two or more races in Nevada is statistically higher than that in North Dakota.

Does it appear that the proportion of Nevadans who are two or more races is higher than the proportion of North Dakotans? Why or why not?

9.4 Practice

Use the following information to answer the next five exercises. A study was conducted to test the effectiveness of a software patch in reducing system failures over a six-month period. Results for randomly selected installations are shown in [link]. The "before" value is matched to an "after" value, and the differences are calculated. The differences have a normal distribution. Test at the 1% significance level.

Installation	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
Before	3	6	4	2	5	8	2	6
After	1	5	2	0	1	0	2	2

What is the random variable?
Solution

the mean difference of the system failures

State the null and alternative hypotheses.

What is the p-value?
Solution

0.0067

Draw the graph of the *p*-value.

What conclusion can you draw about the software patch?
Solution

With a *p*-value 0.0067, we can reject the null hypothesis. There is enough evidence to support that the software patch is effective in reducing the number of system failures.

Use the following information to answer the next five exercises. A study was conducted to test the effectiveness of a juggling class. Before the class started, six subjects juggled as many balls as they could at once. After the class, the same six subjects juggled as many balls as they could. The differences in the number of balls are calculated. The differences have a normal distribution. Test at the 1% significance level.

Subject	**A**	**B**	**C**	**D**	**E**	**F**
Before	3	4	3	2	4	5
After	4	5	6	4	5	7

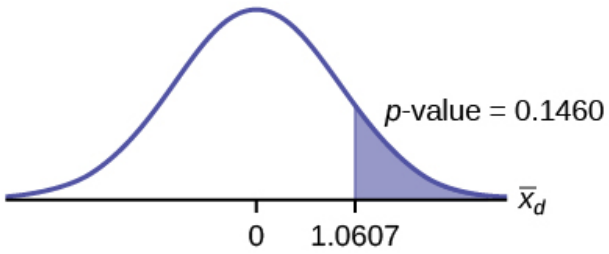
State the null and alternative hypotheses.

What is the *p*-value?
Solution

0.0021

What is the sample mean difference?

Draw the graph of the *p*-value.
Solution



What conclusion can you draw about the juggling class?

Use the following information to answer the next five exercises. A doctor wants to know if a blood pressure medication is effective. Six subjects have their blood pressures recorded. After twelve weeks on the medication,

the same six subjects have their blood pressure recorded again. For this test, only systolic pressure is of concern. Test at the 1% significance level.

Patient	**A**	**B**	**C**	**D**	**E**	**F**
Before	161	162	165	162	166	171
After	158	159	166	160	167	169

State the null and alternative hypotheses.
Solution

H_0: $\mu_d \geq 0$
H_a: $\mu_d < 0$

What is the test statistic?

What is the p-value?
Solution

0.0699

What is the sample mean difference?

What is the conclusion?
Solution

We decline to reject the null hypothesis. There is not sufficient evidence to support that the medication is effective.

CHAPTER 9 HOMEWORK

9.1 Homework

DIRECTIONS: For each of the word problems, use a solution sheet to do the hypothesis test. The solution sheet is found in Appendix E. Please feel free to make copies of the solution sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.

Note

If you are using a Student's *t*-distribution for a homework problem in what follows, including for paired data, you may assume that the underlying population is normally distributed. (When using these tests in a real situation, you must first prove that assumption, however.)

The mean number of English courses taken in a two–year time period by male and female college students is believed to be about the same. An experiment is conducted and data are collected from 29 males and 16 females. The males took an average of three English courses with a standard deviation of 0.8. The females took an average of four English courses with a standard deviation of 1.0. Are the means statistically the same?

A student at a four-year college claims that mean enrollment at four–year colleges is higher than at two–year colleges in the United States. Two surveys are conducted. Of the 35 two–year colleges surveyed, the mean enrollment was 5,068 with a standard deviation of 4,777. Of the 35 four-year colleges surveyed, the mean enrollment was 5,466 with a standard deviation of 8,191.
Solution

Subscripts: 1: two-year colleges; 2: four-year colleges

1. $H_0: \mu_1 \geq \mu_2$

2. $H_a: \mu_1 < \mu_2$
3. $\overline{X}_1 - \overline{X}_2$ is the difference between the mean enrollments of the two-year colleges and the four-year colleges.
4. Student's-*t*
5. test statistic: -0.2480
6. *p*-value: 0.4019
7. Check student's solution.
    1. Alpha: 0.05
    2. Decision: Do not reject
    3. Reason for Decision: *p*-value > alpha
    4. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the mean enrollment at four-year colleges is higher than at two-year colleges.

At Rachel's 11th birthday party, eight girls were timed to see how long (in seconds) they could hold their breath in a relaxed position. After a two-minute rest, they timed themselves while jumping. The girls thought that the mean difference between their jumping and relaxed times would be zero. Test their hypothesis.

Relaxed time (seconds)	Jumping time (seconds)
26	21
47	40
30	28
22	21
23	25
45	43
37	35
29	32

Mean entry-level salaries for college graduates with mechanical engineering degrees and electrical engineering degrees are believed to be approximately the same. A recruiting office thinks that the mean mechanical engineering salary is actually lower than the mean electrical engineering salary. The recruiting office randomly surveys 50 entry level mechanical engineers and 60 entry level electrical engineers. Their mean salaries were $ 46,100 and $ 46,700, respectively. Their standard deviations were $ 3,450 and $ 4,210, respectively. Conduct a hypothesis test to determine if you agree that the mean entry-level mechanical engineering salary is lower than the mean entry-level electrical engineering salary.
Solution

Subscripts: 1: mechanical engineering; 2: electrical engineering

1. $H_0: \mu_1 \geq \mu_2$
2. $H_a: \mu_1 < \mu_2$
3. $\overline{X}_1 - \overline{X}_2$ is the difference between the mean entry level salaries of mechanical engineers and electrical engineers.
4. t_{108}
5. test statistic: $t = -0.82$
6. p-value: 0.2061
7. Check student's solution.
    1. Alpha: 0.05
    2. Decision: Do not reject the null hypothesis.
    3. Reason for Decision: p-value > alpha
    4. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that the mean entry-level salaries of mechanical engineers is lower than that of electrical engineers.

Marketing companies have collected data implying that teenage girls use more ringtones on their cellular phones than teenage boys do. In one particular study of 40 randomly chosen teenage girls and boys (20 of each) with cellular phones, the mean number of ringtones for the girls was 3.2 with a standard deviation of 1.5. The mean for the boys was 1.7 with a standard deviation of 0.8. Conduct a hypothesis test to determine if the means are approximately the same or if the girls' mean is higher than the boys' mean.

Use the information from [link] to answer the next four exercises.

Using the data from Lap 1 only, conduct a hypothesis test to determine if the mean time for completing a lap in races is the same as it is in practices.
Solution

1. $H_0: \mu_1 = \mu_2$
2. $H_a: \mu_1 \neq \mu_2$
3. $\overline{X}_1 - \overline{X}_2$ is the difference between the mean times for completing a lap in races and in practices.
4. $t_{20.32}$
5. test statistic: -4.70
6. p-value: 0.0001
7. Check student's solution.
    1. Alpha: 0.05
    2. Decision: Reject the null hypothesis.
    3. Reason for Decision: p-value < alpha

4. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the mean time for completing a lap in races is different from that in practices.

Repeat the test in Exercise 10.83, but use Lap 5 data this time.

Repeat the test in Exercise 10.83, but this time combine the data from Laps 1 and 5.
Solution

1. $H_0: \mu_1 = \mu_2$
2. $H_a: \mu_1 \neq \mu_2$
3. is the difference between the mean times for completing a lap in races and in practices.
4. $t_{40.94}$
5. test statistic: –5.08
6. p-value: zero
7. Check student's solution.
    1. Alpha: 0.05
    2. Decision: Reject the null hypothesis.
    3. Reason for Decision: p-value < alpha
    4. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the mean time for completing a lap in races is different from that in practices.

In two to three complete sentences, explain in detail how you might use Terri Vogel's data to answer the following question. "Does Terri Vogel drive faster in races than she does in practices?"

Use the following information to answer the next two exercises. The Eastern and Western Major League Soccer conferences have a new Reserve Division that allows new players to develop their skills. Data for a randomly picked date showed the following annual goals.

Western	**Eastern**
Los Angeles 9	D.C. United 9
FC Dallas 3	Chicago 8
Chivas USA 4	Columbus 7
Real Salt Lake 3	New England 6
Colorado 4	MetroStars 5
San Jose 4	Kansas City 3

Conduct a hypothesis test to answer the next two exercises.

The **exact** distribution for the hypothesis test is:

1. the normal distribution
2. the Student's *t*-distribution
3. the uniform distribution
4. the exponential distribution

If the level of significance is 0.05, the conclusion is:

1. There is sufficient evidence to conclude that the **W** Division teams score fewer goals, on average, than the **E** teams
2. There is insufficient evidence to conclude that the **W** Division teams score more goals, on average, than the **E** teams.
3. There is insufficient evidence to conclude that the **W** teams score fewer goals, on average, than the **E** teams score.
4. Unable to determine

Solution

c

Suppose a statistics instructor believes that there is no significant difference between the mean class scores of statistics day students on Exam 2 and statistics night students on Exam 2. She takes random samples from each of the populations. The mean and standard deviation for 35 statistics day students were 75.86 and 16.91. The mean and standard deviation for 37 statistics night students were 75.41 and 19.73. The "day" subscript refers to the statistics day students. The "night" subscript refers to the statistics night students. A concluding statement is:

1. There is sufficient evidence to conclude that statistics night students' mean on Exam 2 is better than the statistics day students' mean on Exam 2.
2. There is insufficient evidence to conclude that the statistics day students' mean on Exam 2 is better than the statistics night students' mean on Exam 2.
3. There is insufficient evidence to conclude that there is a significant difference between the means of the statistics day students and night students on Exam 2.
4. There is sufficient evidence to conclude that there is a significant difference between the means of the statistics day students and night students on Exam 2.

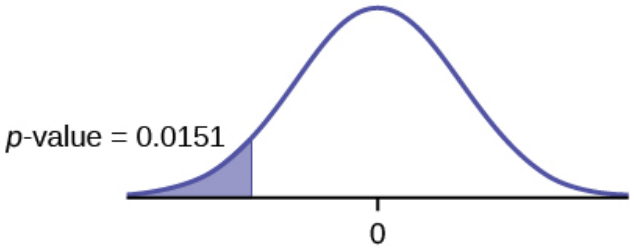
Researchers interviewed street prostitutes in Canada and the United States. The mean age of the 100 Canadian

prostitutes upon entering prostitution was 18 with a standard deviation of six. The mean age of the 130 United States prostitutes upon entering prostitution was 20 with a standard deviation of eight. Is the mean age of entering prostitution in Canada lower than the mean age in the United States? Test at a 1% significance level.
Solution

Test: two independent sample means, population standard deviations unknown.

Random variable: $\overline{X}_1 - \overline{X}_2$

Distribution: H_0: $\mu_1 = \mu_2 H_a$: $\mu_1 < \mu_2$ The mean age of entering prostitution in Canada is lower than the mean age in the United States.



Graph: left-tailed

p-value : 0.0151

Decision: Do not reject H_0.

Conclusion: At the 1% level of significance, from the sample data, there is not sufficient evidence to conclude that the mean age of entering prostitution in Canada is lower than the mean age in the United States.

A powder diet is tested on 49 people, and a liquid diet is tested on 36 different people. Of interest is whether the liquid diet yields a higher mean weight loss than the powder diet. The powder diet group had a mean weight loss of 42 pounds with a standard deviation of 12 pounds. The liquid diet group had a mean weight loss of 45 pounds with a standard deviation of 14 pounds.

Suppose a statistics instructor believes that there is no significant difference between the mean class scores of statistics day students on Exam 2 and statistics night students on Exam 2. She takes random samples from each of the populations. The mean and standard deviation for 35 statistics day students were 75.86 and 16.91, respectively. The mean and standard deviation for 37 statistics night students were 75.41 and 19.73. The "day" subscript refers to the statistics day students. The "night" subscript refers to the statistics night students. An appropriate alternative hypothesis for the hypothesis test is:

1. $\mu_{day} > \mu_{night}$
2. $\mu_{day} < \mu_{night}$
3. $\mu_{day} = \mu_{night}$
4. $\mu_{day} \neq \mu_{night}$

Solution

d

9.2 Homework

DIRECTIONS: For each of the word problems, use a solution sheet to do the hypothesis test. The solution sheet is found in [link]. Please feel free to make copies of the solution sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.

NOTE

If you are using a Student's *t*-distribution for one of the following homework problems, including for paired data, you may assume that the underlying population is normally distributed. (When using these tests in a real situation, you must first prove that assumption, however.)

A study is done to determine if students in the California state university system take longer to graduate, on average, than students enrolled in private universities. One hundred students from both the California state university system and private universities are surveyed. Suppose that from years of research, it is known that the population standard deviations are 1.5811 years and 1 year, respectively. The following data are collected. The California state university system students took on average 4.5 years with a standard deviation of 0.8. The private university students took on average 4.1 years with a standard deviation of 0.3.

Parents of teenage boys often complain that auto insurance costs more, on average, for teenage boys than for teenage girls. A group of concerned parents examines a random sample of insurance bills. The mean annual cost for 36 teenage boys was $679. For 23 teenage girls, it was $559. From past years, it is known that the population standard deviation for each group is $180. Determine whether or not you believe that the mean cost for auto insurance for teenage boys is greater than that for teenage girls.
Solution

Subscripts: 1 = boys, 2 = girls

1. $H_0: \mu_1 \leq \mu_2$
2. $H_a: \mu_1 > \mu_2$

3. The random variable is the difference in the mean auto insurance costs for boys and girls.
4. normal
5. test statistic: $z = 2.50$
6. p-value: 0.0062
7. Check student's solution.
8.
    1. Alpha: 0.05
    2. Decision: Reject the null hypothesis.
    3. Reason for Decision: p-value < alpha
    4. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the mean cost of auto insurance for teenage boys is greater than that for girls.

A group of transfer bound students wondered if they will spend the same mean amount on texts and supplies each year at their four-year university as they have at their community college. They conducted a random survey of 54 students at their community college and 66 students at their local four-year university. The sample means were $ 947 and $ 1,011, respectively. The population standard deviations are known to be $ 254 and $ 87, respectively. Conduct a hypothesis test to determine if the means are statistically the same.

Some manufacturers claim that non-hybrid sedan cars have a lower mean miles-per-gallon (mpg) than hybrid ones. Suppose that consumers test 21 hybrid sedans and get a mean of 31 mpg with a standard deviation of seven mpg. Thirty-one non-hybrid sedans get a mean of 22 mpg with a standard deviation of four mpg. Suppose that the population standard deviations are known to be six and three, respectively. Conduct a hypothesis test to evaluate the manufacturers' claim.
Solution

Subscripts: 1 = non-hybrid sedans, 2 = hybrid sedans

1. $H_0: \mu_1 \geq \mu_2$
2. $H_a: \mu_1 < \mu_2$
3. The random variable is the difference in the mean miles per gallon of non-hybrid sedans and hybrid sedans.
4. normal
5. test statistic: 6.36
6. p-value: 0
7. Check student's solution.
8.
    1. Alpha: 0.05
    2. Decision: Reject the null hypothesis.
    3. Reason for decision: p-value < alpha
    4. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the mean

miles per gallon of non-hybrid sedans is less than that of hybrid sedans.

A baseball fan wanted to know if there is a difference between the number of games played in a World Series when the American League won the series versus when the National League won the series. From 1922 to 2012, the population standard deviation of games won by the American League was 1.14, and the population standard deviation of games won by the National League was 1.11. Of 19 randomly selected World Series games won by the American League, the mean number of games won was 5.76. The mean number of 17 randomly selected games won by the National League was 5.42. Conduct a hypothesis test.

One of the questions in a study of marital satisfaction of dual-career couples was to rate the statement "I'm pleased with the way we divide the responsibilities for childcare." The ratings went from one (strongly agree) to five (strongly disagree). [link] contains ten of the paired responses for husbands and wives. Conduct a hypothesis test to see if the mean difference in the husband's versus the wife's satisfaction level is negative (meaning that, within the partnership, the husband is happier than the wife).

Wife's Score	2	2	3	3	4	2	1	1	2	4
Husband's Score	2	2	1	3	2	1	1	1	2	4

Solution

1. $H_0: \mu_d = 0$
2. $H_a: \mu_d < 0$
3. The random variable X_d is the average difference between husband's and wife's satisfaction level.
4. t_9
5. test statistic: $t = -1.86$
6. p-value: 0.0479
7. Check student's solution
8.
    1. Alpha: 0.05
    2. Decision: Reject the null hypothesis, but run another test.
    3. Reason for Decision: p-value < alpha
    4. Conclusion: This is a weak test because alpha and the p-value are close. However, there is insufficient evidence to conclude that the mean difference is negative.

9.3 Homework

DIRECTIONS: For each of the word problems, use a solution sheet to do the hypothesis test. The solution sheet is

found in [link]. Please feel free to make copies of the solution sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.

Note

If you are using a Student's *t*-distribution for one of the following homework problems, including for paired data, you may assume that the underlying population is normally distributed. (In general, you must first prove that assumption, however.)

A recent drug survey showed an increase in the use of drugs and alcohol among local high school seniors as compared to the national percent. Suppose that a survey of 100 local seniors and 100 national seniors is conducted to see if the proportion of drug and alcohol use is higher locally than nationally. Locally, 65 seniors reported using drugs or alcohol within the past month, while 60 national seniors reported using them.

We are interested in whether the proportions of female suicide victims for ages 15 to 24 are the same for the whites and the blacks races in the United States. We randomly pick one year, 1992, to compare the races. The number of suicides estimated in the United States in 1992 for white females is 4,930. Five hundred eighty were aged 15 to 24. The estimate for black females is 330. Forty were aged 15 to 24. We will let female suicide victims be our population.
Solution

1. H_0: $P_W = P_B$
2. H_a: $P_W \neq P_B$
3. The random variable is the difference in the proportions of white and black suicide victims, aged 15 to 24.
4. normal for two proportions
5. test statistic: –0.1944
6. *p*-value: 0.8458
7. Check student's solution.
    1. Alpha: 0.05
    2. Decision: Reject the null hypothesis.
    3. Reason for decision: *p*-value > alpha
    4. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that the proportions of white and black female suicide victims, aged 15 to 24, are different.

Elizabeth Mjelde, an art history professor, was interested in whether the value from the Golden Ratio formula, $\left(\frac{\text{larger} + \text{smaller dimension}}{\text{larger dimension}}\right)$ was the same in the Whitney Exhibit for works from 1900 to 1919 as

for works from 1920 to 1942. Thirty-seven early works were sampled, averaging 1.74 with a standard deviation of 0.11. Sixty-five of the later works were sampled, averaging 1.746 with a standard deviation of 0.1064. Do you think that there is a significant difference in the Golden Ratio calculation?

A recent year was randomly picked from 1985 to the present. In that year, there were 2,051 Hispanic students at Cabrillo College out of a total of 12,328 students. At Lake Tahoe College, there were 321 Hispanic students out of a total of 2,441 students. In general, do you think that the percent of Hispanic students at the two colleges is basically the same or different?
Solution

Subscripts: 1 = Cabrillo College, 2 = Lake Tahoe College

1. $H_0: p_1 = p_2$
2. $H_a: p_1 \neq p_2$
3. The random variable is the difference between the proportions of Hispanic students at Cabrillo College and Lake Tahoe College.
4. normal for two proportions
5. test statistic: 4.29
6. *p*-value: 0.00002
7. Check student's solution.
    1. Alpha: 0.05
    2. Decision: Reject the null hypothesis.
    3. Reason for decision: *p*-value < alpha
    4. Conclusion: There is sufficient evidence to conclude that the proportions of Hispanic students at Cabrillo College and Lake Tahoe College are different.

Use the following information to answer the next three exercises. Neuroinvasive West Nile virus is a severe disease that affects a person's nervous system. It is spread by the Culex species of mosquito. In the United States in 2010 there were 629 reported cases of neuroinvasive West Nile virus out of a total of 1,021 reported cases and there were 486 neuroinvasive reported cases out of a total of 712 cases reported in 2011. Is the 2011 proportion of neuroinvasive West Nile virus cases more than the 2010 proportion of neuroinvasive West Nile virus cases? Using a 1% level of significance, conduct an appropriate hypothesis test.
"2011" subscript: 2011 group.
"2010" subscript: 2010 group

This is:

1. a test of two proportions

2. a test of two independent means
3. a test of a single mean
4. a test of matched pairs.

An appropriate null hypothesis is:

1. $p_{2011} \leq p_{2010}$
2. $p_{2011} \geq p_{2010}$
3. $\mu_{2011} \leq \mu_{2010}$
4. $p_{2011} > p_{2010}$

Solution

a

The p-value is 0.0022. At a 1% level of significance, the appropriate conclusion is

1. There is sufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuroinvasive West Nile disease is less than the proportion of people in the United States in 2010 who contracted neuroinvasive West Nile disease.
2. There is insufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuroinvasive West Nile disease is more than the proportion of people in the United States in 2010 who contracted neuroinvasive West Nile disease.
3. There is insufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuroinvasive West Nile disease is less than the proportion of people in the United States in 2010 who contracted neuroinvasive West Nile disease.
4. There is sufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuroinvasive West Nile disease is more than the proportion of people in the United States in 2010 who contracted neuroinvasive West Nile disease.

Researchers conducted a study to find out if there is a difference in the use of eReaders by different age groups. Randomly selected participants were divided into two age groups. In the 16- to 29-year-old group, 7% of the 628 surveyed use eReaders, while 11% of the 2,309 participants 30 years old and older use eReaders.
Solution

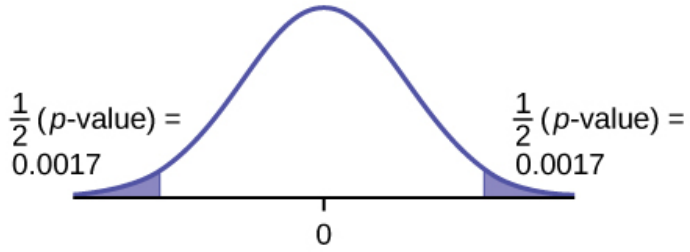
Test: two independent sample proportions.
Random variable: $p'_1 - p'_2$
Distribution:

H_0: $p_1 = p_2$

H_a: $p_1 \neq p_2$

The proportion of eReader users is different for the 16- to 29-year-old users from that of the 30 and older users.

Graph: two-tailed



p-value : 0.0033

Decision: Reject the null hypothesis.

Conclusion: At the 5% level of significance, from the sample data, there is sufficient evidence to conclude that the proportion of eReader users 16 to 29 years old is different from the proportion of eReader users 30 and older.

Adults aged 18 years old and older were randomly selected for a survey on obesity. Adults are considered obese if their body mass index (BMI) is at least 30. The researchers wanted to determine if the proportion of women who are obese in the south is less than the proportion of southern men who are obese. The results are shown in [link]. Test at the 1% level of significance.

	Number who are obese	**Sample size**
Men	42,769	155,525
Women	67,169	248,775

Two computer users were discussing tablet computers. A higher proportion of people ages 16 to 29 use tablets than the proportion of people age 30 and older. [link] details the number of tablet owners for each age group. Test at the 1% level of significance.

	16–29 year olds	**30 years old and older**
Own a Tablet	69	231
Sample Size	628	2,309

Solution

Test: two independent sample proportions

Random variable: $p'_1 - p'_2$

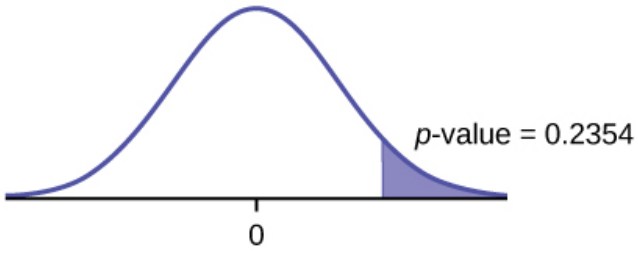
Distribution:

H_0: $p_1 = p_2$

H_a: $p_1 > p_2$

A higher proportion of tablet owners are aged 16 to 29 years old than are 30 years old and older.

Graph: right-tailed



p-value: 0.2354

Decision: Do not reject the H_0.

Conclusion: At the 1% level of significance, from the sample data, there is not sufficient evidence to conclude that a higher proportion of tablet owners are aged 16 to 29 years old than are 30 years old and older.

A group of friends debated whether more men use smartphones than women. They consulted a research study of smartphone use among adults. The results of the survey indicate that of the 973 men randomly sampled, 379 use smartphones. For women, 404 of the 1,304 who were randomly sampled use smartphones. Test at the 5% level of significance.

While her husband spent 2½ hours picking out new speakers, a statistician decided to determine whether the percentage of men who enjoy shopping for electronic equipment is higher than the percentage of women who enjoy shopping for electronic equipment. The population was Saturday afternoon shoppers. Out of 67 men, 24 said they enjoyed the activity. Eight of the 24 women surveyed claimed to enjoy the activity. Interpret the results of the survey.

Solution

Subscripts: 1: men; 2: women

1. H_0: $p_1 \leq p_2$
2. H_a: $p_1 > p_2$
3. $P'_1 - P'_2$ is the difference between the proportions of men and women who enjoy shopping for electronic equipment.

4. normal for two proportions
5. test statistic: 0.22
6. *p*-value: 0.4133
7. Check student's solution.
    1. Alpha: 0.05
    2. Decision: Do not reject the null hypothesis.
    3. Reason for Decision: *p*-value > alpha
    4. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that the proportion of men who enjoy shopping for electronic equipment is more than the proportion of women.

We are interested in whether children's educational computer software costs less, on average, than children's entertainment software. Thirty-six educational software titles were randomly picked from a catalog. The mean cost was $31.14 with a standard deviation of $4.69. Thirty-five entertainment software titles were randomly picked from the same catalog. The mean cost was $33.86 with a standard deviation of $10.87. Decide whether children's educational software costs less, on average, than children's entertainment software.

Joan Nguyen recently claimed that the proportion of college-age males with at least one pierced ear is as high as the proportion of college-age females. She conducted a survey in her classes. Out of 107 males, 20 had at least one pierced ear. Out of 92 females, 47 had at least one pierced ear. Do you believe that the proportion of males has reached the proportion of females?
Solution

1. $H_0: p_1 = p_2$
2. $H_a: p_1 \neq p_2$
3. $P'_1 - P'_2$ is the difference between the proportions of men and women that have at least one pierced ear.
4. normal for two proportions
5. test statistic: –4.82
6. *p*-value: zero
7. Check student's solution.
    1. Alpha: 0.05
    2. Decision: Reject the null hypothesis.
    3. Reason for Decision: *p*-value < alpha
    4. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the proportions of males and females with at least one pierced ear is different.

Use the data sets found in [link] to answer this exercise. Is the proportion of race laps Terri completes slower than 130 seconds less than the proportion of practice laps she completes slower than 135 seconds?

"To Breakfast or Not to Breakfast?" by Richard Ayore

In American society, birthdays are one of those days that everyone looks forward to. People of different ages and peer groups gather to mark the 18th, 20th, ..., birthdays. During this time, one looks back to see what he or she has achieved for the past year and also focuses ahead for more to come.

If, by any chance, I am invited to one of these parties, my experience is always different. Instead of dancing around with my friends while the music is booming, I get carried away by memories of my family back home in Kenya. I remember the good times I had with my brothers and sister while we did our daily routine.

Every morning, I remember we went to the shamba (garden) to weed our crops. I remember one day arguing with my brother as to why he always remained behind just to join us an hour later. In his defense, he said that he preferred waiting for breakfast before he came to weed. He said, "This is why I always work more hours than you guys!"

And so, to prove him wrong or right, we decided to give it a try. One day we went to work as usual without breakfast, and recorded the time we could work before getting tired and stopping. On the next day, we all ate breakfast before going to work. We recorded how long we worked again before getting tired and stopping. Of interest was our mean increase in work time. Though not sure, my brother insisted that it was more than two hours. Using the data in [link], solve our problem.

Work hours with breakfast	Work hours without breakfast
8	6
7	5
9	5
5	4
9	7
8	7
10	7
7	5
6	6
9	5

Solution

1. $H_0: \mu_d = 0$

2. H_a: $\mu_d > 0$
3. The random variable X_d is the mean difference in work times on days when eating breakfast and on days when not eating breakfast.
4. t_9
5. test statistic: 4.8963
6. p-value: 0.0004
7. Check student's solution.
    1. Alpha: 0.05
    2. Decision: Reject the null hypothesis.
    3. Reason for Decision: p-value < alpha
    4. Conclusion: At the 5% level of significance, there is sufficient evidence to conclude that the mean difference in work times on days when eating breakfast and on days when not eating breakfast has increased.

9.4 Homework

DIRECTIONS: For each of the word problems, use a solution sheet to do the hypothesis test. The solution sheet is found in Appendix E. Please feel free to make copies of the solution sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.

Note

If you are using a Student's t-distribution for the homework problems, including for paired data, you may assume that the underlying population is normally distributed. (When using these tests in a real situation, you must first prove that assumption, however.)

Ten individuals went on a low–fat diet for 12 weeks to lower their cholesterol. The data are recorded in [link]. Do you think that their cholesterol levels were significantly lowered?

Starting cholesterol level	Ending cholesterol level
140	140
220	230
110	120
240	220
200	190
180	150
190	200
360	300
280	300
260	240

Solution

p-value = 0.1494

At the 5% significance level, there is insufficient evidence to conclude that the medication lowered cholesterol levels after 12 weeks.

Use the following information to answer the next two exercises. A new AIDS prevention drug was tried on a group of 224 HIV positive patients. Forty-five patients developed AIDS after four years. In a control group of 224 HIV positive patients, 68 developed AIDS after four years. We want to test whether the method of treatment reduces the proportion of patients that develop AIDS after four years or if the proportions of the treated group and the untreated group stay the same.

Let the subscript t = treated patient and ut = untreated patient.

The appropriate hypotheses are:

1. $H_0: p_t < p_{ut}$ and $H_a: p_t \geq p_{ut}$
2. $H_0: p_t \leq p_{ut}$ and $H_a: p_t > p_{ut}$
3. $H_0: p_t = p_{ut}$ and $H_a: p_t \neq p_{ut}$
4. $H_0: p_t = p_{ut}$ and $H_a: p_t < p_{ut}$

If the p-value is 0.0062 what is the conclusion (use $\alpha = 0.05$)?

1. The method has no effect.
2. There is sufficient evidence to conclude that the method reduces the proportion of HIV positive

patients who develop AIDS after four years.
3. There is sufficient evidence to conclude that the method increases the proportion of HIV positive patients who develop AIDS after four years.
4. There is insufficient evidence to conclude that the method reduces the proportion of HIV positive patients who develop AIDS after four years.

Solution

b

Use the following information to answer the next two exercises. An experiment is conducted to show that blood pressure can be consciously reduced in people trained in a "biofeedback exercise program." Six subjects were randomly selected and blood pressure measurements were recorded before and after the training. The difference between blood pressures was calculated (after – before) producing the following results: $\overline{x}_d = -10.2$ $s_d = 8.4$. Using the data, test the hypothesis that the blood pressure has decreased after the training.

The distribution for the test is:

1. t_5
2. t_6
3. $N(-10.2, 8.4)$
4. $N(-10.2, \frac{8.4}{\sqrt{6}})$

If $\alpha = 0.05$, the *p*-value and the conclusion are

1. 0.0014; There is sufficient evidence to conclude that the blood pressure decreased after the training.
2. 0.0014; There is sufficient evidence to conclude that the blood pressure increased after the training.
3. 0.0155; There is sufficient evidence to conclude that the blood pressure decreased after the training.
4. 0.0155; There is sufficient evidence to conclude that the blood pressure increased after the training.

Solution

c

A golf instructor is interested in determining if her new technique for improving players' golf scores is effective. She takes four new students. She records their 18-hole scores before learning the technique and then after having taken her class. She conducts a hypothesis test. The data are as follows.

	Player 1	Player 2	Player 3	Player 4
Mean score before class	83	78	93	87
Mean score after class	80	80	86	86

The correct decision is:

1. Reject H_0.
2. Do not reject the H_0.

A local cancer support group believes that the estimate for new female breast cancer cases in the south is higher in 2013 than in 2012. The group compared the estimates of new female breast cancer cases by southern state in 2012 and in 2013. The results are in [link].

Southern States	2012	2013
Alabama	3,450	3,720
Arkansas	2,150	2,280
Florida	15,540	15,710
Georgia	6,970	7,310
Kentucky	3,160	3,300
Louisiana	3,320	3,630
Mississippi	1,990	2,080
North Carolina	7,090	7,430
Oklahoma	2,630	2,690
South Carolina	3,570	3,580
Tennessee	4,680	5,070
Texas	15,050	14,980
Virginia	6,190	6,280

Solution

Test: two matched pairs or paired samples (t-test)

Random variable: $\overline{X}_d$

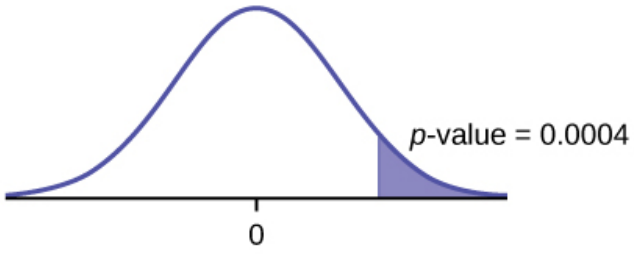
Distribution: t_{12}

H_0: $\mu_d = 0$ H_a: $\mu_d > 0$

The mean of the differences of new female breast cancer cases in the south between 2013 and 2012 is greater than zero. The estimate for new female breast cancer cases in the south is higher in 2013 than in 2012.

Graph: right-tailed

p-value: 0.0004



Decision: Reject H_0

Conclusion: At the 5% level of significance, from the sample data, there is sufficient evidence to conclude that there was a higher estimate of new female breast cancer cases in 2013 than in 2012.

A traveler wanted to know if the prices of hotels are different in the ten cities that he visits the most often. The list of the cities with the corresponding hotel prices for his two favorite hotel chains is in [link]. Test at the 1% level of significance.

Cities	Hyatt Regency prices in dollars	Hilton prices in dollars
Atlanta	107	169
Boston	358	289
Chicago	209	299
Dallas	209	198
Denver	167	169
Indianapolis	179	214
Los Angeles	179	169
New York City	625	459
Philadelphia	179	159
Washington, DC	245	239

A politician asked his staff to determine whether the underemployment rate in the northeast decreased from 2011 to 2012. The results are in [link].

Northeastern States	2011	2012
Connecticut	17.3	16.4
Delaware	17.4	13.7
Maine	19.3	16.1
Maryland	16.0	15.5
Massachusetts	17.6	18.2
New Hampshire	15.4	13.5
New Jersey	19.2	18.7
New York	18.5	18.7
Ohio	18.2	18.8
Pennsylvania	16.5	16.9
Rhode Island	20.7	22.4
Vermont	14.7	12.3
West Virginia	15.5	17.3

Solution

Test: matched or paired samples (*t*-test)

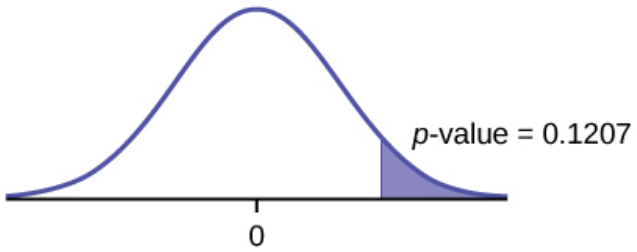
Difference data: {–0.9, –3.7, –3.2, –0.5, 0.6, –1.9, –0.5, 0.2, 0.6, 0.4, 1.7, –2.4, 1.8}

Random Variable: $\overline{X}_d$

Distribution: H_0: $\mu_d = 0$ H_a: $\mu_d < 0$

The mean of the differences of the rate of underemployment in the northeastern states between 2012 and 2011 is less than zero. The underemployment rate went down from 2011 to 2012.

Graph: left-tailed.



p-value: 0.1207

Decision: Do not reject H_0.

Conclusion: At the 5% level of significance, from the sample data, there is not sufficient evidence to conclude that there was a decrease in the underemployment rates of the northeastern states from 2011 to 2012.

Bringing It Together

Use the following information to answer the next ten exercises. indicate which of the following choices best identifies the hypothesis test.

1. independent group means, population standard deviations and/or variances known
2. independent group means, population standard deviations and/or variances unknown
3. matched or paired samples
4. single mean
5. two proportions
6. single proportion

A powder diet is tested on 49 people, and a liquid diet is tested on 36 different people. The population standard deviations are two pounds and three pounds, respectively. Of interest is whether the liquid diet yields a higher mean weight loss than the powder diet.

A new chocolate bar is taste-tested on consumers. Of interest is whether the proportion of children who like the new chocolate bar is greater than the proportion of adults who like it.
Solution

e

The mean number of English courses taken in a two–year time period by male and female college students is believed to be about the same. An experiment is conducted and data are collected from nine males and 16 females.

A football league reported that the mean number of touchdowns per game was five. A study is done to determine if the mean number of touchdowns has decreased.
Solution

d

A study is done to determine if students in the California state university system take longer to graduate than students enrolled in private universities. One hundred students from both the California state university system and private universities are surveyed. From years of research, it is known that the population standard deviations are 1.5811 years and one year, respectively.

According to a YWCA Rape Crisis Center newsletter, 75% of rape victims know their attackers. A study is done to verify this.
Solution

f

According to a recent study, U.S. companies have a mean maternity-leave of six weeks.

A recent drug survey showed an increase in use of drugs and alcohol among local high school students as compared to the national percent. Suppose that a survey of 100 local youths and 100 national youths is conducted to see if the proportion of drug and alcohol use is higher locally than nationally.
Solution

e

A new SAT study course is tested on 12 individuals. Pre-course and post-course scores are recorded. Of interest is the mean increase in SAT scores. The following data are collected:

Pre-course score	Post-course score
1	300
960	920
1010	1100
840	880
1100	1070
1250	1320
860	860
1330	1370
790	770
990	1040
1110	1200
740	850

University of Michigan researchers reported in the JOURNAL OF THE NATIONAL CANCER INSTITUTE that quitting smoking is especially beneficial for those under age 49. In this American Cancer Society study, the risk (probability) of dying of lung cancer was about the same as for those who had never smoked.
Solution

f

Lesley E. Tan investigated the relationship between left-handedness vs. right-handedness and motor

competence in preschool children. Random samples of 41 left-handed preschool children and 41 right-handed preschool children were given several tests of motor skills to determine if there is evidence of a difference between the children based on this experiment. The experiment produced the means and standard deviations shown [link]. Determine the appropriate test and best distribution to use for that test.

	Left-handed	Right-handed
Sample size	41	41
Sample mean	97.5	98.1
Sample standard deviation	17.5	19.2

1. Two independent means, normal distribution
2. Two independent means, Student's-t distribution
3. Matched or paired samples, Student's-t distribution
4. Two population proportions, normal distribution

A golf instructor is interested in determining if her new technique for improving players' golf scores is effective. She takes four (4) new students. She records their 18-hole scores before learning the technique and then after having taken her class. She conducts a hypothesis test. The data are as [link].

	Player 1	**Player 2**	**Player 3**	**Player 4**
Mean score before class	83	78	93	87
Mean score after class	80	80	86	86

This is:

1. a test of two independent means.
2. a test of two proportions.
3. a test of a single mean.
4. a test of a single proportion.

Solution

a

CHAPTER 10: LINEAR REGRESSION AND CORRELATION

INTRODUCTION TO CHAPTER 10: LINEAR REGRESSION AND CORRELATION



Figure 12.1. Linear regression and correlation can help you determine if an auto mechanic's salary is related to his work experience. (credit: Joshua Rothhaas)

Professionals often want to know how two or more numeric variables are related. For example, is there a relationship between the grade on the second math exam a student takes and the grade on the final exam? If there is a relationship, what is the relationship and how strong is it?

In another example, your income may be determined by your education, your profession, your years of experience, and your ability. The amount you pay a repair person for labor is often determined by an initial amount plus an hourly fee.

The type of data described in the examples is **bivariate** data — "bi" for two variables. In reality, statisticians use **multivariate** data, meaning many variables.

In this chapter, you will be studying the simplest form of regression, "linear regression" with one

independent variable (x). This involves data that fits a line in two dimensions. You will also study correlation which measures how strong the relationship is.

Collaborative Exercises

Section 10.3

Collaborative Exercise

If you know a person's pinky (smallest) finger length, do you think you could predict that person's height? Collect data from your class (pinky finger length, in inches).

The independent variable, x, is pinky finger length and the dependent variable, y, is height. For each set of data, plot the points on graph paper. Make your graph big enough and **use a ruler**.

Then "by eye," draw a line that appears to "fit" the data. For your line, pick two convenient points and use them to find the slope of the line. Find the y-intercept of the line by extending your line so it crosses the y-axis.

Using the slopes and the y-intercepts, write your equation of "best fit." Do you think everyone will have the same equation? Why or why not? According to your equation, what is the predicted height for a pinky length of 2.5 inches?

Media Attributions

- Chapter 12 Introduction Figure © OpenStax Introductory Statistics is licensed under a CC BY (Attribution) license

10.1 LINEAR EQUATIONS

Learning Objectives

By the end of this section, the student should be able to:

- Interpret the independent and dependent variable
- Calculate and interpret the slope and y intercept of a linear equation

Linear regression for two variables is based on a linear equation with one independent variable. The equation has the form:

$$y = a + bx$$

where a and b are constant numbers.

The variable x **is the independent variable, and** y **is the dependent variable.** Typically, you choose a value to substitute for the independent variable and then solve for the dependent variable.

Example

The following examples are linear equations.

$$y = 3 + 2x$$
$$y = -0.01 + 1.2x$$

Try It

Is the following an example of a linear equation?

$y = -0.125 - 3.5x$

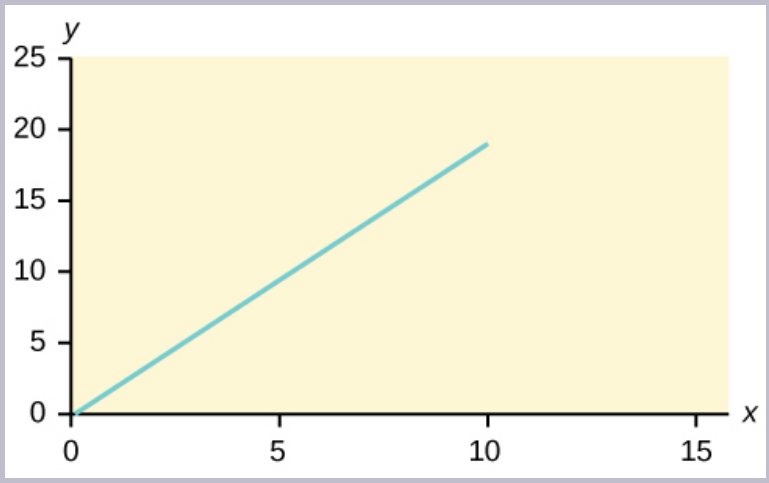
Solution

yes

The graph of a linear equation of the form $y = a + bx$ is a **straight line**. Any line that is not vertical can be described by this equation.

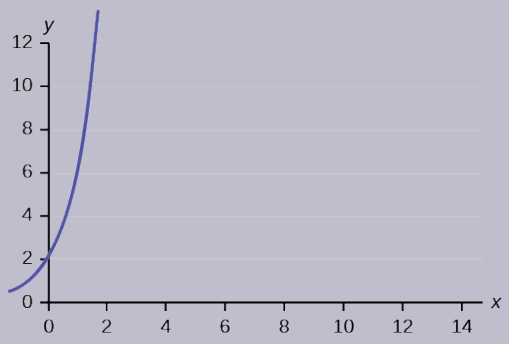
Try It

Graph the equation $y = -1 + 2x$.



Try It

Is the following an example of a linear equation? Why or why not?



Solution

No, the graph is not a straight line; therefore, it is not a linear equation.

Try It

Aaron's Word Processing Service (AWPS) does word processing. The rate for services is $ 32 per hour plus a $ 31.50 one-time charge. The total cost to a customer depends on the number of hours it takes to complete the job.

Find the equation that expresses the **total cost** in terms of the **number of hours** required to complete the job.

Solution
Let x = the number of hours it takes to get the job done.

Let y = the total cost to the customer.

The $ 31.50 is a fixed cost. If it takes x hours to complete the job, then (32)(x) is the cost of the word processing only. The total cost is: $y = 31.50 + 32x$

Try It

Emma's Extreme Sports hires hang-gliding instructors and pays them a fee of $ 50 per class as well as $ 20 per student in the class. The total cost Emma pays depends on the number of students in a class. Find the equation that expresses the total cost in terms of the number of students in a class.

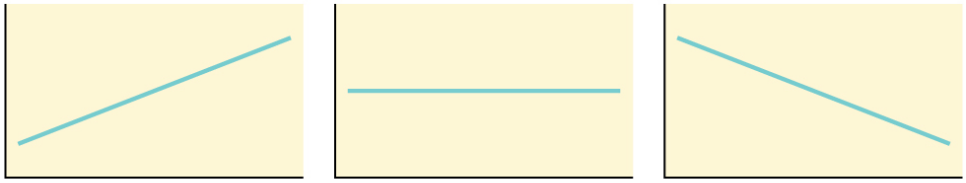
Solution
$y = 50 + 20x$

Slope and *Y*-Intercept of a Linear Equation

For the linear equation $y = a + bx$, b = slope and a = y-intercept. From algebra, recall that the slope is a number that describes the steepness of a line, and the y-intercept is the y coordinate of the point $(0, a)$ where the line crosses the y-axis.

Three possible graphs of $y = a + bx$. (a) If $b > 0$, the line slopes upward to the right. (b) If $b = 0$, the line is horizontal. (c) If $b < 0$, the line slopes downward to the right.

0 and so the line slopes upward to the right. For the second, b = 0 and the graph of the equation is a horizontal line. In the third graph, (c), b < 0 and the line slopes downward to the right.">



(a) (b) (c)

Example

Svetlana tutors to make extra money for college. For each tutoring session, she charges a one-time fee of $25 plus $15 per hour of tutoring. A linear equation that expresses the total amount of money Svetlana earns for each session she tutors is $y = 25 + 15x$.

What are the independent and dependent variables? What is the y-intercept and what is the slope? Interpret them using complete sentences.

Solution

The independent variable (x) is the number of hours Svetlana tutors each session. The dependent variable (y) is the amount, in dollars, Svetlana earns for each session.

The y-intercept is 25 ($a = 25$). At the start of the tutoring session, Svetlana charges a one-time fee of $25 (this is when $x = 0$). The slope is 15 ($b = 15$). For each session, Svetlana earns $15 for each hour she tutors.

Try It

Ethan repairs household appliances like dishwashers and refrigerators. For each visit, he charges $25 plus $20 per hour of work. A linear equation that expresses the total amount of money Ethan earns per visit is $y = 25 + 20x$.

What are the independent and dependent variables? What is the y-intercept and what is the slope? Interpret them using complete sentences.

Solution

The independent variable (x) is the number of hours Ethan works each visit. The dependent variable (y) is the amount, in dollars, Ethan earns for each visit.

The y-intercept is 25 ($a = 25$). At the start of a visit, Ethan charges a one-time fee of $25 (this is when $x = 0$). The slope is 20 ($b = 20$). For each visit, Ethan earns $20 for each hour he works.

References

Data from the Centers for Disease Control and Prevention.

Data from the National Center for HIV, STD, and TB Prevention.

10.2 SCATTER PLOTS

Learning Objectives

By the end of this section, the student should be able to:

- Construct and interpret scatter plots.

Before we take up the discussion of linear regression and correlation, we need to examine a way to display the relation between two variables x and y. The most common and easiest way is a **scatter plot**. The following example illustrates a scatter plot.

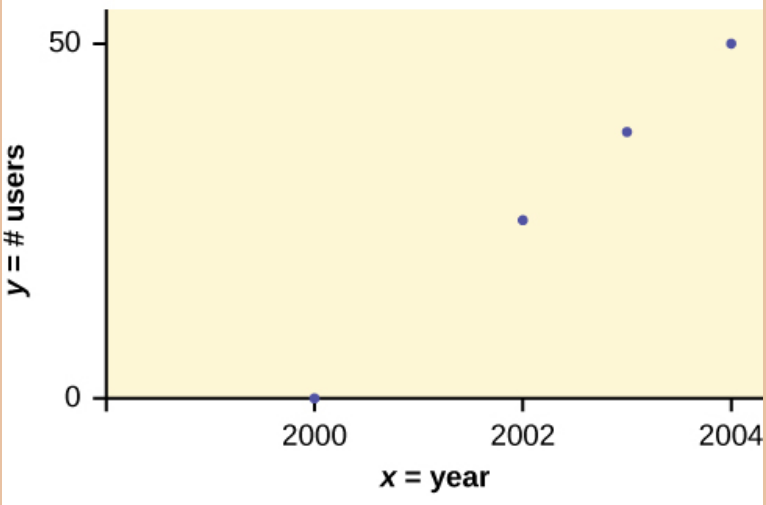
Example

In Europe and Asia, m-commerce is popular. M-commerce users have special mobile phones that work like electronic wallets as well as provide phone and Internet services. Users can do everything from paying for parking to buying a TV set or soda from a machine to banking to checking sports scores on the Internet. For the years 2000 through 2004, was there a relationship between the year and the number of m-commerce users? Construct a scatter plot. Let x = the year and let y = the number of m-commerce users, in millions.

Table showing the number of m-commerce users (in millions) by year.

x (year)	y (# of users)
2000	0.5
2002	20.0
2003	33.0
2004	47.0

Scatter plot showing the number of m-commerce users (in millions) by year.



To create a scatter plot:

1. Enter your X data into list L1 and your Y data into list L2.
2. Press 2nd STATPLOT ENTER to use Plot 1. On the input screen for PLOT 1, highlight On and press ENTER. (Make sure the other plots are OFF.)
3. For TYPE: highlight the very first icon, which is the scatter plot, and press ENTER.
4. For Xlist:, enter L1 ENTER and for Ylist: L2 ENTER.
5. For Mark: it does not matter which symbol you highlight, but the square is the easiest to see. Press ENTER.
6. Make sure there are no other equations that could be plotted. Press Y = and clear any equations out.
7. Press the ZOOM key and then the number 9 (for menu item "ZoomStat"); the calculator will fit the window to the data. You can press WINDOW to see the scaling of the axes.

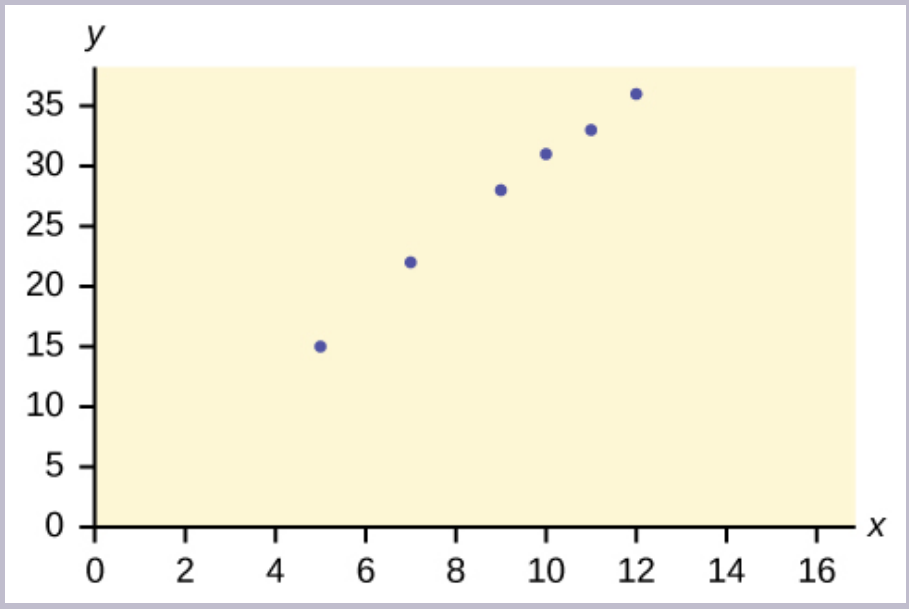
Try It

Amelia plays basketball for her high school. She wants to improve to play at the college level. She notices that the number of points she scores in a game goes up in response to the number of hours she practices her jump shot each week. She records the following data:

X (hours practicing jump shot)	*Y* (points scored in a game)
5	15
7	22
9	28
10	31
11	33
12	36

Construct a scatter plot and state if what Amelia thinks appears to be true.

Solution

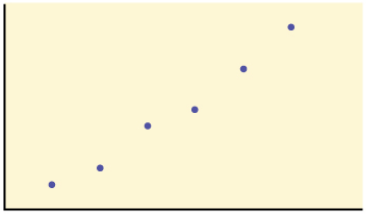


Yes, Amelia's assumption appears to be correct. The number of points Amelia scores per game goes up when she practices her jump shot more.

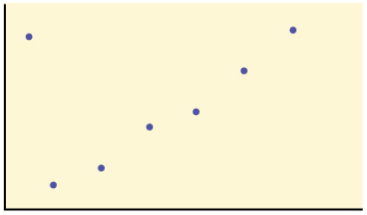
A scatter plot shows the **direction** of a relationship between the variables. A clear direction happens when there is either:
High values of one variable occurring with high values of the other variable or low values of one variable occurring with low values of the other variable.
High values of one variable occurring with low values of the other variable.

You can determine the **strength** of the relationship by looking at the scatter plot and seeing how close the points are to a line, a power function, an exponential function, or some other type of function. For a linear relationship there is an exception. Consider a scatter plot where all the points fall on a horizontal line providing a "perfect fit." The horizontal line would in fact show no relationship.

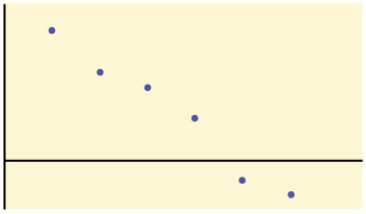
When you look at a scatter plot, you want to notice the **overall pattern** and any **deviations** from the pattern. The following scatter plot examples illustrate these concepts.



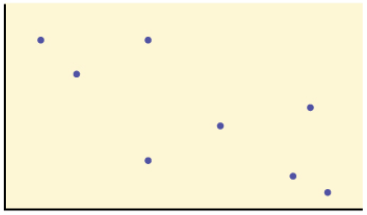
(a) Positive linear pattern (strong)



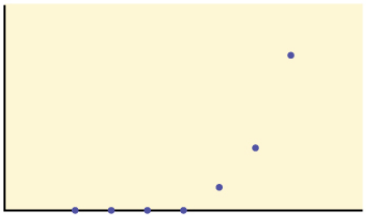
(b) Linear pattern w/ one deviation



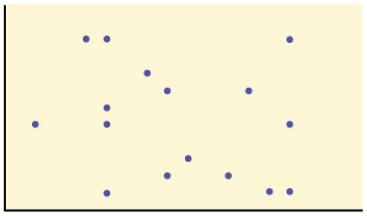
(a) Negative linear pattern (strong)



(b) Negative linear pattern (weak)



(a) Exponential growth pattern



(b) No pattern

In this chapter, we are interested in scatter plots that show a linear pattern. Linear patterns are quite common. The linear relationship is strong if the points are close to a straight line, except in the case of a horizontal line where there is no relationship. If we think that the points show a linear relationship, we would like to draw a line on the scatter plot. This line can be calculated through a process called linear regression. However, we only calculate a regression line if one of the variables helps to explain or predict the other variable. If x is the independent variable and y the dependent variable,
then we can use a regression line to predict y for a given value of x

10.3 THE REGRESSION EQUATION

Learning Objectives

By the end of this section, the student should be able to:

- Create and interpret the line of best fit.
- Calculate and interpret the correlation coefficient.
- Calculate and interpret the coefficient of determination.

Data rarely fit a straight line exactly. Usually, you must be satisfied with rough predictions. Typically, you have a set of data whose scatter plot appears to **"fit"** a straight line. This is called a Line of Best Fit **or** Least-Squares Line.

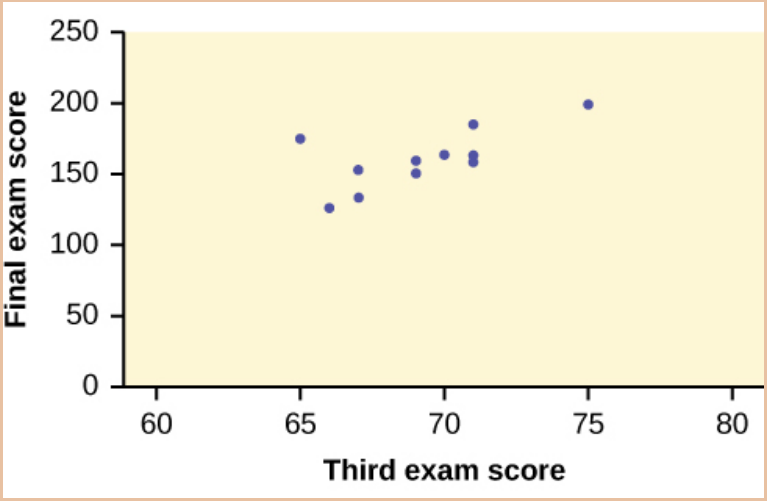
Example

A random sample of 11 statistics students produced the following data, where x is the third exam score out of 80, and y is the final exam score out of 200. Can you predict the final exam score of a random student if you know the third exam score?

Table showing the scores on the final exam based on scores from the third exam.

x (third exam score)	y (final exam score)
65	175
67	133
71	185
71	163
66	126
75	198
67	153
70	163
71	159
69	151
69	159

Scatter plot showing the scores on the final exam based on scores from the third exam.



Try It

SCUBA divers have maximum dive times they cannot exceed when going to different depths. The data in [link] show different depths with the maximum dive times in minutes. Use your calculator to find the least squares regression line and predict the maximum dive time for 110 feet.

***X* (depth in feet)**	***Y* (maximum dive time)**
50	80
60	55
70	45
80	35
90	25
100	22

Solution

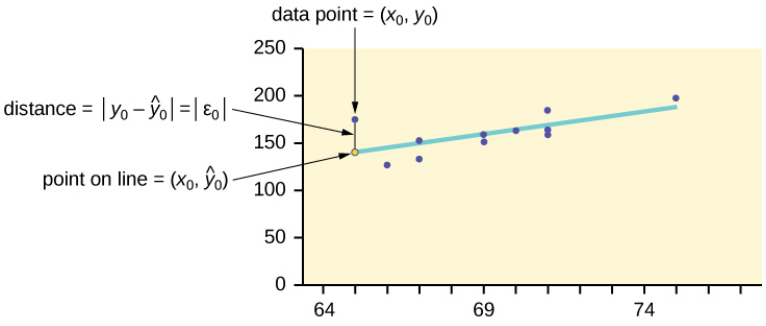
$\hat{y} = 127.24 - 1.11x$

At 110 feet, a diver could dive for only five minutes.

The third exam score, x, is the independent variable, and the final exam score, y, is the dependent variable. We will plot a regression line that best "fits" the data. If each of you were to fit a line "by eye," you would draw different lines. We can use what is called a least-squares regression line to obtain the best fit line.

Consider the following diagram. Each point of data is of the form (x, y) and each point of the line of best fit using least-squares linear regression has the form $(x, \hat{y})$.

The $\hat{y}$ is read **"y hat"** and is the **estimated value of y**. It is the value of y obtained using the regression line. It is not generally equal to y from data.



The term $y_0 - \hat{y}_0 = \varepsilon_0$ is called the **"error" or residual**. It is not an error in the sense of a mistake. The absolute value of a residual measures the vertical distance between the actual value of y and the estimated value of y. In other words, it measures the vertical distance between the actual data point and the predicted point on the line.

If the observed data point lies above the line, the residual is positive, and the line underestimates the actual data value for y. If the observed data point lies below the line, the residual is negative, and the line overestimates that actual data value for y.

In the diagram in [link], $y_0 - \hat{y}_0 = \varepsilon_0$ is the residual for the point shown. Here the point lies above the line and the residual is positive.

ε = the Greek letter **epsilon**

For each data point, you can calculate the residuals or errors, $y_i - \hat{y}_i = \varepsilon_i$ for i = 1, 2, 3, ..., 11.

Each $|\varepsilon|$ is a vertical distance.

For the example about the third exam scores and the final exam scores for the 11 statistics students, there are 11 data points. Therefore, there are 11 ε values. If you square each ε and add, you get

$$(\epsilon_1)^2 + (\epsilon_2)^2 + \ldots + (\epsilon_{11})^2 = \sum_{i=1}^{11} \epsilon^2$$

This is called the Sum of Squared Errors (SSE).

Using calculus, you can determine the values of a and b that make the **SSE** a minimum. When you make the **SSE** a minimum, you have determined the points that are on the line of best fit. It turns out that the line of best fit has the equation:

$$\hat{y} = a + bx$$

where $a = \overline{y} - b\overline{x}$ and $b = \dfrac{\Sigma\,(x - \overline{x})\,(y - \overline{y})}{\Sigma(x - \overline{x})^2}$.

The sample means of the x values and the y values are $\overline{x}$ and $\overline{y}$, respectively. The best fit line always passes through the point $(\overline{x}, \overline{y})$.

The slope b can be written as $b = r\left(\dfrac{s_y}{s_x}\right)$ where s_y = the standard deviation of the y values and s_x = the standard deviation of the x values. r is the **correlation coefficient**, which is discussed in the next section.

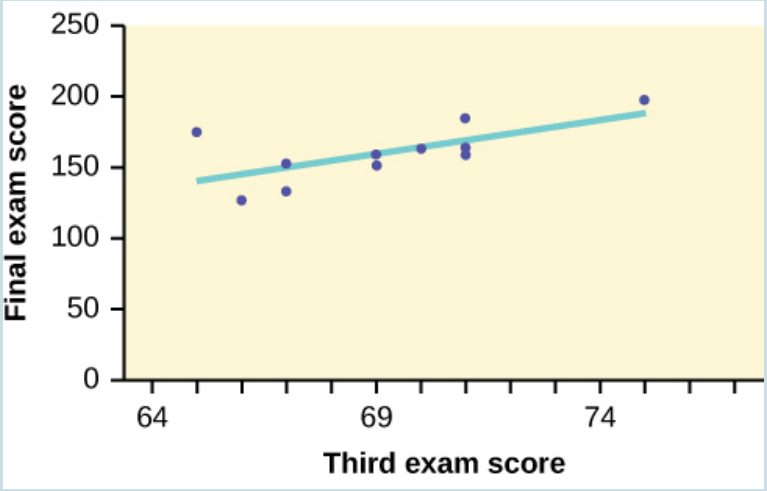
Least Squares Criteria for Best Fit

The process of fitting the best-fit line is called **linear regression**. The idea behind finding the best-fit line is based on the assumption that the data are scattered about a straight line. The criteria for the best fit line is that the sum of the squared errors (SSE) is minimized, that is, made as small as possible. Any other line you might choose would have a higher SSE than the best fit line. This best fit line is called the **least-squares regression line**.

Note

Computer spreadsheets, statistical software, and many calculators can quickly calculate the best-fit line and create the graphs. The calculations tend to be tedious if done by hand. Instructions to use the TI-83, TI-83+, and TI-84+ calculators to find the best-fit line and create a scatterplot are shown at the end of this section.

THIRD EXAM vs FINAL EXAM EXAMPLE:
The graph of the line of best fit for the third-exam/final-exam example is as follows:



The least squares regression line (best-fit line) for the third-exam/final-exam example has the equation:

Formula does not parse

Reminder

Remember, it is always important to plot a scatter diagram first. If the scatter plot indicates that there is a linear relationship between the variables, then it is reasonable to use a best fit line to make predictions for y given x within the domain of x-values in the sample data, **but not necessarily for x-values outside that domain.** You could use the line to predict the final exam score for a student who earned a grade of 73 on the third exam. You should NOT use the line to predict the final exam score for a student who earned a grade of 50 on the third exam, because 50 is not within the domain of the x-values in the sample data, which are between 65 and 75.

Understanding Slope

The slope of the line, b, describes how changes in the variables are related. It is important to interpret the slope of the line in the context of the situation represented by the data. You should be able to write a sentence interpreting the slope in plain English.

INTERPRETATION OF THE SLOPE: The slope of the best-fit line tells us how the dependent variable (y) changes for every one unit increase in the independent (x) variable, on average.

THIRD EXAM vs FINAL EXAM EXAMPLE

Slope: The slope of the line is $b = 4.83$.

Interpretation: For a one-point increase in the score on the third exam, the final exam score increases by 4.83 points, on average.

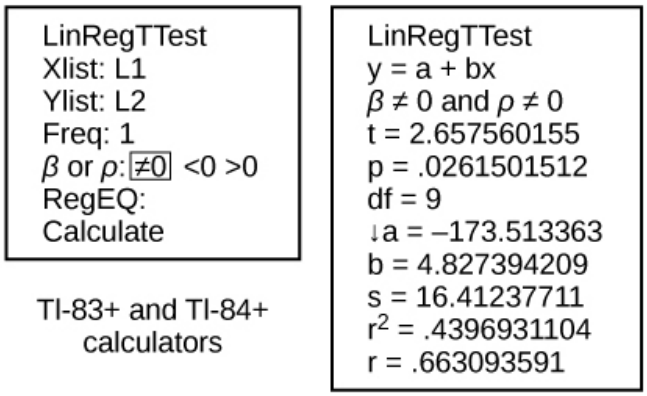
Using the Linear Regression T Test: LinRegTTest

1. In the STAT list editor, enter the X data in list L1 and the Y data in list L2, paired so that the corresponding (x,y) values are next to each other in the lists. (If a particular pair of values is repeated, enter it as many times as it appears in the data.)
2. On the STAT TESTS menu, scroll down with the cursor to select the LinRegTTest. (Be careful to select

LinRegTTest, as some calculators may also have a different item called LinRegTInt.)

3. On the LinRegTTest input screen enter: Xlist: L1; Ylist: L2; Freq: 1
4. On the next line, at the prompt β or ρ, highlight "$\neq 0$" and press ENTER
5. Leave the line for "RegEq:" blank
6. Highlight Calculate and press ENTER.

LinRegTTest Input Screen and Output Screen



The output screen contains a lot of information. For now we will focus on a few items from the output, and will return later to the other items.

The second line says $y = a + bx$. Scroll down to find the values $a = -173.513$, and $b = 4.8273$; the equation of the best fit line is $\hat{y} = -173.51 + 4.83x$

The two items at the bottom are $r_2 = 0.43969$ and $r = 0.663$. For now, just note where to find these values; we will discuss them in the next two sections.

Graphing the Scatterplot and Regression Line

1. We are assuming your X data is already entered in list L1 and your Y data is in list L2.
2. Press 2nd STATPLOT ENTER to use Plot 1.
3. On the input screen for PLOT 1, highlight **On**, and press ENTER.
4. For TYPE: highlight the very first icon which is the scatterplot and press ENTER.
5. Indicate Xlist: L1 and Ylist: L2.
6. For Mark: it does not matter which symbol you highlight.
7. Press the ZOOM key and then the number 9 (for menu item "ZoomStat"); the calculator will fit the window to the data.
8. To graph the best-fit line, press the "Y=" key and type the equation –173.5 + 4.83X into equation Y1. (The X key is immediately left of the STAT key). Press ZOOM 9 again to graph it.
9. Optional: If you want to change the viewing window, press the WINDOW key. Enter your desired

window using Xmin, Xmax, Ymin, Ymax.

Note

Another way to graph the line after you create a scatter plot is to use LinRegTTest.

Make sure you have done the scatter plot. Check it on your screen.
Go to LinRegTTest and enter the lists.
At RegEq: press VARS and arrow over to Y-VARS. Press 1 for 1:Function. Press 1 for 1:Y1. Then arrow down to Calculate and do the calculation for the line of best fit.
Press Y = (you will see the regression equation).
Press GRAPH. The line will be drawn."

The Correlation Coefficient *r*

Besides looking at the scatter plot and seeing that a line seems reasonable, how can you tell if the line is a good predictor? Use the correlation coefficient as another indicator (besides the scatterplot) of the strength of the relationship between x and y.

The **correlation coefficient, *r*,** developed by Karl Pearson in the early 1900s, is numerical and provides a measure of strength and direction of the linear association between the independent variable x and the dependent variable y.

The correlation coefficient is calculated as

$$r = \frac{n\Sigma(xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n\Sigma x^2 - (\Sigma x)^2\right]\left[n\Sigma y^2 - (\Sigma y)^2\right]}}$$

where n = the number of data points.

If you suspect a linear relationship between x and y, then r can measure how strong the linear relationship is.

What the VALUE of r tells us:

The value of r is always between -1 and $+1$: $-1 \leq r \leq 1$.

The size of the correlation r indicates the strength of the linear relationship between x and y. Values of r close to -1 or to $+1$ indicate a stronger linear relationship between x and y.

If $r = 0$ there is absolutely no linear relationship between x and y **(no linear correlation)**.

If $r = 1$, there is perfect positive correlation. If $r = -1$, there is perfect negative correlation. In both these cases, all of the original data points lie on a straight line. Of course, in the real world, this will not generally happen.

What the SIGN of r tells us

A positive value of r means that when x increases, y tends to increase and when x decreases, y tends to decrease **(positive correlation)**.

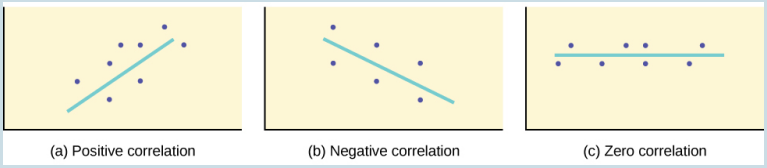
A negative value of r means that when x increases, y tends to decrease and when x decreases, y tends to increase **(negative correlation)**.

The sign of r is the same as the sign of the slope, b, of the best-fit line.

Note

Strong correlation does not suggest that x causes y or y causes x. We say **"correlation does not imply causation."**

(a) A scatter plot showing data with a positive correlation. $0 < r < 1$ (b) A scatter plot showing data with a negative correlation. $-1 < r < 0$ (c) A scatter plot showing data with zero correlation. $r = 0$



The formula for r looks formidable. However, computer spreadsheets, statistical software, and many calculators can quickly calculate r. The correlation coefficient r is the bottom item in the output screens for the LinRegTTest on the TI-83, TI-83+, or TI-84+ calculator (see previous section for instructions).

The Coefficient of Determination

The variable r^2 is called the coefficient of determination and is the square of the correlation coefficient but is usually stated as a percent rather than in decimal form. It has an interpretation in the context of the data:

- r^2, when expressed as a percent, represents the percent of variation in the dependent (predicted) variable y that can be explained by variation in the independent (explanatory) variable x using the regression (best-fit) line.
- $1 - r^2$, when expressed as a percentage, represents the percent of variation in y that is NOT explained by variation in x using the regression line. This can be seen as the scattering of the observed data points about the regression line.

Consider the third exam/final exam example introduced in the previous section
The line of best fit is: $\hat{y} = -173.51 + 4.83x$
The correlation coefficient is $r = 0.6631$
The coefficient of determination is $r^2 = 0.6631^2 = 0.4397$
Interpretation of r^2 in the context of this example:
Approximately 44% of the variation (0.4397 is approximately 0.44) in the final-exam grades can be explained by the variation in the grades on the third exam, using the best-fit regression line.
Therefore, approximately 56% of the variation (1 – 0.44 = 0.56) in the final exam grades can NOT be explained by the variation in the grades on the third exam, using the best-fit regression line. (This is seen as the scattering of the points about the line.)

10.4 TESTING THE SIGNIFICANCE OF THE CORRELATION COEFFICIENT

Learning Objectives

By the end of this section, the student should be able to:

- Test the significance of the correlation coefficient.

The correlation coefficient, r, tells us about the strength and direction of the linear relationship between x and y. However, the reliability of the linear model also depends on how many observed data points are in the sample. We need to look at both the value of the correlation coefficient r and the sample size n, together.

We perform a hypothesis test of the **"significance of the correlation coefficient"** to decide whether the linear relationship in the sample data is strong enough to use to model the relationship in the population.

The sample data are used to compute r, the correlation coefficient for the sample. If we had data for the entire population, we could find the population correlation coefficient. But because we only have sample data, we cannot calculate the population correlation coefficient. The sample correlation coefficient, r, is our estimate of the unknown population correlation coefficient.

- The symbol for the population correlation coefficient is ρ, the Greek letter "rho."
- ρ = population correlation coefficient (unknown)
- r = sample correlation coefficient (known; calculated from sample data)

The hypothesis test lets us decide whether the value of the population correlation coefficient ρ is "close to zero" or "significantly different from zero". We decide this based on the sample correlation coefficient r and the sample size n.

If the test concludes that the correlation coefficient is significantly different from zero, we say that the correlation coefficient is "significant."
Conclusion: There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from zero.
What the conclusion means: There is a significant linear relationship between x and y. We can use the regression line to model the linear relationship between x and y in the population.

If the test concludes that the correlation coefficient is not significantly different from zero (it is close to zero), we say that correlation coefficient is "not significant".
Conclusion: "There is insufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is not significantly different from zero."
What the conclusion means: There is not a significant linear relationship between x and y. Therefore, we CANNOT use the regression line to model a linear relationship between x and y in the population.

Note

- If r is significant and the scatter plot shows a linear trend, the line can be used to predict the value of y for values of x that are within the domain of observed x values.
- If r is not significant OR if the scatter plot does not show a linear trend, the line should not be used for prediction.
- If r is significant and if the scatter plot shows a linear trend, the line may NOT be appropriate or reliable for prediction OUTSIDE the domain of observed x values in the data.

PERFORMING THE HYPOTHESIS TEST

- **Null Hypothesis: H_0: $\rho = 0$**
- **Alternate Hypothesis: H_a: $\rho \neq 0$**

WHAT THE HYPOTHESES MEAN IN WORDS:

- **Null Hypothesis H_0:** The population correlation coefficient IS NOT significantly different from zero.

There IS NOT a significant linear relationship(correlation) between x and y in the population.

- **Alternate Hypothesis H_a:** The population correlation coefficient IS significantly DIFFERENT FROM zero. There IS A SIGNIFICANT LINEAR RELATIONSHIP (correlation) between x and y in the population.

DRAWING A CONCLUSION: There are two methods of making the decision. The two methods are equivalent and give the same result.

- **Method 1: Using the *p*-value**
- **Method 2: Using a table of critical values**

In this chapter of this textbook, we will always use a significance level of 5%, $\alpha = 0.05$

Note

Using the *p*-value method, you could choose any appropriate significance level you want; you are not limited to using $\alpha = 0.05$. But the table of critical values provided in this textbook assumes that we are using a significance level of 5%, $\alpha = 0.05$. (If we wanted to use a different significance level than 5% with the critical value method, we would need different tables of critical values that are not provided in this textbook.)

METHOD 1: Using a *p*-value to make a decision

To calculate the *p*-value using LinRegTTEST:

On the LinRegTTEST input screen, on the line prompt for β or ρ, highlight "$\neq$ **0**"

The output screen shows the p-value on the line that reads "p =".

(Most computer statistical software can calculate the *p*-value.)

If the *p*-value is less than the significance level ($\alpha = 0.05$):

- Decision: Reject the null hypothesis.
- Conclusion: "There is sufficient evidence to conclude that there is a significant linear relationship between *x* and *y* because the correlation coefficient is significantly different from zero."

If the *p*-value is NOT less than the significance level ($\alpha = 0.05$)

- Decision: DO NOT REJECT the null hypothesis.
- Conclusion: "There is insufficient evidence to conclude that there is a significant linear relationship between *x* and *y* because the correlation coefficient is NOT significantly different from zero."

Calculation Notes:

- You will use technology to calculate the *p*-value. The following describes the calculations to compute the test statistics and the *p*-value:
- The *p*-value is calculated using a *t*-distribution with $n - 2$ degrees of freedom.
- The formula for the test statistic is $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$. The value of the test statistic, *t*, is shown in the computer or calculator output along with the *p*-value. The test statistic *t* has the same sign as the correlation coefficient *r*.
- The *p*-value is the combined area in both tails.

An alternative way to calculate the *p*-value **(p)** given by LinRegTTest is the command 2*tcdf(abs(t),10^99, n-2) in 2nd DISTR.

Third-Exam vs. Final-Exam Example: *p*-value method

- Consider the third exam/final exam example.
- The line of best fit is: $\hat{y} = -173.51 + 4.83x$ with $r = 0.6631$ and there are $n = 11$ data points.
- Can the regression line be used for prediction? **Given a third exam score (*x* value), can we use the line to predict the final exam score (predicted *y* value)?**

$H_0: \rho = 0$

$H_a: \rho \neq 0$

$\alpha = 0.05$

- The p-value is 0.026 (from LinRegTTest on your calculator or from computer software).
- The p-value, 0.026, is less than the significance level of $\alpha = 0.05$.
- Decision: Reject the Null Hypothesis H_0
- Conclusion: There is sufficient evidence to conclude that there is a significant linear relationship between the third exam score (x) and the final exam score (y) because the correlation coefficient is significantly different from zero.

Because r is significant and the scatter plot shows a linear trend, the regression line can be used to predict final exam scores.

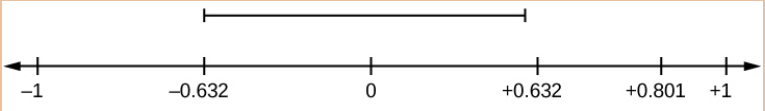
METHOD 2: Using a table of Critical Values to make a decision

The 95% Critical Values of the Sample Correlation Coefficient Table can be used to give you a good idea of whether the computed value of r **is significant or not**. Compare r to the appropriate critical value in the table. If r is not between the positive and negative critical values, then the correlation coefficient is significant. If r is significant, then you may want to use the line for prediction.

Example

Suppose you computed $r = 0.801$ using $n = 10$ data points, $df = n - 2 = 10 - 2 = 8$. The critical values associated with $df = 8$ are -0.632 and + 0.632. If $r <$ negative critical value or $r >$ positive critical value, then r is significant. Since $r = 0.801$ and $0.801 > 0.632$, r is significant and the line may be used for prediction. If you view this example on a number line, it will help you.

r is not significant between -0.632 and +0.632. $r = 0.801 > +0.632$. Therefore, r is significant.



Try It

For a given line of best fit, you computed that $r = 0.6501$ using $n = 12$ data points and the critical value is 0.576. Can the line be used for prediction? Why or why not?

Solution

If the scatter plot looks linear then, yes, the line can be used for prediction, because $r >$ the positive critical value.

Example

Suppose you computed $r = –0.624$ with 14 data points. $df = 14 – 2 = 12$. The critical values are –0.532 and 0.532. Since $–0.624 < –0.532$, r is significant and the line can be used for prediction

r = –0.624-0.532. Therefore, r is significant.



Try It

For a given line of best fit, you compute that $r = 0.5204$ using $n = 9$ data points, and the critical value is 0.666. Can the line be used for prediction? Why or why not?

Solution

No, the line cannot be used for prediction, because $r <$ the positive critical value.

Example

Suppose you computed $r = 0.776$ and $n = 6$. $df = 6 - 2 = 4$. The critical values are –0.811 and 0.811. Since $-0.811 < 0.776 < 0.811$, r is not significant, and the line should not be used for prediction.

$-0.811 < r = 0.776 < 0.811$. Therefore, r is not significant.



Try It

For a given line of best fit, you compute that $r = -0.7204$ using $n = 8$ data points, and the critical value is = 0.707. Can the line be used for prediction? Why or why not?

Solution
Yes, the line can be used for prediction, because r < the negative critical value.

Third-Exam vs. Final-Exam Example: critical value method

Consider the third exam/final exam example.
The line of best fit is: $\hat{y} = -173.51+4.83x$ with $r = 0.6631$ and there are $n = 11$ data points. Can the regression line be used for prediction? **Given a third-exam score (x value), can we use the line to predict the final exam score (predicted y value)?**

- $H_0: \rho = 0$
- $H_a: \rho \neq 0$
- $\alpha = 0.05$

- Use the "95% Critical Value" table for r with $df = n - 2 = 11 - 2 = 9$.
- The critical values are −0.602 and +0.602
- Since 0.6631 > 0.602, r is significant.
- Decision: Reject the null hypothesis.
- Conclusion: There is sufficient evidence to conclude that there is a significant linear relationship between the third exam score (x) and the final exam score (y) because the correlation coefficient is significantly different from zero.

Because r is significant and the scatter plot shows a linear trend, the regression line can be used to predict final exam scores.

Example

Suppose you computed the following correlation coefficients. Using the table at the end of the chapter, determine if r is significant and the line of best fit associated with each r can be used to predict a y value. If it helps, draw a number line.

1. $r = –0.567$ and the sample size, n, is 19. The $df = n – 2 = 17$. The critical value is –0.456. $–0.567 < –0.456$ so r is significant.
2. $r = 0.708$ and the sample size, n, is nine. The $df = n – 2 = 7$. The critical value is 0.666. $0.708 > 0.666$ so r is significant.
3. $r = 0.134$ and the sample size, n, is 14. The $df = 14 – 2 = 12$. The critical value is 0.532. 0.134 is between –0.532 and 0.532 so r is not significant.
4. $r = 0$ and the sample size, n, is five. No matter what the dfs are, $r = 0$ is between the two critical values so r is not significant.

Try It

For a given line of best fit, you compute that $r = 0$ using $n = 100$ data points. Can the line be used for prediction? Why or why not?

Solution

No, the line cannot be used for prediction no matter what the sample size is.

Assumptions in Testing the Significance of the Correlation Coefficient

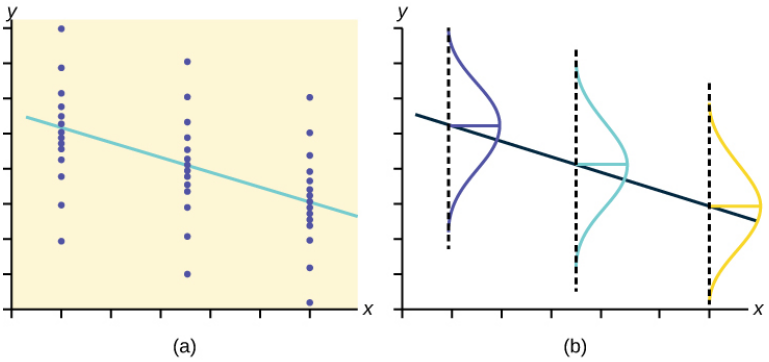
Testing the significance of the correlation coefficient requires that certain assumptions about the data are satisfied. The premise of this test is that the data are a sample of observed points taken from a larger population. We have not examined the entire population because it is not possible or feasible to do so. We are examining the sample to draw a conclusion about whether the linear relationship that we see between x and y in the sample data provides strong enough evidence so that we can conclude that there is a linear relationship between x and y in the population.

The regression line equation that we calculate from the sample data gives the best-fit line for our particular sample. We want to use this best-fit line for the sample as an estimate of the best-fit line for the population. Examining the scatterplot and testing the significance of the correlation coefficient helps us determine if it is appropriate to do this.

The assumptions underlying the test of significance are:

- There is a linear relationship in the population that models the average value of y for varying values of x. In other words, the expected value of y for each particular value lies on a straight line in the population. (We do not know the equation for the line for the population. Our regression line from the sample is our best estimate of this line in the population.)
- The y values for any particular x value are normally distributed about the line. This implies that there are more y values scattered closer to the line than are scattered farther away. Assumption (1) implies that these normal distributions are centered on the line: the means of these normal distributions of y values lie on the line.
- The standard deviations of the population y values about the line are equal for each value of x. In other words, each of these normal distributions of y values has the same shape and spread about the line.
- The residual errors are mutually independent (no pattern).
- The data are produced from a well-designed, random sample or randomized experiment.

The y values for each x value are normally distributed about the line with the same standard deviation. For each x value, the mean of the y values lies on the regression line. More y values lie near the line than are scattered further away from the line.


y
x
(a)
y
x
(b)

10.5 PREDICTION

Learning Objectives

By the end of this section, the student should be able to:

- Use the line of best fit for prediction.

Recall the third exam/final exam example.

We examined the scatterplot and showed that the correlation coefficient is significant. We found the equation of the best-fit line for the final exam grade as a function of the grade on the third-exam. We can now use the least-squares regression line for prediction.

Suppose you want to estimate, or predict, the mean final exam score of statistics students who received 73 on the third exam. The exam scores **(*x*-values)** range from 65 to 75. **Since 73 is between the *x*-values 65 and 75**, substitute $x = 73$ into the equation. Then:

Formula does not parse

We predict that statistics students who earn a grade of 73 on the third exam will earn a grade of 179.08 on the final exam, on average.

Example

Recall the third exam/final exam example.

a. What would you predict the final exam score to be for a student who scored a 66 on the third exam?

Solution
a. 145.27

b. What would you predict the final exam score to be for a student who scored a 90 on the third exam?

Solution

b. The x values in the data are between 65 and 75. Ninety is outside of the domain of the observed x values in the data (independent variable), so you cannot reliably predict the final exam score for this student. (Even though it is possible to enter 90 into the equation for x and calculate a corresponding y value, the y value that you get will not be reliable.)

To understand really how unreliable the prediction can be outside of the observed x values observed in the data, make the substitution $x = 90$ into the equation.

Formula does not parse

The final-exam score is predicted to be 261.19. The highest the final-exam score can be is 200.

Note

The process of predicting inside of the observed x values observed in the data is called interpolation. The process of predicting outside of the observed x values observed in the data is called extrapolation.

Try It

Data are collected on the relationship between the number of hours per week practicing a musical instrument and scores on a math test. The line of best fit is as follows:

$\hat{y} = 72.5 + 2.8x$

What would you predict the score on a math test would be for a student who practices a musical instrument for five hours a week?

Solution
86.5

REFERENCES

Data from the Centers for Disease Control and Prevention.

Data from the National Center for HIV, STD, and TB Prevention.

Data from the United States Census Bureau. Available online at http://www.census.gov/compendia/statab/cats/transportation/motor_vehicle_accidents_and_fatalities.html.

Data from the National Center for Health Statistics.

10.6 OUTLIERS

Learning Objectives

By the end of this section, the student should be able to:

- Find and interpret outliers between two quantitative variables.

In some data sets, there are values **(observed data points)** called **outliers**. **Outliers are observed data points that are far from the least squares line.** They have large "errors", where the "error" or residual is the vertical distance from the line to the point.

Outliers need to be examined closely. Sometimes, for some reason or another, they should not be included in the analysis of the data. It is possible that an outlier is a result of erroneous data. Other times, an outlier may hold valuable information about the population under study and should remain included in the data. The key is to examine carefully what causes a data point to be an outlier.

Besides outliers, a sample may contain one or a few points that are called influential points. Influential points are observed data points that are far from the other observed data points in the horizontal direction. These points may have a big effect on the slope of the regression line. To begin to identify an influential point, you can remove it from the data set and see if the slope of the regression line has changed significantly.

Computers and many calculators can be used to identify outliers from the data. Computer output for regression analysis will often identify both outliers and influential points so that you can examine them.

Identifying Outliers

We could guess at outliers by looking at a graph of the scatter plot and best fit-line. However, we would like some guideline as to how far away a point needs to be in order to be considered an outlier. **As a rough rule of**

thumb, we can flag any point that is located further than two standard deviations above or below the best-fit line as an outlier. The standard deviation used is the standard deviation of the residuals or errors.

We can do this visually in the scatter plot by drawing an extra pair of lines that are two standard deviations above and below the best-fit line. Any data points that are outside this extra pair of lines are flagged as potential outliers. Or we can do this numerically by calculating each residual and comparing it to twice the standard deviation. On the TI-83, 83+, or 84+, the graphical approach is easier. The graphical procedure is shown first, followed by the numerical calculations. You would generally need to use only one of these methods.

Example

In the third exam/final exam example, you can determine if there is an outlier or not. If there is an outlier, as an exercise, delete it and fit the remaining data to a new line. For this example, the new line ought to fit the remaining data better. This means the **SSE** should be smaller and the correlation coefficient ought to be closer to 1 or –1.

Solution

Graphical Identification of Outliers

With the TI-83, 83+, 84+ graphing calculators, it is easy to identify the outliers graphically and visually. If we were to measure the vertical distance from any data point to the corresponding point on the line of best fit and that distance were equal to $2s$ or more, then we would consider the data point to be "too far" from the line of best fit. We need to find and graph the lines that are two standard deviations below and above the regression line. Any points that are outside these two lines are outliers. We will call these lines Y2 and Y3.

As we did with the equation of the regression line and the correlation coefficient, we will use technology to calculate this standard deviation for us. Using the **LinRegTTest** with this data, scroll down through the output screens to find ***s* = 16.412**.

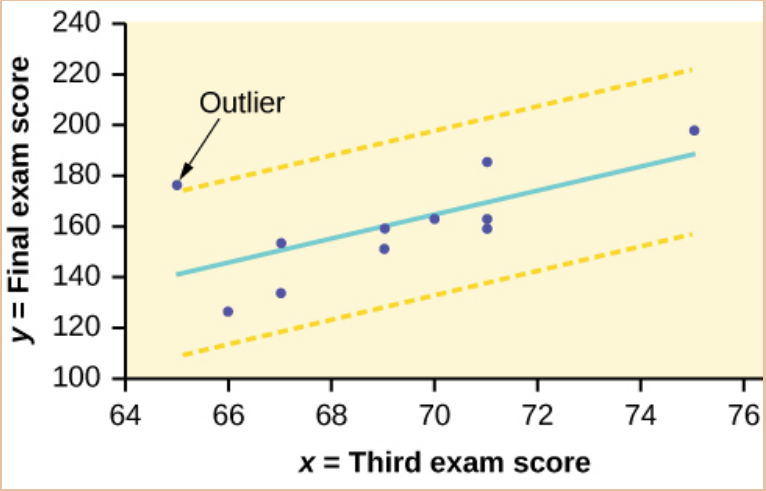
Line Y2 = –173.5 + 4.83x –2(16.4) and line Y3 = –173.5 + 4.83x + 2(16.4)

where $\hat{y}$ = –173.5 + 4.83x is the line of best fit. Y2 and Y3 have the same slope as the line of best fit.

Graph the scatter plot with the best fit line in equation Y1, then enter the two extra lines as Y2 and Y3 in the "Y=" equation editor and press ZOOM 9. You will find that the only data point that is not between lines Y2 and Y3 is the point x = 65, y = 175. On the calculator screen, it is just barely outside

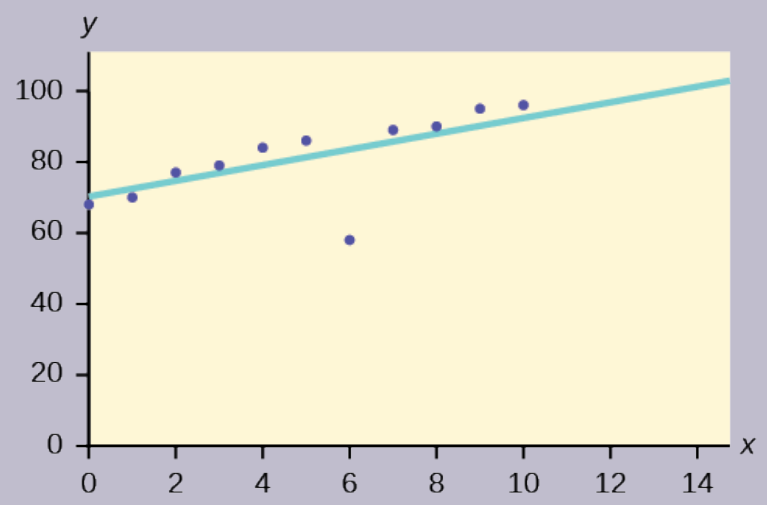
these lines. The outlier is the student who had a grade of 65 on the third exam and 175 on the final exam; this point is farther than two standard deviations away from the best-fit line.

Sometimes a point is so close to the lines used to flag outliers on the graph that it is difficult to tell if the point is between or outside the lines. On a computer, enlarging the graph may help; on a small calculator screen, zooming in may make the graph clearer. Note that when the graph does not give a clear enough picture, you can use the numerical comparisons to identify outliers.



Try It

Identify the potential outlier in the scatter plot. The standard deviation of the residuals or errors is approximately 8.6.



Solution

The outlier appears to be at (6, 58). The expected y value on the line for the point (6, 58) is approximately 82. Fifty-eight is 24 units from 82. Twenty-four is more than two standard deviations ($2s$ = (2)(8.6) = 17.2). So 82 is more than two standard deviations from 58, which makes (6, 58) a potential outlier.

Numerical Identification of Outliers

In [link], the first two columns are the third-exam and final-exam data. The third column shows the predicted $\hat{y}$ values calculated from the line of best fit: $\hat{y} = -173.5 + 4.83x$. The residuals, or errors, have been calculated in the fourth column of the table: observed y value−predicted y value = $y - \hat{y}$.

s is the standard deviation of all the $y - \hat{y} = \varepsilon$ values where n = the total number of data points. If each residual is calculated and squared, and the results are added, we get the SSE. The standard deviation of the residuals is calculated from the SSE as:

$$s = \sqrt{\frac{SSE}{n-2}}$$

Note

We divide by $(n-2)$ because the regression model involves two estimates.

Rather than calculate the value of s ourselves, we can find s using the computer or calculator. For this example, the calculator function LinRegTTest found $s = 16.4$ as the standard deviation of the residuals 35–1716–6–1993–1–10–9–1.

x	y	$\hat{y}$	$y - \hat{y}$
65	175	140	175 – 140 = 35
67	133	150	133 – 150= –17
71	185	169	185 – 169 = 16
71	163	169	163 – 169 = –6
66	126	145	126 – 145 = –19
75	198	189	198 – 189 = 9
67	153	150	153 – 150 = 3
70	163	164	163 – 164 = –1
71	159	169	159 – 169 = –10
69	151	160	151 – 160 = –9
69	159	160	159 – 160 = –1

We are looking for all data points for which the residual is greater than $2s = 2(16.4) = 32.8$ or less than –32.8. Compare these values to the residuals in column four of the table. The only such data point is the student who had a grade of 65 on the third exam and 175 on the final exam; the residual for this student is 35.

How Does the Outlier Affect the Best Fit Line?

Numerically and graphically, we have identified the point (65, 175) as an outlier. We should re-examine the data for this point to see if there are any problems with the data. If there is an error, we should fix the error if

possible, or delete the data. If the data is correct, we would leave it in the data set. For this problem, we will suppose that we examined the data and found that this outlier data was an error. Therefore we will continue on and delete the outlier, so that we can explore how it affects the results, as a learning experience.

Compute a new best-fit line and correlation coefficient using the ten remaining points:
On the TI-83, TI-83+, TI-84+ calculators, delete the outlier from L1 and L2. Using the LinRegTTest, the new line of best fit and the correlation coefficient are:

$\hat{y} = -355.19 + 7.39x$ and $r = 0.9121$

The new line with $r = 0.9121$ is a stronger correlation than the original ($r = 0.6631$) because $r = 0.9121$ is closer to one. This means that the new line is a better fit to the ten remaining data values. The line can better predict the final exam score given the third exam score.

Numerical Identification of Outliers: Calculating *s* and Finding Outliers Manually

If you do not have the function LinRegTTest, then you can calculate the outlier in the first example by doing the following.

First, **square each $|y - \hat{y}|$**

The squares are $35^2 17^2 16^2 6^2 19^2 9^2 3^2 1^2 10^2 9^2 1^2$

Then, add (sum) all the $|y - \hat{y}|$ squared terms using the formula

Formula does not parse (Recall that $y_i - \hat{y}_i = \varepsilon_i$.)

$= 35^2 + 17^2 + 16^2 + 6^2 + 19^2 + 9^2 + 3^2 + 1^2 + 10^2 + 9^2 + 1^2$

$= 2440$ = **SSE**. The result, **SSE**, is the Sum of Squared Errors.

Next, calculate *s*, the standard deviation of all the $y - \hat{y} = \varepsilon$ values where *n* = the total number of data points.

The calculation is $s = \sqrt{\frac{\text{SSE}}{n-2}}$.

For the third exam/final exam problem, $s = \sqrt{\frac{2440}{11-2}} = 16.47$.

Next, multiply *s* by 2:

(2)(16.47) = 32.94

32.94 is 2 standard deviations away from the mean of the $y - \hat{y}$ values.

If we were to measure the vertical distance from any data point to the corresponding point on the line of best fit and that distance is at least 2*s*, then we would consider the data point to be "too far" from the line of best fit. We call that point a potential outlier.

For the example, if any of the $|y - \hat{y}|$ values are **at least** 32.94, the corresponding (x, y) data point is a potential outlier.

For the third exam/final exam problem, all the $|y - \hat{y}|$'s are less than 31.29 except for the first one which is 35.

$35 > 31.29$ That is, $|y - \hat{y}| \geq (2)(s)$

The point which corresponds to $|y - \hat{y}| = 35$ is (65, 175). **Therefore, the data point (65,175) is a potential outlier.** For this example, we will delete it. (Remember, we do not always delete an outlier.)

Note

When outliers are deleted, the researcher should either record that data was deleted, and why, or the researcher should provide results both with and without the deleted data. If data is erroneous and the correct values are known (e.g., student one actually scored a 70 instead of a 65), then this correction can be made to the data.

The next step is to compute a new best-fit line using the ten remaining points. The new line of best fit and the correlation coefficient are:

$\hat{y} = -355.19 + 7.39x$ and $r = 0.9121$

Example

Using this new line of best fit (based on the remaining ten data points in the third exam/final exam example), what would a student who receives a 73 on the third exam expect to receive on the final exam? Is this the same as the prediction made using the original line?

Solution

Using the new line of best fit, $\hat{y} = -355.19 + 7.39(73) = 184.28$. A student who scored 73 points on the third exam would expect to earn 184 points on the final exam.

The original line predicted $\hat{y} = -173.51 + 4.83(73) = 179.08$ so the prediction using the new line with the outlier eliminated differs from the original prediction.

Try It

The data points for the graph from the third exam/final exam example are as follows: (1, 5), (2, 7), (2, 6), (3, 9), (4, 12), (4, 13), (5, 18), (6, 19), (7, 12), and (7, 21). Remove the outlier and recalculate the line of best fit. Find the value of $\hat{y}$ when $x = 10$.

Solution

$\hat{y} = 1.04 + 2.96x$; 30.64

Example

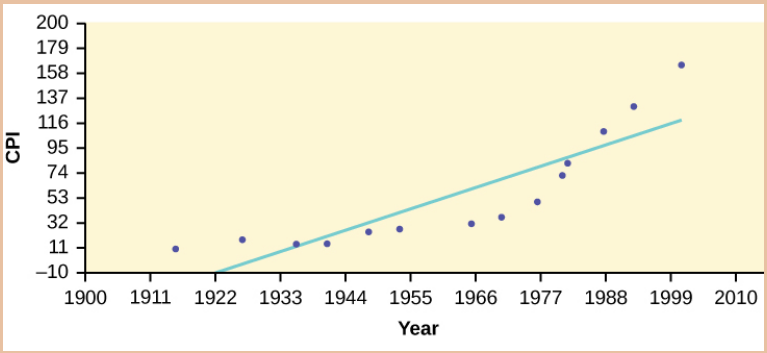
The Consumer Price Index (CPI) measures the average change over time in the prices paid by urban consumers for consumer goods and services. The CPI affects nearly all Americans because of the many ways it is used. One of its biggest uses is as a measure of inflation. By providing information about price changes in the Nation's economy to government, business, and labor, the CPI helps them to make economic decisions. The president, Congress, and the Federal Reserve Board use the CPI's trends to formulate monetary and fiscal policies. In the following table, x is the year and y is the CPI.

Data			
x	y	x	y
1915	10.1	1969	36.7
1926	17.7	1975	49.3
1935	13.7	1979	72.6
1940	14.7	1980	82.4
1947	24.1	1986	109.6
1952	26.5	1991	130.7
1964	31.0	1999	166.6

1. Draw a scatter plot of the data.
2. Calculate the least squares line. Write the equation in the form $\hat{y} = a + bx$.
3. Draw the line on the scatter plot.
4. Find the correlation coefficient. Is it significant?
5. What is the average CPI for the year 1990?

Solution

1. See [link].
2. $\hat{y} = -3204 + 1.662x$ is the equation of the line of best fit.
3. $r = 0.8694$
4. The number of data points is $n = 14$. Use the 95% Critical Values of the Sample Correlation Coefficient table at the end of Chapter 12. $n - 2 = 12$. The corresponding critical value is 0.532. Since $0.8694 > 0.532$, r is significant.
   $\hat{y} = -3204 + 1.662(1990) = 103.4$ CPI
5. Using the calculator LinRegTTest, we find that $s = 25.4$; graphing the lines $Y2 = -3204 + 1.662X - 2(25.4)$ and $Y3 = -3204 + 1.662X + 2(25.4)$ shows that no data values are outside those lines, identifying no outliers. (Note that the year 1999 was very close to the upper line, but still inside it.)



Note

In the example, notice the pattern of the points compared to the line. Although the correlation coefficient is significant, the pattern in the scatter plot indicates that a curve would be a more appropriate model to use than a line. In this example, a statistician should prefer to use other methods to fit a curve to this data, rather than model the data with the line we found. In addition to doing the calculations, it is always important to look at the scatter plot when deciding whether a linear model is appropriate.

If you are interested in seeing more years of data, visit the Bureau of Labor Statistics CPI website ftp://ftp.bls.gov/pub/special.requests/cpi/cpiai.txt; our data is taken from the column entitled "Annual Avg." (third column from the right). For example you could add more current years of data. Try adding the more recent years: 2004: CPI = 188.9; 2008: CPI = 215.3; 2011: CPI = 224.9. See how it affects the model. (Check: $\hat{y} = -4436 + 2.295x$; $r = 0.9018$. Is r significant? Is the fit better with the addition of the new points?)

Try It

The following table shows economic development measured in per capita income PCINC.

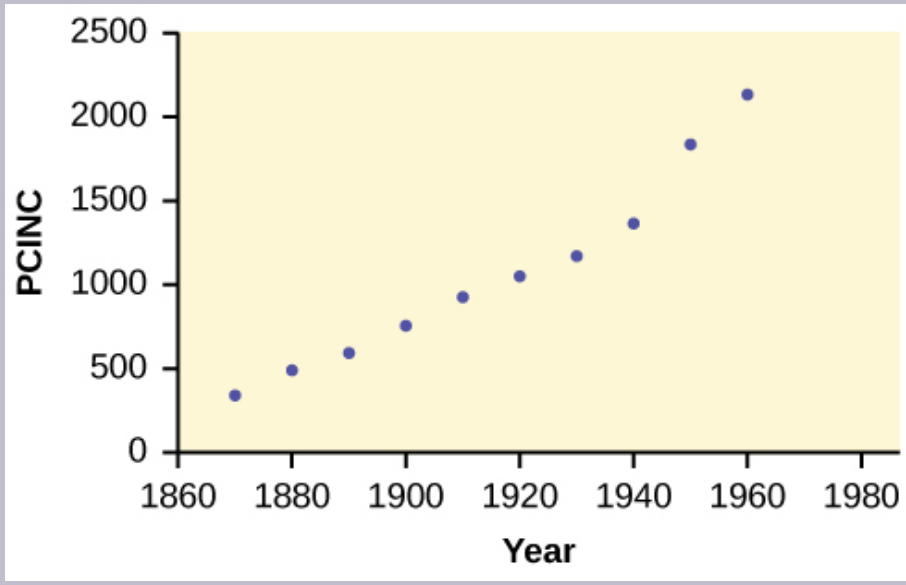
Year	PCINC	Year	PCINC
1870	340	1920	1050
1880	499	1930	1170
1890	592	1940	1364
1900	757	1950	1836
1910	927	1960	2132

1. What are the independent and dependent variables?
2. Draw a scatter plot.
3. Use regression to find the line of best fit and the correlation coefficient.
4. Interpret the significance of the correlation coefficient.
5. Is there a linear relationship between the variables?
6. Find the coefficient of determination and interpret it.
7. What is the slope of the regression equation? What does it mean?
8. Use the line of best fit to estimate PCINC for 1900, for 2000.
9. Determine if there are any outliers.

Solution

a. The independent variable (x) is the year and the dependent variable (y) is the per capita income.

b.



c. $\hat{y} = 18.61x - 34574$; $r = 0.9732$

d. At $df = 8$, the critical value is 0.632. The r value is significant because it is greater than the critical value.

e. There does appear to be a linear relationship between the variables.

f. The coefficient of determination is 0.947, which means that 94.7% of the variation in PCINC is explained by the variation in the years.

g. and h. The slope of the regression equation is 18.61, and it means that per capita income increases by $ 18.61 for each passing year. $\hat{y} = 785$ when the year is 1900, and $\hat{y} = 2{,}646$ when the year is 2000.

i. There do not appear to be any outliers.

95% Critical Values of the Sample Correlation Coefficient Table

Degrees of Freedom: $n - 2$	Critical Values: (+ and –)
1	0.997
2	0.950
3	0.878
4	0.811
5	0.754
6	0.707
7	0.666
8	0.632
9	0.602
10	0.576
11	0.555
12	0.532
13	0.514
14	0.497
15	0.482
16	0.468
17	0.456
18	0.444
19	0.433
20	0.423
21	0.413
22	0.404
23	0.396
24	0.388
25	0.381
26	0.374
27	0.367
28	0.361
29	0.355
30	0.349

Degrees of Freedom: $n - 2$	Critical Values: (+ and –)
40	0.304
50	0.273
60	0.250
70	0.232
80	0.217
90	0.205
100	0.195

References

Data from the House Ways and Means Committee, the Health and Human Services Department.

Data from Microsoft Bookshelf.

Data from the United States Department of Labor, the Bureau of Labor Statistics.

Data from the Physician's Handbook, 1990.

Data from the United States Department of Labor, the Bureau of Labor Statistics.

CHAPTER 10 REVIEW

Chapter Review from 10.1

The most basic type of association is a linear association. This type of relationship can be defined algebraically by the equations used, numerically with actual or predicted data values, or graphically from a plotted curve. (Lines are classified as straight curves.) Algebraically, a linear equation typically takes the form $\boldsymbol{y = mx + b}$, where $\boldsymbol{m}$ and $\boldsymbol{b}$ are constants, $\boldsymbol{x}$ is the independent variable, $\boldsymbol{y}$ is the dependent variable. In a statistical context, a linear equation is written in the form $\boldsymbol{y = a + bx}$, where $\boldsymbol{a}$ and $\boldsymbol{b}$ are the constants. This form is used to help readers distinguish the statistical context from the algebraic context. In the equation $y = a + bx$, the constant b that multiplies the $\boldsymbol{x}$ variable (b is called a coefficient) is called the **slope**. The slope describes the rate of change between the independent and dependent variables; in other words, the rate of change describes the change that occurs in the dependent variable as the independent variable is changed. In the equation $y = a + bx$, the constant a is called the y-intercept. Graphically, the y-intercept is the y coordinate of the point where the graph of the line crosses the y axis. At this point $x = 0$.

The **slope of a line** is a value that describes the rate of change between the independent and dependent variables. The **slope** tells us how the dependent variable (y) changes for every one unit increase in the independent (x) variable, on average. The **y-intercept** is used to describe the dependent variable when the independent variable equals zero. Graphically, the slope is represented by three line types in elementary statistics.

Formula Review

$y = a + bx$ where a is the y-intercept and b is the slope. The variable x is the independent variable and y is the dependent variable.

Use the following information to answer the next three exercises. A vacation resort rents SCUBA equipment to certified divers. The resort charges an up-front fee of $ 25 and another fee of $ 12.50 an hour.

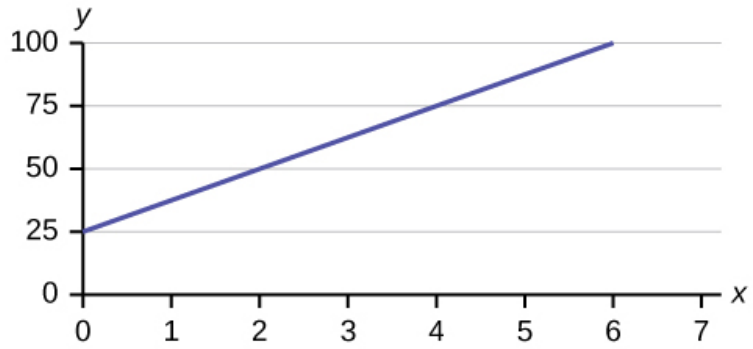
What are the dependent and independent variables?
Solution

dependent variable: fee amount; independent variable: time

Find the equation that expresses the total fee in terms of the number of hours the equipment is rented.

Graph the equation from [link].

Solution

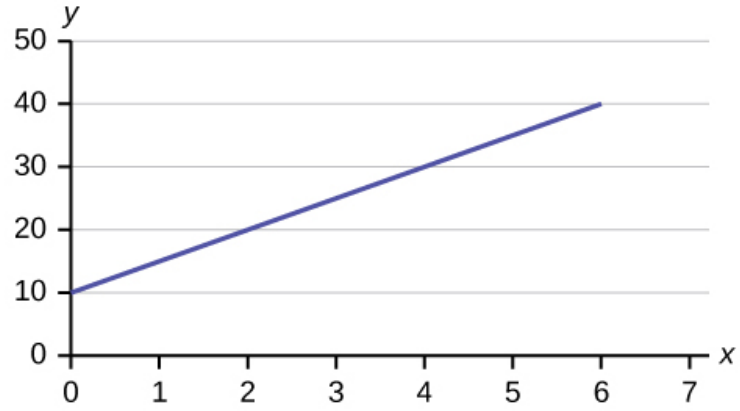


Use the following information to answer the next two exercises. A credit card company charges $ 10 when a payment is late, and $ 5 a day each day the payment remains unpaid.

Find the equation that expresses the total fee in terms of the number of days the payment is late.

Graph the equation from [link].

Solution



Is the equation $y = 10 + 5x - 3x^2$ linear? Why or why not?

Which of the following equations are linear?

a. $y = 6x + 8$

b. $y + 7 = 3x$

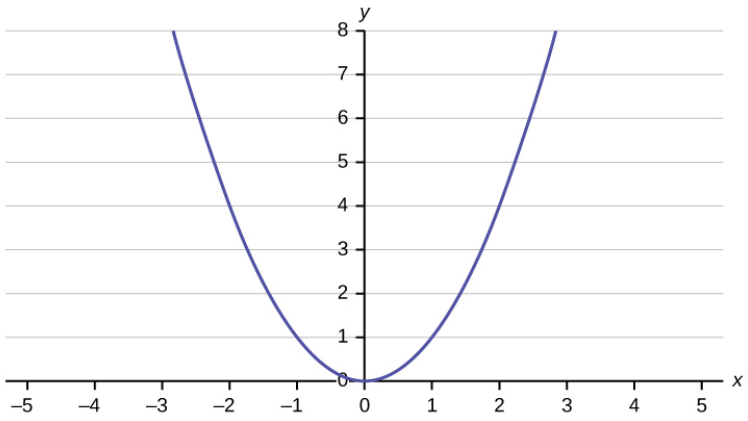
c. $y - x = 8x^2$

d. $4y = 8$

Solution

$y = 6x + 8$, $4y = 8$, and $y + 7 = 3x$ are all linear equations.

Does the graph show a linear equation? Why or why not?



[link] contains real data for the first two decades of AIDS reporting.

Adults and Adolescents only, United States

Year	# AIDS cases diagnosed	# AIDS deaths
Pre-1981	91	29
1981	319	121
1982	1,170	453
1983	3,076	1,482
1984	6,240	3,466
1985	11,776	6,878
1986	19,032	11,987
1987	28,564	16,162
1988	35,447	20,868
1989	42,674	27,591
1990	48,634	31,335
1991	59,660	36,560
1992	78,530	41,055
1993	78,834	44,730
1994	71,874	49,095
1995	68,505	49,456
1996	59,347	38,510
1997	47,149	20,736
1998	38,393	19,005
1999	25,174	18,454
2000	25,522	17,347
2001	25,643	17,402
2002	26,464	16,371
Total	**802,118**	**489,093**

Use the columns “year” and “# AIDS cases diagnosed. Why is “year” the independent variable and “# AIDS cases diagnosed.” the dependent variable (instead of the reverse)?
Solution

The number of AIDS cases depends on the year. Therefore, year becomes the independent variable and the number of AIDS cases is the dependent variable.

Use the following information to answer the next two exercises. A specialty cleaning company charges an equipment fee and an hourly labor fee. A linear equation that expresses the total amount of the fee the company charges for each session is $y = 50 + 100x$.

What are the independent and dependent variables?

What is the y-intercept and what is the slope? Interpret them using complete sentences.
Solution

The y-intercept is 50 ($a = 50$). At the start of the cleaning, the company charges a one-time fee of $ 50 (this is when $x = 0$). The slope is 100 ($b = 100$). For each session, the company charges $ 100 for each hour they clean.

Use the following information to answer the next three questions. Due to erosion, a river shoreline is losing several thousand pounds of soil each year. A linear equation that expresses the total amount of soil lost per year is $y = 12{,}000x$.

What are the independent and dependent variables?

How many pounds of soil does the shoreline lose in a year?
Solution

12,000 pounds of soil

What is the y-intercept? Interpret its meaning.

Use the following information to answer the next two exercises. The price of a single issue of stock can fluctuate throughout the day. A linear equation that represents the price of stock for Shipment Express is $y = 15 - 1.5x$ where x is the number of hours passed in an eight-hour day of trading.

What are the slope and y-intercept? Interpret their meaning.
Solution

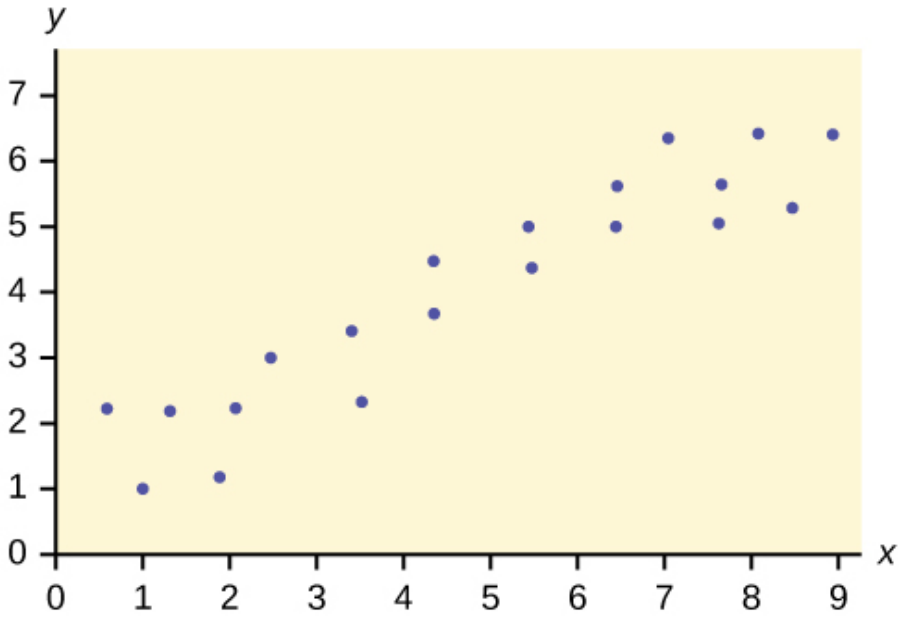
The slope is -1.5 ($b = -1.5$). This means the stock is losing value at a rate of $ 1.50 per hour. The y-intercept is $ 15 ($a = 15$). This means the price of stock before the trading day was $ 15.

If you owned this stock, would you want a positive or negative slope? Why?

Chapter Review from 10.2

Scatter plots are particularly helpful graphs when we want to see if there is a linear relationship among data points. They indicate both the direction of the relationship between the *x* variables and the *y* variables, and the strength of the relationship. We calculate the strength of the relationship between an independent variable and a dependent variable using linear regression.

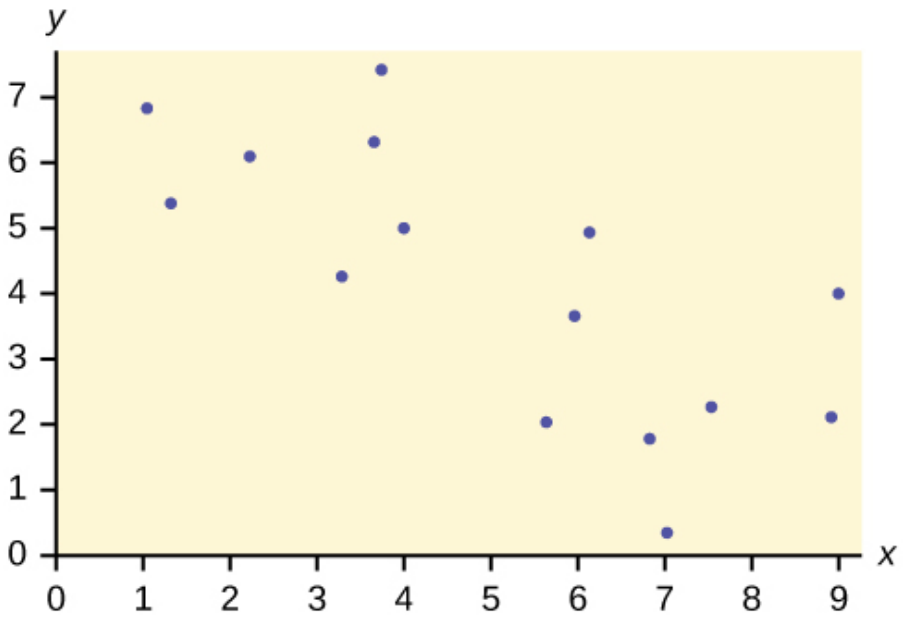
Does the scatter plot appear linear? Strong or weak? Positive or negative?



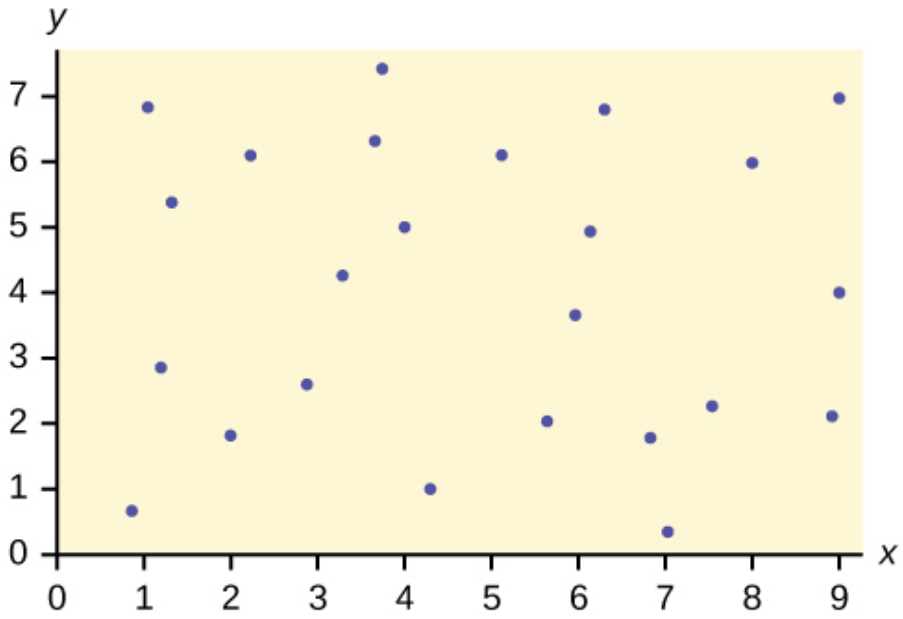
Solution

The data appear to be linear with a strong, positive correlation.

Does the scatter plot appear linear? Strong or weak? Positive or negative?



Does the scatter plot appear linear? Strong or weak? Positive or negative?



Solution

The data appear to have no correlation.

Chapter Review from 10.3

A regression line, or a line of best fit, can be drawn on a scatter plot and used to predict outcomes for the x and y variables in a given data set or sample data. There are several ways to find a regression line, but usually the

least-squares regression line is used because it creates a uniform line. Residuals, also called "errors," measure the distance from the actual value of y and the estimated value of y. The Sum of Squared Errors, when set to its minimum, calculates the points on the line of best fit. Regression lines can be used to predict values within the given set of data, but should not be used to make predictions for values outside the set of data.

The correlation coefficient r measures the strength of the linear association between x and y. The variable r has to be between −1 and +1. When r is positive, the x and y will tend to increase and decrease together. When r is negative, x will increase and y will decrease, or the opposite, x will decrease and y will increase. The coefficient of determination r^2, is equal to the square of the correlation coefficient. When expressed as a percent, r^2 represents the percent of variation in the dependent variable y that can be explained by variation in the independent variable x using the regression line.

Use the following information to answer the next five exercises. A random sample of ten professional athletes produced the following data where x is the number of endorsements the player has and y is the amount of money made (in millions of dollars).

x	y	x	y
0	2	5	12
3	8	4	9
2	7	3	9
1	3	0	3
5	13	4	10

Draw a scatter plot of the data.

Use regression to find the equation for the line of best fit.
Solution

$\hat{y} = 2.23 + 1.99x$

Draw the line of best fit on the scatter plot.

What is the slope of the line of best fit? What does it represent?
Solution

The slope is 1.99 ($b = 1.99$). It means that for every endorsement deal a professional player gets, he gets an average of another $ 1.99 million in pay each year.

What is the y-intercept of the line of best fit? What does it represent?

What does an r value of zero mean?
Solution

It means that there is no correlation between the data sets.

When $n = 2$ and $r = 1$, are the data significant? Explain.

When $n = 100$ and $r = -0.89$, is there a significant correlation? Explain.
Solution

Yes, there are enough data points and the value of r is strong enough to show that there is a strong negative correlation between the data sets.

Chapter Review from 10.5

After determining the presence of a strong correlation coefficient and calculating the line of best fit, you can use the least squares regression line to make predictions about your data.

Use the following information to answer the next two exercises. An electronics retailer used regression to find a simple model to predict sales growth in the first quarter of the new year (January through March). The model is good for 90 days, where x is the day. The model can be written as follows:

$\hat{y} = 101.32 + 2.48x$ where $\hat{y}$ is in thousands of dollars.

What would you predict the sales to be on day 60?
Solution

$ 250,120

What would you predict the sales to be on day 90?

Use the following information to answer the next three exercises. A landscaping company is hired to mow the grass for several large properties. The total area of the properties combined is 1,345 acres. The rate at which one person can mow is as follows:

$\hat{y} = 1350 - 1.2x$ where x is the number of hours and $\hat{y}$ represents the number of acres left to mow.

How many acres will be left to mow after 20 hours of work?
Solution

1,326 acres

How many acres will be left to mow after 100 hours of work?

How many hours will it take to mow all of the lawns? (When is $\hat{y} = 0$?)
Solution

1,125 hours, or when $x = 1{,}125$

[link] contains real data for the first two decades of AIDS reporting.

Adults and Adolescents only, United States

Year	# AIDS cases diagnosed	# AIDS deaths
Pre-1981	91	29
1981	319	121
1982	1,170	453
1983	3,076	1,482
1984	6,240	3,466
1985	11,776	6,878
1986	19,032	11,987
1987	28,564	16,162
1988	35,447	20,868
1989	42,674	27,591
1990	48,634	31,335
1991	59,660	36,560
1992	78,530	41,055
1993	78,834	44,730
1994	71,874	49,095
1995	68,505	49,456
1996	59,347	38,510
1997	47,149	20,736
1998	38,393	19,005
1999	25,174	18,454
2000	25,522	17,347
2001	25,643	17,402
2002	26,464	16,371
Total	**802,118**	**489,093**

Graph "year" versus "# AIDS cases diagnosed" (plot the scatter plot). Do not include pre-1981 data.

Perform linear regression. What is the linear equation? Round to the nearest whole number.
Solution

Check student's solution.

Write the equations:

Linear equation: _____

a = _____

b = _____

r = _____

n = _____

Solve.

When $x = 1985$, $\hat{y}$ = _____

When $x = 1990$, $\hat{y}$ =_____

When $x = 1970$, $\hat{y}$ =_____ Why doesn't this answer make sense?

Solution

When $x = 1985$, $\hat{y} = 25{,}52$

When $x = 1990$, $\hat{y} = 34{,}275$

When $x = 1970$, $\hat{y} = -725$ Why doesn't this answer make sense? The range of x values was 1981 to 2002; the year 1970 is not in this range. The regression equation does not apply, because predicting for the year 1970 is extrapolation, which requires a different process. Also, a negative number does not make sense in this context, where we are predicting AIDS cases diagnosed.

Does the line seem to fit the data? Why or why not?

What does the correlation imply about the relationship between time (years) and the number of diagnosed AIDS cases reported in the U.S.?

Solution

Also, the correlation $r = 0.4526$. If r is compared to the value in the 95% Critical Values of the Sample Correlation Coefficient Table, because $r > 0.423$, r is significant, and you would think that the line could be used for prediction. But the scatter plot indicates otherwise.

<!– From 12.4 Move to 12.8 –>

Plot the two given points on the following graph. Then, connect the two points to form the regression line.



Obtain the graph on your calculator or computer.

<!– From 12.5 MOVE to 12.8 –>

Write the equation: $\hat{y}$= _____
Solution

Formula does not parse = 3,448,225 + 1750x

Hand draw a smooth curve on the graph that shows the flow of the data.

<!– From 12.6 Move to 12.8 –>

Does the line seem to fit the data? Why or why not?
Solution

There was an increase in AIDS cases diagnosed until 1993. From 1993 through 2002, the number of AIDS cases diagnosed declined each year. It is not appropriate to use a linear regression line to fit to the data.

Do you think a linear fit is best? Why or why not?

What does the correlation imply about the relationship between time (years) and the number of diagnosed AIDS cases reported in the U.S.?
Solution

Since there is no linear association between year and # of AIDS cases diagnosed, it is not appropriate to calculate a linear correlation coefficient. When there is a linear association and it is appropriate to calculate a correlation, we cannot say that one variable “causes” the other variable.

<!– From 12.7 MOVE to 12.8 –>

Graph "year" vs. "# AIDS cases diagnosed." Do not include pre-1981. Label both axes with words. Scale both axes.

Enter your data into your calculator or computer. The pre-1981 data should not be included. Why is that so?

Write the linear equation, rounding to four decimal places:

Solution

We don't know if the pre-1981 data was collected from a single year. So we don't have an accurate x value for this figure.

Regression equation: $\hat{y}$ (#AIDS Cases) = −3,448,225 + 1749.777 (year)

	Coefficients
Intercept	−3,448,225
X Variable 1	1,749.777

Calculate the following:

1. a = _____
2. b = _____
3. correlation = _____
4. n = _____

Chapter Review from 10.6

To determine if a point is an outlier, do one of the following:

Input the following equations into the TI 83, 83+,84, 84+:

$y_1 = a + bx$

$y_2 = (2s)\,a + bx$ where s is the standard deviation of the residuals

$y_3 = -(2s)\,a + bx$

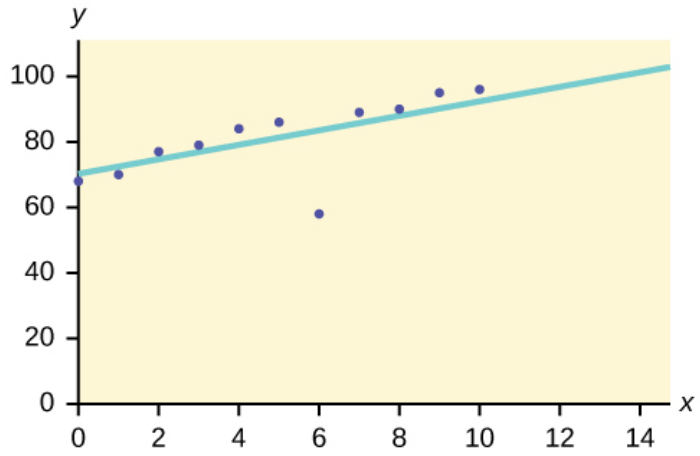
If any point is above $y_2$or below y_3 then the point is considered to be an outlier.

Use the residuals and compare their absolute values to 2s where s is the standard deviation of the residuals. If the absolute value of any residual is greater than or equal to 2s, then the corresponding point is an outlier.

Note: The calculator function LinRegTTest (STATS TESTS LinRegTTest) calculates s.

Use the following information to answer the next four exercises. The scatter plot shows the relationship between

hours spent studying and exam scores. The line shown is the calculated line of best fit. The correlation coefficient is 0.69.



Do there appear to be any outliers?
Solution

Yes, there appears to be an outlier at (6, 58).

A point is removed, and the line of best fit is recalculated. The new correlation coefficient is 0.98. Does the point appear to have been an outlier? Why?

What effect did the potential outlier have on the line of best fit?
Solution

The potential outlier flattened the slope of the line of best fit because it was below the data set. It made the line of best fit less accurate as a predictor for the data.

Are you more or less confident in the predictive ability of the new line of best fit?

The Sum of Squared Errors for a data set of 18 numbers is 49. What is the standard deviation?
Solution

$s = 1.75$

The Standard Deviation for the Sum of Squared Errors for a data set is 9.8. What is the cutoff for the vertical distance that a point can be from the line of best fit to be considered an outlier?

CHAPTER 10 HOMEWORK

Homework from 10.1

For each of the following situations, state the independent variable and the dependent variable.

1. A study is done to determine if elderly drivers are involved in more motor vehicle fatalities than other drivers. The number of fatalities per 100,000 drivers is compared to the age of drivers.
2. A study is done to determine if the weekly grocery bill changes based on the number of family members.
3. Insurance companies base life insurance premiums partially on the age of the applicant.
4. Utility bills vary according to power consumption.
5. A study is done to determine if a higher education reduces the crime rate in a population.

Solution

1. independent variable: age; dependent variable: fatalities
2. independent variable: # of family members; dependent variable: grocery bill
3. independent variable: age of applicant; dependent variable: insurance premium
4. independent variable: power consumption; dependent variable: utility
5. independent variable: higher education (years); dependent variable: crime rates

Piece-rate systems are widely debated incentive payment plans. In a recent study of loan officer effectiveness, the following piece-rate system was examined:

% of goal reached	< 80	80	100	120
Incentive	n/a	$ 4,000 with an additional $ 125 added per percentage point from 81–99%	$ 6,500 with an additional $ 125 added per percentage point from 101–119%	$ 9,500 with an additional $ 125 added per percentage point starting at 121%

If a loan officer makes 95% of his or her goal, write the linear function that applies based on the incentive plan table. In context, explain the y-intercept and slope.

Homework from 10.2

The Gross Domestic Product Purchasing Power Parity is an indication of a country's currency value compared to another country. [link] shows the GDP PPP of Cuba as compared to US dollars. Construct a scatter plot of the data.

Year	Cuba's PPP	Year	Cuba's PPP
1999	1,700	2006	4,000
2000	1,700	2007	11,000
2002	2,300	2008	9,500
2003	2,900	2009	9,700
2004	3,000	2010	9,900
2005	3,500		

Solution

Check student's solution.

The following table shows the poverty rates and cell phone usage in the United States. Construct a scatter plot of the data

Year	Poverty Rate	Cellular Usage per Capita
2003	12.7	54.67
2005	12.6	74.19
2007	12	84.86
2009	12	90.82

Does the higher cost of tuition translate into higher-paying jobs? The table lists the top ten colleges based on mid-career salary and the associated yearly tuition costs. Construct a scatter plot of the data.

School	Midcareer Salary (in thousands)	Yearly Tuition
Princeton	137	28,540
Harvey Mudd	135	40,133
CalTech	127	39,900
US Naval Academy	122	0
West Point	120	0
MIT	118	42,050
Lehigh University	118	43,220
NYU-Poly	117	39,565
Babson College	117	40,400
Stanford	114	54,506

Solution

For graph: check student's solution. Note that tuition is the independent variable and salary is the dependent variable.

If the level of significance is 0.05 and the p-value is 0.06, what conclusion can you draw?

If there are 15 data points in a set of data, what is the number of degrees of freedom?
Solution

13

Homework from 10.3

What is the process through which we can calculate a line that goes through a scatter plot with a linear pattern?

Explain what it means when a correlation has an r^2 of 0.72.
Solution

It means that 72% of the variation in the dependent variable (y) can be explained by the variation in the independent variable (x).

Can a coefficient of determination be negative? Why or why not?

Chapter Review from 10.4

Linear regression is a procedure for fitting a straight line of the form $\hat{y} = a + bx$ to data. The conditions for regression are:

- **Linear** In the population, there is a linear relationship that models the average value of y for different values of x.
- **Independent** The residuals are assumed to be independent.
- **Normal** The y values are distributed normally for any value of x.
- **Equal variance** The standard deviation of the y values is equal for each x value.
- **Random** The data are produced from a well-designed random sample or randomized experiment.

The slope $\boldsymbol{b}$ and intercept $\boldsymbol{a}$ of the least-squares line estimate the slope β and intercept α of the population (true) regression line. To estimate the population standard deviation of y, σ, use the standard deviation of the residuals, s. $\boldsymbol{s} = \sqrt{\frac{SEE}{n-2}}$. The variable ρ (rho) is the population correlation coefficient. To test the null hypothesis
H_0: ρ = *hypothesized value*, use a linear regression t-test. The most common null hypothesis is H_0: $\rho = 0$ which indicates there is no linear relationship between x and y in the population. The TI-83, 83+, 84, 84+ calculator function LinRegTTest can perform this test (STATS TESTS LinRegTTest).

Formula Review

Least Squares Line or Line of Best Fit:

Formula does not parse

where

a = y-intercept

b = slope

Standard deviation of the residuals:

$$\boldsymbol{s} = \sqrt{\frac{SEE}{n-2}}.$$

where

SSE = sum of squared errors

n = the number of data points

When testing the significance of the correlation coefficient, what is the null hypothesis?

When testing the significance of the correlation coefficient, what is the alternative hypothesis?
Solution

$H_a: \rho \neq 0$

If the level of significance is 0.05 and the *p*-value is 0.04, what conclusion can you draw?

Homework from 10.4

If the level of significance is 0.05 and the *p*-value is 0.06, what conclusion can you draw?
Solution

We do not reject the null hypothesis. There is not sufficient evidence to conclude that there is a significant linear relationship between *x* and *y* because the correlation coefficient is not significantly different from zero.

If there are 15 data points in a set of data, what is the number of degrees of freedom?

Homework from 10.5

Recently, the annual number of driver deaths per 100,000 for the selected age groups was as follows:

Age	Number of Driver Deaths per 100,000
17.5	38
22	36
29.5	24
44.5	20
64.5	18
80	28

1. For each age group, pick the midpoint of the interval for the *x* value. (For the 75+ group, use 80.)
2. Using "ages" as the independent variable and "Number of driver deaths per 100,000" as the dependent variable, make a scatter plot of the data.
3. Calculate the least squares (best–fit) line. Put the equation in the form of: $\hat{y} = a + bx$
4. Find the correlation coefficient. Is it significant?
5. Predict the number of deaths for ages 40 and 60.

6. Based on the given data, is there a linear relationship between age of a driver and driver fatality rate?
7. What is the slope of the least squares (best-fit) line? Interpret the slope.

Solution

1.

Age	Number of Driver Deaths per 100,000
16–19	38
20–24	36
25–34	24
35–54	20
55–74	18
75+	28

2. Check student's solution.
3. $\hat{y} = 35.5818045 - 0.19182491x$
4. $r = -0.57874$

   For four df and alpha = 0.05, the LinRegTTest gives p-value = 0.2288 so we do not reject the null hypothesis; there is not a significant linear relationship between deaths and age.

   Using the table of critical values for the correlation coefficient, with four df, the critical value is 0.811. The correlation coefficient $r = -0.57874$ is not less than −0.811, so we do not reject the null hypothesis.
5. if age = 40, $\hat{y}$ (deaths) = 35.5818045 − 0.19182491(40) = 27.9

   if age = 60, $\hat{y}$ (deaths) = 35.5818045 − 0.19182491(60) = 24.1
6. For the entire dataset, there is a linear relationship for the ages up to age 74. The oldest age group shows an increase in deaths from the prior group, which is not consistent with the younger ages.
7. slope = −0.19182491

[link] shows the life expectancy for an individual born in the United States in certain years.

Year of Birth	Life Expectancy
1930	59.7
1940	62.9
1950	70.2
1965	69.7
1973	71.4
1982	74.5
1987	75
1992	75.7
2010	78.7

1. Decide which variable should be the independent variable and which should be the dependent variable.
2. Draw a scatter plot of the ordered pairs.
3. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
4. Find the correlation coefficient. Is it significant?
5. Find the estimated life expectancy for an individual born in 1950 and for one born in 1982.
6. Why aren't the answers to part e the same as the values in [link] that correspond to those years?
7. Use the two points in part e to plot the least squares line on your graph from part b.
8. Based on the data, is there a linear relationship between the year of birth and life expectancy?
9. Are there any outliers in the data?
10. Using the least squares line, find the estimated life expectancy for an individual born in 1850. Does the least squares line give an accurate estimate for that year? Explain why or why not.
11. What is the slope of the least-squares (best-fit) line? Interpret the slope.

The maximum discount value of the Entertainment® card for the "Fine Dining" section, Edition ten, for various pages is given in [link]

Page number	Maximum value ($)
4	16
14	19
25	15
32	17
43	19
57	15
72	16
85	15
90	17

1. Decide which variable should be the independent variable and which should be the dependent variable.
2. Draw a scatter plot of the ordered pairs.
3. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
4. Find the correlation coefficient. Is it significant?
5. Find the estimated maximum values for the restaurants on page ten and on page 70.
6. Does it appear that the restaurants giving the maximum value are placed in the beginning of the "Fine Dining" section? How did you arrive at your answer?
7. Suppose that there were 200 pages of restaurants. What do you estimate to be the maximum value for a restaurant listed on page 200?
8. Is the least squares line valid for page 200? Why or why not?
9. What is the slope of the least-squares (best-fit) line? Interpret the slope.

Solution

1. We wonder if the better discounts appear earlier in the book so we select page as X and discount as Y.
2. Check student's solution.
3. $\hat{y} = 17.21757 - 0.01412x$
4. $r = -0.2752$

   For seven df and alpha = 0.05, using LinRegTTest p-value = 0.4736 so we do not reject; there is not a significant linear relationship between page and discount.

   Using the table of critical values for the correlation coefficient, with seven df, the critical value is 0.666. The correlation coefficient $xi = -0.2752$ is not less than 0.666 so we do not reject.
5. page 10: 17.08 page 70: 16.23

6. There is not a significant linear correlation so it appears there is no relationship between the page and the amount of the discount.
7. page 200: 14.39
8. No, using the regression equation to predict for page 200 is extrapolation.
9. slope = –0.01412

As the page number increases by one page, the discount decreases by $ 0.01412

[link] gives the gold medal times for every other Summer Olympics for the women's 100-meter freestyle (swimming).

Year	Time (seconds)
1912	82.2
1924	72.4
1932	66.8
1952	66.8
1960	61.2
1968	60.0
1976	55.65
1984	55.92
1992	54.64
2000	53.8
2008	53.1

1. Decide which variable should be the independent variable and which should be the dependent variable.
2. Draw a scatter plot of the data.
3. Does it appear from inspection that there is a relationship between the variables? Why or why not?
4. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$.
5. Find the correlation coefficient. Is the decrease in times significant?
6. Find the estimated gold medal time for 1932. Find the estimated time for 1984.
7. Why are the answers from part f different from the chart values?
8. Does it appear that a line is the best way to fit the data? Why or why not?
9. Use the least-squares line to estimate the gold medal time for the next Summer Olympics. Do you think that your answer is reasonable? Why or why not?

State	# letters in name	Year entered the Union	Rank for entering the Union	Area (square miles)
Alabama	7	1819	22	52,423
Colorado	8	1876	38	104,100
Hawaii	6	1959	50	10,932
Iowa	4	1846	29	56,276
Maryland	8	1788	7	12,407
Missouri	8	1821	24	69,709
New Jersey	9	1787	3	8,722
Ohio	4	1803	17	44,828
South Carolina	13	1788	8	32,008
Utah	4	1896	45	84,904
Wisconsin	9	1848	30	65,499

We are interested in whether or not the number of letters in a state name depends upon the year the state entered the Union.

1. Decide which variable should be the independent variable and which should be the dependent variable.
2. Draw a scatter plot of the data.
3. Does it appear from inspection that there is a relationship between the variables? Why or why not?
4. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$.
5. Find the correlation coefficient. What does it imply about the significance of the relationship?
6. Find the estimated number of letters (to the nearest integer) a state would have if it entered the Union in 1900. Find the estimated number of letters a state would have if it entered the Union in 1940.
7. Does it appear that a line is the best way to fit the data? Why or why not?
8. Use the least-squares line to estimate the number of letters a new state that enters the Union this year would have. Can the least squares line be used to predict it? Why or why not?

Solution

1. Year is the independent or x variable; the number of letters is the dependent or y variable.
2. Check student's solution.
3. no
4. $\hat{y} = 47.03 - 0.0216x$
5. −0.4280

6. 6; 5
7. No, the relationship does not appear to be linear; the correlation is not significant.
8. current year: 2013: 3.55 or four letters; this is not an appropriate use of the least squares line. It is extrapolation.

Homework from 10.6

The height (sidewalk to roof) of notable tall buildings in America is compared to the number of stories of the building (beginning at street level).

Height (in feet)	Stories
1,050	57
428	28
362	26
529	40
790	60
401	22
380	38
1,454	110
1,127	100
700	46

1. Using "stories" as the independent variable and "height" as the dependent variable, make a scatter plot of the data.
2. Does it appear from inspection that there is a relationship between the variables?
3. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
4. Find the correlation coefficient. Is it significant?
5. Find the estimated heights for 32 stories and for 94 stories.
6. Based on the data in [link], is there a linear relationship between the number of stories in tall buildings and the height of the buildings?
7. Are there any outliers in the data? If so, which point(s)?
8. What is the estimated height of a building with six stories? Does the least squares line give an accurate estimate of height? Explain why or why not.
9. Based on the least squares line, adding an extra story is predicted to add about how many feet to a

building?

10. What is the slope of the least squares (best-fit) line? Interpret the slope.

Ornithologists, scientists who study birds, tag sparrow hawks in 13 different colonies to study their population. They gather data for the percent of new sparrow hawks in each colony and the percent of those that have returned from migration.

Percent return: 74; 66; 81; 52; 73; 62; 52; 45; 62; 46; 60; 46; 38

Percent new: 5; 6; 8; 11; 12; 15; 16; 17; 18; 18; 19; 20; 20

1. Enter the data into your calculator and make a scatter plot.
2. Use your calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot from part a.
3. Explain in words what the slope and y-intercept of the regression line tell us.
4. How well does the regression line fit the data? Explain your response.
5. Which point has the largest residual? Explain what the residual means in context. Is this point an outlier? An influential point? Explain.
6. An ecologist wants to predict how many birds will join another colony of sparrow hawks to which 70% of the adults from the previous year have returned. What is the prediction?

Solution

a. and b. Check student's solution.

c. The slope of the regression line is -0.3179 with a y-intercept of 32.966. In context, the y-intercept indicates that when there are no returning sparrow hawks, there will be almost 31% new sparrow hawks, which doesn't make sense since if there are no returning birds, then the new percentage would have to be 100% (this is an example of why we do not extrapolate). The slope tells us that for each percentage increase in returning birds, the percentage of new birds in the colony decreases by 0.3179%.

d. If we examine r2, we see that only 50.238% of the variation in the percent of new birds is explained by the model and the correlation coefficient, r = 0.71 only indicates a somewhat strong correlation between returning and new percentages.

e. The ordered pair (66, 6) generates the largest residual of 6.0. This means that when the observed return percentage is 66%, our observed new percentage, 6%, is almost 6% less than the predicted new value of 11.98%. If we remove this data pair, we see only an adjusted slope of -0.2723 and an adjusted intercept of 30.606. In other words, even though this data generates the largest residual, it is not an outlier, nor is the data pair an influential point.

f. If there are 70% returning birds, we would expect to see y = -0.2723(70) + 30.606 = 0.115 or 11.5% new birds in the colony.

The following table shows data on average per capita wine consumption and heart disease rate in a random sample of 10 countries.

Yearly wine consumption in liters	2.5	3.9	2.9	2.4	2.9	0.8	9.1	2.7	0.8	0.7
Death from heart diseases	221	167	131	191	220	297	71	172	211	300

1. Enter the data into your calculator and make a scatter plot.
2. Use your calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot from part a.
3. Explain in words what the slope and y-intercept of the regression line tell us.
4. How well does the regression line fit the data? Explain your response.
5. Which point has the largest residual? Explain what the residual means in context. Is this point an outlier? An influential point? Explain.
6. Do the data provide convincing evidence that there is a linear relationship between the amount of alcohol consumed and the heart disease death rate? Carry out an appropriate test at a significance level of 0.05 to help answer this question.

The following table consists of one student athlete's time (in minutes) to swim 2000 yards and the student's heart rate (beats per minute) after swimming on a random sample of 10 days:

Swim Time	Heart Rate
34.12	144
35.72	152
34.72	124
34.05	140
34.13	152
35.73	146
36.17	128
35.57	136
35.37	144
35.57	148

1. Enter the data into your calculator and make a scatter plot.

2. Use your calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot from part a.
3. Explain in words what the slope and y-intercept of the regression line tell us.
4. How well does the regression line fit the data? Explain your response.
5. Which point has the largest residual? Explain what the residual means in context. Is this point an outlier? An influential point? Explain.

Solution

1. Check student's solution.
2. Check student's solution.
3. We have a slope of –1.4946 with a *y*-intercept of 193.88. The slope, in context, indicates that for each additional minute added to the swim time, the heart rate will decrease by 1.5 beats per minute. If the student is not swimming at all, the *y*-intercept indicates that his heart rate will be 193.88 beats per minute. While the slope has meaning (the longer it takes to swim 2,000 meters, the less effort the heart puts out), the *y*-intercept does not make sense. If the athlete is not swimming (resting), then his heart rate should be very low.
4. Since only 1.5% of the heart rate variation is explained by this regression equation, we must conclude that this association is not explained with a linear relationship.
5. The point (34.72, 124) generates the largest residual of –11.82. This means that our observed heart rate is almost 12 beats less than our predicted rate of 136 beats per minute. When this point is removed, the slope becomes 1.6914 with the y-intercept changing to 83.694. While the linear association is still very weak, we see that the removed data pair can be considered an influential point in the sense that the *y*-intercept becomes more meaningful.

A researcher is investigating whether non-white minorities commit a disproportionate number of homicides. He uses demographic data from Detroit, MI to compare homicide rates and the number of the population that are white males.

White Males	Homicide rate per 100,000 people
558,724	8.6
538,584	8.9
519,171	8.52
500,457	8.89
482,418	13.07
465,029	14.57
448,267	21.36
432,109	28.03
416,533	31.49
401,518	37.39
387,046	46.26
373,095	47.24
359,647	52.33

1. Use your calculator to construct a scatter plot of the data. What should the independent variable be? Why?
2. Use your calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot.
3. Discuss what the following mean in context.
    1. The slope of the regression equation
    2. The y-intercept of the regression equation
    3. The correlation r
    4. The coefficient of determination r2.
4. Do the data provide convincing evidence that there is a linear relationship between the number of white males in the population and the homicide rate? Carry out an appropriate test at a significance level of 0.05 to help answer this question.

School	Mid-Career Salary (in thousands)	Yearly Tuition
Princeton	137	28,540
Harvey Mudd	135	40,133
CalTech	127	39,900
US Naval Academy	122	0
West Point	120	0
MIT	118	42,050
Lehigh University	118	43,220
NYU-Poly	117	39,565
Babson College	117	40,400
Stanford	114	54,506

Using the data to determine the linear-regression line equation with the outliers removed. Is there a linear correlation for the data set with outliers removed? Justify your answer.
Solution

If we remove the two service academies (the tuition is $ 0.00), we construct a new regression equation of $y = -0.0009x + 160$ with a correlation coefficient of 0.71397 and a coefficient of determination of 0.50976. This allows us to say there is a fairly strong linear association between tuition costs and salaries if the service academies are removed from the data set.

CHAPTER 10 PRACTICE

Bring It Together from 10.6

The average number of people in a family that received welfare for various years is given in [link].

Year	Welfare family size
1969	4.0
1973	3.6
1975	3.2
1979	3.0
1983	3.0
1988	3.0
1991	2.9

1. Using "year" as the independent variable and "welfare family size" as the dependent variable, draw a scatter plot of the data.
2. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
3. Find the correlation coefficient. Is it significant?
4. Pick two years between 1969 and 1991 and find the estimated welfare family sizes.
5. Based on the data in [link], is there a linear relationship between the year and the average number of people in a welfare family?
6. Using the least-squares line, estimate the welfare family sizes for 1960 and 1995. Does the least-squares line give an accurate estimate for those years? Explain why or why not.
7. Are there any outliers in the data?
8. What is the estimated average welfare family size for 1986? Does the least squares line give an accurate estimate for that year? Explain why or why not.
9. What is the slope of the least squares (best-fit) line? Interpret the slope.

The percent of female wage and salary workers who are paid hourly rates is given in [link] for the years 1979 to 1992.

Year	Percent of workers paid hourly rates
1979	61.2
1980	60.7
1981	61.3
1982	61.3
1983	61.8
1984	61.7
1985	61.8
1986	62.0
1987	62.7
1990	62.8
1992	62.9

1. Using "year" as the independent variable and "percent" as the dependent variable, draw a scatter plot of the data.
2. Does it appear from inspection that there is a relationship between the variables? Why or why not?
3. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
4. Find the correlation coefficient. Is it significant?
5. Find the estimated percentages for 1991 and 1988.
6. Based on the data, is there a linear relationship between the year and the percent of female wage and salary earners who are paid hourly rates?
7. Are there any outliers in the data?
8. What is the estimated percent for the year 2050? Does the least-squares line give an accurate estimate for that year? Explain why or why not.
9. What is the slope of the least-squares (best-fit) line? Interpret the slope.

Solution

1. Check student's solution.
2. yes
3. $\hat{y} = -266.8863+0.1656x$
4. 0.9448; Yes
5. 62.8233; 62.3265
6. yes
7. yes; (1987, 62.7)

8. 72.5937; no
9. slope = 0.1656.

   As the year increases by one, the percent of workers paid hourly rates tends to increase by 0.1656.

Use the following information to answer the next two exercises. The cost of a leading liquid laundry detergent in different sizes is given in [link].

Size (ounces)	Cost ($)	Cost per ounce
16	3.99	
32	4.99	
64	5.99	
200	10.99	

1. Using "size" as the independent variable and "cost" as the dependent variable, draw a scatter plot.
2. Does it appear from inspection that there is a relationship between the variables? Why or why not?
3. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
4. Find the correlation coefficient. Is it significant?
5. If the laundry detergent were sold in a 40-ounce size, find the estimated cost.
6. If the laundry detergent were sold in a 90-ounce size, find the estimated cost.
7. Does it appear that a line is the best way to fit the data? Why or why not?
8. Are there any outliers in the given data?
9. Is the least-squares line valid for predicting what a 300-ounce size of the laundry detergent would you cost? Why or why not?
10. What is the slope of the least-squares (best-fit) line? Interpret the slope.

1. Complete [link] for the cost per ounce of the different sizes.
2. Using "size" as the independent variable and "cost per ounce" as the dependent variable, draw a scatter plot of the data.
3. Does it appear from inspection that there is a relationship between the variables? Why or why not?
4. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
5. Find the correlation coefficient. Is it significant?
6. If the laundry detergent were sold in a 40-ounce size, find the estimated cost per ounce.
7. If the laundry detergent were sold in a 90-ounce size, find the estimated cost per ounce.
8. Does it appear that a line is the best way to fit the data? Why or why not?
9. Are there any outliers in the data?
10. Is the least-squares line valid for predicting what a 300-ounce size of the laundry detergent would cost

per ounce? Why or why not?

11. What is the slope of the least-squares (best-fit) line? Interpret the slope.

Solution

1.

Size (ounces)	Cost ($)	cents/oz
16	3.99	24.94
32	4.99	15.59
64	5.99	9.36
200	10.99	5.50

2. Check student's solution.
3. There is a linear relationship for the sizes 16 through 64, but that linear trend does not continue to the 200-oz size.
4. $\hat{y} = 20.2368 - 0.0819x$
5. $r = -0.8086$
6. 40-oz: 16.96 cents/oz
7. 90-oz: 12.87 cents/oz
8. The relationship is not linear; the least squares line is not appropriate.
9. no outliers
10. No, you would be extrapolating. The 300-oz size is outside the range of x.
11. slope = –0.08194; for each additional ounce in size, the cost per ounce decreases by 0.082 cents.

According to a flier by a Prudential Insurance Company representative, the costs of approximate probate fees and taxes for selected net taxable estates are as follows:

Net Taxable Estate ($)	Approximate Probate Fees and Taxes ($)
600,000	30,000
750,000	92,500
1,000,000	203,000
1,500,000	438,000
2,000,000	688,000
2,500,000	1,037,000
3,000,000	1,350,000

1. Decide which variable should be the independent variable and which should be the dependent variable.
2. Draw a scatter plot of the data.
3. Does it appear from inspection that there is a relationship between the variables? Why or why not?
4. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$.
5. Find the correlation coefficient. Is it significant?
6. Find the estimated total cost for a next taxable estate of $ 1,000,000. Find the cost for $ 2,500,000.
7. Does it appear that a line is the best way to fit the data? Why or why not?
8. Are there any outliers in the data?
9. Based on these results, what would be the probate fees and taxes for an estate that does not have any assets?
10. What is the slope of the least-squares (best-fit) line? Interpret the slope.

The following are advertised sale prices of color televisions at Anderson's.

Size (inches)	Sale Price ($)
9	147
20	197
27	297
31	447
35	1177
40	2177
60	2497

1. Decide which variable should be the independent variable and which should be the dependent variable.
2. Draw a scatter plot of the data.
3. Does it appear from inspection that there is a relationship between the variables? Why or why not?
4. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
5. Find the correlation coefficient. Is it significant?
6. Find the estimated sale price for a 32 inch television. Find the cost for a 50 inch television.
7. Does it appear that a line is the best way to fit the data? Why or why not?
8. Are there any outliers in the data?
9. What is the slope of the least-squares (best-fit) line? Interpret the slope.

Solution

1. Size is x, the independent variable, price is y, the dependent variable.

2. Check student's solution.
3. The relationship does not appear to be linear.
4. $\hat{y} = -745.252 + 54.75569x$
5. $r = 0.8944$, yes it is significant
6. 32-inch: $ 1006.93, 50-inch: $ 1992.53
7. No, the relationship does not appear to be linear. However, r is significant.
8. yes, the 60-inch TV
9. For each additional inch, the price increases by $ 54.76

[link] shows the average heights for American boys in 1990.

Age (years)	Height (cm)
birth	50.8
2	83.8
3	91.4
5	106.6
7	119.3
10	137.1
14	157.5

1. Decide which variable should be the independent variable and which should be the dependent variable.
2. Draw a scatter plot of the data.
3. Does it appear from inspection that there is a relationship between the variables? Why or why not?
4. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
5. Find the correlation coefficient. Is it significant?
6. Find the estimated average height for a one-year-old. Find the estimated average height for an eleven-year-old.
7. Does it appear that a line is the best way to fit the data? Why or why not?
8. Are there any outliers in the data?
9. Use the least squares line to estimate the average height for a sixty-two-year-old man. Do you think that your answer is reasonable? Why or why not?
10. What is the slope of the least-squares (best-fit) line? Interpret the slope.

State	# letters in name	Year entered the Union	Ranks for entering the Union	Area (square miles)
Alabama	7	1819	22	52,423
Colorado	8	1876	38	104,100
Hawaii	6	1959	50	10,932
Iowa	4	1846	29	56,276
Maryland	8	1788	7	12,407
Missouri	8	1821	24	69,709
New Jersey	9	1787	3	8,722
Ohio	4	1803	17	44,828
South Carolina	13	1788	8	32,008
Utah	4	1896	45	84,904
Wisconsin	9	1848	30	65,499

We are interested in whether there is a relationship between the ranking of a state and the area of the state.

1. What are the independent and dependent variables?
2. What do you think the scatter plot will look like? Make a scatter plot of the data.
3. Does it appear from inspection that there is a relationship between the variables? Why or why not?
4. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
5. Find the correlation coefficient. What does it imply about the significance of the relationship?
6. Find the estimated areas for Alabama and for Colorado. Are they close to the actual areas?
7. Use the two points in part f to plot the least-squares line on your graph from part b.
8. Does it appear that a line is the best way to fit the data? Why or why not?
9. Are there any outliers?
10. Use the least squares line to estimate the area of a new state that enters the Union. Can the least-squares line be used to predict it? Why or why not?
11. Delete "Hawaii" and substitute "Alaska" for it. Alaska is the forty-ninth, state with an area of 656,424 square miles.
12. Calculate the new least-squares line.
13. Find the estimated area for Alabama. Is it closer to the actual area with this new least-squares line or with the previous one that included Hawaii? Why do you think that's the case?
14. Do you think that, in general, newer states are larger than the original states?

Solution

1. Let rank be the independent variable and area be the dependent variable.
2. Check student's solution.
3. There appears to be a linear relationship, with one outlier.
4. $\hat{y}$ (area) = 24177.06 + 1010.478x
5. r = 0.50047, r is not significant so there is no relationship between the variables.
6. Alabama: 46407.576 Colorado: 62575.224
7. Alabama estimate is closer than Colorado estimate.
8. If the outlier is removed, there is a linear relationship.
9. There is one outlier (Hawaii).
10. rank 51: 75711.4; no
11.

Alabama	**7**	**1819**	**22**	**52,423**
Colorado	8	1876	38	104,100
Alaska	6	1959	51	656,424
Iowa	4	1846	29	56,276
Maryland	8	1788	7	12,407
Missouri	8	1821	24	69,709
New Jersey	9	1787	3	8,722
Ohio	4	1803	17	44,828
South Carolina	13	1788	8	32,008
Utah	4	1896	45	84,904
Wisconsin	9	1848	30	65,499

12. $\hat{y}$ = −87065.3 + 7828.532x
13. Alabama: 85,162.404; the prior estimate was closer. Alaska is an outlier.
14. yes, with the exception of Hawaii

CHAPTER 11: THE CHI-SQUARE DISTRIBUTION

INTRODUCTION TO CHAPTER 11: THE CHI-SQUARE DISTRIBUTION



Figure 11.1. The chi-square distribution can be used to find relationships between two things, like grocery prices at different stores. (credit: Pete/flickr)

Have you ever wondered if lottery numbers were evenly distributed or if some numbers occurred with a greater frequency? How about if the types of movies people preferred were different across different age groups? What about if a coffee machine was dispensing approximately the same amount of coffee each time? You could answer these questions by conducting a hypothesis test.

You will now study a new distribution, one that is used to determine the answers to such questions. This distribution is called the chi-square distribution.

In this chapter, you will learn the three major applications of the chi-square distribution:

1. the goodness-of-fit test, which determines if data fit a particular distribution, such as in the lottery

example

2. the test of independence, which determines if events are independent, such as in the movie example
3. the test of a single variance, which tests variability, such as in the coffee example

Note

Though the chi-square distribution depends on calculators or computers for most of the calculations, there is a table available (see [link]). TI-83+ and TI-84 calculator instructions are included in the text.

Collaborative Exercises

Section 11.1

Collaborative Exercise

Look in the sports section of a newspaper or on the Internet for some sports data (baseball averages, basketball scores, golf tournament scores, football odds, swimming times, and the like). Plot a histogram and a box plot using your data. See if you can determine a probability distribution that your data fits. Have a discussion with the class about your choice.

Media Attributions

- Chapter 11 Introduction Figure © OpenStax Introductory Statistics is licensed under a CC BY (Attribution) license

11.1 FACTS ABOUT THE CHI-SQUARE DISTRIBUTION

Learning Objectives

By the end of this section, the student should be able to:

- define Chi-Square Distribution, and apply the concept to problem solving

The notation for the chi-square distribution is:

$$\chi \sim \chi^2_{df}$$

where df = degrees of freedom which depends on how chi-square is being used. (If you want to practice calculating chi-square probabilities then use $df = n - 1$. The degrees of freedom for the three major uses are each calculated differently.)

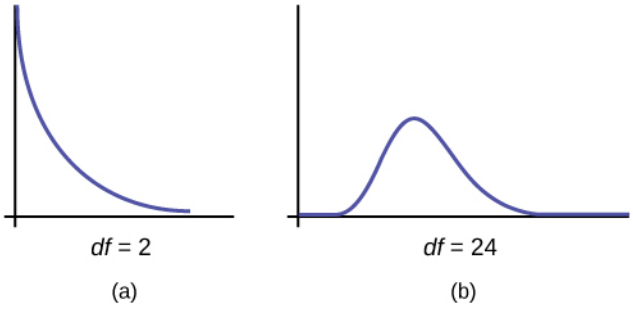
For the χ^2 distribution, the population mean is $\mu = df$ and the population standard deviation is $\sigma = \sqrt{2\,(df)}$.

The random variable is shown as χ^2, but may be any uppercase letter.

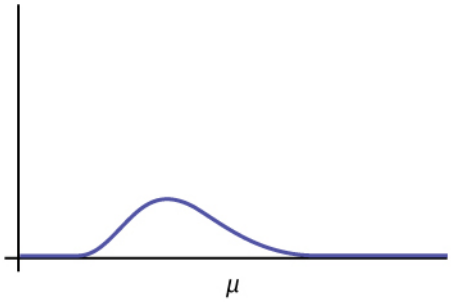
The random variable for a chi-square distribution with k degrees of freedom is the sum of k independent, squared standard normal variables.

$$\chi^2 = (Z_1)^2 + (Z_2)^2 + \ldots + (Z_k)^2$$

1. The curve is nonsymmetrical and skewed to the right.
2. There is a different chi-square curve for each df.



3. The test statistic for any test is always greater than or equal to zero.
4. When $df > 90$, the chi-square curve approximates the normal distribution. For $X \sim \chi^2_{1,000}$ the mean, $\mu = df = 1{,}000$ and the standard deviation, $\sigma = \sqrt{2\,(1,000)} = 44.7$. Therefore, $X \sim N(1{,}000, 44.7)$, approximately.
5. The mean, μ, is located just to the right of the peak.



References

Data from *Parade Magazine*.

"HIV/AIDS Epidemiology Santa Clara County." Santa Clara County Public Health Department, May 2011.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=934#oembed-1

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=934#oembed-2

11.2 GOODNESS-OF-FIT TEST

Learning Objectives

By the end of this section, the student should be able to:

- calculate and interpret the Chi-Square Goodness-of-Fit test

In this type of hypothesis test, you determine whether the data **"fit"** a particular distribution or not. For example, you may suspect your unknown data fit a binomial distribution. You use a chi-square test (meaning the distribution for the hypothesis test is chi-square) to determine if there is a fit or not. **The null and the alternative hypotheses for this test may be written in sentences or may be stated as equations or inequalities.**

The test statistic for a goodness-of-fit test is:

$$\sum_{k} \frac{(O-E)^2}{E}$$

where:

- O = observed values (data)
- E = expected values (from theory)
- k = the number of different data cells or categories

The observed values are the data values and the expected values are the values you would expect to get if the null hypothesis were true. There are n terms of the form $\frac{(O-E)^2}{E}$.

The number of degrees of freedom is df = (number of categories – 1).

The goodness-of-fit test is almost always right-tailed. If the observed values and the corresponding

expected values are not close to each other, then the test statistic can get very large and will be way out in the right tail of the chi-square curve.

Note

The expected value for each cell needs to be at least five in order for you to use this test.

HTTPS://WWW.YOUTUBE.COM/WATCH?V=GG7DFNVBMKI

Example

Absenteeism of college students from math classes is a major concern to math instructors because missing class appears to increase the drop rate. Suppose that a study was done to determine if the actual student absenteeism rate follows faculty perception. The faculty expected that a group of 100 students would miss class according to [link].

Number of absences per term	Expected number of students
0–2	50
3–5	30
6–8	12
9–11	6
12+	2

A random survey across all mathematics courses was then done to determine the actual number (observed) of absences in a course. The chart in [link] displays the results of that survey.

Number of absences per term	Actual number of students
0–2	35
3–5	40
6–8	20
9–11	1
12+	4

Determine the null and alternative hypotheses needed to conduct a goodness-of-fit test.

H0: Student absenteeism fits faculty perception.

The alternative hypothesis is the opposite of the null hypothesis.

Ha: Student absenteeism does not fit faculty perception.

a. Can you use the information as it appears in the charts to conduct the goodness-of-fit test?

Solution

a. No. Notice that the expected number of absences for the "12+" entry is less than five (it is two). Combine that group with the "9–11" group to create new tables where the number of students for each entry are at least five. The new results are in [link] and [link].

Number of absences per term	Expected number of students
0–2	50
3–5	30
6–8	12
9+	8

Number of absences per term	Actual number of students
0–2	35
3–5	40
6–8	20
9+	5

b. What is the number of degrees of freedom (*df*)?

Solution

b. There are four "cells" or categories in each of the new tables.

df = number of cells – 1 = 4 – 1 = 3

Your Turn!

A factory manager needs to understand how many products are defective versus how many are produced. The number of expected defects is listed in [link].

Number produced	Number defective
0–100	5
101–200	6
201–300	7
301–400	8
401–500	10

A random sample was taken to determine the actual number of defects. [link] shows the results of the survey.

Number produced	Number defective
0–100	5
101–200	7
201–300	8
301–400	9
401–500	11

State the null and alternative hypotheses needed to conduct a goodness-of-fit test, and state the degrees of freedom.

Solution

H_0:The number of defaults fits expectations.

H_a:The number of defaults does not fit expectations.

$df = 4$

Example

Employers want to know which days of the week employees are absent in a five-day work week. Most employers would like to believe that employees are absent equally during the week. Suppose a random sample of 60 managers were asked on which day of the week they had the highest number of employee absences. The results were distributed as in [link]. For the population of employees, do the days for the highest number of absences occur with equal frequencies during a five-day work week? Test at a 5% significance level.

Day of the Week Employees Were Most Absent

	Monday	Tuesday	Wednesday	Thursday	Friday
Number of Absences	15	12	9	9	15

Solution
The null and alternative hypotheses are:

- H0: The absent days occur with equal frequencies, that is, they fit a uniform distribution.
- Ha: The absent days occur with unequal frequencies, that is, they do not fit a uniform distribution.

If the absent days occur with equal frequencies, then, out of 60 absent days (the total in the sample: 15 + 12 + 9 + 9 + 15 = 60), there would be 12 absences on Monday, 12 on Tuesday, 12 on Wednesday, 12 on Thursday, and 12 on Friday. These numbers are the expected (E) values. The values in the table are the observed (O) values or data.

This time, calculate the χ^2 test statistic by hand. Make a chart with the following headings and fill in the columns:

- Expected (E) values (12, 12, 12, 12, 12)
- Observed (O) values (15, 12, 9, 9, 15)
- (O – E)
- (O – E)2
- $\frac{(O-E)^2}{E}$

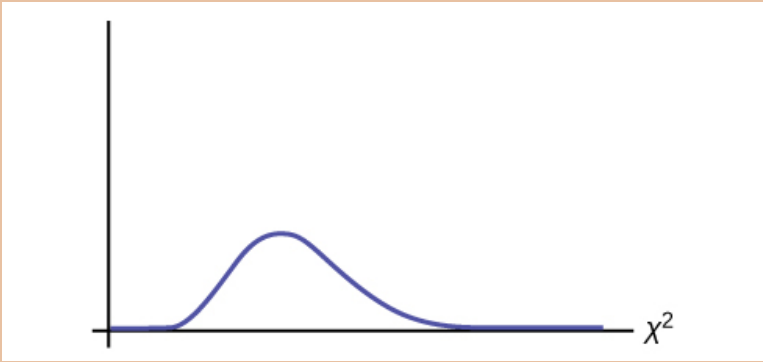
Now add (sum) the last column. The sum is three. This is the χ^2 test statistic.

To find the p-value, calculate $P(\chi^2 > 3)$. This test is right-tailed. (Use a computer or calculator to find the p-value. You should get p-value = 0.5578.)

The dfs are the number of cells – 1 = 5 – 1 = 4

Press 2nd DISTR. Arrow down to χ2cdf. Press ENTER. Enter (3,10^99,4). Rounded to four decimal places, you should see 0.5578, which is the p-value.

Next, complete a graph like the following one with the proper labeling and shading. (You should shade the right tail.)



The decision is not to reject the null hypothesis.

Conclusion: At a 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the absent days do not occur with equal frequencies.

TI-83+ and some TI-84 calculators do not have a special program for the test statistic for the goodness-of-fit test. The next example [link] has the calculator instructions. The newer TI-84 calculators have in STAT TESTS the test Chi2 GOF. To run the test, put the observed values (the data) into a first list and the expected values (the values you expect if the null hypothesis is true) into a second list. Press STAT TESTS and Chi2 GOF. Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press calculate or draw. Make sure you clear any lists before you start. To Clear Lists in the calculators: Go into STAT EDIT and arrow up to the list name area of the particular list. Press CLEAR and then arrow down. The list will be cleared. Alternatively, you can press STAT and press 4 (for ClrList). Enter the list name and press ENTER.

Your Turn!

Teachers want to know which night each week their students are doing most of their homework. Most teachers think that students do homework equally throughout the week. Suppose a random sample of 49 students were asked on which night of the week they did the most homework. The results were distributed as in [link].

	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
Number of Students	11	8	10	7	10	5	5

From the population of students, do the nights for the highest number of students doing the majority of their homework occur with equal frequencies during a week? What type of hypothesis test should you use?

Solution
$df = 6$

p-value = 0.6093

We decline to reject the null hypothesis. There is not enough evidence to support that students do not do the majority of their homework equally throughout the week.

Example

One study indicates that the number of televisions that American families have is distributed (this is the given distribution for the American population) as in [link].

Number of Televisions	Percent
0	10
1	16
2	55
3	11
4+	8

The table contains expected (E) percentages.

A random sample of 600 families in the far western United States resulted in the data in [link].

Number of Televisions	Frequency
0	66
1	119
2	340
3	60
4+	15
	Total = 600

The table contains observed (O) frequency values.

At the 1% significance level, does it appear that the distribution "number of televisions" of far western United States families is different from the distribution for the American population as a whole?

Solution

This problem asks you to test whether the far western United States families distribution fits the distribution of the American families. This test is always right-tailed.

The first table contains expected percentages. To get expected (E) frequencies, multiply the percentage by 600. The expected frequencies are shown in [link].

Number of Televisions	Percent	Expected Frequency
0	10	(0.10)(600) = 60
1	16	(0.16)(600) = 96
2	55	(0.55)(600) = 330
3	11	(0.11)(600) = 66
over 3	8	(0.08)(600) = 48

Therefore, the expected frequencies are 60, 96, 330, 66, and 48. In the TI calculators, you can let the calculator do the math. For example, instead of 60, enter 0.10*600.

H0: The "number of televisions" distribution of far western United States families is the same as the "number of televisions" distribution of the American population.

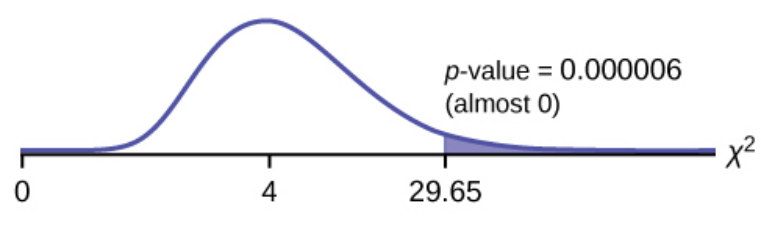
Ha: The "number of televisions" distribution of far western United States families is different from the "number of televisions" distribution of the American population.

Distribution for the test: χ^2_4 where df = (the number of cells) – 1 = 5 – 1 = 4.

Note
df ≠ 600 – 1

Calculate the test statistic:χ2 = 29.65

Graph:



Probability statement: p-value = P(χ2 > 29.65) = 0.000006

Compare α and the p-value:

α = 0.01
p-value = 0.000006
So, α > p-value.

Make a decision: Since α > p-value, reject Ho.

This means you reject the belief that the distribution for the far western states is the same as that of the American population as a whole.

Conclusion: At the 1% significance level, from the data, there is sufficient evidence to conclude that the "number of televisions" distribution for the far western United States is different from the "number of televisions" distribution for the American population as a whole.

Press STAT and ENTER. Make sure to clear lists L1, L2, and L3 if they have data in them (see the note at the end of [link]). Into L1, put the observed frequencies 66, 119, 349, 60, 15. Into L2, put the expected frequencies .10*600, .16*600, .55*600, .11*600, .08*600. Arrow over to list L3 and up to the name area "L3". Enter (L1-L2)^2/L2 and ENTER. Press 2nd QUIT. Press 2nd LIST and arrow over to MATH. Press 5. You should see "sum" (Enter L3). Rounded to 2 decimal places, you should see 29.65. Press 2nd DISTR. Press 7 or Arrow down to 7:χ2cdf and press ENTER. Enter (29.65,1E99,4). Rounded to four places, you should see 5.77E-6 = .000006 (rounded to six decimal places), which is the p-value.

The newer TI-84 calculators have in STAT TESTS the test Chi2 GOF. To run the test, put the observed values (the data) into a first list and the expected values (the values you expect if the null hypothesis is true) into a second list. Press STAT TESTS and Chi2 GOF. Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press calculate or draw. Make sure you clear any lists before you start.

Your Turn!

The expected percentage of the number of pets students have in their homes is distributed (this is the given distribution for the student population of the United States) as in [link].

Number of Pets	Percent
0	18
1	25
2	30
3	18
4+	9

A random sample of 1,000 students from the Eastern United States resulted in the data in [link].

Number of Pets	Frequency
0	210
1	240
2	320
3	140
4+	90

At the 1% significance level, does it appear that the distribution "number of pets" of students in the Eastern United States is different from the distribution for the United States student population as a whole? What is the *p*-value?

Solution

p-value = 0.0036

We reject the null hypothesis that the distributions are the same. There is sufficient evidence to conclude that the distribution "number of pets" of students in the Eastern United States is different from the distribution for the United States student population as a whole.

Example

Suppose you flip two coins 100 times. The results are 20 HH, 27 HT, 30 TH, and 23 TT. Are the coins fair? Test at a 5% significance level.

Solution

This problem can be set up as a goodness-of-fit problem. The sample space for flipping two fair coins is {HH, HT, TH, TT}. Out of 100 flips, you would expect 25 HH, 25 HT, 25 TH, and 25 TT. This is the expected distribution. The question, "Are the coins fair?" is the same as saying, "Does the distribution of the coins (20 HH, 27 HT, 30 TH, 23 TT) fit the expected distribution?"

Random Variable: Let X = the number of heads in one flip of the two coins. X takes on the values 0, 1, 2. (There are 0, 1, or 2 heads in the flip of two coins.) Therefore, the number of cells is three. Since X = the number of heads, the observed frequencies are 20 (for two heads), 57 (for one head), and 23 (for zero heads or both tails). The expected frequencies are 25 (for two heads), 50 (for one head), and 25 (for zero heads or both tails). This test is right-tailed.

H0: The coins are fair.

Ha: The coins are not fair.

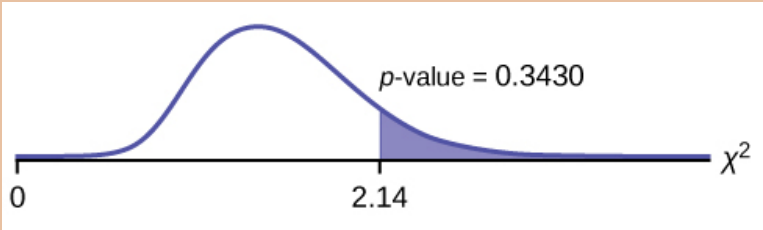
Distribution for the test: χ^2_2

where

df = 3 – 1 = 2.

Calculate the test statistic: $\chi 2 = 2.14$

Graph:



Probability statement: p-value = $P(\chi 2 > 2.14) = 0.3430$

Compare α and the p-value:

α = 0.05

p-value = 0.3430

α < p-value.

Make a decision: Since α < p-value, do not reject H0.

Conclusion: There is insufficient evidence to conclude that the coins are not fair.

Press STAT and ENTER. Make sure you
clear lists L1, L2, and L3 if they have data in them. Into L1, put the observed frequencies 20, 57, 23. Into L2, put the expected frequencies 25, 50, 25. Arrow over to list L3 and up to the name area "L3". Enter (L1-L2)^2/L2 and ENTER. Press 2nd QUIT. Press 2nd LIST and arrow over to MATH. Press 5. You should see "sum". Enter L3. Rounded to two decimal places, you should see 2.14. Press 2nd DISTR. Arrow down to 7:χ2cdf (or press 7). Press ENTER. Enter 2.14,1E99,2). Rounded to four places, you should see .3430, which is the p-value.

The newer TI-84 calculators have in STAT TESTS the test Chi2 GOF. To run the test, put the observed values (the data) into a first list and the expected values (the values you expect if the null hypothesis is true) into a second list. Press STAT TESTS and Chi2 GOF. Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press calculate or draw. Make
sure you clear any lists before you start.

Your Turn!

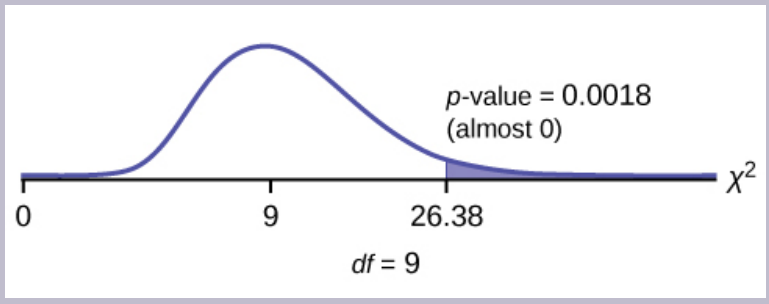
Students in a social studies class hypothesize that the literacy rates across the world for every region are 82%. [link] shows the actual literacy rates across the world broken down by region. What are the test statistics and the degrees of freedom?

MDG Region	Adult Literacy Rate (%)
Developed Regions	99.0
Commonwealth of Independent States	99.5
Northern Africa	67.3
Sub-Saharan Africa	62.5
Latin America and the Caribbean	91.0
Eastern Asia	93.8
Southern Asia	61.9
South-Eastern Asia	91.9
Western Asia	84.5
Oceania	66.4

Solution

degrees of freedom = 9

chi^2 test statistic = 26.38



Press STAT and ENTER. Make sure you clear lists L1, L2, and L3 if they have data in them. Into L1, put the observed frequencies 99, 99.5, 67.3, 62.5, 91, 93.8, 61.9, 91.9, 84.5, 66.4. Into L2, put the expected frequencies 82, 82, 82, 82, 82, 82, 82, 82, 82, 82. Arrow over to list L3 and up to the name area "L3". Enter (L1−L2)^2/L2 and ENTER. Press 2nd QUIT. Press 2nd LIST and arrow over to MATH. Press 5. You should see "sum(". Enter L3). Rounded to two decimal places, you should see 26.38. Press 2nd DISTR. Arrow down to 7:χ2cdf((or press 7). Press ENTER. Enter 26.38,1E99,9). Rounded to four places, you should see .0018, which is the *p*-value.

The newer TI-84 calculators have in STAT TESTS the test Chi2 GOF. To run the test, put the observed values (the data) into a first list and the expected values (the values you expect if the null hypothesis is true) into a second list. Press STAT TESTS and Chi2 GOF. Enter the list names

for the Observed list and the Expected list. Enter the degrees of freedom and press `calculate` or `draw`. Make sure you clear any lists before you start.

References

Data from the U.S. Census Bureau

Data from the College Board. Available online at http://www.collegeboard.com.

Data from the U.S. Census Bureau, Current Population Reports.

Ma, Y., E.R. Bertone, E.J. Stanek III, G.W. Reed, J.R. Hebert, N.L. Cohen, P.A. Merriam, I. S. Ockene, "Association between Eating Patterns and Obesity in a Free-living US Adult Population." *American Journal of Epidemiology* volume 158, no. 1, pages 85-92.

Ogden, Cynthia L., Margaret D. Carroll, Brian K. Kit, Katherine M. Flegal, "Prevalence of Obesity in the United States, 2009–2010." NCHS Data Brief no. 82, January 2012. Available online at http://www.cdc.gov/nchs/data/databriefs/db82.pdf (accessed May 24, 2013).

Stevens, Barbara J., "Multi-family and Commercial Solid Waste and Recycling Survey." Arlington Count, VA. Available online at http://www.arlingtonva.us/departments/EnvironmentalServices/SW/file84429.pdf (accessed May 24,2013).

Use the following information to answer the next nine exercises: The following data are real. The cumulative number of AIDS cases reported for Santa Clara County is broken down by ethnicity as in [link].

11.3 TEST OF INDEPENDENCE

Learning Objectives

By the end of this section, the student should be able to:

- calculate the test of independence and determines whether two factors are independent or not

Tests of independence involve using a contingency table of observed (data) values.

The test statistic for a test of independence is similar to that of a goodness-of-fit test:

$$\sum_{(i \cdot j)} \frac{(O-E)^2}{E}$$

where:

- O = observed values
- E = expected values
- i = the number of rows in the table
- j = the number of columns in the table

There are $i \cdot j$ terms of the form $\frac{(O-E)^2}{E}$.

A test of independence determines whether two factors are independent or not. You first encountered the term independence in Probability Topics. As a review, consider the following example.

Note

The expected value for each cell needs to be at least five in order for you to use this test.



ONE OR MORE INTERACTIVE ELEMENTS HAS BEEN EXCLUDED FROM THIS VERSION OF THE TEXT. YOU CAN VIEW THEM ONLINE HERE:

HTTPS://LOUIS.PRESSBOOKS.PUB/INTRODUCTORYSTATISTICS/?P=944#OEMBED-1

Example

Suppose A = a speeding violation in the last year and B = a cell phone user while driving. If A and B are independent then P(A AND B) = P(A)P(B). A AND B is the event that a driver received a speeding violation last year and also used a cell phone while driving. Suppose, in a study of drivers who received speeding violations in the last year and who used cell phones while driving, that 755 people were surveyed. Out of the 755, 70 had a speeding violation and 685 did not; 305 used cell phones while driving and 450 did not.

Let y = expected number of drivers who used a cell phone while driving and received speeding violations.

If A and B are independent, then P(A AND B) = P(A)P(B). By substitution,

$$\frac{y}{755} = \left(\frac{70}{755}\right)\left(\frac{305}{755}\right)$$

Solve for y: $y = \frac{(70)(305)}{755} = 28.3$

About 28 people from the sample are expected to use cell phones while driving and to receive speeding violations.

In a test of independence, we state the null and alternative hypotheses in words. Since the contingency table consists of two factors, the null hypothesis states that the factors are independent and the alternative hypothesis states that they are not independent (dependent). If we do a test of independence using the example, then the null hypothesis is:

H0: Being a cell phone user while driving and receiving a speeding violation are independent events.

If the null hypothesis were true, we would expect about 28 people to use cell phones while driving and to receive a speeding violation.

The test of independence is always right-tailed because of the calculation of the test statistic. If the expected and observed values are not close together, then the test statistic is very large and way out in the right tail of the chi-square curve, as it is in a goodness-of-fit.

The number of degrees of freedom for the test of independence is:

df = (number of columns – 1)(number of rows – 1)

The following formula calculates the expected number (E):

$$E = \frac{(\text{row total})(\text{column total})}{\text{total number surveyed}}$$

Your Turn!

A sample of 300 students is taken. Of the students surveyed, 50 were music students, while 250 were not. Ninety-seven were on the honor roll, while 203 were not. If we assume being a music student and being on the honor roll are independent events, what is the expected number of music students who are also on the honor roll?

Solution
About 16 students are expected to be music students and on the honor roll.

Example

In a volunteer group, adults 21 and older volunteer from one to nine hours each week to spend time with a disabled senior citizen. The program recruits among community college students, four-year college students, and nonstudents. In [link] is a sample of the adult volunteers and the number of hours they volunteer per week.

Number of Hours Worked per Week by Volunteer Type (Observed). The table contains observed (O) values (data).

Type of Volunteer	1–3 Hours	4–6 Hours	7–9 Hours	Row Total
Community College Students	111	96	48	255
Four-Year College Students	96	133	61	290
Nonstudents	91	150	53	294
Column Total	298	379	162	839

Is the number of hours volunteered independent of the type of volunteer?

Solution

The observed table and the question at the end of the problem, "Is the number of hours volunteered independent of the type of volunteer?" tell you this is a test of independence. The two factors are number of hours volunteered and type of volunteer. This test is always right-tailed.

H0: The number of hours volunteered is independent of the type of volunteer.

Ha: The number of hours volunteered is dependent on the type of volunteer.

The expected results are in [link].

Number of Hours Worked per Week by Volunteer Type (Expected). The table contains expected (E) values (data).

Type of Volunteer	1-3 Hours	4-6 Hours	7-9 Hours
Community College Students	90.57	115.19	49.24
Four-Year College Students	103.00	131.00	56.00
Nonstudents	104.42	132.81	56.77

For example, the calculation for the expected frequency for the top left cell is

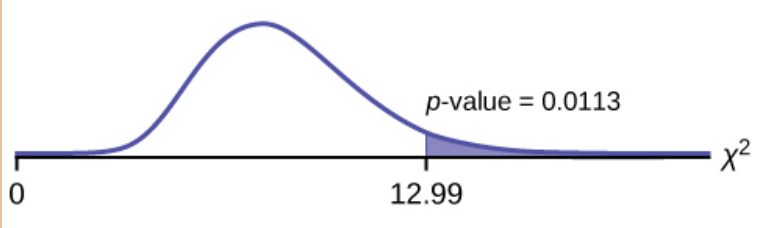
$$E = \frac{(\text{row total})\,(\text{column total})}{\text{total number surveyed}} = \frac{(255)\,(298)}{839} = 90.57$$

Calculate the test statistic:χ2 = 12.99 (calculator or computer)

Distribution for the test:χ^2_4

df = (3 columns – 1)(3 rows – 1) = (2)(2) = 4

Graph:



Probability statement:p-value=P(χ2 > 12.99) = 0.0113

Compare α and the p-value: Since no α is given, assume α = 0.05. p-value = 0.0113. α > p-value.

Make a decision: Since α > p-value, reject H0. This means that the factors are not independent.

Conclusion: At a 5% level of significance, from the data, there is sufficient evidence to conclude that the number of hours volunteered and the type of volunteer are dependent on one another.

For the example in [link], if there had been another type of volunteer, teenagers, what would the degrees of freedom be?

Press the MATRX key and arrow over to EDIT. Press 1:[A]. Press 3 ENTER 3 ENTER. Enter the table values by row from [link]. Press ENTER after each. Press 2nd QUIT. Press STAT and arrow over to TESTS. Arrow down to C:χ2-TEST. Press ENTER. You should see Observed:[A] and Expected:[B]. Arrow down to Calculate. Press ENTER. The test statistic is 12.9909 and the p-value = 0.0113. Do the procedure a second time, but arrow down to Draw instead of calculate.

Your Turn!

The Bureau of Labor Statistics gathers data about employment in the United States. A sample is taken to calculate the number of U.S. citizens working in one of several industry sectors over time. [link] shows the results:

Industry Sector	2000	2010	2020	Total
Nonagriculture wage and salary	13,243	13,044	15,018	41,305
Goods-producing, excluding agriculture	2,457	1,771	1,950	6,178
Services-providing	10,786	11,273	13,068	35,127
Agriculture, forestry, fishing, and hunting	240	214	201	655
Nonagriculture self-employed and unpaid family worker	931	894	972	2,797
Secondary wage and salary jobs in agriculture and private household industries	14	11	11	36
Secondary jobs as a self-employed or unpaid family worker	196	144	152	492
Total	27,867	27,351	31,372	86,590

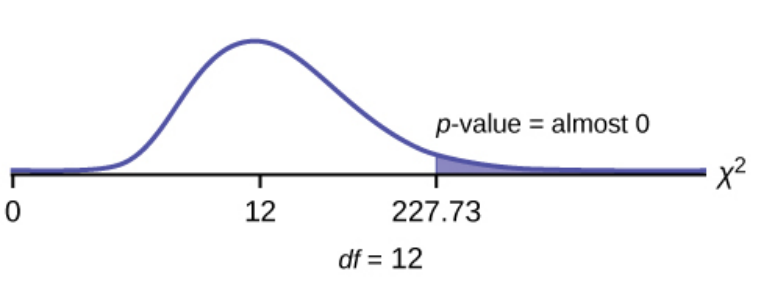
We want to know if the change in the number of jobs is independent of the change in years. State the null and alternative hypotheses and the degrees of freedom.

Solution

H_0 : The number of jobs is independent of the year.

H_a : The number of jobs is dependent on the year.

$df = 12$



Press the MATRX key and arrow over to EDIT. Press 1:[A]. Press 3 ENTER 3 ENTER. Enter the table values by row. Press ENTER after each. Press 2nd QUIT.

Press STAT and arrow over to TESTS. Arrow down to C:χ2-TEST. Press ENTER. You should see Observed:[A] and Expected:[B]. Arrow down to Calculate. Press ENTER. The test statistic is 227.73 and the *p*-value = 5.90E – 42 = 0. Do the procedure a second time but arrow down to Draw instead of calculate.

Example

De Anza College is interested in the relationship between anxiety level and the need to succeed in school. A random sample of 400 students took a test that measured anxiety level and the need to succeed in school. [link] shows the results. De Anza College wants to know if anxiety level and need to succeed in school are independent events.

Need to Succeed in School vs. Anxiety Level

Need to Succeed in School	High Anxiety	Med-high Anxiety	Medium Anxiety	Med-low Anxiety	Low Anxiety	Row Total
High Need	35	42	53	15	10	155
Medium Need	18	48	63	33	31	193
Low Need	4	5	11	15	17	52
Column Total	57	95	127	63	58	400

a. How many high anxiety level students are expected to have a high need to succeed in school?

Solution

a. The column total for a high anxiety level is 57. The row total for high need to succeed in school is 155. The sample size or total surveyed is 400.

$$E = \frac{(\text{row total})(\text{column total})}{\text{total surveyed}} = \frac{155 \cdot 57}{400} = 22.09$$

The expected number of students who have a high anxiety level and a high need to succeed in school is about 22.

b. If the two variables are independent, how many students do you expect to have a low need to succeed in school and a med-low level of anxiety?

Solution
b. The column total for a med-low anxiety level is 63. The row total for a low need to succeed in school is 52. The sample size or total surveyed is 400.

c. $E = \frac{(\text{row total})(\text{column total})}{\text{total surveyed}}$ = _____

Solution

c. $E = \frac{(\text{row total})(\text{column total})}{\text{total surveyed}} = 8.19$

d. The expected number of students who have a med-low anxiety level and a low need to succeed in school is about _____.

Solution
d. 8

Your Turn!

Refer back to the information in [link]. How many service providing jobs are there expected to be in 2020? How many nonagriculture wage and salary jobs are there expected to be in 2020?

Solution
12,727, 14,965

References

DiCamilo, Mark, Mervin Field. "Most Californians See a Direct Linkage between Obesity and Sugary Sodas. Two in Three Voters Support Taxing Sugar-Sweetened Beverages If Proceeds Are Tied to Improving School

Nutrition and Physical Activity Programs." The Field Poll, released Feb. 14, 2013. Available online at http://field.com/fieldpollonline/subscribers/Rls2436.pdf (accessed May 24, 2013).

Harris Interactive. "Favorite Flavor of Ice Cream." Available online at http://www.statisticbrain.com/favorite-flavor-of-ice-cream (accessed May 24, 2013)

"Youngest Online Entrepreneurs List." Available online at http://www.statisticbrain.com/youngest-online-entrepreneur-list (accessed May 24, 2013).

11.4 TEST FOR HOMOGENEITY

Learning Objectives

By the end of this section, the student should be able to:

- calculate the test statistic for a test for homogeneity
- test for homogeneity, and draw a conclusion about whether two populations have the same distribution

The goodness–of–fit test can be used to decide whether a population fits a given distribution, but it will not suffice to decide whether two populations follow the same unknown distribution. A different test, called the test for homogeneity, can be used to draw a conclusion about whether two populations have the same distribution. To calculate the test statistic for a test for homogeneity, follow the same procedure as with the test of independence.

Note

The expected value for each cell needs to be at least five in order for you to use this test.

Hypotheses

H_0: The distributions of the two populations are the same.

H_a: The distributions of the two populations are not the same.

Test Statistic: Use a χ^2 test statistic. It is computed in the same way as the test for independence.

Degrees of Freedom (df)df = number of columns – 1

Requirements: All values in the table must be greater than or equal to five.

Common Uses: Comparing two populations. For example: men vs. women, before vs. after, east vs. west. The variable is categorical with more than two possible response values.



One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=946#oembed-1

Example

Do male and female college students have the same distribution of living arrangements? Use a level of significance of 0.05. Suppose that 250 randomly selected male college students and 300 randomly selected female college students were asked about their living arrangements: dormitory, apartment, with parents, other. The results are shown in [link]. Do male and female college students have the same distribution of living arrangements?

Distribution of Living Arrangements for College Males and College Females

	Dormitory	Apartment	With Parents	Other
Males	72	84	49	45
Females	91	86	88	35

Solution

H0: The distribution of living arrangements for male college students is the same as the distribution of living arrangements for female college students.

Ha: The distribution of living arrangements for male college students is not the same as the distribution of living arrangements for female college students.

Degrees of Freedom (df):

df = number of columns – 1 = 4 – 1 = 3

Distribution for the test: χ^2_3

Calculate the test statistic: $\chi 2 = 10.1287$ (calculator or computer)

Probability statement: p-value = $P(\chi 2 > 10.1287) = 0.0175$

Press the MATRX key and arrow over to EDIT. Press 1:[A]. Press 2 ENTER 4 ENTER. Enter the table values by row. Press ENTER after each. Press 2nd QUIT. Press STAT and arrow over to TESTS. Arrow down to C:$\chi 2$-TEST. Press ENTER. You should see Observed:[A] and Expected:[B]. Arrow down to Calculate. Press ENTER. The test statistic is 10.1287 and the p-value = 0.0175. Do the procedure a second time but arrow down to Draw instead of calculate.

Compare α and the p-value: Since no α is given, assume $\alpha = 0.05$. p-value = 0.0175. $\alpha >$ p-value.

Make a decision: Since $\alpha >$ p-value, reject H0. This means that the distributions are not the same.

Conclusion: At a 5% level of significance, from the data, there is sufficient evidence to conclude that the distributions of living arrangements for male and female college students are not the same.

Notice that the conclusion is only that the distributions are not the same. We cannot use the test for homogeneity to draw any conclusions about how they differ.

Your Turn!

Do families and singles have the same distribution of cars? Use a level of significance of 0.05. Suppose that 100 randomly selected families and 200 randomly selected singles were asked what type of car they drove: sport, sedan, hatchback, truck, van/SUV. The results are shown in [link]. Do families and singles have the same distribution of cars? Test at a level of significance of 0.05.

	Sport	Sedan	Hatchback	Truck	Van/SUV
Family	5	15	35	17	28
Single	45	65	37	46	7

Solution

With a *p*-value of almost zero, we reject the null hypothesis. The data show that the distribution of cars is not the same for families and singles.



One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=946#oembed-2

Example

Both before and after a recent earthquake, surveys were conducted asking voters which of the three candidates they planned on voting for in the upcoming city council election. Has there been a change since the earthquake? Use a level of significance of 0.05. [link] shows the results of the survey. Has there been a change in the distribution of voter preferences since the earthquake?

	Perez	Chung	Stevens
Before	167	128	135
After	214	197	225

Solution

H0: The distribution of voter preferences was the same before and after the earthquake.

Ha: The distribution of voter preferences was not the same before and after the earthquake.

Degrees of Freedom (df):

df = number of columns – 1 = 3 – 1 = 2

Distribution for the test: χ^2_2

Calculate the test statistic: $\chi^2 = 3.2603$ (calculator or computer)

Probability statement: p-value=$P(\chi^2 > 3.2603) = 0.1959$

Press the MATRX key and arrow over to EDIT. Press 1:[A]. Press 2 ENTER 3 ENTER. Enter the table values by row. Press ENTER after each. Press 2nd QUIT. Press STAT and arrow over to TESTS. Arrow down to C:χ^2-TEST. Press ENTER. You should see Observed:[A] and Expected:[B]. Arrow down to Calculate. Press ENTER. The test statistic is 3.2603 and the p-value = 0.1959. Do the procedure a second time but arrow down to Draw instead of calculate.

Compare α and the p-value:$\alpha = 0.05$ and the p-value = 0.1959. $\alpha <$ p-value.

Make a decision: Since $\alpha <$ p-value, do not reject Ho.

Conclusion: At a 5% level of significance, from the data, there is insufficient evidence to conclude that the distribution of voter preferences was not the same before and after the earthquake.

Your Turn!

Ivy League schools receive many applications, but only some can be accepted. At the schools listed in [link], two types of applications are accepted: regular and early decision.

Application Type Accepted	Brown	Columbia	Cornell	Dartmouth	Penn	Yale
Regular	2,115	1,792	5,306	1,734	2,685	1,245
Early Decision	577	627	1,228	444	1,195	761

We want to know if the number of regular applications accepted follows the same distribution as the number of early applications accepted. State the null and alternative hypotheses, the degrees of freedom and the test statistic, sketch the graph of the *p*-value, and draw a conclusion about the test of homogeneity.

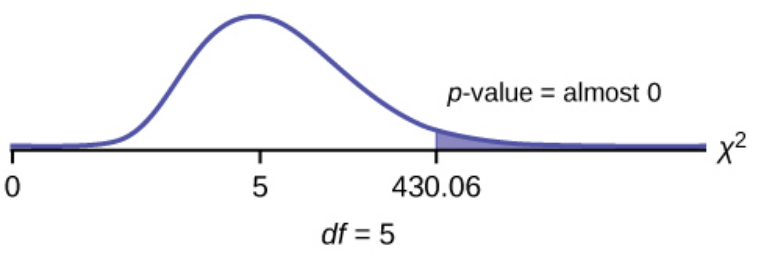
Solution

H_0 : The distribution of regular applications accepted is the same as the distribution of early applications accepted.

H_a : The distribution of regular applications accepted is not the same as the distribution of early applications accepted.

$df = 5$

χ^2 test statistic = 430.06



Press the MATRX key and arrow over to EDIT. Press 1:[A]. Press 3 ENTER 3 ENTER. Enter the table values by row. Press ENTER after each. Press 2nd QUIT. Press STAT and arrow over to TESTS. Arrow down toC:χ2-TEST. Press ENTER. You should see Observed:[A] and Expected:[B]. Arrow down to Calculate. Press ENTER. The test statistic is 430.06 and the *p*-value = 9.80E-91. Do the procedure a second time but arrow down to Draw instead of calculate.

References

Data from the Insurance Institute for Highway Safety, 2013. Available online at www.iihs.org/iihs/ratings (accessed May 24, 2013).

"Energy use (kg of oil equivalent per capita)." The World Bank, 2013. Available online at http://data.worldbank.org/indicator/EG.USE.PCAP.KG.OE/countries (accessed May 24, 2013).

"Parent and Family Involvement Survey of 2007 National Household Education Survey Program (NHES)," U.S. Department of Education, National Center for Education Statistics. Available online at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009030 (accessed May 24, 2013).

"Parent and Family Involvement Survey of 2007 National Household Education Survey Program (NHES)," U.S. Department of Education, National Center for Education Statistics. Available online at http://nces.ed.gov/pubs2009/2009030_sup.pdf (accessed May 24, 2013).

11.5 COMPARISON OF THE CHI-SQUARE TESTS

Learning Objectives

By the end of this section, the student should be able to:

- make comparison of the Chi-Square Tests

You have seen the χ^2 test statistic used in three different circumstances. The following bulleted list is a summary that will help you decide which χ^2 test is the appropriate one to use.

- **Goodness-of-Fit:** Use the goodness-of-fit test to decide whether a population with an unknown distribution "fits" a known distribution. In this case there will be a single qualitative survey question or a single outcome of an experiment from a single population. Goodness-of-Fit is typically used to see if the population is uniform (all outcomes occur with equal frequency), the population is normal, or the population is the same as another population with a known distribution. The null and alternative hypotheses are:

  H_0: The population fits the given distribution.

  H_a: The population does not fit the given distribution.
- **Independence:** Use the test for independence to decide whether two variables (factors) are independent or dependent. In this case there will be two qualitative survey questions or experiments and a contingency table will be constructed. The goal is to see if the two variables are unrelated (independent) or related (dependent). The null and alternative hypotheses are:

  H_0: The two variables (factors) are independent.

H_a: The two variables (factors) are dependent.

- **Homogeneity:** Use the test for homogeneity to decide if two populations with unknown distributions have the same distribution as each other. In this case there will be a single qualitative survey question or experiment given to two different populations. The null and alternative hypotheses are:

  H_0: The two populations follow the same distribution.

  H_a: The two populations have different distributions.



One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=947#oembed-1

11.6 TEST OF A SINGLE VARIANCE

Learning Objectives

By the end of this section, the student should be able to:

- calculate the test of a single variance

A test of a single variance assumes that the underlying distribution is **normal**. The null and alternative hypotheses are stated in terms of the population variance (or population standard deviation). The test statistic is:

$$\frac{(n-1)\,s^2}{\sigma^2}$$

where:

- n = the total number of data
- s^2 = sample variance
- σ^2 = population variance

You may think of s as the random variable in this test. The number of degrees of freedom is $df = n - 1$. **A test of a single variance may be right-tailed, left-tailed, or two-tailed.** [link] will show you how to set up the null and alternative hypotheses. The null and alternative hypotheses contain statements about the population variance.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=950#oembed-1

Example

Math instructors are not only interested in how their students do on exams, on average, but how the exam scores vary. To many instructors, the variance (or standard deviation) may be more important than the average.

Suppose a math instructor believes that the standard deviation for his final exam is five points. One of his best students thinks otherwise. The student claims that the standard deviation is more than five points. If the student were to conduct a hypothesis test, what would the null and alternative hypotheses be?

Solution
Even though we are given the population standard deviation, we can set up the test using the population variance as follows.

- H0: $\sigma^2 = 52$
- Ha: $\sigma^2 > 52$

Your Turn!

A SCUBA instructor wants to record the collective depths each of his students dives during their checkout. He is interested in how the depths vary, even though everyone should have been at the same depth. He believes the standard deviation is three feet. His assistant thinks the standard deviation is less than three feet. If the instructor were to conduct a test, what would the null and alternative hypotheses be?

Solution

H_0: $\sigma^2 = 3^2$

H_a: $\sigma^2 < 3^2$



One or more interactive elements has been excluded from this version of the text. You can view them online here: https://louis.pressbooks.pub/introductorystatistics/?p=950#oembed-2

Example

With individual lines at its various windows, a post office finds that the standard deviation for normally distributed waiting times for customers on Friday afternoon is 7.2 minutes. The post office experiments with a single, main waiting line and finds that for a random sample of 25 customers, the waiting times for customers have a standard deviation of 3.5 minutes.

With a significance level of 5%, test the claim that a single line causes lower variation among waiting times (shorter waiting times) for customers.

Solution

Since the claim is that a single line causes less variation, this is a test of a single variance.
The parameter is the population variance, $\sigma 2$, or the population standard deviation, σ.

Random Variable: The sample standard deviation, s, is the random variable.
Let s = standard deviation for the waiting times.

- H0: σ2 = 7.22
- Ha: σ2 < 7.22

The word "less" tells you this is a left-tailed test.

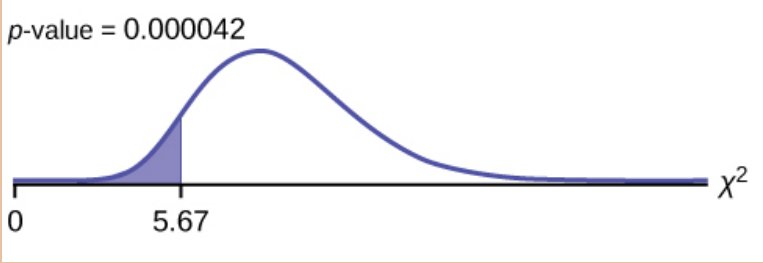
Distribution for the test: χ^2_{24}, where:

n = the number of customers sampled
df = n – 1 = 25 – 1 = 24
Calculate the test statistic:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(25-1)(3.5)^2}{7.2^2} = 5.67$$

where n = 25, s = 3.5, and σ = 7.2.

Graph:



Probability statement: p-value = P (χ2 < 5.67) = 0.000042

Compare α and the p-value:

α = 0.05p-value = 0.000042α > p-value

Make a decision: Since α > p-value, reject H0. This means that you reject σ2 = 7.22. In other words, you do not think the variation in waiting times is 7.2 minutes; you think the variation in waiting times is less.

Conclusion: At a 5% level of significance, from the data, there is sufficient evidence to conclude that a single line causes a lower variation among the waiting times or with a single line, the customer waiting times vary less than 7.2 minutes.

In 2nd DISTR, use 7:χ2cdf. The syntax is (lower, upper, df) for the parameter list.
For [link], χ2cdf(-1E99,5.67,24). The p-value = 0.000042.

Your Turn!

The FCC conducts broadband speed tests to measure how much data per second passes between a consumer's computer and the internet. As of August of 2012, the standard deviation of Internet speeds across Internet Service Providers (ISPs) was 12.2 percent. Suppose a sample of 15 ISPs is taken, and the standard deviation is 13.2. An analyst claims that the standard deviation of speeds is more than what was reported. State the null and alternative hypotheses, compute the degrees of freedom, the test statistic, sketch the graph of the *p*-value, and draw a conclusion. Test at the 1% significance level.

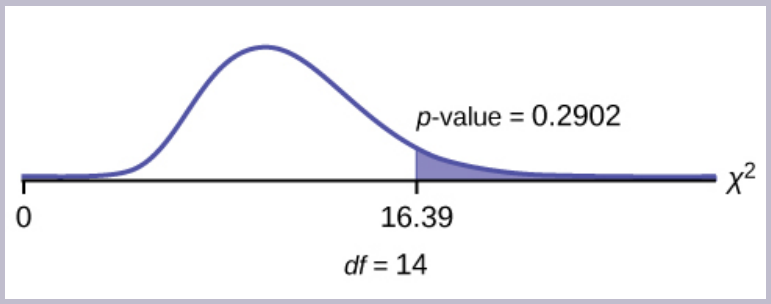
Solution

H_0: $\sigma^2 = 12.2^2$

H_a: $\sigma^2 > 12.2^2$

$df = 14$

chi^2 test statistic = 16.39



The *p*-value is 0.2902, so we decline to reject the null hypothesis. There is not enough evidence to suggest that the variance is greater than 12.2^2.

In `2nd DISTR`, use7:`χ2cdf`. The syntax is `(lower, upper, df)` for the parameter list. `χ2cdf(16.39,10^99,14)`. The *p*-value = 0.2902.

References

"AppleInsider Price Guides." Apple Insider, 2013. Available online at http://appleinsider.com/mac_price_guide (accessed May 14, 2013).

Data from the World Bank, June 5, 2012.

CHAPTER 11 REVIEW

11.1 Review

The chi-square distribution is a useful tool for assessment in a series of problem categories. These problem categories include primarily (i) whether a data set fits a particular distribution, (ii) whether the distributions of two populations are the same, (iii) whether two events might be independent, and (iv) whether there is a different variability than expected within a population.

An important parameter in a chi-square distribution is the degrees of freedom *df* in a given problem. The random variable in the chi-square distribution is the sum of squares of *df* standard normal variables, which must be independent. The key characteristics of the chi-square distribution also depend directly on the degrees of freedom.

The chi-square distribution curve is skewed to the right, and its shape depends on the degrees of freedom *df*. For $df > 90$, the curve approximates the normal distribution. Test statistics based on the chi-square distribution are always greater than or equal to zero. Such application tests are almost always right-tailed tests.

Formula Review

$\chi^2 = (Z_1)^2 + (Z_2)^2 + \ldots (Z_{df})^2$

chi-square distribution random variable

$\mu_{\chi^2} = df$ chi-square distribution population mean

$\sigma_{\chi^2} = \sqrt{2\,(df)}$ Chi-Square distribution population standard deviation

If the number of degrees of freedom for a chi-square distribution is 25, what is the population mean and standard deviation?

Solution

mean = 25 and standard deviation = 7.0711

If $df > 90$, the distribution is _____. If $df = 15$, the distribution is _____.

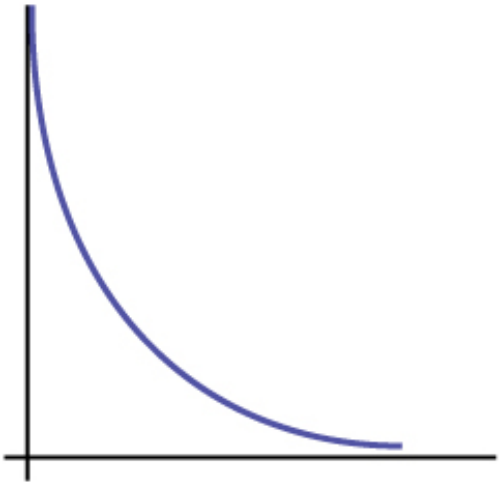
When does the chi-square curve approximate a normal distribution?

Solution

when the number of degrees of freedom is greater than 90

Where is μ located on a chi-square curve?

Is it more likely the df is 90, 20, or two in the graph?



Solution

$df = 2$

11.2 Review

To assess whether a data set fits a specific distribution, you can apply the goodness-of-fit hypothesis test that uses the chi-square distribution. The null hypothesis for this test states that the data come from the assumed distribution. The test compares observed values against the values you would expect to have if your data followed the assumed distribution. The test is almost always right-tailed. Each observation or cell category must have an expected value of at least five.

Formula Review

$$\sum_{k} \frac{(O-E)^2}{E}$$

goodness-of-fit test statistic where:

O: observed values

E: expected values

k: number of different data cells or categories

$df = k - 1$ degrees of freedom

Determine the appropriate test to be used in the next three exercises.

An archeologist is calculating the distribution of the frequency of the number of artifacts she finds in a dig site. Based on previous digs, the archeologist creates an expected distribution broken down by grid sections in the dig site. Once the site has been fully excavated, she compares the actual number of artifacts found in each grid section to see if her expectation was accurate.

An economist is deriving a model to predict outcomes on the stock market. He creates a list of expected points on the stock market index for the next two weeks. At the close of each day's trading, he records the actual points on the index. He wants to see how well his model matched what actually happened.
Solution

a goodness-of-fit test

A personal trainer is putting together a weight-lifting program for her clients. For a 90-day program, she expects each client to lift a specific maximum weight each week. As she goes along, she records the actual maximum weights her clients lifted. She wants to know how well her expectations met with what was observed.

Use the following information to answer the next five exercises: A teacher predicts that the distribution of grades on the final exam will be and they are recorded in [link].

Grade	Proportion
A	0.25
B	0.30
C	0.35
D	0.10

The actual distribution for a class of 20 is in [link].

Grade	Frequency
A	7
B	7
C	5
D	1

$df =$ _____
Solution

3

State the null and alternative hypotheses.

χ^2 test statistic = _____
Solution

2.04

p-value = _____

At the 5% significance level, what can you conclude?
Solution

We decline to reject the null hypothesis. There is not enough evidence to suggest that the observed test scores are significantly different from the expected test scores.

Use the following information to answer the next nine exercises: The following data are real. The cumulative number of AIDS cases reported for Santa Clara County is broken down by ethnicity as in [link].

Ethnicity	Number of Cases
White	2,229
Hispanic	1,157
Black/African-American	457
Asian, Pacific Islander	232
	Total = 4,075

The percentage of each ethnic group in Santa Clara County is as in [link].

Ethnicity	Percentage of total county population	Number expected (round to two decimal places)
White	42.9%	1748.18
Hispanic	26.7%	
Black/African-American	2.6%	
Asian, Pacific Islander	27.8%	
	Total = 100%	

If the ethnicities of AIDS victims followed the ethnicities of the total county population, fill in the expected number of cases per ethnic group.

Perform a goodness-of-fit test to determine whether the occurrence of AIDS cases follows the ethnicities of the general population of Santa Clara County.

H_0: _____
Solution

H_0: the distribution of AIDS cases follows the ethnicities of the general population of Santa Clara County.

H_a: _____

Is this a right-tailed, left-tailed, or two-tailed test?
Solution

right-tailed

degrees of freedom = _____

χ^2 test statistic = _____
Solution

88,621

p-value = _____

Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade in the region corresponding to the *p*-value.



Let $\alpha = 0.05$

Decision: _____

Reason for the Decision: _____

Conclusion (write out in complete sentences): _____

Solution

Graph: Check student's solution.

Decision: Reject the null hypothesis.

Reason for the Decision: p-value < alpha

Conclusion (write out in complete sentences): The make-up of AIDS cases does not fit the ethnicities of the general population of Santa Clara County.

Does it appear that the pattern of AIDS cases in Santa Clara County corresponds to the distribution of ethnic groups in this county? Why or why not?

11.3 Review

To assess whether two factors are independent or not, you can apply the test of independence that uses the chi-square distribution. The null hypothesis for this test states that the two factors are independent. The test compares observed values to expected values. The test is right-tailed. Each observation or cell category must have an expected value of at least 5.

Formula Review

Test of Independence

- The number of degrees of freedom is equal to (number of columns – 1)(number of rows – 1).

- The test statistic is $\sum_{(i \cdot j)} \frac{(O-E)^2}{E}$ where

  O = observed values,

  E = expected values,

  i = the number of rows in the table, and

  j = the number of columns in the table.
- If the null hypothesis is true, the expected number $E = \frac{(\text{row total})(\text{column total})}{\text{total surveyed}}$.

Determine the appropriate test to be used in the next three exercises.

A pharmaceutical company is interested in the relationship between age and presentation of symptoms for a common viral infection. A random sample is taken of 500 people with the infection across different age groups.
Solution

a test of independence

The owner of a baseball team is interested in the relationship between player salaries and team winning percentage. He takes a random sample of 100 players from different organizations.

A marathon runner is interested in the relationship between the brand of shoes runners wear and their run times. She takes a random sample of 50 runners and records their run times as well as the brand of shoes they were wearing.
Solution

a test of independence

Use the following information to answer the next seven exercises: Transit Railroads is interested in the relationship between travel distance and the ticket class purchased. A random sample of 200 passengers is taken. [link] shows the results. The railroad wants to know if a passenger's choice in ticket class is independent of the distance they must travel.

Traveling Distance	Third class	Second class	First class	Total
1–100 miles	21	14	6	41
101–200 miles	18	16	8	42
201–300 miles	16	17	15	48
301–400 miles	12	14	21	47
401–500 miles	6	6	10	22
Total	73	67	60	200

State the hypotheses.

H_0: _____

H_a: _____

df = _____
Solution

8

How many passengers are expected to travel between 201 and 300 miles and purchase second-class tickets?

How many passengers are expected to travel between 401 and 500 miles and purchase first-class tickets?
Solution

6.6

What is the test statistic?

What is the p-value?
Solution

0.0435

What can you conclude at the 5% level of significance?

Use the following information to answer the next eight exercises: An article in the New England Journal of Medicine, discussed a study on smokers in California and Hawaii. In one part of the report, the self-reported ethnicity and smoking levels per day were given. Of the people smoking at most ten cigarettes per day, there were 9,886 African Americans, 2,745 Native Hawaiians, 12,831 Latinos, 8,378 Japanese Americans and 7,650

whites. Of the people smoking 11 to 20 cigarettes per day, there were 6,514 African Americans, 3,062 Native Hawaiians, 4,932 Latinos, 10,680 Japanese Americans, and 9,877 whites. Of the people smoking 21 to 30 cigarettes per day, there were 1,671 African Americans, 1,419 Native Hawaiians, 1,406 Latinos, 4,715 Japanese Americans, and 6,062 whites. Of the people smoking at least 31 cigarettes per day, there were 759 African Americans, 788 Native Hawaiians, 800 Latinos, 2,305 Japanese Americans, and 3,970 whites.

Complete the table.

Smoking Levels by Ethnicity (Observed)

Smoking Level per Day	African American	Native Hawaiian	Latino	Japanese Americans	White	TOTALS
1-10						
11-20						
21-30						
31+						
TOTALS						

Solution

Smoking Level per Day	African American	Native Hawaiian	Latino	Japanese Americans	White	Totals
1-10	9,886	2,745	12,831	8,378	7,650	41,490
11-20	6,514	3,062	4,932	10,680	9,877	35,065
21-30	1,671	1,419	1,406	4,715	6,062	15,273
31+	759	788	800	2,305	3,970	8,622
Totals	18,830	8,014	19,969	26,078	27,559	10,0450

State the hypotheses.

H_0: _____

H_a: _____

Enter expected values in [link]. Round to two decimal places.

Calculate the following values:

Solution

Smoking Level per Day	African American	Native Hawaiian	Latino	Japanese Americans	White
1-10	7777.57	3310.11	8248.02	10771.29	11383.01
11-20	6573.16	2797.52	6970.76	9103.29	9620.27
21-30	2863.02	1218.49	3036.20	3965.05	4190.23
31+	1616.25	687.87	1714.01	2238.37	2365.49

df = _____

χ^2 test statistic = _____
Solution

10,301.8

p-value = _____

Is this a right-tailed, left-tailed, or two-tailed test? Explain why.
Solution

right

Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade in the region corresponding to the p-value.



State the decision and conclusion (in a complete sentence) for the following preconceived levels of α.

α = 0.05

1. Decision: _____
2. Reason for the decision: _____

3. Conclusion (write out in a complete sentence): _____

Solution

1. Reject the null hypothesis.
2. p-value < alpha
3. There is sufficient evidence to conclude that smoking level is dependent on ethnic group.

$\alpha = 0.01$

1. Decision: _____
2. Reason for the decision: _____
3. Conclusion (write out in a complete sentence): _____

11.4 Review

To assess whether two data sets are derived from the same distribution—which need not be known, you can apply the test for homogeneity that uses the chi-square distribution. The null hypothesis for this test states that the populations of the two data sets come from the same distribution. The test compares the observed values against the expected values if the two populations followed the same distribution. The test is right-tailed. Each observation or cell category must have an expected value of at least five.

Formula Review

$\sum_{i \cdot j} \frac{(O-E)^2}{E}$ Homogeneity test statistic where: O = observed values

E = expected values

i = number of rows in data contingency table

j = number of columns in data contingency table

$df = (i-1)(j-1)$ Degrees of freedom

A math teacher wants to see if two of her classes have the same distribution of test scores. What test should she use?

Solution

test for homogeneity

What are the null and alternative hypotheses for [link]?

A market researcher wants to see if two different stores have the same distribution of sales throughout the year. What type of test should he use?
Solution

test for homogeneity

A meteorologist wants to know if East and West Australia have the same distribution of storms. What type of test should she use?

What condition must be met to use the test for homogeneity?
Solution

All values in the table must be greater than or equal to five.

Use the following information to answer the next five exercises: Do private practice doctors and hospital doctors have the same distribution of working hours? Suppose that a sample of 100 private practice doctors and 150 hospital doctors are selected at random and asked about the number of hours a week they work. The results are shown in [link].

	20–30	**30–40**	**40–50**	**50–60**
Private Practice	16	40	38	6
Hospital	8	44	59	39

State the null and alternative hypotheses.

df = _____
Solution

3

What is the test statistic?

What is the p-value?
Solution

0.00005

What can you conclude at the 5% significance level?

11.5 Review

The goodness-of-fit test is typically used to determine if data fits a particular distribution. The test of independence makes use of a contingency table to determine the independence of two factors. The test for homogeneity determines whether two populations come from the same distribution, even if this distribution is unknown.

Which test do you use to decide whether an observed distribution is the same as an expected distribution?
Solution

a goodness-of-fit test

What is the null hypothesis for the type of test from [link]?

Which test would you use to decide whether two factors have a relationship?
Solution

a test for independence

Which test would you use to decide if two populations have the same distribution?

How are tests of independence similar to tests for homogeneity?
Solution

Answers will vary. Sample answer: Tests of independence and tests for homogeneity both calculate the test statistic the same way $\sum_{(ij)} \frac{(O-E)^2}{E}$. In addition, all values must be greater than or equal to five.

How are tests of independence different from tests for homogeneity?

11.6 Review

To test variability, use the chi-square test of a single variance. The test may be left-, right-, or two-tailed, and its hypotheses are always expressed in terms of the variance (or standard deviation).

Formula Review

$\chi^2 = \frac{(n-1) \cdot s^2}{\sigma^2}$ Test of a single variance statistic where:

n: sample size

s: sample standard deviation

σ: population standard deviation

$df = n - 1$ Degrees of freedom

Test of a Single Variance

- Use the test to determine variation.
- The degrees of freedom is the number of samples – 1.
- The test statistic is $\frac{(n-1) \cdot s^2}{\sigma^2}$, where n = the total number of data, s^2 = sample variance, and σ^2 = population variance.
- The test may be left-, right-, or two-tailed.

Use the following information to answer the next three exercises: An archer's standard deviation for his hits is six (data is measured in distance from the center of the target). An observer claims the standard deviation is less.

What type of test should be used?

Solution

a test of a single variance

State the null and alternative hypotheses.

Is this a right-tailed, left-tailed, or two-tailed test?

Solution

a left-tailed test

Use the following information to answer the next three exercises: The standard deviation of heights for students in a school is 0.81. A random sample of 50 students is taken, and the standard deviation of heights of the sample is 0.96. A researcher in charge of the study believes the standard deviation of heights for the school is greater than 0.81.

What type of test should be used?

State the null and alternative hypotheses.
Solution

H_0: $\sigma^2 = 0.81^2$;
H_a: $\sigma^2 > 0.81^2$

df = _____

Use the following information to answer the next four exercises: The average waiting time in a doctor's office varies. The standard deviation of waiting times in a doctor's office is 3.4 minutes. A random sample of 30 patients in the doctor's office has a standard deviation of waiting times of 4.1 minutes. One doctor believes the variance of waiting times is greater than originally thought.

What type of test should be used?
Solution

a test of a single variance

What is the test statistic?

What is the p-value?
Solution

0.0542

What can you conclude at the 5% significance level?

CHAPTER 11 PRACTICE

CHAPTER 11 HOMEWORK

11.1 Homework

Decide whether the following statements are true or false.

As the number of degrees of freedom increases, the graph of the chi-square distribution looks more and more symmetrical.
Solution

true

The standard deviation of the chi-square distribution is twice the mean.

The mean and the median of the chi-square distribution are the same if $df = 24$.
Solution

false

11.2 Homework

For each problem, use a solution sheet to solve the hypothesis test problem. Go to [link] for the chi-square solution sheet. Round expected frequency to two decimal places.

A six-sided die is rolled 120 times. Fill in the expected frequency column. Then, conduct a hypothesis test to determine if the die is fair. The data in [link] are the result of the 120 rolls.

Face Value	Frequency	Expected Frequency
1	15	
2	29	
3	16	
4	15	
5	30	
6	15	

The marital status distribution of the U.S. male population, ages 15 and older, is as shown in [link].

Marital Status	Percent	Expected Frequency
never married	31.3	
married	56.1	
widowed	2.5	
divorced/separated	10.1	

Suppose that a random sample of 400 U.S. young adult males, 18 to 24 years old, yielded the following frequency distribution. We are interested in whether this age group of males fits the distribution of the U.S. adult population. Calculate the frequency one would expect when surveying 400 people. Fill in [link], rounding to two decimal places.

Marital Status	Frequency
never married	140
married	238
widowed	2
divorced/separated	20

Solution

Marital Status	Percent	Expected Frequency
never married	31.3	125.2
married	56.1	224.4
widowed	2.5	10
divorced/separated	10.1	40.4

1. The data fits the distribution.
2. The data does not fit the distribution.
3. 3
4. chi-square distribution with $df = 3$
5. 19.27
6. 0.0002
7. Check student's solution.

1. Alpha = 0.05
2. Decision: Reject null
3. Reason for decision: *p*-value < alpha
4. Conclusion: Data does not fit the distribution.

Use the following information to answer the next two exercises: The columns in [link] contain the Race/Ethnicity of U.S. Public Schools for a recent year, the percentages for the Advanced Placement Examinee Population for that class, and the Overall Student Population. Suppose the right column contains the result of a survey of 1,000 local students from that year who took an AP Exam.

Race/Ethnicity	**AP Examinee Population**	**Overall Student Population**	**Survey Frequency**
Asian, Asian American, or Pacific Islander	10.2%	5.4%	113
Black or African American	8.2%	14.5%	94
Hispanic or Latino	15.5%	15.9%	136
American Indian or Alaska Native	0.6%	1.2%	10
White	59.4%	61.6%	604
Not reported/other	6.1%	1.4%	43

Perform a goodness-of-fit test to determine whether the local results follow the distribution of the U.S. overall student population based on ethnicity.

Perform a goodness-of-fit test to determine whether the local results follow the distribution of the U.S. AP examinee population, based on ethnicity.
Solution

1. H_0: The local results follow the distribution of the U.S. AP examinee population
2. H_a: The local results do not follow the distribution of the U.S. AP examinee population
3. $df = 5$
4. chi-square distribution with $df = 5$
5. chi-square test statistic = 13.4
6. *p*-value = 0.0199
7. Check student's solution.
8. Alpha = 0.05
   Decision: Reject null when $a = 0.05$
   Reason for Decision: *p*-value < alpha

Conclusion: Local data do not fit the AP Examinee Distribution.
Decision: Do not reject null when $a = 0.01$
Conclusion: There is insufficient evidence to conclude that local data do not follow the distribution of the U.S. AP examinee distribution.

The City of South Lake Tahoe, CA, has an Asian population of 1,419 people, out of a total population of 23,609. Suppose that a survey of 1,419 self-reported Asians in the Manhattan, NY, area yielded the data in [link]. Conduct a goodness-of-fit test to determine if the self-reported sub-groups of Asians in the Manhattan area fit that of the Lake Tahoe area.

Race	**Lake Tahoe Frequency**	**Manhattan Frequency**
Asian Indian	131	174
Chinese	118	557
Filipino	1,045	518
Japanese	80	54
Korean	12	29
Vietnamese	9	21
Other	24	66

Use the following information to answer the next two exercises: UCLA conducted a survey of more than 263,000 college freshmen from 385 colleges in fall 2005. The results of students' expected majors by gender were reported in THE CHRONICLE OF HIGHER EDUCATION (2/2/2006). Suppose a survey of 5,000 graduating females and 5,000 graduating males was done as a follow-up last year to determine what their actual majors were. The results are shown in the tables for [link] and [link]. The second column in each table does not add to 100% because of rounding.

Conduct a goodness-of-fit test to determine if the actual college majors of graduating females fit the distribution of their expected majors.

Major	Women – Expected Major	Women – Actual Major
Arts & Humanities	14.0%	670
Biological Sciences	8.4%	410
Business	13.1%	685
Education	13.0%	650
Engineering	2.6%	145
Physical Sciences	2.6%	125
Professional	18.9%	975
Social Sciences	13.0%	605
Technical	0.4%	15
Other	5.8%	300
Undecided	8.0%	420

Solution

1. H_0: The actual college majors of graduating females fit the distribution of their expected majors
2. H_a: The actual college majors of graduating females do not fit the distribution of their expected majors
3. $df = 10$
4. chi-square distribution with $df = 10$
5. test statistic = 11.48
6. p-value = 0.3211
7. Check student's solution.
    1. Alpha = 0.05
    2. Decision: Do not reject null when $a = 0.05$ and $a = 0.01$
    3. Reason for decision: p-value > alpha
    4. Conclusion: There is insufficient evidence to conclude that the distribution of actual college majors of graduating females fits the distribution of their expected majors.

Conduct a goodness-of-fit test to determine if the actual college majors of graduating males fit the distribution of their expected majors.

Major	Men – Expected Major	Men – Actual Major
Arts & Humanities	11.0%	600
Biological Sciences	6.7%	330
Business	22.7%	1130
Education	5.8%	305
Engineering	15.6%	800
Physical Sciences	3.6%	175
Professional	9.3%	460
Social Sciences	7.6%	370
Technical	1.8%	90
Other	8.2%	400
Undecided	6.6%	340

Read the statement and decide whether it is true or false.

In a goodness-of-fit test, the expected values are the values we would expect if the null hypothesis were true.
Solution

true

In general, if the observed values and expected values of a goodness-of-fit test are not close together, then the test statistic can get very large and on a graph will be way out in the right tail.

Use a goodness-of-fit test to determine if high school principals believe that students are absent equally during the week or not.
Solution

true

The test to use to determine if a six-sided die is fair is a goodness-of-fit test.

In a goodness-of fit test, if the *p*-value is 0.0113, in general, do not reject the null hypothesis.
Solution

false

A sample of 212 commercial businesses was surveyed for recycling one commodity; a commodity here means

any one type of recyclable material such as plastic or aluminum. [link] shows the business categories in the survey, the sample size of each category, and the number of businesses in each category that recycle one commodity. Based on the study, on average half of the businesses were expected to be recycling one commodity. As a result, the last column shows the expected number of businesses in each category that recycle one commodity. At the 5% significance level, perform a hypothesis test to determine if the observed number of businesses that recycle one commodity follows the uniform distribution of the expected values.

Business type	Number in class	Observed number that recycle one commodity	Expected number that recycle one commodity
Office	35	19	17.5
Retail/Wholesale	48	27	24
Food/Restaurants	53	35	26.5
Manufacturing/ Medical	52	21	26
Hotel/Mixed	24	9	12

[link] contains information from a survey among 499 participants classified according to their age groups. The second column shows the percentage of obese people per age class among the study participants. The last column comes from a different study at the national level that shows the corresponding percentages of obese people in the same age classes in the USA. Perform a hypothesis test at the 5% significance level to determine whether the survey participants are a representative sample of the USA obese population.

Age Class (Years)	Obese (Percentage)	Expected USA average (Percentage)
20–30	75.0	32.6
31–40	26.5	32.6
41–50	13.6	36.6
51–60	21.9	36.6
61–70	21.0	39.7

Solution

1. H_0: Surveyed obese fit the distribution of expected obese
2. H_a: Surveyed obese do not fit the distribution of expected obese
3. $df = 4$
4. chi-square distribution with $df = 4$
5. test statistic = 54.01

6. p-value = 0
7. Check student's solution.
    1. Alpha: 0.05
    2. Decision: Reject the null hypothesis.
    3. Reason for decision: p-value < alpha
    4. Conclusion: At the 5% level of significance, from the data, there is sufficient evidence to conclude that the surveyed obese do not fit the distribution of expected obese.

11.3 Homework

For each problem, use a solution sheet to solve the hypothesis test problem. Go to Appendix E for the chi-square solution sheet. Round expected frequency to two decimal places.

A recent debate about where in the United States skiers believe the skiing is best prompted the following survey. Test to see if the best ski area is independent of the level of the skier.

U.S. Ski Area	Beginner	Intermediate	Advanced
Tahoe	20	30	40
Utah	10	30	60
Colorado	10	40	50

Car manufacturers are interested in whether there is a relationship between the size of car an individual drives and the number of people in the driver's family (that is, whether car size and family size are independent). To test this, suppose that 800 car owners were randomly surveyed with the results in [link]. Conduct a test of independence.

Family Size	Sub & Compact	Mid-size	Full-size	Van & Truck
1	20	35	40	35
2	20	50	70	80
3–4	20	50	100	90
5+	20	30	70	70

Solution

1. H_0: Car size is independent of family size.

2. H_a: Car size is dependent on family size.
3. $df = 9$
4. chi-square distribution with $df = 9$
5. test statistic = 15.8284
6. p-value = 0.0706
7. Check student's solution.
    1. Alpha: 0.05
    2. Decision: Do not reject the null hypothesis.
    3. Reason for decision: p-value > alpha
    4. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that car size and family size are dependent.

College students may be interested in whether or not their majors have any effect on starting salaries after graduation. Suppose that 300 recent graduates were surveyed as to their majors in college and their starting salaries after graduation. [link] shows the data. Conduct a test of independence.

Major	**< $ 50,000**	**$ 50,000 – $ 68,999**	**$ 69,000 +**
English	5	20	5
Engineering	10	30	60
Nursing	10	15	15
Business	10	20	30
Psychology	20	30	20

Some travel agents claim that honeymoon hot spots vary according to age of the bride. Suppose that 280 recent brides were interviewed as to where they spent their honeymoons. The information is given in [link]. Conduct a test of independence.

Location	**20–29**	**30–39**	**40–49**	**50 and over**
Niagara Falls	15	25	25	20
Poconos	15	25	25	10
Europe	10	25	15	5
Virgin Islands	20	25	15	5

Solution

1. H_0: Honeymoon locations are independent of bride's age.
2. H_a: Honeymoon locations are dependent on bride's age.
3. $df = 9$
4. chi-square distribution with $df = 9$
5. test statistic = 15.7027
6. p-value = 0.0734
7. Check student's solution.
    1. Alpha: 0.05
    2. Decision: Do not reject the null hypothesis.
    3. Reason for decision: p-value > alpha
    4. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that honeymoon location and bride age are dependent.

A manager of a sports club keeps information concerning the main sport in which members participate and their ages. To test whether there is a relationship between the age of a member and his or her choice of sport, 643 members of the sports club are randomly selected. Conduct a test of independence.

Sport	**18 – 25**	**26 – 30**	**31 – 40**	**41 and over**
racquetball	42	58	30	46
tennis	58	76	38	65
swimming	72	60	65	33

A major food manufacturer is concerned that the sales for its skinny french fries have been decreasing. As a part of a feasibility study, the company conducts research into the types of fries sold across the country to determine if the type of fries sold is independent of the area of the country. The results of the study are shown in [link]. Conduct a test of independence.

Type of Fries	**Northeast**	**South**	**Central**	**West**
skinny fries	70	50	20	25
curly fries	100	60	15	30
steak fries	20	40	10	10

Solution

1. H_0: The types of fries sold are independent of the location.
2. H_a: The types of fries sold are dependent on the location.

3. $df = 6$
4. chi-square distribution with $df = 6$
5. test statistic =18.8369
6. p-value = 0.0044
7. Check student's solution.
    1. Alpha: 0.05
    2. Decision: Reject the null hypothesis.
    3. Reason for decision: p-value < alpha
    4. Conclusion: At the 5% significance level, There is sufficient evidence that types of fries and location are dependent.

According to Dan Lenard, an independent insurance agent in the Buffalo, N.Y. area, the following is a breakdown of the amount of life insurance purchased by males in the following age groups. He is interested in whether the age of the male and the amount of life insurance purchased are independent events. Conduct a test for independence.

Age of Males	None	< $ 200,000	$ 200,000– $ 400,000	$ 401,001– $ 1,000,000	$ 1,000,001+
20–29	40	15	40	0	5
30–39	35	5	20	20	10
40–49	20	0	30	0	30
50+	40	30	15	15	10

Suppose that 600 thirty-year-olds were surveyed to determine whether or not there is a relationship between the level of education an individual has and salary. Conduct a test of independence.

Annual Salary	Not a high school graduate	High school graduate	College graduate	Masters or doctorate
< $ 30,000	15	25	10	5
$ 30,000– $ 40,000	20	40	70	30
$ 40,000– $ 50,000	10	20	40	55
$ 50,000– $ 60,000	5	10	20	60
$ 60,000+	0	5	10	150

Solution

1. H_0: Salary is independent of level of education.

2. H_a: Salary is dependent on level of education.
3. $df = 12$
4. chi-square distribution with $df = 12$
5. test statistic = 255.7704
6. p-value = 0
7. Check student's solution.

8. Alpha: 0.05
   Decision: Reject the null hypothesis.
   Reason for decision: p-value < alpha
   Conclusion: At the 5% significance level, there is sufficient evidence to conclude that salary and level of education are dependent.

Read the statement and decide whether it is true or false.

The number of degrees of freedom for a test of independence is equal to the sample size minus one.

The test for independence uses tables of observed and expected data values.
Solution

true

The test to use when determining if the college or university a student chooses to attend is related to his or her socioeconomic status is a test for independence.

In a test of independence, the expected number is equal to the row total multiplied by the column total divided by the total surveyed.
Solution

true

An ice cream maker performs a nationwide survey about favorite flavors of ice cream in different geographic areas of the U.S. Based on [link], do the numbers suggest that geographic location is independent of favorite ice cream flavors? Test at the 5% significance level.

U.S. region/ Flavor	Strawberry	Chocolate	Vanilla	Rocky Road	Mint Chocolate Chip	Pistachio	Row total
West	12	21	22	19	15	8	97
Midwest	10	32	22	11	15	6	96
East	8	31	27	8	15	7	96
South	15	28	30	8	15	6	102
Column Total	45	112	101	46	60	27	391

[link] provides a recent survey of the youngest online entrepreneurs whose net worth is estimated at one million dollars or more. Their ages range from 17 to 30. Each cell in the table illustrates the number of entrepreneurs who correspond to the specific age group and their net worth. Are the ages and net worth independent? Perform a test of independence at the 5% significance level.

Age Group\ Net Worth Value (in millions of US dollars)	1–5	6–24	≥25	Row Total
17–25	8	7	5	20
26–30	6	5	9	20
Column Total	14	12	14	40

Solution

1. H_0: Age is independent of the youngest online entrepreneurs' net worth.
2. H_a: Age is dependent on the net worth of the youngest online entrepreneurs.
3. $df = 2$
4. chi-square distribution with $df = 2$
5. test statistic = 1.76
6. p-value 0.4144
7. Check student's solution.
    1. Alpha: 0.05
    2. Decision: Do not reject the null hypothesis.
    3. Reason for decision: p-value > alpha
    4. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that age and net worth for the youngest online entrepreneurs are dependent.

A 2013 poll in California surveyed people about taxing sugar-sweetened beverages. The results are presented in [link], and are classified by ethnic group and response type. Are the poll responses independent of the participants' ethnic group? Conduct a test of independence at the 5% significance level.

Opinion/Ethnicity	Asian-American	White/Non-Hispanic	African-American	Latino	Row Total
Against tax	48	433	41	160	628
In Favor of tax	54	234	24	147	459
No opinion	16	43	16	19	84
Column Total	118	710	71	272	1171

11.4 Homework

For each word problem, use a solution sheet to solve the hypothesis test problem. Go to [link] for the chi-square solution sheet. Round expected frequency to two decimal places.

A psychologist is interested in testing whether there is a difference in the distribution of personality types for business majors and social science majors. The results of the study are shown in [link]. Conduct a test of homogeneity. Test at a 5% level of significance.

	Open	Conscientious	Extrovert	Agreeable	Neurotic
Business	41	52	46	61	58
Social Science	72	75	63	80	65

Solution

1. H_0: The distribution for personality types is the same for both majors
2. H_a: The distribution for personality types is not the same for both majors
3. $df = 4$
4. chi-square with $df = 4$
5. test statistic = 3.01
6. p-value = 0.5568
7. Check student's solution.
8.
    1. Alpha: 0.05
    2. Decision: Do not reject the null hypothesis.
    3. Reason for decision: p-value > alpha
    4. Conclusion: There is insufficient evidence to conclude that the distribution of personality types is different for business and social science majors.

Do men and women select different breakfasts? The breakfasts ordered by randomly selected men and women at a popular breakfast place is shown in [link]. Conduct a test for homogeneity at a 5% level of significance.

	French Toast	Pancakes	Waffles	Omelets
Men	47	35	28	53
Women	65	59	55	60

A fisherman is interested in whether the distribution of fish caught in Green Valley Lake is the same as the distribution of fish caught in Echo Lake. Of the 191 randomly selected fish caught in Green Valley Lake, 105 were rainbow trout, 27 were other trout, 35 were bass, and 24 were catfish. Of the 293 randomly selected fish caught in Echo Lake, 115 were rainbow trout, 58 were other trout, 67 were bass, and 53 were catfish. Perform a test for homogeneity at a 5% level of significance.
Solution

1. H_0: The distribution for fish caught is the same in Green Valley Lake and in Echo Lake.
2. H_a: The distribution for fish caught is not the same in Green Valley Lake and in Echo Lake.
3. 3
4. chi-square with $df = 3$
5. 11.75
6. p-value = 0.0083
7. Check student's solution.
    1. Alpha: 0.05
    2. Decision: Reject the null hypothesis.
    3. Reason for decision: p-value < alpha
    4. Conclusion: There is evidence to conclude that the distribution of fish caught is different in Green Valley Lake and in Echo Lake

In 2007, the United States had 1.5 million homeschooled students, according to the U.S. National Center for Education Statistics. In [link] you can see that parents decide to homeschool their children for different reasons, and some reasons are ranked by parents as more important than others. According to the survey results shown in the table, is the distribution of applicable reasons the same as the distribution of the most important reason? Provide your assessment at the 5% significance level. Did you expect the result you obtained?

Reasons for Homeschooling	Applicable Reason (in thousands of respondents)	Most Important Reason (in thousands of respondents)	Row Total
Concern about the environment of other schools	1,321	309	1,630
Dissatisfaction with academic instruction at other schools	1,096	258	1,354
To provide religious or moral instruction	1,257	540	1,797
Child has special needs, other than physical or mental	315	55	370
Nontraditional approach to child's education	984	99	1,083
Other reasons (e.g., finances, travel, family time, etc.)	485	216	701
Column Total	5,458	1,477	6,935

When looking at energy consumption, we are often interested in detecting trends over time and how they correlate among different countries. The information in [link] shows the average energy use (in units of kg of oil equivalent per capita) in the USA and the joint European Union countries (EU) for the six-year period 2005 to 2010. Do the energy use values in these two areas come from the same distribution? Perform the analysis at the 5% significance level.

Year	European Union	United States	Row Total
2010	3,413	7,164	10,557
2009	3,302	7,057	10,359
2008	3,505	7,488	10,993
2007	3,537	7,758	11,295
2006	3,595	7,697	11,292
2005	3,613	7,847	11,460
Column Total	45,011	20,965	65,976

Solution

1. H_0: The distribution of average energy use in the USA is the same as in Europe between 2005 and 2010.
2. H_a: The distribution of average energy use in the USA is not the same as in Europe between 2005 and 2010.
3. $df = 4$

4. chi-square with $df = 4$
5. test statistic = 2.7434
6. p-value = 0.7395
7. Check student's solution.
    1. Alpha: 0.05
    2. Decision: Do not reject the null hypothesis.
    3. Reason for decision: p-value > alpha
    4. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that the average energy use values in the US and EU are not derived from different distributions for the period from 2005 to 2010.

The Insurance Institute for Highway Safety collects safety information about all types of cars every year, and publishes a report of Top Safety Picks among all cars, makes, and models. [link] presents the number of Top Safety Picks in six car categories for the two years 2009 and 2013. Analyze the table data to conclude whether the distribution of cars that earned the Top Safety Picks safety award has remained the same between 2009 and 2013. Derive your results at the 5% significance level.

Year \ Car Type	Small	Mid-Size	Large	Small SUV	Mid-Size SUV	Large SUV	Row Total
2009	12	22	10	10	27	6	87
2013	31	30	19	11	29	4	124
Column Total	43	52	29	21	56	10	211

11.5 Homework

For each word problem, use a solution sheet to solve the hypothesis test problem. Go to [link] for the chi-square solution sheet. Round expected frequency to two decimal places.

Is there a difference between the distribution of community college statistics students and the distribution of university statistics students in what technology they use on their homework? Of some randomly selected community college students, 43 used a computer, 102 used a calculator with built in statistics functions, and 65 used a table from the textbook. Of some randomly selected university students, 28 used a computer, 33 used a calculator with built in statistics functions, and 40 used a table from the textbook. Conduct an appropriate hypothesis test using a 0.05 level of significance.
Solution

1. H_0: The distribution for technology use is the same for community college students and university

students.
2. H_a: The distribution for technology use is not the same for community college students and university students.
3. 2
4. chi-square with $df = 2$
5. 7.05
6. p-value = 0.0294
7. Check student's solution.
    1. Alpha: 0.05
    2. Decision: Reject the null hypothesis.
    3. Reason for decision: p-value < alpha
    4. Conclusion: There is sufficient evidence to conclude that the distribution of technology used for statistics homework is not the same for statistics students at community colleges and at universities.

Read the statement and decide whether it is true or false.

If $df = 2$, the chi-square distribution has a shape that reminds us of the exponential.

Bringing It Together

1. Explain why a goodness-of-fit test and a test of independence are generally right-tailed tests.
2. If you did a left-tailed test, what would you be testing?

Solution

1. The test statistic is always positive and if the expected and observed values are not close together, the test statistic is large and the null hypothesis will be rejected.
2. Testing to see if the data fits the distribution "too well" or is too perfect.

11.6 Homework

Use the following information to answer the next twelve exercises: Suppose an airline claims that its flights are consistently on time with an average delay of at most 15 minutes. It claims that the average delay is so consistent that the variance is no more than 150 minutes. Doubting the consistency part of the claim, a disgruntled traveler calculates the delays for his next 25 flights. The average delay for those 25 flights is 22 minutes with a standard deviation of 15 minutes.

Is the traveler disputing the claim about the average or about the variance?

A sample standard deviation of 15 minutes is the same as a sample variance of _____ minutes.
Solution

225

Is this a right-tailed, left-tailed, or two-tailed test?

H_0: _____
Solution

H_0: $\sigma^2 \leq 150$

df = _____

chi-square test statistic = _____
Solution

36

p-value = _____

Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade the p-value.
Solution

Check student's solution.

Let $\alpha = 0.05$

Decision: _____

Conclusion (write out in a complete sentence.): _____

How did you know to test the variance instead of the mean?
Solution

The claim is that the variance is no more than 150 minutes.

If an additional test were done on the claim of the average delay, which distribution would you use?

If an additional test were done on the claim of the average delay, but 45 flights were surveyed, which distribution would you use?
Solution

a Student's *t*– or normal distribution

For each word problem, use a solution sheet to solve the hypothesis test problem. Go to [link] for the chi-square solution sheet. Round expected frequency to two decimal places.

A plant manager is concerned her equipment may need recalibrating. It seems that the actual weight of the 15 oz. cereal boxes it fills has been fluctuating. The standard deviation should be at most 0.5 oz. In order to determine if the machine needs to be recalibrated, 84 randomly selected boxes of cereal from the next day's production were weighed. The standard deviation of the 84 boxes was 0.54. Does the machine need to be recalibrated?

Consumers may be interested in whether the cost of a particular calculator varies from store to store. Based on surveying 43 stores, which yielded a sample mean of $ 84 and a sample standard deviation of $ 12, test the claim that the standard deviation is greater than $ 15.
Solution

1. $H_0: \sigma = 15$
2. $H_a: \sigma > 15$
3. $df = 42$
4. chi-square with $df = 42$
5. test statistic = 26.88
6. p-value = 0.9663
7. Check student's solution.
    1. Alpha = 0.05
    2. Decision: Do not reject null hypothesis.
    3. Reason for decision: p-value > alpha
    4. Conclusion: There is insufficient evidence to conclude that the standard deviation is greater than 15.

Isabella, an accomplished **Bay to Breakers** runner, claims that the standard deviation for her time to run the 7.5 mile race is at most three minutes. To test her claim, Rupinder looks up five of her race times. They are 55 minutes, 61 minutes, 58 minutes, 63 minutes, and 57 minutes.

Airline companies are interested in the consistency of the number of babies on each flight, so that they have adequate safety equipment. They are also interested in the variation of the number of babies. Suppose that

an airline executive believes the average number of babies on flights is six with a variance of nine at most. The airline conducts a survey. The results of the 18 flights surveyed give a sample average of 6.4 with a sample standard deviation of 3.9. Conduct a hypothesis test of the airline executive's belief.
Solution

1. $H_0: \sigma \leq 3$
2. $H_a: \sigma > 3$
3. $df = 17$
4. chi-square distribution with $df = 17$
5. test statistic = 28.73
6. p-value = 0.0371
7. Check student's solution.
    1. Alpha: 0.05
    2. Decision: Reject the null hypothesis.
    3. Reason for decision: p-value < alpha
    4. Conclusion: There is sufficient evidence to conclude that the standard deviation is greater than three.

The number of births per woman in China is 1.6 down from 5.91 in 1966. This fertility rate has been attributed to the law passed in 1979 restricting births to one per woman. Suppose that a group of students studied whether or not the standard deviation of births per woman was greater than 0.75. They asked 50 women across China the number of births they had had. The results are shown in [link]. Does the students' survey indicate that the standard deviation is greater than 0.75?

# of births	Frequency
0	5
1	30
2	10
3	5

According to an avid aquarist, the average number of fish in a 20-gallon tank is 10, with a standard deviation of two. His friend, also an aquarist, does not believe that the standard deviation is two. She counts the number of fish in 15 other 20-gallon tanks. Based on the results that follow, do you think that the standard deviation is different from two? Data: 11; 10; 9; 10; 10; 11; 11; 10; 12; 9; 7; 9; 11; 10; 11
Solution

1. H_0: $\sigma = 2$
2. H_a: $\sigma \neq 2$
3. $df = 14$
4. chi-square distribution with $df = 14$
5. chi-square test statistic = 5.2094
6. p-value = 0.0346
7. Check student's solution.
    1. Alpha = 0.05
    2. Decision: Reject the null hypothesis
    3. Reason for decision: p-value < alpha
    4. Conclusion: There is sufficient evidence to conclude that the standard deviation is different than 2.

The manager of "Frenchies" is concerned that patrons are not consistently receiving the same amount of French fries with each order. The chef claims that the standard deviation for a ten-ounce order of fries is at most 1.5 oz., but the manager thinks that it may be higher. He randomly weighs 49 orders of fries, which yields a mean of 11 oz. and a standard deviation of two oz.

You want to buy a specific computer. A sales representative of the manufacturer claims that retail stores sell this computer at an average price of $ 1,249 with a very narrow standard deviation of $ 25. You find a website that has a price comparison for the same computer at a series of stores as follows: $ 1,299; $ 1,229.99; $ 1,193.08; $ 1,279; $ 1,224.95; $ 1,229.99; $ 1,269.95; $ 1,249. Can you argue that pricing has a larger standard deviation than claimed by the manufacturer? Use the 5% significance level. As a potential buyer, what would be the practical conclusion from your analysis?
Solution

The sample standard deviation is $ 34.29.
$H_0 : \sigma^2 = 25^2$

$H_a : \sigma^2 > 25^2$

$df = n - 1 = 7$.

test statistic: {x}^{2}= {x}_{7}^{2}=\frac{\left(n−1\right){s}^{2}}{{25}^{2}}=\frac{\left(8−1\right){\left(34.29\right)}^{2}}{{25}^{2}}=13.169
;

p-value: $P\left(x_7^2 > 13.169\right) = 1 - P\left(x_7^2 \leq 13.169\right) = 0.0681$

Alpha: 0.05

Decision: Do not reject the null hypothesis.

Reason for decision: *p*-value > alpha

Conclusion: At the 5% level, there is insufficient evidence to conclude that the variance is more than 625.

A company packages apples by weight. One of the weight grades is Class A apples. Class A apples have a mean weight of 150 g, and there is a maximum allowed weight tolerance of 5% above or below the mean for apples in the same consumer package. A batch of apples is selected to be included in a Class A apple package. Given the following apple weights of the batch, does the fruit comply with the Class A grade weight tolerance requirements. Conduct an appropriate hypothesis test.

(a) at the 5% significance level

(b) at the 1% significance level

Weights in selected apple batch (in grams): 158; 167; 149; 169; 164; 139; 154; 150; 157; 171; 152; 161; 141; 166; 172;

CLASS GROUP ACTIVITIES/ PROJECTS

CHAPTER 1 PROJECT I: DATA COLLECTION EXPERIMENT

Data Collection Experiment

Class Time:

Names:

Student Learning Outcomes

- The student will demonstrate the systematic sampling technique.
- The student will construct relative frequency tables.
- The student will interpret results and their differences from different data groupings.

Movie Survey

Ask five classmates from a different class how many movies they saw at the theater last month. Do not include rented movies.

1. Record the data.
2. In class, randomly pick one person. On the class list, mark that person's name. Move down four names on the class list. Mark that person's name. Continue doing this until you have marked 12 names. You may need to go back to the start of the list. For each marked name record the five data values. You now have a total of 60 data values.
3. For each name marked, record the data.

Table 1.17: Table for Data

___	___	___	___	___	___	___	___	___	___	___	___
___	___	___	___	___	___	___	___	___	___	___	___
___	___	___	___	___	___	___	___	___	___	___	___
___	___	___	___	___	___	___	___	___	___	___	___
___	___	___	___	___	___	___	___	___	___	___	___

Order the Data

Complete the two relative frequency tables below using your class data.

Table 1.18: Frequency of Number of Movies Viewed

Number of Movies	Frequency	Relative Frequency	Cumulative Relative Frequency
0			
1			
2			
3			
4			
5			
6			
7+			

Table 1.19: Frequency of Number of Movies Viewed

Number of Movies	Frequency	Relative Frequency	Cumulative Relative Frequency
0–1			
2–3			
4–5			
6–7+			

1. Using the tables, find the percent of data that is at most two. Which table did you use and why?
2. Using the tables, find the percent of data that is at most three. Which table did you use and why?
3. Using the tables, find the percent of data that is more than two. Which table did you use and why?
4. Using the tables, find the percent of data that is more than three. Which table did you use and why?

Discussion Questions:

1. Is one of the tables "more correct" than the other? Why or why not?
2. In general, how could you group the data differently? Are there any advantages to either way of grouping the data?
3. Why did you switch between tables, if you did, when answering the question above?

CHAPTER 1 PROJECT II: SAMPLING EXPERIMENT

Sampling Experiment

Class Time:

Names:

Student Learning Outcomes

- The student will demonstrate the simple random, systematic, stratified, and cluster sampling techniques.
- The student will explain the details of each procedure used.

In this lab, you will be asked to pick several random samples of restaurants. In each case, describe your procedure briefly, including how you might have used the random number generator, and then list the restaurants in the sample you obtained.

Note: The following section contains restaurants stratified by city into columns and grouped horizontally by entree cost (clusters).

Restaurants Stratified by City and Entree Cost

Restaurants Used in Sample

Entree Cost	Under $10	$10 to under $15	$15 to under $20	Over $20
San Jose	El Abuelo Taq, Pasta Mia, Emma's Express, Bamboo Hut	Emperor's Guard, Creekside Inn	Agenda, Gervais, Miro's	Blake's, Eulipia, Hayes Mansion, Germania
Palo Alto	Senor Taco, Olive Garden, Taxi's	Ming's, P.A. Joe's, Stickney's	Scott's Seafood, Poolside Grill, Fish Market	Sundance Mine, Maddalena's, Spago's
Los Gatos	Mary's Patio, Mount Everest, Sweet Pea's, Andale Taqueria	Lindsey's, Willow Street	Toll House	Charter House, La Maison Du Cafe
Mountain View	Maharaja, New Ma's, Thai-Rriffic, Garden Fresh	Amber Indian, La Fiesta, Fiesta del Mar, Dawit	Austin's, Shiva's, Mazeh	Le Petit Bistro
Cupertino	Hobees, Hung Fu, Samrat, Panda Express	Santa Barb. Grill, Mand. Gourmet, Bombay Oven, Kathmandu West	Fontana's, Blue Pheasant	Hamasushi, Helios
Sunnyvale	Chekijababi, Taj India, Full Throttle, Tia Juana, Lemon Grass	Pacific Fresh, Charley Brown's, Cafe Cameroon, Faz, Aruba's	Lion & Compass, The Palace, Beau Sejour	
Santa Clara	Rangoli, Armadillo Willy's, Thai Pepper, Pasand	Arthur's, Katie's Cafe, Pedro's, La Galleria	Birk's, Truya Sushi, Valley Plaza	Lakeside, Mariani's

A Simple Random Sample: Pick a **simple random sample** of 15 restaurants.

1. Describe your procedure.
2. Complete the table with your sample.

1. _____	6. _____	11. _____
2. _____	7. _____	12. _____
3. _____	8. _____	13. _____
4. _____	9. _____	14. _____
5. _____	10. _____	15. _____

A Systematic Sample: Pick a **systematic sample** of 15 restaurants.

1. Describe your procedure.
2. Complete the table with your sample.

1. _____	6. _____	11. _____
2. _____	7. _____	12. _____
3. _____	8. _____	13. _____
4. _____	9. _____	14. _____
5. _____	10. _____	15. _____

A Stratified Sample: Pick a **stratified sample**, by city, of 20 restaurants. Use 25% of the restaurants from each stratum. Round to the nearest whole number.

1. Describe your procedure.
2. Complete the table with your sample.

1. _____	6. _____	11. _____	16. _____
2. _____	7. _____	12. _____	17. _____
3. _____	8. _____	13. _____	18. _____
4. _____	9. _____	14. _____	19. _____
5. _____	10. _____	15. _____	20. _____

A Stratified Sample: Pick a **stratified sample**, by entree cost, of 21 restaurants. Use 25% of the restaurants from each stratum. Round to the nearest whole number.

1. Describe your procedure.
2. Complete the table with your sample.

1. _____	6. _____	11. _____	16. _____
2. _____	7. _____	12. _____	17. _____
3. _____	8. _____	13. _____	18. _____
4. _____	9. _____	14. _____	19. _____
5. _____	10. _____	15. _____	20. _____
			21. _____

A Cluster Sample: Pick a **cluster sample** of restaurants from two cities. The number of restaurants will vary.

1. Describe your procedure.

2. Complete the table with your sample.

1. _____	6. _____	11. _____	16. _____	21. _____
2. _____	7. _____	12. _____	17. _____	22. _____
3. _____	8. _____	13. _____	18. _____	23. _____
4. _____	9. _____	14. _____	19. _____	24. _____
5. _____	10. _____	15. _____	20. _____	25. _____

CHAPTER 2 PROJECT: DESCRIPTIVE STATISTICS

Descriptive Statistics

Class Time:

Names:

Student Learning Outcomes

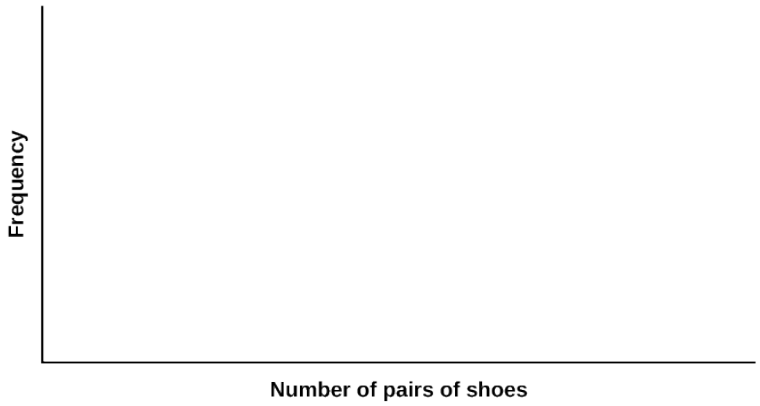
- The student will construct a histogram and a box plot.
- The student will calculate univariate statistics.
- The student will examine the graphs to interpret what the data implies.

Collect the Data: Record the number of pairs of shoes you own.

1. Randomly survey 30 classmates about the number of pairs of shoes they own. Record their values.

Survey Results				
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

2. Construct a histogram. Make five to six intervals. Sketch the graph using a ruler and pencil and scale the axes.



3. Calculate the following values.
   a. $\overline{x}$ = _____
   b. s = _____
4. Are the data discrete or continuous? How do you know?
5. In complete sentences, describe the shape of the histogram.
6. Are there any potential outliers? List the value(s) that could be outliers. Use a formula to check the end values to determine if they are potential outliers.

Analyze the Data:

1. Determine the following values.
   a. Min = _____
   b. M = _____
   c. Max = _____
   d. Q_1 = _____
   e. Q_3 = _____
   f. IQR = _____
2. Construct a box plot of data
3. What does the shape of the box plot imply about the concentration of data? Use complete sentences.
4. Using the box plot, how can you determine if there are potential outliers?
5. How does the standard deviation help you to determine concentration of the data and whether or not there are potential outliers?
6. What does the IQR represent in this problem?
7. Show your work to find the value that is 1.5 standard deviations:
   a. above the mean.
   b. below the mean.

CHAPTER 3 PROJECT: PROBABILITY TOPICS

Probability Topics

Class time:

Names:

Student Learning Outcomes

- The student will use theoretical and empirical methods to estimate probabilities.
- The student will appraise the differences between the two estimates.
- The student will demonstrate an understanding of long-term relative frequencies.

Do the Experiment Count out 40 mixed-color M&Ms® which is approximately one small bag's worth. Record the number of each color in (Figure). Use the information from this table to complete (Figure). Next, put the M&Ms in a cup. The experiment is to pick two M&Ms, one at a time. Do **not** look at them as you pick them. The first time through, replace the first M&M before picking the second one. Record the results in the "With Replacement" column of (Figure). Do this 24 times. The second time through, after picking the first M&M, do **not** replace it before picking the second one. Then, pick the second one. Record the results in the "Without Replacement" column section of (Figure). After you record the pick, put **both** M&Ms back. Do this a total of 24 times, also. Use the data from (Figure) to calculate the empirical probability questions. Leave your answers in unreduced fractional form. Do **not** multiply out any fractions.

Population

Color	Quantity
Yellow (Y)	
Green (G)	
Blue (BL)	
Brown (B)	
Orange (O)	
Red (R)	

Theoretical Probabilities

	With Replacement	Without Replacement
P(2 reds)		
$P(R_1B_2$ OR $B_1R_2)$		
$P(R_1$ AND $G_2)$		
$P(G_2\|R_1)$		
P(no yellows)		
P(doubles)		
P(no doubles)		

Note

G_2 = green on second pick; R_1 = red on first pick; B_1 = brown on first pick; B_2 = brown on second pick; doubles = both picks are the same color.

Empirical Results

With Replacement	Without Replacement
(__, __)(__, __)	(__, __)(__, __)
(__, __)(__, __)	(__, __)(__, __)
(__, __)(__, __)	(__, __)(__, __)
(__, __)(__, __)	(__, __)(__, __)
(__, __)(__, __)	(__, __)(__, __)
(__, __)(__, __)	(__, __)(__, __)
(__, __)(__, __)	(__, __)(__, __)
(__, __)(__, __)	(__, __)(__, __)
(__, __)(__, __)	(__, __)(__, __)
(__, __)(__, __)	(__, __)(__, __)
(__, __)(__, __)	(__, __)(__, __)
(__, __)(__, __)	(__, __)(__, __)

	Empirical Probabilities	
	With Replacement	**Without Replacement**
P(2 reds)		
$P(R_1B_2$ OR $B_1R_2)$		
$P(R_1$ AND $G_2)$		
$P(G_2\|R_1)$		
P(no yellows)		
P(doubles)		
P(no doubles)		

Discussion Questions

1. Why are the "With Replacement" and "Without Replacement" probabilities different?
2. Convert P(no yellows) to decimal format for both Theoretical "With Replacement" and for Empirical "With Replacement". Round to four decimal places.
   a. Theoretical "With Replacement": P(no yellows) = _____
   b. Empirical "With Replacement": P(no yellows) = _____
   c. Are the decimal values "close"? Did you expect them to be closer together or farther apart? Why?
3. If you increased the number of times you picked two M&Ms to 240 times, why would empirical probability values change?
4. Would this change (see part 3) cause the empirical probabilities and theoretical probabilities to be closer together or farther apart? How do you know?
5. Explain the differences in what $P(G_1$ AND $R_2)$ and $P(R_1|G_2)$ represent. Hint: Think about the sample space for each probability.

CHAPTER 4 PROJECT I: DISCRETE DISTRIBUTION (LUCKY DICE EXPERIMENT)

Discrete Distribution (Lucky Dice Experiment)

Class Time:

Names:

Student Learning Outcomes

- The student will compare empirical data and a theoretical distribution to determine if a Tet gambling game fits a discrete distribution.
- The student will demonstrate an understanding of long-term probabilities.

Supplies

- one "Lucky Dice" game or three regular dice

Procedure

Round answers to relative frequency and probability problems to four decimal places.

1. The experimental procedure is to bet on one object. Then, roll three Lucky Dice and count the number of matches. The number of matches will decide your profit.
2. What is the theoretical probability of one die matching the object?
3. Choose one object to place a bet on. Roll the three Lucky Dice. Count the number of matches.
4. Let X = number of matches. Theoretically, $X \sim B(_____,_____)$
5. Let Y = profit per game.

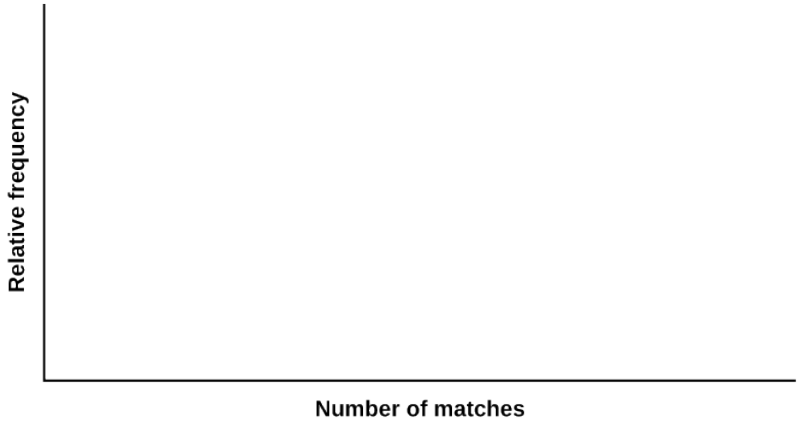
Organize the Data

In (Figure), fill in the y value that corresponds to each x value. Next, record the number of matches picked for your class. Then, calculate the relative frequency.

1. Complete the table.

x	y	Frequency	Relative Frequency
0			
1			
2			
3			

2. Calculate the following:
   a. $\overline{x}$ = _____
   b. s_x = _____
   c. $\overline{y}$ = _____
   d. s_y = _____
3. Explain what $\overline{x}$ represents.
4. Explain what $\overline{y}$ represents.
5. Based upon the experiment:
   a. What was the average profit per game?
   b. Did this represent an average win or loss per game?
   c. How do you know? Answer in complete sentences.
6. Construct a histogram of the empirical data.



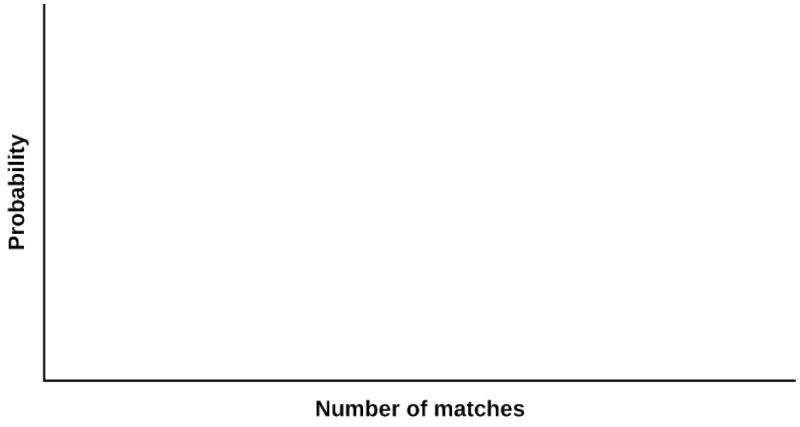
Theoretical Distribution

Build the theoretical PDF chart for *x* and *y* based on the distribution from the Procedure section.

1.

x	y	$P(x) = P(y)$
0		
1		
2		
3		

2. Calculate the following:
   a. $\mu_x =$ _____
   b. $\sigma_x =$ _____
   c. $\mu_x =$ _____
3. Explain what μ_x represents.
4. Explain what μ_y represents.
5. Based upon theory:
   a. What was the expected profit per game?
   b. Did the expected profit represent an average win or loss per game?
   c. How do you know? Answer in complete sentences.
6. Construct a histogram of the theoretical distribution.



Use the Data
Note

RF = relative frequency

Use the data from the Theoretical Distribution section to calculate the following answers. Round your answers to four decimal places.

1. $P(x = 3) =$ _____
2. $P(0 < x < 3) =$ _____

3. $P(x \geq 2) =$ _____

Use the data from the Organize the Data section to calculate the following answers. Round your answers to four decimal places.

1. $RF(x = 3) =$ _____
2. $RF(0 < x < 3) =$ _____
3. $RF(x \geq 2) =$ _____

Discussion Questions

For questions 1 and 2, consider the graphs, the probabilities, the relative frequencies, the means, and the standard deviations.

1. Knowing that data vary, describe three similarities between the graphs and distributions of the theoretical and empirical distributions. Use complete sentences.
2. Describe the three most significant differences between the graphs or distributions of the theoretical and empirical distributions.
3. Thinking about your answers to questions 1 and 2, does it appear that the data fit the theoretical distribution? In complete sentences, explain why or why not.
4. Suppose that the experiment had been repeated 500 times. Would you expect (Figure) or (Figure) to change, and how would it change? Why? Why wouldn't the other table change?

CHAPTER 4 PROJECT II: DISCRETE DISTRIBUTION (PLAYING CARD EXPERIMENT)

Discrete Distribution (Lucky Dice Experiment)

Class Time:

Names:

Student Learning Outcomes

- The student will compare empirical data and a theoretical distribution to determine if a Tet gambling game fits a discrete distribution.
- The student will demonstrate an understanding of long-term probabilities.

Supplies

- one "Lucky Dice" game or three regular dice

Procedure

Round answers to relative frequency and probability problems to four decimal places.

1. The experimental procedure is to bet on one object. Then, roll three Lucky Dice and count the number of matches. The number of matches will decide your profit.
2. What is the theoretical probability of one die matching the object?
3. Choose one object to place a bet on. Roll the three Lucky Dice. Count the number of matches.
4. Let X = number of matches. Theoretically, $X \sim B(_____,_____)$
5. Let Y = profit per game.

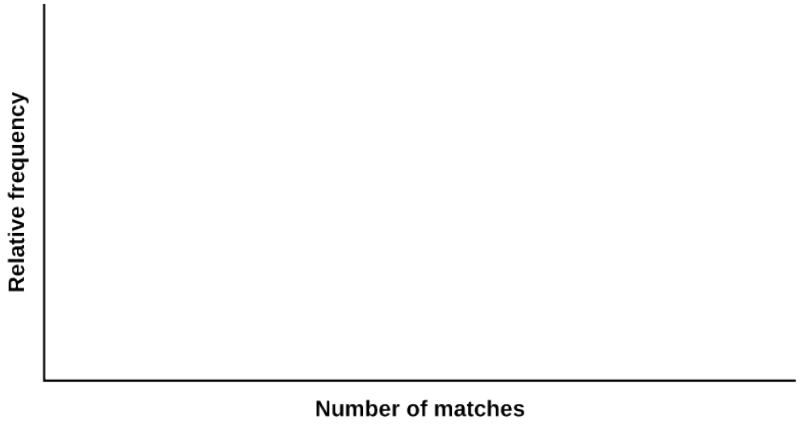
Organize the Data

In (Figure), fill in the y value that corresponds to each x value. Next, record the number of matches picked for your class. Then, calculate the relative frequency.

1. Complete the table.

x	y	Frequency	Relative Frequency
0			
1			
2			
3			

2. Calculate the following:
   a. $\overline{x}$ = _____
   b. s_x = _____
   c. $\overline{y}$ = _____
   d. s_y = _____
3. Explain what $\overline{x}$ represents.
4. Explain what $\overline{y}$ represents.
5. Based upon the experiment:
   a. What was the average profit per game?
   b. Did this represent an average win or loss per game?
   c. How do you know? Answer in complete sentences.
6. Construct a histogram of the empirical data.



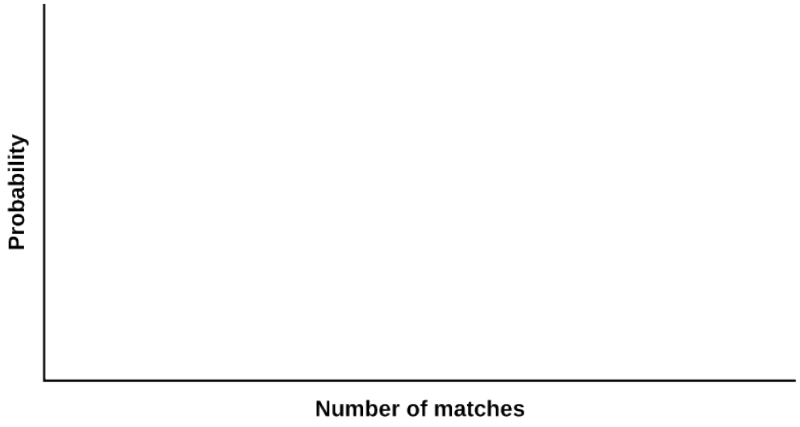
Theoretical Distribution

Build the theoretical PDF chart for *x* and *y* based on the distribution from the Procedure section.

1.

x	y	$P(x) = P(y)$
0		
1		
2		
3		

2. Calculate the following:
   a. $\mu_x =$ _____
   b. $\sigma_x =$ _____
   c. $\mu_x =$ _____
3. Explain what μ_x represents.
4. Explain what μ_y represents.
5. Based upon theory:
   a. What was the expected profit per game?
   b. Did the expected profit represent an average win or loss per game?
   c. How do you know? Answer in complete sentences.
6. Construct a histogram of the theoretical distribution.



Use the Data
Note

RF = relative frequency

Use the data from the Theoretical Distribution section to calculate the following answers. Round your answers to four decimal places.

1. $P(x = 3) =$ _____
2. $P(0 < x < 3) =$ _____

3. $P(x \geq 2) =$ _____

Use the data from the Organize the Data section to calculate the following answers. Round your answers to four decimal places.

1. $RF(x = 3) =$ _____
2. $RF(0 < x < 3) =$ _____
3. $RF(x \geq 2) =$ _____

Discussion Questions

For questions 1 and 2, consider the graphs, the probabilities, the relative frequencies, the means, and the standard deviations.

1. Knowing that data vary, describe three similarities between the graphs and distributions of the theoretical and empirical distributions. Use complete sentences.
2. Describe the three most significant differences between the graphs or distributions of the theoretical and empirical distributions.
3. Thinking about your answers to questions 1 and 2, does it appear that the data fit the theoretical distribution? In complete sentences, explain why or why not.
4. Suppose that the experiment had been repeated 500 times. Would you expect (Figure) or (Figure) to change, and how would it change? Why? Why wouldn't the other table change?

CHAPTER 5 PROJECT: CONTINUOUS DISTRIBUTION

Continuous Distribution

Class Time:

Names:

Student Learning Outcomes

- The student will compare and contrast empirical data from a random number generator with the uniform distribution.

Collect the DataUse a random number generator to generate 50 values between zero and one (inclusive). List them in (Figure). Round the numbers to four decimal places or set the calculator MODE to four places.

1. Complete the table.

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

2. Calculate the following:
   a. $\overline{x}$ = _____
   b. s = _____
   c. first quartile = _____

d. third quartile = _____
e. median = _____

Organize the Data

1. Construct a histogram of the empirical data. Make eight bars.



2. Construct a histogram of the empirical data. Make five bars.



Describe the Data

1. In two to three complete sentences, describe the shape of each graph. (Keep it simple. Does the graph go straight across, does it have a V shape, does it have a hump in the middle or at either end, and so on. One way to help you determine a shape is to draw a smooth curve roughly through the top of the bars.)
2. Describe how changing the number of bars might change the shape.

Theoretical Distribution

1. In words, X = _____.
2. The theoretical distribution of X is $X \sim U(0,1)$.
3. In theory, based upon the distribution $X \sim U(0,1)$, complete the following.
   a. μ = _____

   b. σ = _____
   c. first quartile = _____
   d. third quartile = _____
   e. median = _____
4. Are the empirical values (the data) in the section titled Collect the Data close to the corresponding theoretical values? Why or why not?

Plot the Data

1. Construct a box plot of the data. Be sure to use a ruler to scale accurately and draw straight edges.
2. Do you notice any potential outliers? If so, which values are they? Either way, justify your answer numerically. (Recall that any DATA that are less than $Q_1 - 1.5(IQR)$ or more than $Q_3 + 1.5(IQR)$ are potential outliers. IQR means interquartile range.)

Compare the Data

1. For each of the following parts, use a complete sentence to comment on how the value obtained from the data compares to the theoretical value you expected from the distribution in the section titled Theoretical Distribution.
   a. minimum value: _____
   b. first quartile: _____
   c. median: _____
   d. third quartile: _____
   e. maximum value: _____
   f. width of IQR: _____
   g. overall shape: _____
2. Based on your comments in the section titled Collect the Data, how does the box plot fit or not fit what you would expect of the distribution in the section titled Theoretical Distribution?

Discussion Question

1. Suppose that the number of values generated was 500, not 50. How would that affect what you would expect the empirical data to be and the shape of its graph to look like?

CHAPTER 6 PROJECT I: NORMAL DISTRIBUTION (LAP TIMES)

Normal Distribution (Lap Times)

Class Time:

Names:

Student Learning Outcome

- The student will compare and contrast empirical data and a theoretical distribution to determine if Terry Vogel's lap times fit a continuous distribution.

Directions

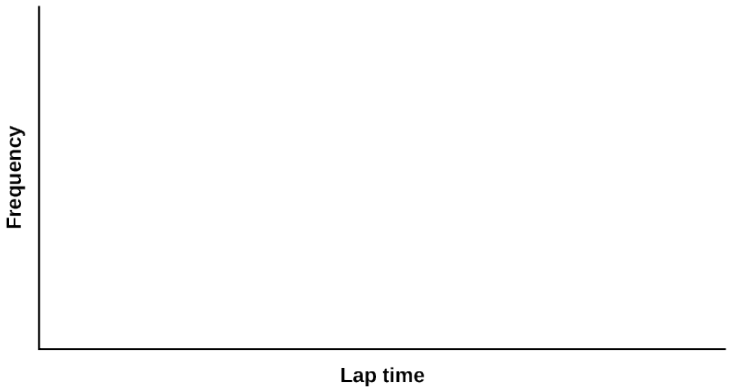
Round the relative frequencies and probabilities to four decimal places. Carry all other decimal answers to two places.

Collect the Data

1. Use the data from Appendix C. Use a stratified sampling method by lap (races 1 to 20) and a random number generator to pick six lap times from each stratum. Record the lap times below for laps two to seven.

_____	_____	_____	_____	_____	_____
_____	_____	_____	_____	_____	_____
_____	_____	_____	_____	_____	_____
_____	_____	_____	_____	_____	_____
_____	_____	_____	_____	_____	_____
_____	_____	_____	_____	_____	_____

2. Construct a histogram. Make five to six intervals. Sketch the graph using a ruler and pencil. Scale the axes.



3. Calculate the following:
   a. $\overline{x}$ = _____
   b. s = _____
4. Draw a smooth curve through the tops of the bars of the histogram. Write one to two complete sentences to describe the general shape of the curve. (Keep it simple. Does the graph go straight across, does it have a v-shape, does it have a hump in the middle or at either end, and so on?)

Analyze the Distribution Using your sample mean, sample standard deviation, and histogram to help, what is the approximate theoretical distribution of the data?

- X ~ _____(_____,_____)
- How does the histogram help you arrive at the approximate distribution?

Describe the Data

Use the data you collected to complete the following statements.

- The *IQR* goes from _____ to _____.
- *IQR* = _____. ($IQR = Q_3 - Q_1$)
- The 15^{th} percentile is _____.
- The 85^{th} percentile is _____.
- The median is _____.
- The empirical probability that a randomly chosen lap time is more than 130 seconds is _____.
- Explain the meaning of the 85^{th} percentile of this data.

Theoretical Distribution

Using the theoretical distribution, complete the following statements. You should use a normal approximation based on your sample data.

- The *IQR* goes from _____ to _____.

- *IQR* = _____.
- The 15th percentile is _____.
- The 85th percentile is _____.
- The median is _____.
- The probability that a randomly chosen lap time is more than 130 seconds is _____.
- Explain the meaning of the 85th percentile of this distribution.

Discussion Questions

Do the data from the section titled Collect the Data give a close approximation to the theoretical distribution in the section titled Analyze the Distribution? In complete sentences and comparing the result in the sections titled Describe the Data and Theoretical Distribution, explain why or why not.

CHAPTER 6 PROJECT II: NORMAL DISTRIBUTION (PINKIE LENGTH)

Normal Distribution (Pinkie Length)

Class Time:

Names:

Student Learning Outcomes

- The student will compare empirical data and a theoretical distribution to determine if data from the experiment follow a continuous distribution.

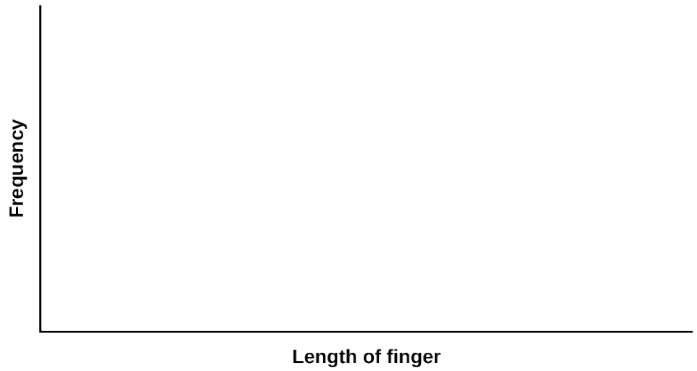
Collect the Data

Measure the length of your pinky finger (in centimeters).

1. Randomly survey 30 adults for their pinky finger lengths. Round the lengths to the nearest 0.5 cm.

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

2. Construct a histogram. Make five to six intervals. Sketch the graph using a ruler and pencil. Scale the axes.



3. Calculate the following.
   a. $\overline{x}$ = _____
   b. s = _____
4. Draw a smooth curve through the top of the bars of the histogram. Write one to two complete sentences to describe the general shape of the curve. (Keep it simple. Does the graph go straight across, does it have a v-shape, does it have a hump in the middle or at either end, and so on?)

Analyze the Distribution

Using your sample mean, sample standard deviation, and histogram, what was the approximate theoretical distribution of the data you collected?

- X ~ _____(_____,_____)
- How does the histogram help you arrive at the approximate distribution?

Describe the Data

Using the data you collected complete the following statements. (Hint: order the data)
Remember

($IQR = Q_3 - Q_1$)

- IQR = _____
- The 15th percentile is _____.
- The 85th percentile is _____.
- Median is _____.
- What is the theoretical probability that a randomly chosen pinky length is more than 6.5 cm?
- Explain the meaning of the 85th percentile of this data.

Theoretical Distribution

Using the theoretical distribution, complete the following statements. Use a normal approximation based on the sample mean and standard deviation.

- *IQR* = _____
- The 15^{th} percentile is _____.
- The 85^{th} percentile is _____.
- Median is _____.
- What is the theoretical probability that a randomly chosen pinky length is more than 6.5 cm?
- Explain the meaning of the 85^{th} percentile of this data.

Discussion Questions

Do the data you collected give a close approximation to the theoretical distribution? In complete sentences and comparing the results in the sections titled Describe the Data and Theoretical Distribution, explain why or why not.

CHAPTER 6 PROJECT III: CENTRAL LIMIT THEOREM (POCKET CHANGE)

Central Limit Theorem (Pocket Change)

Class Time:

Names:

Student Learning Outcomes

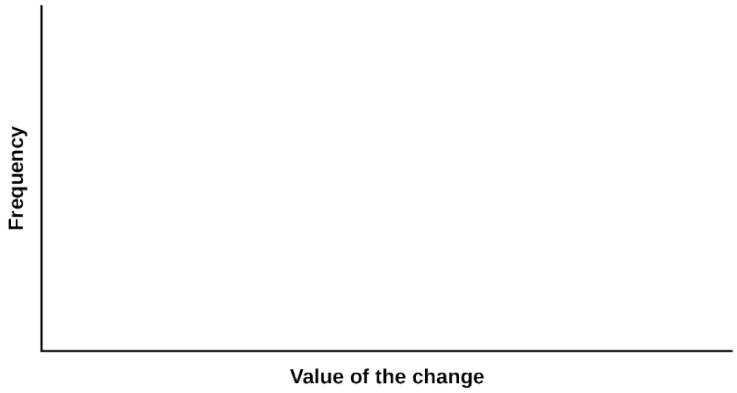
- The student will demonstrate and compare properties of the central limit theorem.

Note

This lab works best when sampling from several classes and combining data.
Collect the Data

1. Count the change in your pocket. (Do not include bills.)
2. Randomly survey 30 classmates. Record the values of the change in (Figure).

3. Construct a histogram. Make five to six intervals. Sketch the graph using a ruler and pencil. Scale the axes.



4. Calculate the following (n = 1; surveying one person at a time):
   a. $\overline{x}$ = _____
   b. s = _____
5. Draw a smooth curve through the tops of the bars of the histogram. Use one to two complete sentences to describe the general shape of the curve.

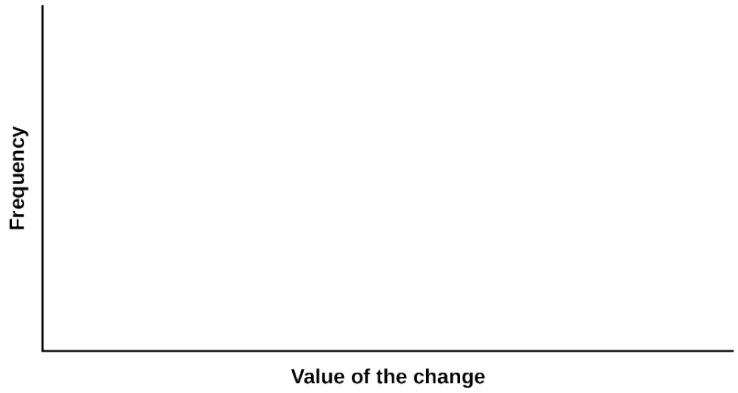
Collecting Averages of Pairs

Repeat steps one through five of the section Collect the Data. with one exception. Instead of recording the change of 30 classmates, record the average change of 30 pairs.

1. Randomly survey 30 **pairs** of classmates.
2. Record the values of the average of their change in (Figure).

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

3. Construct a histogram. Scale the axes using the same scaling you used for the section titled Collect the Data. Sketch the graph using a ruler and a pencil.



4. Calculate the following (n = 2; surveying two people at a time):
   a. $\overline{x}$ = _____
   b. s = _____
5. Draw a smooth curve through the tops of the bars of the histogram. Use one to two complete sentences to describe the general shape of the curve.

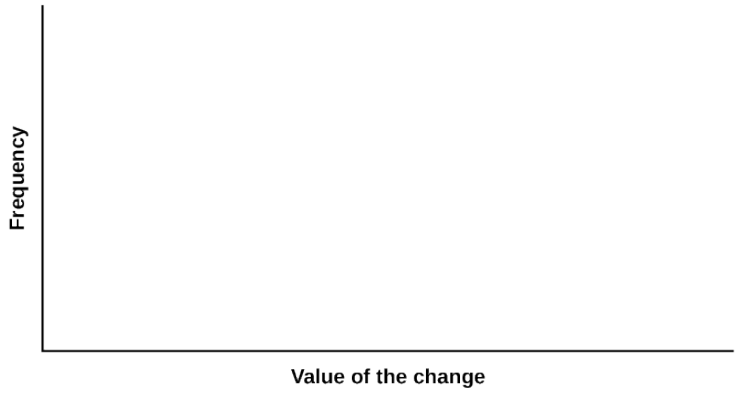
Collecting Averages of Groups of Five

Repeat steps one through five (of the section titled Collect the Data) with one exception. Instead of recording the change of 30 classmates, record the average change of 30 groups of five.

1. Randomly survey 30 **groups of five** classmates.
2. Record the values of the average of their change.

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

3. Construct a histogram. Scale the axes using the same scaling you used for the section titled Collect the Data. Sketch the graph using a ruler and a pencil.



4. Calculate the following (n = 5; surveying five people at a time):
   a. $\overline{x}$ = _____
   b. s = _____
5. Draw a smooth curve through the tops of the bars of the histogram. Use one to two complete sentences to describe the general shape of the curve.

Discussion Questions

1. Why did the shape of the distribution of the data change, as n changed? Use one to two complete sentences to explain what happened.
2. In the section titled Collect the Data, what was the approximate distribution of the data? X ~ _____(_____,_____)
3. In the section titled Collecting Averages of Groups of Five, what was the approximate distribution of the averages? $\overline{X}$ ~ _____(_____,_____)
4. In one to two complete sentences, explain any differences in your answers to the previous two questions.

CHAPTER 6 PROJECT IV: CENTRAL LIMIT THEOREM (COOKIE RECIPES)

Central Limit Theorem (Cookie Recipes)

Class Time:

Names:

Student Learning Outcomes

- The student will demonstrate and compare properties of the central limit theorem.

Given X = length of time (in days) that a cookie recipe lasted at the Olmstead Homestead. (Assume that each of the different recipes makes the same quantity of cookies.)

Recipe #	*X*	Recipe #	*X*	Recipe #	*X*	Recipe #	*X*
1	1	16	2	31	3	46	2
2	5	17	2	32	4	47	2
3	2	18	4	33	5	48	11
4	5	19	6	34	6	49	5
5	6	20	1	35	6	50	5
6	1	21	6	36	1	51	4
7	2	22	5	37	1	52	6
8	6	23	2	38	2	53	5
9	5	24	5	39	1	54	1
10	2	25	1	40	6	55	1
11	5	26	6	41	1	56	2
12	1	27	4	42	6	57	4
13	1	28	1	43	2	58	3
14	3	29	6	44	6	59	6
15	2	30	2	45	2	60	5

Calculate the following:

a. μ_x = _____
b. σ_x = _____

Collect the Data

Use a random number generator to randomly select four samples of size $n = 5$ from the given population. Record your samples in (Figure). Then, for each sample, calculate the mean to the nearest tenth. Record them in the spaces provided. Record the sample means for the rest of the class.

1. Complete the table:

Sample 1	Sample 2	Sample 3	Sample 4	Sample means from other groups:
Means: $\overline{x}$ = ____	$\overline{x}$ = ____	$\overline{x}$ = ____	$\overline{x}$ = ____	

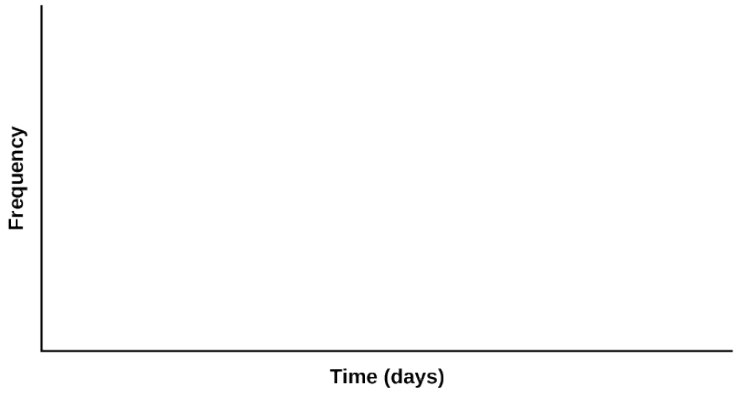
2. Calculate the following:
   a. $\overline{x}$ = _____
   b. $s_{\overline{x}}$ = _____
3. Again, use a random number generator to randomly select four samples from the population. This time, make the samples of size $n = 10$. Record the samples in (Figure). As before, for each sample, calculate the mean to the nearest tenth. Record them in the spaces provided. Record the sample means for the rest of the class.

Sample 1	Sample 2	Sample 3	Sample 4	Sample means from other groups
Means: $\overline{x}$ = ____	$\overline{x}$ = ____	$\overline{x}$ = ____	$\overline{x}$ = ____	

4. Calculate the following:
   a. $\overline{x}$ = _____

b. $s_{\overline{x}}$ = _____

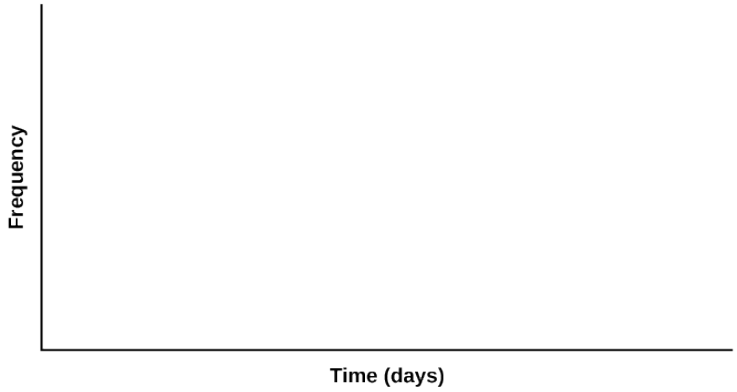
5. For the original population, construct a histogram. Make intervals with a bar width of one day. Sketch the graph using a ruler and pencil. Scale the axes.



6. Draw a smooth curve through the tops of the bars of the histogram. Use one to two complete sentences to describe the general shape of the curve.

Repeat the Procedure for $n = 5$

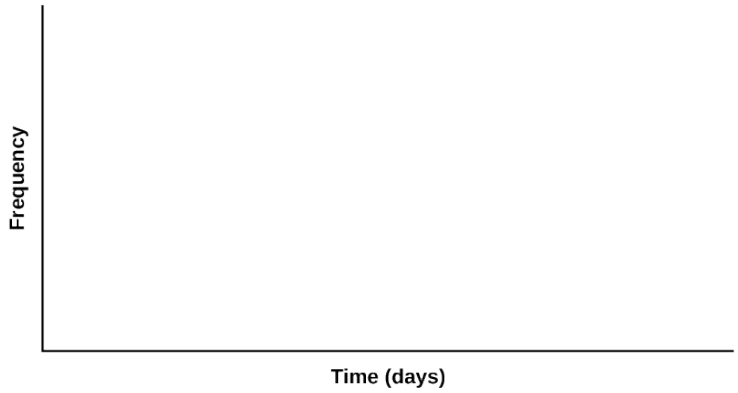
1. For the sample of $n = 5$ days averaged together, construct a histogram of the averages (your means together with the means of the other groups). Make intervals with bar widths of $\frac{1}{2}$ a day. Sketch the graph using a ruler and pencil. Scale the axes.



2. Draw a smooth curve through the tops of the bars of the histogram. Use one to two complete sentences to describe the general shape of the curve.

Repeat the Procedure for $n = 10$

1. For the sample of $n = 10$ days averaged together, construct a histogram of the averages (your means together with the means of the other groups). Make intervals with bar widths of $\frac{1}{2}$ a day. Sketch the graph using a ruler and pencil. Scale the axes.



2. Draw a smooth curve through the tops of the bars of the histogram. Use one to two complete sentences to describe the general shape of the curve.

Discussion Questions

1. Compare the three histograms you have made, the one for the population and the two for the sample means. In three to five sentences, describe the similarities and differences.
2. State the theoretical (according to the CLT) distributions for the sample means.
   a. $n = 5$: $\overline{x}$ ~ _____(_____,_____)
   b. $n = 10$: $\overline{x}$ ~ _____(_____,_____)
3. Are the sample means for $n = 5$ and $n = 10$ "close" to the theoretical mean, μ_x? Explain why or why not.
4. Which of the two distributions of sample means has the smaller standard deviation? Why?
5. As n changed, why did the shape of the distribution of the data change? Use one to two complete sentences to explain what happened.

CHAPTER 7 PROJECT I: CONFIDENCE INTERVAL (HOME COSTS)

Confidence Interval (Home Costs)

Class Time:

Names:

Student Learning Outcomes

- The student will calculate the 90% confidence interval for the mean cost of a home in the area in which this school is located.
- The student will interpret confidence intervals.
- The student will determine the effects of changing conditions on the confidence interval.

Collect the Data

Check the Real Estate section in your local newspaper. Record the sale prices for 35 randomly selected homes recently listed in the county.

Note

Many newspapers list them only one day per week. Also, we will assume that homes come up for sale randomly.

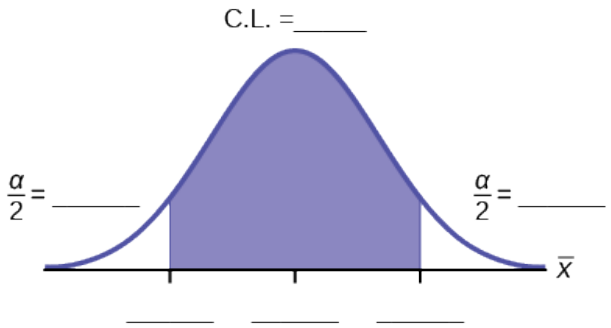
1. Complete the table:

Describe the Data

1. Compute the following:
   a. $\overline{x}$ = _____
   b. s_x = _____
   c. n = _____
2. In words, define the random variable $\overline{X}$.
3. State the estimated distribution to use. Use both words and symbols.

Find the Confidence Interval

1. Calculate the confidence interval and the error bound.
   a. Confidence Interval: _____
   b. Error Bound: _____
2. How much area is in both tails (combined)? α = _____
3. How much area is in each tail? $\frac{\alpha}{2}$ = _____
4. Fill in the blanks on the graph with the area in each section. Then, fill in the number line with the upper and lower limits of the confidence interval and the sample mean.



5. Some students think that a 90% confidence interval contains 90% of the data. Use the list of data on the first page and count how many of the data values lie within the confidence interval. What percent is this? Is this percent close to 90%? Explain why this percent should or should not be close to 90%.

Describe the Confidence Interval

1. In two to three complete sentences, explain what a confidence interval means (in general), as if you were talking to someone who has not taken statistics.
2. In one to two complete sentences, explain what this confidence interval means for this particular study.

Use the Data to Construct Confidence Intervals

1. Using the given information, construct a confidence interval for each confidence level given.

Confidence level	EBM/Error Bound	Confidence Interval
50%		
80%		
95%		
99%		

2. What happens to the EBM as the confidence level increases? Does the width of the confidence interval increase or decrease? Explain why this happens.

CHAPTER 7 PROJECT II: CONFIDENCE INTERVAL (PLACE OF BIRTH)

Confidence Interval (Place of Birth)

Class Time:

Names:

Student Learning Outcomes

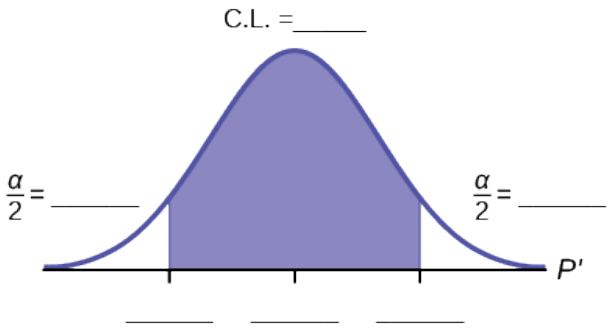
- The student will calculate the 90% confidence interval the proportion of students in this school who were born in this state.
- The student will interpret confidence intervals.
- The student will determine the effects of changing conditions on the confidence interval.

Collect the Data

1. Survey the students in your class, asking them if they were born in this state. Let X = the number that were born in this state.
   a. n = _____
   b. x = _____
2. In words, define the random variable P'.
3. State the estimated distribution to use.

Find the Confidence Interval and Error Bound

1. Calculate the confidence interval and the error bound.
   a. Confidence Interval: _____
   b. Error Bound: _____
2. How much area is in both tails (combined)? α = _____
3. How much area is in each tail? $\frac{\alpha}{2}$ = _____
4. Fill in the blanks on the graph with the area in each section. Then, fill in the number line with the upper and lower limits of the confidence interval and the sample proportion.



Describe the Confidence Interval

1. In two to three complete sentences, explain what a confidence interval means (in general), as though you were talking to someone who has not taken statistics.
2. In one to two complete sentences, explain what this confidence interval means for this particular study.
3. Construct a confidence interval for each confidence level given.

Confidence level	EBP/Error Bound	Confidence Interval
50%		
80%		
95%		
99%		

4. What happens to the EBP as the confidence level increases? Does the width of the confidence interval increase or decrease? Explain why this happens.

CHAPTER 7 PROJECT III: CONFIDENCE INTERVAL (WOMEN'S HEIGHTS)

Confidence Interval (Women's Heights)

Class Time:

Names:

Student Learning Outcomes

- The student will calculate a 90% confidence interval using the given data.
- The student will determine the relationship between the confidence level and the percentage of constructed intervals that contain the population mean.

Given:

Heights of 100 Women (in Inches)

59.4	71.6	69.3	65.0	62.9	66.5	61.7	55.2
67.5	67.2	63.8	62.9	63.0	63.9	68.7	65.5
61.9	69.6	58.7	63.4	61.8	60.6	69.8	60.0
64.9	66.1	66.8	60.6	65.6	63.8	61.3	59.2
64.1	59.3	64.9	62.4	63.5	60.9	63.3	66.3
61.5	64.3	62.9	60.6	63.8	58.8	64.9	65.7
62.5	70.9	62.9	63.1	62.2	58.7	64.7	66.0
60.5	64.7	65.4	60.2	65.0	64.1	61.1	65.3
64.6	59.2	61.4	62.0	63.5	61.4	65.5	62.3
65.5	64.7	58.8	66.1	64.9	66.9	57.9	69.8
58.5	63.4	69.2	65.9	62.2	60.0	58.1	62.5
62.4	59.1	66.4	61.2	60.4	58.7	66.7	67.5
63.2	56.6	67.7	62.5				

1. (Figure) lists the heights of 100 women. Use a random number generator to select ten data values

randomly.

2. Calculate the sample mean and the sample standard deviation. Assume that the population standard deviation is known to be 3.3 inches. With these values, construct a 90% confidence interval for your sample of ten values. Write the confidence interval you obtained in the first space of (Figure).
3. Now write your confidence interval on the board. As others in the class write their confidence intervals on the board, copy them into (Figure).

90% Confidence Intervals

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

Discussion Questions

1. The actual population mean for the 100 heights given (Figure) is $\mu = 63.4$. Using the class listing of confidence intervals, count how many of them contain the population mean μ; i.e., for how many intervals does the value of μ lie between the endpoints of the confidence interval?
2. Divide this number by the total number of confidence intervals generated by the class to determine the percent of confidence intervals that contains the mean μ. Write this percent here: _____.
3. Is the percent of confidence intervals that contain the population mean μ close to 90%?
4. Suppose we had generated 100 confidence intervals. What do you think would happen to the percent of confidence intervals that contained the population mean?
5. When we construct a 90% confidence interval, we say that we are **90% confident that the true population mean lies within the confidence interval.** Using complete sentences, explain what we mean by this phrase.
6. Some students think that a 90% confidence interval contains 90% of the data. Use the list of data given (the heights of women) and count how many of the data values lie within the confidence interval that you generated based on that data. How many of the 100 data values lie within your confidence interval? What percent is this? Is this percent close to 90%?
7. Explain why it does not make sense to count data values that lie in a confidence interval. Think about the random variable that is being used in the problem.

8. Suppose you obtained the heights of ten women and calculated a confidence interval from this information. Without knowing the population mean μ, would you have any way of knowing **for certain** if your interval actually contained the value of μ? Explain.

CHAPTER 8 PROJECT: HYPOTHESIS TESTING OF A SINGLE MEAN AND SINGLE PROPORTION

Hypothesis Testing of a Single Mean and Single Proportion

Class Time:

Names:

Student Learning Outcomes

- The student will select the appropriate distributions to use in each case.
- The student will conduct hypothesis tests and interpret the results.

Television Survey

In a recent survey, it was stated that Americans watch television on average four hours per day. Assume that $\sigma = 2$. Using your class as the sample, conduct a hypothesis test to determine if the average for students at your school is lower.

1. H_0: _____
2. H_a: _____
3. In words, define the random variable. _____ = _____
4. The distribution to use for the test is _____.
5. Determine the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.
    a. Graph:



   b. Determine the *p*-value.
7. Do you or do you not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

Language Survey

About 42.3% of Californians and 19.6% of all Americans over age five speak a language other than English at home. Using your class as the sample, conduct a hypothesis test to determine if the percent of the students at your school who speak a language other than English at home is different from 42.3%.

1. H_0: _____
2. H_a: _____
3. In words, define the random variable. _____ = _____
4. The distribution to use for the test is _____
5. Determine the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.
   a. Graph:



   b. Determine the *p*-value.
7. Do you or do you not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

Jeans Survey

Suppose that young adults own an average of three pairs of jeans. Survey eight people from your class to determine if the average is higher than three. Assume the population is normal.

1. H_0: _____
2. H_a: _____
3. In words, define the random variable. _____ = _____
4. The distribution to use for the test is _____.
5. Determine the test statistic using your data.

6. Draw a graph and label it appropriately. Shade the actual level of significance.
    a. Graph:



    b. Determine the *p*-value.
7. Do you or do you not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

CHAPTER 9 PROJECT: HYPOTHESIS TESTING FOR TWO MEANS AND TWO PROPORTIONS

Hypothesis Testing for Two Means and Two Proportions

Class Time:

Names:

Student Learning Outcomes

- The student will select the appropriate distributions to use in each case.
- The student will conduct hypothesis tests and interpret the results.

Supplies:

- the business section from two consecutive days' newspapers
- three small packages of M&Ms®
- five small packages of Reese's Pieces®

Increasing Stocks Survey

Look at yesterday's newspaper business section. Conduct a hypothesis test to determine if the proportion of New York Stock Exchange (NYSE) stocks that increased is greater than the proportion of NASDAQ stocks that increased. As randomly as possible, choose 40 NYSE stocks, and 32 NASDAQ stocks and complete the following statements.

1. H_0:_____
2. H_a:_____
3. In words, define the random variable.
4. The distribution to use for the test is _____.
5. Calculate the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.
    1. Graph:



2. Calculate the *p*-value.

7. Do you reject or not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

Decreasing Stocks Survey

Randomly pick eight stocks from the newspaper. Using two consecutive days' business sections, test whether the stocks went down, on average, for the second day.

1. H_0: _____
2. H_a: _____
3. In words, define the random variable.
4. The distribution to use for the test is _____.
5. Calculate the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.
    1. Graph:



    2. Calculate the *p*-value:
7. Do you reject or not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

Candy Survey

Buy three small packages of M&Ms and five small packages of Reese's Pieces (same net weight as the M&Ms). Test whether or not the mean number of candy pieces per package is the same for the two brands.

1. H_0: _____
2. H_a: _____
3. In words, define the random variable.
4. What distribution should be used for this test?
5. Calculate the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.
    1. Graph:



    2. Calculate the *p*-value.
7. Do you reject or not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

Shoe Survey

Test whether women have, on average, more pairs of shoes than men. Include all forms of sneakers, shoes, sandals, and boots. Use your class as the sample.

1. H_0: _____
2. H_a: _____
3. In words, define the random variable.
4. The distribution to use for the test is _____.
5. Calculate the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.
    1. Graph:



2. Calculate the *p*-value.

7. Do you reject or not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

CHAPTER 10 PROJECT I: REGRESSION (DISTANCE FROM SCHOOL)

Regression (Distance from School)

Class Time:

Names:

Student Learning Outcomes

- The student will calculate and construct the line of best fit between two variables.
- The student will evaluate the relationship between two variables to determine if that relationship is significant.

Collect the Data

Use eight members of your class for the sample. Collect bivariate data (distance an individual lives from school, the cost of supplies for the current term).

1. Complete the table.

Distance from school	Cost of supplies this term

2. Which variable should be the dependent variable and which should be the independent variable? Why?
3. Graph "distance" vs. "cost." Plot the points on the graph. Label both axes with words. Scale both axes.



Analyze the Data

Enter your data into your calculator or computer. Write the linear equation, rounding to four decimal places.

1. Calculate the following:
    a. a = _____
    b. b = _____
    c. correlation = _____
    d. n = _____
    e. equation: $\hat{y}$ = _____
    f. Is the correlation significant? Why or why not? (Answer in one to three complete sentences.)
2. Supply an answer for the following scenarios:
    a. For a person who lives eight miles from campus, predict the total cost of supplies this term:
    b. For a person who lives eighty miles from campus, predict the total cost of supplies this term:
3. Obtain the graph on your calculator or computer. Sketch the regression line.



Discussion Questions

1. Answer each question in complete sentences.
    a. Does the line seem to fit the data? Why?

   b. What does the correlation imply about the relationship between the distance and the cost?
2. Are there any outliers? If so, which point is an outlier?
3. Should the outlier, if it exists, be removed? Why or why not?

CHAPTER 10 PROJECT II: REGRESSION (TEXTBOOK COST)

Regression (Textbook Cost)

Class Time:

Names:

Student Learning Outcomes

- The student will calculate and construct the line of best fit between two variables.
- The student will evaluate the relationship between two variables to determine if that relationship is significant.

Collect the Data

Survey ten textbooks. Collect bivariate data (number of pages in a textbook, the cost of the textbook).

1. Complete the table.

Number of pages	Cost of textbook

2. Which variable should be the dependent variable and which should be the independent variable? Why?
3. Graph "pages" vs. "cost." Plot the points on the graph in Analyze the Data. Label both axes with words. Scale both axes.

Analyze the Data

Enter your data into your calculator or computer. Write the linear equation, rounding to four decimal places.

1. Calculate the following:
   a. a = _____

   b. b = _____
   c. correlation = _____
   d. n = _____
   e. equation: y = _____
   f. Is the correlation significant? Why or why not? (Answer in complete sentences.)
2. Supply an answer for the following scenarios:
   a. For a textbook with 400 pages, predict the cost.
   b. For a textbook with 600 pages, predict the cost.
3. Obtain the graph on your calculator or computer. Sketch the regression line.



Discussion Questions

1. Answer each question in complete sentences.
   a. Does the line seem to fit the data? Why?
   b. What does the correlation imply about the relationship between the number of pages and the cost?
2. Are there any outliers? If so, which point(s) is an outlier?
3. Should the outlier, if it exists, be removed? Why or why not?

CHAPTER 10 PROJECT III: REGRESSION (FUEL EFFICIENCY)

Regression (Fuel Efficiency)

Class Time:

Names:

Student Learning Outcomes

- The student will calculate and construct the line of best fit between two variables.
- The student will evaluate the relationship between two variables to determine if that relationship is significant.

Collect the Data

Find a reputable source that provides information on total fuel efficiency (in miles per gallon) and weight (in pounds) of new model cars with automatic transmissions. We will use this data to determine the relationship, if any, between the fuel efficiency of a car and its weight.

1. Using your random number generator, randomly select 20 cars from the list and record their weights and fuel efficiency into (Figure).

Weight	Fuel Efficiency

2. Which variable should be the dependent variable and which should be the independent variable? Why?

3. By hand, do a scatterplot of "weight" vs. "fuel efficiency". Plot the points on graph paper. Label both axes with words. Scale both axes accurately.



Analyze the Data

Enter your data into your calculator or computer. Write the linear equation, rounding to 4 decimal places.

1. Calculate the following:
   a. a = _____
   b. b = _____
   c. correlation = _____
   d. n = _____
   e. equation: $\hat{y}$ = _____
2. Obtain the graph of the regression line on your calculator. Sketch the regression line on the same axes as your scatter plot.

Discussion Questions

1. Is the correlation significant? Explain how you determined this in complete sentences.
2. Is the relationship a positive one or a negative one? Explain how you can tell and what this means in terms of weight and fuel efficiency.
3. In one or two complete sentences, what is the practical interpretation of the slope of the least squares line in terms of fuel efficiency and weight?
4. For a car that weighs 4,000 pounds, predict its fuel efficiency. Include units.
5. Can we predict the fuel efficiency of a car that weighs 10,000 pounds using the least squares line? Explain why or why not.
6. Answer each question in complete sentences.
   a. Does the line seem to fit the data? Why or why not?
   b. What does the correlation imply about the relationship between fuel efficiency and weight of a car? Is this what you expected?

7. Are there any outliers? If so, which point is an outlier?

CHAPTER 11 PROJECT I: CHI-SQUARE GOODNESS-OF-FIT

Chi-Square Goodness-of-Fit

Class Time:

Names:

Student Learning Outcomes

- The student will evaluate data collected to determine if they fit either the uniform or exponential distributions.

Collect the Data

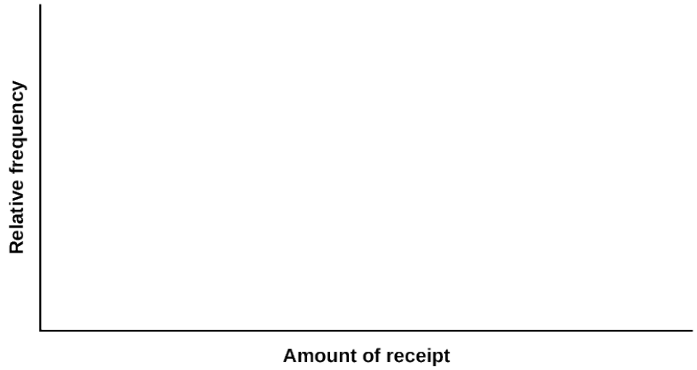
Go to your local supermarket. Ask 30 people as they leave for the total amount on their grocery receipts. (Or, ask three cashiers for the last ten amounts. Be sure to include the express lane, if it is open.)

NOTE

You may need to combine two categories so that each cell has an expected value of at least five.

1. Record the values.

2. Construct a histogram of the data. Make five to six intervals. Sketch the graph using a ruler and pencil. Scale the axes.



3. Calculate the following:
    1. $\overline{x}$ =_____
    2. s = _____
    3. s^2 = _____

Uniform Distribution

Test to see if grocery receipts follow the uniform distribution.

1. Using your lowest and highest values, $X \sim U$(_____, _____)
2. Divide the distribution into fifths.
3. Calculate the following:
    1. lowest value = _____
    2. 20^{th} percentile = _____
    3. 40^{th} percentile = _____
    4. 60^{th} percentile = _____
    5. 80^{th} percentile = _____
    6. highest value = _____
4. For each fifth, count the observed number of receipts and record it. Then determine the expected number of receipts and record that.

Fifth	Observed	Expected
1st		
2nd		
3rd		
4th		
5th		

5. H_0: _____
6. H_a: _____
7. What distribution should you use for a hypothesis test?
8. Why did you choose this distribution?
9. Calculate the test statistic.
10. Find the *p*-value.
11. Sketch a graph of the situation. Label and scale the *x*-axis. Shade the area corresponding to the *p*-value.



12. State your decision.
13. State your conclusion in a complete sentence.

Exponential Distribution

Test to see if grocery receipts follow the exponential distribution with decay parameter $\frac{1}{\overline{x}}$.

1. Using $\frac{1}{\overline{x}}$ as the decay parameter, $X \sim Exp$(_____).
2. Calculate the following:
    1. lowest value = _____
    2. first quartile = _____
    3. 37th percentile = _____
    4. median = _____

5. 63^{rd} percentile = _____
6. 3^{rd} quartile = _____
7. highest value = _____

3. For each cell, count the observed number of receipts and record it. Then determine the expected number of receipts and record that.

Cell	Observed	Expected
1^{st}		
2^{nd}		
3^{rd}		
4^{th}		
5^{th}		
6^{th}		

4. H_0: _____
5. H_a: _____
6. What distribution should you use for a hypothesis test?
7. Why did you choose this distribution?
8. Calculate the test statistic.
9. Find the *p*-value.
10. Sketch a graph of the situation. Label and scale the *x*-axis. Shade the area corresponding to the *p*-value.



11. State your decision.
12. State your conclusion in a complete sentence.

Discussion Questions

1. Did your data fit either distribution? If so, which?

2. In general, do you think it's likely that data could fit more than one distribution? In complete sentences, explain why or why not.

CHAPTER 11 PROJECT II: CHI-SQUARE TEST OF INDEPENDENCE

Chi-Square Test of Independence

Class Time:

Names:

Student Learning Outcomes

- The student will evaluate if there is a significant relationship between favorite type of snack and gender.

Collect the Data

1. Using your class as a sample, complete the following chart. Ask each other what your favorite snack is, then total the results.

NOTE

You may need to combine two food categories so that each cell has an expected value of at least five.

	Favorite type of snack				
	sweets (candy & baked goods)	**ice cream**	**chips & pretzels**	**fruits & vegetables**	**Total**
male					
female					
Total					

2. Looking at [link], does it appear to you that there is a dependence between gender and favorite type of snack food? Why or why not?

Hypothesis Test

Conduct a hypothesis test to determine if the factors are independent:

1. H_0: _____
2. H_a: _____
3. What distribution should you use for a hypothesis test?
4. Why did you choose this distribution?
5. Calculate the test statistic.
6. Find the *p*-value.
7. Sketch a graph of the situation. Label and scale the *x*-axis. Shade the area corresponding to the *p*-value.



8. State your decision.
9. State your conclusion in a complete sentence.

Discussion Questions

1. Is the conclusion of your study the same as or different from your answer to answer to question two under Collect the Data?
2. Why do you think that occurred?

GROUP AND PARTNER PROJECTS

Univariate Data

Student Learning Objectives

- The student will design and carry out a survey.
- The student will analyze and graphically display the results of the survey.

Instructions

As you complete each task below, check it off. Answer all questions in your summary.
____ Decide what data you are going to study.

Here are two examples, but you may **NOT** use them: number of M&M's per bag, number of pencils students have in their backpacks.

____ Are your data discrete or continuous? How do you know?
____ Decide how you are going to collect the data (for instance, buy 30 bags of M&M's; collect data from the World Wide Web).
____ Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random (using a random number generator) sampling. Do not use convenience sampling. Which method did you use? Why did you pick that method?
____ Conduct your survey. **Your data size must be at least 30.**
____ Summarize your data in a chart with columns showing **data value, frequency, relative frequency and cumulative relative frequency.**
Answer the following (rounded to two decimal places):

a. $\overline{x}$ = _____
b. s = _____
c. First quartile = _____
d. Median = _____
e. 70^{th} percentile = _____

____ What value is two standard deviations above the mean?

____ What value is 1.5 standard deviations below the mean?

____ Construct a histogram displaying your data.

____ In complete sentences, describe the shape of your graph.

____ Do you notice any potential outliers? If so, what values are they? Show your work in how you used the potential outlier formula to determine whether or not the values might be outliers.

____ Construct a box plot displaying your data.

____ Does the middle 50% of the data appear to be concentrated together or spread apart? Explain how you determined this.

____ Looking at both the histogram and the box plot, discuss the distribution of your data.

Assignment Checklist

You need to turn in the following typed and stapled packet, with pages in the following order:

- **Cover sheet**: name, class time, and name of your study
- **Summary page**: This should contain paragraphs written with complete sentences. It should include answers to all the questions above. It should also include statements describing the population under study, the sample, a parameter or parameters being studied, and the statistic or statistics produced.
- **URL** for data, if your data are from the World Wide Web
- **Chart of data, frequency, relative frequency, and cumulative relative frequency**
- **Page(s) of graphs:** histogram and box plot

Continuous Distributions and Central Limit Theorem

Student Learning Objectives

- The student will collect a sample of continuous data.
- The student will attempt to fit the data sample to various distribution models.
- The student will validate the central limit theorem.

Instructions

As you complete each task below, check it off. Answer all questions in your summary.

Part I: Sampling

____ Decide what **continuous** data you are going to study. (Here are two examples, but you may NOT use

them: the amount of money a student spent on college supplies this term, or the length of time a long-distance telephone call lasts.)

____ Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random (using a random number generator) sampling. Do not use convenience sampling. What method did you use? Why did you pick that method?

____ Conduct your survey. Gather **at least 150 pieces of continuous, quantitative data**.

____ Define (in words) the random variable for your data. X = _____

____ Create two lists of your data: (1) unordered data, (2) in order of smallest to largest.

____ Find the sample mean and the sample standard deviation (rounded to two decimal places).

a. $\overline{x}$ = _____
b. s = _____

____ Construct a histogram of your data containing five to ten intervals of equal width. The histogram should be a representative display of your data. Label and scale it.

Part II: Possible Distributions

____ Suppose that X followed the following theoretical distributions. Set up each distribution using the appropriate information from your data.

____ Uniform: $X \sim U$ _____ Use the lowest and highest values as a and b.

____ Normal: $X \sim N$ _____ Use $\overline{x}$ to estimate for μ and s to estimate for σ.

____ **Must** your data fit one of the above distributions? Explain why or why not.

____ **Could** the data fit two or three of the previous distributions (at the same time)? Explain.

____ Calculate the value k(an X value) that is 1.75 standard deviations above the sample mean. k = _____ (rounded to two decimal places) Note: $k = \overline{x} + (1.75)s$

____ Determine the relative frequencies (RF) rounded to four decimal places.

Note

$$RF = \frac{\text{frequency}}{\text{total number surveyed}}$$

a. $RF(X < k)$ = _____
b. $RF(X > k)$ = _____
c. $RF(X = k)$ = _____

Note

You should have one page for the uniform distribution, one page for the exponential distribution, and one page for the normal distribution.

_____ State the distribution: $X \sim$ _____
_____ Draw a graph for each of the three theoretical distributions. Label the axes and mark them appropriately.

_____ Find the following theoretical probabilities (rounded to four decimal places).

a. $P(X < k) =$ _____
b. $P(X > k) =$ _____
c. $P(X = k) =$ _____

_____ Compare the relative frequencies to the corresponding probabilities. Are the values close?
_____ Does it appear that the data fit the distribution well? Justify your answer by comparing the probabilities to the relative frequencies, and the histograms to the theoretical graphs.

Part III: CLT Experiments

_____ From your original data (before ordering), use a random number generator to pick 40 samples of size five. For each sample, calculate the average.
_____ On a separate page, attached to the summary, include the 40 samples of size five, along with the 40 sample averages.
_____ List the 40 averages in order from smallest to largest.
_____ Define the random variable, $\overline{X}$, in words. $\overline{X}$ = _____
_____ State the approximate theoretical distribution of $\overline{X}$. $\overline{X} \sim$ _____
_____ Base this on the mean and standard deviation from your original data.
_____ Construct a histogram displaying your data. Use five to six intervals of equal width. Label and scale it.
Calculate the value $\overline{k}$ (an $\overline{X}$ value) that is 1.75 standard deviations above the sample mean. $\overline{k}$ = _____ (rounded to two decimal places)
Determine the relative frequencies (*RF*) rounded to four decimal places.

a. $RF(\overline{X} < \overline{k}) =$ _____
b. $RF(\overline{X} > \overline{k}) =$ _____
c. $RF(\overline{X} = \overline{k}) =$ _____

Find the following theoretical probabilities (rounded to four decimal places).

a. $P(\overline{X} < \overline{k}) =$ _____

b. $P(\overline{X} > \overline{k})$ = _____
c. $P(\overline{X} = \overline{k})$ = _____

_____ Draw the graph of the theoretical distribution of X.
_____ Compare the relative frequencies to the probabilities. Are the values close?
_____ Does it appear that the data of averages fit the distribution of $\overline{X}$ well? Justify your answer by comparing the probabilities to the relative frequencies, and the histogram to the theoretical graph.
In three to five complete sentences for each, answer the following questions. Give thoughtful explanations.
_____ In summary, do your original data seem to fit the uniform, exponential, or normal distributions? Answer why or why not for each distribution. If the data do not fit any of those distributions, explain why.
_____ What happened to the shape and distribution when you averaged your data? **In theory,** what should have happened? In theory, would "it" always happen? Why or why not?
_____ Were the relative frequencies compared to the theoretical probabilities closer when comparing the X or $\overline{X}$ distributions? Explain your answer.

Assignment Checklist

You need to turn in the following typed and stapled packet, with pages in the following order:
____ **Cover sheet**: name, class time, and name of your study
____ **Summary pages**: These should contain several paragraphs written with complete sentences that describe the experiment, including what you studied and your sampling technique, as well as answers to all of the questions previously asked questions
____ **URL** for data, if your data are from the World Wide Web
____ **Pages, one for each theoretical distribution**, with the distribution stated, the graph, and the probability questions answered
____ **Pages of the data requested**
____ **All graphs required**

Hypothesis Testing Article

Student Learning Objectives

- The student will identify a hypothesis testing problem in print.
- The student will conduct a survey to verify or dispute the results of the hypothesis test.
- The student will summarize the article, analysis, and conclusions in a report.

Instructions

As you complete each task, check it off. Answer all questions in your summary.

____**Find an article** in a newspaper, magazine, or on the internet which makes a claim about **ONE** population mean or **ONE** population proportion. The claim may be based upon a survey that the article was reporting on. Decide whether this claim is the null or alternate hypothesis.

____**Copy or print out the article** and include a copy in your project, along with the source.

____**State how you will collect your data.** (Convenience sampling is not acceptable.)

____**Conduct your survey. You must have more than 50 responses in your sample.** When you hand in your final project, attach the tally sheet or the packet of questionnaires that you used to collect data. Your data must be real.

____**State the statistics** that are a result of your data collection: sample size, sample mean, and sample standard deviation, OR sample size and number of successes.

____**Make two copies of the appropriate solution sheet.**

____**Record the hypothesis test** on the solution sheet, based on your experiment. **Do a DRAFT solution** first on one of the solution sheets and check it over carefully. Have a classmate check your solution to see if it is done correctly. Make your decision using a 5% level of significance. Include the 95% confidence interval on the solution sheet.

____**Create a graph that illustrates your data.** This may be a pie or bar graph or may be a histogram or box plot, depending on the nature of your data. Produce a graph that makes sense for your data and gives useful visual information about your data. You may need to look at several types of graphs before you decide which is the most appropriate for the type of data in your project.

____**Write your summary** (in complete sentences and paragraphs, with proper grammar and correct spelling) that describes the project. The summary **MUST** include:

a. Brief discussion of the article, including the source
b. Statement of the claim made in the article (one of the hypotheses).
c. Detailed description of how, where, and when you collected the data, including the sampling technique; did you use cluster, stratified, systematic, or simple random sampling (using a random number generator)? As previously mentioned, convenience sampling is not acceptable.
d. Conclusion about the article claim in light of your hypothesis test; this is the conclusion of your hypothesis test, stated in words, in the context of the situation in your project in sentence form, as if you were writing this conclusion for a non-statistician.
e. Sentence interpreting your confidence interval in the context of the situation in your project

Assignment Checklist

Turn in the following typed (12 point) and stapled packet for your final project:

____**Cover sheet** containing your name(s), class time, and the name of your study
____**Summary**, which includes all items listed on summary checklist
____**Solution sheet** neatly and completely filled out. The solution sheet does not need to be typed.
____**Graphic representation of your data**, created following the guidelines previously discussed; include only graphs which are appropriate and useful.
____**Raw data collected AND a table summarizing the sample data** (n, $\overline{x}$ and s; or x, n, and p', as appropriate for your hypotheses); the raw data does not need to be typed, but the summary does. Hand in the data as you collected it. (Either attach your tally sheet or an envelope containing your questionnaires.)

Bivariate Data, Linear Regression, and Univariate Data

Student Learning Objectives

- The students will collect a bivariate data sample through the use of appropriate sampling techniques.
- The student will attempt to fit the data to a linear model.
- The student will determine the appropriateness of linear fit of the model.
- The student will analyze and graph univariate data.

Instructions

1. As you complete each task below, check it off. Answer all questions in your introduction or summary.
2. Check your course calendar for intermediate and final due dates.
3. Graphs may be constructed by hand or by computer, unless your instructor informs you otherwise. All graphs must be neat and accurate.
4. All other responses must be done on the computer.
5. Neatness and quality of explanations are used to determine your final grade.

Part I: Bivariate Data

Introduction

____State the bivariate data your group is going to study.

Here are two examples, but you may **NOT** use them: height vs. weight and age vs. running distance.

____Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random sampling (using a random number generator) sampling. Convenience sampling is **NOT** acceptable.

_____Conduct your survey. Your number of pairs must be at least 30.

_____Print out a copy of your data.

Analysis

_____On a separate sheet of paper construct a scatter plot of the data. Label and scale both axes.

_____State the least squares line and the correlation coefficient.

_____On your scatter plot, in a different color, construct the least squares line.

_____Is the correlation coefficient significant? Explain and show how you determined this.

_____Interpret the slope of the linear regression line in the context of the data in your project. Relate the explanation to your data, and quantify what the slope tells you.

_____Does the regression line seem to fit the data? Why or why not? If the data does not seem to be linear, explain if any other model seems to fit the data better.

_____Are there any outliers? If so, what are they? Show your work in how you used the potential outlier formula in the Linear Regression and Correlation chapter (since you have bivariate data) to determine whether or not any pairs might be outliers.

Part II: Univariate Data

In this section, you will use the data for **ONE** variable only. Pick the variable that is more interesting to analyze. For example: if your independent variable is sequential data such as year with 30 years and one piece of data per year, your *x*-values might be 1971, 1972, 1973, 1974, ..., 2000. This would not be interesting to analyze. In that case, choose to use the dependent variable to analyze for this part of the project.

_____Summarize your data in a chart with columns showing data value, frequency, relative frequency, and cumulative relative frequency.

_____Answer the following question, rounded to two decimal places:

a. Sample mean = _____
b. Sample standard deviation = _____
c. First quartile = _____
d. Third quartile = _____
e. Median = _____
f. 70th percentile = _____
g. Value that is 2 standard deviations above the mean = _____
h. Value that is 1.5 standard deviations below the mean = _____

_____Construct a histogram displaying your data. Group your data into six to ten intervals of equal width. Pick regularly spaced intervals that make sense in relation to your data. For example, do NOT group data by age as 20-26,27-33,34-40,41-47,48-54,55-61 . . . Instead, maybe use age groups 19.5-24.5, 24.5-29.5, . . . or 19.5-29.5, 29.5-39.5, 39.5-49.5, . . .

_____In complete sentences, describe the shape of your histogram.

_____Are there any potential outliers? Which values are they? Show your work and calculations as to how you used the potential outlier formula in Descriptive Statistics (since you are now using univariate data) to determine which values might be outliers.

_____Construct a box plot of your data.

_____Does the middle 50% of your data appear to be concentrated together or spread out? Explain how you determined this.

_____Looking at both the histogram AND the box plot, discuss the distribution of your data. For example: how does the spread of the middle 50% of your data compare to the spread of the rest of the data represented in the box plot; how does this correspond to your description of the shape of the histogram; how does the graphical display show any outliers you may have found; does the histogram show any gaps in the data that are not visible in the box plot; are there any interesting features of your data that you should point out.

Due Dates

- Part I, Intro: _____ (keep a copy for your records)
- Part I, Analysis: _____ (keep a copy for your records)

- Entire Project, typed and stapled: _____
  - ____ Cover sheet: names, class time, and name of your study
  - ____ Part I: label the sections "Intro" and "Analysis."
  - ____ Part II:
  - ____ Summary page containing several paragraphs written in complete sentences describing the experiment, including what you studied and how you collected your data. The summary page should also include answers to ALL the questions asked above.
  - ____ All graphs requested in the project
  - ____ All calculations requested to support questions in data
  - ____ Description: what you learned by doing this project, what challenges you had, how you overcame the challenges

Note

Include answers to ALL questions asked, even if not explicitly repeated in the items above.

GLOSSARY

"AND" Event

An outcome is in the event A AND B if the outcome is in both A AND B at the same time.

"OR" Event

An outcome is in the event A OR B if the outcome is in A or is in B or is in both A and B.

variability in samples

how much an estimate varies between samples

Analysis of variance

also referred to as ANOVA, is a method of testing whether or not the means of three or more populations are equal. The method is applicable if:

all populations of interest are normally distributed.

the populations have equal standard deviations.

samples (not necessarily of the same size) are randomly and independently selected from each population.

The test statistic for analysis of variance is the F-ratio.

average

a number that describes the central tendency of the data; there are a number of specialized averages, including the arithmetic mean, weighted mean, median, mode, and geometric mean.

Bernoulli Trial

an experiment with the following characteristics:

1. There are only two possible outcomes called “success” and “failure” for each trial.
2. The probability p of a success is the same for any trial (so the probability q = 1 − p of a failure is the same for any trial).

binomial distribution

A discrete random variable which arises from Bernoulli trials; there are a fixed number, n, of independent

trials with two outcomes called success and failure with probability p and q respectively. The binomial random variable X is the number of successes in n trials, denoted $X \sim B(n, p)$. The mean is Formula does not parse. The probability of exactly x successes in n trials is $P(X = x) = \binom{n}{x} p^x q^{n-x}$.

binomial experiment

A statistical experiment that satisfies three conditions:
1. There are a fixed number of trials, n.
2. There are only two possible outcomes, called "success" and "failure," for each trial. The letter p denotes the probability of a success on one trial, and q denotes the probability of a failure on one trial.
3. The n trials are independent and are repeated using identical conditions.

binomial formula

A random variable that counts the number of successes in a fixed number (n) of independent Bernoulli trials each with probability of a success (p)

Blinding

not telling participants which treatment a subject is receiving

box plot

a graph that gives a quick picture of the middle 50% of the data

Central Limit Theorem

Given a random variable (RV) with known mean μ and known standard deviation σ. We are sampling with size n and we are interested in two new RVs - the sample mean, $\overline{X}$, and the sample sum, ΣX.

If the size n of the sample is sufficiently large, then $\overline{X}\ N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\Sigma X\ N\left(n\mu, \sqrt{n}\sigma\right)$. If the size n of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

cluster sample

a method for selecting a random sample and dividing the population into groups (clusters); use simple random sampling to select a set of clusters. Every individual in the chosen clusters is included in the sample.

complement of event

The complement of event A consists of all outcomes that are NOT in A.

conditional probability

the likelihood that an event will occur given that another event has already occurred

confidence intervals

an interval estimate for an unknown population parameter. This depends on: (1) the desired confidence level, (2) information that is known about the distribution (for example, known standard deviation), and (3) the sample and its size.

confidence level

the percent expression for the probability that the confidence interval contains the true population parameter; for example, if the CL = 90%, then in 90 out of 100 samples the interval estimate will enclose the true population parameter.

contingency tables

the method of displaying a frequency distribution as a table with rows and columns to show how two variables may be dependent (contingent) upon each other; the table provides an easy way to calculate conditional probabilities.

continuity correction

When statisticians add or subtract .5 to values to improve approximation

continuous

a random variable (RV) whose outcomes are measured; the height of trees in the forest is a continuous RV.

control group

a group in a randomized experiment that receives an inactive treatment but is otherwise managed exactly as the other groups

Convenience sampling

a nonrandom method of selecting a sample; this method selects individuals that are easily accessible and may result in biased data.

correlation coefficient

a measure developed by Karl Pearson (early 1900s) that gives the strength of association between the independent variable and the dependent variable; the formula is:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}}$$

where n is the number of data points. The coefficient cannot be more than 1 and less than –1. The closer the coefficient is to ±1, the stronger the evidence of a significant linear relationship between x and y.

Cumulative relative frequency

The term applies to an ordered set of observations from smallest to largest. The cumulative relative frequency is the sum of the relative frequencies for all values that are less than or equal to the given value.

data

a set of observations (a set of possible outcomes); most data can be put into two groups: qualitative (an attribute whose value is indicated by a label) or quantitative (an attribute whose value is indicated by a number). Quantitative data can be separated into two subgroups: discrete and continuous. Data is discrete if it is the result of counting (such as the number of students of a given ethnic group in a class or the number of books on a shelf). Data is continuous if it is the result of measuring (such as distance traveled or weight of luggage)

Datum

a piece of information

decay parameter

The decay parameter describes the rate at which probabilities decay to zero for increasing values of x. It is the value m in the probability density function f(x) = me(-mx) of an exponential random variable. It is also equal to m = $\frac{1}{\mu}$, where μ is the mean of the random variable.

degrees of freedom

the number of objects in a sample that are free to vary

dependent

If two events are NOT independent, then we say that they are dependent.

discrete

a random variable (RV) whose outcomes are counted

double-blind experiment

the act of blinding both the subjects of an experiment and the researchers who work with the subjects

Empirical Rule

Roughly 68% of values are within 1 standard deviation of the mean, roughly 95% of values are within 2 standard deviations of the mean, and 99.7% of values are within 3 standard deviations of the mean

equally likely

Each outcome of an experiment has the same probability.

Error Bound (EBM)

the margin of error; depends on the confidence level, sample size, and known or estimated population standard deviation.

Error Bound (EBP)

the margin of error; depends on the confidence level, sample size, and the estimated (from the sample) proportion of successes.

event

a subset of the set of all outcomes of an experiment; the set of all outcomes of an experiment is called a sample space and is usually denoted by S. An event is an arbitrary subset in S. It can contain one outcome, two outcomes, no outcomes (empty subset), the entire sample space, and the like. Standard notations for events are capital letters such as A, B, C, and so on.

expected value

expected arithmetic average when an experiment is repeated many times; also called the mean. Notations: μ. For a discrete random variable (RV) with probability distribution function P(x),the definition can also be written in the form $\mu = \sum xP(x)$.

experiment

a planned activity carried out under controlled conditions

experimental units

any individual or object to be measured

explanatory variable

the independent variable in an experiment; the value controlled by researchers

Exponential Distribution

a continuous random variable (RV) that appears when we are interested in the intervals of time between some random events, for example, the length of time between emergency arrivals at a hospital; the notation is $X \sim Exp(m)$. The mean is $\mu = 1m$ and the standard deviation is $\sigma = 1m$. The probability density function is $f(x) = me-mx$, $x \geq 0$ and the cumulative distribution function is $P(X \leq x) = 1 - e-mx$.

first quartile

the value that is the median of the of the lower half of the ordered data set

frequency

the number of times a value of the data occurs

frequency polygon

looks like a line graph but uses intervals to display ranges of large amounts of data

frequency table

a data representation in which grouped data is displayed along with the corresponding frequencies

geometric distribution

a discrete random variable (RV) that arises from the Bernoulli trials; the trials are repeated until the first success. The geometric variable X is defined as the number of trials until the first success. Notation: X ~ G(p). The mean is $\mu = \frac{1}{p}$ and the standard deviation is $\sigma = \sqrt{\frac{1}{p}\left(\frac{1}{p} - 1\right)}$. The probability of exactly x failures before the first success is given by the formula: P(X = x) = p(1 – p)^(x – 1)

geometric experiment

a statistical experiment with the following properties:

1. There are one or more Bernoulli trials with all failures except the last one, which is a success.
2. In theory, the number of trials could go on forever. There must be at least one trial.
3. The probability, p, of a success and the probability, q, of a failure do not change from trial to trial.

histogram

a graphical representation in x-y form of the distribution of data in a data set; x represents the data and y represents the frequency, or relative frequency. The graph consists of contiguous rectangles.

hypergeometric

hypergeometric experiment

a statistical experiment with the following properties:

1. You take samples from two groups.
2. You are concerned with a group of interest, called the first group.
3. You sample without replacement from the combined groups.
4. Each pick is not independent, since sampling is without replacement.
5. You are not dealing with Bernoulli Trials.

hypergeometric probability distribution

a discrete random variable (RV) that is characterized by:
1. A fixed number of trials.
2. The probability of success is not the same from trial to trial.
We sample from two groups of items when we are interested in only one group. X is defined as the number of successes out of the total number of items chosen. Notation: X ~ H(r, b, n), where r = the number of items in the group of interest, b = the number of items in the group not of interest, and n = the number of items chosen.

hypothesis

a statement about the value of a population parameter, in case of two hypotheses, the statement assumed to be true is called the null hypothesis (notation H0) and the contradictory statement is called the alternative hypothesis (notation Ha).

independent

The occurrence of one event has no effect on the probability of the occurrence of another event. Events A and B are independent if one of the following is true:
• P(A|B) = P(A)
• P(B|A) = P(B)
• P(A AND B) = P(A)P(B)

inferential statistics

also called statistical inference or inductive statistics; this facet of statistics deals with estimating a population parameter based on a sample statistic. For example, if four out of the 100 calculators sampled are defective we might infer that four percent of the production is defective.

informed consent

Any human subject in a research study must be cognizant of any risks or costs associated with the study. The subject has the right to know the nature of the treatments included in the study, their potential risks, and their potential benefits. Consent must be given freely by an informed, fit participant.

Institutional Review Boards (IRB)

a committee tasked with oversight of research programs that involve human subjects

interquartile range

or IQR, is the range of the middle 50 percent of the data values; the IQR is found by subtracting the first quartile from the third quartile.

intersection

The shared outcomes of two events

intervals

also called a class interval; an interval represents a range of data and is used when displaying large data sets

Law of Large Numbers

As the number of trials in a probability experiment increases, the difference between the theoretical probability of an event and the relative frequency probability approaches zero.

lurking variables

a variable that has an effect on a study even though it is neither an explanatory variable nor a response variable

mean

a number that measures the central tendency; a common name for mean is "average." The term "mean" is a shortened form of "arithmetic mean." By definition, the mean for a sample (denoted by $\overline{x}$) is \overline{x}\text{ = }\frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}
, and the mean for a population (denoted by μ) is \mu
\text{ = }\frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}
.

mean of the distribution

the long-term average of many trials of a statistical experiment

median

a number that separates ordered data into halves; half the values are the same number or smaller than the

median and half the values are the same number or larger than the median. The median may or may not be part of the data.

memoryless property

For an exponential random variable X, the memoryless property is the statement that knowledge of what has occurred in the past has no effect on future probabilities. This means that the probability that X exceeds x + k, given that it has exceeded x, is the same as the probability that X would exceed k if we had no knowledge about it. In symbols we say that $P(X > x + k|X > x) = P(X > k)$.

midpoint

the mean of an interval in a frequency table

mode

the value that appears most frequently in a set of data

mutually exclusive events

Two events are mutually exclusive if the probability that they both happen at the same time is zero. If events A and B are mutually exclusive, then P(A AND B) = 0.

nonsampling errors

an issue that affects the reliability of sampling data other than natural variation; it includes a variety of human errors including poor study design, biased sampling methods, inaccurate information provided by study participants, data entry errors, and poor analysis.

normal distribution

a continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$, where μ is the mean of the distribution and σ is the standard deviation, notation: $X \sim N(\mu,\sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called the standard normal distribution.

Numerical variables

variables that take on values that are indicated by numbers

one-way ANOVA

a method of testing whether or not the means of three or more populations are equal; the method is applicable if:

all populations of interest are normally distributed.

the populations have equal standard deviations.

samples (not necessarily of the same size) are randomly and independently selected from each population.

The test statistic for analysis of variance is the F-ratio.

outcome

a particular result of an experiment

outliers

an observation that does not fit the rest of the data

p-value

the probability that an event will happen purely by chance assuming the null hypothesis is true. The smaller the p-value, the stronger the evidence is against the null hypothesis.

parameter

a numerical characteristic of a population

percentiles

a number that divides ordered data into hundredths; percentiles may or may not be part of the data. The median of the data is the second quartile and the 50th percentile. The first and third quartiles are the 25th and the 75th percentiles, respectively.

placebo treatment

an inactive treatment that has no real effect on the explanatory variable

point estimate

a single number computed from a sample and used to estimate a population parameter.

Poisson distribution

A discrete random variable that counts the number of times a certain event will occur in a specific interval; characteristics of the variable:

• The probability that the event occurs in a given interval is the same for all intervals.

• The events occur with a known mean and independently of the time since the last event.

The distribution is defined by the mean μ of the event in the interval. Notation: X ~ P(μ). The mean is μ = np. The standard deviation is $\sigma = \sqrt{\mu}$. The probability of having exactly x successes in r trials is P(X = x) = $\left(e^{-\mu}\right) \frac{\mu^x}{x!}$. The Poisson distribution is often used to approximate the binomial distribution, when n is "large" and p is "small" (a general rule is that n should be greater than or equal to 20 and p should be less than or equal to 0.05).

Poisson probability distribution

A discrete random variable that counts the number of times a certain event will occur in a specific interval; characteristics of the variable:

• The probability that the event occurs in a given interval is the same for all intervals.

• The events occur with a known mean and independently of the time since the last event.

The distribution is defined by the mean μ of the event in the interval. Notation: X ~ P(μ). The mean is μ = np. The standard deviation is $\sigma = \sqrt{\mu}$. The probability of having exactly x successes in r trials is P(X = x) = $\left(e^{-\mu}\right) \frac{\mu^x}{x!}$. The Poisson distribution is often used to approximate the binomial distribution, when n is "large" and p is "small" (a general rule is that n should be greater than or equal to 20 and p should be less than or equal to 0.05).

Probability

a number between zero and one, inclusive, that gives the likelihood that a specific event will occur; the foundation of statistics is given by the following 3 axioms (by A.N. Kolmogorov, 1930's): Let S denote the sample space and A and B are two events in S. Then:

• 0 ≤ P(A) ≤ 1

• If A and B are any two mutually exclusive events, then P(A OR B) = P(A) + P(B).

• P(S) = 1

probability density function

A function that defines a continuous random variable, and the likelihood of an outcome

probability experiment

A random experiment where the result is not predetermined

proportion

a part, share, or number considered in comparative relation to a whole

Qualitative

an attribute whose value is indicated by a label

quantile

Points in a distribution that relate to the rank order of values in that distribution

Quantitative

an attribute whose value is indicated by a number

quartiles

the numbers that separate the data into quarters; quartiles may or may not be part of the data. The second quartile is the median of the data.

random assignment

the act of organizing experimental units into treatment groups using random methods

random sampling

a method of selecting a sample that gives every member of the population an equal chance of being selected.

Random Variable

A probability model quantifying a situation; can equal the values of the possible outcomes. Represented using a capital letter.

relative frequency

the ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes to the total number of outcomes

representative sample

A representative sample is a subset of a population that seeks to accurately reflect the characteristics of the larger group

response variable

the dependent variable in an experiment; the value that is measured for change at the end of an experiment

sample space

the set of all possible outcomes of an experiment

sampling bias

not all members of the population are equally likely to be selected

sampling errors

the natural variation that results from selecting a sample to represent a larger population; this variation decreases as the sample size increases, so selecting larger samples reduces sampling error.

simple random sampling

a straightforward method for selecting a random sample; give each member of the population a number. Use a random number generator to select a set of labels. These randomly selected labels identify the members of your sample.

skewed

used to describe data that is not symmetrical; when the right side of a graph looks "chopped off" compared to the left side, we say it is "skewed to the left." When the left side of the graph looks "chopped off" compared to the right side, we say the data is "skewed to the right." Alternatively: when the lower values of the data are more spread out, we say the data are skewed to the left. When the greater values are more spread out, the data are skewed to the right.

standard deviation

a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: s for sample standard deviation and σ for population standard deviation

standard deviation of a probability distribution

a number that measures how far the outcomes of a statistical experiment are from the mean of the distribution

standard error of the mean

the standard deviation of the distribution of the sample means, or $\frac{\sigma}{\sqrt{n}}$.

standard normal distribution

a continuous random variable (RV) with a mean of 0 and a standard deviation of 1 which z-scores follow: X ~ N(0, 1); when X follows the standard normal distribution, it is often noted as Z ~ N(0, 1).

statistics

the practice or science of collecting and analyzing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample.

stratified sample

a method for selecting a random sample used to ensure that subgroups of the population are represented adequately; divide the population into groups (strata). Use simple random sampling to identify a proportionate number of individuals from each stratum.

Student's t-distribution

investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student; the major characteristics of the random variable (RV) are: (1)It is continuous and assumes any real values; (2) The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution; (3) It approaches the standard normal distribution as n gets larger. There is a "family of t–distributions: each representative of the family is completely defined by the number of degrees of freedom, which is one less than the number of data.

systematic sample

a method for selecting a random sample; list the members of the population. Use simple random sampling to select a starting point in the population. Let k = (number of individuals in the population)/(number of individuals needed in the sample). Choose every kth individual in the list

starting with the one that was randomly selected. If necessary, return to the beginning of the population list to complete your sample.

treatments

different values or components of the explanatory variable applied in an experiment

tree diagram

the useful visual representation of a sample space and events in the form of a "tree" with branches marked by possible outcomes together with associated probabilities (frequencies, relative frequencies)

Type I Error

The decision is to reject the null hypothesis when, in fact, the null hypothesis is true.

Type II Error

The decision is not to reject the null hypothesis when, in fact, the null hypothesis is false.

Uniform Distribution

a continuous random variable (RV) that has equally likely outcomes over the domain, a < x < b; it is often referred to as the rectangular distribution because the graph of the pdf has the form of a rectangle. Notation: X ~ U(a,b). The mean is $\mu = \frac{a+b}{2}$ and the standard deviation is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$. The probability density function is $f(x) = \frac{1}{b-a}$ for a < x < b or a ≤ x ≤ b. The cumulative distribution is $P(X \leq x) = \frac{x-a}{b-a}$.

union

The set of all outcomes in two events; outcomes can be in either or both events

variance

mean of the squared deviations from the mean; the square of the standard deviation. For a set of data, a deviation can be represented as $x - \overline{x}$ where x is a value of the data and $\overline{x}$ is the sample mean. The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and one

variation

mean of the squared deviations from the mean, or the square of the standard deviation; for a set of data, a deviation can be represented as x – $\overline{x}$ where x is a value of the data and $\overline{x}$ is the sample mean. The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and one.

Venn diagram

the visual representation of a sample space and events in the form of circles or ovals showing their intersections

with replacement

Once a member of the population is selected for inclusion in a sample, that member is returned to the population for the selection of the next individual.

without replacement

A member of the population may be chosen for inclusion in a sample only once. If chosen, the member is not returned to the population before the next selection.

z-scores

the linear transformation of the form z = \frac{x\text{ }–\text{ }\mu }{\sigma }; if this transformation is applied to any normal distribution X ~ N(μ, σ) the result is the standard normal distribution Z ~ N(0,1). If this transformation is applied to any specific value x of the RV with mean μ and standard deviation σ, the result is called the z-score of x. The z-score allows us to compare data that are normally distributed but scaled differently.

SUMMARY OF ADAPTATIONS

Full Resource Adaptations

Across all chapters of the resource, the following changes were made (where necessary):

- Edited all mathematical notations to have them appear in LaTex
- Created and added H5P activities for select content and exercises
  - Some H5P activities were added from *Significant Statistics: An Introduction to Statistics*
- Added solution drop-downs (HTML or Accordion H5P) for Exercises and Your Turn
- Added embedded videos from YouTube
  - Other videos added from *Introduction to Statistics: An Excel-Based Approach (7.1 (one video), 7.2 (two videos), 7.3 (two videos), 8.1 (one video), 8.2 (one video), 8.5 (five videos), 9.1 (two videos), 9.2 (two videos), 9.3 (one video))*
- Removed Chapter Objectives from Introduction pages
- Added Learning Objectives for each Section/Chapter
- Enabled glossary term functionality
- Consistent colored text boxes throughout
- Added Heading 1 to main parts for Table of Contents
- Minor removal and reordering of content for flow and clarity
  - Try-It was renamed to "Your Turn" (like in Significant Statistics)
  - Merged Chapter 6 (The Normal Distribution) and Chapter 7 (The Central Limit Theorem) into one Part
  - Moved Collaborative Exercises to Introduction pages
  - Made Chapter Review, Practice, and Homework into their own Chapters
  - Moved end of Chapter Projects into own their own Part titled Class Group Activities/Projects
  - Chapter 12: F Distribution and One-Way ANOVA – removed from this version of book
- Adjusted language for clarity
- Images on Introductions captioned for accessibility – same captions provided from source; alt-tag used as provided from source

Chapter-Specific Changes

Chapter 3: Probability Topics

- 3.1: Updated last example to reference freshman course enrollment
- The LaTeX formula within the conditional statement was adjusted.

Chapter 4: Discrete Random Variables

- Introduction: Changed example for long-distance phone calls to emails
- Moved "Random Variables" heading and content from the Introduction to Section 4.1 Probability Distribution Function (PDF) for a Discrete Random Variable
- Section 4.1 and 4.2: Changed to neutral pronouns for parents, students, and the university soccer team
- Section 4.2: Changed betting on earthquakes in Japan example to purchasing travel insurance
- Section 4.3: Added text regarding Binomial PDF from Significant Statistics
- Section 4.3: Removed example regarding cancer rates to shorten section slightly, added new figure for Workers example

- The LaTeX formula within the mean or expected value was adjusted.

Chapter 5: Continuous Random Variables

- Moved "Properties of Continuous Probability Distributions" heading and content from the Introduction to Section 5.1 Continuous Probability Functions

Chapter 6: The Normal Distribution and The Central Limit Theorem

- Moved "Normal Distribution parameters" content from the Introduction to Section 6.1 The Standard Normal Distribution
- Content brought in from 5.2 Normal Distribution in Significant Statistics for use in our Sections 6.1 and 6.2
- Removed "For that reason, as well as the fact that there are various table formats, we are not including table instructions." from 6.1
    - Added my own information on using the Z-Score Table in 6.2 (in text and solutions for examples)

    - Added Z Table link provided by OpenStax from Back Matter in 6.2
    - Added section just for using the technology (graphing calculator)
- Rearranged examples and try-its in 6.1 and 6.2
- Rearranged content for better flow
- Images in 6.1, 6.2, 6.3, and 6.5 from Significant Statistics
- H5P Solutions brought in from Significant Statistics for 6.1, 6.2, 6.3, and 6.5
- Added content from 6.2 Significant Statistics to to 6.3 for CLT with Sample Mean
- Section 6.5 (which was Section 7.3 in Original Text):
    - Moved examples into Sections 6.3 and 6.4 for more CLT practice, removed other CLT content in beginning
    - Changed Section to be just The Normal Approximation to the Binomial – took most content from Significant Statistics Section 5.3
    - Added my own content in places, to fill in gaps.
- Videos added from Khan Academy and OpenIntroOrg

Class Group Activities/Projects

- Titled each to contain "Chapter # Project" for quick reference
- Edited tables to have captions for accessibility
- Group and Partner Projects (from Appendix) made into separate projects.