Інститут прикладного системного аналізу
НАН України та Міністерства освіти і науки України

П. І. Бідюк, О. М. Терентьєв, Т. І. Просянкіна-Жарова

ПРИКЛАДНА СТАТИСТИКА

Навчальний посібник

Вінниця
ПП «ТД»Едельвейс і К»
2013

УДК (519.248: 004)(075.8)
ББК (22.172:32.97)я731
Б 59

Рекомендовано до друку Методичною Радою Інституту Прикладного Системного Аналізу НАН та Міністерства освіти і науки України (протокол № 2 від 3 жовтня 2013 р.)

Рецензенти: **В. М. Азарсков**, д. т. н, професор, директор інституту аерокосмічних систем керування Національного Авіаційного Університету, Лауреат державної премії в галузі науки і техніки, Лауреат премії імені академіка М.К. Янгеля НАН України;
С. Ф. Теленік, д. т. н., професор, завідувач кафедри автоматики та управління в технічних системах Національного Технічного Університету України "Київський Політехнічний Інститут;

Б 59 Бідюк, П. І. Прикладна статистика / П. І. Бідюк, О. М. Терентьєв, Т. І. Просянкіна-Жарова. – Вінниця : ПП "ТД"Едельвейс і К", 2013. – 304 с.
ISBN 978-966-2462-21-0

Пропонується висвітлення сучасних методів статистичного аналізу даних, представлених часовими перерізами або часовими рядами. Наведено узагальнену методику збору і статистичного аналізу даних у вигляді чотирьох етапів. Розглянуто задачі обчислення описових статистик, опису положення окремого спостереження в ряду розподілу і методику групування даних. Докладно розглянуто нормальний розподіл даних, його статистичні характеристики, можливості використання таблиць нормального розподілу, а також розподіл вибіркових середніх. Наведено аналіз процесу прийняття статистичних рішень, поняття ризику і його моделювання, формулювання і застосування процедур перевірки гіпотез. Розглянуто деякі стандартні процедури перевірки гіпотез та інтервальне оцінювання. Наведено основи регресійного моделювання процесів довільної природи за багатокроковою методикою. Запропоновано методику діагностування побудованих регресійних моделей з прикладами їх побудови за реальними статистичними даними.

Окремий розділ відведено використанню сучасної системи статистичного аналізу даних SAS для розв'язання задач поглибленої статистичної обробки, дослідженню розподілів, змінних аналізу та використанню ANOVA аналізу.

Наведено приклади застосування методів обробки статистичних даних до розв'язування реальних практичних задач.

Для студентів, аспірантів та викладачів, а також інженерів, які спеціалізуються в галузі розв'язання задач статистичного аналізу даних, математичного моделювання і прогнозування динаміки фінансово-економічних та процесів іншої природи, представлених статистичними даними.

УДК (519.248: 004)(075.8)
ББК (22.172:32.97)я731

ISBN 978-966-2462-21-0

ЗМІСТ

Стор.

Вступ

Розділ 1 Чотири етапи статистичного аналізу

1.1 Задачі, які розв'язують методами прикладної статистики . . . . . .
1.2 Планування збору даних . . . . . .
1.3 Попередня обробка і дослідження даних . . . . . .
1.4 Оцінювання параметрів статистичних і математичних моделей . . . .
1.5 Перевірка сформульованих гіпотез . . . . . .
1.6 Суцільний та вибірковий методи збору даних . . . . . .
1.7 Зловживання статистикою . . . . . .
1.8 Контрольні запитання і вправи . . . . . .

Розділ 2 Мода, медіана, середнє і варіація

2.1 Означення . . . . . .
2.2 Дві властивості середнього . . . . . .
2.3 Вплив зміни значень ряду розподілу на середнє . . . . . .
2.4 Деякі приклади застосування середнього, медіани і моди . . . . . .
2.5 Позначення та обчислення середнього арифметичного . . . . . .
2.6 Обчислення поточного середнього . . . . . .
2.7 Математичне сподівання дискретної випадкової змінної . . . . . .
2.8 Інші види середнього . . . . . .
2.9 Варіація . . . . . .
2.10 Міри варіації: середнє відхилень та дисперсія . . . . . .
2.11 Обчислення незміщеної оцінки дисперсії . . . . . .
2.12 Властивості дисперсії . . . . . .
2.13 Стандартне або середнє квадратичне відхилення . . . . . .
2.14 Вплив зміни значень елементів ряду на дисперсію . . . . . .
2.15 Застосування дисперсії . . . . . .
2.16 Однаково розподілені взаємно незалежні випадкові величини . . . . .
2.17 Початкові і центральні моменти . . . . . .
2.18 Поняття групових і загальних статистичних характеристик . . . . . .
2.19 Контрольні запитання і вправи . . . . . .

Розділ 3 Опис положення окремого спостереження в ряду розподілу, групування даних

3.1 Процентильні ранги і процентилі . . . . . .
3.2 Нормовані відхилення (z-оцінки) . . . . . .
3.3 Середні і z-оцінки . . . . . .
3.4 Порівняння z-оцінок і процентилів . . . . . .
3.5 Середнє і стандартне відхилення ряду, утвореного із z-оцінок . . . . . .
3.6 Стандартні або *T*-оцінки . . . . . .
3.7 Застосування процентильних рангів і стандартних оцінок . . . . . .
3.8 Часові ряди та часові перерізи даних . . . . . .
3.9 Дискретні і неперервні величини . . . . . .
3.10 Табулювання і графічне зображення дискретних величин . . . . . .
3.11 Округлювання значень . . . . . .
3.12 Групування даних і побудова групової таблиці частот . . . . . .
3.13 Гістограма і полігон ряду розподілу . . . . . .
3.14 Застосування стовпчикових діаграм і гістограм . . . . . .

3.15 Контрольні запитання і вправи . . . . . .

Розділ 4 Статистичні параметри згрупованих даних
4.1 Визначення медіани і процентилів з гістограми . . . . . .
4.2 Обчислення медіани та процентилів . . . . . .
4.3 Квартилі і децилі . . . . . .
4.4 Кумулятивна крива . . . . . .
4.5 Визначення моди згрупованих даних . . . . . .
4.6 Визначення середнього згрупованих даних . . . . . .
4.7 Обчислення дисперсії і стандартного відхилення для згрупованих даних
4.8 Контрольні запитання і вправи . . . . . .

Розділ 5 Нормальний розподіл
5.1 Теоретичні розподіли . . . . . .
5.2 Нормальний розподіл і його графік . . . . . .
5.3 Чотири властивості нормального розподілу . . . . . .
5.4 Функція розподілу і числові характеристики неперервних величин . .
5.5 Статистичні характеристики нормального розподілу . . . . . .
5.6 Дослідження кривої нормального розподілу . . . . . .
5.7 Вплив параметрів нормального розподілу на форму нормальної кривої
5.8 Ймовірність попадання випадкової нормальної величини у заданий інтервал
5.9 Ймовірність отримання випадковою нормальною величиною заданого значення відхилення
5.10 Правило трьох сигм . . . . . .
5.11 Оцінка відхилення теоретичного розподілу від нормального; асиметрія і ексцес
5.12 Використання таблиці площ нормального розподілу . . . . . .
5.13 Поняття про теорему Ляпунова. Формулювання центральної граничної теореми
5.14 Властивості ряду розподілу середніх . . . . . .
5.15 Процентильний ранг і z-оцінки вибіркового розподілу середніх . . .
5.16 Ймовірність і нормальний розподіл . . . . . .
5.17 Квантилі нормованого розподілу . . . . . .
5.18 Контрольні запитання і вправи . . . . . .

Розділ 6 Аналіз процесу прийняття статистичних рішень, ризик і перевірка гіпотез
6.1 Роль теорії ймовірностей і математичної статистики в прийнятті рішень . . . . . .
6.2 Основні принципи перевірки гіпотез . . . . . .
6.3 Статистична перевірка гіпотез . . . . . .
6.4 Перевірка гіпотез в задачах трьох типів . . . . . .
6.5 Зауваження стосовно термінології . . . . . .
6.6 Односторонні критерії . . . . . .
6.7 Приклади застосування процедур перевірки гіпотез . . . . . .
6.8 Альтернативне формулювання процедури перевірки гіпотез . . . . . .
6.9 Методи побудови процедури перевірки гіпотез . . . . . .
6.10 Зауваження щодо односторонньої і двосторонньої перевірки гіпотез

6.11 Оцінювання надійності правил перевірки гіпотез . . . . . .
6.12 Контрольні запитання і вправи . . . . . .

Розділ 7 Деякі стандартні процедури перевірки гіпотез та інтервальне оцінювання
7.1 Перевірка гіпотез щодо середніх . . . . . .
7.2 Поняття інтервальних оцінок . . . . . .
7.3 Інтервальна оцінка середнього значення розподілу . . . . . .
7.4 Контрольні запитання і вправи . . . . . .

Розділ 8 Побудова регресійних моделей
8.1 Методика побудови математичних моделей часових рядів
8.2 Аналіз процесу . . . . . .
8.3 Попередня обробка даних . . . . . .
8.4 Аналіз рядів на наявність нелінійностей . . . . . .
8.5 Формування інших елементів структури моделі . . . . . .
8.6 Оцінювання коефіцієнтів моделей-кандидатів . . . . . .
8.7 Діагностика моделі – вибір кращої з множини кандидатів . . . . . .
8.8 Приклади побудови моделей за статистичними даними . . . . . .
8.9 Контрольні запитання і вправи . . . . . .

Розділ 09 SAS – сучасний інструмент статистичної обробки даних
9.1 Історичні відомості щодо створення системи SAS . . . . . .
9.2 Класифікація світових брендів бізнес інтелектуальних платформ за рейтингом Gartner і місце SAS у ньому
9.3 Короткий огляд продуктів системи SAS . . . . . .
9.4 Базові процедури статистичної обробки даних в SAS . . . . . .
9.5 Вимоги до даних для обчислення базових характеристик . . . . . .
9.6 Створення набору даних в системі SAS . . . . . .
9.7 Дослідження розподілів за допомогою PROC UNIVARIATE . . . . . .
9.8 Обчислення статистик за допомогою PROC MEANS . . . . . .
9.9 Дослідження категоріальних даних з використанням PROC ANOVA .
9.10 Дослідження кореляції за процедурою PROC CORR . . . . . .
9.11 Використання PROC REG для однофакторногорегресійного аналізу .
9.12 Використання PROC ANOVA для однофакторного дисперсійного аналізу
9.13 Досвід використання програмного забезпечення SAS у ННК ІПСА НТУУ КПІ . . . . . .
9.14 Контрольні запитання і вправи . . . . . .
9.15 Відповіді до контрольних запитань і вправ

Перелік посилань

ПЕРЕЛІК СКОРОЧЕНЬ

АКФ – автокореляційна функція
АР – авторегресія
АРКС – авторегресія з ковзним середнім
АРІКС – авторегресія з інтегрованим ковзним середнім
АРУГ – авторегресія з умовною гетероскедастичністю
ВП – випадковий процес
ВВП – валовий внутрішній продукт
ЕПП – економіка перехідного періоду
ІМ – імітаційна модель
МАП – максимальна абсолютна похибка
МВК – модель випадкового кроку
МіАП – мінімальна абсолютна похибка
МДВ – модель двійкового вибору
МКМЛ – метод Монте- Карло для марковських ланцюгів
МКП – модель корегування похибки
ММП – метод максимальної правдоподібності
МНК – метод найменших квадратів
МОП – метод опорних векторів
МР – множинна регресія
МС – математичне сподівання
НКФ – нелінійна кореляційна функція
НМНК – нелінійний метод найменших квадратів
ПС – простір станів
РМНК – рекурсивний метод найменших квадратів
САП – середнє абсолютне значення похибки
СЕ – сезонні ефекти
САПП – середня абсолютна похибка у відсотках
СКП – середня квадратична похибка
ЧАКФ – часткова автокореляційна функція
ETS – Econometrics and Time-Series analysis
GIS – Geographic Information System
IML – Interactive Matrix Language
OR – Operative Research
SAS – Statistical Analysis System

ВСТУП

Статистичний аналіз даних виконується з метою побудови прогнозуючих моделей та прийняття рішень на основі оцінок прогнозів. Його застосовують у соціальних, фінансово-економічних, екологічних, технічних, технологічних та інших системах для коротко- і середньострокового прогнозування процесів ціноутворення, обсягів виробництва та накопичення продукції на складах, оцінювання альтернативних економічних стратегій, формування бюджетів підприємств, галузей і держави в цілому, оцінювання, прогнозування та менеджменту ризиків довільної природи, а також для розв'язання багатьох інших практичних завдань. Сьогодні методи статистичного аналізу даних використовуються в усіх без винятку сферах діяльності.

У спеціальній літературі описано широку множину методів аналізу статистичних даних з метою обробки даних різних типів, прогнозування лінійних стаціонарних та нелінійних нестаціонарних процесів на основі використання даних у вигляді часових рядів. Найбільш поширеними серед них є лінійна та нелінійна множинна регресія, квантильна регресія, регресійні дерева, метод групового урахування аргументів, нейромережі, байєсівські моделі і мережі, нечіткі множини, нечіткі нейромережі та інші. В деякій мірі "універсальними" методами моделювання і прогнозування можна вважати метод групового врахування аргументів (МГВА) і нечіткі нейромережі. Однак практика показує, що одного, навіть відносно універсального методу, як правило, недостатньо для досягнення повноти аналізу досліджуваного процесу і прийняття коректного рішення.

Кожний метод аналізу даних має свої недоліки і переваги стосовно обчислювальних витрат, характеристик точності моделей і оцінок прогнозів. Так, висока точність прогнозу за допомогою МГВА або нейромережі іноді досягається за рахунок значних обчислювальних витрат, які можуть виявитись неприйнятними в окремих випадках. Це особливо стосується застосування отриманої моделі у системі керування реального часу, де модель необхідна для оцінювання прогнозу і синтезу керуючого впливу. Суттєвий виграш стосовно обчислювальних витрат можна досягти в такому випадку за допомогою набагато простішої моделі авторегресії з ковзним середнім, перевагами якої є простота оцінювання структури і параметрів та можливості її оперативної адаптації до характеристик процесу у реальному часі.

Метою даної роботи є: (1) – представлення основних етапів збору і статистичного аналізу даних; (2) – обчислення описових статистик: мода, медіана, середнє, варіація, дисперсія; (3) – опис положення окремого спостереження в ряду розподілу: процентилі, *z*-оцінки і *T*-оцінки; (4) – аналіз згрупованих даних; (5) – аналіз процесу прийняття статистичних рішень; (6) – представлення елементів теорії перевірки гіпотез; (7) – аналіз нормального розподілу та можливостей його практичного використання; (8) – формулювання багатокрокової методики побудови математичних (регресійних) моделей на основі статистичних даних; (9) – опис системи SAS як високорозвиненого сучасного інструменту статистичної обробки даних.

В цілому економетричні статистичні дані, а також дані з інших сфер діяльності містять множину структурних компонент, математичний опис яких потребує застосування сучасних методів аналізу, наведених у посібнику.

Автори сподіваються, що книга буде корисною для студентів, аспірантів, інженерів, викладачів та всіх, хто займається практичними задачами поглибленого статистичного аналізу даних, математичного моделювання і прогнозування на основі статистичних даних у різних галузях техніки, економіки, фінансів, екології, соціальних дослідженнях і т. ін.

Розділ 1

Чотири етапи статистичного аналізу

Прикладна статистика – це мистецтво і наука збору, обробки даних і аналізу отриманих результатів з метою прийняття коректних обґрунтованих рішень технічного, ділового, політичного, персонального або іншого характеру.

Статистичні методи необхідно розглядати як важливу трудомістку частину процесу прийняття рішень, яка дає можливість приймати обґрунтовані тактичні і стратегічні рішення. Ці методи ґрунтуються на знаннях та інтуїції фахівців і глибокому аналізі наявної інформації.

Коректне використання статистичних методів надає значні переваги практично в усіх напрямах людської діяльності: в конкуренції за підвищення якості та зростання продаж нової продукції, дає можливість отримати високоякісні оцінки прогнозів, обґрунтувати мікро- і макроекономічні рішення, а також рішення стосовно раціонального ведення домашнього господарства і розв’язати багато інших задач, у тому числі персонального характеру.

1.1 Задачі, які розв’язують методами прикладної статистики

1. Аналіз стану ринку і прийняття рішення стосовно номенклатури та об’єму випуску продукції виробничого підприємства.
2. Аналіз, прогнозування і управління соціально-економічними процесами і системами (регіональний, галузевий та державний рівні).
3. Управління якістю продукції на виробництві.
4. Управління якістю навчання на всіх рівнях підготовки персоналу.
5. Автоматичне керування технологічними процесами і технічними системами - побудова адекватних моделей, застосування методів статистичного керування, прогнозування розвитку процесів і аналіз якості керування за допомогою множини відповідних статистичних параметрів.
6. Прогнозування і планування розвитку процесів різної природи на всіх рівнях ієрархії прийняття управлінських рішень.
7. Статистична підтримка прийняття експертних рішень з використанням експертних систем.
8. Керування фізичними експериментами, аналіз даних і поглиблене дослідження отриманих результатів.
9. Виконання соціальних досліджень.
10. Підтримка прийняття особистих рішень та рішень стосовно ведення домашнього господарства.

11. Контроль стану навколишнього середовища з використанням двох типів моделей – на основі диференціальних рівнянь в частинних похідних і рівнянь авторегресії (АР), авторегресії з ковзним середнім (АРКС), множинної регресії.

12. Аналіз даних і прийняття рішень в генетиці, біології (біостатистика), психології (наприклад, сучасний *Journal of Experimental Psychology* в США).

На **початковій стадії** виконання статистичного аналізу, власне даних для аналізу, як правило, немає або ще навіть не прийнято рішення стосовно того, які дані необхідно аналізувати. Це питання вирішують на етапі *планування збору даних* таким чином, щоб отримати дійсно корисні, максимально повні та інформативні дані.

На **другому етапі** *висувають гіпотези* та виконують *попередню обробку* і *дослідження* даних.

Третій етап – *оцінювання необхідних статистичних параметрів та інших невідомих величин*, параметрів математичних моделей і т. ін.

На останньому, **четвертому, етапі** виконується *перевірка висунутих гіпотез* – дані використовують для прийняття рішення щодо відповідності апріорно висунутого припущення дійсній ситуації. Розглянемо згадані етапи статистичного аналізу докладніше.

Чотири етапи статистичного аналізу даних

1. Планування експерименту і збір даних. Результат – корисні інформативні та повні дані стосовно функціонування (протікання) процесу.

2. Попередня обробка та дослідження даних, формулювання гіпотез стосовно типів розподілів, значущості оцінок, якості (адекватності) моделей, настання можливих ситуацій і т. ін.

3. Оцінювання параметрів математичних і статистичних моделей. Результат – адекватні процесу математичні і статистичні моделі, що підтверджується відповідними статистичними параметрами якості.

4. Перевірка раніше сформульованих гіпотез, прийняття рішень стосовно управління процесами.

Розглянемо деякі терміни, що стосуються статистичних даних і моделей, які оцінюються на їх основі.

Математичні моделі, які ми будемо розглядати, це моделі у вигляді рівнянь різних типів: диференціальних різницевих, алгебраїчних.

Статистичні моделі – це моделі у формі розподілів ймовірностей випадкових величин.

Статистичні дані у формі вимірів, прив'язаних до конкретних моментів часу, тобто *часових рядів* розглядаються у прикладній статистиці як випадкові процеси з детермінованими складовими:

$$y(k) = константа + детермінована\ складова + випадкова\ складова.$$

Детермінована складова процесу надає можливість будувати регресійні моделі та моделі деяких інших типів. Так, наявність автокореляції часового ряду дає можливість будувати авторегресійні моделі, порядок яких визначається за допомогою автокореляційної і часткової автокореляційної функцій.

Статистичні дані у формі вимірів, які стосуються деякого вибраного моменту часу (наприклад, конкретного місяця року), називають ***часовими перерізами.***

Поява випадкової складової у вимірах змінних досліджуваних процесів зумовлена такими причинами:

- наявність випадкових збурень, які, як правило, неможливо виміряти;
- некоректне формування структури моделі: в праву частину рівняння можуть бути введені "зайві" змінні, тобто такі, що формально корельовані із залежною змінною, але фактично на неї не впливають або впливають дуже слабо;
- можливо, що у праву частину рівняння помилково не введені незалежні змінні, які фактично мають досить сильний вплив на залежну;
- похибки обчислень, які завжди виникають при оцінюванні параметрів моделей.

1.2 Планування збору даних (планування експерименту) – перший етап статистичного аналізу даних

Планування збору даних в соціально-економічних, фінансових та маркетингових дослідженнях називають ***плануванням вибіркового дослідження***.

У процесі планування вибіркового дослідження розв'язують задачі, які можна розділити на групи, зазначені нижче.

<u>Група 1</u>: Визначення цілей дослідження та їх представлення у конкретному вигляді

До цієї групи відносять такі завдання**:**

- визначення області (галузі) дослідження (економіка, політика, соціологія і т. ін.);
- визначення головної мети; наприклад, оцінювання та прогнозування

стану галузі промисловості або стійкості виробничого підприємства (його близькість до банкрутства);

- аналіз структури досліджуваного процесу – з яких елементів він складається (наприклад, університет складається з факультетів, інститутів та кафедр, а підприємство може мати цехи, відділи, філії і т. ін.);
- встановлення переліку змінних і параметрів, які необхідно виміряти чи обчислити;
- визначення показників (змінні), які будуть представляти остаточний результат дослідження, тобто, що буде отримано після закінчення роботи (наприклад, оцінки коротко- або середньострокового прогнозу розвитку процесу, рейтинг кандидата у президенти або середня успішність студентів університету для вибраних спеціальностей та курсів).

<u>Група 2</u>: Створення загального плану дослідження

Для створення загального плану дослідження розв’язують такі задачі:

- вибирають одиницю, стосовно якої буде виконуватись дослідження (наприклад, сім’я, окремі особи або групи, підприємство, район чи область);
- визначають, чи будуть вибрані об’єкти досліджуватись за вибіркою чи за генеральною сукупністю даних (виконання збору максимального можливого об’єму даних);
- обирають метод отримання інформації стосовно об'єкта: за допомогою інтерв’ю, поштою, з Інтернету або з використанням спеціальної апаратури чи шляхом простого спостереження;
- вибирають тип бази даних для накопичення інформації;
- оцінюють вартість кожної операції;
- встановлюють вимоги до точності результатів дослідження (як правило, збільшення об’єму вибірки сприяє підвищенню точності результату, але підвищує вартість дослідження).

<u>Група 3</u>: Планування характеру вибірки

У процесі планування характеру вибірки даних необхідно:

- встановити необхідність стратифікації, тобто необхідно досліджувати об’єкт на одному рівні чи більше (наприклад, успішність в університеті може досліджуватись на таких рівнях: молодші курси, старші курси, аспіранти, друга вища освіта та курси підвищення кваліфікації);
- визначити вибіркову одиницю, її характер і величину (наприклад, вибірковою одиницею може бути один студент або група студентів);
- встановити кількість і типи стадій збору даних (наприклад, збір даних по семестрах);

- оцінити кількість і типи фаз виконання кожної стадії (наприклад, послідовність збору даних по групах);
- встановити кількість вибіркових одиниць, які будуть взяті на кожній стадії і фазі;
- оцінити вартість виконання операцій стосовно збору даних;
- встановити метод відбору одиниць спостережень – ймовірнісний або на основі суджень, експертний; тобто якість даних можна оцінювати на основі обчислення ймовірностей появи некоректних значень (наприклад, відомо, що процент недостовірних даних може сягати 10-15 %) або ж якість даних оцінюють експерти;
- встановити способи подолання труднощів, які можуть виникнути в процесі отримання вибіркових даних (дані не можна отримати у випадку небажання їх давати або у випадку захворювання суб'єкта дослідження і т. ін.);
- вибрати (розробити) методи обчислення необхідних статистичних оцінок за вибірковими даними.

<u>Група 4</u>: Безпосередній збір даних

Безпосередній збір даних передбачає виконання таких операцій:

- розробка форми реєстрації даних (анкета, таблиця, виборчий бюлетень, щоденник, цифровий реєстратор і т. ін.);
- обрання методу вибору, навчання та контролю роботи дослідника (інтерв'юера, спостерігача і т. ін.);
- вибір альтернативного (відмінного від вибіркового) методу збору спостережень у випадку виникнення труднощів із отриманням даних стосовно досліджуваного об'єкта;
- безпосередній збір даних прямим або непрямим (опосередкованим) методом.

Дослідження **технічних об'єктів** і **технологічних процесів** має свої особливості. Перший етап називають *плануванням та виконанням експерименту*. На цьому етапі виконують такі дії:

- визначають мету дослідження;
- визначають об'єкт дослідження та можливості виконання експерименту з цим об'єктом;
- вибирають змінні для вимірювання, їх граничні значення та створюють методику збору даних;
- визначають тип бази даних та інструментальну систему для її створення;
- визначають необхідний для дослідження процесу об'єм даних;
- визначають число учасників і час проведення дослідження;
- вибирають технічне устаткування для виконання дослідження, засоби обчислювальної техніки;

– визначають вартість дослідження.

Вибіркові статистичні дослідження особливо корисні тоді, коли є великі групи людей, підприємств або інших об'єктів чи процесів (***генеральна сукупність***), які необхідно дослідити, але повне дослідження провести неможливо. Для того, щоб отримати не ідеальне, але практично корисне розуміння ситуації щодо генеральної сукупності, можна відібрати невелику групу (***вибірку***) даних, яка складається із деяких, але не всіх, об'єктів генеральної сукупності.

Процес узагальнення результатів дослідження на всю генеральну сукупність називають статистичним висновком або рішенням.

Випадкова вибірка з генеральної сукупності є одним із кращих способів накопичення даних, оскільки генеральна сукупність, як правило, занадто велика, щоб її вивчати повністю.

1.3 Попередня обробка і дослідження даних – другий етап статистичного аналізу даних

Попередня обробка і дослідження даних – важлива частина вибіркового дослідження, оскільки кваліфікований попередній аналіз дає можливість уникнути багатьох трудомістких обчислювальних процедур, скоротити час виконання дослідження.

Метою попередньої обробки і дослідження даних є:

– приведення даних до формату, який сприяє підвищенню їх якості з точки зору оцінювання параметрів моделей, тобто підвищення ступеня обумовленості матриць вимірів;

– попереднє оцінювання структури моделі.

Попередня обробка передбачає виконання таких операцій над зібраними статистичними даними.

1. Візуальне дослідження даних

За допомогою графіків, діаграм та описових статистик попередньо виявляють наявність можливих трендів (інтегрованість та коінтегрованість процесів), сезонних ефектів, екстремальних значень.

Тренд – поточне середнє значення процесу, яка вказує на його довгострокові зміни. Як правило аналіз тренду виконують при розв'язанні задачі довгострокового прогнозування. Короткострокові зміни визначаються коливаннями, які накладаються на тренд.

Називають нестаціонарний процес із трендом по-різному:

процес із трендом = інтегрований процес = процес із одиничними коренями.

Термін „*інтегрований*” походить від зовнішньої подібності графіка тренду з виходом електронного пристрою – *інтегратора*. При подачі на вхід інтегратора сигналу у вигляді константи на його виході формується сигнал у вигляді прямої з додатнім нахилом (лінійний тренд).

Термін „процес з одиничними коренями” означає, що серед коренів характеристичного рівняння, записаного для рівняння АР(p) або АРКС (*p*,*q*), є хоча б один одиничний корінь. Кількість одиничних коренів відповідає порядку тренду. Як приклад, розглянемо рівняння авторегресії першого порядку, АР(1):

$$y(k) = a_0 + a_1 y(k-1) + \varepsilon(k).$$

Однорідне рівняння для цієї моделі має вигляд:

$$y(k) - a_1 y(k-1) = 0.$$

Підставимо в це однорідне рівняння однорідний розв’язок загального вигляду $y^h(k) = A\lambda^k$ і отримаємо:

$$A\lambda^k - a_1 A\lambda^{k-1} = 0.$$

Поділимо обидві частини останнього рівняння на $A\lambda^{k-1}$ і отримаємо характеристичне рівняння першого порядку для однорідного рівняння АР(1):

$$\lambda - a_1 = 0.$$

Звідси корінь характеристичного рівняння: $\lambda = a_1$. Якщо $a_1 = 1$, то таке рівняння АР(1) описує процес з лінійним трендом (тренд першого порядку), який позначають через $I(1)$. Це можна показати, якщо знайти розв’язок такого рівняння:

$$y(k) = a_1 + y(k-1) + \varepsilon(k).$$

Знайдемо розв’язок ітераційним методом при початкових умовах: $y(0) = y_0$ і $\varepsilon(0) = \varepsilon_0$:

$$y(1) = a_0 + y_0 + \varepsilon_1;$$

$$y(2) = a_0 + y(1) + \varepsilon(2) = a_0 + a_0 + y_0 + \varepsilon(1) + \varepsilon(2).$$

За індукцією для довільного моменту часу k отримаємо розв’язок:

$$y(k) = y_0 + a_0 k + \sum_{i=1}^{k} \varepsilon(i),$$

який містить лінійний тренд $a_0 k$.

Візуальний аналіз дає можливість (наближено) встановити наявність *гетероскедастичності* – тобто дисперсія процесу змінюється у часі за лінійним або складнішим законом. Різні рівні дисперсії на різних ділянках ряду даних свідчать про наявність гетероскедастичності і необхідності застосування

спеціальних моделей для опису самих даних і дисперсії процесу. Моделі дисперсії необхідно будувати для того, щоб отримати можливість обчислювати оцінки прогнозів дисперсії і стандартного відхилення (стандартне відхилення називають ще волатильністю, тобто ступенем мінливості досліджуваного процесу).

За допомогою візуального аналізу можна наближено оцінити тип розподілу за гістограмою та описовими статистичними параметрами (ексцес, асиметрія (куртозис), статистика Жак-Бера, статистика Колмогорова-Смірнова та інші); а також методами, які ґрунтуються на використанні функцій-ядер.

Деякі формули для розрахунку описових статистик

Приклади статистичних параметрів, які використовуються з метою встановлення наближення розподілу до нормального.

Критерій узгодженості χ^2 – це критерій загального типу, який ґрунтується на порівнянні емпіричної гістограми розподілу випадкової величини з її теоретичною щільністю. Діапазон зміни експериментальних даних розбивають на m інтервалів і обчислюють статистику:

$$\chi^2 = \sum_{k=1}^{m} \frac{(n_k - N\,p_k)^2}{N\,p_k},$$

де n_k – кількість значень випадкової величини, які попадають в $k-$й інтервал; $m-$ кількість інтервалів; $N = \sum_{k=1}^{m} n_k$ – об'єм вибірки; $p_k = F(x_{k+1}) - F(x_k)$ – теоретична ймовірність попадання випадкової величини в $k-$й інтервал; $F(x)-$ гіпотетичний теоретичний закон розподілу ймовірностей випадкової величини.

Дисперсія статистики χ^2 визначається за виразом:

$$\mathrm{var}(\chi^2) = 2(m-1) + \frac{1}{N}\left(\sum_{k=1}^{m}\frac{1}{p_k} - m^2 - 2m + 2)\right).$$

Якщо $\sum_{k=1}^{m}\frac{1}{p_k} << N$ і $m << N$, то $\mathrm{var}(\chi^2) = 2(k-1)$, тобто співпадає з дисперсією випадкової величини, яка має розподіл χ^2. На цій основі прийнято вважати, що статистика χ^2 має розподіл, близький до розподілу хі-квадрат.

Коефіцієнт асиметрії (коефіцієнт скісності або *skewness*) – характеризує симетричність (хвостів) розподілу і розраховується за виразом:

$$S = \frac{1}{N}\sum_{k=1}^{N}\left[\frac{y(k) - \bar{y}}{\sigma}\right]^3,$$

де $\bar{y}$ – вибіркове середнє; σ – стандартне відхилення процесу.

Якщо $S > 0$, то правий хвіст розподілу довший, а при $S < 0$ довшим є лівий хвіст розподілу; якщо $S = 0$, то розподіл симетричний.

Ексцес (коефіцієнт гостровершинності або kurtosis) *– характеризує відмінність форми розподілу від нормального і розраховується за виразом:*

$$K = \frac{1}{N}\sum_{k=1}^{N}\left[\frac{y(k)-\bar{y}}{\sigma}\right]^4;$$

$K = 3$ для нормального розподілу; якщо $K > 3$, то форма розподілу буде "гострішою" від нормального; при $K < 3$ форма розподілу буде "пласкішою" від нормального.

Статистика Жак-Бера *(Jarque-Bera)* – тестова статистика, яка показує наскільки близьким є ряд до нормального розподілу; фактично це різниця між значеннями S і K для досліджуваного ряду та нормального розподілу:

$$JB = \frac{N-p}{6}\left[S^2 + \frac{1}{4}(K-3)^2\right],$$

де p – кількість коефіцієнтів, використаних для побудови моделі ряду даних; при нуль-гіпотезі щодо нормальності розподілу статистика Жак-Бера має розподіл χ^2 з двома степенями свободи. Ймовірність, пов'язана із статистикою Жак-Бера, показує ймовірність справедливості нуль-гіпотези.

Мала ймовірність (близька до нуля) свідчить про те, що нуль-гіпотезу щодо нормальності розподілу необхідно відхилити.

- Шляхом візуального аналізу можна наближено встановити наявність екстремальних значень (значення, які суттєво перевищують всі інші значення вибірки, але не відносяться до похибок вимірів). Необхідно встановити походження наявних екстремальних значень і вибрати метод їх обробки. Як правило, за вибраною методикою зменшують амплітуди екстремальних значень з метою наближення процесу до стаціонарного.
- Також виявляють наявність нелінійностей у даних; нелінійні процеси вимагають застосування спеціальних моделей для опису та оцінювання параметрів.
- Візуальний аналіз дає можливість визначити наявність перехідних та усталених режимів функціонування процесу. Перехідний – це, як правило, відносно короткостроковий режим функціонування процесу, який може характеризуватись значними коливаннями значень змінних. Перехідні процеси у макроекономіці можуть тривати десятиліття, а перехідні процеси у технічних системах можуть тривати від десятків мікросекунд до десятків хвилин. Усталений режим – це режим, у якому досліджуваний об'єкт може функціонувати тривалий час (декілька годин, днів, років, десятиліть – для макроекономіки).

- Встановлюється наявність коливань – слабкі, що зумовлені похибками вимірів і сильні коливання, зумовлені різноманітними випадковими впливами (збуреннями).

2. Заповнення пропусків даних, якщо вони є

Для заповнення пропусків даних використовують такі методи: *просте усереднення, екстраполяція, інтерполяція, прогнозування* та інші.

Приклад прогнозування без розв'язку рівнянь

Структура різницевого рівняння така, що воно дозволяє виконувати прогнозування на один крок (один період дискретизації вимірів) без додаткових перетворень. Тобто в праву частину необхідно підставити минулі значення змінних і обчислити оцінку прогнозу головної змінної в лівій частині. Але для того, щоб знайти оцінку прогнозу на більше число кроків, необхідно застосувати деякі попередні перетворення різницевих рівнянь. Розглянемо деякі можливі підходи до формування функцій прогнозування та обчислення оцінок прогнозованих значень.

Як приклад, розглянемо рівняння АР(1):

$$y(k) = a_0 + a_1 y(k-1) + \varepsilon(k), \;\; E[\varepsilon(k)] = 0 \qquad (1.3.1).$$

Збільшимо незалежну змінну *час* на одну одиницю і запишемо рівняння знову:

$$y(k+1) = a_0 + a_1 y(k) + \varepsilon(k+1) \qquad (1.3.2).$$

Якщо коефіцієнти a_0, a_1 відомі, то можна знайти умовне математичне сподівання на основі відомої інформації до моменту *k* включно:

$$\begin{aligned} E_k[y(k+1)] &= E_k[y(k+1)\,|\,y(k), y(k-1), \ldots, \varepsilon(k), \varepsilon(k-1), \ldots] = \\ &= a_0 + a_1 E_k[y(k)] = a_0 + a_1 y(k), \end{aligned} \qquad (1.3.3)$$

оскільки $y(k)$ в момент *k* є відомою константою.

По аналогії запишемо рівняння (1.3.1) для моменту $k+2$:

$$y(k+2) = a_0 + a_1 y(k+1) + \varepsilon(k+2) \qquad (1.3.4)$$

і знайдемо умовне математичне сподівання

$$\begin{aligned} E_k[y(k+2)] &= a_0 + a_1 E_k[y(k+1)] = a_0 + a_1 E_k[a_0 + a_1 y(k)] = \\ &= a_0 + a_0 a_1 + a_1^{\,2} y(k). \end{aligned}$$

Для наступного моменту часу маємо:

$$E_k[y(k+3)] = a_0 + a_0 a_1 + a_0 a_1^{\,2} + a_1^{\,3} y(k).$$

Таким чином, для загального випадку прогнозування на *s* кроків можна записати:

$$E_S[y(k+s)] = a_0\left(\sum_{i=0}^{S-1} a_1^{\,i}\right) + a_1^{\,S} y(k) = a_0 \sum_{i=0}^{S-1} a_1^{\,i} + a_1^{\,S} y(k) \qquad (1.3.5).$$

Отримане рівняння називають функцією прогнозування на довільне число кроків. Прогноз представляє собою збіжний процес, якщо $|a_1| < 1$, тобто

$$\lim_{s\to\infty} E_k[y(k+s)] = \frac{a_0}{1-a_1}, \qquad |a_1| < 1, \qquad (1.3.6)$$

де a_1 – знаменник геометричної прогресії. Вираз (1.3.6) свідчить про те, що для будь-якого стаціонарного процесу АР чи АРКС оцінка умовного прогнозу асимптотично $(s \to \infty)$ збігається до безумовного середнього.

3. Обробка екстремальних значень

Екстремальними називають значення, які суттєво відрізняються від середнього значення вибірки, але не відносяться до помилкових.

Мета виявлення та обробки екстремальних значень – наблизити процес до стаціонарного для того, щоб можна було застосувати теорію аналізу стаціонарних процесів.

Методи виявлення та опису екстремальних значень:

- за розподілами середніх для екстремальних значень;
- за допомогою характеристичних екстремальних значень;
- з використанням розподілів вимірів типу експоненціального, логістичного, логнормального.
- шляхом вибору та оцінювання параметрів спеціальних функцій для опису розподілу, наприклад копул.

4. Нормування даних

Нормування даних можна виконувати за допомогою різних процедур, але при цьому важливо, щоб перетворення вимірів було лінійним.

Розглянемо деякі ***популярні підходи до нормування даних***.

1. Нормування додатніх значень шляхом їх ***логарифмування***:

$$y_l(k) = \ln[y(k)].$$

2. Нормування ***діленням на максимальне значення***:

$$y_{норм}(k)=\frac{y(k)}{|Y_{max}|} \;\Rightarrow\; -1\le y(k)\le +1. \qquad (1.3.7)$$

Нормоване значення може бути визначене також і в іншому діапазоні:

$$-10\le y(k)\le 10.$$

3. Досить хороші результати ***нормування при оцінюванні множинної регресії***

$$y(k)=\beta_0+\beta_1 x_1(k)+\beta_2 x_2(k)+...+\beta_p x_p(k)+\varepsilon(k) \qquad (1.3.8).$$

можна досягти завдяки одночасному нормуванню і центруванню даних таким чином:

$$x_{iH}(k)=\frac{x_i(k)-\bar{x}_i}{\sqrt{S_i}}=\frac{x_i(k)-\bar{x}_i}{\left(\sum_{k=1}^{N}(x_i(k)-\bar{x}_i)^2\right)^{1/2}},\quad k=1,...,N,\ i=1,...,p;$$
$$y_H(k)=\frac{y(k)-\bar{y}}{\sqrt{S_y}}=\frac{y(k)-\bar{y}}{\left(\sum_{k=1}^{N}(y(k)-\bar{y})^2\right)^{1/2}}, \qquad (1.3.9)$$

де $x_i(k)-$ значення $i-$го стовпчика матриці вимірів (виміри незалежних змінних); $y(k)-$ виміри залежної змінної; $x_{iH}(k), y_H(k)-$ нормовані значення змінних; $\bar{x}_i, \bar{y}-$ вибіркові середні значення незалежних і залежної змінних, відповідно; $N-$ кількість вимірів; $p-$ кількість незалежних змінних (регресорів) x_i.

Якщо ввести позначення для центрованих змінних

$$\tilde{x}_i(k)=x_i(k)-\bar{x}_i;\quad \tilde{y}(k)=y(k)-\bar{y}, \qquad (1.3.10)$$

то регресія для центрованих змінних матиме вигляд:

$$\tilde{y}(k)=\beta_1\tilde{x}_1(k)+\beta_2\tilde{x}_2(k)+...+\beta_p\tilde{x}_p(k)+\varepsilon(k) \qquad (1.3.11)$$

Якщо підставити (1.3.9) в (1.3.11), то рівняння множинної регресії матиме вигляд:

$$y_H(k)\,S_y^{1/2}=\beta_1 S_1^{1/2}x_{1H}(k)+\beta_2 S_2^{1/2}x_{2H}(k)+...+\beta_p S_p^{1/2}x_{pH}(k)+\varepsilon'(k) \qquad (1.3.12).$$

Тепер поділимо ліву і праву частини на $S_y^{1/2}$:

$$y_H(k)=\alpha_1 x_{1H}(k)+\alpha_2 x_{2H}(k)+...+\alpha_p x_{pH}(k)+\varepsilon''(k), \qquad (1.3.13)$$

де $\alpha_1=\beta_1(S_1/S_y)^{1/2},\ ...,\alpha_p=\beta_p(S_p/S_y)^{1/2}.$

Отримане рівняння (1.3.13) – це рівняння для нормованих вимірів.

В результаті центрування і нормування покращується ступінь обумовленості матриці вимірів, яка вимірюється відношенням:

$$\eta = \left| \frac{\lambda_{max}}{\lambda_{min}} \right|,$$

де $\lambda_{max}, \lambda_{min}$ – максимальне і мінімальне власні числа матриці вимірів. Для забезпечення належних умов оцінювання параметрів необхідно задовольнити умову: $\eta < 10$ (емпірично встановлена рекомендація).

5. Фільтрація даних

За необхідності, виконується ***фільтрація даних*** від шумів. Для розв'язання цієї важливої задачі застосовують *цифрові* або *оптимальні* фільтри.

Цифровий фільтр можна представити, наприклад, рівнянням типу АР(р):

$$y(k) = a_1 y(k-1) + a_2 y(k-2) + \ldots + a_p y(k-p).$$

Фільтр має амплітудно-частотну характеристику (АЧХ), яка визначається значеннями коефіцієнтів рівняння.

Мета застосування цифрового фільтра: пропустити корисну частину спектру і затримати шумову або просто непотрібну для аналізу складову.

Оптимальний фільтр потребує використання моделі процесу, представленої у просторі станів. Основна задача застосування оптимального фільтра полягає у обчисленні оптимальних оцінок стану досліджуваного процесу із врахуванням впливу на його функціонування випадкових збурень та похибок (шумів) вимірів.

Так, нестаціонарна лінійна система описується у дискретному часі рівняннями із змінними у часі коефіцієнтами (тобто коефіцієнти або параметри моделі залежать від часу k):

$$\mathbf{x}(k) = \mathbf{A}(k, k-1)\mathbf{x}(k-1) + \mathbf{B}(k, k-1)\mathbf{u}(k-1) + \mathbf{w}(k),$$

де $\mathbf{x}(k)$ – n-вимірний вектор станів системи; $\mathbf{u}(k-1)$ – m- вимірний вектор детермінованих вхідних величин (сигнали керування); $\mathbf{w}(k-1)$ – n-вимірний вектор випадкових зовнішніх збурень; $\mathbf{A}(k, k-1)$ – $(n \times n)$ матриця динаміки системи (вона містить коефіцієнти, що характеризують динаміку, тобто швидкість зміни станів у часі); $\mathbf{B}(k, k-1)$ – $(n \times m)$ матриця коефіцієнтів керування.

Подвійний часовий аргумент у вигляді $(k, k-1)$ означає, що величина з цим аргументом використовується в момент k, але її значення ґрунтується на

попередніх даних, які відомі на момент $k-1$, включно.

Далі будемо записувати для простоти матриці **A** і **B** з одним аргументом, тобто $\mathbf{A}(k)$ та $\mathbf{B}(k)$. Очевидно, що стаціонарна система описується матрицями з постійними коефіцієнтами, які записують просто **A** і **B**. Оскільки матриця **A** зв'язує поточний стан із попереднім, то її називають ще перехідною матрицею станів. Нагадаємо, що дискретний час k зв'язаний з неперервним часом t періодом дискретизації вимірів T_s: $t = kT_s$. Використання поняття дискретного часу відповідає характеру зібраних даних і дає можливість спростити обчислення.

У класичній постановці задачі оптимальної фільтрації послідовність зовнішніх збурень $\mathbf{w}(k)$ задовольняє властивостям білого гаусового шуму з нульовим середнім значенням і коваріаційною матрицею **Q**, тобто статистики шуму мають вигляд:

$$E[\mathbf{w}(k)] = 0, \quad \forall k,$$
$$E[\mathbf{w}(k)\mathbf{w}^T(j)] = \mathbf{Q}(k)\,\delta_{kj},$$

де δ_{kj} – дельта-функція Кронекера, що визначається так: $\delta_{kj} = \begin{cases} 0 \text{ для } k \neq j \\ 1 \text{ для } k = j \end{cases}$; $\mathbf{Q}(k)$ – додатно визначена коваріаційна матриця зовнішніх збурень стану розмірності $(n \times n)$. Діагональні елементи матриці представляють собою дисперсії компонент вектора збурень $\mathbf{w}(k)$.

Початковим станом системи $\mathbf{x}_0$ будемо вважати випадкові змінні з відомими статистиками:

$$E[\mathbf{x}_0] = \bar{\mathbf{x}}_0; \quad E[\mathbf{x}_0\mathbf{x}_0^T] = \mathbf{M}; \quad E[\mathbf{w}(k)\mathbf{x}_0^T] = 0, \ \forall k.$$

Нехай вектор вимірів $\mathbf{z}(k)$ вихідних змінних доступний в будь-який момент часу t_k, а його компоненти лінійно зв'язані з вектором стану і на них впливає шум вимірів, тобто

$$\mathbf{z}(k) = \mathbf{H}(k)\mathbf{x}(k) + \mathbf{v}(k),$$

де $\mathbf{H}(k)$ – матриця спостережень вимірності $(r \times n)$, $\mathbf{v}(k)$ – r-вимірний вектор випадкових величин шуму вимірів з відомими статистиками:

$$E[\mathbf{v}(k)] = 0, \quad E[\mathbf{v}(k)\mathbf{v}^T(j)] = \mathbf{R}(k)\,\delta_{kj},$$

де $\mathbf{R}(k)$ – додатно визначена коваріаційна матриця шумів вимірів вимірності $(r \times r)$, діагональні елементи якої є дисперсіями адитивного шуму в кожному каналі вимірів.

Шум вимірів також задовольняє властивостям білого гаусового шуму. Він вважається некорельованим із зовнішнім збуренням $\mathbf{w}(k)$ і початковим станом системи, тобто

$$E[\mathbf{v}(k)\mathbf{w}^T(j)] = 0 \quad \forall\ k, j;$$
$$E[\mathbf{v}(k)\mathbf{x}_0^T] = 0 \quad \forall\ k.$$

Для визначеної вище системи з вектором стану $\mathbf{x}(k)$ необхідно знайти оцінку стану $\hat{\mathbf{x}}(k)$ в момент t_k як лінійну комбінацію оцінки $\hat{\mathbf{x}}(k-1)$ в момент t_{k-1} і самого останнього виміру (статистичних даних) $\mathbf{z}(k)$.

Оцінка $\hat{\mathbf{x}}(k)$ повинна обчислюватися як найкраща за мінімумом середнього значення суми квадратів оцінок похибок. Іншими словами, оцінка повинна бути такою, щоб

$$E[(\hat{\mathbf{x}}(k) - \mathbf{x}(k))^T(\hat{\mathbf{x}}(k) - \mathbf{x}(k))] = \min_K,$$

де $\mathbf{x}(k)$ – точне значення вектора стану, яке може бути обчислене за допомогою детермінованої складової математичної моделі процесу; $\mathbf{K}$ – оптимальний матричний коефіцієнт фільтра, який необхідно обчислити в результаті розв'язання оптимізаційної задачі.

Таким чином, фільтр необхідно будувати та використовувати для уточнення оцінок стану процесу в умовах впливу випадкових зовнішніх збурень та наявності шумів (похибок) вимірів.

Сьогодні оптимальні фільтри – це невід'ємна складова частина комп'ютерних систем обробки експериментальних даних.

6. Кореляційний аналіз

Кореляційний аналіз даних необхідно виконувати для встановлення наявних зв'язків між значеннями однієї вибірки даних та між значеннями декількох вибірок.

Кореляція характеризує наявність (відсутність) лінійної або нелінійної залежності між змінними.

Коефіцієнт кореляції, а в загальному випадку кореляційна функція, дозволяють встановити зв'язок між змінними, наприклад r_{yx} – коефіцієнт кореляції між змінними y та x. Кореляція може бути лінійною або нелінійною залежно від типу взаємозв'язку, який фактично існує між змінними. Досить часто на практиці розглядають тільки лінійну кореляцію, але більш глибокий аналіз потребує використання для дослідження функціонування процесів нелінійних залежностей. Складну нелінійну залежність часто можна спростити (лінеаризувати), але знати про її існування необхідно для того, щоб побудувати адекватну модель процесу.

Вибірковий коефіцієнт кореляції між двома змінними обчислюється за формулою:

$$r_{yx} = \frac{1}{N-1} \frac{\sum_{k=1}^{N} \{[y(k) - \bar{y}][x(k) - \bar{x}]\}}{\sigma_x \sigma_y},$$

або

$$r_{yx} = \frac{\sum_{k=1}^{N} \{[y(k) - \bar{y}][x(k) - \bar{x}]\}}{\sqrt{\sum_{k=1}^{N} [y(k) - \bar{y}]^2} \sqrt{\sum_{k=1}^{N} [x(k) - \bar{x}]^2}}.$$

де $-1 < r_{yx} < +1$; σ_x, σ_y – стандартні відхилення для змінних x і y, відповідно.

Якщо необхідно одночасно обчислити кореляцію між декількома змінними (наприклад, між ендогенною та декількома екзогенними), то формують кореляційну матрицю.

7. Попереднє визначення структури математичних моделей

Однією із складових прикладної статистики є регресійний аналіз даних, тобто побудова математичних моделей на основі статистичних даних і їх використання для прогнозування і керування процесами різної природи.

Важливою задачею процесу моделювання є коректний вибір структури моделі. Поняття структури математичної моделі включає в себе такі елементи:

розмірність моделі (кількість рівнянь, що описують процес або об’єкт);

порядок моделі (максимальний порядок різницевого або диференціального рівняння, яке входить у модель);

нелінійність та її тип (може бути нелінійність стосовно змінних або нелінійність стосовно параметрів);

час (лаг) запізнення (по входу) та його оцінка;

збурення стану та його тип (детерміноване або випадкове);

можливі обмеження на змінні і параметри моделі.

Оцінювання структури моделі процесу називають ще структурною ідентифікацією. Воно виконується на основі всієї наявної інформації про функціонування процесу, а також на основі результатів кореляційного аналізу наявних експериментальних (статистичних) даних.

1.4 Оцінювання параметрів статистичних і математичних моделей

На цьому (третьому) етапі статистичного аналізу оцінюють параметри вибраних типів моделей, а також інших невідомих величин. Наприклад, об’єми

продажу деякого продукту, реакцію населення на новий продукт, зміну продуктивності виробничого підприємства, рівень браку у виробничому процесі.

Відповідно до типів моделей (лінійні та нелінійні) вибирають метод оцінювання параметрів. Найбільш поширені – *метод найменших квадратів* (МНК) і *метод максимальної правдоподібності* (ММП).

Звичайний метод найменших квадратів для оцінювання лінійних та псевдолінійних моделей ґрунтується на використанні квадратичного критерію:

$$\min_{\hat{\theta}} J = \sum_{k=1}^{N} [y(k) - \hat{y}(k)]^2 = \sum_{k=1}^{N} e^2(k),$$

де $\hat{y}(k)-$ оцінка залежної змінної по побудованій моделі, наприклад, по АР(2): $\hat{y}(k) = \hat{a}_0 + \hat{a}_1 y(k-1) + \hat{a}_2 y(k-2)$.

Формула МНК має вигляд***:***

$$\hat{\theta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y},$$

де $\mathbf{X}-$ матриця вимірів екзогенних змінних у правій частині рівняння; $\mathbf{y}(k)-$ вектор вимірів ендогенної змінної. Наприклад, для регресійної моделі множинної регресії

$$y(k) = a_0 + a_1 x_1(k) + a_2 x_2(k) + \ldots + a_m x_m(k) + \varepsilon(k)$$

матриця вимірів має вигляд:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1(1) & \cdots & x_m(1) \\ 1 & x_1(2) & \cdots & x_m(2) \\ \vdots & & & \vdots \\ 1 & x_1(N) & \cdots & x_m(N) \end{bmatrix}.$$

Виміри регресорів (стовпчики матриці $\mathbf{X}$) не повинні бути висококорельованими, оскільки їх висока корельованість може призвести до виродженості і неможливості обчислити обернену матрицю (проблема мультиколінеарності). Для зменшення корельованості регресорів застосовують спеціальні методи ортогоналізації, наприклад, метод головних компонент (МГК).

Метод максимальної правдоподібності (ММП), що ґрунтується на максимізації функції правдоподібності:

$$L = (2\pi\sigma^2)^{-N/2} \prod_{k=1}^{N} \exp\left(-\frac{[y(k) - \mathbf{X}(k)\beta]^2}{2\sigma^2} \right);$$

або логарифмованої функції правдоподібності:

$$\ell = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\prod_{k=1}^{N}[y(k) - \mathbf{X}(k)\beta]^2 =$$

$$= -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{k=1}^{N}(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) =$$

$$= -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\mathbf{e}^T\mathbf{e},$$

де $\mathbf{e} = \mathbf{y} - \mathbf{X}\beta$.

Застосовують і ***рекурсивні версії МНК і ММП.***

Формулювання гіпотез стосовно оцінок параметрів

Статистичні оцінки – це тільки *припущення щодо можливих значень*, а тому вони часто бувають неточними. Однак, якщо вони достатньо близькі до істини, то служать поставленій меті. Якщо відома їх точність, то можна вирішити в якій мірі їх варто брати до уваги.

Верхнє і нижнє значення оцінки невідомої величини дає ***довірчий інтервал,*** який дає нам впевненість, що оцінка лежить у визначеному діапазоні значень.

Для отриманих статистичних оцінок параметрів моделі формулюють гіпотези стосовно їх статистичної значущості.

Наприклад, для оцінок регресійної моделі формулюють таку гіпотезу:

$H_0 : \hat{a}_i = 0$ – оцінка статистично незначуща;

$H_0 : \hat{a}_i \neq 0$ – оцінка статистично значуща.

Кожна гіпотеза формулюється як твердження, яке може бути вірним або невірним. В результаті перевірки гіпотези (на основі обробки даних) ми приймаємо або відкидаємо попередньо висунуту гіпотезу.

Існує стандартна процедура перевірки гіпотез, якою користуються в техніці, психології, освіті, суспільних науках та багатьох інших областях. Ця процедура буде розглянута в цьому розділі.

Необхідно підкреслити той важливий факт, що *результат експерименту, який підтверджує справедливість висунутої гіпотези, майже ніколи не може бути основою для прийняття цієї гіпотези*. В той же час ***результат, несумісний з висунутою гіпотезою, є цілком достатнім для її відхилення як не правильної***. Очевидно, що наведене твердження потребує обґрунтування.

Приклад 1.1. Припустимо, що середній коефіцієнт розумового розвитку (КРР) деякої генеральної сукупності людей $\mu = 100$ (це гіпотетичне значення). Результат взятої випадкової вибірки дає результат $\mu_i = 102$, який є сумісним з висунутою гіпотезою. Однак цей результат є також сумісним з припущенням, що $\mu = 101$ або $\mu = 99$ і, звичайно, є

сумісним з гіпотезою, що $\mu = 102$. Таким чином, на основі даного результату не можна віддати перевагу гіпотезі, що $\mu = 100$.

Припустимо тепер, що інша випадкова вибірка дала середнє $\mu = 135$. Якщо об'єм вибірки був достатньо великим, то можна показати таке: якщо початкова гіпотеза правильна, то ми практично ніколи не отримали б подібного результату. На основі цього висновку отриманий результат цілком обґрунтовано можна використати як підтвердження неправильності висунутої гіпотези і ризик помилки при цьому буде мінімальним.

Все сказане вище ґрунтується на тому факті, що результат експерименту, який є сумісним з висунутою гіпотезою, виявляється також сумісним з іншими гіпотезами. Це призводить до того, що подібний результат не може бути використано для обґрунтування вибору деякої гіпотези порівняно з іншими. Однак ми завжди можемо отримати результат, який розбігається з висунутою гіпотезою і може спричинити значні сумніви щодо її правильності або достовірності.

При перевірці гіпотез остаточний висновок можна зробити тільки в тому випадку, якщо ми можемо відхилити висунуту гіпотезу. Таким чином, мета експерименту повинна полягати в тому, щоб відхилити сформульовану гіпотезу.

Мета експерименту повинна полягати у тому, щоб відхилити сформульовану гіпотезу.

А це означає, що початкова гіпотеза повинна формулюватись як альтернатива тому, в що ми віримо і що ми хочемо отримати.

Якщо ми зможемо відхилити висунуту гіпотезу (довести її неправильність), то тим самим продемонструємо справедливість того твердження, в яке дійсно віримо.

1.5 Перевірка сформульованих гіпотез (4-й етап статистичного аналізу)

На даному (четвертому) етапі ми перевіряємо істинність гіпотез, висунутих на другому етапі. Це можуть бути гіпотези стосовно статистичної значущості отриманих оцінок невідомих величин, тобто, оцінок параметрів статистичних і математичних моделей.

Іншими прикладами гіпотез, які можна було б перевірити за допомогою статистичних даних, можуть бути такі:

– Ви переможете на виборах, які відбудуться через 10 днів.

– Похибка оцінки параметра є меншою деякої величини.

– Рівень виробничого браку є меншим, ніж його очікують споживачі.

При перевірці гіпотез велике значення має мінімізація можливості появи ситуацій, коли випадкова збіжність ряду факторів може призвести до

відхилення правильної гіпотези. Тому перед тим як відхилити гіпотезу експериментатор вимагає, щоб ймовірність отримання відповідного вибіркового значення була дуже малою.

В деяких областях науки прийнято відхиляти гіпотезу тільки у тих випадках, коли випадкове вибіркове значення може зустрічатись не частіше ніж 5 разів на 100 експериментів. В інших областях гіпотези відхиляються, якщо ймовірність появи відповідного вибіркового значення не перевищує 0,01. Очевидно, що експериментатор повинен прямувати до того, щоб ймовірність появи вибіркового значення, яке вказує на неправильність висунутої гіпотези, була досить малою.

Ймовірність появи вибіркового значення, яке вказує на неправильність сформульованої гіпотези, вибирається експериментатором і називається *рівнем значущості* експерименту.

Вибір рівня значущості *повинен відбуватись до збору експериментальних даних*, оскільки результати експерименту не повинні впливати на величину вибраного критерію. Після визначення рівня значущості і обробки даних він припускає, що його гіпотеза є правильною і визначає – більшою чи меншою вибраного рівня значущості буде ймовірність отриманого результату.

Якщо ймовірність отриманого результату перевищить рівень значущості, то експериментатор не зможе відхилити висунуту гіпотезу, справедливо вважаючи, що даний результат є в достатній мірі сумісним з нею. Наприклад, припустимо, що вибрано рівень значущості 0,05, а в результаті експерименту отримано 52 орли при 100 підкиданнях монети. Якщо експериментатор встановить, що відхилення в два орли від очікуваного значення зустрічається більше ніж в 5% випадків, то він не зможе відхилити висунуту гіпотезу.

Якщо розраховане значення ймовірності деякого результату виявилось меншим рівня значущості, то висунуту гіпотезу можна відхилити. Наприклад, припустимо, що при використанні рівня значимості 0,05 підкидання монети призвело до 96 орлів, а розрахунок показав, що відхилення цього значення від очікуваного має ймовірність появи менше 0,05.

Логіка, якою керується експериментатор, полягає в такому: “*Після формулювання даної гіпотези я отримав такий неймовірний результат, що не можу в нього повірити*.” Після цього експериментатор може вважати своє початкове припущення (теорію) доведеною і, таким чином, може сподіватись, що завжди буде отримувати результати, ймовірність появи яких при даних обставинах буде меншою рівня значущості.

Вибір рівня значущості означає також, що введено визначене правило прийняття і неприйняття гіпотез. *Рівень значущості показує ймовірність відхилення деякого показника від його очікуваного значення, при якому*

дослідник може відхилити висунуту гіпотезу.

Вибраний *рівень значущості вказує на точне значення ймовірності помилки першого роду* у випадку, якщо гіпотеза дійсно правильна. Наприклад, якщо рівень значущості дорівнює 0,01, то це означає, що у випадку правильності гіпотези в одному випадку із 100 буде робитись помилка першого роду на основі отриманого результату.

Іншими словами, *рівень значущості означає величину ризику зробити помилку першого роду*. Чим менший рівень значущості, тим меншою є ймовірність припущення помилки першого роду (відхилити правильну гіпотезу). Однак, чим менший рівень значущості, тим більшою є ймовірність помилки другого роду, якщо гіпотеза виявиться помилковою. Таким чином, *вибір рівня значущості означає вибір правила прийняття рішення при перевірці гіпотези*.

Сформулюємо процедуру перевірки гіпотез у вигляді наведених нижче кроків.

1. *Формулювання нульової гіпотези, яку необхідно перевірити*. Відхилення сформульованої гіпотези дає можливість вважати початкові припущення (теорію) правильною. (У прикладі з монетою метою може бути встановлення факту неповноцінності монети. Таким чином, необхідно перевірити гіпотезу про те, що монета є повноцінною.)
2. *Вибір рівня значущості*. (Наприклад, рівень значущості 5% означає, що ймовірність припуститись помилки першого роду складає 0,05.)
3. Виконання експерименту і обчислення необхідного статистичний параметру.
4. Визначення ймовірності відхилення обчисленого значення статистичного параметра від його очікуваного значення, припускаючи, що сформульована гіпотеза є вірною.
5. Якщо за припущення істинності гіпотези розрахунки показують, що ймовірність відхилення отриманого вибіркового статистичного показника від очікуваного значення перевищує рівень значущості, то *відхилити висунуту гіпотезу неможливо*.

Якщо за припущення стосовно істинності гіпотези розрахунки показують, що ймовірність відхилення отриманого вибіркового статистичного показника від очікуваного значення є меншою рівня значущості, то *висунута гіпотеза відхиляється*.

Приклад 1.2 (задача другого типу). Чи буде середній коефіцієнт розумового розвитку (КРР) 64-х *десятилітніх хлопчиків, які збираються стати інженерами*, відрізнятись від середнього генеральної сукупності. Нехай дослідження виконується в місті, в якому середній КРР десятилітніх хлопчиків $\mu = 100$, а стандартне відхилення $\sigma = 20$.

Нехай необхідно показати, що десятилітні хлопчики, які збираються стати інженерами, відрізняються за рівнем свого інтелектуального розвитку від середнього генеральної сукупності. За нуль-гіпотезу прийнято:

$$H_0 : \textit{вибрана група} - \textit{типовий представник генеральної сукупності};$$

$$H_1 : \textit{вибрана група відрізняється від генеральної сукупності}.$$

Виберемо за рівень значущості 0,05 і будемо сподіватись, що зможемо відхилити висунуту гіпотезу. Нехай із генеральної сукупності вибрано 64 хлопчики, які збираються стати інженерами. Встановлено, що середній КРР для них складає: $\mu_{гр} = 108$.

Ми вважаємо, що нуль-гіпотеза є вірною, тобто, їхні 64 КРР можна розглядати як випадкову вибірку із великої сукупності спостережень і таку, що представляє собою одну з численних вибірок потужністю 64, кожна. З теорем для середнього ряду розподілу, утвореного з середніх значень (будуть розглянуті нижче) випливає, що при таких умовах розподіл середніх значень цих вибірок буде мати середнє $\mu = 100$, а стандартне відхилення

$$\sigma_{гр} = \frac{\sigma}{\sqrt{N}} = \frac{20}{\sqrt{64}} = 2{,}5.$$

Цей розподіл представлено на рис. 1.1, на якому середнє $\mu_{гр} = 108$ позначено хрестиком.

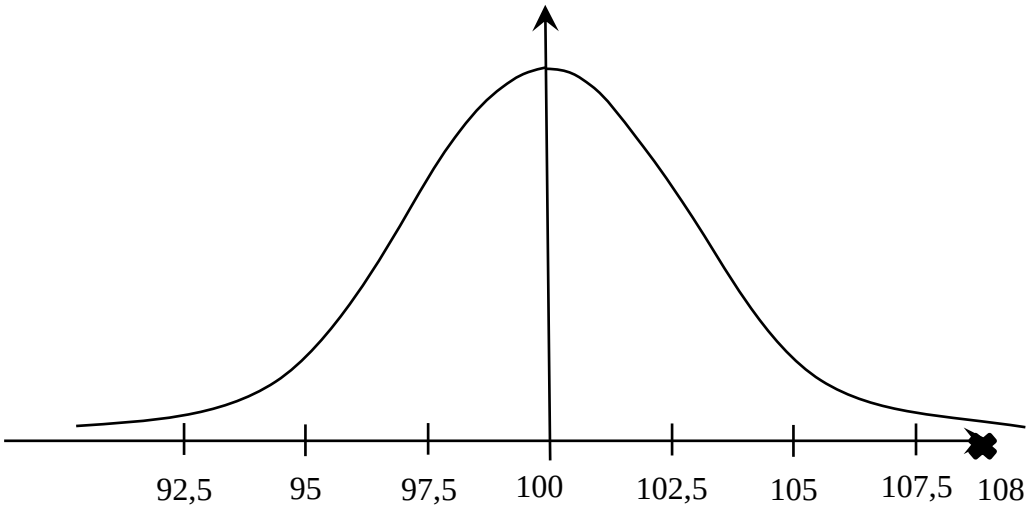


Рис. 1.1 Форма розподілу для КРР при $\mu_{гр} = 108$

З рис. 1.1 видно, що за припущення істинності нульової гіпотези ми отримали вибіркове середнє, величина якого перевищує точку рівноваги нормального розподілу на 8 пунктів, а стандартне відхилення випадкової вибірки складає 2,5. Іншими словами, припускаючи істинність нульової гіпотези, ми повинні зробити висновок, що отримане вибіркове середнє на 3,2 стандартного відхилення перевищує очікуване значення середнього, тобто 100.

З таблиці для площ нормального розподілу знайдемо, що ймовірність отримання значення, z – оцінка якого відрізняється більше, ніж на 3 одиниці (в обидва боки) від точки рівноваги, складає менше 0,001. Таким чином, припустивши істинність нульової гіпотези, ми отримали вибіркове значення, яке настільки сильно відрізняється від очікуваної величини, що ймовірність його появи є меншою 0,001. Очевидно, що такий результат свідчить про те, що ми повинні відхилити нульову гіпотезу і прийняти альтернативну. Тобто хлопчики, які у майбутньому збираються працювати інженерами, в середньому відрізняються за рівнем свого інтелектуального розвитку від інших хлопчиків їхнього віку.

В даному випадку було б корисно застосувати формулу обчислення $z-$оцінки вибіркового середнього розподілу вибіркових середніх. Ця формула дозволяє швидше розв'язувати подібні задачі. Так, $z-$оцінка вибіркового середнього обчислюється за виразом:

$$z = \frac{\textit{Оцінка відхилення середнього}}{\textit{Стандартне відхилення середніх}} = \frac{\bar{x} - \mu}{\sigma / \sqrt{N}}, \qquad (1.4.1)$$

де $\bar{x}-$ значення вибіркового середнього (в нашому прикладі $\bar{x} = 108$); $\mu-$ середнє розподілу вибіркових середніх (ми припускали, що $\mu = 100$); $\sigma-$ стандартне відхилення КРР для генеральної сукупності (в нашому прикладі $\sigma = 20$).

Підставляючи конкретні значення в формулу (1.4.1), отримаємо:

$$z = \frac{\bar{x} - \mu}{\sigma\sqrt{N}} = \frac{108 - 100}{20\sqrt{64}} = 3{,}2.$$

Звертаючись тепер до таблиці площ нормального розподілу (при $z = 3{,}2$), відхиляємо нуль-гіпотезу на рівні значимості 0,05.

Приклад 1.3. (задача третього типу). *Нехай, необхідно перевірити гіпотезу стосовно того, що студенти п'ятого курсу ІПСА НТУ КПІ є типовими представниками всіх студентів п'ятого курсу за вмінням виконувати курсові проекти* з проектування комп'ютерних інформаційних систем.

Перевіримо нульову гіпотезу для даного випадку на рівні значущості 0,01 (1% процент). За основу взято результати виконання курсових проектів у всіх університетах Києва. Середня оцінка за курсовий проект по місту виявилась рівною $\mu = 72$ (при максимальному значенні 100); стандартне відхилення $\sigma = 12$.

В групі ІПСА, яка аналізувалась, налічувалось $N = 36$ студентів при середньому значення оцінки $\mu_{гр} = 74$. Тобто, необхідно визначити чи суттєво відрізняється середнє для вибраної групи від середнього генеральної сукупності.

Припустимо, що сформульована гіпотеза вірна (студенти ІПСА є типовими представниками студентів 5-го курсу по Києву) і що отримані 36 оцінок за курсові проекти можна розглядати як випадкову вибірку із всієї сукупності оцінок. Це означає, що ми можемо розглядати вибіркове середнє для групи студентів, $\mu_{гр} = 74$, як одне із значень теоретичного розподілу середніх різних вибірок об'ємом $N = 36$, кожна. У відповідності з наведеною нижче теоремою 1.1 (див. нижче) розподіл таких вибіркових середніх є нормальним. При цьому середнє розподілу середніх дорівнює 72 (таке ж значення має середнє генеральної сукупності), а стандартне відхилення:

$$\sigma_{рс} = \frac{\sigma}{\sqrt{N}} = \frac{12}{\sqrt{36}} = 2.$$

Таким чином, отримане вибіркове середнє для вибраної групи студентів, $\mu_{гр} = 74$, на $z = 1$ перевищує теоретично очікуване значення, тобто:

$$z_{гр} = \frac{\bar{x} - \mu}{\sigma / \sqrt{N}} = \frac{74 - 72}{12 / \sqrt{36}} = 1{,}0.$$

У відповідності з таблицею площ нормального розподілу, *ймовірність розбіжності між середнім генеральної сукупності і середнім вибраної групи студентів* є більшою вибраного нами рівня значимості, а тому *нуль-гіпотеза приймається*. Тобто середнє генеральної сукупності і середнє вибраної групи відрізняються несуттєво.

Таким чином, на основі виконаного експерименту *не можна зробити висновок, що студенти ІПСА НТУУ КПІ відрізняються за рівнем виконання курсових проектів* з проектування інформаційних систем від студентів інших університетів.

Теорема 1.1. *Ряд розподілу, утвореного з середніх значень великої кількості випадкових вибірок потужністю* N, *взятих з однієї нескінченної сукупності, має стандартне відхилення* (*стандартне відхилення середніх або стандартна похибка середнього*), *яке дорівнює стандартному відхиленню генеральної сукупності, діленому на корінь квадратний з* N, *тобто*:

$$\sigma_{рс} = \frac{\sigma_{гс}}{\sqrt{N}} = \frac{\sigma}{\sqrt{N}},$$

де $\sigma_{гс} = \sigma -$ *стандартне відхилення генеральної сукупності, з якої взято велику кількість вибірок об'ємом* N.

Приклад 1.4. Розглянемо, наприклад, коефіцієнти розумового розвитку (КРР) чоловіків, які проживають в місті Києві, середнє цієї сукупності $\mu = 100$, а стандартне відхилення $\sigma = 15$. Припустимо, що з цієї генеральної сукупності взято велику кількість вибірок по 10 елементів у кожній.

Сформуємо новий розподіл із середніх цих вибірок і знайдемо стандартне відхилення вибіркового розподілу середніх:

$$\sigma_{рс} = \frac{\sigma}{\sqrt{N}} = \frac{15}{\sqrt{10}} = 4{,}7.$$

Графік розподілу середніх є компактнішим, оскільки стандартне відхилення середніх $\sigma_{рс} = 4{,}7$, а стандартне відхилення генеральної сукупності $\sigma = 15$ (рис. 1.2).

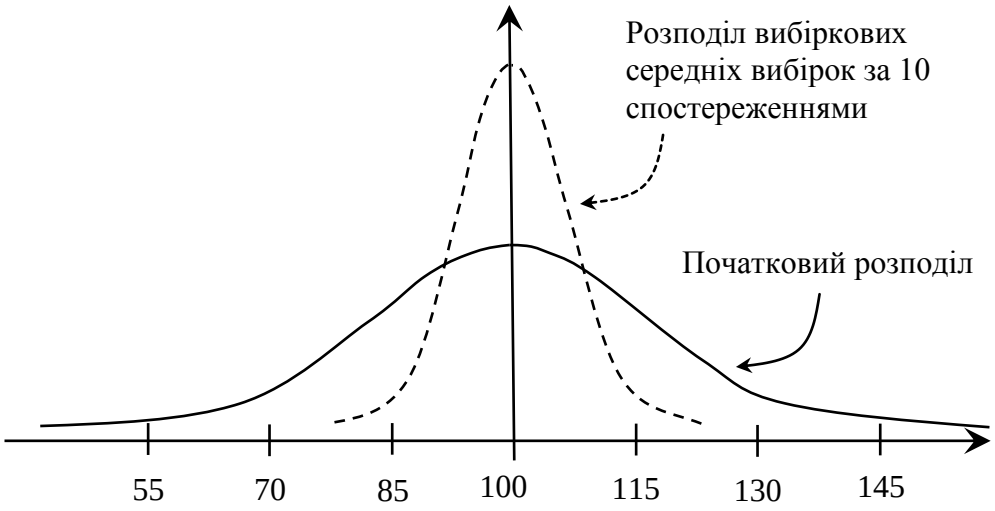


Рис. 1.2. Криві розподілу генеральної сукупності і розподілу середніх

Іншими словами, середні значення вибірок виразніше групуються навколо точки рівноваги ніж індивідуальні коефіцієнти розумового розвитку. Таким чином, середні значення вибірок в меншій мірі відрізняються одне від одного, ніж індивідуальні КРР.

Це явище можна пояснити тим, що завдяки усередненню відбувається процес "балансування" вибірки, в результаті якого середнє вибірки приймає значення, яке наближається до 100, тобто, до середнього генеральної сукупності. При цьому, чим більшою буде потужність вибірки, тим ближчими одне до одного будуть значення середніх, тобто, зменшується стандартне відхилення розподілу середніх σ_{pc}.

На рис. 1.3 показано три ряди розподілу. Перший ряд є рядом розподілу індивідуальних КРР із середнім 100 і стандартним відхиленням 15. Другий ряд представляє собою вибірковий розподіл середніх значень вибірок при $N = 10$, взятих з генеральної сукупності КРР. Цей розподіл має $\mu_{pc} = 100$ і $\sigma_{pc} = 4{,}7$.

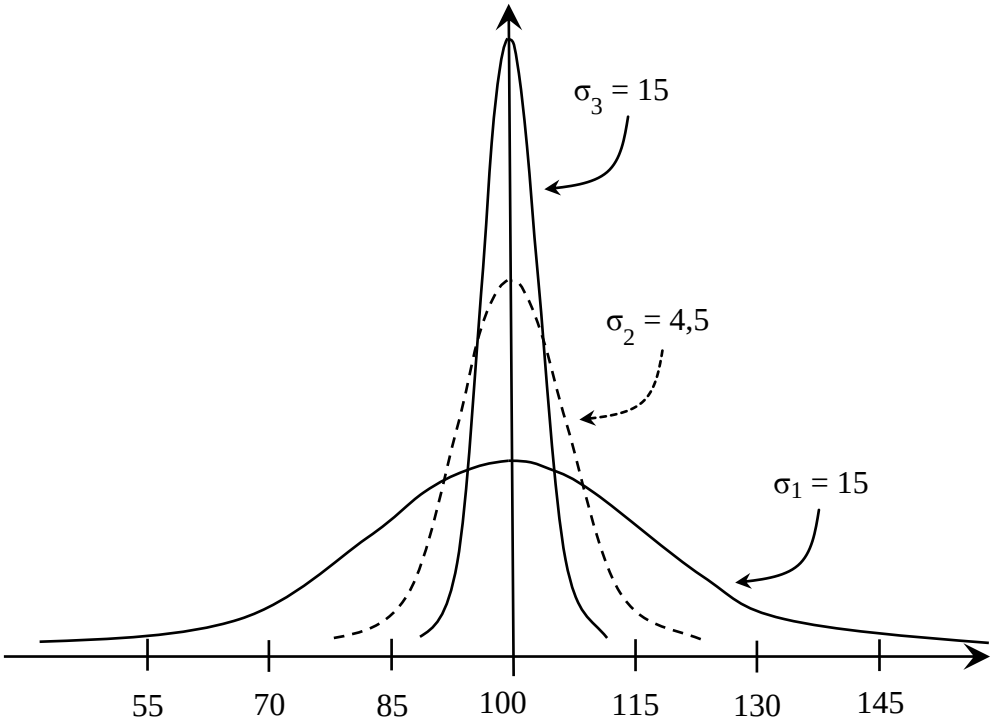


Рис. 1.3. Криві розподілу при різних стандартних відхиленнях

Природно, що даний розподіл є компактнішим порівняно з розподілом індивідуальних КРР. Друга крива ілюструє розподіл середніх значень вибірок потужністю $N = 100$ кожна. Стандартне відхилення складає $\sigma_{pc} = 1{,}5$; цей розподіл самий "гострий", тобто середні значення вибірок потужністю $N = 100$, кожна, є дуже близькими одне до одного.

1.6 Суцільний та вибірковий методи збору даних

Будемо називати набір значень деякої змінної ***рядом розподілу*** або розподілом значень цієї змінної. Наприклад, множина оцінок, отриманих на екзамені або ряд числових значень, що характеризують зміну в часі рейтингу кандидата у президенти. В останньому випадку кажуть, що існує часовий ряд даних, що характеризує вибрану змінну на деякому (вибраному) часовому інтервалі. При цьому дані можуть накопичуватись через однакові (постійні) або

різні (змінні) часові інтервали, які називають ***періодами дискретизації*** даних (вимірів). У більшості випадків статистичні (експериментальні) дані накопичують через однаковий часовий інтервал, що спрощує їх подальшу обробку. Якщо ж період дискретизації змінюється у часі, то такі дані потребують застосування спеціальних методів обробки і спеціальних моделей для опису.

Існує *два основних підходи до збору експериментальних або статистичних даних*:

- метод суцільних спостережень;
- вибірковий метод.

Метод суцільних спостережень

Метод *суцільних спостережень* полягає у визначенні максимально можливого об'єму спостережень та формуванні бази вимірів із включенням в неї всіх елементів даної сукупності. Таку базу даних називають *генеральною сукупністю*. Наприклад, для визначення успішності студентів з прикладної статистики в КПІ формується база даних, яка вміщує оцінки всіх студентів всіх факультетів і груп, які вивчають прикладну статистику.

Теорія вибіркового методу передбачає формування статистичного висновку щодо всієї можливої сукупності даних на основі аналізу деякої обмеженої вибірки, взятої з генеральної сукупності.

Наприклад, статистичний висновок стосовно успішності студентів з прикладної статистики (ПС) в університеті формується на основі вибірки оцінок студентів, взятих з тих факультетів, де вивчають ПС. Можна взяти випадкову вибірку, яка включає по 10 студентів з кожної групи і знайти середню оцінку для сформованої таким чином вибірки.

Вибірковий метод є основним методом статистичного аналізу, оскільки, як правило, формування генеральної сукупності є фізично неможливим або занадто дорогим. Крім того, результати дослідження вибірок із генеральної сукупності можна коректно узагальнити на генеральну сукупність за допомогою відповідних обчислювальних процедур.

Випадковий і типовий відбори даних

Для того, щоб правильно виконувати статистичні дослідження, необхідно розуміти можливості статистичних методів. Якщо необхідно зробити узагальнюючі висновки щодо аналізу на основі деякої вибірки даних, остання повинна бути *представницькою* (репрезентативною), тобто, *вона повинна представляти*:

- *всі соціальні групи населення*, якщо, наприклад, виконується визначення рейтингу політичної фігури;
- *представників студентів всіх курсів і всіх спеціальностей*, якщо аналізується успішність студентів університету в цілому;

– *повні дані щодо визначеного режиму роботи об'єкта*, який досліджується (наприклад, перехідного режиму).

Синонімом слова представницька є *репрезентативна* від англійського *representative*. Наприклад, вибірка студентів не буде представницькою, якщо ми включимо тільки половину факультетів, де вивчають ПС. Може виявитись, що друга половина, яка не включена в вибірку, суттєво відрізняється від першої. Якщо виконується аналіз рейтингу кандидатів у президенти, то у вибірку необхідно включити всі наявні соціальні групи населення. Інакше опитування може призвести до так званих зміщених оцінок, тобто оцінок, далеких від реальності.

У випадках, коли необхідно виконати статистичний аналіз роботи *технічного об'єкта*, дані повинні повністю охоплювати спостереженнями ті режими його функціонування, що цікавлять дослідника. Наприклад, якщо це *перехідний режим*, то необхідно так спланувати експеримент, що перехідний процес був повністю відображений зібраними числовими даними.

Існує два підходи до формування вибірок даних з генеральної сукупності:

– *власне випадкова вибірка*, коли кожна одиниця сукупності має однакові шанси бути відібраною для аналізу;

– типовий відбір.

Випадкову вибірку студентів можна сформувати наступним чином: згенерувати випадкові числа в діапазоні 1-30 (за числом студентів в групі) і взяти перших кілька чисел або написати на однакових листочках прізвища всіх студентів факультету, які вивчають деякий предмет, і наугад витягати ці листочки. В такому випадку кожна одиниця сукупності має однаковий шанс щодо включення в вибірку.

Різновидом такої техніки відбору є *типовий відбір*. Техніка цього методу гарантує, що вибірка буде містити такі ж пропорції елементів різних груп, як і в генеральній сукупності, з якої вони взяті. Наприклад, нехай нас цікавить думка студентів університету щодо загального рівня навчання (підготовки) в даному університеті.

Типовий відбір передбачає, що у сформованій вибірці будуть представлені частини (пропорції) студентів з кожного курсу і факультету і що *ці пропорції будуть такими ж, як і в генеральній сукупності*.

Тобто, якщо на першому курсі навчається 20% студентів університету, то першокурсники повинні скласти 20% вибірки даних. Якщо рівень компетентності студентів різний на різний на різних курсах, то це також необхідно врахувати у вибірці.

1.7 Зловживання статистикою

Статистика – *надзвичайно могутній інструмент аналізу даних різної природи і її силу можна використати коректно або некоректно*. Є багато

прикладів некоректного застосування статистичних методів, що проявилось у формулюванні висновків, протилежних до дійсності. Тому коректність застосування методів статистичного аналізу даних – це ключовий момент, який забезпечує правильність остаточного результату.

Наприклад, у 1936 році в США журнал "*Literary Digest*" помилково визначив вибірку даних для аналізу. Журнал виконав телефонне *опитування своїх передплатників* і спрогнозував, що *Альфред Лендон* легко переможе свого суперника *Франкліна Рузвельта*. Однак, Рузвельт переміг Лендона в 46 із 48 штатів. Справа в тому, що 1936 рік – це один з років економічної депресії і дозволити собі телефон і передплату журналу могли тільки люди з хорошим матеріальним станом. Тобто вибірка, визначена журналом, не представляла всіх прошарків виборців в США – вона *не була представницькою*. Передплатники журналу у своїй більшості збирались голосувати за Лендона, в той час як основна частина генеральної сукупності була за Рузвельта. Після виборів журнал "*Literary Digest*" швидко втратив свою популярність і досить швидко перестав існувати.

Звичайно, можна навести багато інших прикладів некорректного застосування статистичних методів і не тільки у минулому, а й сьогодні. На жаль, бувають випадки, коли некоректне застосування цих методів зумовлене не тільки поганими знаннями, але й цілеспрямованим навмисним формуванням некоректних висновків на основі зібраних даних.

Очевидно, що майже всі статистичні процедури можна виконати некоректно як в результаті незнання, так і навмисно. Однак, як свідчить практика, навмисне спотворення методів формування статистичного висновку призводить тільки до негативних результатів. Якщо ми будемо стверджувати, що зростання ВВП України складає 12% на рік, а фактичне зростання складає 4-5%, то добробут населення і довір'я до влади від цього не підвищаться.

Зазначимо, що невірні статистичні дані і некоректно виконаний статистичний аналіз можуть призвести до прийняття принципово неправильних тактичних і стратегічних рішень з катастрофічними наслідками.

Іноді говорять, що за допомогою статистики можна довести, що завгодно. Однак, якраз це "що завгодно" пов'язане з випадками некоректного застосування статистичних методів. Деякі компанії виконують дослідження, за допомогою яких доводять, що їхній продукт переважає за своїми якостями аналогічні продукти, що випускаються конкуруючими фірмами. Очевидно, що висновки будь-якого дослідження можна підігнати під бажаний результат за рахунок зміни техніки збору даних або неправильної інтерпретації встановлених залежностей. У повсякденному житті кожний з нас зустрічається із статистичними даними та висновками, що знаходяться в протиріччі між собою. Тому немає нічого дивного у тому, що багато людей висловлюють недовіру до статистики. Зазначимо, що те ж саме може стосуватись будь-якої науки.

Однак, у випадках зловживання статистикою *винні не цифри і не використані методи, а ті виконавці*, які мають схильність обманювати суспільство за їх допомогою. Чесні наміри і знання статистичних методів завжди дають корисні змістовні результати.

Іноді можна зустріти заперечення проти застосування статистичних методів у зв'язку з тим, що отримувані за ними висновки носять узагальнений характер і не відображають індивідуальних якостей людей. Так, висновок стосовно того, що індивідууми з вищою освітою матимуть деякий рівень середньої зарплати, не сприймається деякими людьми. Вони вважають, що їх це не стосується і що висновок невірний. Так само як і у випадку голосування за деякого визначеного кандидата в президенти не всі представники однієї соціальної групи голосуватимуть за одного кандидата.

З цього приводу можна сказати, що статистичні висновки є узагальнюючими, але вони *не абсолютні*. Загальне судження ніколи не залишається незмінним. Якщо ми будемо наводити приклади, що будуть у протиріччі з висновком, то тим самим будемо надавати висновку *форму абсолютного судження*. Розв'язок цієї проблеми полягає у тому, що кожний індивідуум повинен сприймати справедливість статистичного судження, але в той же час повинен розуміти, що ніхто не забирає у нього свободи вибору за допомогою цього судження.

Загалом можна сказати, *що основною метою застосування статистичного аналізу є обґрунтоване прийняття рішень на основі аналізу даних*. Якщо ми хочемо, щоб рішення були об'єктивними і якісними, то необхідно знати і вміти користуватись методами статистичного аналізу та іншими методами прийняття рішень, що доповнюють його. Доповнюючими методами є *методи оптимізації*, *експертне оцінювання альтернатив, нечіткі множини, нейромережі, байєсівський підхід* та інші методи. Зазначимо, що сукупне застосування кількох методів аналізу даних і прийняття рішень, а також відповідних критеріїв аналізу їх якості значно підвищує ймовірність отримання кращих рішень з множини можливих альтернатив.

1.8 Контрольні запитання і вправи

1. Які задачі можна розв'язувати методами прикладної статистики?
2. Назвіть чотири етапи статистичного аналізу даних?
3. Що означає генеральна сукупність даних?
4. Яку вибірку даних називають випадковою?
5. Які моделі називають математичними і які статистичними?
6. Назвіть які існують два основних види статистичних даних?
7. Які існують причини появи випадкової складової у вимірах (статистичних даних)?
8. Яка складова вимірів дає можливість будувати математичні моделі досліджуваних процесів?

9. Які задачі необхідно розв’язують на етапі планування збору даних?
10. Яка мета попередньої обробки і дослідження даних?
11. Дайте означення тренду? Що означають терміни "інтегрований" і "процес з одиничними коренями"?
12. Для чого виконується візуальний аналіз даних?
13. Які процеси називають гетероскедастичними? Що означає волатильність процесу?
14. Поясніть такі статистики: критерій узгодженості χ^2, асиметрію, ексцес?
15. Як розраховується статистика Жак-Бера? Що означають обчислені значення?
16. Дайте означення прогнозу? Запишіть функцію прогнозування для заповнення пропусків даних з використанням рівняння авторегресії з ковзним середнім АРКС(1, 1):

$$y(k) = a_0 + a_1\, y(k-1) + b_1\, \varepsilon(k-1) + \varepsilon(k).$$

17. Які значення відносять до екстремальних? Як вони впливають на статистичні характеристики вибірки даних?
18. Охарактеризуйте особливості перехідного та усталеного режимів функціонування досліджуваних об’єктів?
19. Які методи заповнення пропусків даних ви знаєте?
20. Що таке нормування даних і яким чином воно виконується? Чи можна виконувати нормування даних з використанням нелінійних перетворень?
21. Яка мета кореляційного аналізу даних? Яким чином розраховується вибіркова дискретна функція взаємної кореляції двох змінних?
22. Що таке автокореляція? Яка причина існування ненульової автокореляції? Як розраховується автокореляційна функція?
23. Як розраховується часткова автокореляційна функція? У чому полягає відмінність часткової АКФ від звичайної?
24. Які елементи включає у себе поняття структури математичної моделі? Яким чином можна оцінити порядок моделі авторегресії за умови наявності статистичних даних у вигляді часового ряду?
25. Які методи оцінювання параметрів математичних і статистичних моделей широко вживаються на практиці?
26. Який критерій оптимальності використовують при оцінюванні параметрів за методом найменших квадратів (МНК)?
27. Наведіть вираз для оцінювання параметрів за МНК, поясніть всі змінні, які в нього входять? Як формується матриця вимірів при оцінюванні моделі множинної регресії?
28. Оцінки параметрів (коефіцієнтів) моделі, які отримують за допомогою МНК, – це випадкові чи детерміновані величини? Дайте ґрунтовне пояснення.
29. Який критерій оптимальності використовують при застосуванні методу максимальної правдоподібності? Наведіть приклад функції правдоподібності?
30. Як потрібно формулювати нуль-гіпотезу при застосуванні теорії перевірки гіпотез для аналізу отриманих результатів статистичного дослідження?

31. Опишіть і поясніть послідовність дій при перевірці гіпотез у загальному випадку?
32. Опишіть і поясніть послідовність дій при перевірці гіпотези стосовно значущості оцінок параметрів регресійної моделі у статистичному смислі?
33. Який смисл рівня значущості при перевірці гіпотез?
34. Що означають помилки першого і другого роду при перевірці гіпотез? Поясніть на прикладі.
35. Що означає термін "стандартна похибка" у регресійному аналізі даних? Як вона обчислюється? Чи пов’язана стандартна похибка з точністю оцінювання параметрів регресійних моделей?
36. Що означає "рівень значущості експерименту"? Виходячи з яких міркувань його вибирають?
37. Як пов’язаний рівень значущості експерименту з помилками першого і другого роду?

Розділ 2

Мода, медіана, середнє і варіація

2.1 Означення

Найбільш відомими статистичними параметрами, що характеризують середину ряду розподілу, є *середнє*, *медіана* і *мода*. Розглянемо простий набір із семи елементів або спостережень:

4 5 5 7 8 10 17.

О з н а ч е н н я 2.1. ***Мода*** *– це значення елемента, яке зустрічається найчастіше*. У даному випадку мода дорівнює 5, оскільки це значення зустрічається 2 рази, тобто,

Мода = Мд = 5.

Деякі ряди розподілу не мають моди, оскільки всі елементи зустрічаються в них тільки по одному разу.

Разом з тим ряд може бути *бімодальним* і *багатомодальним* (мультимодальним), наприклад, ряд

9 3 3 8 7 9 2

має дві моди: 3 і 9.

О з н а ч е н н я 2.2. ***Медіана*** *– це значення елемента, який більше або дорівнює і одночасно менше або дорівнює половині всіх інших елементів ряду розподілу*.

У даному прикладі медіаною є 7, оскільки воно перевищує значення перших трьох елементів і одночасно є меншим трьох останніх:

Медіана = Ме = 7.

Значення медіани і моди *не залежать від того*, *упорядковані значення* в ряду розподілу, чи ні.

У випадку парної кількості елементів не існує дійсної медіани, тобто немає елемента, який був би одночасно більше і менше рівно половини інших елементів ряду.

Незалежно від того, який елемент ми виберемо, залишиться непарне число елементів. У такому випадку можна знайти *два центральних елементи* і визначити медіану, як їх середнє.

О з н а ч е н н я 2.3. ***Середнім арифметичним*** *є значення, отримане в результаті додавання всіх елементів ряду з наступним діленням суми на загальну кількість елементів*:

$$\text{Середнє} = \frac{4+5+5+7+8+10+17}{7} = 8.$$

Середнє – це найважливий статистичний параметр (статистика) деякої множини значень. Покажемо, що це так, на простому прикладі. Збільшимо 5-й елемент наведеного вище ряду на 7:

4 5 5 7 15 10 17.

Тепер ряд має такі параметри: Медіана = 7; Мода = 5; Середнє = 9.

Таким чином, мода і медіана залишились незмінними, а середнє збільшилось на одиницю. Якщо додати 1 до 4-го елемента останнього ряду, то Медіана = 8, але і середнє при цьому збільшиться.

Таким чином, середнє реагує на *кожну зміну* значень ряду, в той час як мода і медіана тільки на *деякі*.

2.2 Дві властивості середнього

1. Середнє значення будь-якої вибірки даних – це його *єдина точка рівноваги*.

Якщо, наприклад, числові значення вибірки даних представляють собою вагу деякого набору гирьок, то середнє буде точкою рівноваги. В деяких випадках в економіці середнє вважають середньостроковим або довгостроковим прогнозом (умова довгострокової рівноваги).

У випадку, коли розглядають *теоретичне середнє* деякого випадкового процесу, то його називають *математичним сподіванням*. Тобто для економічного процесу, що описується змінною $y(k)$, прогнозом буде його *математичне сподівання*:

$$E[y(k)] = \mu_y,$$

де E – оператор математичного сподівання або визначення середнього; k – дискретний час; μ_y – математичне сподівання змінної $y(k)$, яке можна обчислити різними способами.

Одним із способів є, наприклад, математичний опис ряду даних рівнянням авторегресії другого порядку

$$y(k) = a_0 + a_1 y(k-1) + a_2 y(k-2) + \varepsilon(k)$$

і знаходження розв'язку цього рівняння, який дасть можливість обчислити математичне сподівання.

2. *Сума відхилень від середнього дорівнює нулю*. Це означає, що ніяке інше число не може бути точкою рівноваги ряду розподілу.

Таблиця 2.1. Знаходження відхилень від середнього

Елементи ряду розподілу	Середнє	Різниці
4	8	-4
5	8	-3
5	8	-3
7	8	-1
8	8	0
10	8	2
17	8	9
Сума різниць		0

Теорема 2.1. *Для будь-якої множини даних суми різниць між їх значеннями та середнім дорівнює нулю. Навпаки, якщо сума різниць між цими елементами і деяким числом дорівнює нулю, то це число є середнім*.

Доведення:

$$\sum_{k=1}^{N}[x(k)-\mu_X]=\sum_{k=1}^{N}x(k)-\sum_{k=1}^{N}\mu_X=\sum_{k=1}^{N}x(k)-N\frac{1}{N}\sum_{k=1}^{N}x(k)=0,$$

оскільки $\sum_{k=1}^{N}\mu_X = N\mu_X$.

Можна легко встановити, що сума квадратів відхилень від середнього також буде мати мінімальне значення у порівнянні з іншими можливими відхиленнями. Це можна сформулювати такою теоремою.

Теорема 2.2. *Для будь-якої множини елементів сума квадратів їх відхилень від середнього приймає найменше можливе значення у порівнянні з сумою квадратів відхилень від інших точок. І навпаки, точка, в якій мінімізується сума квадратів відхилень, є середнім значенням*.

Теорему 2.2 не можна застосувати для знаходження середнього. Однак, наведений результат дуже важливий для теорії оцінювання. Його називають властивістю "*найменших квадратів*". Мінімізація суми квадратів відхилень від середнього дає можливість отримати конкретні вирази для обчислення параметрів випадкових процесів. В першу чергу це відомий метод найменших квадратів (МНК), який широко застосовується для оцінювання параметрів (коефіцієнтів) лінійних та псевдолінійних моделей (до псевдолінійних відносять моделі, нелінійні стосовно змінних, наприклад, моделі поліноміального типу).

2.3 Вплив зміни значень ряду розподілу на середнє

Теорема 2.3. *Якщо кожний елемент ряду розподілу збільшити або зменшити на деяку константу, то і середнє відповідно збільшиться або зменшиться на цю ж константу.*

Т е о р е м а 2.4. *Якщо кожний елемент ряду розподілу помножити на деяку константу, то і значення середнього помножиться на цю константу* (табл. 2.2). *Якщо кожний елемент ряду розподілу поділити на деяку константу, то і значення середньої поділиться на цю ж константу*.

Таблиця 2.2. Знаходження відхилень від середнього

Елементи ряду розподілу	Середнє	Різниці
4 x 3 = 12	8 x 3 = 24	- 12
5 x 3 = 15	24	- 9
5 x 3 = 15	24	- 9
7 x 3 = 21	24	- 3
8 x 3 = 24	24	0
10 x 3 = 30	24	6
17 x 3 = 51	24	27
Сума різниць		0

2.4 Деякі приклади застосування середнього, медіани і моди

Внаслідок чутливості середнього до зміни значень елементів ряду його можна використовувати як міру, що характеризує всі елементи досліджуваної сукупності.

Середнє можна використати як міру, що відображає наскільки великими є у своїй масі елементи даної сукупності. Одним із самих поширених методів порівняння елементів двох різних груп полягає у порівнянні їх середніх.

Так, коли ми говоримо, що середнє число попадань у гравців команди **А** в середньому є більшим, ніж у команди **Б**, то можна стверджувати, що гравці команди **А** в цілому мають вищу майстерність, ніж гравці команди **Б**.

Якщо середнє число дітей в сім'ї в одній країні складає 2,5, а в іншій 1,5, то це важливе і об'єктивне свідчення того, що в першій країні демографічна ситуація є кращою, ніж в другій. Якщо в першій країні населення поступово збільшується, то в другій воно зменшується.

Тобто, *середнє використовують для порівняння між собою груп даних* і визначення того, яка група є переважаючою за даним параметром.

В економетричному аналізі *безумовне математичне сподівання* використовують як довгостроковий прогноз. Його можна знайти, наприклад, як математичне сподівання розв'язку рівняння, що описує динаміку процесу, при $k \to \infty$ (асимптотичне значення середнього).

Умовне математичне сподівання використовують для визначення короткострокового і середньострокового прогнозу.

Середнє часто використовують для *визначення ступеня зміщення набору даних сигналу від нуля*. Так, математичне сподівання шумової складової вказує на те, чи виконується припущення щодо її центрованості.

Середнє визначають також *для того, щоб видалити його з даних і працювати тільки з відхиленнями*. Такий підхід, як правило, сприяє покращенню якості оцінок математичних моделей завдяки покращенню ступеня обумовленості матриць вимірів.

Медіана (або середній елемент вибірки) ніяк не залежить від крайніх (за значеннями) елементів що іноді робить її дуже важливим показником. Медіана дає особливо важливу інформацію щодо ряду розподілу у тих випадках, коли *відносно невелике число елементів суттєво відрізняється від загальної маси спостережень*.

Наприклад, нехай в містечку проживає 5000 жителів. Припустимо, що всі вони фактично заробляють не більше 8000 гривень на рік. Однак, власники магазинів та фірм з виробництва м'ясопродуктів і круп'яних виробів, розміщених в цьому містечку, отримують майже стільки, скільки всі інші жителі містечка разом взяті.

Медіана ряду розподілу доходів дає реальнішу картину рівня життя населення містечка, ніж середнє. Те, що на медіану не впливають великі доходи, отримувані окремими особами, перетворює її у *представницький* або *репрезентативний* показник.

Так, якщо річний медіанний рівень доходу складає 5000 гривень, то це означає, що половина жителів заробляє менше 5000 гривень на рік, а інша половина – більше 5000. При цьому середнє, на величину якого впливають великі доходи власників магазинів та підприємств, може складати 10000 гривень на рік. Таким чином, лише невелике число жителів може мати середній доход.

Значення *моди* дає важливу інформацію для виробників різних товарів, конструкторів, власників магазинів, які поставляють свої товари на конкретний ринок. Так, наприклад, виробник годинників повинен знати, за якою ціною найчастіше купляють його продукцію для того, щоб приділити увагу виробництву годинників саме такої вартості.

Власник магазину одягу повинен знати, які розміри костюмів є найбільш розповсюдженими в даному районі для того, щоб відповідно формувати запаси на складі.

Смисл показника моди полягає в тому, що він вказує на максимальну частоту значень деякого показника, яка нерідко свідчить про найбільшу популярність того чи іншого виробу або послуги.

2.5 Позначення та обчислення середнього арифметичного

В подальшому будемо користуватись такими позначеннями:

$X = \{x_1, x_2, ..., x_N\}$ або $\{x(k)\}, k = 1,...,N$ – ряд спостережень довжиною N (або потужністю N); $x_1, x_2, ..., x_N$ – елементи ряду;

μ_X – середнє значення ряду елементів *X*;

$\sum X = \sum_{k=1}^{N} x(k)$ – сума значень елементів ряду;

N_X – число значень ряду розподілу *X*; якщо розглядається один ряд елементів, то його потужність позначається просто через *N*.

Таким чином, середнє арифметичне будемо розраховувати за формулою:

$$\mu_X = \frac{\sum X}{N} = \frac{\sum_{k=1}^{N} x(k)}{N} = \frac{1}{N}\sum_{k=1}^{N} x(k).$$

2.6 Обчислення поточного середнього

Досить часто виникає необхідність обчислення так званого поточного значення середнього. Наприклад, у випадку, коли середнє змінюється у часі (середнє, що є функцією часу, називають *трендом*). Формулу для обчислення поточного середнього можна легко отримати за допомогою формули для арифметичного середнього:

$$\vec{x} = \frac{1}{N}\sum_{k=1}^{N} x(k) = \frac{1}{N}\sum_{k=1}^{N-1} x(k) + \frac{1}{N}x(N) = \frac{1}{N}\frac{N-1}{N-1}\sum_{k=1}^{N-1} x(k) + \frac{1}{N}x(N) =$$

$$= \frac{N-1}{N}\frac{1}{N-1}\sum_{k=1}^{N-1} x(k) + \frac{1}{N}x(N) = \bar{x}(k-1) - \frac{1}{N}\bar{x}(k-1) + \frac{1}{N}x(N) =$$

$$= \bar{x}(k-1) + \frac{1}{N}\left[x_N - \bar{x}(k-1)\right].$$

Таким чином, остаточно рекурсивна формула для обчислення поточного середнього має вигляд:

$$\bar{x}(k) = \bar{x}(k-1) + \frac{1}{k}\left[x(k) - \bar{x}(k-1)\right].$$

Очевидно, що цю формулу не можна застосовувати на відносно великих відрізках часу, оскільки величина $1/k$ буде прямувати до нуля при $k \to \infty$.

2.7 Математичне сподівання дискретної випадкової змінної

Математичне сподівання (МС) випадкової змінної наближено дорівнює її середньому значенню, фактично, *це очікуване значення середнього змінної*. Для розв'язку багатьох задач достатньо знати математичне сподівання.

Означення. *Математичним сподіванням дискретної випадкової змінної називають суму добутків її можливих значень на ймовірності цих значень*.

Якщо випадкова змінна X може приймати тільки значення $x_1, x_2, ..., x_n$ з відповідними ймовірностями $p_1, p_2, ..., p_n$, то математичне сподівання цієї змінної $E[X] = E[x(k)]$ визначається за виразом:

$$E[x(k)] = x_1 p_1 + x_2 p_2 + ... + x_n p_n,$$

де E – оператор математичного сподівання.

Якщо дискретна випадкова змінна X приймає злічену множину значень, то її математичне сподівання

$$E[x(k)] = \sum_{k=1}^{\infty} x(k) p(k)$$

існує, якщо ряд в правій частині є абсолютно збіжним.

З визначення випливає, що математичне сподівання дискретної випадкової змінної є не випадковою величиною, але не обов'язково постійною.

Приклад 2.1. Знайти математичне сподівання числа появ події A в одному випробуванні (досліді), якщо ймовірність події A дорівнює p, тобто $p(A) = p$.

Розв'язок. Випадкова величина X – число появ події A в одному випробуванні – може приймати тільки два значення: $x_1 = 1$ (подія A наступила) з імовірністю p і $x_2 = 0$ (подія A не наступила) з імовірністю $q = 1 - p$. Таким чином математичне сподівання числа появ події A:

$$E[X] = 1 \cdot p + 0 \cdot q = p.$$

Таким чином, *математичне сподівання числа появ події в одному випробуванні дорівнює ймовірності цієї події*. Цим результатом ми скористаємось нижче.

Математичне сподівання є більшим найменшого значення і меншим найбільшого значення випадкової величини. Тобто, її можливі значення на числовій осі знаходяться зліва і справа від математичного сподівання. В цьому смислі МС *характеризує розміщення розподілу*, а тому його часто називають *центром розподілу*.

Властивості математичного сподівання

1. Математичне сподівання постійної величини дорівнює цій величині:

$$E[c] = c.$$

2. Постійний множник можна виносити за знак математичного сподівання:

$$E[cX] = c\,E[X].$$

3. Математичне сподівання добутку двох незалежних випадкових змінних дорівнює добутку їх математичних сподівань:

$$E[XY] = E[X]\,E[Y].$$

4. Математичне сподівання суми двох випадкових змінних дорівнює сумі математичних сподівань доданків:

$$E[X+Y]=E[X]+E[Y].$$

Наслідок: *математичне сподівання суми кількох випадкових змінних дорівнює сумі математичних сподівань доданків*.

Приклад 2.2. Виконується три постріли з ймовірностями попадання в ціль, рівними відповідно $p_1=0{,}4;\ p_2=0{,}3\ i\ p_3=0{,}6$. Знайти математичне сподівання загального числа попадань в ціль.

Розв'язок. Числом попадань при першому пострілі є випадкова величина X_1, яка може приймати тільки два значення: 1 (попадання) з ймовірністю $p_1=0{,}4$ і 0 (промах) з ймовірністю $q=1-0{,}4=0{,}6$.

Математичне сподівання числа попадань при першому пострілі дорівнює ймовірності попадання (попередній приклад), тобто, $E[X_1]=0{,}4$. Аналогічно знайдемо математичне сподівання числа попадань при другому і третьому пострілі: $E[X_2]=0{,}3$ і $E[X_3]=0{,}6$.

Загальним числом попадань також є випадкова величина, яка суми чисел попадань в кожному з трьох пострілів:

$$X=X_1+X_2+X_3.$$

Таким чином, шукане математичне сподівання знайдемо за властивістю математичного сподівання суми:

$$E[X]=E[X_1+X_2+X_3]=E[X_1]+E[X_2]+E[X_3]=$$
$$=0{,}4+0{,}3+0{,}6=1{,}3 \text{ (попадань)}.$$

2.8 Інші види середнього

Далеко не у всіх випадках типові ознаки процесів можна характеризувати за допомогою середнього арифметичного. Тому застосовують інші форми середнього, наприклад, *середнє квадратичне* і *середнє кубічне*:

$$\mu_2=\sqrt{\frac{1}{N}\sum_{k=1}^{N}x^2(k)},\qquad \mu_3=\sqrt[3]{\frac{1}{N}\sum_{k=1}^{N}x^3(k)};$$

середнє геометричне

$$G=\mu_{геом}=\sqrt[n]{x_1\cdot x_2\cdot\ldots\cdot x_n}$$

та ***середнє гармонічне***

$$H=\frac{N}{\sum_{k=1}^{N}[x(k)]^{-1}}.$$

При визначенні середніх величин для значень площі і об'єму необхідно

узгоджувати середні значення для лінійних розмірів та середні для отриманих за ними значень площі і об'єму. Обчислене за лінійними розмірами середнє арифметичне, як правило, дає менші значення площі та об'єму ніж ті, які отримують за фактичними лінійними розмірами. В таких випадках обчислюють середнє квадратичне для значень лінійного параметра поверхні і середнє кубічне для значень лінійного параметра об'єму.

Приклад 2.3. Розглянемо приклад *використання середнього квадратичного* при визначенні залежності площі листків рослини *примули* від їх лінійного параметра. Форма ділянок поверхні листків може бути довільною, але однаковою або близькою в рамках вибірки, що усереднюється. Значення площі листків в рамках цієї вибірки повинні визначатись за однією формулою. Вихідні дані для аналізу наведені в таблиці 2.3.

Таблиця 2.3. Розміри листків примули

Довжина листка, см, x	Фактична площа листка, см^2, y
3	7
5	20
6	30
8	55
$\sum x = 22;\ \mu_x = 5{,}50;\ \ \mu_{x2} = 5{,}788$	$\sum y = 112;\ \mu_y = 28{,}00$

Покажемо, що користуватись середнім арифметичним при обчисленні площі листка буде некоректно. Площа листків примули з хорошим наближенням може бути обчислена за наступною емпіричною формулою:

$$y = 0{,}6728 \cdot x^{2{,}1229},$$

де y – площа листка; x – довжина листка.

Середнє арифметичне рядів x і y дорівнюють:

$$\mu_x = 5{,}50 \text{ см} \quad \text{і} \quad \mu_y = 28{,}00\ см^2.$$

Середня площа листка з використанням середнього арифметичного довжини листка буде дорівнювати:

$$y = 0{,}6728 \cdot 5{,}50^{2{,}1229} = 25{,}00\ \ см^2,$$

тобто, середнє арифметичне довжини листка не дає правильного середнього значення площі листка. Середнє квадратичне довжини листка:

$$\mu_{x2} = \sqrt{\frac{3^2 + 5^2 + 6^2 + 8^2}{4}} = 5{,}7879 \approx 5{,}788.$$

Тепер визначимо середню площу листка через середнє квадратичне:

$$y = 0{,}6728 \cdot 5{,}78790^{2{,}1229} = 28{,}000\ \ см^2,$$

тобто, це значення повністю відповідає фактичному значенню площі листків. Таким чином, середнє квадратичне точніше передає залежність площі від лінійного параметра об'єкта дослідження.

Приклад 2.4. Розглянемо приклад *використання середнього геометричного*. Наприклад, нехай нас цікавить середній за рік щомісячний темп приросту ваги птахів, яких розводять на фермі. При цьому вимірюється відносний приріст ваги за кожний місяць:

- за перший місяць – в 1,23 рази;
- за другий місяць – в 1,65 рази;
- за третій – в 1,08 рази і т.д.

В такому випадку більш адекватно середньомісячний темп приросту буде виражатись не середнім арифметичним, а середнім геометричним. Адекватність описання потрібно розуміти наступним чином. Якщо мова йде про *абсолютний приріст*, то адекватною середньою характеристикою процесу буде середнє арифметичне тому, що воно дорівнює загальному приросту за рік, поділеному на число місяців; і навпаки, для знаходження повного абсолютного значення приросту необхідно помножити середній приріст за місяць на число місяців.

Якщо мова йде про *відносний приріст*, то характер обчислень змінюється. Дійсно, нехай на початку року загальна вага птахів на фермі дорівнює 1000 *кг*. Оскільки за перший місяць вага збільшилась в 1,23 рази, то на кінець першого місяця загальна вага складе 1230 *кг*. За другий місяць вага збільшилась в 1,65 рази. Тобто, на кінець другого місяця загальна вага птахів складе $1230 \cdot 1{,}65 \approx 2030$ *кг*, а на кінець третього місяця вона складе $2030 \cdot 1{,}08 \approx 2190$ *кг*.

Обчислюючи середній за квартал щомісячний темп приросту ваги як середнє геометричне, отримаємо:

$$\mu_{геом} = \sqrt[3]{1{,}23 \cdot 1{,}65 \cdot 1{,}08} \approx 1{,}30,$$

а середнє арифметичне буде таким:

$$\mu_X = \frac{1{,}23 + 1{,}65 + 1{,}08}{3} = 1{,}32.$$

Тепер правильний загальний приріст ваги за квартал отримаємо тільки у випадку, якщо скористаємось середнім геометричним:

$$\textit{Вага в кінці кварталу} = 1000 \cdot 1{,}30 \cdot 1{,}30 \cdot 1{,}30 = 2190\,кг.$$

Приклад 2.5. *Середнє гармонічне* застосовують, зокрема, при усередненні величин, які представляють собою зміни швидкості протікання досліджуваних процесів. Наприклад, при дослідженні приросту довжини або діаметру пагінців, при усередненні індексів інфляції, або при дослідженні зміни концентрації речовин в міліграмах на літр.

Розглянемо визначення середнього гармонічного для концентрації деякої шкідливої речовини в воді озера біля промислового підприємства в *мг/літр*. Нехай виміри концентрації речовини, виконані в чотирьох кутах та посередині озера, виявились наступними:

$$X = [5 \;\; 7 \;\; 10 \;\; 13 \;\; 15].$$

Середнє арифметичне цих концентрацій складає:

$$\mu_X = \frac{5 + 7 + 10 + 13 + 15}{5} = \frac{50}{5} = 10 \;\; \textit{мг/літр}.$$

Середнє гармонічне цих же чисел дорівнює:

$$H = \frac{5}{1/5 + 1/7 + 1/10 + 1/13 + 1/15} = \frac{5}{0{,}587} = 8{,}52 \;\; \textit{мг/літр}.$$

Прямим підрахунком можна показати, що в даному випадку точніше значення має середнє гармонічне Так, загальний об’єм п’яти зразків води (по одному зразку з кожної

вказаної вище точки озера) при умові, що в кожному зразку буде міститись 100 *мг* речовини, дорівнює:

$$\frac{100}{5}+\frac{100}{7}+\frac{100}{10}+\frac{100}{13}+\frac{100}{15}=58{,}7 \quad \textit{літрів}.$$

Таким чином, фактична маса речовини в даному об'ємі дорівнює: 500 *мг*.

Загальна маса речовини, визначена в даному об'ємі 58,7 *літрів* через середнє гармонічне H, в цьому об'ємі складає:

$$58{,}7 \cdot 8{,}52 = 500{,}1 \quad \textit{мг},$$

а маса, визначена через середнє арифметичне, дорівнює:

$$58{,}7 \cdot 10 = 587 \quad \textit{мг}.$$

Тобто, середнє гармонічне дає похибку 0,1 мг, а середнє арифметичне призводить до похибки 87 *мг*.

2.9 Варіація

Елементи ряду розподілу можуть суттєво різнитись між собою і можуть бути дуже схожими. Наприклад, якщо ми виміряємо артеріальний тиск кожного з нас, то отримаємо ряд, значення якого будуть досить близькими між собою. Однак, якщо утворити ряд значень, який описує формування цін на біржові акції, то ці значення можуть дуже значно різнитись між собою. Можна сказати, що *варіація або розсіювання* даних в другому випадку є набагато більшою, ніж у першому.

Варіація є надзвичайно важливим поняттям, яке характеризує взаємозв'язок між всіма елементами ряду розподілу, взятого в цілому. Далі покажемо, що зміна значення одного елемента ряду призводить до зміни варіації ряду в цілому. Величину варіації або ступеня розсіювання елементів ряду можна, в деякій мірі, оцінити візуально, але це буде тільки якісна суб'єктивна оцінка. Необхідно мати об'єктивну числову міру варіації.

2.10 Міри варіації: середнє відхилень та дисперсія

Міра варіації повинна вказувати на степінь розсіювання елементів ряду. Вона повинна задовольняти певним вимогам. Сформулюємо ці вимоги або властивості показника.

- Значення показника розсіювання повинне бути *не дуже великим*, якщо значення елементів ряду, на основі яких він розраховується, *не дуже значно відрізняються* одне від одного. І навпаки, його значення повинне бути великим, якщо значення елементів ряду значно розсіяні.
- Друга властивість показника розсіювання полягає в тому, що його значення *не повинне залежати від числа елементів ряду*. Точніше кажучи, нам не потрібен показник розсіювання, значення якого зростало б тільки внаслідок *збільшення* числа елементів ряду. Він повинен

відображати лише подібність або розбіжність між самими числами і не залежати від їх кількості.

- Третя властивість даного показника полягає в наступному: оскільки він характеризує тільки степінь розсіювання даних, то він *не повинен залежати від значення середнього*. Показник середнього ніяк не пов'язаний з варіацією елементів і його величина не повинна впливати на величину розсіювання даних.

Перша міра варіації – середнє значення відхилень від середнього

Якщо знайти суму всіх абсолютних відхилень від середнього і поділити її на число елементів ряду, то знайдемо міру варіації яку називають *середнім відхиленням*.

Однак середнє відхилення використовується досить рідко внаслідок того, що ця міра є недостатньо інформативною. Існує інший підхід до визначення міри варіації, який ґрунтується на використанні квадратів відхилень.

Друга міра варіації – дисперсія

Для того, щоб сума квадратів відхилень не була чутливою до числа елементів ряду, суму квадратів відхилень необхідно розділити на число елементів. Отриманий показник називають дисперсією.

Означення 2.4. *Дисперсією називають середнє квадратів різниць між елементами ряду розподілу та їх середнім*.

Таким чином, дисперсію можна знайти за виразом:

$$Var[y(k)] = \sigma_y^2 = \frac{1}{N-1} \sum_{k=1}^{N} [y(k) - \mu_y]^2 ,$$

де Var – позначення дисперсії; μ_y – середнє значення ряду розподілу $\{y(k)\}$. Ділення суми квадратів відхилень від середнього на $N-1$ забезпечує незміщеність оцінки дисперсії. Якщо, скажемо, ряд *А* має більшу дисперсію ніж ряд *Б*, то варіація ряду *А* є вищою.

Скориставшись математичним сподіванням, отримаємо наступний вираз для дисперсії:

$$Var[y(k)] = E[y(k) - E(y(k))]^2 = E[Y - E(Y)]^2 .$$

Розрахункову формулу для дисперсії можна представити також в іншому вигляду:

$$\sigma_y^2 = \frac{\sum_{k=1}^{N} [y(k) - \mu_y]^2}{N-1} = \frac{\sum y^2(k) - 2\mu_y \sum y(k) + \sum \mu_y^2}{N-1} =$$

$$= \frac{\sum y^2(k)}{N-1} - 2\mu_y^2 + \left(\frac{\sum y(k)}{N-1}\right)^2 = \frac{\sum y^2(k)}{N-1} - 2\mu_y^2 + \mu_y^2 =$$

$$= \frac{\sum y^2(k)}{N-1} - \mu_y^2$$

або

$$\sigma_y^2 = \frac{\sum y^2(k)}{N-1} - \frac{(\sum y(k))^2}{(N-1)^2}.$$

Використовуючи оператор математичного сподівання, останню формулу для дисперсії можна записати у вигляді:

$$Var[y(k)] = \sigma_y^2 = E[y^2(k)] - \{E[y(k)]\}^2,$$

тобто, *дисперсія дорівнює різниці між математичним сподіванням квадрату випадкової величини* $y(k)$ *та квадратом її математичного сподівання*.

Приклад 2.6. Знайти дисперсію випадкової величини X, яка задана таким розподілом:

$$X = [2 \;\; 3 \;\; 5]; \quad p = [0{,}1 \;\; 0{,}6 \;\; 0{,}3].$$

Р о з в ' я з о к . Математичне сподівання $E[X] = E[x(k)]$:

$$E[X] = 2 \cdot 0{,}1 + 3 \cdot 0{,}6 + 5 \cdot 0{,}3 = 3{,}5.$$

Знайдемо квадрати значень змінної $X^2 = [4 \;\; 9 \;\; 25]$ і математичне сподівання $E[X^2]$:

$$E[X^2] = 4 \cdot 0{,}1 + 9 \cdot 0{,}6 + 25 \cdot 0{,}3 = 13{,}3.$$

Шукана дисперсія:

$$Var[X] = E[X^2] - \{E[X]\}^2 = 13{,}3 - (3{,}5)^2 = 1{,}05.$$

2.11 Обчислення незміщеної оцінки дисперсії

Коректне обчислення значень статистичних параметрів вибірок даних розглядається в теорії оцінювання, яка є самостійною дисципліною для вивчення. Однак деякі положення цієї теорії розглядають в курсі прикладної статистики з метою ознайомлення з методами коректного оцінювання статистичних параметрів (статистик). *Оцінити* – означає обчислити значення статистики шляхом застосування процедур оцінювання до експериментальних даних.

Наприклад, можна взяти вибірку коефіцієнтів розумового розвитку (КРР) для вибраної соціальної групи при $N = 50$ і обчислити середнє $\bar{x} = 115$. Очевидно, що знайдена оцінка є хорошим наближенням лише для даної вибірки. Якщо вибірку зменшити або розширити, то середнє арифметичне може

отримати інше значення. Можна було б, також, обчислити середнє для мінімального і максимального значень вибірки. Воно може відрізнятись від 115. Існують різні критерії вибору кращих оцінок, але найбільш важливим серед них є критерій *незміщеності*. Загалом незміщеність оцінок можна сформулювати так:

Поняття незміщеності оцінок

Оператор оцінювання параметра має властивість незміщеності, якщо середнє вибіркових оцінок, отриманих з незалежних випадкових вибірок, наближається до істинного значення параметра при необмеженому зростанні кількості вибірок.

Нехай, є 5 різних вибірок значень КРР із генеральної сукупності. Мета полягає в тому, щоб оцінити невідоме середнє значення розподілу. Для наявних п'яти вибірок вибіркові середні мають наступні значення: $\bar{x}_1 = 108$, $\bar{x}_2 = 107$, $\bar{x}_3 = 113$, $\bar{x}_4 = 115$ і $\bar{x}_5 = 105$. Середнє значення цих оцінок дорівнює 109,6.

Якщо збільшувати число вибірок, то середнє оцінок середнього буде наближатись до істинного значення середнього розподілу. Так, середнє для 5000 вибіркових оцінок середніх буде відрізнятись від істинного середнього не більше, ніж на ½ стандартного відхилення розподілу. Для того, щоб переконатись в цьому, розглянемо середнє 5000 значень середніх як середнє вибірки потужністю $N = 25000$. Якщо навіть стандартне відхилення значень КРР перевищує 20, то існує невеликий шанс, що середнє ряду, що складається з 25000 елементів, буде відрізнятись від істинного середнього значення розподілу більше, ніж на ½ стандартного відхилення, що доводиться наступною теоремою.

Теорема 2.4: *Припустимо, що з однієї нескінченної сукупності формується нескінченне число випадкових вибірок однакової потужності N. Розподіл середніх значень цих вибірок має стандартне відхилення (стандартне відхилення середніх), що дорівнює стандартному відхиленню вихідної сукупності, діленому на $\sqrt{N}$*, тобто

$$\sigma_M = \frac{\sigma}{\sqrt{N}},$$

де σ – стандартне відхилення генеральної сукупності, з якої сформовано вибірки для обчислення середніх; σ_M – стандартне відхилення середніх.

На основі сказаного можна зробити висновок, що чим більше ми маємо оцінок середніх для різних вибірок, тим ближчим буде середнє цих оцінок до істинного значення середнього розподілу. Тому говорять, що арифметичне середнє дає оцінку середнього розподілу, яка має властивість незміщеності.

Незміщена оцінка дисперсії

Розглянемо приклад з визначенням статистичних параметрів для значень КРР, наведених в таблиці 2.4. Нехай істинне середнє розподілу відоме: $\mu = 110$ (наприклад, його можна знайти за допомогою моделі процесу).

Таблиця 2.4. КРР і варіації, обчислені за допомогою відомого середнього

№ пп	КРР	$x-\mu$	$(x-\mu)^2$	Сума 5-и квадратів різниць
1	90	–20	400	
2	100	–10	100	
3	105	– 5	25	
4	145	35	1225	
5	100	–10	100	1850
6	120	10	100	
7	110	0	0	
8	100	–10	100	
9	115	5	25	
10	90	–20	400	625
11	105	– 5	25	
12	120	10	100	
13	95	–15	225	
14	130	20	400	
15	115	5	25	775

Квадрати різниць, необхідні для обчислення дисперсії, наведено в четвертому стовпчику. Якщо для оцінювання дисперсії взяти першу вибірку з 5-и значень, то оцінка дисперсії приймає значення: 1850/5=370.

Однак реальна ситуація відрізняється від наведеної тим, що істинне середнє значення розподілу невідоме. Таким чином, знову розглянемо першу вибірку КРР із 5-и значень і знайдемо для неї вибіркове арифметичне середнє: $\mu_1 = 108$. Тепер дисперсія першої вибірки з 5-и значень:

$$\sigma_1^2 = \frac{1}{5}\sum_{i=1}^{5}[x_i - 108]^2 = \frac{1830}{5} = 366,$$

тобто, отримане значення оцінки дисперсії є меншим, ніж при використанні істинного значення середнього. Цей результат не випадковий. Згідно з наступною теоремою: *для кожної вибірки сума квадратів відхилень від середнього розподілу повинна бути меншою суми квадратів відхилень від будь-якої іншої точки*.

Для наступних п’яти членів другого стовпчика таблиці 2.4 вибіркове середнє: $\mu_2 = 107$, а сума квадратів відхилень від цього вибіркового середнього дорівнює 580, тобто, є меншою 625.

Таким чином, ми переконались, що неможливо отримати незміщену оцінку дисперсії розподілу за допомогою звичайної середньої суми квадратів

відхилень від вибіркової оцінки середнього. Така оцінка буде завжди зміщеною.

Для подальшого розгляду необхідно ввести поняття *числа ступенів свободи*, яке відіграє важливу роль в теорії оцінювання.

Поняття кількості ступенів свободи

Необхідно розглянути різницю між числом незалежних фактів і загальним числом фактів, що містяться в експериментальних даних. Наприклад, нехай абітурієнт отримав 85 балів із 100 можливих, тобто, втратив 15 балів. В даному випадку у нас є три числових факти: 85, 100 і 15. Однак, з них тільки два є незалежними, оскільки третій можна обчислити з двох інших.

Факт називають незалежним від іншого або від групи інших фактів, якщо він несе нову інформацію, яку неможливо отримати з групи фактів, що порівнюється з ним.

В подальшому будемо використовувати термін *ступені свободи* для *позначення незалежної інформації*, тобто інформації, яку неможливо отримати (вивести) ні з якої іншої групи даних, що аналізуються. Таким чином, інформація щодо абітурієнта містить два ступені свободи.

Розглянемо тепер коректну процедуру оцінювання вибіркової дисперсії. Нехай є вибірка з двох даних: 90 і 100. Для них вибіркове середнє складає 95, а відхилення від середнього: –5 і +5. Нагадаємо, що сума відхилень повинна дорівнювати нулю. Таким чином, відхилення є залежними величинами, оскільки $(x_1 - \bar{x}) + (x_2 - \bar{x}) = 0$. Звідси можна зробити висновок, що вибірка з двох чисел дає тільки один факт щодо значення дисперсії. Хоча вибірка містить два елементи, оцінка дисперсії базується тільки на одному ступені свободи.

Якщо вибірка складається з N елементів і *відоме середнє розподілу* (генеральної вибірки), то для кожного з них також можна обчислити значення відхилення від середнього (наприклад, середнє можна обчислити за моделлю процесу). Таким чином, ми будемо мати N незалежних значень відхилень для обчислень дисперсії, тобто N ступенів свободи.

Якщо середнє розподілу невідоме (реальна ситуація), то обчислюючи вибіркове середнє (центральну точку), ми можемо обчислити тільки $N-1$ незалежних відхилень від цієї точки. Це зумовлено тим, що сума всіх відхилень повинна дорівнювати нулю незалежно від довжини вибірки. Наприклад, якщо сума будь-яких $N-1$ відхилень дорівнює –5, то виключене з розгляду відхилення буде дорівнювати +5. Таким чином, в загальному випадку, N різниць (відхилень) містять $N-1$ ступінь свободи і для визначення оцінки дисперсії також ґрунтується на $N-1$ ступені свободи. Отримана таким чином оцінка дисперсії буде незміщеною. Сказане може бути сформульоване у вигляді теореми.

Теорема 2.5. *Незміщена оцінка дисперсії випадкової вибірки з N елементів визначається як сума квадратів відхилень всіх членів вибірки від вибіркового середнього, поділена на $N-1$:*

$$s_x^2 = \frac{1}{N-1}\sum_{k=1}^{N}[x(k)-\bar{x}]^2,$$

де s_x^2 – вибіркова дисперсія випадкової змінної x, тобто це незміщена оцінка дисперсії генеральної сукупності σ_x^2.

Зазначимо ще раз, що дільник $N-1$ необхідно застосовувати коректно. Якщо експериментальні дані представляють собою вибірку з деякого розподілу, дисперсію якого необхідно оцінити, то для обчислення оцінки дисперсії необхідно застосовувати дільник $N-1$. В такому випадку обчислене значення s^2 являє собою вибіркову дисперсію, а s – вибіркове стандартне відхилення.

Якщо ж набір даних являє собою всю вибірку повністю, то необхідно застосовувати дільник N, а обчисленою величиною буде σ^2, а не s^2.

2.12 Властивості дисперсії

Оскільки константа не має розсіювання, то її дисперсія повинна дорівнювати нулю.

Властивість 1. *Дисперсія постійної величини дорівнює нулю*:

$$Var[c]=0.$$

Доведення. За визначенням дисперсії

$$Var[c]=E\{[c-E(c)]^2\}.$$

Оскільки математичне сподівання константи (перша властивість МС) дорівнює самій константі, то

$$Var[c]=E[(c-c)^2]=E[0]=0.$$

Властивість 2. *Постійний множник, піднесений до квадрату, виноситься за знак дисперсії*:

$$Var[cX]=c^2 Var[X].$$

Доведення. За визначенням дисперсії маємо:

$$Var[cX]=E\{[cX-E(cX)]^2\}.$$

Скористаємось другою властивістю математичного сподівання – постійний множник можна виносити за символ оператора. В результаті отримаємо:

$$Var[cX]=E\{[cX-cE(X)]^2\}=E\{c^2[X-E(X)]^2\}=$$
$$=c^2E\{[X-E(X)]^2\}=c^2Var[X],$$

тобто, $Var[cX]=c^2Var[X]$.

Властивість 3. *Дисперсія суми двох незалежних випадкових величин дорівнює сумі дисперсій цих величин*:

$$Var[X+Y]=Var[X]+Var[Y].$$

Доведення. За формулою для обчислення дисперсії маємо:

$$Var[X+Y]=E[X+Y]^2-[E(X+Y)]^2.$$

Розкриємо дужки і скористаємось властивостями математичного сподівання щодо суми кількох величин та добутку двох незалежних випадкових величин:

$$Var[X+Y]=E[X^2+2XY+Y^2]-[E(X)+E(Y)]^2=$$
$$=E[X^2]+2E[X]E[Y]+E[Y^2]-E^2[X]-2E[X]E[Y]-E^2[Y]=$$
$$=\{E(X^2)-[E(X)]^2\}+\{E(Y^2)-[E(Y)]^2\}=$$
$$=Var[X]+Var[Y].$$

Наслідок 1. *Дисперсія суми декількох взаємно незалежних випадкових величин дорівнює сумі дисперсій цих величин*.

Наслідок 2. *Дисперсія суми постійної і випадкової величини дорівнює дисперсії випадкової величини*.

Властивість 4. *Дисперсія різниці двох незалежних випадкових величин дорівнює сумі їх дисперсій*:

$$Var[X-Y]=Var[X]+Var[Y].$$

Доведення. За третьою властивістю

$$Var[X-Y]=Var[X]+Var[-Y].$$

Згідно з другою властивістю маємо:

$$Var[X-Y]=Var[X]+(-1)^2 Var[Y]=Var[X]+Var[Y].$$

2.13 Стандартне або середнє квадратичне відхилення

Ще одним показником варіації ряду є *стандартне* або *середнє квадратичне відхилення*, яке обчислюється за допомогою дисперсії.

Означення 2.5. *Стандартним відхиленням називають показник, що дорівнює квадратному кореню з дисперсії, взятому із знаком плюс*.

В наших позначеннях це буде σ_y.

Стандартне відхилення має ті ж властивості, що і дисперсія. Його застосування буде розглянуте нижче, але можна сказати, що його, як і дисперсію, часто використовують, наприклад, як міру ризику при аналізі фінансових та економічних процесів. Воно часто використовується в системах статистичного аналізу і контролю якості продукції, а також в багатьох інших випадках. Розраховується стандартне відхилення за формулою:

$$\sigma_y = \sqrt{\frac{\sum_{k=1}^{N}[y(k)-\mu_y]^2}{N-1}}.$$

Середнє квадратичне відхилення суми незалежних випадкових змінних

В процесі виконання статистичного аналізу досить часто необхідно розглядати кілька випадкових змінних. Нехай відомі стандартні відхилення для кількох взаємно незалежних випадкових змінних. Необхідно знайти стандартне відхилення суми значень цих змінних. Відповідь на це запитання дає наступна теорема.

Теорема 2.6. *Стандартне відхилення суми скінченого числа взаємно незалежних випадкових величин дорівнює квадратному кореню з суми квадратів стандартних відхилень цих величин*:

$$\sigma_{X_1+X_2+\ldots+X_n} = \sqrt{\sigma_{X_1}^2+\sigma_{X_2}^2+\ldots+\sigma_{X_n}^2}.$$

Доведення. Позначимо через X суму випадкових величин:

$$X = X_1 + X_2 + \ldots + X_n$$

і знайдемо дисперсію цієї суми:

$$Var[X] = Var[X_1] + Var[X_2] + \ldots + Var[X_n],$$

а звідси стандартне відхилення:

$$\sigma_X = \sqrt{Var[X_1] + Var[X_2] + \ldots + Var[X_n]},$$

або остаточно

$$\sigma_X = \sqrt{\sigma_{X_1}^2+\sigma_{X_2}^2+\ldots+\sigma_{X_n}^2}.$$

2.14 Вплив зміни значень елементів ряду на дисперсію

Нагадаємо, що додавання константи до всіх значень ряду призводить до зміни середнього на таку ж величину. Однак, додавання константи до кожного елемента ряду не впливає на значення дисперсії ряду розподілу, оскільки при обчисленні варіації середнє вираховується з кожного елемента.

Теорема 2.7. *Збільшення або зменшення кожного елемента ряду розподілу на деяку константу не впливає на значення варіації ряду розподілу і , відповідно, не впливає на значення дисперсії та стандартного відхилення*.

З іншого боку, множення кожного елемента ряду на константу впливає на варіацію цього ряду. Приклад наведено в таблиці 2.5. В результаті множення кожного елемента ряду на 5, дисперсія збільшилась в 25 разів, а стандартне відхилення – в 5 разів.

Таблиця 2.5. Вплив множення на константу елементів ряду на дисперсію

Початковий ряд розподілу			Характеристики ряду, отриманого множенням кожного елемента на 5		
Елементи ряду	Середнє	Відхилення від середнього	Елементи ряду	Середнє	Відхилення від середнього
1	2	3	1	2	3
4	8	-4	20	40	- 20
5	8	-3	25	40	-15
5	8	-3	25	40	- 15
7	8	-1	35	40	- 5
8	8	0	40	40	0
10	8	2	50	40	10
17	8	9	85	40	45
Дисперсія		17,143	Дисперсія		428,57
Стандартне відхилення		4,14	Стандартне відхилення		20,70

Теорема 2.8. *При множенні кожного елемента ряду розподілу на деяку константу, стандартне відхилення множиться на абсолютну величину цієї константи, а дисперсія – на квадрат цієї константи*. Аналогічні наслідки мають місце при діленні значень ряду на константу.

2.15 Застосування дисперсії

Дисперсія і стандартне відхилення знаходять дуже широке застосування в прикладних і теоретичних задачах. Наведемо декілька прикладів їх практичного використання.

1. Дисперсія є важливою *мірою неоднорідності* даних. Так, дисперсія оцінок за тест, який щоденно виконується студентом на протязі 10 днів, свідчить про різну ступінь його зосередженості в процесі відповіді на запитання. Якщо аналізуються доходи жителів міста, то висока дисперсія буде свідчити про велику різницю в рівнях доходів.

2. Дисперсія часто використовується як *міра ризику* у фінансах. Наприклад, високе значення дисперсії доходів від цінних паперів свідчить про те, що існує високий ризик не отримати ці доходи і навіть втратити кошти, витрачені на придбання цінних паперів.

3. Дисперсію використовують також як *міру інформативності* сигналів та рядів даних. Так, нульове значення дисперсії свідчить про відсутність інформації (дані представляють собою константу або нулі). Збільшення дисперсії говорить про те, що дані містять інформацію про зміни, що відбуваються в процесі чи об'єкті.

4. Дисперсію використовують також для опису поведінки нестаціонарних процесів. Існує клас процесів, нестаціонарних стосовно дисперсії (значення дисперсії є функцією часу), які отримали назву *гетероскедастичні* процеси, тобто, це процеси для яких $\mathrm{var}[y(k)] \neq const$. Використовуючи *дисперсію як змінну, що характеризує поведінку процесу*, можна побудувати математичну модель, яка описує динаміку дисперсії. Така модель дозволяє прогнозувати значення дисперсії на задане число кроків і приймати рішення на основі цього прогнозу. Наприклад, це може бути рішення щодо купівлі цінних паперів. Для прогнозування значення дисперсії при виконанні аналізу часових рядів користуються *умовною дисперсією*:

$$\sigma^2(k) = \frac{1}{k-1}\sum_{i=1}^{k}[y(k) - \mu_y]^2, \quad k = 2,3,....$$

5. Дисперсію також використовують в *системах статистичного аналізу якості* продукції. Системи аналізу і контролю якості – це невід'ємна частина сучасного виробництва, оскільки вона забезпечує неперервний контроль та підвищення якості продукції. При цьому однією з основних мір якості є стандартне відхилення розмірів деталей від нормативів.

6. Дисперсія може служити за *міру подібності процесів*, які описуються вибірками даних однакового змісту. Наприклад, якщо необхідно порівняти коефіцієнти розумового розвитку для різних груп студентів чи населення, то однією з мір може бути дисперсія стандартного відхилення.

2.16 Однаково розподілені взаємно незалежні випадкові величини

Відомо, що за допомогою закону розподілу можна знайти числові характеристики випадкової величини. Розглянемо випадок, коли аналізуються кілька випадкових величин, що мають однакові розподіли з однаковими числовими характеристиками.

Нехай $X_1, X_2, ..., X_n$ – взаємно незалежні випадкові змінні з однаковими законами розподілу і однаковими характеристиками (математичне сподівання, дисперсія та інші можливі параметри). З практичної точки зору, як правило, є цікавим дослідження середнього арифметичного цих величин.

Позначимо середнє арифметичне вказаної множини змінних через $\overline{X}$:

$$\overline{X} = (X_1 + X_2 + ... + X_n)/n.$$

Три наступні положення встановлюють зв'язок між числовими характеристиками середнього арифметичного $\overline{X}$ та відповідними характеристиками кожної з випадкових величин.

1. *Математичне сподівання середнього арифметичного однаково розподілених (тобто, з однаковими статистичними характеристиками) взаємно незалежних випадкових величин дорівнює математичному сподіванню кожної з величин, тобто,* $\bar{x}_1 = \bar{x}_2 = \ldots = \bar{x}_n = \bar{x}_0$, де

$$\bar{x}_0 = E[\overline{X}].$$

Доведення. Винесемо постійний множник за знак математичного сподівання:

$$E[\overline{X}] = E\left[\frac{X_1 + X_2 + \ldots + X_n}{n}\right] = \frac{E[X_1] + E[X_2] + \ldots + E[X_n]}{n}.$$

Оскільки математичні сподівання всіх величин однакові і дорівнюють $\bar{x}_0$, то

$$E[\overline{X}] = \frac{n\bar{x}_0}{n} = \bar{x}_0.$$

2. *Дисперсія середнього арифметичного* n *однаково розподілених взаємно незалежних випадкових величин є в* n *разів меншою дисперсії кожної з величин, тобто,* $\sigma^2_{X_1} = \sigma^2_{X_2} = \ldots = \sigma^2_{X_n} = \sigma^2_0$:

$$Var[\overline{X}] = \sigma^2_{\overline{X}} = \frac{\sigma^2_0}{n}.$$

Доведення. Винесемо постійний множник в квадраті за знак дисперсії:

$$Var[\overline{X}] = Var\left[\frac{X_1 + X_2 + \ldots + X_n}{n}\right] = \frac{Var[X_1] + Var[X_2] + \ldots + Var[X_n]}{n^2}$$

або

$$Var[\overline{X}] = \frac{n\sigma^2_0}{n^2} = \frac{\sigma^2_0}{n}.$$

3. *Стандартне відхилення середнього арифметичного* n *однаково розподілених взаємно незалежних випадкових величин є в* $\sqrt{n}$ *разів меншим стандартного відхилення* σ_0 *кожної з величин*:

$$\sigma[\overline{X}] = \frac{\sigma_0}{\sqrt{n}}.$$

Доведення. Оскільки $Var[\overline{X}] = \sigma^2_0 / n$, то середнє квадратичне відхилення для $\overline{X}$ дорівнює:

$$\sigma[\overline{X}] = \sqrt{Var[\overline{X}] = \sqrt{\sigma^2_0 / n}} = \sigma_0 / n.$$

Оскільки дисперсія і середнє квадратичне представляють собою міри розсіювання випадкової величини, то із наведених вище результатів можна зробити висновок, що середнє арифметичне суми кількох взаємно незалежних випадкових величин має значно менше розсіювання, ніж кожна величина окремо.

Приклад 2.7. При дослідженні фізичних та хімічних процесів для визначення оцінок змінних роблять декілька вимірів, а потім знаходять середнє арифметичне отриманих чисел, яке приймають за наближене значення (оцінку) вимірюваної величини. За припущення, що виміри виконуються в одних і тих же умовах, необхідно довести наступне:

А) середнє арифметичне дає надійніший результат, ніж окремі виміри;

Б) збільшення числа вимірів призводить до підвищення надійності результату.

Р о з в ' я з о к .

А) Відомо, що окремо взяті виміри однієї і тієї ж змінної дають неоднакові значення. Результат кожного виміру залежить від багатьох випадкових впливів (зміна температури, вплив випадкових шумових сигналів, вплив сторонніх магнітних полів і т.ін.). Всі випадкові впливи неможливо врахувати наперед.

Тому, ми розглядаємо результати n окремих вимірів як випадкові величини $x_1, x_2,\ldots,x_n$. Оскільки виміри виконуються за допомогою одного і того приладу і за однією методикою, то вважаємо, що вони мають однаковий розподіл ймовірностей. Крім того, вони взаємно незалежні, тобто, результат кожного окремого виміру не залежить від інших вимірів.

Вище було показано, що середнє арифметичне випадкових величин має менше розсіювання ніж, кожна величина взята окремо. Тобто, середнє арифметичне виявляється ближчим до істинного значення сигналу, ніж результат окремого виміру. З цього випливає, що середнє арифметичне кількох вимірів є надійнішим результатом, ніж окремо взятий вимір.

Б) Вище також було встановлено, що із зростанням числа окремих випадкових величин, розсіювання середнього арифметичного зменшується. Це означає, що із збільшенням числа вимірів їх середнє арифметичне буде все менше відрізнятись від істинного значення, що підвищує надійність результату.

Наприклад, якщо стандартне відхилення одного виміру становить 6 одиниць, а всього виконано 36 вимірів, то стандартне відхилення середнього арифметичного цих вимірів дорівнює тільки 1, дійсно:

$$\sigma(\overline{X}) = \sigma_0 / \sqrt{n} = \frac{6}{\sqrt{36}} = 1.$$

Таким чином, середнє арифметичне кількох вимірів є набагато ближчим до істинного значення, ніж результат окремо взятого виміру.

2.17 Початкові і центральні моменти

Розглянемо дискретну випадкову величину X, задану таким табличним законом розподілу:

X	1	2	5	100
P	0,6	0,2	0,19	0,01

Знайдемо для неї математичне сподівання:

$$E[X] = 1 \cdot 0{,}6 + 2 \cdot 0{,}2 + 5 \cdot 0{,}19 + 100 \cdot 0{,}01 = 2{,}95.$$

Запишемо тепер закон розподілу для X^2:

X^2	1	4	25	10000
P	0,6	0,2	0,19	0,01

і знайдемо математичне сподівання для X^2:

$$E[X^2] = 1 \cdot 0{,}6 + 4 \cdot 0{,}2 + 25 \cdot 0{,}19 + 10000 \cdot 0{,}01 = 106{,}15.$$

Видно, що математичне сподівання $E[X^2] >> E[X]$. Це пояснюється тим, що після піднесення до квадрату значення 100 призвело до появи значення 10000, хоча ймовірність цього значення досить незначна (0,01).

Таким чином, перехід від $E[X]$ до $E[X^2]$ дозволив краще врахувати вплив на математичне сподівання великого можливого значення, яке має незначну ймовірність. Очевидно, що якби змінна X мала декілька великих малоймовірних значень, то перехід до X^2, а тим більше до X^3, X^4 і так далі, дозволив би ще більше "підсилити" роль великих, але малоймовірних значень. Саме з цієї точки зору корисно обчислювати математичне сподівання цілої додатної степені (дискретної і неперервної) випадкової величини.

Початковим моментом порядку p випадкової величини X називають математичне сподівання величини X^p:

$$m_p = E[X^p].$$

Зокрема,

$$m_1 = E[X], \quad m_2 = E[X^2].$$

Користуючись визначенням моментів, запишемо вираз для обчислення дисперсії у вигляді:

$$Var[X] = m_2 - m_1^2.$$

Крім моментів випадкової змінної X доцільно розраховувати моменти відхилення $X - E[X]$, тобто розглядати статистичні характеристики відхилень від середнього.

Центральним моментом порядку p випадкової змінної X називають математичне сподівання величини $(X - E[X])^p$:

$$\mu_p = E[(X - E[X])^p].$$

Зокрема,

$$\mu_1 = E(X - E[X]) = 0,$$

$$\mu_2 = E[(X - E[X])^2] = Var[X].$$

Можна легко отримати співвідношення, які зв'язують початкові і центральні моменти. Наприклад, порівнюючи два останні вирази, отримаємо:

$$\mu_2 = m_2 - m_1^2.$$

Якщо скористатись визначенням центрального моменту і властивостями математичного сподівання, то можна отримати наступні формули:

$$\mu_3 = m_3 - 3m_2m_1 + 2m_1^3,$$
$$\mu_4 = m_4 - 4m_3m_1 + 6m_2m_1^2 - 3m_1^4.$$

Зазначимо, що моменти вищих порядків в теорії і практиці статистичного аналізу застосовують досить рідко.

Примітка. Моменти, розглянуті у даному параграфі, називають теоретичними. Моменти, які обчислюють на основі експериментальних спостережень, називають вибірковими або емпіричними.

2.18 Поняття групових і загальних статистичних характеристик

Групова і загальна середні

Нехай всі значення деякої кількісної змінної X (немає значення – генеральної сукупності чи вибіркової) розділені на кілька груп (саме такі ситуації часто зустрічаються на практиці). Розглядаючи кожну групу як самостійну сукупність, можна знайти її середнє арифметичне.

Груповою середньою *називають середнє арифметичне значень змінної, що належать одній групі*.

Тепер введемо спеціальний термін для середнього всієї сукупності даних.

Загальним середнім $\bar{x}$ *називають середнє арифметичне значень змінної, які належать всій сукупності*.

Якщо відомі групові середні і об'єми груп, то можна знайти загальне середнє: *загальне середнє дорівнює середньому арифметичному групових середніх, зваженому за об'ємами груп*.

Приклад 2.8. Знайти загальне середнє сукупності, яка складається з двох наступних груп:

Група	Перша		Друга	
Значення змінної	1	6	1	5
Частота	10	15	20	30
Об'єм	10+15=25		20+30=50	

Р о з в ' я з о к . Обчислимо групові середні:

$$\bar{x}_1 = (10 \cdot 1 + 15 \cdot 6)/25 = 4,$$

$$\bar{x}_2 = (20 \cdot 1 + 30 \cdot 5)/50 = 3{,}4.$$

Загальне середнє на основі групових середніх:

$$\bar{x} = (25 \cdot 4 + 50 \cdot 3{,}4)/(25 + 50) = 3{,}6.$$

П р и м і т к а. Для спрощення розрахунку загального середнього сукупності великого об'єму доцільно розбити її на кілька груп, знайти групові середні і за ними – загальне середнє.

Відхилення від загального середнього та його властивості

Розглянемо сукупність значень деякої кількісної змінної X (немає значення – генеральної сукупності чи вибіркової) об'єму N.

Значення змінної	x_1	x_2	…	x_l
Частоти	n_1	n_2	…	n_l

При цьому $\sum_{i=1}^{l} n_i = N$. Знайдемо загальне середнє:

$$\bar{x} = \frac{\sum_{i=1}^{l} n_i x_i}{N},$$

а звідси маємо:

$$\sum_{i=1}^{l} n_i x_i = N \bar{x}. \qquad (2.9.1)$$

Зазначимо, що оскільки $x -$ постійна величина, то

$$\sum_{i=1}^{l} n_i x = x \sum_{i=1}^{l} n_i = x N. \qquad (2.9.2)$$

Відхиленням називають різницю $x_i - \bar{x}$ між значенням змінної та загальним середнім.

Т е о р е м а 2.9. *Сума добутків відхилень на відповідні частоти дорівнює нулю*:

$$\sum_{i=1}^{l} n_i (x_i - x) = 0.$$

Д о в е д е н н я. Враховуючи (2.9.1) і (2.9.2), отримаємо:

$$\sum_{i=1}^{l} n_i (x_i - \bar{x}) = \sum_{i=1}^{l} n_i x_i - \sum_{i=1}^{l} n_i \bar{x} = \bar{x} N - \bar{x} N = 0.$$

Н а с л і д о к. *Середнє значення відхилення дорівнює нулю*. Дійсно,

$$\frac{\sum_{i=1}^{l} n_i (x_i - x)}{\sum_{i=1}^{l} n_i} = \frac{0}{N} = 0.$$

2.19 Контрольні запитання і вправи

1. Дайте означення моди і медіани? Наведіть приклади на вибірці даних.
2. Які дві властивості має середнє арифметичне?
3. Як впливають зміни (збільшення, зменшення, множення і ділення) значень деякого ряду розподілу на середнє?
4. Запишіть формулу для обчислення вибіркового середнього? Наведіть приклади використання середнього?
5. Виведіть рекурсивну формулу для обчислення поточного середнього. Яку перевагу і недолік має отриманий вираз?
6. Чому дорівнює математичне сподівання добутку двох незалежних випадкових змінних? Доведіть результат.
7. Чому дорівнює математичне сподівання суми двох випадкових змінних? Доведіть результат.
8. Запишіть формули для обчислення середнього квадратичного і середнього кубічного? У яких випадках їх доцільно застосовувати?
9. Наведіть вираз для обчислення середнього геометричного і приклад його практичного використання?
10. У якому випадку необхідно використовувати середнє гармонічне? Як воно обчислюється?
11. Які є міри варіації елементів вибірки?
12. Запишіть формулу для обчислення вибіркової дисперсії?
13. Яким чином забезпечується незміщеність оцінок параметрів, які обчислюються за статистичними даними?
14. Запишіть і поясніть вираз для обчислення стандартного відхилення середніх значень випадкових вибірок однакової потужності N.
15. Який факт називають незалежним? Поясніть на прикладі.
16. Поясніть термін «кількість ступенів свободи»?
17. Які властивості має дисперсія? Доведіть, що
$$\mathrm{var}[X + Y] = \mathrm{var}[X] + \mathrm{var}[Y].$$
18. Дайте означення стандартного відхилення? Які воно має властивості?
19. Як впливають зміни значень елементів вибірки на дисперсію?
20. Які процеси називають гетероскедастичними? Наведіть приклад гетероскедастичного процесу.
21. Чи можна використовувати дисперсію як формальну міру інформативності? Поясніть на прикладі.
22. Яке застосування має дисперсія у фінансах?
23. Яким чином зв'язана дисперсія середнього арифметичного n однаково розподілених взаємно незалежних випадкових змінних з дисперсією кожної змінної?
24. Дайте означення початкового моменту порядку p випадкової змінної X?
25. Дайте означення центрального моменту порядку p випадкової змінної X?

Розділ 3

ОПИС ПОЛОЖЕННЯ ОКРЕМОГО СПОСТЕРЕЖЕННЯ В РЯДУ РОЗПОДІЛУ, ГРУПУВАННЯ ДАНИХ

3.1 Процентильні ранги і процентилі

Необхідність описати положення окремого спостереження в ряду розподілу зустрічається досить часто. Наприклад, необхідно встановити, яке місце за успішністю посідає один із студентів в своїй групі. В даному випадку, елементами розподілу є числові значення оцінок учнів конкретної групи. Можна сказати, наприклад, що студент *A* отримав другу за величиною оцінку в групі. Однак, така інформація не дає можливості визначити, наскільки успішно навчається студент *A*, тому що в групі могло бути і 5 студентів, і 35. Для того, щоб встановити, яке місце посідає в групі окремий студент, необхідно знати число студентів в цій групі.

Місцеположення деякого спостереження в розподілі визначається шляхом присвоєння йому значення, яке називають *процентильним рангом*.

Наприклад, нехай у групі всього 20 студентів і студент *A* отримав вищу оцінку ніж п’ятнадцятеро його товаришів по групі, а інші четверо студентів отримали оцінки, вищі ніж *A*. Таким чином, 75% оцінок (15 з 20) нижчі ніж в *A*. Сама оцінка студента *A* складає 5% від загального числа оцінок, а 20% оцінок є вищими ніж у *A*. Ці процентні співвідношення показані на рис. 3.1.

75%	5%	20%

Рис. 3.1. Процентні співвідношення оцінок в групі

Місце студента *A* в групі вкажемо наступним чином: додамо до 75% (оцінки, які є меншими ніж у студента *A*) ще 2,5%, тобто, половину від процентного значення оцінки *A*. Для цього можна умовно провести вертикальну лінію через центр області, що відноситься до оцінки *A*. В результаті знайдемо, що процентильний ранг студента *A* дорівнює 77,5%:

$$Пр = 75 + 2{,}5 = 77{,}5\%.$$

Означення 3.1: *Процентильний ранг деякого спостереження ряду розподілу дорівнює сумі процентів, що відносяться до спостережень, які стоять в розподілі нижче нього, і половини процентів, які відносяться безпосередньо до нього*.

Нехай в іншій групі, яка налічує 50 студентів, студент *Б* отримав оцінку, яка є вищою, ніж у 17 його товаришів, тобто, 34% оцінок в розподілі є нижчими від оцінки *Б*. Оскільки на одного студента припадає 2%, то процентильний ранг *Б* складає:

$$Пр = 34 + 1 = 35\%.$$

Можна сказати, що студент *А* навчається в своїй групі краще ніж студент *Б* у своїй, оскільки його процентильний ранг 77,5% набагато перевищує 35%.

Процентильні ранги дають можливість порівнювати успішність студентів, не зважаючи на те, що суворість вчителів та їх методи оцінювання можуть бути різними. Таким чином, *процентильні ранги дають можливість співставляти між собою елементи різних розподілів*.

Показник студента *А* є вищим тому, що він випередив більше студентів-одногрупників, ніж студент *Б*. При цьому припускається, що рівень конкуренції в обох групах однаковий, хоча на практиці це припущення часто не виконується. Зазначимо, що при порівнянні студентів *А* і *Б* ми не користувались самими числовими значеннями членів рядів розподілу, тобто, дійсними значеннями оцінок, отриманих студентами.

Процентильний ранг елемента розподілу часто виявляється змістовнішим, ніж власне значення цього елемента.

Наприклад, нехай випускник КПІ претендує на місце в аналітичному відділі банку. В результаті проходження спеціального тестування для виявлення здібностей до роботи в аналітичному відділі встановлено, що його коефіцієнт розумового розвитку склав 115 балів (КРР = *IQ* = *Intelligence Quotient*). Це велике значення, але воно ще не дає підстав для прийняття позитивного рішення щодо прийняття випускника на роботу; для цього потрібна додаткова інформація. Якщо буде надана додаткова інформація, що випускник має 94-й процентиль по успішності серед студентів, які вивчали банківську справу, то, судячи з результату тестування, він має перспективу влаштуватись на роботу в банк.

Слово *процентиль відноситься безпосередньо до елемента розподілу*, тобто, до значення, яке знаходиться між двома елементами. Елемент розподілу з процентильним рангом 35 називають тридцять п’ятим процентилем, а елемент з процентильним рангом 77 – сімдесят сьомим процентилем.

3.2 Нормовані відхилення (z – оцінки)

Процентильні ранги не дають відповіді на деякі запитання. Наприклад, припустимо, що студенти *А* і *Б* є кращими в своїх групах і мають однакові процентильні ранги. При цьому один з них може значно переважати своїх конкурентів, а перевага іншого в своєму класі є незначною. Оскільки процентильні ранги обох студентів однакові, то виявити між ними різницю

неможливо. Для того, щоб її виявити, необхідно ввести новий спосіб визначення положення елемента в ряду розподілу, який буде показувати, наскільки далеко знаходиться даний елемент від точки рівноваги (середнього) і по який бік від нього.

Розглянемо розподіл з шести елементів (оцінки студентів): 4, 6, 7, 10, 15, 18. Середнє дорівнює:

$$\mu_x = (4+6+7+10+15+18)/6 = 10.$$

Знайдемо відхилення для елементів ряду і стандартне відхилення (табл. 3.1).

Таблиця 3.1. Відхилення оцінок від середнього для 1-ї групи

№ п/п	Оцінка	Середнє	Відхилення
1	4	10	- 6
2	6	10	- 4
3	7	10	- 3
4	10	10	0
5	15	10	5
6	18	10	8
	Стандартне відхилення = 5		

Відхилення оцінки кожного студента вказує на місце його оцінки в розподілі для його групи, але відхиленням не можна скористатись для порівняння студентів двох різних груп. Причина полягає в тому, що в іншій групі оцінки можуть ставитись по іншому принципу. Таким чином, відхилення від середнього на кілька балів може мати різний смисл в різних групах. Розглянемо для прикладу оцінки, отримані в іншій групі (табл. 3.2).

В другій групі оцінки розсіяні більше ніж в першій, про що свідчать відхилення оцінок.

Таблиця 3.2. Відхилення оцінок від середнього для 2-ї групи

№ п/п	Оцінка	Середнє	Відхилення
1	20	50	- 30
2	30	50	- 20
3	35	50	- 15
4	50	50	0
5	75	50	25
6	90	50	40
	Стандартне відхилення = 25		

З порівняння таблиць 3.1 і 3.2 видно, що оцінки в другій групі є набагато вищими. Так, краща оцінка в першій групі, яка дорівнює 18, в другій групі буде

меншою найнижчої. Відповідні відхилення оцінок від середнього також є набагато більшими в другій групі.

Для порівняння успішності студентів цих груп *не можна використовувати відхилення тому, що вони мають різну „вагу" для різних груп*. Однак, для цієї мети можна скористатись двома параметрами – *відхиленням і стандартним відхиленням*.

Означення 3.2. *Відношення відхилення від середнього до стандартного відхилення називають* $z-$ *оцінкою*. Вона розраховується за формулою:

$$z_k = \frac{x_k - \mu_x}{\sigma_x}, \quad k = 1,\ldots,N .$$

Так, $z-$оцінка 6-го студента першої групи: $z_{16} = 8/5 = 1{,}6$, а $z-$оцінка 5-го студента другої групи: $z_{25} = 25/25 = 1$. Оскільки $z-$оцінка 6-го студента першої групи є вищою, ніж 5-го студента другої групи, то можна сказати, що перший з них займає вище місце в своїй групі, ніж другий в своїй групі.

Таким чином, $z-$*оцінка показує на скільки одиниць стандартних відхилень конкретний елемент є більшим або меншим середнього його ряду розподілу*.

Наприклад, якщо $z-$оцінка елемента дорівнює 1,5, то він перевищує середнє на 1,5 одиниці стандартного відхилення. Елемент, який має $z-$оцінку – 0,5, є на 5/10 одиниці стандартного відхилення меншим середнього його розподілу. Таким чином, $z-$оцінки можна використовувати для порівняння елементів різних розподілів. На відміну від значень відхилень $z-$оцінки не залежать від „цінності" балів, що виставляються студентам. Причина цього полягає у тому, що $z-$оцінка характеризує відносне місцеположення елемента із врахуванням ступеня коливання всього ряду розподілу.

3.3 Середні і $z-$оцінки

У випадках, коли в різних розподілах змінних використовують різні одиниці виміру, для знаходження їх середніх виявляються коректними $z-$оцінки. Нехай група студентів отримує два завдання: *А* і *Б*. Максимальне число балів за виконання завдання *А* складає 100, а за завдання *Б* – 10. Нехай студент *С1* отримує за завдання *А* 76 балів, а за завдання *Б* – 8. *С2* отримує 83 бали за завдання *А* і отримує 1 бал за завдання *Б*, тобто, не виконує його. *С1* отримує вищу оцінку, тому що він виконав два завдання, а *С2* – тільки одне.

Формально можна підійти до проблеми оцінювання виконаних завдань студентами наступним чином. Середня оцінка *С2* за два завдання складає: $(83+1)/2 = 42$, а середня оцінка *С1* складає: $(76+8)/2 = 42$, тобто формальні середні арифметичні однакові. Однак, чи буде ця оцінка правильною?

Інтуїтивне відчуття того, що *С1* заслуговує вищої оцінки не підтверджується середнім арифметичним отриманих балів. Фактично, усереднювати в даному випадку не можна тому, що "вага" або "ціна" одного бала в одному тесті була вищою, ніж в іншому. *С2* випередив *С1* в тесті *А* на сім балів, але *С1* набрав на сім балів більше в тесті *Б*, в якому кожний бал цінiться набагато вище (за рахунок того, що максимальна оцінка складає 10).

Якщо ж перетворити оцінки кожного студента в $z-$оцінки, то можна отримати рівноцінні одиниці виміру, незважаючи на те, що початкові одиниці виміру оцінок можуть бути різними. Такий підхід можливий завдяки тому, що $z-$оцінки мають однаковий смисл для будь-якого ряду розподілу; фактично, це величини, які не мають одиниць виміру.

В загальному випадку, коли елементи різних розподілів оцінюються за допомогою різних одиниць виміру, застосування звичайних середніх втрачає смисл.

3.4 Порівняння $z-$оцінок і процентилів

Процентильні ранги і $z-$оцінки безпосередньо не зв'язані між собою і не можна отримати значення іншого, якщо відомий один з них. Теоретично, елемент ряду розподілу, який має визначений процентильний ранг, може мати майже будь-яке значення $z-$оцінки. Розглянемо приклад, наведений в таблиці 3.3. Розподіл складається з елементів: 2, 3, 6, 8, 11 із середнім 6.

Таблиця 3.3. Порівняння $z-$оцінок з процентильними рангами

Елементи	$z-$оцінки	Процентильні ранги
2	– 1,21	10
3	– 0,91	30
6	0,00	50
8	0,61	70
11	1,52	90

Два елементи цього ряду (2 і 3) розташовані зліва від середнього, а тому вони мають від'ємні $z-$оцінки. Два інших (8 і 11) розташовані справа від середнього, а тому мають додатні $z-$оцінки. Один елемент (6) дорівнює середньому, а тому його $z-$оцінка дорівнює нулю.

Елемент 8 має процентильний ранг 70, а $z-$оцінка цього елемента дорівнює 0,61. Додатна $z-$оцінка даного елемента свідчить про те, що його значення перевищує середнє. Однак, в іншому випадку може мати місце ситуація, коли елемент з процентильним рангом 70 буде знаходитись нижче середнього і мати від'ємну $z-$оцінку (табл. 3.4, середнє дорівнює 4).

Таблиця 3.4. Порівняння $z-$оцінок з процентильними рангами

Елементи	$z-$оцінки	Процентильні ранги
0	– 0,78	10
1	– 0,59	30
2	– 0,39	50
3	– 0,20	70
14	1,96	90

Середнє = 4.

Елемент 3 має процентильний ранг 70, оскільки він перевищує 70% елементів розподілу, але його $z-$оцінка від'ємна, тому що він є меншим середнього. Існують ще більше "незбалансовані" ряди, в яких елемент може мати процентильний ранг 90, але все-таки бути меншим середнього.

Висновок: *Не існує строгих правил для визначення залежності між $z-$оцінкою і процентильним рангом. Кожний з цих показників є окремою характеристикою ряду. Процентильний ранг вказує на те, скільки елементів ряду розташовано нижче даного спостереження, але не дає інформації щодо його зв'язку з середнім. В свою чергу, $z-$оцінка вказує на те, як розташований даний елемент по відношенню до середнього значення ряду, але не містить інформації щодо проценту елементів ряду, розташованих нижче або вище нього*.

Примітка. В більшості рядів розподілу, які мають місце на практиці, елемент з процентильним рангом 70 – 80 буде мати додатну $z-$оцінку. Аналогічно, якщо для елемента x_i $z-$оцінка $z_i \geq 1$, то він, як правило, буде більшим щонайменше половини інших елементів.

3.5 Середнє і стандартне відхилення ряду, утвореного із $z-$оцінок

Теорема 3.1. *Середнє ряду розподілу $z-$оцінок завжди дорівнює нулю, а стандартне відхилення – одиниці, тобто* $\mu_z = 0,\ \sigma_z = 1.$

Доведення: $$\mu_z = E\left[\frac{x_k - \mu_x}{\sigma_x}\right] = \frac{E[x_k] - \mu_x}{\sigma_x} = \frac{\mu_x - \mu_x}{\sigma_x} = 0;$$

$$Var[z_k] = E\left[\left(\frac{x_k - \mu_x}{\sigma_x}\right)^2\right] = \frac{E[(x_k - \mu_x)^2]}{\sigma_x^2} = \frac{\sigma_x^2}{\sigma_x^2} = 1.$$

Приклад 3.1. Розглянемо ряд розподілу $X = [4\ \ 7\ \ 8\ \ 10\ \ 11\ \ 20]$ з середнім $\mu_x = 10$ і стандартним відхиленням $\sigma_x = 5$.

Розрахункові значення $z-$оцінок та їхні характеристики наведені в таблиці 3.5.

Таблиця 3.5. $z-$оцінки та їх характеристики

Елементи	Відхилення	$z-$оцінки
4	– 6,0	– 1,2
7	– 3,0	– 0,6
8	– 2,0	– 0,4
10	0	0,0
11	1,0	0,2
20	10,0	2,0
$\mu_x = 10,\ \sigma_x = 5$		$\mu_z = 0,\ \sigma_z = 1$

Зазначимо, що *отриманий результат не залежить від випадкових характеристик ряду розподілу*. Так, середнє відхилень завжди дорівнює нулю, а їх стандартне відхилення – стандартному відхиленню вихідного ряду розподілу.

Очевидно, що середнє ряду, отриманого в результаті ділення відхилень на стандартне відхилення також буде нульовим. Оскільки всі відхилення від середнього ряду розподілу діляться на їх власне стандартне відхилення, то і стандартне відхилення середнього ряду буде поділеним само на себе. В результаті стандартне відхилення $z-$оцінок буде дорівнювати 1.

3.6 Стандартні або $T-$оцінки

Оскільки $z-$оцінки містять десяткові знаки і можуть бути додатними або від'ємними, то це не зовсім зручно при аналізі цих оцінок. Часто буває зручніше замінити $z-$оцінки додатними числами. Такі оцінки називають стандартними або $T-$оцінками, а розраховують їх за виразом:

$$T_k = 50 + 10\, z_k = 50 + 10 \cdot \frac{x_k - \mu_x}{\sigma_x}.$$

Тобто, $z-$оцінки множать на 10 і заокруглюють до цілого, а додавання числа 50 перетворює всі значення $z-$оцінок в додатні числа. Теоретично $T-$оцінка може бути від'ємною, але таких $z-$оцінок практично немає. Так, $T_k = 0$ при $z_k = -5{,}0$, а від'ємними стандартні оцінки будуть при $z_k < -5{,}0$.

$T-$оцінки можна застосовувати для розв'язування тих же задач, що і $z-$оцінки. Наприклад, за їх допомогою можна порівняти успішність студентів різних груп. Очевидно, що середнє $T-$оцінок $\mu_T = 50$, а стандартне відхилення $\sigma_T = 10$.

3.7 Застосування процентильних рангів і стандартних оцінок

А) *Процентилі і процентильні ранги* використовують як *характеристики рівня освіти, для визначення різниці між індивідуумами в психології* та інших

областях, де застосовують бали та оцінки. Наприклад, приймальним комісіям університетів було б доцільно розглядати спочатку процентильні ранги абітурієнтів, а вже потім числові значення оцінок. Процентильний ранг вказує на місце, яке займає учень в своєму класі, а оцінка вказує на рівень успішності та на строгість визначення оцінок в школі.

Б) *Економісти* розраховують процентильні ранги рівнів *доходу на жителя країни* з метою порівняння з іншими країнами.

Можна порівнювати також процентильні ранги рівня *промислового виробництва* різних міст в масштабах однієї країни.

В) $z-$*оцінки зручніше трансформувати в стандартні Т-оцінки, оскільки вони містить суттєву інформацію*. Так, якщо студент отримав стандартну оцінку за тест $T=60$, то (на основі виразу для її обчислення) можна зробити висновок, що його оцінка є на одне стандартне відхилення вищою середньої оцінки класу.

3.8 Часові ряди і часові перерізи даних

Всі статистичні дані можна розділити на два великих класи: часові ряди і часові перерізи. Обидва типи даних зустрічаються дуже часто на практиці.

Часовий ряд *– це послідовність значень випадкової величини, яка отримана з деяким визначеним часовим інтервалом* (періодом дискретизації T_s).

Період дискретизації частіше всього постійний, але в окремих випадках він може бути перемінним, тобто, $T_s \neq const$. Він може бути перемінним у випадках, коли неможливо виконати збір даних з постійним періодом або у даних є пропуски. Крім того, період збору даних може бути взагалі *нерегулярним*, тобто, значення T_s носить не прогнозований характер. В таких випадках кажуть, що *дані носять нерегулярний характер* і для того щоб робити статистичний висновок на основі таких даних, необхідно застосовувати спеціальні методи.

Прикладами часових рядів є статистичні дані, які характеризують зміну в часі економічних процесів, виміри змінних, що характеризують поведінку технічної системи (швидкість, прискорення, амплітуди коливань) або технологічного процесу (температура протікання реакції, концентрація розчину, рівень рідини в ємності і т.ін.).

Часові перерізи *– це дані, що характеризують вибраний процес в деякий конкретний момент або відрізок часу* (цим відрізком може бути день, тиждень, місяць чи навіть рік).

Наприклад, може представляти цікавість дослідження успішності студентів на кінець семестру, тобто, на кінець січня або червня. В такому випадку дані щодо успішності збирають протягом кількох днів в кінці визначеного місяця і вони характеризують вибрану групу студентів саме на

кінець конкретного періоду часу. Якщо визначається рейтинг того чи іншого кандидата в президенти, то дані також збирають на протязі кількох днів. Результати опитування представляють як опитування на конкретну дату. Перепис населення виконують, наприклад, на протязі місяця і отриманий результат перепису вважається дійсним на кінець конкретного місяця і року.

3.9 Дискретні і неперервні величини

В загальному випадку, змінні можуть приймати різні за своєю природою та величиною значення. Змінна "вік" за своєю природою є неперервною, оскільки час є неперервною категорією, але ми користуємось дискретними значеннями цієї величини. Коли досліджують зміни функціонування довгострокової чи короткострокової пам'яті людини, то вибирають індивідуумів конкретного віку, наприклад, 10, 20, 30, 40, 50, 60 і 70 років. Можна вважати, що в даному випадку період дискретизації експериментальних даних буде складати 10 років ($T_s = 10$ років).

В технічних системах період дискретизації може складати дуже короткі проміжки часу, наприклад, десятки мікросекунд (*мкс*) або мілісекунд (*мс*). Так, в системах автоматичного керування двигунами внутрішнього згоряння він складає біля 20 *мс*. В соціально-економічних системах процеси мають, в основному, неперервний характер, але дані завжди збирають дискретно. Інтервал між будь-якими двома сусідніми числами включає всі значення, які знаходяться між ними. Однак, є змінні, які приймають тільки дискретні значення. Наприклад, число деталей, які виробляє робітник за одиницю часу.

Означення 3.3. *Якщо всередині деякого інтервалу змінна теоретично може приймати тільки деякі, визначені тим чи іншим способом, значення, то така змінна на даному інтервалі змінюється дискретно*.

Наприклад, число голосів, поданих за того чи іншого кандидата в президенти. Сімейний стан може визначатись чотирма якісними змінними: *одинокий*, *одружений*, *розлучений*, *вдівець*. Змінна, яка відображає відношення до військової служби: *військовозобов'язаний* і *не військовозобов'язаний*.

Означення 3.4. *Якщо всередині деякого інтервалу змінна теоретично може приймати будь-яке значення, то ця величина змінюється в даному інтервалі неперервно*.

Змінна "вік" теоретично є неперервною, оскільки може приймати будь-яке значення в інтервалі від нуля до нескінченності. Для людини ця нескінченність обмежена значенням приблизно 150 років, а для деяких дерев це може бути кілька тисяч років. Вік Сонячної Системи обчислюється мільярдами років.

На практиці ми користуємось тільки деякими обмеженими значеннями з інтервалу визначення неперервної змінної і округлюємо фактичні значення до зручних для користування. Наприклад, *вік людини* заокруглюють до *років*.

Здібності людини також являють собою неперервну змінну, але коефіцієнт розумового розвитку завжди округлюють до цілого числа.

В технічних та багатьох інших системах ми можемо покращити точність вимірювань завдяки використанню нових точних приладів.

3.10 Табулювання і графічне зображення дискретних величин

Ряд значень дискретної змінної зручно представляти у вигляді *таблиці частот*. Що означає термін "*частота*"?

О з н а ч е н н я 3.5. *Частотою будь-якого значення, яке може приймати змінна в ряду розподілу, називається число, яке показує, скільки разів дане значення зустрічається у вибірці*.

В таблиці 3.6 наведено приклад значень частот для ряду розподілу сімейного стану жінок на невеликій фабриці.

Таблиця 3.6. Ряд розподілу сімейного стану жінок

Сімейний стан	Частоти
Одинокі	15
Заміжні	25
Розлучені	20
Вдови	5

Зображення ряду розподілу можна спростити, якщо скористатись *стовпчиковою діаграмою*.

Стовпчикова діаграма – один з найбільш зручних та популярних видів графіків, які використовують для графічного представлення ряду розподілу. Зазначимо, що між окремими стовпчиками є проміжки. На рис. 3.2 наведено стовпчикову діаграму для ряду з таблиці 3.6.

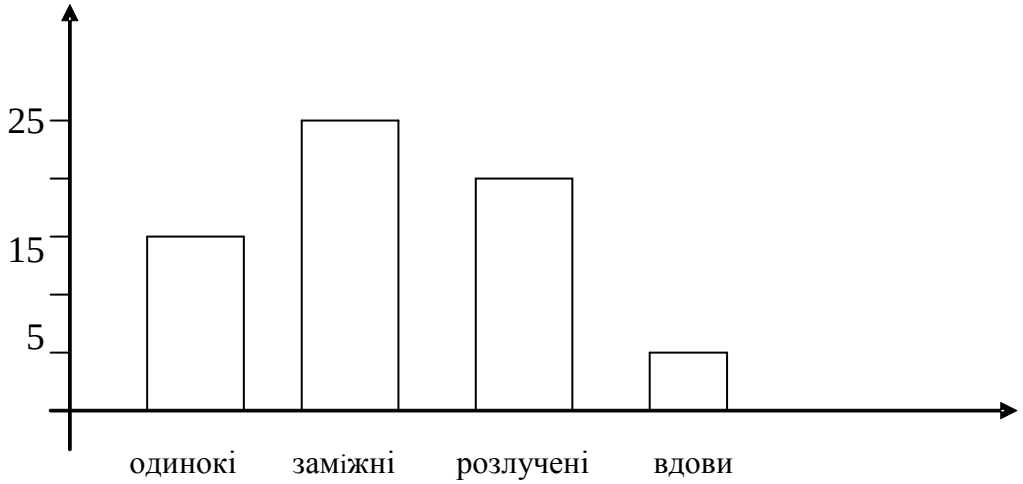


Рис. 3.2. Стовпчикова діаграма для ряду розподілу сімейного стану жінок

3.11 Округлення значень

Оскільки вимірювання значень змінних неможливо здійснити з абсолютною точністю, то їх округлюють. Це дозволяє суттєво спростити необхідні розрахунки.

Округлення означає, що змінній приписується значення найближчої, вибраної нами точки.

Так, всю область визначення змінної можна розбити на інтервали і в подальшому аналізі, значенню, яке попадає в конкретний інтервал, приписувати значення середини цього інтервалу. Таким чином, при округлюванні, змінні приймають значення з обмеженого (скінченого числа значень). Розглянемо, наприклад, ряд значень зросту людини (рис. 3.3).

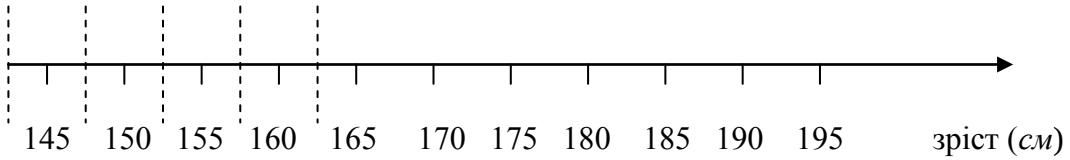


Рис. 3.3. Інтервали та їх середні для значень змінної "зріст".

Так, всім величинам, які попадають в інтервал $142{,}5 \div 147{,}5$ (верхня границя не включається в інтервал), приписується значення 145. Тим, що попадають в інтервал $147{,}5 \div 152{,}5$ приписують 150 і т. ін.

Таким чином, округлювання передбачає групування значень в інтервали і присвоєння згрупованим даним значень середини інтервалу.

Техніку роботи з округленими значеннями змінної називають *методом групування*.

3.12 Групування даних і побудова групової таблиці частот

Групування і табулювання сукупності спостережень складається з чотирьох наступних процедур.

1. Розрахувати *варіаційний розмах*, який дорівнює різниці між найбільшим і найменшим значеннями ряду спостережень. Наприклад, якщо в групі студентів максимальна оцінка дорівнює 100, а мінімальна 3, то варіаційний розмах дорівнює: $100 - 3 = 97$.

2. Визначити кількість *інтервалів* та *їх довжину*. Зручніше довжину інтервалу вибирати рівною непарному числу для того, щоб середина кожного інтервалу дорівнювала цілому значенню.

Так, в прикладі з оцінками групи студентів зручно взяти інтервали довжиною 9 балів. Всього буде 11 інтервалів, які охоплюють 99 балів. Важливо, щоб сумарна довжина всіх інтервалів дорівнювала, щонайменше, варіаційному розмаху. Таким чином, в інтервалі $3 \div 100$ знаходиться 98 цілих чисел, тобто, на

одне число більше від варіаційного розмаху.

3. Визначити *граничне значення самого верхнього інтервалу*. В нашому прикладі верхній інтервал охоплює значення $92 \div 100$. Визначаючи крайні точки інтервалу, ми включаємо їх в даний інтервал. Тобто, ширина інтервалу $92 \div 100$ складає 9 одиниць. Наступний інтервал буде охоплювати значення $83 \div 91$. Продовжуючи процедуру визначення інтервалів, встановимо, що самий нижній інтервал включає значення: $2 \div 10$ (табл. 3.7).

Таблиця 3.7. Приклад згрупованого ряду розподілу

Інтервали	Центри
92 – 100	96
83 – 91	87
74 – 82	78
65 – 73	69
56 – 64	60
47 – 55	51
38 – 46	42
29 – 37	33
20 – 28	24
11 – 19	15
2 – 10	6

4. *Табулювання частот інтервалів*, тобто, визначення числа значень, які відносяться до кожного інтервалу. Наприклад, таблиця частот для ряду розподілу оцінок, розглянутих вище, може мати вигляд, наведений в табл. 3.8.

Таблиця 3.8. Групова таблиця частот ряду розподілу

Інтервали	Центри	Частоти
92 – 100	96	60
83 – 91	87	140
74 – 82	78	160
65 – 73	69	120
56 – 64	60	140
47 – 55	51	80
38 – 46	42	119
29 – 37	33	81
20 – 28	24	50
11 – 19	15	32
2 – 10	6	18
		Всього = 1000

З табл. 3.8 видно, що 60 студентів мають оцінки в інтервалі $92 \div 100$ (включаючи оцінки 92 і 100), а оцінку в інтервалі $83 \div 91$ (включаючи 83) мають 140 студентів.

Тепер ми можемо оперувати з таким рядом, що має 60 елементів із значенням 96, 140 елементів із значенням 87 і так далі. Фактично, ми округлили значення змінної до найближчого, наперед визначеного числа.

3.13 Гістограма і полігон ряду розподілу

В цьому параграфі розглянемо метод графічного зображення неперервного ряду розподілу на основі інформації, яка міститься в груповій таблиці частот. Повернемось ще раз до таблиці 3.8. Оцінку за тест можна розглядати як неперервну змінну, значення якої можуть відрізнятись як завгодно мало. Теоретично, оцінка може приймати значення 91,6 або 92,4 і т. ін.

Фактично, оцінки, які перевищують 91,5, відображають як 92 і вище. Таким чином, нижньою границею верхнього інтервалу буде 91,5. Аналогічно 91,5 є і верхньою границею інтервалу $83 \div 91$, а нижньою границею цього інтервалу є значення 82,5. Звідси випливає, що показники знань, які були вміщені в інтервал $83 \div 91$, при більш точному вимірюванні можуть потрапити в інтервал $82{,}5 \div 91{,}5$.

По аналогії можна визначити границі всіх інтервалів, представлених в табл. 3.8. Значення цих границь наведено в табл. 3.9.

Таблиця 3.9. Групова таблиця частот

Інтервали	Границі	Центри	Частоти
92 – 100	91,5 – 100,5	96	60
83 – 91	82,5 – 91,5	87	140
74 – 82	73,5 – 82,5	78	160
65 – 73	64,5 – 73,5	69	120
56 – 64	55,5 – 64,5	60	140
47 – 55	46,5 – 55,5	51	80
38 – 46	37,5 – 46,5	42	119
29 – 37	28,5 – 37,5	33	81
20 – 28	19,5 – 28,5	24	50
11 – 19	10,5 – 19,5	15	32
2 – 10	1,5 – 10,5	6	18
		Всього = 1000	

Ряд розподілу оцінок за тест можна представити тепер в графічній формі - рис. 3.4.

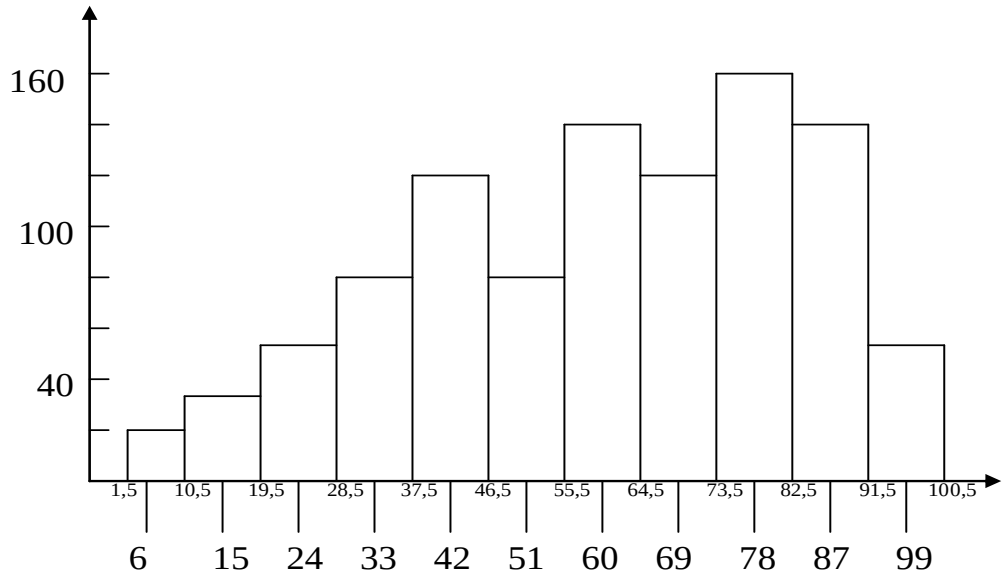


Рис. 3.4. Гістограма ряду розподілу оцінок за тест

Якщо з цього графіка *прибрати вертикальні лінії, то отримаємо гістограму*. Гістограма схожа на стовпчикову діаграму, але на стовпчиковій діаграмі прямокутники розташовані окремо один від одного, що вказує на дискретний характер розподілу. Прямокутники на гістограмі стоять впритул, що свідчить про те, що ряд розподілу є неперервним.

Ще одним способом графічного представлення неперервного ряду є *частотний полігон*.

Частоти інтервалів позначають точками, розташованими над центрами інтервалів, які з'єднують між собою прямими лініями (рис. 3.5).

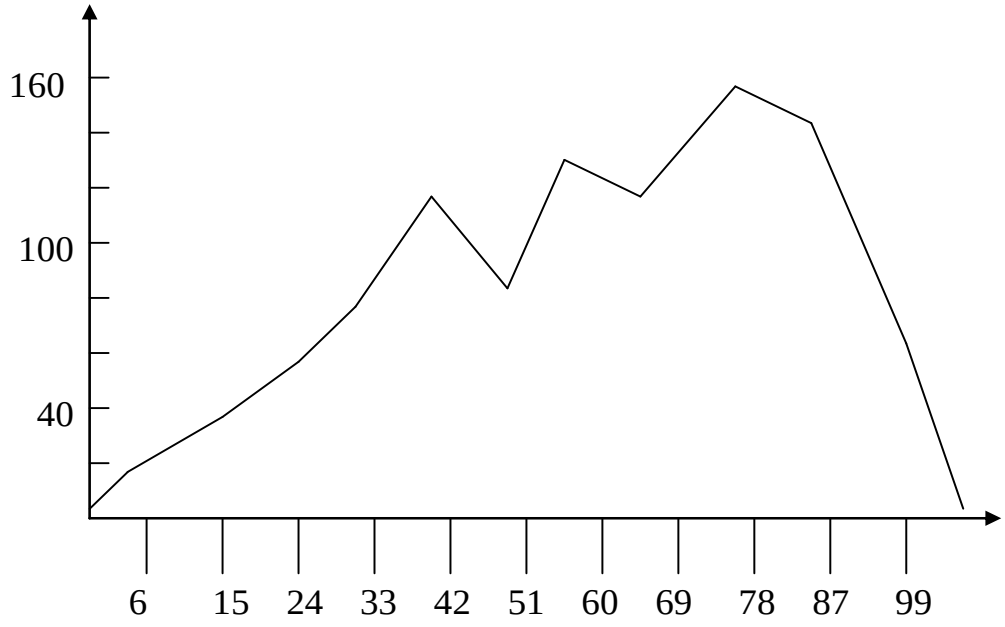


Рис. 3.5. Частотний полігон ряду розподілу оцінок за тест

Якщо при побудові гістограми зменшувати довжину інтервалів, то число прямокутників буде збільшуватись, а зміна висоти прямокутників, що стоять поряд, буде більш рівномірною. Гістограма буде нагадувати криву лінію. Подальше зменшення довжини інтервалу призводить до того, що різниця між площею гістограми і площею фігури, до якої прямує графік, стане довільно малою. Таким чином, *можна перейти до неперервного ряду розподілу*. Однак, для реалізації такої процедури необхідно мати досить велику кількість спостережень.

3.14 Застосування стовпчикових діаграм і гістограм

Стовпчикові діаграми та гістограми широко *застосовуються для* графічного представлення та візуального аналізу практично *будь-яких типів даних*. Частіше застосовують гістограми. Так, якщо на гістограмі представити зріст чоловіків у вибраній країні, то відразу можна визначити, який зріст має половина або якась інша частина чоловіків.

Однак, якщо на одному графіку зображають два або більше рядів, то зручніше користуватись полігонами, оскільки в даному випадку зображення в меншій мірі накладаються одне на одне.

Дослідження таблиць частот та гістограм дозволяє оперативно створити розумове уявлення щодо вигляду ряду розподілу і, відповідно, швидше вибрати для нього статистичне описання.

3.15 Контрольні запитання і вправи

1. Дайте означення процентильного рангу, наведіть приклад? Поясніть термін «процентиль»?
2. Які значення називають $z-$оцінками? Порівняйте середнє і $z-$оцінку деякого ряду розподілу?
3. Які статистичні характеристики (середнє і дисперсію) має ряд, утворений з $z-$оцінок? Доведіть результат.
4. Яким чином обчислюються $T-$оцінки? Чим вони зручніші у використанні?
5. Наведіть приклади застосування процентильних рангів?
6. Наведіть приклади застосування $T-$оцінок?
7. Дайте означення часового ряду і часового перерізу статистичних даних?
8. Які змінні відносять до дискретних, наведіть приклади? Чи можна віднести до дискретних змінних квартальні значення валового внутрішнього продукту?
9. Які змінні називають неперервними, наведіть приклади? Змінна *вік людини* – це дискретна чи неперервна змінна?
10. Як розраховується варіаційний розмах ряду спостережень, наведіть приклад?
11. З яких кроків складається процедура групування і табулювання сукупності спостережень?

12. Поясніть на прикладі побудову гістограми?
13. Яким чином будується частотний полігон ряду розподілу?
14. Як будується стовпчикова діаграма? Наведіть приклади застосування стовпчикових діаграм?
15. Наведіть приклади застосування гістограм?
16. Побудуйте стовпчикову діаграму для успішності студентів у Вашій групі.

РОЗДІЛ 4

СТАТИСТИЧНІ ПАРАМЕТРИ ДЛЯ ЗГРУПОВАНИХ ДАНИХ

4.1 Визначення медіани і процентилів з гістограми

Нагадаємо, що медіана – це число, яке більше або дорівнює і одночасно менше або дорівнює половині значень ряду розподілу. Таким чином, *медіана є 50-м процентилем*. *Медіана не обов'язково повинна співпадати з конкретним значенням ряду розподілу*. Про це необхідно пам'ятати при визначенні медіани для ряду розподілу згрупованих даних, коли окремі значення вихідного (початкового) ряду в деякій мірі втратили свою індивідуальність.

Розглянемо наведений вище приклад з оцінками за тест (табл. 4.1 та відповідна гістограма на рис. 4.1). Відомо, що медіана є більшою і одночасно меншою половини з 1000 оцінок. Для того, щоб її знайти, необхідно вказати на горизонтальній осі (рис. 4.1) число, яке перевищує кількість оцінок – 500.

Таблиця 4.1. Групова таблиця частот

Інтервали	Границі	Центри	Частоти
92 – 100	91,5 – 100,5	96	60
83 – 91	82,5 – 91,5	87	140
74 – 82	73,5 – 82,5	78	160
65 – 73	64,5 – 73,5	69	120
56 – 64	55,5 – 64,5	60	140
47 – 55	46,5 – 55,5	51	80
38 – 46	37,5 – 46,5	42	119
29 – 37	28,5 – 37,5	33	81
20 – 28	19,5 – 28,5	24	50
11 – 19	10,5 – 19,5	15	32
2 – 10	1,5 – 10,5	6	18
		Всього = 1000	

Знайдемо суму частот для шести і семи нижніх інтервалів: 380 – для шести і 520 – для семи. Іншими словами, 380 оцінок розташовані нижче значення 55,5 і 520 оцінок розташовані нижче значення 64,5. Медіана повинна задовольняти нерівності $55{,}5 < Me < 64{,}5$, але все-таки бути ближчою до 64,5.

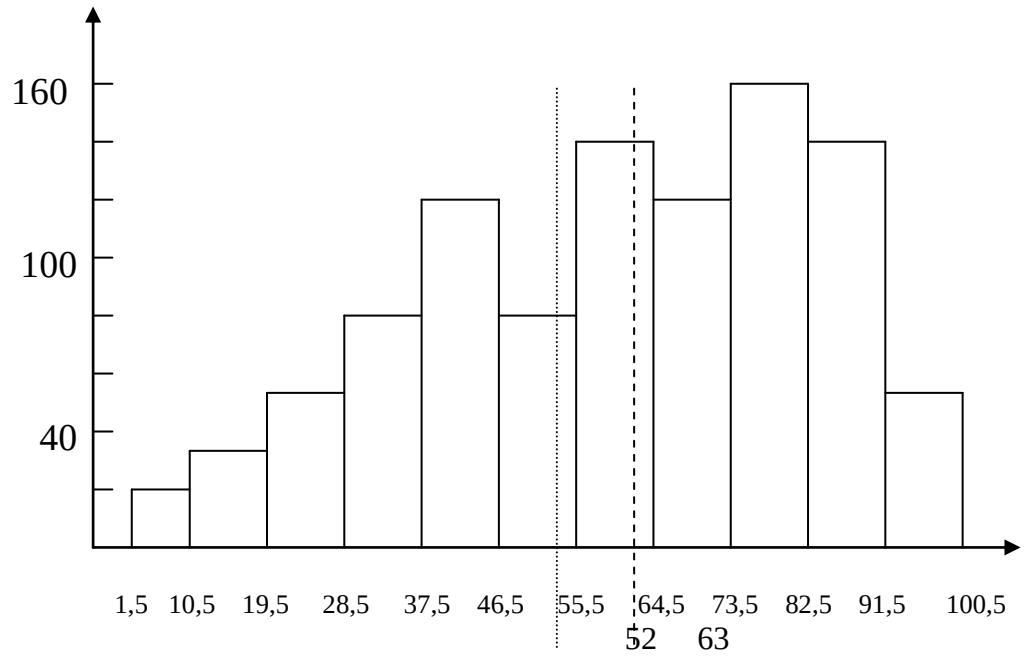


Рис. 4.1. Гістограма ряду розподілу оцінок за тест

Площа прямокутника, побудованого над інтервалом $55{,}5 \div 64{,}5$, відповідає 140 оцінкам, які були згруповані в даному інтервалі шляхом округлювання їх значень до 60. Для того, щоб знайти медіану, необхідно визначити на горизонтальній осі точку, через яку можна провести вертикальну лінію, зліва від якої залишиться 120 із 140 оцінок даного інтервалу. В результаті 500 оцінок будуть знаходитись зліва від цієї лінії, а знайдена точка перетину з віссю вкаже на значення медіани. Таким чином, 120/140 або 6/7 площі прямокутника залишиться зліва, а 1/7 – справа. Проведемо, відповідно, вертикальну лінію, яка перетне горизонтальну вісь в точці приблизно 63.

Оскільки вершина кожного стовпчика розташована горизонтально, це означає, що при побудові гістограми і визначенні медіани припускалось, що 140 оцінок (округлених до 60) мають рівномірний розподіл в даному інтервалі. Зазначимо, що половина загальної площі прямокутників гістограми розташована зліва і справа від пунктирної лінії.

Аналогічно, гістограму можна використати для *визначення процентилів* ряду. *Нагадаємо, що процентиль може бути окремим значенням ряду або ж представляти собою величину, яка лежить між двома значеннями*. Так, 49-й процентиль ряду оцінок за тест дорівнює оцінці, яка перевищує 49% оцінок + половина значення (в %) даної оцінки. Для того, щоб визначити 49-й процентиль, необхідно знайти на горизонтальній осі таке число, яке було б більшим 49% оцінок. Іншими словами, необхідно так провести вертикальну лінію на графіку, щоб 49% площі знаходилось зліва від неї і 51% – справа.

Зазначимо, що 49% від 1000 складає 490 оцінок. В шести нижніх інтервалах знаходиться 380 оцінок, а в шести верхніх – 520. Це означає, що 49-й процентиль знаходиться між значеннями $55{,}5 \div 64{,}5$. Очевидно, що 110/140 або 11/14 площі прямокутника над цим інтервалом необхідно розташувати зліва

від лінії і 3/14 – справа. Ця лінія показана на рис. 6.1 крапками. Вона перетинає горизонтальну вісь приблизно в точці 62,57. Таким чином, 49% загальної площі гістограми лежить зліва від цієї лінії і 51% – справа. Оскільки медіана є 50-м процентилем, то її можна знайти аналогічно розглянутій процедурі.

4.2 Обчислення медіани і процентилів

Обчислення значення процентиля для інтервального ряду розподілу є просто чисельною версією розглянутої вище процедури. Розглянемо для початку смисл показників останнього стовпчика табл. 4.2.

Таблиця 4.2. Накопичені частоти для ряду оцінок за тест

Границі	Частоти	Накопичені частоти
91,5 – 100,5	60	1000
82,5 – 91,5	140	940
73,5 – 82,5	160	800
64,5 – 73,5	120	640
55,5 – 64,5	140	520
46,5 – 55,5	80	380
37,5 – 46,5	119	300
28,5 – 37,5	81	181
19,5 – 28,5	50	100
10,5 – 19,5	32	50
1,5 – 10,5	18	18

В останньому стовпчику наведено так звані накопичені частоти, які представляють собою суму частот сусідніх інтервалів, починаючи знизу. Накопичені частоти інтерпретують наступним чином: є 18 оцінок, менших 10,5; 50 оцінок, менших 19,5; 800 оцінок, менших ніж 82,5. Тобто, відношення „*менше ніж*” відноситься до верхньої границі інтервалу.

Для того щоб знайти медіану, необхідно визначити точку, в якій сума накопичених частот буде дорівнювати 500. Ця точка знаходиться в інтервалі $55{,}5 \div 64{,}5$, в якому знаходиться 140 оцінок. З цього числа необхідно 500 – 380 = 120 необхідно додати до попереднього значення накопиченої частоти. Очевидно, що точку, яка відповідає медіані знайдемо наступним чином:

$$Me = 55{,}5 + \frac{120}{140} \cdot 9 = 55{,}5 + \frac{6}{7} \cdot 9 = 63{,}2.$$

Для того, щоб знайти 35-й процентиль, необхідно визначити точку, в якій

сума накопичених частот складає 35% від 1000, тобто, 350. Із стовпчика накопичених частот табл. 4.2 видно, що вона знаходиться в інтервалі $46{,}5 \div 55{,}5$. Менше значення 46,5 знаходиться 300 оцінок. Таким чином, до цього значення необхідно додати ще 50 (з 80-ти оцінок, що знаходяться в інтервалі $46{,}5 \div 55{,}5$). Тепер тридцять п'ятий процентиль знайдемо як

$$Пр(35) = 46{,}5 + \frac{50}{80} \cdot 9 = 52{,}1.$$

Загальна формула для обчислення процентилів має вигляд:

$$Пр(n) = L + \frac{s}{f} I, \qquad (4.2.1)$$

де L – нижня границя інтервалу, в який попадає значення $Пр(n)$; f – число елементів в даному інтервалі; s – число елементів, яке необхідно для попадання в точку на горизонтальній осі, що відповідає даному процентилю; I – ширина інтервалу.

4.3 Квартилі і децилі

Крім процентиля використовують такі поняття, як *квартиль* і *дециль*. Поняття квартиля відноситься до четвертої частини сукупності даних, а дециля – до десятої частини.

Для попереднього прикладу з оцінками тестування перший квартиль, або Q_1, дорівнює значенню, нижче якого розташовано 25% оцінок, а вище – 75%. Іншими словами, це ще одна назва для $Пр(25)$. Другий квартиль збігається з медіаною і $Пр(50)$, а третій, Q_3 – це $Пр(75)$.

Таким чином, оцінка точно попадає в *перший* (нижній) *квартиль*, якщо вона *перевищує* 25% оцінок. Аналогічно, оцінка попадає в *верхній квартиль*, Q_3, якщо вона *перевищує рівно* 75% оцінок. І, насамкінець, оцінка знаходиться у „верхній чверті", якщо її процентильний ранг попадає в інтервал $75 \div 100$.

Поняття *дециля* відноситься, відповідно, до десятих часток сукупності даних. Перший дециль, або D_1, відповідає значенню оцінки, менше якого знаходяться рівно 10% оцінок. Очевидно, що $D_1 = Пр(10)$, $D_2 = Пр(20)$ і так далі.

4.4 Кумулятивна крива

Якщо по осі абсцис відкласти значення інтервалів оцінок, а по осі ординат – значення накопичених частот, то отримаємо графік кумулятивних (накопичених) частот. Графік кумулятивних частот для розглянутого вище прикладу з оцінками за тест наведено на рис. 4.2.

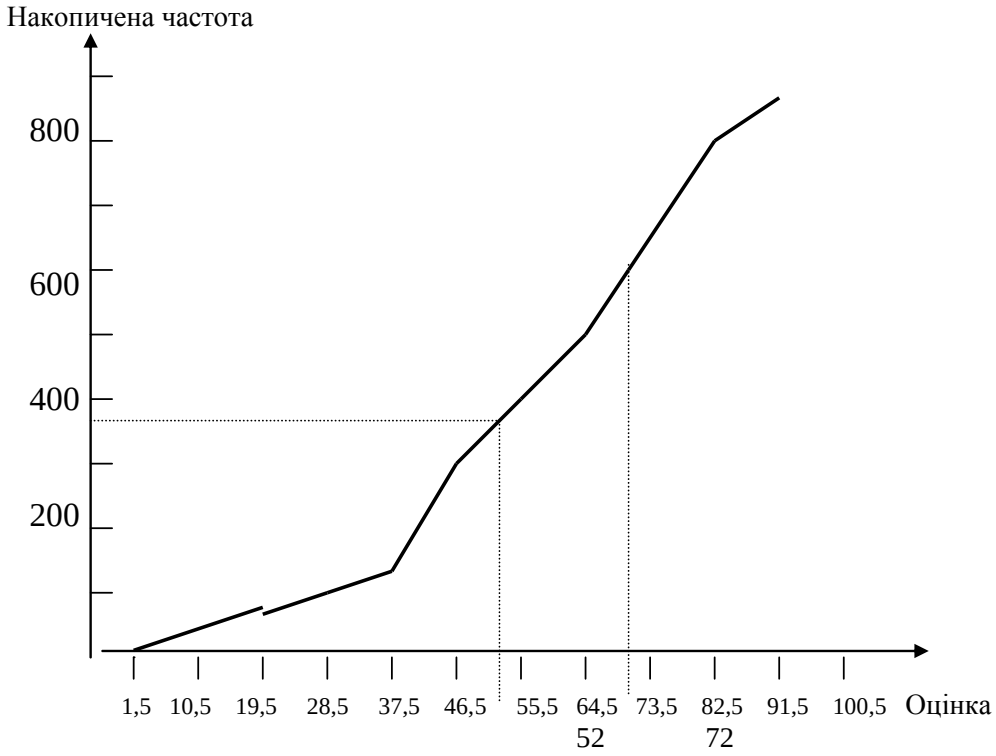


Рис. 4.2. Кумулятивна крива для оцінок за тест

Значення накопичених частот строго відповідають верхнім границям відповідних інтервалів. Для того, щоб показати, що немає менших оцінок ніж, 1,5, нанесемо відповідну точку на горизонтальній осі.

Кумулятивну криву накопичених частот можна використати для визначення процентилів і процентильних рангів. Нехай, необхідно знайти 60-й процентиль $Пр(60)$. На вертикальній осі знаходимо точку, яка відповідає 60% від 1000, тобто 600 (оскільки всього було 1000 оцінок). З графіка кумулятивної кривої знаходимо, що $Пр(60) = 70{,}5$. Для того, щоб визначити процентильний ранг, необхідно йти в зворотному порядку. Так, для оцінок 70 і 71 частоти дорівнюють приблизно 600, що складає 60%, тобто процентильний ранг складає 60.

4.5 Визначення моди згрупованих даних

Розглянемо задачу визначення моди спостережень, представлених у вигляді інтервального ряду розподілу. Раніше було встановлено, що ряд може мати дві і більше мод. Така ж ситуація може мати місце і для згрупованих даних. Спочатку необхідно вирішити, яке число ми будемо розглядати як моду.

У випадку згрупованих даних *інтервал, який характеризується найбільшою частотою, будемо називати модальним інтервалом, а його середнє значення – наближеною модою*.

Так, у випадку з оцінками за тест (табл. 4.1) модальним інтервалом є інтервал $74 \div 82$, оскільки його частота 160 є найбільшою. Наближеним значенням моди є (74+82)/2=78. Якби було два модальних інтервали, то було б

дві наближених моди. Зазначимо, що наближеною модою в подальшому ми не будемо користуватись.

4.6 Визначення середнього згрупованих даних

Раніше було встановлено, що віднімання константи від кожного значення ряду зменшує середнє на цю ж величину, а ділення кожного значення на константу приводить до зменшення середнього в таке ж число разів. В даному параграфі розглянемо розрахунок середнього на основі таблиці згрупованих частот.

При розрахунку середнього інтервального ряду необхідно взяти тільки одне значення, яке представляє всі оцінки інтервалу. При рівномірному розподілі оцінок в рамках інтервалу таким значенням є центр інтервалу. Частоти інтервалів стають частотами їх центрів.

Розрахуємо середнє розподілу методом групування для ряду оцінок за тест, який складається з 1000 елементів. Результати попередніх розрахунків, необхідних для визначення середнього, наведені в табл. 4.3. Таблиця містить добутки частот на середні значення кожного інтервалу, які необхідні для визначення середнього згрупованих даних.

Таблиця 4.3. До розрахунку середнього

Інтервали	Центри (x_i)	Частоти (f_i)	Добутки ($f_i x_i$)
92 – 100	96	60	5760
83 – 91	87	140	12180
74 – 82	78	160	12480
65 – 73	69	120	8280
56 – 64	60	140	8400
47 – 55	51	80	4080
38 – 46	42	119	4998
29 – 37	33	81	2673
20 – 28	24	50	1200
11 – 19	15	32	480
2 – 10	6	18	108
		$\sum f_i$ = 1000	$\sum f_i x_i = 60639$

Таким чином, формула для середнього ряду розподілу при розрахунку методом групування має вигляд:

$$\mu = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}, \qquad (4.6.1)$$

де x_i – центри інтервалів; f_i – значення частот інтервалів; n – число інтервалів; очевидно, що $\sum f_i = 1000$ – загальне число оцінок за тест. Для даних, наведених в таблиці 4.2, середнє для згрупованих даних: $\mu = 60639/1000 = 60{,}639$ або 60,6.

Якщо *центри інтервалів приймають дуже великі значення*, то середнє для згрупованих даних можна розрахувати також за формулою:

$$\mu = m \frac{\sum_{i=1}^{n} f_i u_i}{\sum_{i=1}^{n} f_i} + R, \qquad (4.6.2)$$

де $u_i = (x_i - R)/m$ (їх називають ще $u-$*оцінками або умовними відхиленнями*); $R-$ константа, яку віднімають від кожного x_i з метою зменшення їх значень (її вибирають, як правило, рівною одному із серединних значень інтервалів x_i); $m-$ масштабний коефіцієнт, який вибирають рівним відстані між сусідніми серединними значеннями x_i (тобто, ширині інтервалу). Наприклад, якщо користуватись цією формулою для даних, наведених в таблиці 4.3, то для цієї мети можна вибрати $R = 60$, а ширина $m = 9$.

4.7 Обчислення дисперсії і стандартного відхилення для згрупованих даних

Раніше ми отримали формулу для обчислення дисперсії не згрупованих даних у вигляді:

$$\sigma_x^2 = \frac{1}{N-1}\sum_{i=1}^{N} x_i^2 - \frac{1}{(N-1)^2}\left(\sum_{i=1}^{N} x_i\right)^2. \qquad (4.7.1)$$

Вище було показано, що віднімання константи від кожного елемента ряду не впливає на величину дисперсії і стандартного відхилення. Друга властивість полягає в тому, що ділення кожного елемента ряду на константу призводить до зменшення стандартного відхилення (за абсолютною величиною) в таке ж число разів. Для ілюстрації застосування модифікованої формули (4.7.1) скористаємось прикладом аналізу оцінок за тест (табл. 4.4).

Таблиця 4.4. Розрахунок дисперсії

Інтервали	Центри, x_i	Частоти, f_i	Добутки, $f_i x_i$	Квадрати, x_i^2	Добутки, $f_i x_i^2$
92 – 100	96	60	5760	9216	552960
83 – 91	87	140	12180	7569	1059660
74 – 82	78	160	12480	6084	973440
65 – 73	69	120	8280	4761	571320
56 – 64	60	140	8400	3600	504000
47 – 55	61	80	4080	2601	208080
38 – 46	42	119	4998	1764	209915
29 – 37	33	81	2673	1089	88209
20 – 28	24	50	1200	576	28800
11 – 19	15	32	480	225	7200
2 – 10	6	18	108	36	648
		$\sum f_i = 1000$	$\sum f_i x_i = 60639$		$\sum f_i x_i^2 = 4204233$

Дисперсію для даних, згрупованих так, як показано в табл. 4.4, розраховують за формулою:

$$\sigma_x^2 = \frac{1}{N-1}\sum_{i=1}^{n} f_i x_i^2 - \frac{1}{(N-1)^2}\left(\sum_{i=1}^{n} f_i x_i\right)^2, \qquad (4.7.2)$$

де $N = \sum_{i=1}^{n} f_i$ – загальне число елементів розподілу; n – число груп елементів (інтервалів).

Для наведених в табл. 4.4 даних дисперсія складає:

$$\sigma_x^2 = \frac{4204233}{999} - \frac{1}{999^2}(60639)^2 = 524{,}0,$$

а стандартне відхилення дорівнює: $\sigma_x = \sqrt{524} = 22.96 \approx 23$.

Для обчислення дисперсії групових даних можна скористатись також u – оцінками або *умовними відхиленнями*:

$$\sigma_u^2 = \frac{1}{N-1}\sum_{i=1}^{n} f_i u_i^2 - \frac{1}{(N-1)^2}\left(\sum_{i=1}^{n} f_i u_i\right)^2, \qquad (4.7.3)$$

де u_i – оцінки.

Значення u – оцінок для прикладу з оцінками тестування наведені в табл. 4.5.

Таблиця 4.5. Розрахунок дисперсії з використанням умовних відхилень

Центри, x_i	Частоти, f_i	Значення u_i	Добутки, $f_i u_i$	Квадрати, u_i^2	Добутки, $f_i u_i^2$
96	60	4	240	16	960
87	140	3	420	9	1260
78	160	2	320	4	640
69	120	1	120	1	120
60	140	0	0	0	0
61	80	– 1	– 80	1	80
42	119	– 2	– 238	4	476
33	81	– 3	– 243	9	729
24	50	– 4	– 200	16	800
15	32	– 5	– 160	25	800
6	18	– 6	– 108	36	648
	$\sum f_i = 1000$		$\sum f_i u_i =$ 71		$\sum f_i x_i^2 =$ 6513

Значення u_i формуються наступним чином:

- центри x_i зменшують на деяке значення, вибране з середини стовпчика значень x_i (в даному випадку вибрано 60);
- значення $u_7 = 0$ (сьомий стовпчик знизу), оскільки центр $x_7 = 60$, а $x_7' = 60 - 60 = 0$;
- для інтервалів, серединні значення яких перевищують 60, в стовпчик u_i послідовно записують числа 1, 2, 3 і так далі, а для інтервалів, серединні значення яких є меншими 60, записують числа –1, –2, –3 і так далі.

Фактично, $u-$оцінки отримано шляхом віднімання 60 від серединних значень інтервалів з наступним діленням результату на 9, тобто, на ширину інтервалів (як це мало місце при обчисленні середнього за виразом 4.6.2).

Оскільки при отриманні $u-$оцінок всі елементи початкового ряду були поділені на 9 (ширину інтервалу), то для того щоб обчислити дисперсію початкового ряду на основі σ_u^2, необхідно помножити σ_u^2 на $9^2 = 81$. Підставимо дані з табл. 4.5 в (4.7.3):

$$\sigma_u^2 = \frac{6513}{999} - \left(\frac{71}{999}\right)^2 = 6{,}465 - (0{,}071)^2 = 6{,}46,$$

а стандартне відхилення: $\sigma_u = 2{,}54$. Статистичні характеристики для початкового ряду:

$$\sigma_x = 9\sigma_u = 9 \cdot 2{,}54 = 22{,}89 \approx 23,$$
$$\sigma_x^2 = (22{,}89)^2 = 524.$$

Таким чином, остаточна формула для розрахунку стандартного відхилення ряду згрупованих оцінок з використанням умовних відхилень має вигляд:

$$\sigma_x = m\sqrt{\frac{1}{N-1}\sum_{i=1}^{n} f_i u_i^2 - \frac{1}{(N-1)^2}\left(\sum_{i=1}^{n} f_i u_i\right)^2}, \qquad (4.7.4)$$

де $m-$ довжина інтервалу. Зазначимо, що використання умовного відхилення значно прискорює і спрощує всі обчислення.

4.8 Контрольні запитання і вправи

1. Поясніть, чи обов’язково медіана повинна збігатись із конкретним значенням ряду розподілу?
2. Як визначити медіану за гістограмою?
3. Яким чином визначаються процентилі за гістограмою?
4. Поясніть такі терміни: 50-й процентиль і 88-й процентиль?
5. Запишіть і поясніть формулу для обчислення процентилів?

6. Поясніть терміни *квартиль* і *децильь*? Поясніть на числовому прикладі.
7. Яким чином будується кумулятивна крива частот? Поясніть це на прикладі отриманих балів за тест.
8. Що таке модальний інтервал і як він визначається для згрупованих даних?
9. Поясніть на числовому прикладі обчислення середнього для згрупованих даних?
10. Поясніть термін u – оцінка і наведіть приклад практичного використання цих оцінок?
11. Запишіть і поясніть вираз для обчислення дисперсії і стандартного відхилення для згрупованих даних?

Розділ 5

НОРМАЛЬНИЙ РОЗПОДІЛ

5.1 Теоретичні розподіли

Вже встановлено, що розширення ряду розподілу якомога більшою кількістю значень дає можливість точніше відтворити форму його графіка. Однак, після досягнення деякої границі додавання нових значень до ряду розподілу випадкової величини (ВВ) перестає суттєво впливати на покращення форми графіка. Після того, як форма графіка перестала суттєво змінюватись, можна досліджувати властивості даного розподілу, не враховуючи кількість його елементів. В такому випадку кажуть, що кількість спостережень є "досить великою".

У подальшому будемо використовувати вираз "*нескінченна кількість спостережень*" тоді, коли ми хочемо сказати, що додавання додаткових спостережень до ряду не може надати нам нову інформацію щодо його властивостей.

Теоретичний розподіл *– це розподіл, який складається з нескінченної кількості спостережень*.

Іншими словами, це розподіл, характеристики якого вже сформувались і про його властивості можна вже сказати, що вони стали фіксованими. В теоретичному розподілі на кожне окреме значення змінної припадає такий незначний процент від загальної кількості спостережень, що процентильний ранг цього значення можна розглядати просто як процент спостережень, що знаходяться нижче нього в розподілі.

5.2 Нормальний розподіл і його графік

Одним з конкретних видів розподілів є *нормальний або Гаусів розподіл* (НР), який відіграє дуже велику роль у статистиці, теорії оцінювання та багатьох інших науках. Більша частина досягнень статистики базується на дослідженні властивостей нормального розподілу.

Крива нормального (гаусового) розподілу представлена на рис. 5.1. На горизонтальній осі відкладено значення нормованих відхилень ($z-$оцінок). Форма кривої нормального розподілу залежить від масштабу, вибраного для $z-$оцінок.

Наприклад, крива буде мати іншу ("гострішу") форму, якщо зменшити відстань між значеннями $z-$оцінок при постійному масштабі для частот спостережень на осі ординат. Крива НР є *симетричною*, а її найвища точка знаходиться над значенням $z=0$.

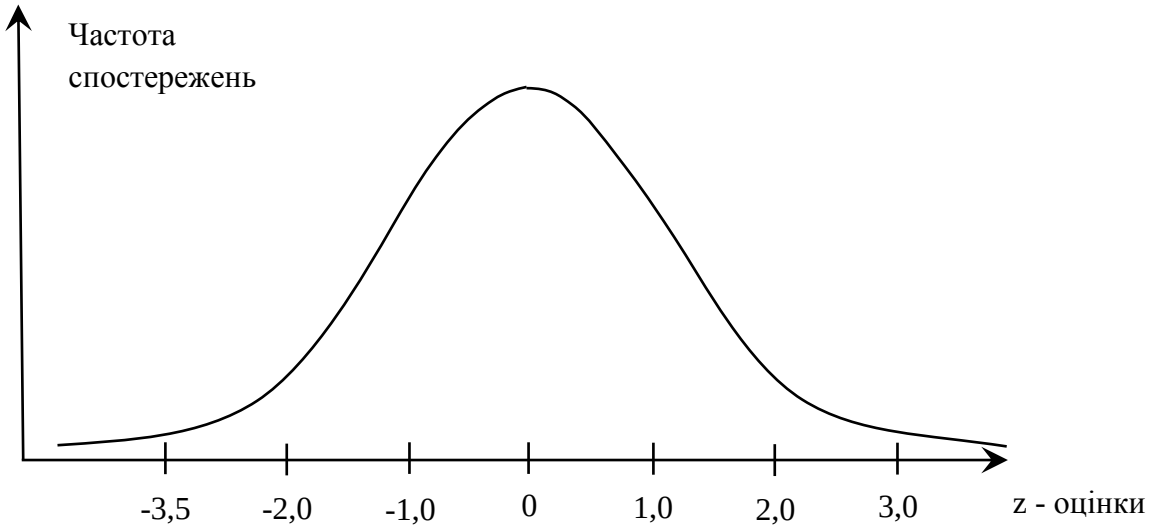


Рис. 5.1. Крива нормального розподілу

З попереднього розділу відомо, що таблиця, яка містить значення $z-$оцінок і відповідних їм процентильних рангів, є числовим описом розподілу. Така таблиця дає можливість нарисувати графік нормального розподілу без додаткової інформації.

В таблиці для нормального розподілу (вона, як правило, наводиться у додатках до підручників або у довідниках) наведено значення $z-$оцінок від – 4 до + 4. Якщо відома $z-$оцінка деякого спостереження, яке належить нормальному розподілу, то з цієї таблиці можна знайти частку спостережень, які за своєю величиною є меншими даного спостереження. Наприклад, з таблиці випливає, що спостереження із $z_i = -1{,}0$ перевищує приблизно 0,159 або 15,9% процентів спостережень розподілу. Іншими словами, спостереження із $z_i = -1{,}0$ є наближено 16-м процентилем. Спостереження, яке має $z_i = 0$, перевищує приблизно половину спостережень нормального розподілу і є, таким чином, 50-м процентилем. Спостереження із $z_i = 1{,}5$ є приблизно 93-м процентилем, оскільки воно перевищує 93,3% всіх спостережень. Чим детальнішою є така таблиця, тим точніше вона визначає нормальний розподіл.

Таким чином, твердження, що розподіл є нормальним, означає, що залежності між $z-$оцінками спостережень та їх процентильними рангами відповідає таблиці нормального розподілу.

5.3 Чотири властивості нормального розподілу

Властивості нормального розподілу сформульовано наступним чином.

1. *Значення змінної мають властивість концентруватись навколо точки* $z = 0$.

2. *Нормальна крива є симетричною відносно вертикальної осі*. Звідси випливає, що половина спостережень розташовується зліва від вертикальної осі

при $z = 0$ і має від'ємні $z-$оцінки.

3. *Значення спостережень, які належать нормальному розподілу, не обмежені за своєю величиною*. Теоретично спостереження можуть мати будь-які $z-$оцінки, наприклад, – 1000. Ця властивість пояснюється тим, що крива нормального розподілу не перетинається з віссю абсцис.

4. *Середнє, медіана і мода нормального розподілу мають одне і те ж значення*. Спостереження, яке має $z_i = 0$, якраз і є цим значенням. Крива є симетричною відносно прямої, що проходить через цю точку і в ній крива має максимальну висоту.

5.4 Функція розподілу і числові характеристики неперервних величин

Функцією розподілу (або *кумулятивною функцією розподілу*) називають функцію $F(x)$, яка визначає ймовірність того, що випадкова величина x в результаті виконання деякого досліду прийме значення менше a, тобто

$$F(x) = p(x \le a).$$

Властивості функції розподілу

Властивість 1. *Значення функції розподілу належить відрізку* [0,1]:

$$0 \le F(x) \le 1. \qquad (5.4.1)$$

Ця властивість випливає із визначення функції розподілу як ймовірності: *ймовірність* – це невід'ємне число, яке не перевищує одиниці.

Властивість 2. $F(x)-$*ниспадаюча функція, тобто,*

$$F(x_2) \ge F(x_1), \text{ якщо } x_2 > x_1. \qquad (5.4.2)$$

Наслідок 1. *Ймовірність того, що випадкова величина набуде значення, яке знаходиться в інтервалі* (a, b)*, дорівнює приросту функції розподілу на цьому інтервалі*:

$$p(a \le x < b) = F(b) - F(a). \qquad (5.4.3)$$

Наслідок 2. *Ймовірність того, що неперервна випадкова величина* x *прийме одне конкретне значення, дорівнює нулю*.

Похідну від функції розподілу називають *щільністю або густиною розподілу*, тобто

$$f(x) = F'(x).$$

Функції розподілів випадкових величин називають *статистичними моделями* випадкових процесів. Нижче будуть розглянуті статистичні моделі нормального та деяких інших розподілів.

Ймовірність попадання неперервної випадкової величини в заданий інтервал

Т е о р е м а 5.1. *Ймовірність того, що неперервна випадкова величина x прийме значення, яке належить інтервалу (a, b), дорівнює визначеному інтегралу від густини розподілу, взятому в границях від a до b*:

$$p\,(a < x < b) = \int_a^b f(x)\,dx\,.$$

Д о в е д е н н я. Скористаємось відомим співвідношенням:

$$p(a \le x < b) = F(b) - F(a)\,.$$

За формулою Лейбніца: $F(b) - F(a) = \int_a^b F'(x)\,dx = \int_a^b f(x)\,dx$. Таким чином,

$$p(a \le x < b) = \int_a^b f(x)\,dx\,.$$

Оскільки $p(a \le x < b) = p(a < x < b)$, то остаточно отримаємо:

$$p(a < x < b) = \int_a^b f(x)\,dx\,.$$

Отриманий результат геометрично можна трактувати так: ймовірність того, що неперервна випадкова величина набуде значення, яке належить інтервалу (a,b), дорівнює площі криволінійної трапеції, обмеженої віссю $0x$, кривою розподілу $f(x)$ і вертикальними прямими $x = a$ і $x = b$.

Статистичні характеристики неперервних випадкових величин

При необмеженому збільшенні кількості спостережень для дослідження розподілів випадкових величин, зручно застосовувати характеристики неперервних випадкових величин. Почнемо з математичного сподівання.

Нехай неперервна випадкова величина X задана щільністю розподілу $f(x)$. Припустимо, що X визначена на відрізку $[a, b]$. Розіб'ємо цей відрізок на n менших відрізків довжиною $\Delta x_1, \Delta x_2, \ldots, \Delta x_n$ і виберемо в кожному з них довільну точку x_i, $i = 1, \ldots, n$. Необхідно визначити математичне сподівання неперервної величини. По *аналогії з розподілом дискретних величин* складемо суму добутків можливих значень x_i на ймовірності їх попадання в інтервал Δx_i (нагадаємо, що добуток $f(x)\Delta x$ наближено дорівнює ймовірності попадання значення X в інтервал Δx):

$$\sum_{i=1}^{n} x_i\ f(x_i)\,\Delta x_i\,.$$

Якщо спрямувати довжину найбільшого із часткових відрізків до нуля, то

$$\lim_{\Delta x_i \to 0} \sum_{i=1}^{n} x_i\ f(x_i)\,\Delta x_i = \int_a^b x\ f(x)\,dx\,.$$

Означення 5.1. *Математичним сподіванням неперервної випадкової величини X, визначеної на інтервалі $[a,b]$, називають визначений інтеграл*

$$E[X]=\int_a^b x\,f(x)\,dx. \tag{5.4.4}$$

Якщо X визначена на всій осі $0x$, то

$$E[X]=\int_{-\infty}^{\infty} x\,f(x)\,dx. \tag{5.4.5}$$

В даному випадку припускається, що невласний інтеграл сходиться абсолютно, тобто існує інтеграл $\int_{-\infty}^{\infty}|x|\,f(x)\,dx$. Якби ця вимога не виконувалась, то значення інтегралу залежало б від швидкості наближення (окремо) нижньої границі до $-\infty$, а верхньої – до $+\infty$.

Означення 5.2. *Дисперсією неперервної випадкової величини називають математичне сподівання квадрату її відхилення.*

Якщо $X\in[a,b]$, то

$$Var[X]=\int_a^b [x-E(X)]^2\,f(x)\,dx. \tag{5.4.6}$$

Якщо ж $X\in[-\infty,\infty]$, то

$$Var[X]=\int_{-\infty}^{\infty} [x-E(X)]^2\,f(x)\,dx. \tag{5.4.7}$$

Примітка 1. Можна показати, що властивості математичного сподівання і дисперсії дискретних величин зберігаються і для неперервних величин.

Примітка 2. Для обчислення дисперсії можна легко отримати наступні формули:

$$Var[X]=\int_a^b x^2 f(x)\,dx-[E(X)]^2, \tag{5.4.8}$$

$$Var[X]=\int_{-\infty}^{\infty} x^2 f(x)\,dx-[E(X)]^2, \tag{5.4.9}$$

Приклад 5.1. Визначити математичне сподівання і дисперсію випадкової величини X, заданої функцією розподілу:

$$F(x)=\begin{cases}0, & x\le 0,\\ x, & 0<x\le 1,\\ 1, & x>1.\end{cases}$$

Р о з в ' я з о к : Знайдемо щільність розподілу:

$$f(x)=F'(x)=\begin{cases}0, & x\le 0,\\ 1, & 0<x\le 1,\\ 0, & x>1.\end{cases}$$

Математичне сподівання:

$$E[X]=\int_0^1 x\cdot 1\cdot dx=\frac{x^2}{2}\Big|_0^1=\frac{1}{2}.$$

Дисперсія:

$$Var[X]=\int_0^1 x^2\cdot 1\cdot dx-\left(\frac{1}{2}\right)^2=\frac{x^3}{3}\Big|_0^1-\frac{1}{4}=\frac{1}{12}.$$

Приклад 5.2. Визначити математичне сподівання і дисперсію неперервної випадкової величини X , розподіленої рівномірно в інтервалі (a,b).

Р о з в ' я з о к . Враховуючи, що густина рівномірного розподілу $f(x)=\frac{1}{b-a}$, знайдемо математичне сподівання:

$$E[X]=\int_a^b x\cdot f(x)\,dx=\frac{1}{b-a}\int_a^b x\,dx=\frac{1}{b-a}\cdot\frac{x^2}{2}\Big|_a^b=\frac{(b^2-a^2)}{2(b-a)}=\frac{a+b}{2}.$$

Дисперсія:

$$Var[X]=\int_a^b x^2\,f(x)\,dx-[E(X)]^2=\frac{1}{b-a}\int_a^b x^2dx-\left[\frac{a+b}{2}\right]^2=\frac{(b-a)^2}{12}.$$

5.5 Статистичні характеристики нормального розподілу

Статистична модель (щільність, густина) нормального розподілу, має вигляд:

$$f(x)=\frac{1}{\sigma_x\sqrt{2\pi}}\exp\left(-\frac{1}{2}\frac{(x-\mu_x)^2}{\sigma_x^2}\right), \qquad (5.5.1)$$

де σ_x^2- дисперсія ряду розподілу змінної X ; μ_x- середнє ряду; $f(x)-$ густина розподілу. Покажемо, що σ_x^2 і μ_x мають вказаний статистичний смисл.

А) За означенням математичного сподівання неперервної випадкової величини,

$$E[X]=\int_{-\infty}^{\infty} x\,f(x)\,dx=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} x\cdot\exp\left(-(x-\mu_x)^2/2\sigma_x^2\right)dx. \qquad (5.5.2)$$

Введемо нову змінну $z=(x-\mu_x)/\sigma_x$, а звідси $x=\sigma_x z+\mu_x$ і $dx=\sigma_x\,dz$. Оскільки нові границі інтегрування залишаються незмінними, то

$$E[X]=\frac{\sigma_x}{\sigma_x\sqrt{2\pi}}\int_{-\infty}^{\infty}(\sigma_x z+\mu_x)\,e^{-z^2/2}\,dz=$$

$$=\frac{\sigma_x}{\sqrt{2\pi}}\int_{-\infty}^{\infty}z\,e^{-z^2/2}\,dz+\frac{\mu_x}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-z^2/2}\,dz.$$

Оскільки під знаком інтеграла в першому доданку *непарна функція* (z може бути меншим і більшим нуля), а *границі інтегрування симетричні стосовно початку координат*, то він дорівнює нулю. Другий доданок дорівнює μ_x, оскільки

$$\int_{-\infty}^{\infty}\exp\left(-z^2/2\right)d\,z=\sqrt{2\pi}$$

є інтегралом Пуассона.

Таким чином, $E[X]=\mu_x$.

Б) За означенням дисперсії неперервної випадкової величини маємо:

$$Var[X]=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}(x-\mu_x)^2\cdot\exp\left(-(x-\mu_x)^2/2\sigma_x^2\right)dx.$$

Знову введемо змінну $z=(x-\mu_x)/\sigma_x$, а звідси $x=\sigma_x z+\mu_x$ і $dx=\sigma_x\,dz$. І враховуючи, що границі інтегрування залишаються незмінними, отримаємо:

$$Var[X]=\frac{\sigma_x}{\sigma_x\sqrt{2\pi}}\int_{-\infty}^{\infty}(\sigma_x z+\mu_x-\mu_x)^2\cdot\exp\left(-z^2/2\right)dz=.$$

$$=\frac{\sigma_x^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty}z\cdot z\cdot\exp\left(-z^2/2\right)dz=\frac{\sigma_x^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty}z\;d(-\exp\left(-z^2/2\right).$$

Покладемо: $u=z$, $dv=z\,e^{-z^2/2}\,dz$ $\left(v=\int z\,e^{-z^2/2}dz=-e^{-z^2/2}\right)$ і скористаємось методом інтегруванням за частинами: $\int u\,dv=u\,v\,|-\int v\,du$:

$$Var[X]=\frac{\sigma_x^2}{\sqrt{2\pi}}\left[u\,v\,\Big|_{-\infty}^{\infty}-\int_{-\infty}^{\infty}v\,du\right],$$

тобто,

$$Var[X]=\frac{\sigma_x^2}{\sqrt{2\pi}}\left[-z\cdot\exp\left(-\frac{z^2}{2}\right)\right]_{-\infty}^{\infty}+\frac{\sigma_x^2}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\exp\left(-\frac{z^2}{2}\right)dz=\sigma_x^2,$$

оскільки $z\cdot\exp\left(-\frac{z^2}{2}\right)\to 0$ при $|u|\to\infty$, а $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\exp\left(-\frac{z^2}{2}\right)dz=1$.

Тобто, перша компонента як і раніше (при визначенні середнього) дорівнює нулю.

Примітка 1. *Загальним нормальним розподілом* називають розподіл з довільними значеннями параметрів μ_x і σ_x^2.

Нормованим називають нормальний розподіл з параметрами $\mu = 0$ і $\sigma^2 = 1$ (відповідно, стандартне відхилення $\sigma = 1$).

Наприклад, якщо $x -$ нормально розподілена величина з довільними параметрами μ_x і σ_x, то $u = (x - \mu_x)/\sigma_x$ – нормована величина з $\mu_u = 0$ і $\sigma_u = 1$.

Нормована густина нормального розподілу:

$$\varphi(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-u^2/2\right). \qquad (5.5.3)$$

Ця функція табульована (таблиці наведені в підручниках та довідниках).

Густина розподілу для довільної випадкової змінної X зв'язана з нормованою густиною виразом:

$$p(x) = \frac{1}{\sigma}\varphi(u) \quad \text{при} \quad u = \frac{x - \mu_x}{\sigma}.$$

Примітка 2. Функція $F(x)$ загального нормального розподілу визначається за формулою:

$$F(y) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-(y-\mu_y)^2/2\sigma_y^2\right) dy, \qquad (5.5.4)$$

де $y -$ змінна, яка має нормальний розподіл.

Нормована функція нормованого розподілу (функція Лапласа) визначається як

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-z^2/2\right) dz. \qquad (5.5.5)$$

Значення функції $F_0(x)$ зведені в таблицю. Можна легко перевірити, що $F(x) = \Phi((x - \mu_x)/\sigma_x)$:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(t-\mu)^2/2\sigma^2} dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-u^2/2} du = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Примітка 3. Ймовірність попадання нормованої нормальної величини X в інтервал $(0, x)$ можна знайти за допомогою функції Лапласа

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{0}^{x} \exp\left(-z^2/2\right) dz.$$

Дійсно,

$$p(0 < X < x) = \int_{0}^{x} \varphi(x)\, dx = \frac{1}{\sqrt{2\pi}} \int_{0}^{x} \exp\left(-z^2/2\right) dz = \Phi(x). \qquad (5.5.6)$$

Примітка 4. Враховуючи, що $\int_{-\infty}^{\infty} \varphi(x)\,dx = 1$, а в силу симетрії $\varphi(x)$ відносно нуля

$$\int_{-\infty}^{0} \varphi(x)\,dx = 0{,}5, \quad \text{а значить і} \quad p(-\infty < X < 0) = 0{,}5,$$

то легко визначити, що

$$F(x) = 0{,}5 + \Phi(x).$$

Дійсно,

$$F(x) = p(-\infty < X < x) = p(-\infty < X < 0) + p(0 < X < x) = \\ = 0{,}5 + \Phi(x).$$

5.6 Дослідження кривої нормального розподілу

Виконаємо дослідження функції

$$y = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-(x-a)^2 / 2\sigma^2\right]$$

методами диференціального числення. Для спрощення запису середнє μ_x позначено буквою a.

1. *Функція визначена на всій осі* x.
2. При всіх значеннях x *функція приймає додатні значення* (парна функція), тобто, нормальна крива розташована над віссю $0x$.
3. *Границя функції* при необмеженому зростанні модуля x дорівнює нулю:

$$\lim_{|x|\to\infty} y = 0,$$

тобто вісь $0x$ служить горизонтальною асимптотою графіка нормального розподілу.

4. *Дослідимо функцію на екстремум*. Запишемо першу похідну:

$$y' = -\frac{x-a}{\sigma^3\sqrt{2\pi}} \exp\left[-(x-a)^2 / 2\sigma^2\right].$$

Видно, що $y' = 0$ при $x = a$; $y' > 0$ при $x < a$; $y' < 0$ при $x > a$.

Тобто, функція має максимум $\frac{1}{\sigma\sqrt{2\pi}}$ при $x = a$.

5. Різниця $x - a$ в квадраті міститься в аналітичному виразі функції, тобто, *графік функції є симетричним стосовно прямої* $x = a$.

6. *Дослідимо функцію на точки перегину*. Знайдемо другу похідну:

$$y'' = -\frac{1}{\sigma^3\sqrt{2\pi}} e^{-(x-a)^2/2\sigma^2}\left[1-\frac{(x-a)^2}{\sigma^2}\right].$$

Видно, що друга похідна дорівнює нулю при $x = a+\sigma$ і $x = a-\sigma$, а при переході через ці точки вона змінює знак (в обох цих точках значення функції дорівнює $\frac{1}{\sigma\sqrt{2\pi e}}$).

Таким чином, точками перегину графіка є наступні:

$$\left(a-\sigma,\ \frac{1}{\sigma\sqrt{2\pi e}}\right) \quad \text{і} \quad \left(a+\sigma,\ \frac{1}{\sigma\sqrt{2\pi e}}\right).$$

5.7 Вплив параметрів нормального розподілу на форму нормальної кривої

Знову позначимо середнє розподілу $\mu_x = a$ і встановимо, як впливають параметри a і σ на форму і розташування нормальної кривої.

Відомо, що форма графіка нормального розподілу не залежить від значення середнього a. Тобто, графіки функцій $f(x)$ і $f(x-a)$ мають однакову форму. При $a>0$ крива буде розташована справа від осі ординат, а при $a<0$ крива буде розташована зліва від цієї осі.

Звідси можна зробити *висновок*: *зміна величини математичного сподівання не змінює форми нормального розподілу, а призводить лише до її зсуву вздовж осі* $0x$ *– вправо при зростанні значення середнього, і вліво – при зменшенні значення середнього*.

Тепер розглянемо вплив стандартного відхилення. Максимум функції нормального розподілу дорівнює $1/(\sigma\sqrt{2\pi})$. Звідси випливає *висновок*: *зростання* σ *призводить до зменшення максимальної ординати нормальної кривої, а сама крива стає більш пологою; при зменшенні* σ *вершина нормальної кривої загострюється і розтягується в додатному напрямку осі* $0y$.

Зазначимо, що при будь-яких значеннях параметрів a і σ площа, обмежена нормальною кривою та віссю $0x$, залишається рівною одиниці (друга властивість густини розподілу).

5.8 Ймовірність попадання випадкової нормальної величини у заданий інтервал

З попереднього аналізу відомо, що ймовірність попадання значення випадкової величини X в інтервал (α, β) визначається за виразом:

$$p(\alpha < X < \beta) = \int_{\alpha}^{\beta} f(x)\,dx\,. \quad (5.8.1)$$

Нехай випадкова величина X розподілена за нормальним законом. Ймовірність того, що X прийме значення, яке належить інтервалу (α, β), дорівнює:

$$p(\alpha < X < \beta) = \frac{1}{\sigma\sqrt{2\pi}} \int_{\alpha}^{\beta} \exp\left[-(x-a)^2 / 2\sigma^2\right] dx. \qquad (5.8.2)$$

Виконаємо перетворення цієї формули так, щоб можна було скористатись готовими таблицями. Введемо нову змінну $z = (x-a)/\sigma$. Звідси маємо: $x = \sigma z + a$ і $dx = \sigma dz$. Нові границі інтегрування будуть такими: якщо $x = \alpha$, то $z = (\alpha - a)/\sigma$; а якщо $x = \beta$, то $z = (\beta - a)/\sigma$.

Таким чином, маємо:

$$p(\alpha < X < \beta) = \frac{1}{\sigma\sqrt{2\pi}} \int_{(\alpha-a)/\sigma}^{(\beta-a)/\sigma} \exp\left(-z^2/2\right) dz =$$

$$= \frac{1}{\sqrt{2\pi}} \int_{(\alpha-a)/\sigma}^{0} \exp\left(-z^2/2\right) dz + \frac{1}{\sqrt{2\pi}} \int_{0}^{(\beta-a)/\sigma} \exp\left(-z^2/2\right) dz =$$

$$= \frac{1}{\sqrt{2\pi}} \int_{0}^{(\beta-a)/\sigma} \exp\left(-z^2/2\right) dz - \frac{1}{\sqrt{2\pi}} \int_{0}^{(\alpha-a)/\sigma} \exp\left(-z^2/2\right) dz.$$

Користуючись функцією Лапласа

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{0}^{x} e^{-z^2/2}\, dz,$$

остаточно запишемо:

$$p(\alpha < X < \beta) = \Phi\left(\frac{\beta - a}{\sigma}\right) - \Phi\left(\frac{\alpha - a}{\sigma}\right). \qquad (5.8.3)$$

Приклад 5.3. Випадкова величина X має нормальний розподіл. Математичне сподівання і стандартне відхилення цієї величини дорівнюють 30 і 10, відповідно. Знайти ймовірність того, що X прийме значення в інтервалі (10, 50).

Р о з в ' я з о к . Скористаємось формулою (5.8.3) за умови, що $\alpha = 10,\ \beta = 50,\ a = 30$ і $\sigma = 10$:

$$p(10 < X < 50) = \Phi\left(\frac{50-30}{10}\right) - \Phi\left(\frac{10-30}{10}\right) = 2\Phi(2).$$

З таблиці для функції Лапласа знайдемо, що $\Phi(2) = 0{,}4772$, а тому

$$p(10 < X < 50) = 2 \cdot 0{,}4772 = 0{,}9544.$$

5.9 Ймовірність отримання випадковою нормальною величиною заданого значення відхилення

В прикладних дослідженнях часто виникає задача обчислення

ймовірності того, що відхилення випадкової нормальної величини X є меншим за модулем заданого додатного числа δ, тобто, $|X-\mu_x|<\delta$.

Позначимо середнє $\mu_x = a$ і замінимо цю нерівність рівносильною нерівністю:

$$-\delta < X-a < \delta \quad \text{або} \quad a-\delta < X < a+\delta.$$

У відповідності до виразу (5.8.3) можна отримати:

$$p(|X-a|<\delta) = p(a-\delta<X<a+\delta) =$$
$$= \Phi\left[\frac{(a+\delta)-a}{\sigma}\right] - \Phi\left[\frac{(a-\delta)-a}{\sigma}\right] = \Phi\left(\frac{\delta}{\sigma}\right) - \Phi\left(-\frac{\delta}{\sigma}\right).$$

Оскільки функція Лапласа Φ –непарна, то $\Phi(-\delta/\sigma) = -\Phi(\delta/\sigma)$. Остаточно можна записати, що

$$p(|X-a|<\delta) = 2\Phi(\delta/\sigma). \qquad (5.9.1)$$

Зокрема, при $a=0$:

$$p(|X|<\delta) = 2\Phi(\delta/\sigma). \qquad (5.9.2)$$

Нормальна крива стає “гострішою” при зменшенні значення σ. Звідси випливає, що для двох центрованих процесів ($a=0$) ймовірність прийняття значення в інтервалі $(-\delta, \delta)$ є більшою у тієї величини, яка має менше σ. Це пояснюється тим, що частина графіка в інтервалі $(-\delta, \delta)$ буде мати більшу площу. Крім того, цей факт повністю відповідає ймовірнісному смислу параметра σ – чим меншим є розсіювання випадкової величини, тим більшою буде ймовірність попадання в інтервал $(-\delta, \delta)$.

Примітка. Очевидно, що події, які зв’язані з нерівностями $|X-a|<\delta$ і $|X-a|\geq\delta$, носять взаємно виключний характер. Тому, якщо ймовірність виконання нерівності $|X-a|<\delta$ дорівнює p, то ймовірність виконання нерівності $|X-a|\geq\delta$ буде $(1-p)$.

Приклад 5.4. Випадкова нормальна величина X має $\mu_x = 20$ і $\sigma_x = 10$. Знайти ймовірність того, що абсолютне значення відхилення буде меншим 3 $(\delta = 3)$.

Р о з в ’ я з о к . Скористаємось формулою (5.9.1):

$$p(|X-a|<\delta) = 2\Phi(\delta/\sigma) = p(|X-20|<3) = 2\Phi(3/10) =$$
$$= 2\Phi(0{,}33) = 2\cdot 0{,}1179 = 0{,}2358.$$

5.10 Правило трьох сигм

У виразі (5.9.1), тобто $p(|X-a|<\delta) = 2\Phi(\delta/\sigma)$, покладемо $\delta = \sigma q$:

$$p(|X-a|<\sigma q) = 2\Phi(q).$$

Якщо $q = 3$, тобто $\sigma q = 3\sigma$, то

$$p(|X - a| < 3\sigma) = 2\Phi(3) = 2 \cdot 0{,}49865 = 0{,}9973, \qquad (5.10.1)$$

тобто, ймовірність того, що абсолютне відхилення буде меншим трьох стандартних відхилень, дорівнює 0,9973.

Іншими словами, ймовірність того, що абсолютна величина відхилення випадкової величини *перевищить три стандартних відхилення* (СВ), є дуже малою, 0,0027. Це означає, що така ситуація може мати місце тільки у 0,27% випадків. Якщо скористатись *принципом неможливості малоймовірних подій*, то можна сказати, що такі ситуації є практично неможливими. Таким чином, суть правила трьох сигм зводиться до такого:

Абсолютна величина відхилення нормальної випадкової величини від її математичного сподівання не перевищує трьох стандартних відхилень.

На практиці правило трьох сигм застосовують так: якщо розподіл випадкової величини не відомий, але умова (5.9.2) виконується, то можна припустити, що ця величина розподілена нормально; інакше, вона має інший розподіл.

5.11 Оцінка відхилення теоретичного розподілу від нормального; асиметрія і ексцес

Емпіричним називають *розподіл відносних частот даних*. Його досліджують методами математичної статистики.

Теоретичним називають такий розподіл випадкової величини, подальше збільшення спостережень якої не впливає на форму кривої розподілу.

При дослідженні розподілів, які є відмінними від нормального, виникає необхідність знайти кількісну міру цієї відмінності. З цією метою введено спеціальні характеристики: *асиметрію, ексцес, статистику Жак-Бера*.

Для нормального розподілу асиметрія і ексцес дорівнюють нулю. Якщо асиметрія і ексцес мають для досліджуваного процесу невеликі значення, то можна припустити близькість цього розподілу до нормального.

Оцінювання асиметрії

Можна показати, що для симетричного розподілу кожний центральний момент непарного порядку дорівнює нулю.

Нагадаємо, що момент порядку p визначається як $m_p = E[X^p]$, а центральний момент порядку p – як $\mu_p E[(X - E[X])^p]$.

Для несиметричних розподілів центральні моменти непарного порядку є відмінними від нуля. Тому будь-який з цих центральних моментів (крім моменту

першого порядку, який дорівнює нулю для будь-якого розподілу) може служити для оцінки величини асиметрії.

Логічно вибрати для цієї мети самий простий з них, тобто, момент третього порядку μ_3. Однак, приймати цей момент для оцінювання асиметрії незручно, тому що його величина залежить від одиниць, в яких вимірюється випадкова величина. Для того, щоб позбутись цього недоліку, μ_3 ділять на σ^3 і , таким чином, отримують безрозмірну величину.

Асиметрією (skewness) розподілу називають відношення центрального моменту третього порядку до куба стандартного відхилення (він характеризується симетричність хвостів розподілу):

$$S = \frac{\mu_3}{\sigma^3} \tag{5.11.1}$$

або

$$S = \frac{1}{N}\sum_{k=1}^{N}\left[\frac{y(k)-\bar{y}}{\sigma}\right]^3 \tag{5.11.2}$$

Якщо $S > 0$, то правий хвіст розподілу довший, а при $S < 0$ довшим є лівий хвіст розподілу; якщо $S = 0$, то розподіл симетричний.

Практично знак S визначають за розміщенням кривої розподілу відносно моди, тобто, точки максимуму диференціальної функції: якщо "довга частина" кривої розташована правіше моди, то асиметрія додатна, а якщо зліва, то від'ємна.

Для оцінювання величини нахилу ("крутизни") кривої розподілу у порівнянні з кривою теоретичного розподілу користуються ексцесом.

Ексцес (kurtosis) – характеризує відмінність форми розподілу від нормального і розраховується за виразом:

$$E_k = \frac{\mu_4}{\sigma^4} - 3. \tag{5.11.3}$$

Для нормального розподілу $\mu_4 / \sigma^4 = 3$, а тому ексцес $E_k = 0$. Якщо ексцес додатний, то крива має вищу і "гострішу" вершину, ніж нормальний розподіл; якщо ексцес від'ємний, то досліджувана крива має нижчу і "плоскішу" вершину, ніж нормальна крива.

В англомовній літературі аналогом терміну ***ексцес*** є ***куртозис*** (*kurtosis*):

$$K = \frac{1}{N}\sum_{k=1}^{N}\left[\frac{y(k)-\bar{y}}{\sigma}\right]^4. \tag{5.11.4}$$

$K = 3$ для нормального розподілу; якщо $K > 3$, то форма розподілу буде "*гострішою*" від нормального; при $K < 3$ форма розподілу буде "*плоскішою*" від нормального.

Статистика Жак-Бера (Jarque-Bera) – тестова статистика, яка показує наскільки близьким є ряд до нормального розподілу; фактично це різниця між значеннями S і K для досліджуваного ряду та нормального розподілу:

$$JB = \frac{N-n}{6}\left[S^2 + \frac{1}{4}(K-3)^2\right], \qquad (5.11.5)$$

де n – число коефіцієнтів, використаних для побудови моделі ряду; при нуль-гіпотезі щодо нормальності розподілу статистика Жак-Бера має розподіл χ^2 з двома степенями свободи. Ймовірність, зв'язана із статистикою Жак-Бера, показує ймовірність справедливості нуль-гіпотези.

Мала ймовірність свідчить про те, що нуль-гіпотезу щодо нормальності розподілу необхідно відхилити.

5.12 Використання таблиці площ нормального розподілу

В таблиці для площі нормального розподілу наведено частини спостережень, які знаходяться між оцінкою $z = 0$ та деяким іншим вибраним значенням $z-$оцінки.

Припустимо, що необхідно визначити, яка частина спостережень має $z-$оцінки, які знаходяться між $z = 0$ і $z = 0{,}7$. За допомогою лівої колонки при $z = 0{,}7$ та другого стовпчика $z = 0{,}0$ знайдемо значення 0,2580. Іншими словами, 25,8% спостережень нормального розподілу мають $z-$оцінки в інтервалі $0 \div 0{,}7$.

Оскільки нормальний розподіл симетричний, то в інтервалі $z = -0{,}7 \div 0{,}7$ буде знаходитись 51,6% спостережень.

Нехай, необхідно знайти частку спостережень нормального розподілу, які мають $z-$оцінки в інтервалі: $z = -1{,}96 \div 1{,}96$. Між $z = -1{,}96$ і $z = 0$ знаходяться 47,5% спостережень. Таким чином, всього в інтервалі $z = -1{,}96 \div 1{,}96$ знаходиться 95% спостережень.

Таким чином, всього в інтервалі $z = -1{,}96 \div 1{,}96$ знаходиться 95% спостережень.

Якщо необхідно знайти, яка частина спостережень має $z-$оцінки в інтервалі $z = 0{,}5 \div 1{,}0$, то необхідно спочатку знайти відсоток спостережень в інтервалі $z = 0 \div 1{,}0$, а потім з нього відняти відсоток спостережень, які знаходяться в інтервалі $z = 0 \div 0{,}5$. Тобто, отримаємо таке значення:

$$34{,}13\,\% - 19{,}15\,\% = 14{,}98\,\%.$$

Приклад 5.5. Розглянемо розподіл значень коефіцієнтів розумового розвитку (КРР) чоловіків в США. Його середнє складає 100, а стандартне відхилення 15 пунктів за тестом Векслера, тобто, $\mu = 100,\ \sigma = 15$.

Таким чином, якщо КРР = 115, то його $z = 1{,}0$, а якщо КРР = 130, то його $z = 2$. Процентильний ранг для КРР = 115 можна визначити так: в інтервалі $z = 0 \div 1{,}0$ знаходяться 34,13% оцінок, а зліва від $z = 0$ знаходяться 50% оцінок. Таким чином, процентильний ранг для КРР = 115 складає:

$$Пр(115) = 50{,}0\% + 34{,}13\ \% = 84{,}13\ \%.$$

Цей результат свідчить про те, що всі чоловіки, які мають КРР = 115, перевищують оцінки приблизно 84% всіх інших індивідуумів.

Для КРР = 70 $z-$оцінка складає $z = -2$. Тобто, ті чоловіки, які отримали 70 пунктів за тест, перевищують оцінки 2% індивідуумів, тобто, вони мають 2-й процентиль. Як визначити, скільки чоловіків мають КРР в інтервалі КРР = $90 \div 110$? Фактично, це задача визначення частки спостережень, $z-$оцінки, яких лежать в інтервалі від $z = -0{,}67$ ($z_{90} = (90-100)/15 \approx -0{,}67$) до $z = +0{,}67$ ($z_{110} = (110-100)/15 \approx 0{,}67$). Це буде

$$x_{90-110} = 24{,}86\% + 24{,}86\ \% = 49{,}72\ \%.$$

Визначення $z-$оцінок на основі процентильних рангів

Ця задача є оберненою до попередньої. Припустимо, що необхідно встановити, яка $z-$оцінка має процентильний ранг 35. Це означає, що 35% площі під графіком розташовано зліва від шуканої оцінки і 15% між цією оцінкою і $z = 0$. Найближчим до 0,15 табличним значенням є 0,1517, а відповідна йому $z-$оцінка складає: $z = 0{,}39$.

Практичне значення цієї задачі може бути наступним. Припустимо, що ми маємо інформацію тільки щодо 10% спостережень, які є найбільшими за своєю величиною. Відомо, що ці 10% спостережень належать нормальному розподілу із середнім $\mu = 80$ і стандартним відхиленням $\sigma = 5$. Необхідно визначити $z-$оцінку граничної (зліва) точки. Тобто, необхідно знайти $z-$оцінку для процентильного рангу 90. Між $z = 0$ та шуканою оцінкою знаходиться 40% площі. Найближчим до 0,4 значенням є 0,3997. Йому відповідає $z = 1{,}28$, що складає 1,28 стандартного відхилення, а шуканим значенням граничної точки буде

$$x_{гран} = 80 + 1{,}28 \cdot 5 = 86{,}4.$$

Застосування таблиці площ нормального розподілу

Розташування спостережень в нормальному розподілі може бути визначене за допомогою значень середнього та стандартного відхилення. Дійсно, можна дати повне описання нормально розподілених спостережень тальки за допомогою μ і σ.

Так, педагоги і психологи часто застосовують тести, результати яких потім узагальнюються на велике число спостережень. Розподіл отриманих оцінок дуже часто буває нормальним. Розраховані значення середнього та стандартного відхилення фіксують і в подальшому використовують для визначення точного розташування будь-якого значення оцінки в розподілі. Процедура полягає в перетворенні спостереження в $z-$оцінку з наступним

визначенням положення цієї оцінки в нормальному розподілі за допомогою таблиці.

Якщо таблицю ввести в базу даних, то процес статистичного аналізу даних можна автоматизувати і, таким чином, суттєво прискорити його реалізацію. Таблиця може бути також частиною системи підтримки прийняття рішень при виконанні аналізу даних.

5.13 Поняття про теорему Ляпунова. Формулювання центральної граничної теореми (ЦГТ)

Широке розповсюдження нормально розподілених величин на практиці можна пояснити за допомогою ***теореми Ляпунова***:

Якщо випадкова величина X представляє собою суму дуже великого числа взаємно незалежних випадкових величин і при цьому вплив кожної з них на всю суму є незначним, то X має близький до нормального розподіл.

Існує дещо **інше формулювання центральної граничної теореми**:

Припустимо, що із нескінченної сукупності значень сформовано деяку множину випадкових вибірок. Для кожної вибірки обчислено суму її значень, а з отриманих сум сформовано достатньо великий новий ряд розподілу, який є нормальним.

Приклад 5.5. На функціонування людського організму впливає багато факторів – температура навколишнього середовища, атмосферний тиск, хімічний склад їжі, час приймання їжі та її об'єм, характер спілкування з іншими індивідуумами, шум, ситуація з транспортом та інші. Кожний з цих факторів спричиняє деякі незначні порушення функціонування нервової системи та організму в цілому. Однак, оскільки загальне число цих факторів є великим, то їх сукупна дія породжує помітний сумарний вплив. Цей вплив проявляється у втомі, дратівливості, зниженню працездатності і т. ін.

Таким чином, ми можемо розглядати сумарний негативний вплив на функціонування організму як суму великого числа взаємно незалежних часткових впливів. Це дає підставу стверджувати, що сумарний негативний вплив (якщо його коректно виміряти) має розподіл, близький до нормального.

Подібних прикладів утворення нормального розподілу можна навести безліч. Класичним прикладом є стрільба по мішені. Точність попадання залежить від освітленості місця, наявності вітру, стану зору, випадкових коливань карабіна, випадкових звуків, тіней від навколишніх дерев, будівель і т. ін.

Наведемо формулювання центральної граничної теореми, яка встановлює умови, при яких сума великого числа незалежних доданків має близький до нормального розподіл.

Теорема 5.1. *Нехай $X_1, X_2, ..., X_n, ...$ – послідовність незалежних випадкових величин, кожна з яких має скінченне математичне сподівання і дисперсію:*

$$E[X_k] = \mu_k, \quad Var[X_k] = \sigma_k^2.$$

Введемо позначення:

$$S_n = X_1 + X_2 + \ldots + X_n, \quad A_n = \sum_{k=1}^{n} \mu_k, \quad B^2 = \sum_{k=1}^{n} \sigma_k^2 .$$

Позначимо функцію розподілу нормованої суми через

$$F_n(x) = p\left(\frac{S_n - A_n}{B_n} < x \right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-z^2/2\right) dz .$$

До послідовності $X_1, X_2, \ldots$ можна застосувати центральну граничну теорему, якщо при будь-якому значенні x функція розподілу нормованої суми прямує до функції нормального розподілу при $n \to \infty$:

$$\lim_{n\to\infty} p\left[\frac{S_n - A_n}{B_n} < x \right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-z^2/2\right) dz . \qquad (5.13.1)$$

Зокрема, якщо всі випадкові величин $X_1, X_2, \ldots$ мають однаковий розподіл, то до цієї послідовності можна застосувати центральну граничну теорему при умові, що дисперсії всіх величин скінченні і відмінні від нуля.

А.М. Ляпунов довів наступну теорему: якщо для $\delta > 0$ при $n \to \infty$ відношення Ляпунова

$$L_n = \frac{C_n}{B_n^{2+\delta}}, \ де \ C_n = \sum_{k=1}^{n} E\left| X_k - \mu_k \right|^{2+\delta} \qquad (5.13.2)$$

прямує до нуля (*умова Ляпунова*), то до послідовності $X_1, X_2, \ldots$ можна застосувати центральну граничну теорему.

Умова Ляпунова полягає в наступному: вона висуває вимогу до того, щоб кожний доданок у незначній мірі впливав на суму $(S_n - A_n)/B_n$.

Примітка. Для доведення центральної граничної теореми А.М. Ляпунов скористався апаратом характеристичних функцій. Характеристичною функцією випадкової величини X називають функцію $\gamma(t) = E[e^{itX}]$.

Для дискретної випадкової величини X з можливими значеннями x_k та їх ймовірностями p_k характеристична функція

$$\gamma(t) = \sum_{k} e^{itx_k} \, p_k .$$

Для неперервної випадкової величини X з щільністю розподілу $f(x)$ характеристична функція

$$\gamma(t) = \int_{-\infty}^{\infty} e^{itx} \, f(x)\, dx .$$

Можна довести, що характеристична функція суми незалежних випадкових величин дорівнює добутку характеристичних функцій доданків.

Центральна гранична теорема для вибіркових середніх

Припустимо, що сформовано множину випадкових вибірок по 200 значень в кожній. Замість суми значень кожної вибірки обчислимо їх середні і сформуємо новий розподіл із цих середніх. Наступна теорема стверджує, що отриманий розподіл буде нормальним.

Теорема 5.2. (**ЦГТ для вибіркових середніх**). *Середні значення деякої достатньо великої множини випадкових вибірок, сформованих з нескінченної сукупності спостережень, утворюють нормальний розподіл*.

Нагадаємо, що кожна вибірка повинна складатись з однакового числа елементів. Вибірковий розподіл середніх представляє собою нормований розподіл сум, а тому він точно відображає їх вибірковий розподіл. Це можна довести наступним чином: відомо, що ділення кожного елемента ряду розподілу на одну й ту ж константу не впливає ні на $z-$оцінку ні на процентильний ранг елементів. Таким чином операція ділення всіх N елементів розподілу (в складі суми) на N залишає без зміни пари значень $z-$оцінок і процентильних рангів. З цього випливає, що вибірковий розподіл середніх також буде нормальним.

5.14 Властивості ряду розподілу середніх

Припустимо, що розподіл, сформований із середніх, є нескінченним рядом. Це означає, що із початкової (генеральної) сукупності було утворено нескінченне число вибірок. Існують важливі зв'язки між значеннями *середнього і стандартного відхилення початкової* (*генеральної*) *сукупності*, а також *середнім і стандартним відхиленням розподілу середніх*, взятих з цієї сукупності. При цьому середнє розподілу середніх дорівнює середньому початкової (генеральної) сукупності.

Теорема 5.2. *Середнє ряду розподілу, утвореного з середніх значень великої кількості випадкових вибірок, взятих з однієї нескінченної сукупності, дорівнює середньому нескінченної* (*генеральної*) *сукупності*.

Таким чином, $\mu_{рс}=\mu_{гс}=\mu$, де $\mu_{рс}-$ середнє ряду середніх; $\mu_{гс}=\mu-$ середнє початкової (генеральної) сукупності.

Зв'язок стандартного відхилення середніх із стандартним відхиленням генеральної сукупності відображає наступна теорема.

Теорема 5.3. *Ряд розподілу, утвореного з середніх значень великої кількості випадкових вибірок потужністю N, взятих з однієї нескінченної сукупності, має стандартне відхилення* (*стандартне відхилення середніх або стандартна похибка середнього*), *яке дорівнює стандартному відхиленню генеральної сукупності, діленому на N, тобто*

$$\sigma_{рс}=\frac{\sigma_{гс}}{\sqrt{N}}=\frac{\sigma}{\sqrt{N}} \qquad (5.14.1)$$

де $\sigma_{гс}=\sigma-$ стандартне відхилення генеральної сукупності, з якої взято

велике число вибірок потужністю N.

Розглянемо, наприклад, коефіцієнти розумового розвитку (КРР) чоловіків, які проживають в місті Києві, середнє цієї сукупності $\mu = 100$, а стандартне відхилення $\sigma = 15$.

Припустимо, що з цієї генеральної сукупності взято велике (нескінченне) число вибірок по 10 елементів в кожній. Сформуємо новий розподіл із середніх цих вибірок і знайдемо стандартне відхилення вибіркового розподілу середніх:

$$\sigma_{pc} = \frac{\sigma}{\sqrt{N}} = \frac{15}{\sqrt{10}} = 4{,}7 .$$

Графік розподілу середніх є більш компактним, оскільки стандартне відхилення середніх $\sigma_{pc} = 4{,}7$, а стандартне відхилення генеральної сукупності $\sigma = 15$.

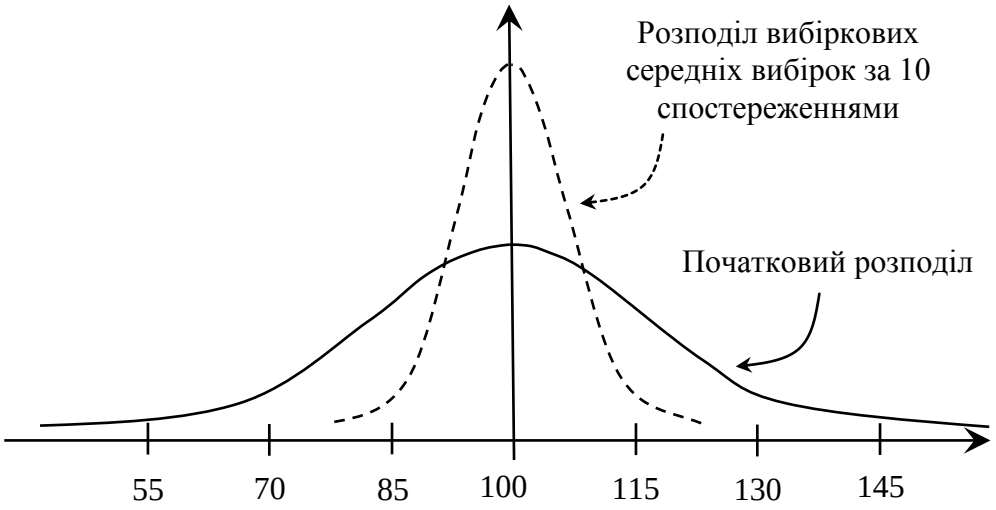


Рис. 5.2. Криві розподілу генеральної сукупності і розподілу середніх

Іншими словами, середні значення вибірок сильніше групуються навколо точки рівноваги, ніж індивідуальні коефіцієнти розумового розвитку. Таким чином, середні значення вибірок в меншій мірі відрізняються одне від одного, ніж індивідуальні КРР.

Це явище можна пояснити тим, що завдяки усередненню відбувається процес “балансування” вибірки, в результаті якого середнє вибірки приймає значення, яке наближається до 100, тобто, до середнього генеральної сукупності. При цьому, чим більшою буде потужність вибірки, тим ближчими одне до одного будуть значення середніх, тобто, зменшується стандартне відхилення розподілу середніх σ_{pc}.

На рис. 5.3 показано три ряди розподілу. Перший ряд – це ряд розподілу індивідуальних КРР із середнім 100 і стандартним відхиленням 15. Другий ряд представляє собою вибірковий розподіл середніх значень вибірок при $N = 10$, взятих з генеральної сукупності КРР. Цей розподіл має $\mu_{pc} = 100$ і $\sigma_{pc} = 4{,}7$.

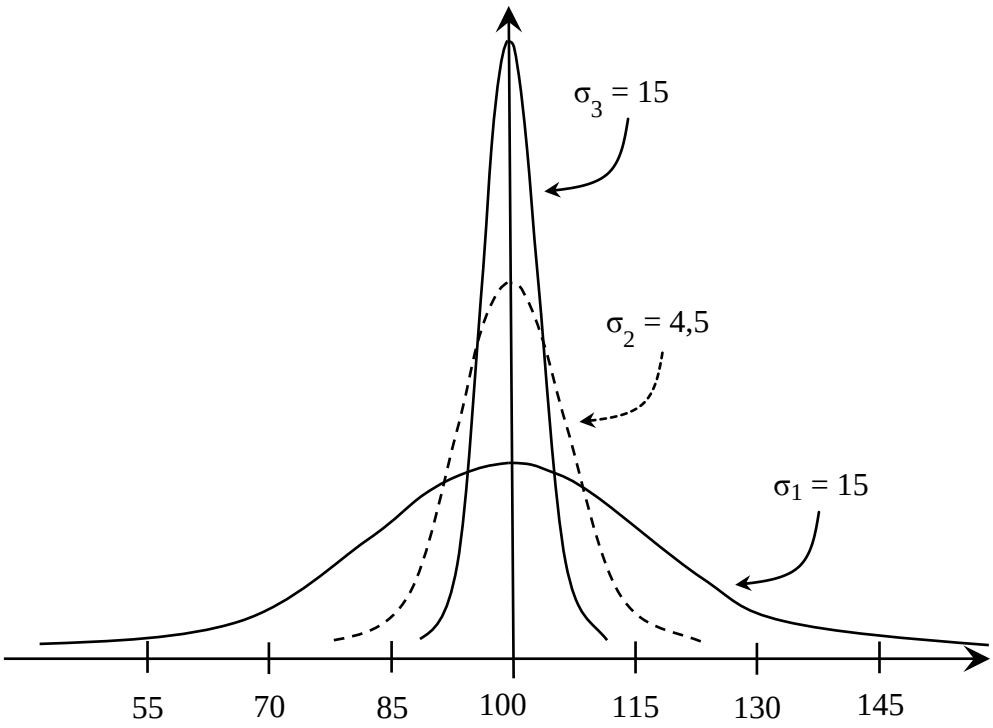


Рис. 5.3. Криві розподілу при різних стандартних відхиленнях

Природно, що цей розподіл є компактнішим порівняно з розподілом індивідуальних КРР. Крива *B* ілюструє розподіл середніх значень вибірок потужністю $N = 100$, кожна. Стандартне відхилення складає $\sigma_{pc} = 1{,}5$. Цей розподіл самий "гострий", тобто, середні значення вибірок потужністю $N = 100$, кожна, є дуже близькими одне до одного.

5.15 Процентильний ранг і $z-$оцінки вибіркового розподілу середніх

Як розраховуються $z-$оцінки і процентильний ранг окремого елемента нормального розподілу вже відомо. Розглянемо тепер, як визначити $z-$оцінку і процентильний ранг окремого значення нормального розподілу середніх. Будемо вважати, що потужності вибірок є достатньо великими для того, щоб розподіл середніх був нормальним.

Також припустимо, що відомі середнє і стандартне відхилення (СВ) для генеральної сукупності, з якої сформовані вибірки. Доступною є тільки одна випадкова вибірка з генеральної сукупності, для якої можна розрахувати середнє.

Необхідно розрахувати $z-$оцінку і процентильний ранг значення середнього для доступної вибірки з вибіркового розподілу середніх, не розглядаючи інших вибірок.

Відомо, що вибірковий розподіл середніх є нормальним. На першому кроці можна розрахувати середнє і стандартне відхилення вибіркового

розподілу середніх, а потім можна перейти до розрахунку $z-$оцінки і процентильного рангу.

Повернемось до прикладу з розподілом КРР, для якого $\mu=100$, $\sigma=15$. Нехай з цієї сукупності отримано нескінченне число випадкових вибірок довжиною $N=25$. Із середніх цих вибірок сформовано деякий розподіл. Але ми маємо інформацію тільки щодо одного розподілу і визначили, що її середнє $\mu_i=106$. Необхідно знайти $z-$оцінку і процентильний ранг вибіркового середнього $\mu_i=106$ у вибірковому розподілі середніх.

Знайдемо стандартне відхилення вибіркового розподілу:

$$\sigma_i=\frac{\sigma}{\sqrt{N}}=\frac{15}{\sqrt{25}}=3.$$

Перший крок в знаходженні $z-$оцінки деякого вибіркового середнього полягає в знаходженні відстані від цього вибіркового середнього до точки рівноваги. Оскільки вибіркове середнє $\mu_i=106$, то відхилення від рівноваги складе: $\mu_i-\mu=106-100=6$, а $\sigma_i=3$. Таким чином, $z-$оцінка вибіркового середнього буде така:

$$z_i=\frac{\mu_i-\mu}{\sigma_i}=\frac{106-100}{3}=2.$$

За допомогою $z-$оцінки можна визначити процентильний ранг вибіркового середнього, яке дорівнює 106. З таблиці площ нормального розподілу знаходимо, що елемент із $z-$оцінкою $z_i=2$, має процентильний ранг $Пр(z_i=2)\approx 98$. Тобто, вибіркове середнє, $\mu_i=106$, перевищує приблизно 98% значень вибіркових середніх.

Наведемо ще одну ілюстрацію процедури визначення процентильного рангу. Нехай середнє і стандартне відхилення КРР для генеральної сукупності чоловіків призовного віку складають: $\mu=100$, $\sigma=15$. Після чергового призову випадково сформовано множину підрозділів по 400 солдат в кожному.

Середній КРР для деякого підрозділу B склав $\sigma_i=99{,}25$. Необхідно знайти місце, яке займає за цим показником підрозділ B серед інших підрозділів. Іншими словами, який відсоток всіх інших підрозділів має середній КРР нижчий 99,25.

Стандартне відхилення середніх:

$$\sigma_{pc}=\frac{\sigma}{\sqrt{N}}=\frac{3}{4}=0{,}75,$$

а його $z-$оцінка: $z_i=\frac{\mu_i-\mu}{\sigma}=\frac{99{,}25-100}{0{,}75}=\frac{-0{,}75}{0{,}75}=-1.$

З таблиці площ нормального розподілу знаходимо, що елемент розподілу із $z_i=-1$ перевищує приблизно 16% елементів цього розподілу. Таким чином,

середнє підрозділу *B* має процентильний ранг 16. Фактично, ми виконали порівняння характеристик деякої випадкової вибірки з іншими вибірками за допомогою $z-$оцінки її середнього з наступним перетворенням середнього в процентильний ранг.

5.16 Ймовірність і нормальний розподіл

Знання характеристик нормального розподілу дозволяє зв'язати ймовірнісні судження з даним видом розподілу. Розглянемо, наприклад, ймовірність того, що елемент нормального розподілу буде мати $z-$оцінку більше нуля.

В нормальному розподілі ймовірність такого результату дорівнює частині елементів генеральної сукупності, які мають $z>0$. Таким чином, ймовірність того, що вибраний елемент розподілу буде мати $z>0$, буде дорівнювати 0,5, тобто, $p(z_i>0)=0{,}5$.

Яка буде ймовірність того, що $z-$оцінка вибраного *i*-го елемента буде знаходитись в інтервалі між –1 і +1, тобто, $p(-1<z_i<+1)=?$. Для цього необхідно знайти частину елементів нормального розподілу, які знаходяться в інтервалі між –1 і +1. З таблиці площ нормального розподілу знаходимо, що в цьому інтервалі знаходяться біля 68% всіх елементів. Тобто, ймовірність $p(-1<z_i<+1)=0{,}68$.

Відповідно, ймовірність того, що випадково вибраний елемент нормального розподілу не потрапляє у вказаний інтервал, складає 0,32.

Раніше було встановлено, що біля 95% елементів нормального розподілу знаходяться в інтервалі між –2 і +2. Таким чином, ймовірність того, що $z-$оцінка випадково вибраного елемента знаходиться в інтервалі між значеннями –2 і +2 складає 0,95.

Біля 2,5% елементів розподілу мають $z-$оцінки, які перевищують +2 (вони знаходяться в правому хвості графіка розподілу). Так само, біля 2,5% елементів розподілу мають $z-$оцінки, які є меншими ніж –2 (вони знаходяться в лівому хвості графіка розподілу).

Розглянутий підхід до статистичного аналізу даних можна використати при розв'язування прикладних задач. Припустимо, що зріст чоловіків, які проживають у великому місті, має нормальний розподіл із середнім 175 *см* і стандартним відхиленням 10 *см*. Вибирається випадково чоловік із даної генеральної сукупності. Яка ймовірність того, що його зріст має значення в інтервалі: $175\div185\,см$?

Трансформуємо інтервал $175\div185\,см$ в $z-$оцінки. Значенню 175 см відповідає $z=0$. Значення 185 *см* має $z=1$. Таким чином, необхідно визначити ймовірність попадання $z-$оцінки випадково вибраного елемента розподілу в інтервал $z=0\div1$.

В таблиці площі нормального розподілу знаходимо, що ймовірність

такого випадку складає 34/100 або 0,34, тобто,

$$p(0 < z_i < 1) = 0,34.$$

Можна визначити також ймовірність того, що випадково вибраний чоловік матиме зріст більший 185 *см*. Іншими словами необхідно визначити ймовірність того, що його $z-$оцінка буде перевищувати 1. Оскільки ми встановили ймовірність попадання $z-$оцінки в інтервал $z = 0 \div 1$, то шукана ймовірність буде дорівнювати:

$$p(z_i > 1) = 0,5 - 0,34 = 0,16,$$

або ж значення 0,16 можна знайти безпосередньо з таблиці.

Останнє завдання полягає у визначенні ймовірності того, що зріст випадково вибраного чоловіка буде меншим 155 *см* чи більшим 195 *см*. Після трансформування значень в $z-$оцінки, задача полягає в наступному: визначити ймовірність того, що $z-$оцінка випадково вибраного елемента буде меншою –2 або більшою +2. Ймовірність такої ситуації складає приблизно 5/100 або 0,05.

Задача визначення ймовірності середнього

При виконанні експериментів вибірка складається, як правило, з великого числа елементів, вибраних випадково з деякої сукупності. В процесі аналізу даних знаходять середнє цієї сукупності. Як і раніше, будемо вважати, що при виконанні експерименту був визначений деякий інтервал, який є цікавим для дослідження. Задача полягає в тому, щоб визначити ймовірність попадання вибіркового середнього у визначений інтервал.

Розглянемо *дані з попереднього прикладу*. Середнє розподілу значень зросту чоловіків складає 175 *см*, а стандартне відхилення – 10 *см*. Нехай випадкова вибірка містить 100 спостережень і необхідно знайти середній зріст чоловіків, які потрапили в цю вибірку. Яка ймовірність того, що середній зріст чоловіків з цієї вибірки знаходиться між $171\frac{2}{3}см$ і $178\frac{1}{3}см$?

Таким чином, необхідно визначити ймовірність того, що середнє вибірки із 100 спостережень потрапить в зазначений інтервал. Інакше кажучи, якщо сформувати велике число однакових за об'ємом вибірок, то яка частина цих вибірок буде мати середнє, яке знаходиться всередині вказаного інтервалу? Для того, щоб відповісти на це запитання, уявимо, що із деякої сукупності формується нескінченне число вибірок по 100 значень. Розподіл середніх цих вибірок буде нормальним, а його стандартне відхилення:

$$\sigma_{pc} = \frac{\sigma}{\sqrt{N}} = \frac{10}{\sqrt{100}} = 1см.$$

Необхідно визначити ймовірність того, що середнє деякої випадкової вибірки знаходиться в межах $171\frac{2}{3} \div 178\frac{1}{3}$. В розподілі середніх значенню

$171\frac{2}{3}см$ відповідає $z(171\frac{2}{3}) = -3{,}3$, а значенню $178\frac{1}{3}см$ відповідає $z(178\frac{1}{3}) = 3{,}3$.

З таблиці площ нормального розподілу знаходимо, що випадкова вибірка попаде у визначений інтервал з імовірністю 0,999, тобто, майже з одиничною ймовірністю.

Зазначимо, що вибіркове середнє, яке дорівнює $176\,см$, має $z(176) = 1{,}0$. З таблиці знаходимо, що наближено 34% вибіркових середніх даного розподілу знаходяться між значеннями 175 *см* і 176 *см*. Таким чином, ймовірність того, що середній зріст випадкової вибірки чоловіків розташований між 175 *см* і 176 *см*, складає наближено 0,34.

Тепер розглянемо *приклад, який стосується задачі середньої тривалості демонстрації фільмів*. Припустимо, що середня тривалість демонстрації фільмів, випущених за останнє десятиліття, складає 100 хвилин, а стандартне відхилення $\sigma = 24$ хвилини.

Якщо вибрати випадково 36 різних фільмів, то якою буде ймовірність того, що середня тривалість фільму з цієї вибірки буде меншою 92 хвилин або більшою 108 хвилини?

Припустимо, що сформовано множину вибірок об'ємом по 36 фільмів кожна і визначено середню тривалість фільму для кожної вибірки. Із середніх значень сформовано ряд розподілу. Отриманий розподіл середніх є нормальним із середнім $\mu_{pc} = 100$ і стандартним відхиленням

$$\sigma_{pc} = \frac{\sigma}{\sqrt{N}} = \frac{24}{\sqrt{36}} = 4.$$

Цей розподіл представлено на рис. 5.4.

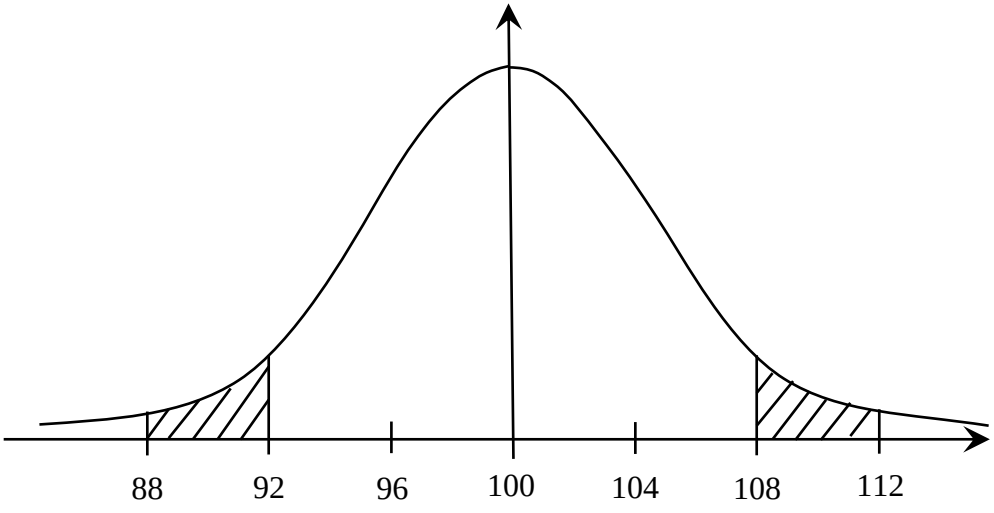


Рис. 5.4. Крива розподілу для задачі з фільмами

Значенню 92 відповідає $z(92) = -2$, а значення 108 має $z(108) = +2$. Необхідно визначити ймовірність того, що взяте випадково вибіркове середнє буде меншим 92 або більшим 108. Іншими словами, необхідно визначити ймовірність того, що взяте випадково вибіркове середнє буде мати $z_i < -2$ або $z_i > +2$. Така ймовірність складає наближено 5/100 або 0,05. Тобто, тільки в 5 вибірках із 100 середнє буде відрізнятись від середнього генеральної сукупності більше ніж на 8 хвилин.

Розглянута задача може бути сформульована, також, наступним чином: яка ймовірність того, що середня тривалість фільму для деякої вибірки потужністю 36 елементів буде відрізнятись від середньої тривалості демонстрації всіх випущених фільмів (100 хвилин) щонайменше на 8 хвилин? Очевидно, що відповідь буде такою ж, як і в попередньому формулюванні задачі.

5.17 Квантилі нормованого розподілу

Квантилі – це ще одна назва z –оцінок випадкових змінних. На практиці квантилі використовують у вигляді зміщених z –оцінок, які розглядаються нижче. Квантилі z_p нормованого нормального розподілу визначають з рівняння:

$$\Phi(z_p) = p, \qquad 0 \le p \le 1. \tag{5.17.1}$$

Значення z_p можна знайти з графіка функції розподілу як абсцису, яка відповідає координаті p. Квантиль z_p зростає від $-\infty$ до $+\infty$ при збільшенні p від нуля до 1; при $p = 0{,}5$ значення $z_p = 0$.

Та обставина, що z_p приймає як додатні, так і від'ємні значення, часто призводить до труднощів з обчисленнями. Цих труднощів можна уникнути, якщо вибрати іншу початкову точку відліку значень квантилів. Оскільки z надзвичайно рідко приймає значення менші –5, то можна встановити початкову точку відліку $\alpha = 5$. Таким чином, на практиці користуються зміщеними квантилями $z_p + 5$, тобто:

$$\textit{Зміщений квантиль} = \textit{квантиль} + 5. \tag{5.17.2}$$

На рис. 5.4 показано, яким чином зв'язані значення z_p і $z_p + 5$ з ймовірністю p (функція $z_p + 5$ також табульована). Дослідження асимптотичної поведінки кривої при $p \to 0$ і при $p \to 1$ показують, що зміні p на 0,01 відповідає Δz_p, яке є значно більшим при $p \to 0$ або $p \to 1$, ніж при значеннях $p \to 0{,}5$. Наприклад, з таблиці для $z_p + 5$ знаходимо, що

$$(z_{0{,}995} + 5) - (z_{0{,}985} + 5) = 7{,}576 - 7{,}170 = 0{,}406$$

і

$$(z_{0{,}505} + 5) - (z_{0{,}495} + 5) = 5{,}013 - 4{,}987 = 0{,}026.$$

Із симетрії нормального розподілу випливає, що

$$z_p + z_{1-p} = 0, \tag{5.17.3}$$

або

$$(z_p + 5) + (z_{1-p} + 5) = 10, \tag{5.17.4}$$

Квантилі випадкової величини із середнім μ і стандартним відхиленням σ можна знайти з рівняння:

$$p\{x_p\} = p.$$

Оскільки

$$p\{x_p\} = \Phi\left(\frac{x_p - \mu}{\sigma}\right), \tag{5.17.5}$$

то

$$z_p = \frac{x_p - \mu}{\sigma} \tag{5.17.6}$$

або

$$x_p = \mu + z_p\sigma. \tag{5.17.7}$$

Таким чином, випадкова величина x є лінійною функцією відповідного значення z в рівнянні $p\{x\} = \Phi(z)$. Якщо в прямолінійній системі координат значення x відкласти по осі абсцис, а значення z – по осі ординат, то отримаємо пряму лінію з кутовим коефіцієнтом $1/\sigma$, яка проходить через точку $(\mu, 0)$.

Розглянемо m точок деякого розподілу: $(x_1, p_1),...,(x_m,...,p_m)$. При цьому форма відповідної статистичної моделі невідома. Чи можна ці m точок вважати такими, що належать нормальному розподілу? Для розв'язку цієї задачі необхідно знайти з таблиці квантилі z_p або зміщені квантилі $z_p + 5$ і нанесемо m значень (x, z_p) на міліметровий або інший папір. Якщо ці m точок лежать на прямій лінії, то можна вважати, що точки $(x_1, p_1),...,(x_m,...,p_m)$ належать нормальному розподілу.

5.18 Контрольні запитання і вправи

1. Який розподіл називають теоретичним?
2. Що означає твердження, що розподіл є нормальним?
3. Яку функцію називають (кумулятивною) функцією розподілу?
4. Які властивості має функція розподілу?
5. Як називають похідну кумулятивної функції розподілу? Запишіть цю похідну для нормального розподілу.
6. Запишіть і поясніть формулу Лейбніца?
7. Дайте означення математичного сподівання неперервної випадкової величини і запишіть вираз для його обчислення?

8. Дайте означення дисперсії неперервної випадкової величини і запишіть вираз для його обчислення?
9. Запишіть і поясніть статистичну модель (щільність) нормального розподілу?
10. Запишіть і поясніть застосування інтеграла Пуасона?
11. Запишіть і поясніть вираз для обчислення нормованої щільності нормального розподілу? У чому полягає перевага його використання?
12. Як обчислити ймовірність нормованої нормальної величини X в інтервал $(0, x)$?
13. Зробіть дослідження кривої нормального розподілу на екстремум?
14. Як впливають зміни значення математичного сподівання випадкових величин, що утворюють розподіл, на форму нормального розподілу?
15. Як впливають зміни значення стандартного відхилення випадкових величин, що утворюють розподіл, на форму нормального розподілу?
16. За яким виразом обчислюється ймовірність попадання нормованої випадкової величини в заданий інтервал? Наведіть приклад обчислення цієї ймовірності.
17. Як обчислити ймовірність отримання випадковою величиною заданого значення відхилення?
18. Сформулюйте правило трьох сигм? Яка ймовірність того, що абсолютна величина відхилення випадкової величини перевищить три стандартних відхилення?
19. Який розподіл називають емпіричним? Поясніть на прикладі.
20. Який розподіл називають теоретичним і чому?
21. Поясніть такі статистичні характеристики нормального розподілу як асиметрія і ексцес?
22. Запишіть і поясніть вираз для статистики Жак-Бера? Які значення може приймати ця статистика і як їх інтерпретують?
23. Скільки спостережень знаходиться в інтервалі значень $z-$оцінок: $-1{,}96 \le z \le 1{,}96$?
24. Як визначити $z-$оцінку за допомогою процентильних рангів:
25. Поясніть на прикладі можливість практичного використання площ нормального розподілу?
26. Сформулюйте центральну граничну теорему Ляпунова?
27. Сформулюйте центральну граничну теорему Ляпунова для вибіркових середніх?
28. Чому дорівнює середнє ряду розподілу, утвореного із середніх значень великої кількості випадкових вибірок, взятих з однієї нескінченної сукупності?
29. Як обчислити стандартне відхилення ряду, утвореного із середніх значень великої кількості випадкових вибірок потужності N , взятих з однієї нескінченної сукупності?

30. Як розрахувати z – оцінки і процентильний ранг значення середнього для доступної вибірки з вибіркового розподілу середніх, не розглядаючи інші вибірки?
31. Як обчислити ймовірність того, що елемент нормального розподілу буде мати z – оцінку більшу нуля?
32. Як визначаються квантилі нормального розподілу?

Розділ 6

АНАЛІЗ ПРОЦЕСУ ПРИЙНЯТТЯ РІШЕНЬ, РИЗИК І ПЕРЕВІРКА ГІПОТЕЗ

6.1 Роль теорії ймовірностей і математичної статистики в прийнятті рішень

Прийняття рішень та виконання дій у відповідності до них – основний напрямок нашої цілеспрямованої (розумної) діяльності. Причини, які спонукають прийняти те чи інше рішення, можуть бути навіть не завжди цілком зрозумілими і наші дії можуть здаватись рефлекторними. Це справедливо, наприклад, у випадку, коли водій чує позаду шум потужного автомобіля, який наздоганяє його, і повертає вправо, ближче до узбіччя. Подібна ситуація має місце у випадку, коли ми вибираємо вранці одяг або думаємо, де можна пообідати. Однак, в більш серйозних ситуаціях, виникає необхідність докладного вивчення проблеми, аналізу можливих варіантів розв'язку задачі та наслідків вибраних варіантів.

Важливу роль в прийнятті багатьох рішень відіграють *теорія ймовірностей* і *математична статистика*, які пропонують обґрунтовані процедури аналізу доступних даних та прийняття рішень на основі виконаного аналізу.

При прийнятті рішень необхідно приймати до уваги два таких моменти:

- необхідно визначити ймовірність того, що висновки, на основі яких ми збираємось виконати інші дії, є правильними;
- необхідно визначити можливі варіанти розв'язання задачі і зважити можливі наслідки наших дій.

Статистика дозволяє визначити ймовірність того, що наші висновки виявились правильними. Однак, за її допомогою, як правило, не можна встановити ціну допущеної помилки, якщо прийняте рішення виявиться невірним.

Якщо існує ряд альтернатив щодо розв'язання задачі, то часто можна припустити, що помилка при прийнятті деякого рішення буде мати приблизно такі ж наслідки, як і при прийнятті будь-якого іншого рішення. У такому випадку розумніше всього зупинитись на рішенні, яке має найбільшу ймовірність того, що воно виявиться правильним.

Іноді помилка у прийнятті деякого рішення може обійтись нам набагато дорожче, ніж наслідки помилки, допущеної при прийнятті інших рішень. В такому випадку дії, які потенційно можуть обійтись нам дуже

дорого, повинні реалізовуватись тільки у тому випадку, якщо ймовірність того, що вони виявляться єдино правильними, є досить високою.

Звідси можна зробити висновок, що одне і те ж значення ймовірності дозволяє виконати деякі визначені дії в одній ситуації і не дає достатнього підгрунтя для їх виконання в іншій. Наприклад, гравець на біржі, який має солідний пакет акцій, може дозволити собі купити цінні папери, знаючи, що ймовірність підвищення їх ціни найближчим часом складає 0,7. Однак, уявімо собі, що деякий індивідуум обвинувачується в убивстві і свідоцтва проти нього такі, що в аналогічній ситуації 7 з 10 звинувачуваних виявлялись дійсно винними. Таким чином, ймовірність того, що даний звинувачений є також винним, складає 0,7. Однак присяжні, в подібних ситуаціях, проголосують, скоріше всього, за виправдання.

Причиною прийняття такого рішення є характер наслідків, до яких можуть призвести дії, які є можливими в даній ситуації. Помилка, яка є наслідком винесення несправедливого вироку, буде коштувати життя невинній людині, що значно перевищує негативні наслідки помилки, припущеної у випадку виправдання злочинця. Враховуючи подібну нерівноцінність наслідків, вирок буде, скоріше всього, буде виправдовувальним, хоча ймовірність справедливого звинувачення підсудного складає 0,7.

Наступний простий ілюстративний приклад дозволить нам ще більше зрозуміти важливість різних ставок у процесі прийняття рішень. Нехай, десятирічний хлопчик попав у чуже місто і для того, щоб доїхати додому йому потрібно 10 гривень, але він має тільки 9. Він занадто гордий, щоб просити гривню у незнайомих людей, а тому сердитий та сумний прямує до передмістя. Хлопчикові завжди подобалось грати на невеликі суми грошей за допомогою монетки. Коли він зустрів по дорозі ровесника, то сказав йому: "*Я підкину монету, а ти вгадай, що випаде – герб чи лице. Хто виграє, той отримає гривню*." Однак, зустрічний ровесник відмовився грати, так само, як і всі інші.

Хлопчик вже зібрався шукати місце для ночівлі, але тут йому в голову прийшов сміливий план. Останній із ровесників, які зустрілись йому, здавалось, трохи вагався щодо пропозиції зіграти і хлопчик звернувся до нього ще раз.

На цей раз його пропозиція була такою: "*Якщо ти виграєш, то отримаєш 9 гривень, а якщо програєш, то віддаси мені тільки одну гривню*." Пропозиція була прийнята – 9 гривень хлопчика і одна гривня його ровесника були покладені на камінь. Монетку підкинуто вверх і ровесник крикнув: "Герб!". Але монетка впала лицевою стороною вверх і хлопчик швидко забрав всі 10 гривень і побіг до автобуса.

Рішення хлопчика ризикнути своїми 9 гривнями було цілком вірним. У звичайних умовах воно було б авантюристичним, але життя наповнило його реальністю. Ймовірність того, що хлопчик виграє парі, складала 0,5, але з його точки зору цього було цілком достатньо. Втрата 9 гривень ніяк не погіршувала ситуацію, в якій він опинився, але виграш гривні дозволив йому поїхати

додому. Звичайно, що умови гри цілком влаштовували і його партнера, оскільки у випадку виграшу він отримував 9 гривень, ризикуючи тільки однією.

У сфері бізнесу також часто виникають ситуації, коли ризикові дії є вигідними для обох партнерів, оскільки ставки, якими вони ризикують, мають різне значення. Аналіз проблем, що стосуються подібних аспектів, приводить у світ азартних ігор та фінансових операцій. *Наша мета у даному випадку полягає у тому, щоб підкреслити значення фактора величини та важливості ставки у конкретній ситуації*. Вплив цих факторів завжди великий, навіть у галузі науки.

В експериментальній роботі необхідність врахування фактора ризику відчувається постійно. Так, великі ризики пов'язані з космічними польотами, але необхідність дослідження космосу і тяжіння людини до нових знань примушують ризикувати. Наприклад, можливі тяжкі наслідки лікування хворого деякими ліками, але, в той же час, вони для нього необхідні. Ризик полягає у тому, що лікар може припуститись помилки з тяжкими наслідками. З іншого боку, ризик виникає внаслідок того, що ми не використовуємо можливість вилікувати тяжку хворобу наявними ліками.

Загалом, у процесі експериментування не можна виключити повністю появу помилок. Фактично, чим більше експериментатор зменшує ймовірність появи однієї помилки, тим більше підвищується ймовірність появи іншої. Необхідно докладно вивчати можливі наслідки всіх можливих рішень (наслідки помилок) і тільки після цього остаточно вибирати конкретні дії.

6.2 Основні принципи перевірки гіпотез

Вірогідність тієї чи іншої гіпотези необхідно оцінювати в будь-якій галузі науки. Гіпотези щодо вірогідності тих чи інших припущень підтверджують, як правило, експериментально.

При прийнятті рішень завжди розглядають дві можливості. Припустимо, що наше рішення полягає у тому, що ми відхиляємо гіпотезу.

1. Перша можливість полягає у тому, що наше рішення виявилось правильним. Це означає, що в дійсності висунута (сформульована) гіпотеза невірна і повинна бути відхилена.

2. Друга можливість полягає у тому, що наше рішення відхилити гіпотезу виявилось помилковим. В дійсності висунута гіпотеза виявилась вірною, але результат, отриманий на основі деякої експериментальної вибірки, ввів особу, що приймає рішення (ОПР), в оману. Таку помилку називають *помилкою першого роду*.

Означення 6.1. *Якщо особа, що приймає рішення відхиляє вірну гіпотезу, то вона робить помилку першого роду*.

Тепер припустимо, що рішення полягає у прийнятті гіпотези.

1. Перша можливість полягає у тому, що наше рішення виявилось правильним. Це означає, що в дійсності висунута гіпотеза правильна і вона повинна бути прийнята.

2. Друга можливість полягає у тому, що наше рішення прийняти гіпотезу виявилось помилковим. Гіпотеза невірна і особа, що приймала рішення прийняла її помилково. Таку помилку називають *помилкою другого роду*.

Означення 6.2. *Якщо особа, що приймає рішення помилково приймає невірну гіпотезу, то вона робить помилку другого роду*.

Розглянемо приклад з підкиданням монети. У кожному досліді ми очікуємо один з двох можливих результатів, тобто маємо справу з *дихотомічною змінною*. Якщо монета вважається повноцінною, то ймовірність випадання орла і лицьового боку однакова: $p(орла) = p(решки) = 0{,}5$. Вид розподілу, який утворюється в результаті ряду послідовних експериментів з монетою, можна визначити за допомогою такої теореми.

Теорема 6.1. *Якщо у випадку дихотомічної змінної ймовірність появи події складає* p *і при цьому формується нескінченна кількість вибірок однакової потужності* N, *то кожна вибірка* X *такої потужності асимптотично наближається до нормального розподілу з параметрами*:

$$\mu = N\,p;\ \ \sigma = \sqrt{N\,p(1-p)},\ \text{тобто},\ \ \{X\} \leftrightarrow N(Np, Np(1-p)).$$

Нехай, для прикладу з підкиданням монети $p = 0{,}5$ при $N = 100$. Теорема 6.1 стверджує, що розподіл отриманих результатів нормальний з параметрами:

$$\mu = Np = 100 \cdot 0{,}5 = 50;\quad \sigma = \sqrt{Np(1-p)} = \sqrt{100 \cdot 0{,}5 \cdot (1-0{,}5)} = 5.$$

Припустимо, що після 100 дослідів з підкиданням монети отримано такі результати:

орли: 55 разів; *лицьовий бік*: 45 разів.

Після цього ставиться запитання: „*Чи можна на основі цієї інформації прийняти або відхилити гіпотезу стосовно того, що монета повноцінна*?" Тобто, H_0 : *монета повноцінна.*

Нехай, *перше правило* прийняття рішення полягає у такому: *гіпотеза стосовно повноцінності монети вважається правильною, якщо кількість орлів, що випали в результаті експерименту, знаходиться в інтервалі* $10 \div 90$. У відповідності до цього правила гіпотеза стосовно повноцінності відхиляється тільки у тому випадку, коли кількість орлів буде меншою 10 або більшою 90. Це правило ілюструє рис. 6.1.

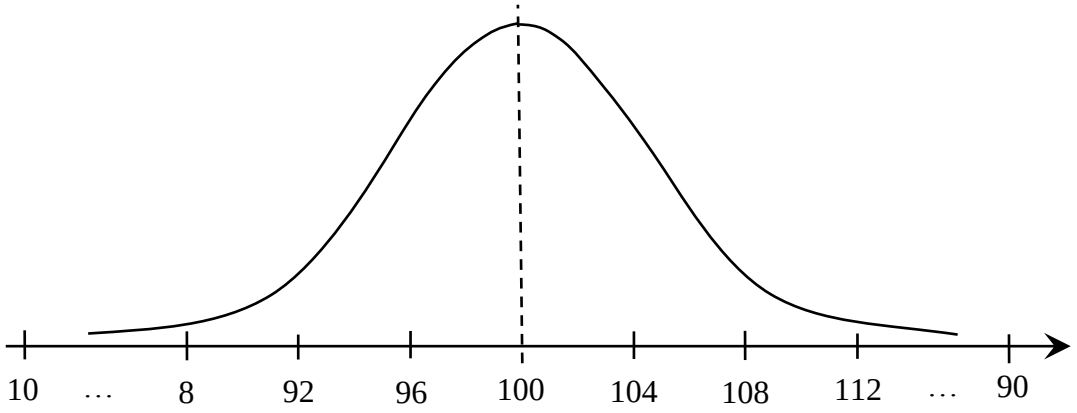


Рис. 6.1. Ілюстрація першого правила прийняття рішень

Таким чином, *якщо монета повноцінна, то результат експерименту практично завжди буде знаходитись в інтервалі* $10 \div 90$. Оскільки зона, в якій даний критерій припускає істинність висунутої гіпотези, обмежена в кожний бік величиною 8-и стандартних відхилень, то результат підкидання повноцінної монети майже ніколи не відхилиться від значення 50 настільки, щоб можна було відхилити гіпотезу у відповідності до правила 1. Іншими словами, *використання правила* 1 *навряд чи призведе до помилки першого роду*.

Припустимо, що сформульовано *друге правило*: *монета буде повноцінною, якщо кількість орлів попадає в інтервал* $45 \div 55$. Це означає, що гіпотеза стосовно повноцінності монети буде прийнята, якщо це співвідношення буде лежати в дуже вузьких границях. Рис. 6.2 ілюструє, що відбудеться, якщо при підкиданні монети рішення будуть прийматись на основі другого правила.

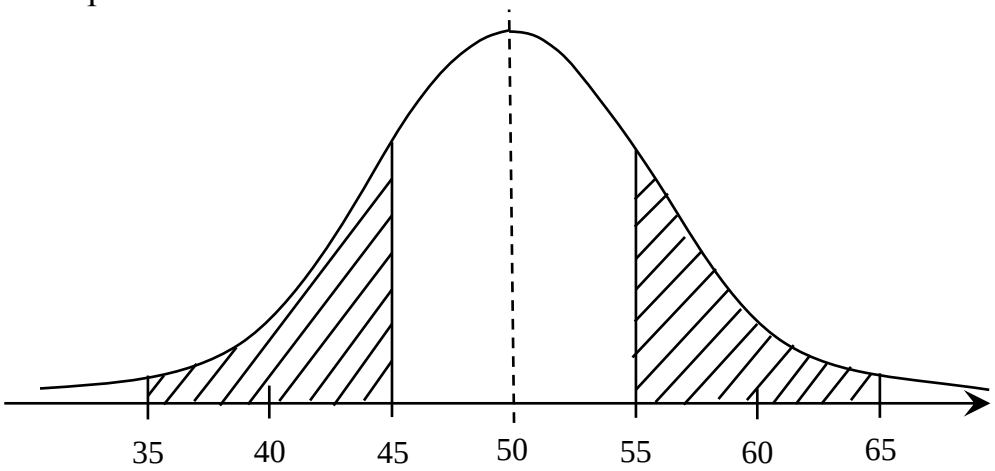


Рис. 6.2. Ілюстрація другого правила прийняття рішень

За допомогою теореми 6.1 можна розрахувати ймовірність того, що результати підкидання повноцінної монети попадуть у виділений нами інтервал. Цей інтервал обмежений значеннями 45 і 55. Кожне з цих значень відрізняється від середнього, що дорівнює 50, на одне стандартне відхилення. Використовуючи таблицю площ нормального розподілу визначимо, що на область прийняття гіпотези приходиться у даному випадку біля 68% загальної площі графіка. Відповідно, затінена область складає біля 32%. Таким чином,

ймовірність того, що при підкиданні монети зустрінеться вибірка, в якій кількість орлів попаде у виділену область, складає 0,68.

Важливо підкреслити те, що *ймовірність появи вибірки, в якій кількість орлів попаде в область, де гіпотеза стосовно повноцінності монети відхиляється, складає* 0,32. Тобто, це ймовірність того, що за правилом 2 буде помилково відхилена висунута гіпотеза, навіть якщо вона правильна – це ймовірність помилки першого роду.

Тепер можна зробити *висновок стосовно результатів застосування першого і другого правил* прийняття рішень. Згідно з правилом 1, навряд чи можна буде помилково відхилити висунуту гіпотезу або зробити помилку першого роду. Згідно з правилом 2, в 32% випадків гіпотеза щодо повноцінності монети буде відхилятись помилково і, таким чином, ймовірність припущення помилки першого роду, у випадку повноцінності монети, складає приблизно 0,32%.

Припустимо тепер, що для експерименту вибрали *неповноцінну монету*, тобто, ймовірність випадання орла відрізняється від величини 0,5. Наприклад, ймовірність випадання орла при кожному окремому підкиданні може складати 0,2 або 0,6 або 0,9 (або іншу величину).

Нехай ймовірність випадання орла складає 0,6 для $N = 100$. У відповідності з теоремою 6.1 розподіл результатів експерименту має середнє і стандартне відхилення:

$$\mu = N\,p = 100 \cdot 0{,}6 = 60;$$

$$\sigma = \sqrt{N\,p(1-p)} = \sqrt{100 \cdot 0{,}6(1-0{,}6)} = 4{,}9.$$

На рис. 6.3 показано розподіл результатів експерименту при використанні неповноцінної монети.

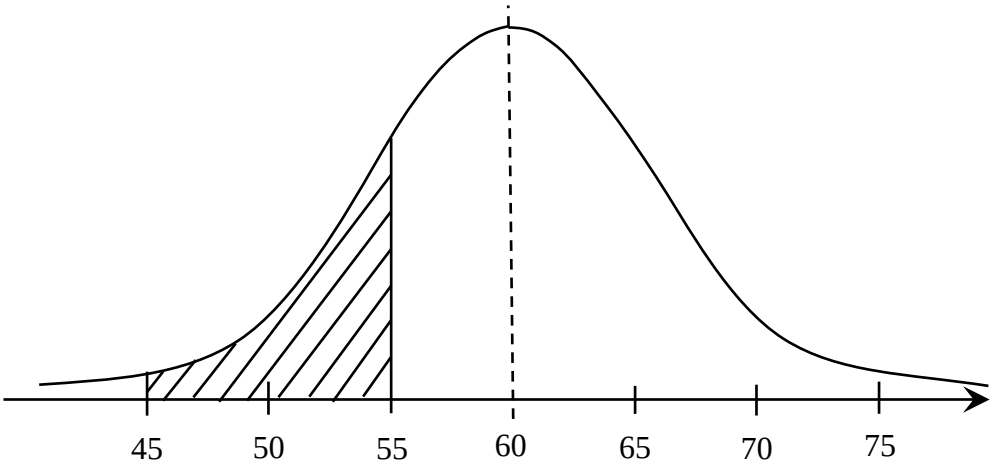


Рис. 6.3. Розподіл результатів експерименту при використанні неповноцінної монети

Розглядаючи експеримент з монетою, про неповноцінність якої відомо наперед, можна визначити наслідки застосування правил стосовно відхилення

гіпотези про повноцінність монети. Нагадаємо, що *перше правило* таке: *гіпотеза щодо повноцінності приймається, якщо кількість орлів попадає в інтервал* $10 \div 90$. З рис. 6.3 видно, що навіть у випадку використання неповноцінної монети результати практично кожного експерименту будуть попадати в область прийняття гіпотези. Такий же результат можна знайти за допомогою $z-$оцінок і таблиць площ нормального розподілу. Іншими словами, *у випадку, коли ймовірність випадання орла складає 0,6, за першим правилом практично завжди буде помилково прийматись висунута гіпотеза*. Тим самим робиться помилка другого роду.

За *другим правилом,* висунута гіпотеза приймається тільки у тому випадку, коли число орлів, що випадає, попадає в інтервал $45 \div 55$. Прийнявши помилкову гіпотезу (що монета повноцінна), за другим правилом можна припуститись помилки другого роду у випадку, якщо кількість орлів при підкиданні неповноцінної монети буде попадати в указаний інтервал. Затінена площа на рис. 6.3 відображає кількість випадків, у яких буде припущена помилка другого роду за другим правилом. За допомогою $z-$оцінок для чисел 45 і 55 і таблиці площ нормального розподілу знайдемо, що площа затіненої частини графіка складає 15% від всієї площі під графіком.

Іншими словами, *у випадку використання неповноцінної монети* із ймовірністю випадання орла, $p(орла) = 0{,}6$, ймовірність того, що за другим правилом буде зроблена помилка другого роду, складає:

$$p(помилки\ другого\ роду) = 0{,}15.$$

Звідси можна зробити висновок, що при експериментуванні з неповноцінною монетою ймовірність того, що за другим правилом буде справедливо відхилено висунуту гіпотезу, складає 0,85, тобто:

$$p(справедливого\ відхилення) = 0{,}85.$$

Цей факт має важливе значення. Використовуючи одну і ту ж монету, за першим правилом ніколи не можна буде відхилити неправильну гіпотезу, а за другим правилом ймовірність справедливого відхилення неправильної гіпотези складає 0,85. Тобто *потужність другого правила є значною*.

Узагальнення отриманих результатів

Правило 1 дозволяє відхилити деяку гіпотезу тільки в рідкісних випадках. Якщо гіпотеза правильна, то за правилом 1 вона ніколи помилково не буде відхилена, тобто, не буде припущено помилки першого роду. В той час, як за правилом 2 помилку першого роду буде зроблено у 32 випадках із 100.

Якщо ж гіпотеза виявиться неправильною і ймовірність випадання орла при підкиданні монети складає 0,6, то за правилом 1 практично в усіх випадках буде робитись помилка другого роду, тобто прийматись неправильна гіпотеза. Критерій цього правила є значно меншим, ніж критерій правила 2, яке дозволяє зменшити ймовірність припущення помилки другого роду до 0,15. Потужність цього критерію складає для описаного випадку 0,85.

Необхідно пам’ятати, що отримані числові результати справедливі для конкретного характеру неповноцінності монети. Вище було розглянуто випадок, коли ймовірність випадання орла складає 0,6. Якби розглядався інший випадок, то ми прийшли б до іншого результату, при якому ризик прийняття неправильної гіпотези мав би інше значення.

Можна очікувати, що чим більшою є неповноцінність монети, тим меншою буде ймовірність попадання результату експерименту в область прийняття гіпотези. Розглянемо крайню ситуацію: нехай ймовірність випадання орла складає 0,95. В такому випадку гіпотеза стосовно повноцінності буде відхилена за обома правилами практично завжди. Відповідний розподіл представлено на рис. 6.4.

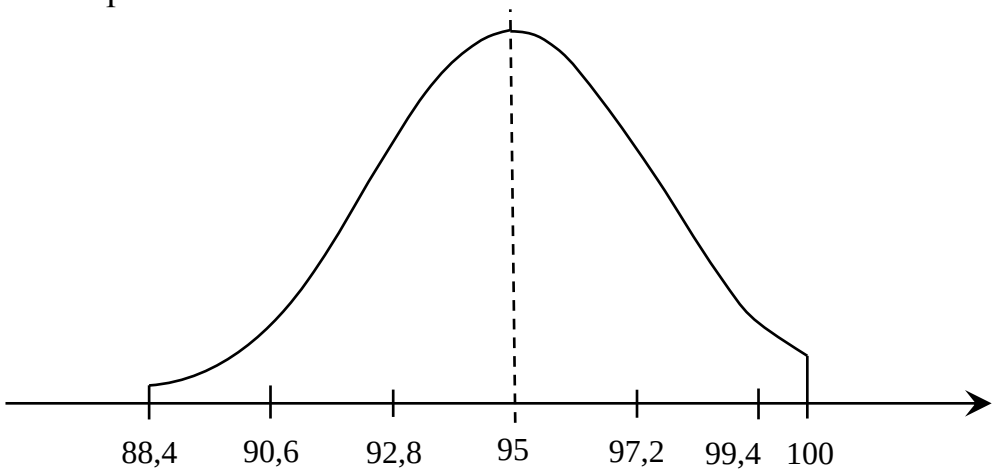


Рис. 6.4. Випадок, коли ймовірність випадання орла складає 0,95

Зазначимо, що ймовірність помилки другого роду залежить від конкретних умов виконання експерименту, тобто від ступеня неповноцінності монети. За обома правилами гіпотеза стосовно повноцінності монети буде відхилена скоріше у випадку, коли дійсність сильно відрізняється від припущень, ніж у випадку, коли дійсність незначно відрізняється від висунутої гіпотези.

Іншими словами, якщо неповноцінність монети є незначною, то ймовірність прийняття гіпотези щодо її повноцінності буде вищою за обома правилами. Тепер очевидно, що чим більше гіпотеза відділена від дійсності, тим швидше вона буде відхилена.

В реальних ситуаціях при перевірці гіпотез експериментатор потрапляє в складне становище, якщо він не знає, чи правий, коли приймає або відхиляє гіпотезу. Звичайно, він завжди намагається прийняти правильну гіпотезу і відхилити неправильну. При цьому важливо знати ступінь ризику, який виникає в результаті прийняття того чи іншого рішення. Тобто, *необхідно визначити деяку область прийняття сформульованої гіпотези і пам’ятати, що навіть у випадку правильної гіпотези може статись так, що отриманий результат не попаде у визначену область*.

Крім того, навіть якщо гіпотеза є неправильною, то є деякі шанси, що отриманий результат попаде в область прийняття гіпотези. Виникає запитання – як визначити границі області прийняття висунутих гіпотез?

Чим вужчою є ця область, тим більшим буде ризик помилки першого роду. Збільшуючи область прийняття гіпотези, ми тим самим зменшуємо ймовірність помилки першого роду і одночасно збільшуємо ризик помилки другого роду.

Збільшуючи область прийняття гіпотези, ми зменшуємо ймовірність виникнення помилки першого роду і одночасно збільшуємо ризик помилки другого роду. Очевидно, що *ми не можемо одночасно зменшувати ризик помилки одного роду без одночасного збільшення помилки іншого роду*.

Рішення щодо величини області прийняття висунутої гіпотези залежить від того, наслідки якої помилки будуть гіршими. У випадку з монетою критерії ризику відрізнялись суттєво. *Перше правило* було спрямоване на те, щоб не забракувати повноцінну монету. Велика площа його області прийняття гіпотези свідчить про те, що воно практично ніколи не допоможе виявити неповноцінну монету.

За другим правилом, область прийняття гіпотези була досить вузькою, що попереджало прийняття монети з дефектом за повноцінну. Очевидно, що за другим правилом багато хороших монет можуть бути забракованими.

Загалом, при визначенні області прийняття висунутої гіпотези кожний експериментатор керується власними критеріями відносно можливості помилкового відхилення правильної гіпотези, а також, в деякій мірі, помилкового прийняття неправильної гіпотези. Величина вибраної області фактично відображає компроміс між цими двома вимогами.

Таким чином, при прийнятті рішень, завжди існує ризик. В наступному розділі буде показано, як на практиці визначаються границі допустимого ризику при перевірці гіпотез. Вище ми розглянули принципи, які лежать в основі процедури перевірки гіпотез, а тепер необхідно перейти безпосередньо до самої процедури перевірки гіпотез.

6.3 Статистична перевірка гіпотез

Існує стандартна процедура перевірки гіпотез, якою користуються в техніці, психології, освіті, суспільних науках та багатьох інших областях. Ця процедура буде розглянута в даній главі.

Необхідно підкреслити той важливий факт, що результат експерименту, який підтверджує справедливість висунутої гіпотези, майже ніколи не може бути основою для прийняття цієї гіпотези. В той же час *результат, несумісний з висунутою гіпотезою, є цілком достатнім для її відхилення як неправильної*. Очевидно, що наведене твердження потребує обґрунтування.

Причиною того, що результат, який підтверджує висунуту гіпотезу, не обов'язково може бути основою для її прийняття, полягає в тому, що отриманий результат може бути сумісним і з іншими гіпотезами. Тобто, він не обов'язково може служити доведенням справедливості даної гіпотези проти інших сформульованих альтернатив. Наприклад, випадання 51 орла при 100 підкиданнях монети є сумісним з гіпотезою про те, що монета є повноцінною. Разом з тим, цей результат є сумісним також з припущенням, що монета має деякий дефект і при її підкиданні орел випадає дещо частіше. Таким чином, результат, сумісний з гіпотезою, висунутою на початку, не може бути 100%-м доведенням її правильності. Навіть випадання 50 орлів при підкиданні монети не дає підґрунтя для висновку, що монета не має хоча б незначного дефекту.

З іншого боку, випадання 80-и орлів при 100 підкиданнях монети могло б служити для відхилення гіпотези щодо її повноцінності. Оскільки існує лише один шанс з 10000, що подібний результат може мати місце, то цілком обґрунтовано, і з мінімальним ризиком зробити помилку, його можна використати як основу для висновку щодо неповноцінності монети.

Приклад 6.1. Припустимо, що середній коефіцієнт розумового розвитку (КРР) деякої генеральної сукупності людей $\mu = 100$. Результат випадкової вибірки дає результат $\mu_i = 102$, який є сумісним з висунутою гіпотезою. Однак цей результат є також сумісним з припущенням, що $\mu = 101$ або $\mu = 99$ і, звичайно, є сумісним з гіпотезою, що $\mu = 102$. Таким чином, на основі даного результату не можна віддати перевагу гіпотезі, що $\mu = 100$.

Припустимо тепер, що інша випадкова вибірка дала середнє $\mu = 135$. Якщо об'єм вибірки був достатньо великим, то можна показати таке: якщо початкова гіпотеза правильна, то ми практично ніколи не отримали б подібного результату. На основі цього висновку отриманий результат цілком обґрунтовано можна використати як підтвердження неправильності висунутої гіпотези і ризик помилки при цьому буде мінімальним.

Все сказане вище ґрунтується на тому факті, що результат експерименту, який є сумісним з висунутою гіпотезою, виявляється також сумісним з іншими гіпотезами. Це призводить до того, що подібний результат не може бути використано для обґрунтування вибору деякої гіпотези порівняно з іншими. Однак, ми завжди можемо отримати результат, який розбігається з висунутою гіпотезою і може спричинити значні сумніви щодо її правильності або достовірності.

Гіпотезу можна порівняти із свідченнями звинуваченого в суді. Він не може довести істинність своїх слів. Разом з тим, деякі факти, наведені ним, залишають відкритою можливість для висування припущення, що він міг діяти не так, як говорить. Тому прокурор може піддати сумніву правдивість його розповіді і надати іншу інтерпретацію наведених фактів.

Нульова або нуль-гіпотеза

Відмова відхилити гіпотезу означає, що вона може бути правильною. *Відхилення сформульованої гіпотези означає, що ми робимо висновок щодо її неправильності*. Це положення є дуже важливим.

Таким чином, при перевірці гіпотез остаточний висновок можна зробити тільки в тому випадку, якщо ми можемо відхилити висунуту гіпотезу. Таким чином, мета експерименту повинна полягати в тому, щоб відхилити сформульовану гіпотезу.

Мета експерименту повинна полягати у тому, щоб відхилити сформульовану гіпотезу.

А це означає, що початкова гіпотеза повинна формулюватись як альтернатива тому, в що ми віримо.

Якщо ми зможемо відхилити висунуту гіпотезу (довести її неправильність), то тим самим продемонструємо справедливість того твердження, в яке дійсно віримо.

Наприклад, якщо ми хочемо показати, що зріст чоловіків є більшим зросту жінок, то висуваємо гіпотезу щодо відсутності відмінностей в їх зрості. Потім пробуємо відхилити цю гіпотезу. Для того щоб довести існування відмінностей між різними партіями, необхідно перевірити гіпотезу, яка стверджує, *що між партіями немає відмінностей*. Відхиляючи цю гіпотезу, ми встановлюємо істинність вихідного припущення.

Означення 6.3. *Гіпотезу, у відповідності до якої немає відмінностей між різними сукупностями (або варіантами), називають нуль-гіпотезою*.

Тобто у нуль-гіпотезі формулюється результат (висновок), протилежний до очікуваного. Відкидання нуль-гіпотези після проведення експерименту свідчить, що ми отримали очікуваний результат.

Гіпотеза, яку ми зможемо перевірити, *не може бути сформульована на основі будь-якого судження*. Так, судження щодо того, що монета є неповноцінною, є недостатньо визначеним для того, щоб на його основі можна було сформулювати конкретну визначену гіпотезу. І подібні випадки зустрічаються досить часто. Із сказаного випливає наступне: *експериментатор повинен формулювати альтернативу тому, що він намагається довести, у вигляді чітко визначеної гіпотези*.

Експериментатор повинен формулювати альтернативу тому, що він намагається довести, у вигляді чітко визначеної гіпотези.

Тільки у тому випадку, коли це можливо, він може спробувати відхилити її з метою доведення істинності початкових припущень.

Таким чином, перший крок експериментатора повинен полягати у формулюванні статистичної гіпотези, яку він сподівається *відхилити з метою доведення істинності свого вихідного (початкового) припущення*. Після цього він може застосувати процедуру перевірки гіпотези.

Відхилення гіпотези та рівень значущості

Припустимо, що експериментатор вже сформулював гіпотезу, яку він сподівається відхилити з метою довести істинність свого вихідного

припущення. У прикладі з монетою це може відповідати тому, що він вирішив перевірити гіпотезу щодо повноцінності монети, сподіваючись її при цьому відхилити. Експериментатор приступає до формування вибірки і розрахунку на її основі числа орлів при підкиданні монети (наприклад, 100 разів) та інших параметрів.

Наступним кроком експериментатора є формулювання наступного питання:

"*Якщо моя нульова гіпотеза вірна, то яка ймовірність того, що мною буде зроблена вибірка, показник якої відрізняється від очікуваного результату так само, як і отримане мною значення*?"

Припустимо, що монета дійсно є повноцінною і орли повинні випадати при її підкиданні в середньому в половині зробленого числа дослідів. Наскільки ймовірним є те, що може випасти число орлів, яке відрізняється від очікуваного значення 50 настільки, наскільки відрізняється результат, отриманий в дійсності?

Зверніть увагу на сформульоване тут запитання – воно є ключем для розуміння подальшого матеріалу даного розділу.

На наступному кроці з'ясовується, наскільки ймовірним буде отриманий результат вибірки за умови, що висунута гіпотеза є правильною. Ймовірність отримання вибірки з характеристиками, які відповідають сформульованій гіпотезі, може бути високою або низькою і цей факт визначає рішення щодо прийняття або відхилення гіпотези.

Припустимо, розрахунки показали, що у випадку правильності сформульованої гіпотези ймовірність отримання вибірки з характеристиками, які відповідають цій гіпотезі, є високою. В такому випадку отримана *вибірка повинна розглядатись як представницька* сукупність (наприклад, при випаданні 52 орлів при підкиданні монети). Це означає, що отримане значення має надто високу ймовірність появи, щоб дозволити нам відхилити висунуту гіпотезу.

Представимо тепер, що отримане в результаті виконання досліду значення так сильно відрізняється від очікуваного (у відповідності до висунутої гіпотези), що ймовірність його появи є дуже малою. (Наприклад, такий випадок буде мати місце, якщо при 100 підкиданнях монети випаде 93 орли). Іншими словами, залишається припустити, що отримане вибіркове значення має таке велике відхилення, що (як свідчать відповідні розрахунки) практично неможливо отримати подібний випадковий результат при справедливості нульової гіпотези. У такому випадку рішення буде полягати у відхиленні висунутої нами гіпотези. Тобто, швидше ми припустимо, що висунута гіпотеза виявилась неправильною, ніж допустимо ймовірність появи надзвичайно неправдоподібного результату.

Для експериментатора велике значення має мінімізація можливостей появи ситуацій, коли випадкова збіжність ряду факторів може призвести до

відхилення правильної гіпотези. Тому, перед тим як відхилити гіпотезу, він вимагає, щоб ймовірність отримання відповідного вибіркового значення була дуже малою.

В деяких областях науки прийнято відхиляти гіпотезу тільки у тих випадках, коли випадкове вибіркове значення може зустрічатись не частіше ніж 5 разів на 100 експериментів. В інших областях гіпотези відхиляються, якщо ймовірність появи відповідного вибіркового значення не перевищує 0,01. Очевидно, що експериментатор повинен прямувати до того, щоб ймовірність появи вибіркового значення, яке вказує на неправильність висунутої гіпотези, була досить малою.

Ймовірність появи вибіркового значення, яке вказує на неправильність висунутої гіпотези, вибирається експериментатором і називається *рівнем значущості* експерименту.

Вибір рівня значущості повинен вибиратись до збору експериментальних даних, оскільки результати експерименту не повинні впливати на величину вибраного критерію. Після визначення рівня значущості і обробки даних він припускає, що його гіпотеза є правильною і визначає – більшою чи меншою вибраного рівня значущості буде ймовірність отриманого результату.

Якщо ймовірність отриманого результату перевищить рівень значущості, то експериментатор не зможе відхилити висунуту гіпотезу, справедливо вважаючи, що даний результат є в достатній мірі сумісним з нею. Наприклад, припустимо, що вибрано рівень значущості 0,05, а в результаті експерименту отримано 52 орли при 100 підкиданнях монети. Якщо експериментатор встановить, що відхилення в два орли від очікуваного значення зустрічається більше, ніж в 5% випадків, то він не зможе відхилити висунуту гіпотезу.

Якщо розраховане значення ймовірності деякого результату виявилось меншим рівня значущості, то висунуту гіпотезу можна відхилити. Наприклад, припустимо, що при використанні рівня значущості 0,05, підкидання монети призвело до 96 орлів, а розрахунок показав, що відхилення цього значення від очікуваного має ймовірність появи менше 0,05. Логіка, якою керується експериментатор, полягає у такому: "*Після формулювання даної гіпотези я отримав такий неймовірний результат, що не можу в нього повірити*." Після цього експериментатор може вважати свою початкову теорію доведеною і, таким чином, може сподіватись, що завжди будуть отримувати результати, ймовірність появи яких при даних обставинах буде меншою рівня значущості.

Вибір рівня значущості означає, що введено визначене правило прийняття і неприйняття гіпотез. *Рівень значущості показує ймовірність відхилення деякого показника від його очікуваного значення, при якому дослідник може відхилити висунуту гіпотезу*.

Вибраний *рівень значущості вказує на точне значення ймовірності помилки першого роду* у випадку, якщо гіпотеза дійсно правильна. Наприклад, якщо рівень значущості дорівнює 0,01, то це означає, що у випадку

правильності гіпотези в одному випадку із 100 буде робитись помилка першого роду на основі отриманого результату.

Іншими словами, рівень значущості означає величину ризику помилки першого роду. Чим меншим є рівень значущості, тим меншою є ймовірність припущення помилки першого роду. Однак, чим меншим є рівень значущості, тим більшою є ймовірність помилки другого роду, якщо гіпотеза виявиться помилковою. Таким чином, вибір рівня значущості означає вибір правила прийняття рішення при перевірці гіпотези.

Сформулюємо процедуру перевірки гіпотез у вигляді наступних етапів:

1. Формулювання нульової гіпотези, яку необхідно перевірити. Відхилення сформульованої гіпотези дає можливість вважати початкову теорію правильною. (В прикладі з монетою метою може бути встановлення факту неповноцінності монети. Таким чином, необхідно перевірити гіпотезу про те, що монета є повноцінною.)
2. Вибір рівня значущості. (Наприклад, рівень значущості 5% означає, що ймовірність помилки першого роду складає 0,05.)
3. Виконати експеримент і обчислити необхідний статистичний параметр.
4. Припускаючи, що сформульована гіпотеза є вірною, необхідно визначити ймовірність відхилення обчисленого значення статистичного параметра від його очікуваного значення.
5. Якщо, в припущенні істинності гіпотези, розрахунки показують, що ймовірність відхилення отриманого вибіркового статистичного показника від очікуваного значення перевищує рівень значущості, то *відхилити висунуту гіпотезу неможливо*.
6. Якщо, в припущенні щодо істинності гіпотези, розрахунки показують, що ймовірність відхилення отриманого вибіркового статистичного показника від очікуваного значення є меншим рівня значущості, то *висунута гіпотеза відхиляється*.

Нехай у експерименті з монетою випало 55 орлів. Гіпотеза щодо повноцінності монети буде перевірятись на рівні значущості 0,05. Припустимо, що монета є дійсно повноцінною, а отримана вибірка є однією з великої множини випадкових вибірок об'ємом 100 підкидань кожна.

Розподіл результатів великої множини експериментів з повноцінною монетою є нормальним. Цей розподіл представлено на рис. 6.5 ($\mu = 50, \sigma = 5$). Місце, де знаходиться оцінка 55, позначено хрестиком. Виникає наступне запитання: якщо гіпотеза правильна, то *яка ймовірність відхилення числа орлів, які випали у випадковій вибірці із 100 спостережень, від очікуваного значення на 5 одиниць*? В розподілі, представленому на рис. 6.5, значення 55 знаходиться на відстані 5 одиниць від очікуваного значення. Таким чином, z – оцінка значення 55 дорівнює: $z(55) = (55 - 50)/5 = 1$.

Тепер можна записати вираз для визначення $z-$оцінки деякого результату експерименту для вибірки дихотомічної змінної при $N=100$:

$$z=\frac{оцінка\ відхилення}{стандартне\ відхилення}=\frac{r-N\,p}{\sqrt{N\,p(1-p)}}, \qquad (6.3.1)$$

де $r-$ значення отриманого результату експерименту (тобто, скільки разів дана подія зустрічається у вибірці); $N-$ об'єм вибірки; $p-$ ймовірність того, що дана подія буде мати місце.

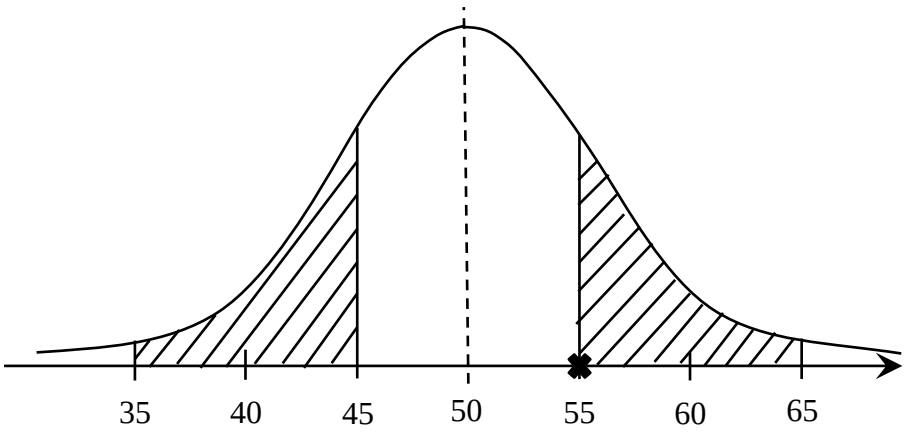


Рис. 6.5. Розподіл результатів експерименту з повноцінною монетою ($\mu=50,\ \sigma=5$)

Зазначимо, що $N\,p-$ величина гіпотетичного очікування, яке дорівнює теоретичній ймовірності появи події, що визначається висунутою гіпотезою. Очевидно, що навіть у випадку, коли гіпотеза є вірною, можна очікувати деякого відхилення отриманого результату від його теоретичного значення. Однак, вирішальну роль при прийнятті рішення відіграє ступінь цього відхилення, виражена через $z-$оцінку.

Для експерименту з повноцінною монетою $r=55,\ n=100,\ p=0{,}5$. Таким чином:

$$z=\frac{r-N\,p}{\sqrt{N\,p(1-p)}}=\frac{55-100\cdot 0{,}5}{\sqrt{100\cdot 0{,}5\cdot 0{,}5}}=1{,}0.$$

Суть цієї формули полягає у такому: відповідно до нульової гіпотези ми припускаємо, що $p=0{,}5$, і при цих умовах отриманий нами результат має $z=1{,}0$ в розподілі результатів великої множини ідентичних експериментів.

Нагадаємо запитання, яке поставив перед собою експериментатор: *якщо гіпотеза правильна, то яка ймовірність отримання результату, який відхиляється від очікуваного значення на* 5 *одиниць*? Це ж запитання, але записане за допомогою формули, звучить так: *яка ймовірність отримання випадкового результату,* $z-$*оцінка якого є на одну одиницю більшою від*

нульового значення? За допомогою таблиці площ нормального розподілу знаходимо, що приблизно 32% спостережень нормального розподілу мають z – оцінки, що відрізняються від очікуваного значення щонайменше на одну одиницю. А це означає, що всі спостереження, які належать заштрихованим областям графіка, представленого на рис. 6.5, або 32% всіх спостережень, відхиляються від очікуваного значення більше, ніж отриманий нами результат – 55 орлів.

Таким чином, результат 55 орлів означає наступне: якщо гіпотеза вірна, то ймовірність появи подібного результату або результату, який ще більше відрізняється від очікуваного значення, складає 32%. Оскільки дане значення ймовірності перевищує вибраний рівень значущості 0,05, то ми не можемо відхилити гіпотезу щодо повноцінності монети. Можна сказати, що отриманий результат не є настільки неправдоподібним, щоб дати нам підґрунтя відкинути висунуту гіпотезу.

Припустимо далі, що при підкиданні монети випало 65 орлів. На цей раз відхилення від очікуваного значення складає 15 одиниць і можна припустити, що ймовірність отримання подібного результату в експерименті з повноцінною монетою є значно меншою. З рис. 6.5 можна зробити висновок: *якщо висунута гіпотеза є вірною, то отримане на цей раз значення буде відрізнятись від очікуваного результату на три одиниці стандартного відхилення*.

За формулою (6.1) знайдемо:

$$z = \frac{r - N\,p}{\sqrt{N\,p(1-p)}} = \frac{65 - 100 \cdot 0{,}5}{\sqrt{100 \cdot 0{,}5 \cdot 0{,}5}} = 3{,}0,$$

що є формальним підтвердженням сформульованого висновку. З таблиці площ нормального розподілу випливає, що ймовірність отримання результату, z – оцінка якого так суттєво відрізняється від очікуваного значення, приблизно дорівнює 0,001.

Таким чином, отриманий нами результат експерименту є настільки рідкісним, що ймовірність його появи складає приблизно 0,001. Оскільки це значення є меншим рівня значущості, то наше рішення буде полягати, в даному випадку, у відхиленні гіпотези щодо повноцінності монети, а це означає підтвердження початкового припущення щодо її неповноцінності.

Таким чином, вище розглянуто ідею методу перевірки гіпотез, логіка (послідовність реалізації) якого полягає у такому:

- точне формулювання гіпотези, яка буде перевірятись;
- вибір рівня значущості і визначення результату (параметра), який нас цікавить, на основі вибірки;
- у припущенні істинності висунутої гіпотези визначити ймовірність того, що отриманий результат буде відрізнятись від очікуваного значення на величину відхилення, отриманого в експерименті;

– якщо ймовірність подібного відхилення буде меншою рівня значущості, то висунута гіпотеза відхиляється;

– якщо ймовірність подібного відхилення перевищить рівень значущості, то висунута гіпотеза приймається.

Логічна основа розглянутого методу залишається однаковою при розв'язуванні багатьох задач подібного типу. Гіпотеза стосовно середнього, медіани, стандартного відхилення, а також багато інших гіпотез перевіряються за допомогою ідентичної процедури. Очевидно, що розрахунки у кожному окремому випадку будуть мати свої особливості.

6.4 Перевірка гіпотез в задачах трьох типів

Розглянемо три типи задач та шляхи їх розв'язку. Перший тип задач відноситься до дихотомічних змінних (приклад розглянуто у попередньому параграфі). В двох інших задачах буде сформульовано гіпотезу щодо невідомого середнього деякої сукупності.

Задача першого типу

Припустимо, що виборці деякого округу не збираються в цьому році голосувати за кандидатів тих партій, яким вони віддавали свої голоси в минулому. Як правило, 2/3 з них голосували за зелених, а 1/3 – за помаранчевих. Однак, мали місце події, які дають підґрунтя припускати, що виборці змінять свій вибір. При цьому визначити точно величину цього впливу неможливо. Це означає, що у відповідності до нашого припущення зазначені події будуть мати велике значення для результатів голосування, але вгадати, яка з партій від цього виграє, неможливо.

Відповідно, можна висунути гіпотезу, що звичне співвідношення між голосами, поданими за обидві партії, не зміниться, і необхідно перевірити цю гіпотезу на рівні значущості 0,05. При цьому ми сподіваємось, що зможемо цю гіпотезу відкинути. В дійсності, ми просто сподіваємось показати, що звичне для нас співвідношення 2/3 до 1/3 буде порушено.

Нехай вибірка складається з 450 спостережень і встановлено, що 225 виборців будуть голосувати за зелених. Таким чином, маємо:

$$N = 450, \quad p = 2/3, \quad r = 225.$$

Звідси отримаємо: $N \cdot p = 300$. Це значення відповідає очікуваному результату експерименту за умови, що висунута гіпотеза є вірною. Іншими словами, якщо гіпотеза, у відповідності до якої 2/3 виборців будуть голосувати за зелених, є вірною, то очікується, що 2/3 членів вибірки віддадуть свої голоси за зелених.

Припускаючи, що дана гіпотеза вірна, отримаємо:

$$z = \frac{r - N\,p}{\sqrt{N\,p(1-p)}} = \frac{225 - 300}{\sqrt{450 \cdot (2/3) \cdot (1/3)}} = \frac{-75}{10} = -7{,}5.$$

Оцінку $z = -7{,}5$ отримано у припущенні, що гіпотеза щодо зберігання співвідношення між голосами виборців, поданими за кандидатів різних партій, є вірною.

Можна припустити, що отримано нескінченне число випадкових вибірок потужністю $N = 450$, кожна, і на цій основі сформульовано новий розподіл, утворений із показників, які представляють собою число голосів, поданих за зелених в кожній вибірці.

Вибірка, яку ми фактично зібрали, містить тільки 225 голосів, поданих за зелених. Якщо розглядати її як одну із великої множини незалежних вибірок, то ми отримаємо для неї значення $z = -7{,}5$. На рис. 6.6 показано, який вигляд буде мати графік нашого розподілу.

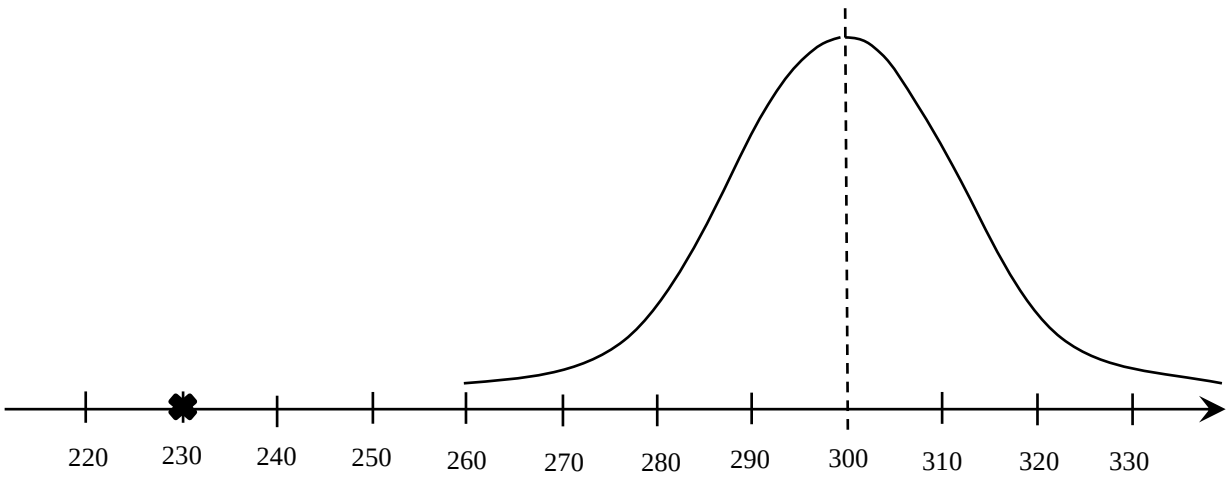


Рис. 6.6. Графік розподілу кількості виборців

Із сказаного випливає таке: якщо гіпотеза вірна, то у отриманій вибірці число голосів, поданих за зелених, відрізняється від очікуваного значення на 7,5 одиниць стандартного відхилення. Ймовірність отримання такого результату визначається частиною спостережень, для яких $z > 7{,}5$ або $z < -7{,}5$. З таблиці знаходимо, що це значення ймовірності є таким: $p < 0{,}00001$.

Таким чином, гіпотеза відносно того, що співвідношення між голосами, поданими за зелених та оранжевих, дійсно складає 2/3 до 1/3 повинна бути відхилена. Хоча в даному випадку і не було доведено, які саме події стали вирішальним фактором впливу на результати голосування, все таки можна вважати, що співвідношення між голосами змінилось. При цьому необхідно розуміти, що розраховані значення $z-$оцінок відповідають $z-$оцінкам, отриманим за умови, що сформульована гіпотеза є вірною. *Логіка тут така*: *якщо значення $z-$оцінки буде відповідати події, ймовірність появи якої є незначною, то рішенням буде відхилення гіпотези*.

Аналогічна процедура буде застосована також у випадку перевірки гіпотези щодо середнього значення деякої сукупності. В цьому випадку результатом експерименту буде значення вибіркового середнього, а не просто число повторень результату експерименту, який нас цікавить.

Задача другого типу: перевірка гіпотези щодо середнього значення деякої сукупності даних

Припустимо, що відоме стандартне відхилення деякої сукупності, яка є однією із великої множини випадкових вибірок однакового об'єму. Середні цих вибірок утворюють новий розподіл. Як і раніше, починаємо з нульової гіпотези і визначення рівня значущості. Необхідно визначити величину розбіжності між фактичним вибірковим середнім і величиною, яка очікується у відповідності з нашою гіпотезою. Оскільки ми припускаємо, що висунута гіпотеза є вірною, то за допомогою центральної граничної теореми для середніх можна встановити вигляд даного розподілу середніх.

Після формування розподілу визначимо z – оцінку обчисленого значення вибіркового середнього. На основі цієї оцінки встановимо наскільки близькою є величина отриманого середнього до очікуваного значення. Якщо ймовірність розбіжності між теоретичним і фактичним значеннями середнього перевищує рівень значущості, то гіпотеза приймається. Якщо ж вона буде меншою рівня значущості, то гіпотеза відхиляється.

Приклад 6.2. Чи буде середній коефіцієнт розумового розвитку (КРР) *десятилітніх хлопчиків, які збираються стати водіями автомобілів*, відрізнятись від середнього генеральної сукупності. Нехай дослідження виконується в місті, в якому середній КРР десятилітніх хлопчиків $\mu = 100$, а стандартне відхилення $\sigma = 20$.

Нехай необхідно показати, що десятилітні хлопчики, які збираються стати водіями автомобілів, відрізняються за рівнем свого інтелектуального розвитку від середнього генеральної сукупності. За нуль-гіпотезу прийнято:

H_0: *вибрана група – типовий представник генер. сукупності*;

H_1: *вибрана група відрізняється від генеральної сукупності*.

Виберемо за рівень значущості 0,05 і будемо сподіватись, що зможемо відхилити висунуту гіпотезу. Нехай із генеральної сукупності вибрано 64 хлопчики, які збираються стати водіями. Встановлено, що середній КРР для них складає: $\mu_{гр} = 108$.

Ми вважаємо, що нуль-гіпотеза є вірною, тобто, їхні 64 КРР можна розглядати як випадкову вибірку із великої сукупності спостережень і яке представляє собою одну з численних вибірок потужністю 64, кожна. З теорем 6.2 і 6.3 (для середнього ряду розподілу, утвореного з середніх значень) випливає, що при таких умовах розподіл середніх значень цих вибірок буде мати середнє $\mu = 100$, а стандартне відхилення

$$\sigma_{гр} = \frac{\sigma}{\sqrt{N}} = \frac{20}{\sqrt{64}} = 2{,}5.$$

Цей розподіл представлено на рис. 6.7, на якому середнє $\mu_{гр} = 108$ позначено хрестиком.

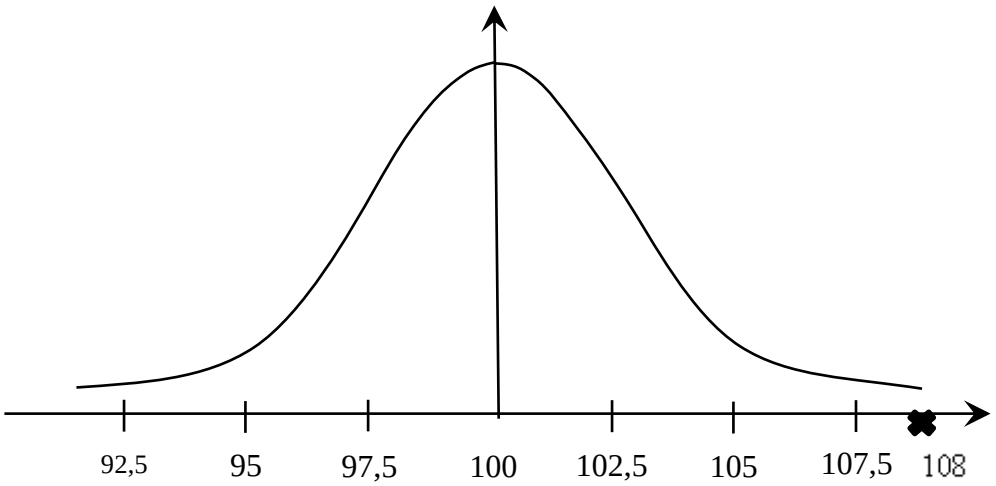


Рис. 6.7. Форма розподілу для КРР при $\mu_{гр} = 108$

З рис. 6.7 видно, що в припущенні істинності нульової гіпотези ми отримали вибіркове середнє, величина якого на 8 пунктів перевищує точку рівноваги нормального розподілу, а стандартне відхилення випадкової вибірки складає 2,5. Іншими словами, припускаючи істинність нульової гіпотези, ми повинні зробити висновок, що отримане нами вибіркове середнє на 3,2 стандартного відхилення перевищує очікуване значення середнього, тобто, 100.

З таблиці для площ нормального розподілу знайдемо, що ймовірність отримання значення, z −оцінка якого відрізняється більше ніж на 3 одиниці (в обидва боки) від точки рівноваги, складає <0,001. Таким чином, припустивши істинність нульової гіпотези, ми отримали вибіркове значення, яке настільки сильно відрізняється від очікуваної величини, що ймовірність його появи є меншою 0,001. Очевидно, що такий результат свідчить про те, що ми повинні відхилити нульову гіпотезу і прийняти альтернативну. Тобто, хлопчики, які в майбутньому збираються працювати водіями, в середньому відрізняються за рівнем свого інтелектуального розвитку від інших хлопчиків їхнього віку.

В такому випадку було б корисно застосувати формулу обчислення z −оцінки вибіркового середнього розподілу вибіркових середніх. Ця формула дозволяє швидше розв'язувати подібні задачі. Так, z −оцінка вибіркового середнього обчислюється за виразом:

$$z = \frac{Оцінка\ відхилення\ середнього}{Стандартне\ відхилення\ середніх} = \frac{\bar{x} - \mu}{\sigma / \sqrt{N}}, \qquad (6.4.1)$$

де $\bar{x}$ − значення вибіркового середнього (в нашому прикладі $\bar{x} = 108$); μ − середнє розподілу вибіркових середніх (ми припускали, що $\mu = 100$); σ − стандартне відхилення КРР для генеральної сукупності (в нашому прикладі $\sigma = 20$).

Підставляючи конкретні значення в формулу (6.4.1), отримаємо:

$$z = \frac{\bar{x} - \mu}{\sigma\sqrt{N}} = \frac{108 - 100}{20\sqrt{64}} = 3{,}2.$$

Звертаючись тепер до таблиці площ нормального розподілу (при $z = 3{,}2$), відхиляємо нуль-гіпотезу на рівні значущості 0,05.

Задача третього типу

Нехай, *необхідно перевірити гіпотезу стосовно того, що студенти п'ятого курсу ІПСА НТУ КПІ є типовими представниками всіх студентів п'ятого курсу за вмінням виконувати курсові проекти* з проектування комп'ютерних інформаційних систем. Перевіримо нульову гіпотезу для даного випадку на рівні значущості 0,01. За основу були взяті результати виконання курсових проектів у всіх університетах Києва. Середня оцінка за курсовий проект по місту виявилась рівною $\mu = 72$ (при максимальному значенні 100); стандартне відхилення $\sigma = 12$. В групі ІПСА, яка аналізувалась, налічувалось $N = 36$ студентів при середньому значення оцінки $\mu_{гр} = 74$. Тобто, необхідно визначити чи суттєво відрізняється середнє для вибраної групи від середнього генеральної сукупності.

Припустимо, що наша гіпотеза є вірною і що отримані 36 оцінок за курсові проекти можна розглядати як випадкову вибірку із всієї сукупності оцінок. Це означає, що ми можемо розглядати вибіркове середнє для групи студентів, $\mu_{гр} = 74$, як одне із значень теоретичного розподілу середніх різних вибірок об'ємом $N = 36$, кожна. У відповідності з теоремою 6.1 розподіл таких вибіркових середніх є нормальним. При цьому середнє розподілу середніх дорівнює 72 (таке ж значення має середнє генеральної сукупності), а стандартне відхилення:

$$\sigma_{рс} = \frac{\sigma}{\sqrt{N}} = \frac{12}{\sqrt{36}} = 2.$$

Таким чином, отримане вибіркове середнє для вибраної групи студентів, $\mu_{гр} = 74$, на $z = 1$ перевищує теоретично очікуване значення, тобто,

$$z_{гр} = \frac{\bar{x} - \mu}{\sigma / \sqrt{N}} = \frac{74 - 72}{12 / \sqrt{36}} = 1{,}0.$$

У відповідності із таблицею площ нормального розподілу, *ймовірність розбіжності між середнім генеральної сукупності і середнім вибраної групи студентів* є більшою вибраного нами рівня значущості, а тому *нуль-гіпотеза приймається*. Тобто середнє генеральної сукупності і середнє вибраної групи відрізняються не суттєво.

Таким чином, на основі виконаного експерименту не можна зробити висновок, що студенти ІПСА НТУ КПІ відрізняються за рівнем виконання курсових проектів з проектування інформаційних систем від студентів інших університетів.

6.5 Зауваження стосовно термінології

Необхідно строго розрізняти показники генеральної сукупності і показники, отримані для деякої конкретної вибірки. Так, наприклад, середній

КРР генеральної сукупності хлопчиків може дорівнювати 100 при $\sigma = 12$, а значення вибіркового середнього конкретної вибірки, взятої з цієї генеральної сукупності, може складати 95, 110 чи 132. Середнє генеральної сукупності, $\mu = 100$, і її стандартне відхилення, $\sigma = 12$, називають *параметрами*, а показник, розрахований для окремої випадкової вибірки, називають *статистикою*.

Зазначимо, що будь-яке судження, сформульоване у вигляді гіпотези, відноситься до параметра. Для перевірки гіпотези формується конкретна вибірка і на її основі обчислюється статистика. Обчислене значення статистики використовується для прийняття або відхилення висунутої гіпотези.

В процесі перевірки гіпотез ми часто звертаємось до нормального розподілу і легко запам'ятовуємо основні відповідності між значеннями $z-$оцінок та ймовірностями. Наприклад, ймовірність отримання нормального випадкового спостереження, $z-$оцінка якого відрізняється від середнього більше, ніж на 1,96, дорівнює 0,05. Таким чином, при перевірці гіпотез на рівні значущості $\alpha = 0{,}05$ можна автоматично зробити висновок, що гіпотезу не можна відхилити, якщо отримане значення оцінки попадає в інтервал $-1{,}96 \le z_i \le 1{,}96$. Якщо ж воно є меншим –1,96 або більшим +1,96, то гіпотеза відхиляється.

6.6 Односторонні критерії

Нехай наша мета полягає в тому, щоб відхилити гіпотезу відносно того, що деякий параметр, наприклад, середнє генеральної сукупності, дорівнює деякому значенню. Таким чином, необхідно перевірити гіпотезу, *що дане генеральне середнє дорівнює цьому значенню* і спробувати відхилити її на основі вибіркового значення.

Наприклад, припустимо, що КРР восьмилітніх хлопчиків, які збираються стати пожежниками, відрізняються від генеральної сукупності за рівнем свого інтелекту. Висувається гіпотеза, що ці хлопчики мають середній КРР=100. На основі результатів деякої вибірки хлопчиків, які планують стати пожежниками, ми спробуємо відхилити гіпотезу, у відповідності до якої їхній середній КРР виявиться значно меншим або значно більшим 100.

Розглянемо тепер ситуацію, коли необхідно показати, що члени деякої окремої групи відрізняються від генеральної сукупності. Це означає, що робиться припущення не тільки щодо відмінності вибіркового показника від параметра, але і щодо характеру цієї відмінності. Наприклад, припустимо, що хлопчики, які збираються стати пожежниками, в цілому повинні мати вищий рівень інтелектуального розвитку, ніж представники великої групи хлопчиків їхнього віку. Відповідно, ми спрогнозуємо, що середній КРР вибірки хлопчиків, які збираються стати пожежниками, буде перевищувати 100. Тепер ми маємо справу *не просто із встановленням деякої відмінності, але і з прогнозуванням того, що в статистиці називають "напрямом" даної відмінності*. Отримання

вибіркового середнього, яке виявиться навіть на 0,1 стандартного відхилення меншим від параметра, буде означати, що наше припущення було помилковим.

Сформулюємо нульову гіпотезу так: середній КРР групи з 64 хлопчиків, які збираються стати пожежниками, дорівнює відповідному показнику генеральної сукупності, тобто, $\mu_{гр} = \mu = 100$ при $\sigma = 20$.

Перевіримо цю гіпотезу на рівні значущості 0,05. Будемо також вважати, що 64 хлопчики вибрані випадково і таких груп існує велика множина. Таким чином, середнє цієї групи можна вважати одним із множини вибіркових середніх. Якщо сформульована гіпотеза вірна, то розподіл вибіркових середніх буде мати вигляд, як показано на рис. 6.8.

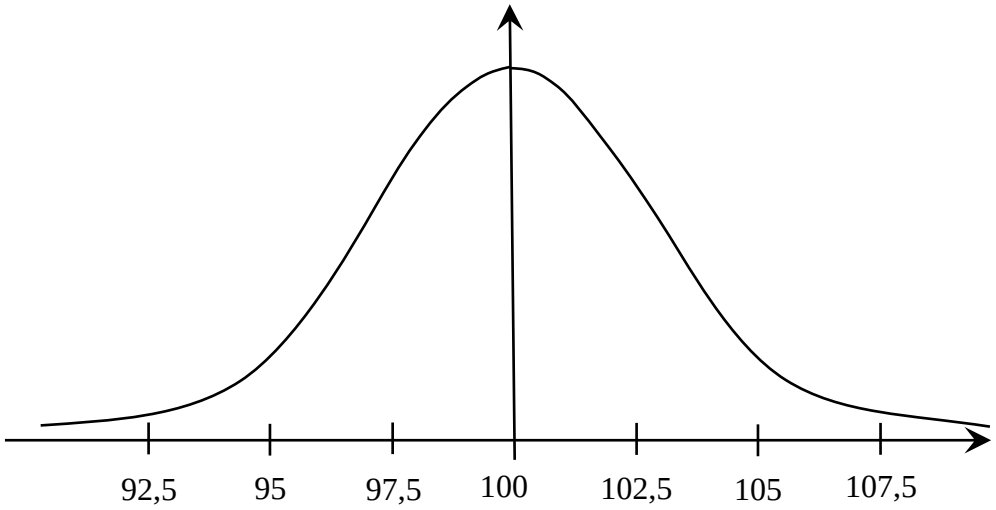


Рис. 6.8. Розподіл вибіркових середніх

Тепер питання полягає в тому, яке значення отриманого нами вибіркового середнього дозволить відхилити висунуту гіпотезу? Це дозволить підтвердити наше початкове припущення. Оскільки вибрано рівень значущості 0,05, це означає, що у випадку справедливості висунутої гіпотези ми ризикуємо зробити помилку першого роду тільки у 5 випадках із 100. Тепер необхідно визначити такий інтервал, в який попаде тільки 5/100 вибіркових середніх, якщо висунута нами гіпотеза вірна.

Нашим початковим припущенням було те, що середній КРР хлопчиків, які збираються стати пожежниками, перевищує 100. Сформулюємо і перевіримо нуль-гіпотезу, що КРР=100. Таким чином, ми не зможемо відхилити дану гіпотезу, якщо отримаємо будь-яке значення середнього, яке є меншим 100. Необхідно вибрати такий інтервал значень КРР, щоб шанси на успіх були максимальними, якщо наше початкове припущення виявиться правильним. Тобто необхідно вибрати 5%-й інтервал, в якому будуть міститись найбільші значення вибіркових середніх.

У відповідності з таблицею площ нормального розподілу, 5% найбільших спостережень нормального розподілу повинні щонайменше на 1,64

стандартного відхилення перевищувати середнє генеральної сукупності. Для розподілу вибіркових середніх (рис. 6.9), 1,64 стандартного відхилення (тобто, для вибіркових середніх $z = 1{,}64$) складають 4,1:

$$z \cdot \frac{\sigma}{\sqrt{N}} = 1{,}64 \cdot \frac{20}{\sqrt{64}} = 1{,}64 \cdot 2{,}5 = 4{,}1.$$

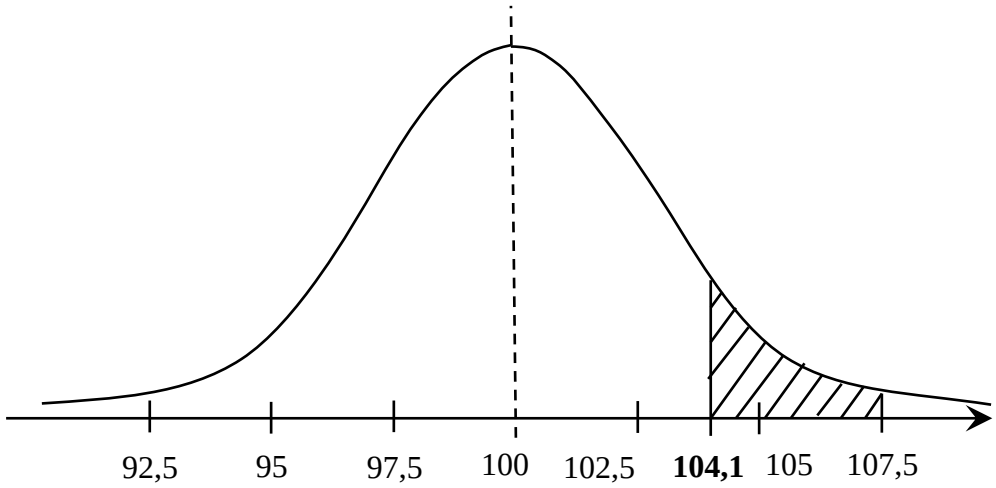


Рис. 6.9. Форма розподілу вибіркових середніх для КРР; 5/100 найбільших значень перевищують рівень 104,1

Таким чином, 5% самих великих значень вибіркових середніх, які були б отримані за умови, що висунута гіпотеза є вірною, повинні перевищувати більше, ніж на 4,1 середнє генеральної сукупності. Іншими словами, 5% найбільших вибіркових середніх повинні мати величину, яка перевищує 104,1. Отриманий результат представлено на рис. 6.9 заштрихованою областю.

Нульова гіпотеза буде відхилена, якщо отримане значення вибіркового середнього перевищить 104,1. Тим самим ми підтвердимо початкове припущення, що середній КРР хлопчиків, які збираються стати пожежниками, перевищує 100. Інакше, гіпотеза буде прийнята. Вважається, що прийнятий рівень значущості 0,05 найкраще відповідає вимогам початкового припущення.

Такий метод перевірки називають *одностороннім тестом (критерієм)* тому, що область відхилення гіпотези повністю міститься в одному з хвостів розподілу.

Нехай необхідно довести, що середній КРР хлопчиків, які збираються стати пожежниками, є меншим 115. В такому випадку необхідно відхилити гіпотезу, що середнє дорівнює 115. Для цього необхідно, щоб вибіркове середнє значно відрізнялось від цього значення, але тепер вже в іншому напрямку. Застосовуючи аналогічну аргументацію, можна встановити, що область відхилення гіпотези відповідає заштрихованій площі на рис. 6.10. В даному прикладі ми також вважаємо, що $\sigma = 20$, а $N = 64$.

Легко визначити, що на рівні значущості 0,05 граничне значення складає 110,9, тобто:

$$\mu_{гр} - z \cdot \frac{\sigma}{\sqrt{N}} = 115 - 1{,}64 \cdot \frac{20}{\sqrt{64}} = 115 - 1{,}64 \cdot 2{,}5 = 110{,}9 .$$

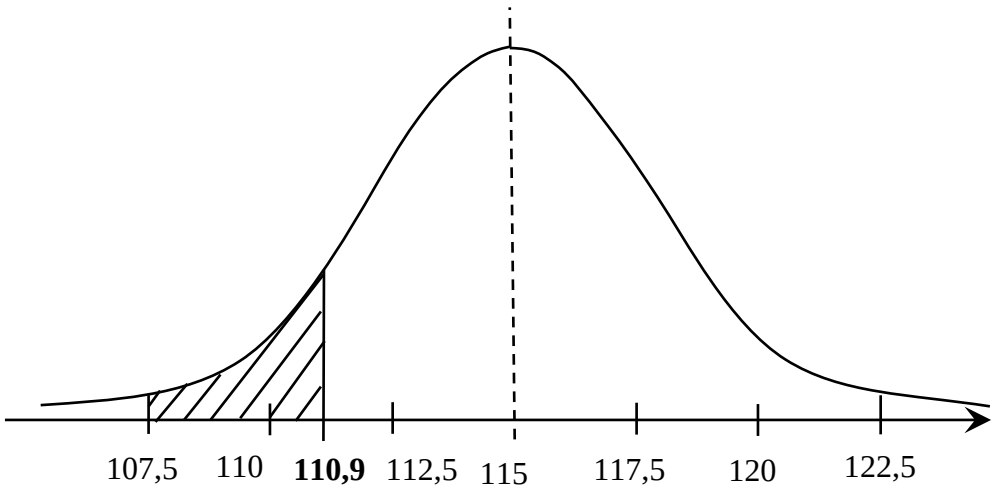


Рис. 6.10. Приклад одностороннього тестування середнього

Коли необхідно зробити перевірку, що середнє просто відрізняється від деякого значення, використовується *двосторонній тест (критерій)*. При цьому вибіркове середнє порівнюється з теоретично очікуваною величиною і визначається максимальне відхилення від неї в обидва боки. Це дасть нам можливість відхилити висунуту гіпотезу. Однак, якщо необхідно довести, що параметр відрізняється від очікуваного значення в *деякому визначеному напрямку*, то для цього використовується односторонній тест. Як і раніше, область відхилення гіпотези визначається рівнем значущості, але логіка задачі є такою, що в даному випадку необхідно брати до уваги тільки одну область, яка розташована в одному із хвостів розподілу.

Якщо необхідно довести, що середнє перевищує деяке значення, то область, в якій відхиляється гіпотеза, знаходиться в правому хвості. Якщо ж необхідно довести, що середнє вихідної (початкової) сукупності є меншим деякого значення, то така область буде розташована в лівому хвості.

Якщо гіпотеза припускає, *яким чином вибіркове середнє буде відрізнятись від очікуваної величини*, то необхідно користуватись одностороннім тестом. В такому випадку є більший шанс довести справедливість теорії, ніж при застосуванні двостороннього тесту.

Так, якби в попередньому прикладі було застосовано двосторонній тест, то область відхилення гіпотези була б такою, як це показано на рис. 6.11.

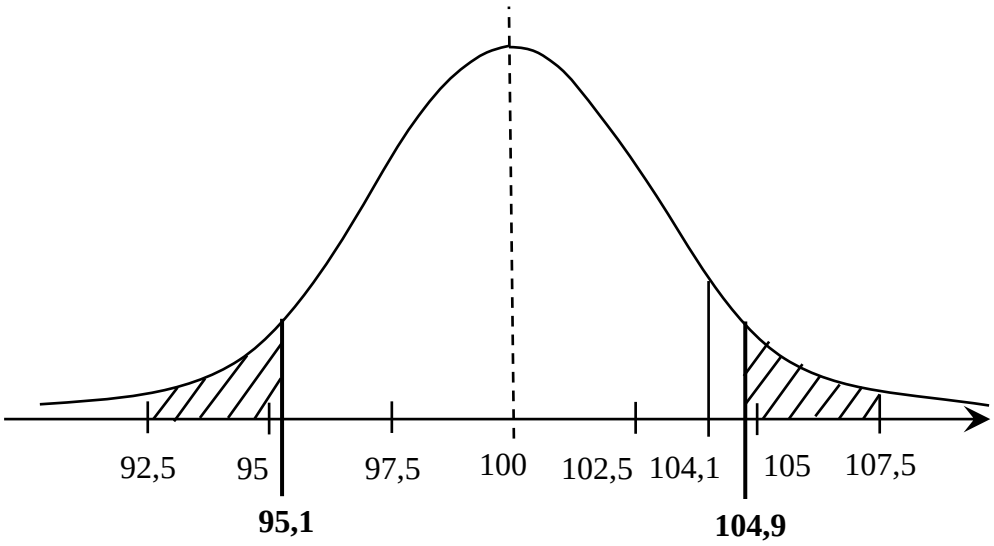


Рис. 6.11. Приклад двостороннього тестування

Значення, які знаходяться між 104,1 і 104,9 в даному випадку будуть відрізнятись від очікуваного недостатньо щоб довести вихідне припущення за допомогою двостороннього тесту. Однак, якщо отримане значення попаде в цей інтервал, то при застосуванні одностороннього тесту ми могли б відхилити висунуту гіпотезу.

По суті, у випадку застосування двостороннього тесту теорія просто прогнозує, що вибіркове середнє буде відрізнятись від очікуваного значення. Припущення вважається підтвердженим, якщо це середнє буде фактично відрізнятись від очікуваного значення. У випадку застосування одностороннього тесту гіпотеза припускає, що вибіркове середнє буде знаходитись з визначеного боку по відношенню до генерального середнього. Той факт, що таке припущення виявилось правильним, вже сам по собі є частковим підтвердженням висунутої гіпотези.

Іншими словами, він свідчить на користь даної гіпотези і, таким чином, величина розбіжності між очікуваним і фактичним показниками, яка необхідна для повного підтвердження висунутої гіпотези, буде меншою, ніж у випадку двостороннього тесту. Це зумовлено тим, що підтвердилось припущення щодо напрямку цієї відмінності.

Односторонній тест застосовують також при перевірці гіпотез стосовно інших параметрів. Наприклад, цей тест іноді застосовують для доведення того факту, що дисперсія вихідної сукупності перевищує деяке конкретне значення. В подальшому викладенні матеріалу будуть наведені інші приклади застосування одностороннього тесту.

6.7 Приклади застосування процедур перевірки гіпотез

Приклад 6.1. Строк служби (випадкова змінна X) дисплейних панелей,

виготовлених за стандартною технологією є нормальною випадковою величиною з параметрами розподілу: $\{X\} \sim N\{1000, (200)^2\}$.

Після впровадження нової технології необхідно експериментально встановити чи будуть мати нові панелі довший середній строк служби ніж старі?

Сформулюємо *гіпотези*:

$$H_0: \mu = 1000 \, проти \, H_1: \mu > 1000,$$

де μ – середнє генеральної сукупності.

Для перевірки гіпотези взяли випадкову вибірку потужністю $N = 25$ дисплейних панелей, виготовлених за новою технологією.

Тобто, вибірковий простір у даному випадку можна записати так:

$$\Omega = \{(x_1, x_2, \ldots, x_{25}) \mid 0 < x_i < \infty, \; i = 1, 2, \ldots, 25\}.$$

Також припустимо, що ми відхилимо нуль-гіпотезу, якщо $\mu_1 > 1080$ годин, де μ_1 – середній строк служби нових кінескопів.

Приймемо наступні правила прийняття рішень (ППР):

– *Відхилити* H_0 *і прийняти* H_1, *якщо* $\bar{x} > 1080$;
– *Відхилити* H_1 *і прийняти* H_0, *якщо* $\bar{x} \le 1080$,

де $\bar{x}$ – середнє вибірки, взятої для дослідження.

Таким чином, *область відхилення* H_0 визначається так:

$$c = \left\{ (x_1, x_2, \ldots, x_{25}) \mid \bar{x} = \frac{x_1 + x_2 + \ldots + x_{25}}{25} > 1080 \right\}.$$

Будемо перевіряти гіпотези шляхом оцінювання ймовірностей припущення помилок 1-го та 2-го роду. Ймовірність помилки першого роду:

$$\alpha = p(\textit{перевірка призведе до помилки } 1-\textit{го роду}) =$$

$$= p(\textit{буде відхилена гіпотеза } H_0, \textit{ коли } H_0 \textit{ вірна}) =$$

$$= p\,(\bar{x} > 1080 \;\; \textit{при} \;\; \mu = 1000) =$$

$$= p\left(\frac{\bar{x} - \mu}{\sigma / \sqrt{N}} > \frac{1080 - 1000}{200 / \sqrt{25}}\right) = p\left(\frac{\bar{x} - \mu}{\sigma / \sqrt{N}} > 2\right) = 0{,}02.$$

На рис. 6.12 наведено криву розподілу для цього випадку.

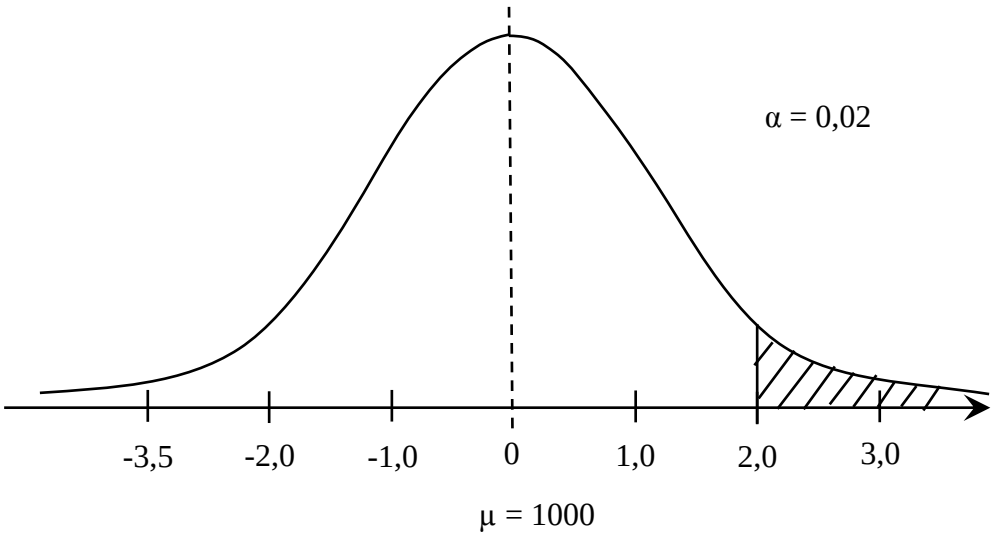


Рис. 6.12. Крива нормального розподілу при перевірці гіпотез

Обчислимо ймовірність помилки другого роду:

$$\beta = p(\textit{перевірка призведе до помилки } 2-\textit{го роду}) =$$
$$= p(\textit{буде прийнята гіпотеза } H_0, \textit{ коли } H_0 \textit{ невірна}) =$$

$$= p\ (\bar{x} \le 1080 \ \textit{ при } \ \mu > 1000) =$$

$$= p\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{N}} \le \frac{1080-\mu}{200/\sqrt{25}} \ \textit{ при } \ \mu > 1000 \right) = \ ?$$

Ми не можемо закінчити обчислення помилки другого роду, тому що не задано значення μ. У всіх прикладах, в яких альтернативна гіпотеза не задає одне значення для параметра, обчислити β неможливо.

Повернемось до головного питання: чи будуть мати нові панелі довший середній строк служби, ніж старі? Припустимо, що для випадкової вибірки $N = 25$ нових панелей середній строк служби склав:

$$\bar{x} = 1100.$$

У відповідності до правила прийняття рішення ми повинні відхилити H_0, оскільки $\bar{x} > 1080$. Іншими словами, панелі, виготовлені за новою технологією, мають довший середній строк служби.

Якщо ми відхилимо H_0, то як ми взнаємо, що прийняли правильне рішення? Пояснення наступне: в такому випадку єдино можливою помилкою є помилка 1-го роду (тобто, відхилити H_0, коли H_0 вірна). Раніше ми визначили, що ймовірність помилки 1-го роду $\alpha = 0{,}02$. Така низька ймовірність помилки 1-го роду надає впевненості у правильності прийнятого рішення (відхиляючи її, ми помиляємось тільки у 2% випадків).

6.8 Альтернативне формулювання процедури перевірки гіпотези

Припустимо, що необхідно сформулювати процедуру перевірки нуль-гіпотези, сформульованої вище (для $N = 25$), але при цьому ймовірність помилки першого роду повинна дорівнювати 0,01 ($\alpha = 0{,}01$). Таку процедуру можна сформулювати наступним чином.

Скористаємось сформульованими гіпотезами:

$$H_0 : \mu = 1000 \, \textit{проти} \, H_1 : \mu > 1000,$$

де μ – середнє нормальної генеральної сукупності при стандартному відхиленні $\sigma = 200$. Нехай $X = (x_1, x_2, ..., x_{25})$ – випадкова вибірка потужністю $N = 25$ з цієї генеральної сукупності; і нехай

$$\bar{x} = \frac{x_1 + x_2 + ... + x_{25}}{25}$$

– середній строк служби 25 панелей, виготовлених за новою технологією. Оскільки значення $\bar{x}$ очікується великим, якщо середнє генеральної сукупності μ велике, і значення $\bar{x}$ очікується малим, якщо середнє генеральної сукупності μ мале, то логічно відхилити H_0 (і можливо прийняти H_1), якщо обчислене на основі експерименту $\bar{x}$ буде значно більшим 1000 ($\bar{x} > 1000$). Тепер можемо сформулювати *правило прийняття рішення*:

– *відхилити* H_0, *якщо* $\bar{x} > c$.

Визначимо граничне (критичне) значення константи c так, щоб ймовірність помилки 1-го роду дорівнювала 0,01 ($\alpha = 0{,}01$):

$$\alpha = p(\textit{перевірка призведе до помилки } 1 - \textit{го роду}) =$$

$$= p(\textit{буде відхилена гіпотеза } H_0, \textit{ коли } H_0 \textit{ вірна}) =$$

$$= p(\bar{x} > c \;\; \textit{при} \;\; \mu = 1000) =$$

$$= p\left(\frac{\bar{x} - \mu}{\sigma / \sqrt{N}} > \frac{c - 1000}{200/5} \right) = p\left(\frac{\bar{x} - \mu}{\sigma / \sqrt{N}} > \frac{c - 1000}{40} \right).$$

З таблиці площ нормального розподілу знайдемо, що при $\alpha = 0{,}01$ константу можна знайти з рівняння:

$$\frac{c - 1000}{40} = 2{,}327.$$

Таким чином, $c = 1093$. Наведену рівність можна інтерпретувати так: z – оцінка $z = \frac{c - 1000}{40} = 2{,}327$ відповідає лівій границі області нормального розподілу, яка займає 1% від загальної площі і знаходиться в правому хвості розподілу.

Тепер можна сформулювати правило прийняття рішень, яке грунтується на аналізі випадкової вибірки потужністю $N = 25$ при ймовірності припуститись помилки першого роду $\alpha = 0{,}01$:

$$- \textit{ відхилити } H_0 \textit{ якщо } \bar{x} > 1093.$$

На основі розглянутого прикладу можна сформулювати *альтернативний метод перевірки гіпотез*, який може бути застосований до всіх випадків з простою нуль-гіпотезою проти простої або композиційної альтернативної гіпотези.

Зазначимо, що гіпотеза називається **простою**, якщо вона задає одне значення для параметра, що аналізується (наприклад, $\mu = 1000;\ \theta = 0{,}5$). Гіпотеза називається **композиційною**, якщо названий параметр може приймати більше одного можливого значення (наприклад, $\mu > 1000$).

6.9 Метод побудови процедури перевірки гіпотез

Виконання процедури перевірки гіпотез передбачає наступні кроки:

Крок 1. Знайти такий статистичний параметр (тобто, функцію випадкової змінної), розподіл якої повністю визначений, якщо H_0 вірна. (Це дозволяє обчислити α, якщо ми використовуємо даний статистичний параметр як тестову статистику.)

В попередньому прикладі цією статистикою було вибіркове середнє $\bar{x}$.

Крок 2. Скористатись статистикою, знайденою на попередньому кроці, як тестовою (перевірочною). Іншими словами, використати цю статистику для формулювання такого правила прийняття рішень, для якого помилка 1-го роду складає α. (Тобто, за допомогою вибраної статистики необхідно визначити критичну область розподілу розміром α.)

В попередньому прикладі ми визначили правило прийняття рішення після обчислення константи c з умови: $p(\bar{x} > c \textit{ коли } H_0 \textit{ вірна}) = 0{,}01$.

Крок 3. Формалізувати правило прийняття рішення, сформульоване на попередньому кроці, із врахуванням конкретного експериментального значення тестової статистики.

В попередньому прикладі таким формальним правилом було:

$$- \textit{ відхилити } H_0 \textit{ якщо } \bar{x} > 1093.$$

Приклад 6.2. Припустимо, що середнє генеральної сукупності може приймати одне з двох значень: 1 або 2, але в обох випадках стандартне відхилення складає $\sigma = 0{,}5$. Використовуючи наведену вище трикрокову процедуру, необхідно побудувати процедуру перевірки гіпотез:

$$H_0 :\ \mu = 1 \textit{ проти } H_1 :\ \mu = 2.$$

Для цього скористатись випадковою вибіркою потужністю $N = 9$ та ймовірністю помилки першого роду $\alpha = 0{,}05$.

Побудова процедури перевірки гіпотез

***Крок** 1*. (Знайти такий статистичний параметр (тобто, функцію випадкової змінної), розподіл якої повністю визначений, якщо H_0 вірна.)

В даному випадку можна скористатись вибірковим середнім $\bar{x}$. Це можливо завдяки тому, що у випадку справедливості H_0 (тобто, $\mu = 1$), $\bar{x}$ має нормальний розподіл із середнім $\mu = 1$ (при великому числі вибірок потужністю $N = 9$) і стандартним відхиленням $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{N}} = \dfrac{0{,}5}{\sqrt{9}} = \dfrac{0{,}5}{3}$. Іншими словами, випадкову змінну $\bar{x}$ можна використати як тестову статистику, оскільки її розподіл (параметри розподілу) повністю відомий, якщо припустити, що H_0 вірна.

***Крок** 2*. (Використати цю статистику для формулювання такого правила прийняття рішень, для якого помилка 1-го роду складає α.)

Оскільки значення μ є більшим для H_1 ніж для H_0, то логічно очікувати більші значення $\bar{x}$ при справедливості гіпотези H_1, а не H_0. Таким чином, буде логічно відхилити гіпотезу H_0, якщо значення тестової статистики $\bar{x}$ будуть відносно великими. Тобто, правило прийняття рішення можна сформулювати у вигляді:

$$\textit{відхилити}\, H_0, \quad \textit{якщо} \quad \bar{x} > k,$$

де k – деяка константа. Знайдемо значення цієї константи, скориставшись визначенням помилки першого роду (вона задана на рівні $\alpha = 0{,}05$):

$$\begin{gathered}\alpha = p(\textit{перевірка призведе до помилки}\, 1-\textit{го роду}) = \\ = p(\textit{буде відхилена гіпотеза}\ H_0, \textit{коли}\ H_0\ \textit{вірна}) = \\ = p(\bar{x} > k \quad \textit{при}\ \ \mu = 1) = \\ = p\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{N}} > \frac{k-1}{0{,}5/\sqrt{9}}\right).\end{gathered}$$

З таблиці площ нормального розподілу знайдемо, що значенню $\alpha = 0{,}05$ відповідає ліва границя області в правому хвості розподілу, яка дорівнює: 1,645, тобто:

$$\frac{k-1}{0{,}5 \cdot 3} = 1{,}645.$$

Звідси маємо: $k = 1{,}2742$.

***Крок** 3*. Формалізувати правило прийняття рішення, сформульоване на попередньому кроці, із врахуванням конкретного експериментального значення

тестової статистики. Правило прийняття рішень, сформульоване на другому кроці, має вигляд:

$$- \textit{відхилити}\, H_0, \quad \textit{якщо}\, \bar{x} > 1{,}2742.$$

На рис. 6.13 представлені елементи процедури перевірки гіпотез до прикладу 6.2.

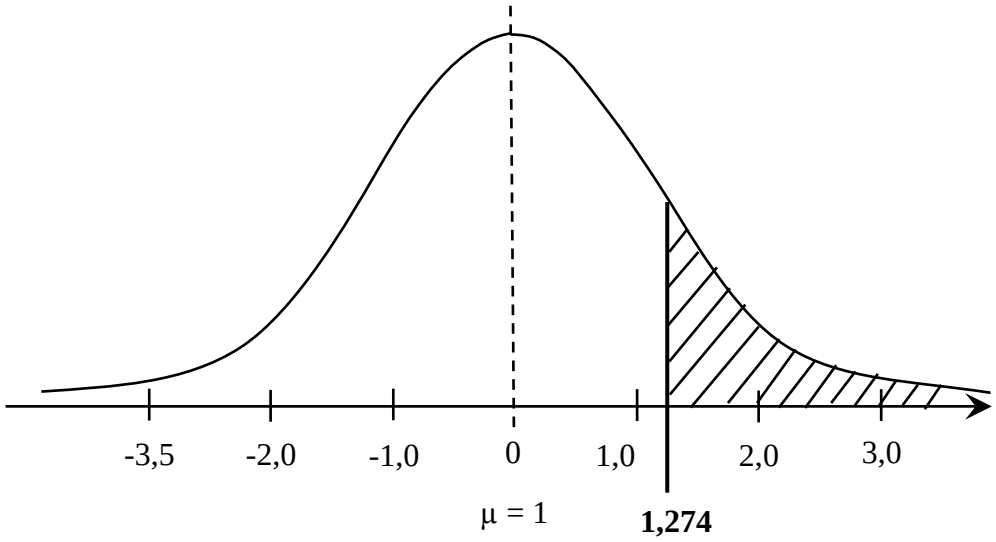


Рис. 6.13. Крива нормального розподілу при перевірці гіпотез

Зробимо резюме щодо основних елементів процедури перевірки гіпотез для розглянутого прикладу:

1. Гіпотези: $\textit{відхилити}\, H_0 :\ \mu = 1 \textit{проти}\, H_1 :\ \mu > 2.$
2. Суцільна крива нормального розподілу (рис. 6.13) відповідає щільності розподілу $\bar{x}$, якщо справедлива нуль-гіпотеза: $\mu = 1.$
3. Крива, зображена пунктиром, – щільність розподілу $\bar{x}$, якщо справедлива альтернативна гіпотеза: $\mu = 2.$
4. Правило прийняття рішення: відхилити H_0, якщо $\bar{x} > 1{,}2742.$

   (Імовірніше, що значення $\bar{x}$ будуть перевищувати 1,2742, якщо справедливою буде H_1, а не H_0.)
5. Площа правого хвоста під кривою розподілу (суцільна лінія) в межах від 1,2742 до $+\infty$ дорівнює $\alpha = p(\textit{помилки}\ 1 - \textit{го}\ \textit{роду}) = 0{,}05.$

Приклад 6.3. Побудуйте трикрокову процедуру перевірки гіпотез (по аналогії з попереднім прикладом) для перевірки таких гіпотез:

$$H_0 :\ \mu = \mu_0 \textit{проти}\, H_1 :\ \mu < \mu_0,$$

де $\mu -$ невідоме середнє нормальної генеральної сукупності із стандартним відхиленням $\sigma;\ \mu_0 -$ константа.

А) Побудуйте процедуру перевірки гіпотез, яка грунтується на випадковій вибірці розміром n при ймовірності зробити помилку 1-го роду $\alpha = 0{,}05$.

Б) Побудуйте процедуру перевірки гіпотез, яка грунтується на випадковій вибірці розміром n при ймовірності зробити помилку 1-го роду α.

Розв'язок

А) *Крок 1*. (Знайти такий статистичний параметр, розподіл якого повністю визначений, якщо H_0 вірна.)

Розглянемо випадкову змінну $\dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$, де $\bar{x}$ – вибіркове середнє. Чи можна використати цю випадкову змінну як тестову статистику? Так, можна, тому, що у випадку, коли H_0 вірна (тобто, якщо $\mu = \mu_0$), то наведена випадкова змінна має стандартний (нормований) нормальний розподіл.

Крок 2. (Використати статистику, вибрану на кроці 1, для формулювання такого правила прийняття рішень, для якого помилка 1-го роду складає $0{,}05$.)

Оскільки спостереження $\bar{x}$ групуються навколо істинного значення середнього генеральної сукупності μ, то логічно відхилити H_0 у тому випадку (і, можливо, прийняти H_1), коли спостережуване значення $\bar{x}$ буде значно меншим μ_0.

Оскільки вибрана змінна $\dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$ має нульове середнє, то логічно відхилити H_0, *коли спостережуване значення цієї змінної буде значно меншим нуля* (тобто, значно меншим середнього). Іншими словами, правило прийняття рішення можна сформулювати так:

$$- \textit{ відхилити } H_0, \textit{ якщо } \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} < c,$$

де $c < 0$ – константа. Знайдемо значення цієї константи, за умови, що при застосуванні наведеного правила прийняття рішення можна зробити помилку 1-го роду з імовірністю 0,05:

$$\alpha = p(\textit{перевірка призведе до помилки } 1 - \textit{го роду}) =$$
$$= p(\textit{буде відхилена гіпотеза } H_0, \textit{ коли } H_0 \textit{ вірна}) =$$
$$= p\left(\frac{\bar{x} - \mu_0}{\sigma / \sqrt{N}} < c \textit{ при } \mu = \mu_0\right) =$$
$$= p\left(\frac{\bar{x} - \mu}{\sigma / \sqrt{N}} < c\right).$$

З таблиці площ нормального розподілу знайдемо, що при $\alpha = 0{,}05$ значення константи c повинно дорівнювати: $c = -1{,}645$.

Крок 3. (Формалізувати правило прийняття рішення, сформульоване на попередньому кроці, із врахуванням конкретного експериментального значення тестової статистики.)

$$- \textit{відхилити } H_0, \textit{ якщо } \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} < -1{,}645.$$

Б) Формулювання правила прийняття рішення при умові, що ймовірність помилки 1-го роду складає α, зводиться до заміни числа 1,645 на загальну величину: $-z_\alpha$:

$$p\left(\frac{\bar{x} - \mu}{\sigma / \sqrt{N}} < -z_\alpha \right) = \alpha .$$

Примітка: значення z_α визначається з рівняння:

$$\int_{z_\alpha}^{\infty} \frac{1}{2\pi} e^{-x^2/2}\, dx = \alpha .$$

Тобто, частина площі нормального розподілу зліва від z_α буде дорівнювати α. Таким чином, правило прийняття рішення набуває вигляду:

$$- \textit{відхилити } H_0, \textit{ якщо } \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} < -z_\alpha .$$

Основні елементи цієї процедури тестування гіпотези ілюструє рис. 6.14.

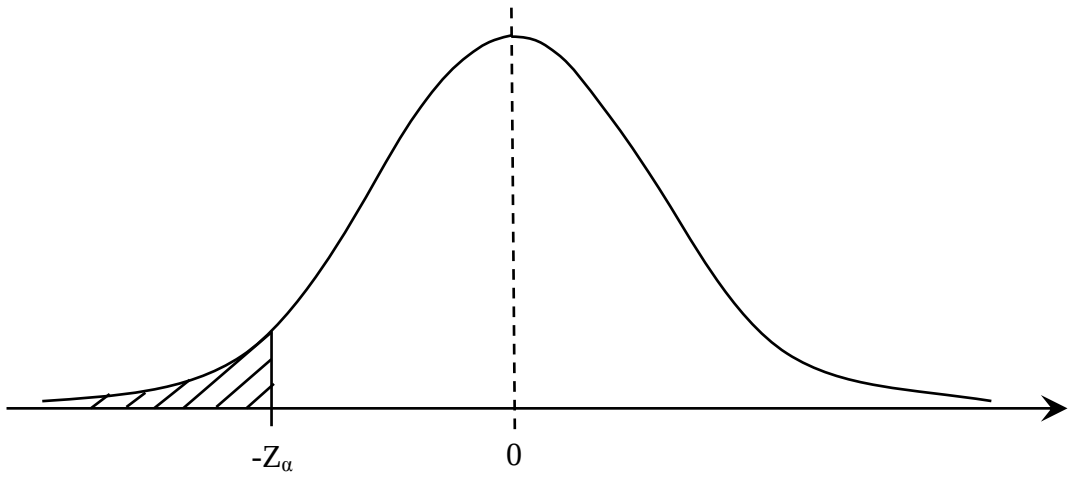


Рис. 6.14. Елементи процедури тестування для прикладу 3. Площа під кривою від $-\infty$ до $-z_\alpha$ дорівнює $\alpha = p(\textit{імовірність помилки } 1-\textit{го роду})$

Приклад 6.4. Випадкова вибірка чашок з кавою ($N = 25$), взята з кавоварного автомату, має середній вміст кави $\bar{x} = 93$ грамів на чашку. Скористайтесь тестом, розробленим в попередньому прикладі, для перевірки нуль-гіпотези, що середнє генеральної сукупності для цього автомата складає $\mu = 95$ проти альтернативної гіпотези: $\mu < 95$ (рівень значимості $\alpha = 0{,}05$).

Припустимо, що маса кави в чашках має нормальний розподіл з дисперсією $\sigma^2 = 1$.

Розв'язок

Процедура тестування, розроблена в попередньому прикладі, визначена наступним правилом прийняття рішень:

$$- \textit{відхилити } H_0, \textit{ якщо } \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} < -1{,}645.$$

В даному прикладі $\sigma = 1,\ N = 25$ і $\mu = 95,$ а звідси значення тестової статистики:

$$\frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{93 - 95}{1/\sqrt{25}} = -\frac{2}{0{,}2} = -10.$$

Оскільки -10 є меншим від $-1{,}645$, то нуль-гіпотеза повинна бути відхилена. Таким чином, середня маса кави, яка надається споживачеві автоматом, є суттєво меншою 95 грамів, тому, що ми відхилили нуль-гіпотезу на рівні значимості 5%.

Приклад 6.5. Використовуючи наведену вище трикрокову методику побудови гіпотез, побудуйте процедуру для тестування таких гіпотез:

$$H_0 : \mu = \mu_0 \textit{ проти } H_1 : \mu \neq \mu_0,$$

де $\mu -$ невідоме середнє генеральної сукупності з нормальним розподілом; стандартне відхилення σ цієї сукупності відоме. Необхідно побудувати процедуру перевірки гіпотез з використанням випадкової вибірки потужністю N за припущення, що помилка першого роду можлива з імовірністю α.

Розв'язок

Крок 1. (Знайти такий статистичний параметр, розподіл якого повністю визначений, якщо H_0 вірна.)

Оскільки генеральна сукупність та нуль-гіпотеза даного прикладу є ідентичними генеральній сукупності та нуль-гіпотезі прикладу 6.3, то в якості перевірочної статистики можна скористатись аналогічною випадковою змінною:

$$\frac{\bar{x} - \mu_0}{\sigma / \sqrt{N}}.$$

Крок 2. (Використати статистику, вибрану на кроці 1, для формулювання такого правила прийняття рішень, для якого помилка 1-го роду складає α.)

Оскільки цілком ймовірно, що спостереження $\bar{x}$ будуть концентруватись навколо μ, то буде логічно відхилити H_0 (і можливо прийняти H_1), коли

спостережуване значення перевірочної статистики $(\bar{x}-\mu_0)/(\sigma/N^{1/2})$ суттєво відрізняється від нуля (тобто, від його середнього).

Іншими словами, логічно скористатись наступним правилом прийняття рішення:

$$-\ \textit{відхилити}\ H_0,\ \textit{якщо}\ \left|\frac{\bar{x}-\mu_0}{\sigma/\sqrt{N}}\right| > c,$$

де $c = const > 0$. Для визначення цієї константи скористаємось визначенням помилки 1-го роду:

$$\alpha = p(\textit{перевірка призведе до помилки}\ 1-\textit{го роду}) =$$

$$= p(\textit{буде відхилена гіпотеза}\ H_0, \textit{коли}\ H_0\ \textit{вірна}) =$$

$$= p\left(\left|\frac{\bar{x}-\mu_0}{\sigma/\sqrt{N}}\right| > c\ \textit{при}\ \mu=\mu_0\right) = p\left(\left|\frac{\bar{x}-\mu}{\sigma/\sqrt{N}}\right| > c\right).$$

Оскільки змінна $(\bar{x}-\mu_0)/(\sigma/N^{1/2})$ є розподіленою за стандартним (нормованим) нормальним розподілом, то константа c задовольняє рівнянню:

$$c = z_{\alpha/2}.$$

Крок 3. (Формалізувати правило прийняття рішення, сформульоване на попередньому кроці, із врахуванням конкретного значення тестової статистики.)

Правило прийняття рішення із врахуванням значення константи c:

$$-\ \textit{відхилити}\ H_0,\ \textit{якщо}\ \left|\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right| > z_{\alpha/2}.$$

Основні елементи розглянутої процедури перевірки гіпотез наведено на рис. 6.15.

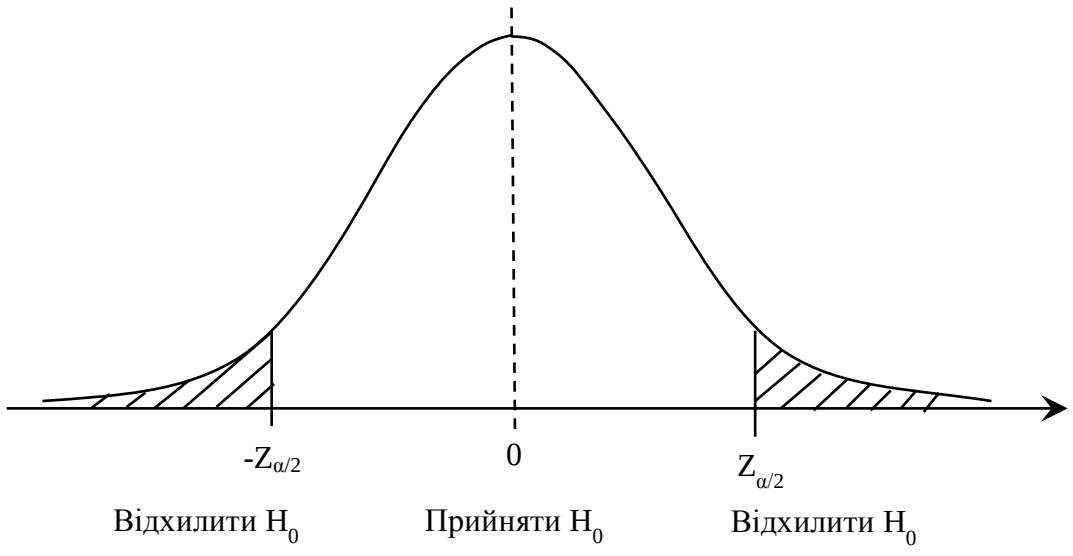


Рис. 6.15. Елементи процедури тестування для прикладу 6.5

Резюме до процедури перевірки.

1. Сформульовані гіпотези: $H_0: \mu = \mu_0$ *проти* $H_1: \mu \neq \mu_0$.
2. Крива щільності розподілу наведена для змінної $(\bar{x} - \mu_0)/(\sigma / N^{1/2})$ при $\mu = \mu_0$.
3. Правило прийняття рішень: *– відхилити* H_0, *якщо* $\left| \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \right| > z_{\alpha/2}$. (Якщо вірна H_0, то ймовірніше, що значення $(\bar{x} - \mu_0)/(\sigma / N^{1/2})$ будуть біля 0.)
4. Площа під кривою від $-\infty$ до $-z_{\alpha/2}$ і від $z_{\alpha/2}$ до $+\infty$ дорівнює $\alpha = p(\textit{помилки 1}-\textit{го роду})$.

6.10 Зауваження щодо односторонньої і двосторонньої перевірки гіпотез

В даному випадку перевірка гіпотез виконується із використанням обох хвостів розподілу. Таку перевірку називають **двосторонньою**, або перевіркою з двома хвостами розподілу. Двосторонні тести, як правило, застосовують до перевірки гіпотез наступного типу:

$$H_0: \theta = \theta_0 \textit{ проти } H_1: \theta \neq \theta_0.$$

Перевірка гіпотез, виконана в прикладах 6.1 - 6.3, називається **односторонньою** перевіркою (або перевіркою з використанням одного хвоста розподілу). При односторонній перевірці застосовують правила прийняття рішень, які передбачають відхилення нуль-гіпотези у випадку, коли тестова статистика приймає значення тільки зліва від деякого числа (порогового значення) або тільки справа від нього. Такі процедури перевірки, як правило, виконують перевірку гіпотез вигляду:

$$H_0: \theta = \theta_0 \textit{ проти } H_1: \theta > \theta_0$$

або

$$H_0: \theta = \theta_0 \textit{ проти } H_1: \theta < \theta_0.$$

6.11 Оцінювання надійності правил (процедур) перевірки статистичних гіпотез

– Наскільки хорошою (надійною) є конкретна процедура перевірки статистичної гіпотези?

– Чи є вона кращою від іншої процедури тестування тієї ж гіпотези?

– Чи є вона найкращою серед множини всіх процедур перевірки гіпотез?

– Які є стандартні критерії оцінювання якості (надійності) процедур перевірки гіпотез?

Відповіді на ці запитання залежать від типу конкретної процедури, що аналізується. Процедури тестування *простих нуль-гіпотез проти простих альтернативних* гіпотез оцінюються їх *потужністю*. Процедури перевірки *простих гіпотез проти альтернативних композиційних* оцінюють за їх *функціями потужності* або за *операційними характеристичними кривими*. Нижче розглянемо деякі з цих характеристик статистичних тестів і скористаємось ними для вибору *найкращих тестів* та *найбільш потужних тестів*.

Потужність тесту – це міра його здатності відхилити нуль-гіпотезу, якщо нуль-гіпотеза невірна. Формально це записується так.

Означення 6.4. Розглянемо процедуру перевірки простих гіпотез:

$$H_0: \theta = \theta_0 \ \textit{проти}\ H_1: \theta = \theta_1,$$

де θ – параметр розподілу генеральної сукупності.

Потужністю процедури (правила) перевірки називають ймовірність того, що тестування призведе до відхилення нуль-гіпотези $\theta = \theta_0$, якщо вірна альтернативна гіпотеза $\theta = \theta_1$. Тобто потужність тесту визначається величиною $1-\beta$, де β – ймовірність помилки другого роду, тобто ймовірність прийняття нуль-гіпотези, якщо вірною є альтернативна.

Якщо виконується перевірка простої гіпотези $H_0: \theta = \theta_0$ проти деякої композиційної гіпотези стосовно θ, то потужність тесту при $\theta = \theta_1$ визначається ймовірністю того, що тест призведе до відхилення нуль-гіпотези, $\theta = \theta_0$, якщо дійсним значенням параметра θ є θ_1.

Приклад 6.6. Нехай $\Re$ – популяція, яка має нормальний розподіл з дисперсією $\sigma^2 = 0{,}25$, але її середнє μ – невідоме. Необхідно знайти α, β, а також потужності кожного з наведених нижче правил щодо перевірки гіпотез:

$$H_0: \mu = 1 \ \textit{проти}\ H_1: \mu = 2.$$

Кожне правило ґрунтується на спостережуваному значенні середнього $\bar{x}$ випадкової вибірки (з генеральної сукупності) потужністю $N = 9$. Дано три правила:

- правило 1: *відхилити* H_0, *якщо* $\bar{x} > 1{,}3878$;
- правило 2: *відхилити* H_0, *якщо* $\bar{x} < 0{,}6122$;
- правило 3: *відхилити* H_0, *якщо* $\bar{x} > 2$ *або якщо* $\bar{x} < 0{,}6122$.

Розв'язок:

Для правила 1:

$$\alpha = p(\textit{відхилення}\ H_0,\ \textit{якщо}\ H_0\ \textit{вірна}) =$$
$$= p(\bar{x} > 1{,}3878,\ \textit{якщо}\ \mu = 1) =$$

$$= p\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{N}} > \frac{1{,}3878-1}{0{,}5/\sqrt{9}}\right) = p\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{N}} > 2{,}327\right) = 0{,}01.$$

Примітка: при обчисленнях було використано той факт, що змінна $(\bar{x}-\mu)/(\sigma/\sqrt{N})$ має стандартний нормальний розподіл.

$$\beta = p(\textit{прийняття } H_0,\ \textit{якщо } H_0 \textit{ невірна}) =$$
$$= p(\textit{прийняття } H_0,\ \textit{якщо } H_1 \textit{ вірна}) =$$
$$= p(\bar{x} \le 1{,}3878,\ \textit{якщо}\ \mu = 2) =$$
$$= p\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{N}} \le \frac{1{,}3878-2}{0{,}5/\sqrt{9}}\right) = p\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{N}} \le -3{,}6732\right) = 0{,}001.$$

Потужність 1-го правила *Тр*1:

$$Тр1 = p(\textit{відхилення } H_0,\ \textit{якщо вірна } H_1) =$$
$$= 1 - p(\textit{прийняття } H_0,\ \textit{якщо вірна } H_1) =$$
$$= 1-\beta = 1-0{,}001 = 0{,}999.$$

Для правила 2:

$$\alpha = p(\textit{відхилення } H_0,\ \textit{якщо } H_0 \textit{ вірна}) =$$
$$= p(\bar{x} < 0{,}6122,\ \textit{якщо}\ \mu = 1) =$$
$$= p\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{N}} < \frac{0{,}6122-1}{0{,}5/\sqrt{9}}\right) = p\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{N}} < -2{,}3268\right) = 0{,}01.$$
$$\beta = p(\textit{прийняття } H_0,\ \textit{якщо } H_1 \textit{ вірна}) =$$
$$= p(\bar{x} \ge 0{,}6122,\ \textit{якщо}\ \mu = 2) =$$
$$= p\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{N}} \ge \frac{0{,}6122-2}{0{,}5/\sqrt{9}}\right) = p\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{N}} \ge -8{,}3268\right) = 1{,}0.$$

Потужність 2-го правила:

$$Тр2 = 1 - p(\textit{прийняття } H_0,\ \textit{якщо вірна } H_1) =$$
$$= 1-\beta = 1-1 = 0.$$

Для правила 3:

$$\alpha = p(\bar{x} > 2 \ \textit{ або }\ \bar{x} < 0{,}6122,\ \textit{якщо}\ \mu = 1) =$$
$$= p(\bar{x} > 2,\ \textit{якщо}\ \mu = 1) + p(\bar{x} < 0{,}6122,\ \textit{якщо}\ \mu = 1) = 0{,}01;$$
$$\beta = p(0{,}6122 \le \bar{x} \le 2,\ \textit{якщо}\ \mu = 2) = 0{,}5.$$

Потужність 3-го правила: $Тр3 = 1-\beta = 1-0{,}5 = 0{,}5.$

Таким чином, всі три правила прийняття рішень характеризуються однаковою ймовірністю відхилення нуль-гіпотези, коли вона вірна. Ймовірності прийняття нуль-гіпотези, якщо вірна альтернативна, складають 0,001; 1,0 і 0,5, відповідно. Потужності ППР (*ймовірності відхилення нуль-гіпотез, якщо вірна альтернативна*) дорівнюють: 0,999; 0,0 і 0,5. Тобто з трьох

правил перше характеризується найвищою ймовірністю прийняття правильного рішення; воно має найбільшу потужність і є, відповідно, найкращим з трьох наведених. Формальною мовою статистичного висновку найкраще правило (тест) визначається наступним чином.

Означення 6.5. При тестуванні простих нуль-гіпотез проти простих альтернативних гіпотез правило, що характеризується ймовірністю помилки 1-го роду α, буде *найкращим серед всіх правил*, які характеризуються ймовірністю помилки 1-го роду α, що має найбільшу потужність (іншими словами, воно має найменше β).

Примітка: Означення 6.5 передбачає, що всі правила прийняття рішень ґрунтуються на випадкових вибірках однакового розміру N.

Порівняння правил прийняття рішень не завжди виконується просто. Розглянемо приклад перевірки гіпотез та порівняння ППР у випадку простої нуль-гіпотези і композиційної альтернативної.

Приклад 6.7. Нехай необхідно тестувати гіпотези:

$$H_0: \mu = 0 \quad \textit{проти} \quad H_1: \mu \neq 0,$$

де μ – середнє нормальної генеральної сукупності з дисперсією $\sigma^2 = 4$. Для наведених нижче ППР необхідно знайти: α, потужність правила при $\mu = 0{,}5$ і потужність правила при $\mu = -0{,}5$. Позначимо вибіркове середнє через $\bar{x}$ для випадкової вибірки розміром $N = 25$ з генеральної сукупності.

– Правило 1: *відхилити* H_0, *якщо* $|\bar{x}| > 0{,}784$.
– Правило 2: *відхилити* H_0, *якщо* $\bar{x} > 0{,}658$.
– Правило 2: *відхилити* H_0, *якщо* $\bar{x} < -0{,}658$.

Розв'язок

Для правила 1:

$\alpha = 0{,}05;$
$Tp1(\textit{при}\ \mu = 0{,}5) = 0{,}239;$
$Tp1(\textit{при}\ \mu = -0{,}5) = 0{,}239.$

Для правила 2:

$\alpha = 0{,}05;$
$Tp2(\textit{при}\ \mu = 0{,}5) = 0{,}347;$
$Tp2(\textit{при}\ \mu = -0{,}5) = 0{,}002.$

Для правила 3:

$\alpha = 0{,}05;$
$Tp3(\textit{при}\ \mu = 0{,}5) = 0{,}002;$
$Tp3(\textit{при}\ \mu = -0{,}5) = 0{,}347;$

Всі три ППР мають однакову ймовірність відхилення нуль-гіпотези, коли вона вірна ($\alpha = 0{,}05$). При $\mu = 0{,}5$ потужності ППР мають такі значення: 0,239; 0,347 і 0,02. Таким чином, якби ми перевіряли нуль-гіпотезу $\mu = 0$ проти альтернативної $\mu = 0{,}5$, то найкращим було б правило 2, а на другому місці – перше правило.

Однак, якби ми перевіряли нуль-гіпотезу $\mu = 0$ проти альтернативної $\mu = -0{,}5$, то найкращим було б 3-є ППР, на другому місці – 1-е, а 2-е було б найгіршим (відповідні потужності: 0,239; 0,02 і 0,347).

Оскільки конкретне значення потужності характеризує правило тільки при конкретному значенні тестової статистики, то для отримання повної оцінки ППР необхідно обчислити його потужність для кожного можливого значення статистики. Це приводить до визначення *операційної характеристичної кривої* ППР та його *функції потужності*.

Означення 6.6. Розглянемо задачу перевірки простої нуль-гіпотези, що стосується параметра, проти альтернативної композиційної гіпотези стосовно того ж параметра.

Нехай $\beta(\theta)$ – ймовірність прийняття нуль-гіпотези, якщо значення параметра дорівнює θ. Значення $\beta(\theta)$ для кожного значення θ називають *операційною характеристикою* ППР, а криву, яка визначається функцією $\beta(\theta)$, називають *операційною характеристичною кривою* ППР.

Примітка: у випадку перевірки простої нуль-гіпотези $(\theta = \theta_0)$ проти простої альтернативної гіпотези $(\theta = \theta_1)$ функція $\beta(\theta)$ спрощується до $\beta(\theta_1) = \beta$, де β – ймовірність помилки 2-го роду.

Означення 6.7. Розглянемо перевірку простої нуль-гіпотези стосовно параметра проти композиційної альтернативної гіпотези стосовно того ж параметра.

Нехай $\pi(\theta)$ – ймовірність відхилення нуль-гіпотези, якщо значенням параметра є θ. Для кожного конкретного значення параметра θ значення $\pi(\theta)$ називають потужністю тесту. Множина значень $\pi(\theta)$ утворює *функцію потужності* тесту. Функція потужності та операційна характеристична крива тесту зв'язані виразом:

$$\pi(\theta) = 1 - \beta(\theta).$$

Примітка: якщо нуль-гіпотезою є $H_0: \theta = \theta_0$, то $\pi(\theta_0) = \alpha$, де α – ймовірність помилки 1-го роду.

Операційні характеристичні криві та функції потужності використовують для оцінювання ППР щодо простих нуль-гіпотез проти композиційних альтернативних гіпотез. Наприклад, якщо перевіряють гіпотези

$$H_0: \theta = \theta_0 \quad \textit{проти} \quad H_1: \theta \neq \theta_1,$$

то бажано мати високу ймовірність прийняття нуль-гіпотези, якщо $\theta = \theta_0$, і низьку ймовірність прийняття гіпотези, якщо $\theta \neq \theta_0$. Тобто, бажано мати велике значення для $\beta(\theta_0)$, але якщо $\theta \neq \theta_0$, то бажано мати малі значення для $\beta(\theta)$. Іншими словами, бажаним тестом (ППР) є такий, операційна характеристична крива якого має високе значення при $\theta = \theta_0$ і низькі значення при інших значеннях θ.

6.12 Контрольні запитання і вправи

1. Поясніть помилки першого і другого роду, які можуть виникати при прийнятті рішень?
2. Чи можна одночасно зменшувати ризик припуститись помилки одного роду без одночасного збільшення помилки іншого роду?
3. Який результат експерименту є цілком достатнім для відхилення як неправильної висунутої гіпотези?
4. Яку мету стосовно перевірки гіпотез необхідно ставити при постановці експерименту?
5. Яку величину називають рівнем значущості експерименту?
6. Сформулюйте загальну багатокрокову процедуру перевірки гіпотез і продемонструйте її застосування на прикладі.
7. Поясніть величини, що входять у вираз для обчислення z – оцінок результату експерименту для вибірки дихотомічної змінної і наведіть приклад оцінювання:

8. $$z = \frac{оцінка\ відхилення}{стандартне\ відхилення} = \frac{r - N\,p}{\sqrt{N\,p(1-p)}}.$$

9. До змінних якого характеру відноситься перевірка гіпотез у задачах першого типу.
10. При перевірці гіпотез які задачі відносять до другого типу? Наведіть ілюстративний приклад.
11. Наведіть приклад задачі третього типу при перевірці гіпотез.
12. Поясніть суть використання одностороннього критерію при перевірці гіпотез? Скористайтесь графіком кривої нормального розподілу.
13. У чому полягає суть використання двостороннього критерію при перевірці гіпотез? Поясніть на прикладі.
14. Строк служби (випадкова змінна X) мобільних телефонів, виготовлених за стандартною технологією, є нормальною випадковою величиною з параметрами розподілу: $\{X\} \sim N\{5000, (100)^2\}$. Після впровадження нової технології необхідно експериментально встановити чи будуть мати нові панелі довший середній строк служби ніж старі?
15. Які гіпотези називають простими і які композиційними? Наведіть приклади обох видів гіпотез.
16. Сформулюйте трикроковий метод побудови процедури перевірки гіпотез?

17. Припустимо, що середнє генеральної сукупності може приймати одне з двох значень: 0,5 або 1, але в обох випадках стандартне відхилення складає $\sigma = 0,25$. Використовуючи наведену вище трикрокову процедуру, необхідно побудувати процедуру перевірки гіпотез:

18. $H_0 : \mu = 0,5$ *проти* $H_1 : \mu = 1.$

19. Для цього скористатись випадковою вибіркою потужністю $N = 20$ та ймовірністю помилки першого роду $\alpha = 0,05$.
20. Побудуйте трикрокову процедуру перевірки гіпотез (за аналогією з попереднім прикладом) для перевірки таких гіпотез:

21. $H_0 : \mu = \mu_0$ *проти* $H_1 : \mu < \mu_0$,

22. де μ – невідоме середнє нормальної генеральної сукупності із стандартним відхиленням σ; μ_0 – константа.
23. А) Побудуйте процедуру перевірки гіпотез, яка ґрунтується на випадковій вибірці розміром 100 при ймовірності зробити помилку 1-го роду $\alpha = 0,05$.
24. Б) Побудуйте процедуру перевірки гіпотез, яка ґрунтується на випадковій вибірці розміром n при ймовірності зробити помилку 1-го роду α.
25. Випадкова вибірка стаканчиків з соком ($N = 50$), взята з торговельного автомата, має середній вміст соку $\bar{x} = 98$ грамів на чашку. Скористайтесь тестом, розробленим у попередньому прикладі, для перевірки нуль-гіпотези, що середнє генеральної сукупності для цього автомата складає $\mu = 100$ проти альтернативної гіпотези: $\mu < 100$ (рівень значущості $\alpha = 0,05$). Припустимо, що маса соку в стаканчиках має нормальний розподіл з дисперсією $\sigma^2 = 3$.
26. Сформулюйте означення потужності тесту?
27. Сформулюйте потужність правила перевірки гіпотез?
28. Яке правило, що характеризується ймовірністю помилки 1-го роду α, при тестуванні простих нуль-гіпотез проти простих альтернативних гіпотез, буде *найкращим серед всіх правил*, які характеризуються ймовірністю помилки 1-го роду α?
29. Дайте означення функції потужності тесту? Яким виразом зв'язані між собою функція потужності тесту і характеристична крива?
30. Наведіть приклади застосування функції потужності тесту і характеристичної кривої?

Розділ 7

ДЕЯКІ СТАНДАРТНІ ПРОЦЕДУРИ ПЕРЕВІРКИ ГІПОТЕЗ ТА ІНТЕРВАЛЬНЕ ОЦІНЮВАННЯ

7.1 Перевірка гіпотез щодо середніх

На практиці досить часто виникають задачі перевірки середніх у випадках, коли необхідно порівняти характеристики вибраної групи елементів із деякими прийнятими або стандартизованими характеристиками. Подібні приклади були розглянуті в попередньому розділі. Вони стосувались середнього строку служби плазмових панелей, виготовлених за новою технологією; середнього об'єму кави, що розливається в стаканчики торговельним автоматом, тощо.

Перевірка гіпотези щодо середнього нормальної сукупності з відомою дисперсією

Нехай $\Re -$ генеральна сукупність з нормальним розподілом, який має невідоме середнє μ і відому дисперсію σ^2, тобто: $\{\Re\} \sim (\mu, \sigma^2)$. Стандартна процедура перевірки гіпотези

$$H_0 : \mu = \mu_0 \quad \textit{проти} \quad H_1 : \mu \neq \mu_0$$

ґрунтується на випадковій вибірці даних об'ємом N і на рівні значимості α. За тестову статистику вибирають випадкову змінну:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{N}},$$

яка має стандартний нормальний розподіл при $\mu = \mu_0$ (якщо $\mu \neq \mu_0$, то розподіл буде нормальний але не стандартний).

Правило прийняття рішень можна записати так:

$$-\textit{відхилити} \quad H_0, \quad \textit{якщо} \quad |z| > z_{\alpha/2}.$$

Примітка: значення z_α визначається з рівняння:

$$\frac{1}{\sqrt{2\pi}} \int_{z_\alpha}^{\infty} e^{-(x^2/2)} dx = \alpha \ .$$

Іншими словами, $z_\alpha -$ це значення на осі абсцис таке, що площа справа від нього (під кривою стандартного нормального розподілу) дорівнює α.

Якщо альтернативну гіпотезу змінити на $\mu > \mu_0$ або на $\mu < \mu_0$, то належним чином змінюється і правило прийняття рішення:

відхилити H_0, *якщо* $z > z_\alpha$ і *відхилити* H_0, *якщо* $z < z_\alpha$, відповідно.

Розглянуту процедуру перевірки гіпотез можна резюмувати наступним чином (табл. 7.1):

Таблиця 7.1. Стандартна процедура перевірки гіпотез щодо середнього генеральної сукупності з відомою дисперсією

<table>
<tr><td colspan="2">Дана нормальна генеральна сукупність з відомою дисперсією σ^2.
$H_0: \mu = \mu_0$; рівень значимості α.
Тестова статистика: $z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{N}}$ (має стандартний нормальний розподіл, якщо $\mu = \mu_0$).</td></tr>
<tr><td>Альтернативні гіпотези
H_1</td><td>Правило прийняття рішень:
– відхилити H_0, якщо</td></tr>
<tr><td>$\mu \neq \mu_0$</td><td>$\lvert z \rvert > z_{\alpha/2}$</td></tr>
<tr><td>$\mu > \mu_0$</td><td>$z > z_\alpha$</td></tr>
<tr><td>$\mu < \mu_0$</td><td>$z < -z_\alpha$</td></tr>
</table>

Приклад 7.1. Виробник корму для бройлерів запевняє керівника птахофабрики, що його новий корм дешевший, але не менш ефективний ніж той, що використовується зараз. Керівник птахофабрики знає, що при використанні попереднього корму маса птахів віком до 3-х місяців має нормальний розподіл з середнім $\mu = 1{,}3$ кг і стандартним відхиленням $\sigma = 0{,}1$ кг.

Для того, щоб перевірити запевнення виробника корму взяли випадкову вибірку з 10 новонароджених курчат і годували їх новим кормом протягом 3-х місяців. На кінець 3-го місяця встановлено, що середня вибіркова маса 10 курчат складає: $\bar{x} = 1{,}2$ кг. Побудуйте процедуру перевірки гіпотез на рівні значущості $\alpha = 0{,}05$ з метою перевірки запевнення виробника корму за припущення, що вибіркове стандартне відхилення також складає 0,1 кг ($\sigma_в = 0{,}1$).

Розв'язок.

1. *Формулювання гіпотез*. Позначимо через μ середню масу птахів після трьох місяців годування новим кормом. Оскільки для 10 птахів вибіркове середнє $\bar{x} = 1{,}2$ кг, то виникає сумнів щодо того, що середнє генеральної сукупності буде $\mu = 1{,}3$ кг. Для перевірки запевнення виробника корму можна сформулювати наступні гіпотези:

$$H_0: \mu = 1{,}3 \quad \textit{проти} \quad H_1: \mu < 1{,}3.$$

2. *Вибір статистики і формулювання правила прийняття рішення*. Ми тестуємо середнє нормально розподіленої генеральної сукупності, яка має дисперсію $\sigma^2 = (0{,}1)^2 = 0{,}01$. Перевірочна статистика:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{N}} = \frac{\bar{x} - 1{,}3}{0{,}1 / \sqrt{10}}.$$

Правило прийняття рішення:

$$-\textit{відхилити}\ \ H_0,\ \textit{якщо}\ \ z<-z_\alpha.$$

3. *Застосування ППР до даних*. Дані для цього прикладу:

$$\bar{x}=1{,}2;\quad \mu_0=1{,}2;\quad \sigma=0{,}1;\quad N=10;\quad \alpha=0{,}05;\quad -z_\alpha=-1{,}645.$$

Спостережуване значення тестової статистики z:

$$z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{N}}=\frac{1{,}2-1{,}3}{0{,}1/\sqrt{10}}=-3{,}162.$$

Застосування правила прийняття рішень: оскільки

$$z=-3{,}162<-1{,}645=z_\alpha,$$

то H_0 необхідно *відхилити*. Таким чином, на рівні значимості $\alpha=0{,}05$ керівник птахофабрики повинен відхилити запевнення виробника корму в тому, що новий корм є таким же ефективним, як і попередній, тобто, він є менш ефективним.

Приклад 7.2. Розглянемо вихідні дані, наведені у попередньому прикладі. На основі отриманого результату не можна стверджувати, що стандартне відхилення σ маси бройлерів, відгодованих новим кормом, залишається таким же, як і σ маси бройлерів, відгодованих старим кормом. Чи можна розв'язати цю задачу без зробленого припущення?

Розв'язок.

Хоча на птахофабриці не знають стандартне відхилення для генеральної сукупності бройлерів, відгодованих новим кормом, але їм відомо, що $\sigma=0{,}1$ у випадку, якщо запевнення виробника корму відповідають дійсності. Тому, якщо нуль-гіпотезу попереднього прикладу розширити таким чином, щоб врахувати запевнення виробника корму щодо σ, то H_0 приймає таку форму:

$$H_0:\ \mu=1{,}3\ i\ \sigma=0{,}1\quad \textit{проти}\quad H_1:\ \mu<1{,}3.$$

В такому випадку, якщо нуль-гіпотеза вірна, то перевірочна статистика

$$z=\frac{\bar{x}-1{,}3}{0{,}1/\sqrt{10}}$$

є стандартною нормальною випадковою змінною $\dfrac{\bar{x}-\mu_0}{\sigma/\sqrt{N}}$, а це означає, що ми можемо скористатись в даному випадку тією ж тестовою статистикою, що і в попередньому прикладі і розв'язок буде ідентичним попередньому прикладу.

Підсумовуючи можна сказати, що навіть у випадку, коли значення σ відоме тільки при вірній нуль-гіпотезі, можна скористатись таблицею 7.1 при умові, що H_0 включає твердження відносно σ.

Перевірки гіпотези щодо середнього нормальної сукупності у випадку невідомої дисперсії

Нехай $\Re$ – генеральна сукупність з відомим середнім μ і невідомою дисперсією. Якщо сформулювати нуль-гіпотезу як $H_0:\ \mu=\mu_0$, то стандартна

процедури перевірки буде ґрунтуватись на використанні статистики:

$$t = \frac{\bar{x} - \mu_0}{S / \sqrt{N}},$$

де випадкова змінна t буде мати $t-$розподіл з $N-1$ ступенями свободи при умові, що $\mu = \mu_0$. Відомо, що при $N > 29$ $t-$розподіл з N ступенями свободи можна апроксимувати стандартним нормальним розподілом. Процедура перевірки гіпотези щодо середнього нормальної генеральної сукупності з невідомою дисперсією сформульована в табл. 7.2.

Таблиця 7.2. Стандартна процедура перевірки гіпотез щодо середнього генеральної сукупності з невідомою дисперсією

<table>
<tr><td colspan="2">Дана нормальна генеральна сукупність з невідомою дисперсією σ^2.
$H_0: \mu = \mu_0$; рівень значимості α.
Тестова статистика: $t = \frac{\bar{x} - \mu_0}{S / \sqrt{N}}$ (має $t-$розподіл з $N-1$ ступенями свободи, якщо $\mu = \mu_0$).</td></tr>
<tr><td>Альтернативні гіпотези
H_1
$\mu \neq \mu_0$
$\mu > \mu_0$
$\mu < \mu_0$</td><td>Правило прийняття рішень:
– відхилити H_0, якщо
$\lvert t \rvert > t_{\alpha/2, N-1}$
$t > t_{\alpha, N-1}$
$t < -t_{\alpha, N-1}$</td></tr>
</table>

Приклад 7.3. Власник ресторану продає свій ресторан і запевняє потенційного покупця, що середня кількість відвідувачів по суботах (без врахування святкових днів) складає 100 ($\mu = 100$). Для того, щоб перевірити запевнення власника ресторану, потенційний покупець підрахував число відвідувачів по 9-и випадково вибраних суботах. В результаті він встановив, що вибіркове середнє $\bar{x} = 95$, а вибіркове стандартне відхилення складає $S = 10$. Необхідно встановити:

а) Чи повинен відхилити потенційний покупець запевнення власника ресторану на рівні значимості $\alpha = 0{,}05$? (За припущення, що розподіл кількості відвідувачів є нормальним.)

б) Яка буде відповідь при $S = 4$?

в) Якщо на запитання **а) і б)** ви дали різні відповіді, то поясніть різницю?

Розв'язок.

Завдання **а**)

1. *Формулювання гіпотез*. Оскільки $\bar{x} = 95$, то потенційний покупець має підозру, що фактичне середнє число відвідувачів $\mu < 100$ (для генеральної сукупності). Тому логічно сформулювати такі гіпотези:

$$H_0: \mu = 100 \quad \textit{проти} \quad H_1: \mu < 100.$$

2. *Вибір перевірочної статистики і формулювання ППР*. В даному прикладі виконується перевірка середнього нормальної сукупності елементів з невідомою дисперсією, а тому необхідно скористатись табл. 9.2. Тестова статистика:

$$t = \frac{\bar{x} - \mu_0}{S / \sqrt{N}}$$

і правило прийняття рішення:

$$- \textit{відхилити } H_0, \textit{ якщо } t < t_{\alpha, N-1}.$$

3. *Застосування ППР до даних*. В даному випадку дані мають вигляд:

$$\bar{x} = 95; \quad \mu_0 = 100; \quad S = 10; \quad N = 9; \quad \alpha = 0{,}05;$$
$$-t_{\alpha, N-1} = -t_{0{,}05;8} = -1{,}860.$$

Спостережуване значення перевірочної статистики:

$$t = \frac{\bar{x} - \mu_0}{S / \sqrt{N}} = \frac{95 - 100}{10 / \sqrt{9}} = -1{,}5.$$

Оскільки $t = -1{,}5$, а $-t_{\alpha, N-1} = -1{,}86$, тобто $t > -t_{\alpha, N-1}$, то H_0 не може бути відхилена. Таким чином, потенційний покупець не може відхилити твердження власника ресторану на рівні значимості $\alpha = 0{,}05$ відносно того, що середнє число відвідувачів по суботах складає 100.

Завдання **б)**

Якщо $S = 4$, то тестова статистика приймає значення:

$$t = \frac{\bar{x} - \mu_0}{S / \sqrt{N}} = \frac{95 - 100}{4 / \sqrt{9}} = -3{,}75.$$

Оскільки $t = -3{,}75 < -1{,}86 = -t_{\alpha, N-1}$, то H_0 необхідно відхилити на рівні значимості $\alpha = 0{,}05$.

Завдання **в**) Вибіркові стандартні відхилення в пунктах завданнь а) і б) складали $S = 10$ і $S = 4$, відповідно. Це свідчить про вищу ступінь розсіювання числа відвідувачів у випадку а) ніж у випадку б). Тому, припускаючи, що $\mu = 100$, спостережуване значення $\bar{x} = 95$ ймовірніше буде отримане у випадку а).

Перевірка гіпотези щодо середнього генеральної сукупності на основі вибірок великої потужності

Спочатку розглянемо випадок коли дисперсія σ^2 відома. Нехай $\Re$ – генеральна сукупність з невідомим розподілом (тобто, не обов'язково нормальним). Позначимо через μ і σ^2 середнє і дисперсію генеральної сукупності.

Припустимо, що необхідно перевірити нуль-гіпотезу $H_0 : \mu = \mu_0$ проти деякої належної альтернативної.

Якщо відоме σ^2, і якщо потужність вибірки N велика, то можна скористатись тестовою статистикою:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{N}}.$$

Зазначимо, що при $\mu = \mu_0$ і при великому N випадкова змінна z наближено має стандартний нормальний розподіл. Процедура перевірки гіпотез для даного випадку (велика вибірка з відомою дисперсією) наведена в табл. 7.3.

Таблиця 7.3. Процедура перевірки гіпотез щодо середнього генеральної сукупності з відомою дисперсією на основі великої вибірки

<table>
<tr><td colspan="2">Дана будь-яка генеральна сукупність; відома дисперсія σ^2;
$H_0: \mu = \mu_0$; велика вибірка даних N; рівень значимості α.
Тестова статистика: $z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{N}}$ (наближено має стандартний нормальний розподіл, якщо $\mu = \mu_0$).</td></tr>
<tr><td>Альтернативні гіпотези
H_1</td><td>Правило прийняття рішень:
– відхилити H_0, якщо</td></tr>
<tr><td>$\mu \neq \mu_0$</td><td>$\left| z \right| > z_{\alpha/2}$</td></tr>
<tr><td>$\mu > \mu_0$</td><td>$z > z_\alpha$</td></tr>
<tr><td>$\mu < \mu_0$</td><td>$z < -z_\alpha$</td></tr>
</table>

Якщо дисперсія σ^2 генеральної сукупності невідома, то необхідно скористатись тестовою статистикою:

$$z = \frac{\bar{x} - \mu_0}{S / \sqrt{N}}$$

при великому значенні N. Якщо N велике, то z має наближено стандартний нормальний розподіл. Процедура перевірки гіпотез для даного випадку (велика вибірка з невідомою дисперсією) наведена в табл. 7.4.

Таблиця 7.4. Процедура перевірки гіпотез щодо середнього генеральної сукупності з невідомою дисперсією на основі великої вибірки

<table>
<tr><td colspan="2">Дана будь-яка генеральна сукупність; дисперсія σ^2 невідома;
$H_0: \mu = \mu_0$; велика вибірка даних N; рівень значимості α.
Тестова статистика: $z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{N}}$ (наближено має стандартний нормальний розподіл, якщо $\mu = \mu_0$).</td></tr>
<tr><td>Альтернативні гіпотези
H_1</td><td>Правило прийняття рішень:
– відхилити H_0, якщо</td></tr>
<tr><td>$\mu \neq \mu_0$</td><td>$\left| z \right| > z_{\alpha/2}$</td></tr>
<tr><td>$\mu > \mu_0$</td><td>$z > z_\alpha$</td></tr>
<tr><td>$\mu < \mu_0$</td><td>$z < -z_\alpha$</td></tr>
</table>

Приклад 7.4. В обласне управління освіти надійшла інформація, що діти перших класів проводять біля телевізора 15 годин на тиждень. Однак, з іншого джерела надійшла інформація, що в деяких районах це число годин є меншим. Для перевірки цієї гіпотези були випадково вибрані прізвища 49 дітей перших класів і їх батьки надали інформацію щодо того, скільки часу проводять їхні діти біля телевізорів.

а) Нехай вибіркове середнє значення часу, який діти проводять біля телевізорів, складає $\bar{x} = 10$ годин, а вибіркове стандартне відхилення – $S = 5$ годин. Чи повинен відділ освіти відхилити $H_0 : \mu = 15$ на рівні значимості $\alpha = 0{,}05$ і прийняти альтернативну: $H_1 : \mu < 15$?

б) Чи можна на основі вказаних вище вибіркових даних (на рівні значимості $\alpha = 0{,}05$) стверджувати, що діти перших класів (для досліджуваного району) в середньому проводять біля телевізора менше 11 годин на тиждень?

Розв'язок.

1. *Формулювання гіпотез*

$$H_0 : \mu = 15 \quad \textit{проти} \quad H_1 < 15.$$

2. *Вибір тестової статистики і формулювання правила прийняття рішення*. Хоча дисперсія σ^2 невідома, але вибірка досить велика і тому можна скористатись процедурою перевірки, наведеною в таблиці 7.4. Тестова статистика:

$$z = \frac{\bar{x} - \mu_0}{S / \sqrt{N}}.$$

Правило прийняття рішення:

$$- \textit{відхилити} \quad H_0, \ \textit{якщо} \quad z < -z_\alpha.$$

3. *Застосування правила прийняття рішення до даних*.

Дані: $\bar{x} = 10; \quad \mu_0 = 15; \quad S = 5; \quad N = 49; \quad \alpha = 0{,}05; \quad -z_\alpha = -1{,}645.$

Спостережуване значення тестової статистики:

$$z = \frac{\bar{x} - \mu_0}{S / \sqrt{N}} = \frac{10 - 15}{5 / \sqrt{49}} = -7.$$

Оскільки

$$z = -7 < -1.645 = -z_\alpha,$$

то згідно з правилом прийняття рішення H_0 необхідно відхилити, тобто приймається альтернативна гіпотеза: $\mu < 15$.

б) Гіпотези:

$$H_0 : \mu = 11 \quad \textit{проти} \quad H_1 < 11.$$

Правило прийняття рішення:

$$- \textit{відхилити} \quad H_0, \ \textit{якщо} \quad z < -z_\alpha.$$

В даному випадку:

$$z = \frac{\bar{x} - \mu_0}{S / \sqrt{N}} = \frac{10 - 11}{5 / \sqrt{49}} = -1{,}4 \quad \text{і} \quad -z_\alpha = -z_{0{,}5} = -1{,}645.$$

Оскільки $z = -1{,}4 > -1{,}645 = -z_{\alpha}$, то H_0 не може бути відхилена. Тобто не можна відхилити нуль-гіпотезу, що діти перших класів проводять (у вибраному районі) в середньому біля телевізорів 11 годин на тиждень.

Приклад 7.5. Нехай $\Re -$ генеральна сукупність, яка має нормальний розподіл з невідомим середнім μ. Перевірте наступні гіпотези для значень α, величини вибірки N, вибіркового середнього $\bar{x}$ та вибіркового стандартного відхилення S, наведених нижче.

а) $\alpha = 0{,}05;\ N = 9;\ \bar{x} = 22{,}2;\ S = 3{,}3;$

$$H_0 :\ \mu = 20 \quad проти \quad H_1 :\ \mu \neq 20.$$

б) $\alpha = 0{,}05;\ N = 9;\ \bar{x} = 22{,}2;\ S = 3{,}3;$

$$H_0 :\ \mu = 20 \quad проти \quad H_1 :\ \mu > 20.$$

в) $\alpha = 0{,}01;\ N = 16;\ \bar{x} = 3{,}9;\ S = 1{,}6;$

$$H_0 :\ \mu = 5 \quad проти \quad H_1 :\ \mu < 5.$$

г) $\alpha = 0{,}01;\ N = 16;\ \bar{x} = 3{,}9;\ S = 1{,}6;$

$$H_0 :\ \mu = 5 \quad проти \quad H_1 :\ \mu \neq 5.$$

Розв'язок.

а) Тестова статистика:

$$t = \frac{\bar{x} - \mu}{S/\sqrt{N}} = \frac{22{,}2 - 20}{3{,}3/\sqrt{9}} = \frac{2{,}2}{1{,}1} = 2{,}0.$$

Оскільки $\alpha = 0{,}05$, то $\alpha/2 = 0{,}025$; звідси $t_{0{,}025;\,8} = 2{,}306$ при $N = 9$. Таким чином, $t = 2{,}0 < 2{,}306 = z_{\alpha/2}$, тобто, H_0 не можна відхилити.

б) В даному випадку ми тестуємо гіпотезу:

$$H_0 :\ \mu = 20 \quad проти \quad H_1 :\ \mu > 20$$

На рівні значимості $\alpha = 0{,}05$ тестова статистика: $t_{0{,}05;\,8} = 1{,}86$.

Оскільки $z > z_{\alpha}$, то H_0 відхиляється і приймається альтернативна гіпотеза: $H_1 :\ \mu > 20$.

в) Тестова статистика:

$$t = \frac{\bar{x} - \mu}{S/\sqrt{N}} = \frac{3{,}9 - 5}{1{,}6/\sqrt{16}} = -\frac{1{,}1}{0{,}4} = -2{,}75,$$

а критичне значення $t_{0{,}01;\,15} = -2{,}602$.

Оскільки $t = -2{,}75 < -2{,}602 = t_{\alpha,\,N-1}$, то H_0 необхідно відхилити і прийняти альтернативну.

г) У даному випадку $t = -2{,}75;\ t_{0{,}005;\,15} = 2{,}947$, то H_0 приймається.

7.2 Поняття інтервальних оцінок

Вище (у третьому розділі) розглянуто, як за допомогою випадкової вибірки даних оцінити середнє і дисперсію вихідного (генерального) розподілу. Однак, навіть найкраща оцінка може відрізнятись від істинного значення статистичного параметра. Тому на практиці бажано знати, наскільки обчислена оцінка є близькою до істинного значення параметра. В зв'язку з цим введемо деякі поняття, які дозволяють виносити судження щодо точності оцінок з імовірнісної точки зору.

Оцінюючи статистичний параметр (наприклад, середнє розподілу) не можна говорити про ймовірність того, що оцінка є точною. Наприклад, нехай на основі деякої вибірки стверджують, що середній ріст чоловіків в Україні складає 175 см. Очевидно, що таке твердження необхідно перевірити за допомогою даних щодо зросту чоловіків по всій країні. Тобто, без такої перевірки ми не можемо говорити про ймовірність того, що оцінка є точною.

На практиці замість однієї (точкової) оцінки розглядають деякий інтервал і визначають ймовірність того, що значення невідомого параметра буде лежати між граничними точками цього інтервалу. Такі оцінки називають *інтервальними*. Як правило, спочатку вибирають бажану ймовірність попадання в інтервал (наприклад 0,95 або 0,99), а потім знаходять два значення (точки) між якими буде знаходитись значення невідомого параметра із вибраною ймовірністю. Твердженням подібного типу може бути наступне: середній зріст дорослого чоловіка в США знаходиться в інтервалі 173-175 см з імовірністю 0,95.

Деякі статистики говорять, що застосування слова "ймовірність" є в даному випадку некоректним, оскільки параметр – конкретне число і говорити про те, що воно міститься в деякому інтервалі немає сенсу. Вони вважають, що поняття "ймовірність" можна застосовувати тільки до випадкових величин, а тому часто дають визначення ймовірності, яке відрізняється за змістом від того, що використовується тут. Нижче наведемо визначення ймовірності, яке є коректним не у всіх ситуаціях, але є цілком прийнятним для подальшого викладення.

Раніше ймовірність випадкової події визначалась як частота її появи, коли дослід повторюється багато разів. Іншими словами, ймовірність падіння монети визначеною стороною догори є пропорційною числу появи цього боку зверху з зростанням числа підкидань.

Таким чином, що ж означає твердження, що зріст дорослих чоловіків коливається в межах 173-175 см з імовірністю 0,95. Зокрема, що в даному випадку є тим результатом, що нас цікавить і які події можуть повторюватись багато разів? Зазначимо, що ми маємо справу з однією випадковою вибіркою та інтервалом. Ми прогнозуємо, що невідомий параметр знаходиться в даному інтервалі.

Можна навести ще одну аналогію тієї ситуації, коли ймовірність відноситься до інтервальної оцінки. Наприклад, на борту військового літака є лише одна бомба, якою необхідно знищити ворожий об'єкт, місцезнаходження якого закрите хмарами. Припустимо, що ударна хвиля діє на відстані 10 кілометрів. Таким чином, льотчик знає, що якщо бомба упаде в радіусі 10 км від місцезнаходження об'єкта, то він буде знищений. Інакше мета не буде досягнута. В даному випадку інтервальній оцінці параметра відповідає діаметр дії ударної хвилі вибуху, а місцеположення об'єкта представляє собою невідомий статистичний параметр. Вибух бомби, який охоплює групу своєю ударною хвилею, подібний інтервальній оцінці, яка включає істинне значення параметра між своїми граничними точками.

Згідно з визначенням, для визначення ймовірності, з якою бомба досягне ціль, необхідно повторити бомбування велике число разів. Припустимо, що в результаті здійснення 100 вильотів льотчик зміг би 95 разів знищити ворожу групу. В такому випадку буде справедливим твердження, що пілот досягне мети за допомогою лише однієї бомби з ймовірністю 0,95. Можна також сказати, що інтервальна оцінка включає невідомий параметр з імовірністю 0,95.

7.3 Інтервальна оцінка середнього значення розподілу

Розглянемо процедуру визначення інтервальної оцінки середнього розподілу. Ця процедура повинна забезпечити включення в інтервал найкращої оцінки середнього значення розподілу. В той же час інтервал навколо цієї найкращої оцінки будується таким чином, щоб істинне значення середнього розподілу знаходилось всередині цього інтервалу з визначеною імовірністю. Наприклад, замість того, щоб говорити, що середнє значення КРР з даного розподілу дорівнює 115, ми можемо сказати, що середнє розподілу повинне бути в інтервалі між 111 і 119 з імовірністю 0,95.

Звичайно, що завжди бажано мати інтервальну оцінку по-можливості вужчою. Ширина інтервальної оцінки – це одна з її характеристик. Наприклад, порівняємо дві наступні інтервальні оцінки:

1 – середня маса людей, які страждають хворобою L, знаходиться в інтервалі 50-80 кілограмів з імовірністю 0,95;

2 – середня маса людей, які страждають хворобою L, знаходиться в інтервалі 60-70 кілограмів з імовірністю 0,95.

В кожному припущенні бере участь однакове імовірнісне відношення. Однак, в другому припущенні інтервал є вужчим, що робить інтервальну оцінку більш визначеною.

Також, чим більшою є імовірність, тим точнішою буде інтервальна оцінка. Порівняємо два таких твердження:

1 – середня тривалість життя собак породи фокстер'єр складає 10-14 років з імовірністю 0,9;

2 – середня тривалість життя собак породи фокстер'єр складає 10-14 років з імовірністю 0,99.

Можна сказати, що друге припущення є більш строгим в ймовірнісному смислі, ніж перше. Тобто другому твердженню можна довіряти більше, ніж першому.

Повернемось до процедури інтервального оцінювання середнього розподілу. Припустимо, що необхідно визначити інтервальну оцінку середнього коефіцієнтів розумового розвитку студентів з вибраної спеціальності. Необхідно визначити границі інтервалу, в який з імовірністю 0,95 входить середнє значення КРР цих студентів. Нехай, відоме стандартне відхилення КРР складає 16 одиниць. Припустимо, що середнє розподілу оцінено на основі вибірки, яка є в нашому розпорядженні, і дорівнює 118.

В даному випадку ми виходимо з того, що є множина дослідників, кожний з яких має вибірку з 64 елементів, на основі якої він оцінює середнє розподілу і обчислює його інтервальну оцінку. Розглянемо процедуру, яка в 95 випадках із 100 дасть для всіх дослідників інтервальні оцінки, які включають невідоме середнє значення розподілу.

Для розв'язання цієї задачі спочатку необхідно зробити деякі узагальнення. Середнє значення КРР невідоме. Стандартне відхилення розподілу відоме і складає 16 одиниць. Всі дослідники обчислюють вибіркове середнє арифметичне на основі вибірок, які є у їх розпорядженні. У відповідності з теоремою про розподіл вибіркових оцінок середнього, вони мають нормальний розподіл. Їх середнє невідоме, але стандартне відхилення відоме: $\sigma_M = 2$, тобто,

$$\sigma_M = \frac{\sigma}{\sqrt{N}} = \frac{16}{\sqrt{64}} = 2,$$

де σ_M – стандартне відхилення для середніх або стандартна похибка оцінок вибіркових середніх.

В розподілі вибіркових середніх кожний член представляє собою оцінку середнього для 64 значень КРР, тобто, кожне вибіркове середнє – це найкраща оцінка середнього для КРР даної групи студентів. Іноді найкращу оцінку називають *точковою оцінкою* для того, щоб відрізнити її від більш невизначеної інтервальної.

Оскільки вибіркові середні мають нормальний розподіл, а стандартне відхилення дорівнює 2, то за допомогою таблиці площ під кривою нормального розподілу знайдемо, що 68 із 100 вибіркових середніх відхиляються від істинного значення менше, ніж на 2 одиниці.

З цієї ж таблиці знайдемо, що для нормального розподілу 95 із 100 його членів відхиляються від середнього менше, ніж на 1,96 стандартного відхилення. Таким чином, 95 із 100 вибіркових середніх знаходяться в області 1,96 стандартного відхилення. Точніше можна сказати так:

Точкові оцінки, знайдені різними дослідниками, відхиляються від невідомого середнього розподілу в обидва боки не більше, ніж на 3,92 стандартного відхилення з ймовірністю 0,95.

Подивимось тепер на результати з точки зору кожного дослідника окремо. Він може сказати: "Більшість моїх колег отримали вибіркове середнє, яке відхиляється від невідомого середнього не більше, ніж на 3,92. Я можу бути одним із 95, які отримали такі оцінки або одним із 5, чиї вибіркові середні відхиляються від невідомого середнього більше, ніж на 3,92. Оскільки в 95 випадків із 100 дослідник попадає в першу категорію, то із ймовірністю 0,95, я також попаду в цю категорію. Іншими словами, моє вибіркове середнє буде відрізнятись від невідомого середнього розподілу не більше, ніж на 3,92 одиниці з імовірністю 0,95."

Припустимо далі, що за результатами всіх дослідників, було обчислене середнє, яке склало 118 одиниць. Тепер можна сказати, що з імовірністю 0,95 вибіркове середнє буде відрізнятись від середнього розподілу не більше, ніж на 3,92 одиниці. Всі вибіркові середні будуть знаходитись в області $118 \pm 3{,}92$ з імовірністю 0,95. Іншими словами, середнє значення КРР студентів знаходиться між 114,08 і 121,92 з імовірністю 0,95. Таким чином, ми побудували інтервальну оцінку, яка містить всі числові значення в області $\pm 3{,}92$ від отриманої вибіркової оцінки середнього. Ця інтервальна оцінка називається *95%-м довірчим інтервалом* для середнього.

Зазначимо, що процедуру оцінювання можна повторювати кожний день і отримувати при цьому нові вибіркові оцінки. Таким чином, центр інтервальної оцінки кожний день буде новим. Можна сказати лише те, що кожні 95 днів із 100 побудований нами інтервал буде містити невідоме середнє розподілу коефіцієнтів розумового розвитку для студентів з вибраної спеціальності.

Можна побудувати також таку інтервальну оцінку, яка буде містити невідоме середнє з імовірністю 0,99. В такому випадку 99 із 100 елементів нормального розподілу відрізняються від його середнього не більше, ніж на 2,58 стандартного відхилення. Таким чином, обчислене вибіркове середнє буде знаходитись в інтервалі $\pm 2{,}58$ стандартного відхилення від істинного значення середнього з імовірністю 0,99. А це означає, що з імовірністю 0,99 вибіркове середнє, яке дорівнює 118, не більше, ніж на 5,16 відрізняється від невідомого середнього розподілу. Насамкінець, можна стверджувати, що невідоме середнє значення КРР студентів вибраної спеціальності знаходиться в інтервалі між 112,84 і 123,16 з імовірністю 0,99. Цей інтервал називають *99%-м довірчим інтервалом*.

Повторимо тепер всі етапи процесу побудови інтервальної оцінки середнього. Для цього необхідно виконати наступне:

1. обчислити стандартне відхилення для середнього;

2. якщо вибрана імовірність попадання оцінки в інтервал складає 0,95, то значення множника для інтервальної оцінки складає 1,96 і вираз для 95%-го

довірчого інтервалу має вигляд:

$$\left[\bar{x}-1{,}96\frac{\sigma}{\sqrt{N}}\right]<\mu<\left[\bar{x}+1{,}96\frac{\sigma}{\sqrt{N}}\right],$$

де $\bar{x}$ – вибіркове середнє; σ – стандартне відхилення розподілу (для всієї вибірки); μ – невідоме середнє розподілу; N – кількість елементів у вибірці; $\sigma/\sqrt{N}$ – стандартне відхилення середнього; для розглянутої вище задачі: $\bar{x}=118,\ \sigma=16,\ N=64$;

3. якщо вибрана імовірність попадання оцінки в інтервал складає 0,99, то значення множника для інтервальної оцінки складає 2,58 і вираз для 99%-го довірчого інтервалу має вигляд:

$$\left[\bar{x}-2{,}58\frac{\sigma}{\sqrt{N}}\right]<\mu<\left[\bar{x}+2{,}58\frac{\sigma}{\sqrt{N}}\right].$$

Таким чином, для розглянутої вище задачі дослідження КРР для студентів вибраної спеціальності можна обчислити наступні інтервальні оцінки:

– 95%-й довірчий інтервал:

$$\left[118-1{,}96\frac{16}{\sqrt{64}}\right]<\mu<\left[118+1{,}96\frac{16}{\sqrt{64}}\right],$$

$$114{,}1<\mu<121{,}9.$$

– 99%-й довірчий інтервал:

$$\left[118-2{,}58\frac{16}{\sqrt{64}}\right]<\mu<\left[118+2{,}58\frac{16}{\sqrt{64}}\right],$$

$$112{,}2<\mu<123{,}2.$$

Очевидно, що 95%-й довірчий інтервал буде завжди меншим 99%-го довірчого інтервалу. Тобто, чим меншою вимагається імовірнісна точність, тим точнішим може бути твердження. Чим більшою вимагається імовірність того, що побудований інтервал містить невідомий параметр, тим ширшим буде цей інтервал.

7.4 Контрольні запитання і вправи

1. Поясніть стандартну процедуру перевірки гіпотези стосовно середнього генеральної сукупності з відомою дисперсією?
2. Виробник мінеральних добрив стверджує, що його нове мінеральне добриво коштує менше, але не менш ефективне ніж те, що вже використовується. Керівник агрофірми знає, що при використанні попереднього добрива урожайність з одного гектара має нормальний розподіл з

середнім $\mu = 35$ центнерів і стандартним відхиленням $\sigma = 1,0$ центнер. Для того щоб перевірити твердження виробника добрив взяли випадкову вибірку із 20 гектарів і застосували на ній нове добриво. Після збирання врожаю було встановлено, що середня врожайність на 20 га склала $\bar{x} = 34,5$ центнера. Побудуйте процедуру перевірки гіпотез на рівні значущості $\alpha = 0,05$ з метою перевірки твердження виробника міндобрив за припущення, що вибіркове стандартне відхилення також складає 1,0 центнер ($\sigma_в = 1,0$).

3. Опишіть стандартну процедуру перевірки гіпотези стосовно середнього генеральної сукупності з невідомою дисперсією?
4. Поясність процедуру перевірки гіпотези стосовно середнього генеральної сукупності з відомою дисперсією на основі великої вибірки?
5. Опишіть процедуру перевірки гіпотези стосовно середнього генеральної сукупності з невідомою дисперсією на основі великої вибірки?
6. Зібрані статистичні дані свідчать, що діти старших класів проводять біля комп'ютера по 20 годин на тиждень. Однак, з іншого джерела надійшла інформація, що в деяких районах це число годин є меншим. Для перевірки цієї гіпотези були випадково вибрані прізвища 60 дітей
старших класів і їх батьки надали інформацію щодо того, скільки часу проводять їхні діти біля комп'ютерів.

а) Нехай вибіркове середнє значення часу, який діти проводять біля комп'ютерів, складає $\bar{x} = 16$ годин, а вибіркове стандартне відхилення – $S = 6$ годин. Чи можна відхилити $H_0 : \mu = 20$ на рівні значущості $\alpha = 0,05$ і прийняти альтернативну: $H_1 : \mu < 20$?

б) Чи можна на основі вказаних вище вибіркових даних (на рівні значимості $\alpha = 0,05$) стверджувати, що діти старших класів (для досліджуваного району) в середньому проводять біля телевізора менше 18 годин на тиждень?

7. Які оцінки називають точковими і які інтервальними? Наведіть числові приклади.
8. Опишіть процедуру і наведіть приклад обчислення інтервальної оцінки середнього вибірки даних?
9. Яким чином формуються довірчі інтервали? Які довірчі інтервали часто використовуються на практиці?

Розділ 8

ПОБУДОВА РЕГРЕСІЙНИХ МОДЕЛЕЙ

8.1. Методика побудови математичних моделей часових рядів

Розглянемо методику побудови математичних моделей часових рядів, яка більше орієнтована на побудову моделей фінансово-економічних, соціальних та екологічних процесів, для яких, як правило, набагато складніше поставити експеримент (або неможливо взагалі) та отримати інформативні експериментальні дані у достатньому об'ємі. У наведеному вигляді ця методика також може бути успішно застосована і до побудови моделей динаміки технічних систем і технологічних процесів при належному плануванні та реалізації експерименту з метою збору даних, необхідних для побудови високоякісних моделей.

Основи методики побудови моделей і аналізу часових рядів запропоновані Боксом і Дженкінсом [17]. Модифікована авторами методика побудови математичної моделі процесу з використанням даних у вигляді часового ряду і часового перерізу складається з таких кроків.

– Виконання аналізу процесу, для якого будується модель, на основі спеціальних літературних джерел, експертних оцінок перебігу процесу, візуального дослідження графіків вимірів вхідних і вихідних змінних, представлених часовими рядами або часовими перерізами, та іншої доступної інформації.

– Належна попередня обробка експериментальних даних з метою їх приведення до форми, найбільш придатної для оцінювання параметрів (коефіцієнтів) моделі.

– Аналіз часових рядів на можливу наявність нестаціонарності і нелінійності за допомогою множини статистичних критеріїв.

– Вибір (оцінювання) структури моделей-кандидатів, для чого необхідно виконати такі дії: (1) обчислити та виконати аналіз кореляційної матриці для часових рядів залежної та незалежних змінних з метою визначення тих екзогенних змінних, які необхідно включити в модель; (2) обчислити автокореляційну (АКФ) та часткову автокореляційну функції (ЧАКФ) залежної змінної з метою визначення оцінки порядку авторегресійної частини моделі та ковзного середнього; (3) оцінити характеристики інших елементів структури математичної моделі, що буде розглянуто нижче.

– Вибір методу (методів) оцінювання параметрів математичних моделей вибраних структур і обчислення оцінки векторів їх параметрів. Найчастіше це метод найменших квадратів (МНК), метод максимальної правдоподібності (ММП) та їх модифікації (рекурсивні та нелінійні). В окремих випадках

застосовують метод Монте-Карло для марковських ланцюгів (МКМЛ), який придатний для оцінювання нелінійних моделей.

– Вибір кращої з оцінених моделей-кандидатів за допомогою множини статистичних критеріїв адекватності (якості) моделі. Застосувати модель до розв'язання основної задачі – прогнозування, синтезу системи керування або поглибленого дослідження процесу і остаточно встановити її придатність.

Тепер розглянемо докладніше кожний з етапів побудови моделі.

8.2 Аналіз процесу

Аналіз процесу – це надзвичайно важливий етап, коректне виконання якого потребує досвіду дослідження реальних процесів різної природи. Ігнорування цього етапу призводить до неможливості побудови моделі високого ступеня адекватності процесу та її придатності для розв'язання задач, згаданих вище. Аналіз процесу спрямований на вирішення таких завдань:

- визначення кількості входів і виходів, тобто визначення вимірності процесу; як правило, вимірність визначається кількістю виходів процесу, кожний з яких описують окремим рівнянням;
- встановлення логічних зв'язків між змінними та аналіз можливостей їх спільного математичного опису (коректного об'єднання в одному математичному виразі); для цього необхідно використати всю наявну інформацію про процес із спеціальної літератури, наукових звітів та від експертів;
- визначення кількості зовнішніх збурень та їх типу (детерміноване чи стохастичне) та попереднє встановлення можливості їх статистичного опису за допомогою конкретних типів розподілів випадкових величин;
- встановлення можливості декомпозиції процесу на окремі підпроцеси, які є простішими як з точки зору їх функціонування, так і з точки зору математичного опису; декомпозиція – це досить складний процес, який ґрунтується на спеціальних математичних методах;
- якщо процес має ієрархічну структуру (верхній та нижній рівень функціонування або більшу кількість рівнів), то необхідно чітко розмежувати ці рівні, визначити функції кожного з них і встановити, які типи зв'язків існують між ними; наприклад, технологічні процеси часто можна розмежувати на два і більше рівнів, які зв'язані між собою інформаційними потоками, за логічними ознаками і т. ін.;
- використання знань зі спеціальної літератури стосовно особливостей функціонування процесу, відомих законів та закономірностей його перебігу, виявлення існуючих моделей процесу та досвіду його теоретичного чи експериментального дослідження;
- при наявності розроблених моделей досліджуваного процесу необхідно

встановити їх недоліки та переваги, а також визначити можливість подальшого використання (модифікації); аналіз і використання існуючих моделей надає можливість дослідникам суттєво скоротити час та інші витрати на побудову і використання моделі.

Отриману інформацію максимально використовують для попереднього оцінювання структури моделі або декількох моделей-кандидатів, параметри яких оцінюють за допомогою експериментальних даних. У процесі виконання аналізу функціонування досліджуваного процесу доцільно використовувати та порівнювати інформацію з різних джерел. Це особливо стосується фінансово-економічних процесів, щодо яких може надходити інформація з суттєвими протиріччями, пропусками і похибками.

8.3 Попередня обробка даних

Процес попередньої обробки експериментальних (статистичних) даних складається, як правило, з таких операцій:

- нормування та візуальна перевірка даних і, при необхідності, їх корегування; нормування даних означає їх логарифмування або приведення до зручного діапазону їх зміни, наприклад, від 0 до 1; від –1 до +1; від +10 до –10 і т. ін.;
- корегування даних полягає у заповненні пропусків та зменшенні рівнів викидів (екстремальних імпульсних значень), що виходять за основний діапазон значень змінних;
- бутстреп-аналіз з метою збільшення об’ємів вибірок (розмноження вибірки);
- заміна некоректних вимірів інтерпольованими або усередненими значеннями;
- формування перших та різниць вищих порядків, які необхідні для аналізу відповідних складових процесу, представленого часовом рядом;
- ортогональні перетворення і цифрова фільтрація даних з метою вилучення шумових складових.

Поширеним методом нормування даних є їх логарифмування з наступним формуванням додаткових часових рядів з перших чи других різниць. Нагадаємо, що перші різниці представляють собою наближений дискретний аналог першої похідної, а другі різниці – другої похідної. Використання різниць дає можливість будувати моделі для швидкості та прискорення основної змінної. Часто із значень ряду віднімають його середнє для того, щоб отримати можливість працювати з відхиленнями, а не повними значеннями змінних. Такий підхід застосовують, наприклад, при побудові моделей у просторі станів з їх подальшим використанням для оптимальної фільтрації або оптимального керування процесом.

Досить хороші результати нормування при оцінюванні множинної регресії

$$y(k)=\beta_0+\beta_1 x_1(k)+\beta_2 x_2(k)+...+\beta_p x_p(k)+\varepsilon(k) \qquad (8.2.1)$$

можна досягти завдяки одночасному нормуванню і центруванню даних таким чином:

$$x_{iH}(k)=\frac{x_i(k)-\bar{x}_i}{\sqrt{S_i}}=\frac{x_i(k)-\bar{x}_i}{\left(\sum_{k=1}^{N}(x_i(k)-\bar{x}_i)^2\right)^{1/2}},\quad k=1,...,N,\ i=1,...,p;$$

$$y_H(k)=\frac{y(k)-\bar{y}}{\sqrt{S_y}}=\frac{y(k)-\bar{y}}{\left(\sum_{k=1}^{N}(y(k)-\bar{y})^2\right)^{1/2}}, \qquad (8.2.2)$$

де $x_i(k)-$ значення $i-$го стовпчика матриці вимірів (виміри незалежних змінних); $y(k)-$ виміри залежної змінної; $x_{iH}(k)$, $y_H(k)-$ нормовані значення змінних; $\bar{x}_i$, $\bar{y}-$ вибіркові середні значення незалежних і залежної змінних, відповідно; $N-$ кількість вимірів; $p-$ кількість незалежних змінних (регресорів) x_i.

Якщо ввести позначення для центрованих змінних

$$\tilde{x}_i(k)=x_i(k)-\bar{x}_i;\qquad \tilde{y}(k)=y(k)-\bar{y}, \qquad (8.2.3)$$

то регресія для центрованих змінних матиме вигляд

$$\tilde{y}(k)=\beta_1\tilde{x}_1(k)+\beta_2\tilde{x}_2(k)+...+\beta_p\tilde{x}_p(k)+\varepsilon(k). \qquad (8.2.4)$$

Якщо підставити (8.2.2) в (8.2.4), то рівняння множинної регресії набуде вигляду:

$$y_H(k)\,S_y^{1/2}=\beta_1 S_1^{1/2}x_{1H}(k)+\beta_2 S_2^{1/2}x_{2H}(k)+...+\beta_p S_p^{1/2}x_{pH}(k)+\varepsilon'(k). \qquad (8.2.5)$$

Тепер поділимо ліву і праву частини на $S_y^{1/2}$:

$$y_H(k)=\alpha_1 x_{1H}(k)+\alpha_2 x_{2H}(k)+...+\alpha_p x_{pH}(k)+\varepsilon''(k), \qquad (8.2.6)$$

де $\alpha_1=\beta_1(S_1/S_y)^{1/2}$, ..., $\alpha_p=\beta_p(S_p/S_y)^{1/2}$.

Отримане рівняння (8.2.6) – це рівняння для нормованих вимірів.

У результаті центрування і нормування покращується ступінь обумовленості матриці вимірів, який вимірюється відношенням

$$\eta=\left|\frac{\lambda_{\max}}{\lambda_{\min}}\right|,$$

де $\lambda_{\max}, \lambda_{\min}-$ максимальне і мінімальне власні числа матриці вимірів. Для забезпечення належних умов оцінювання параметрів необхідно задовольнити умову: $\eta<10$.

Застосування того чи іншого методу підготовки даних для моделювання визначається у кожному випадку по-різному.

Обробка екстремальних (аномальних) значень

Хоча виявлення та обробка екстремальних значень – це велика окрема тема для дослідження, розглянемо деякі можливості щодо розв'язання цієї проблеми. У подальшому будемо вважати дані *аномальними*, якщо вони виникли внаслідок впливу значних похибок вимірів або похибок, пов'язаних з некоректним збором статистичних даних. Якщо можна встановити факт наявності аномальних даних, то їх просто видаляють з розподілу.

Екстремальні значення – це правильно виміряні (зібрані) дані, які характеризують фактичні раптові (стрибкоподібні) зміни процесу. Підхід до розв'язання задачі дослідження екстремальних значень спостережень залежить від поставленої мети. Якщо дослідника цікавить тільки факт наявності таких значень (наприклад, з метою виявлення умов, що призводять до появи екстремальних значень), то досить мати надійний критерій для виявлення таких спостережень.

Якщо ж ставиться задача виявлення і виключення екстремальних значень (наприклад, з метою покращення оцінок статистичних параметрів і моделей), то виникає задача – як правильно виконати обробку даних. Так, спираючись на критерій для визначення екстремальних значень, можна визначити величину зміщення оцінок параметрів.

Критерії аналізу екстремальних значень застосовують з метою:

- вирівняти спостереження перед аналізом (як правило, суттєво зменшити великі значення);
- переконатись, що дані містять аномальні значення, що свідчить про необхідність перегляду процедури отримання даних;
- виділити спостереження, які є цікавими з точки зору їх аномальності та, по-можливості, описати встановлений ефект математично.

Класичний підхід до виявлення аномальних спостережень полягає в тому, що вибіркові спостереження розглядають як випадкові, нормально розподілені величини. При цьому для аналізу (виявлення екстремальних значень) формується статистика (статистичний тест, який ґрунтується на статистичних даних), яка є чутливою до різких відхилень такого типу. Необхідно встановити розподіл цієї статистики при нульовій гіпотезі, що всі спостереження належать нормальній сукупності, а потім відхилити цю гіпотезу, якщо виявиться, що обчислена статистика їй протирічить.

Розглянемо можливий критерій відкидання екстремальних значень. Нехай, дана деяка вибірка $\{x_1, x_2, ..., x_N\}$, $N \geq 3$, яка, за припущенням, є випадковою для випадкової змінної X з нормальним розподілом: $\{X\} \sim (\mu_x, \sigma_x^2)$.

Позначимо відхилення від середнього через

$$\tilde{x}_i = x_i - \bar{x}, \quad i = 1,2,\ldots,N, \quad \text{де} \quad \bar{x} = \frac{1}{N}\sum_{k=1}^{N} x(k).$$

Якщо виділити одне значення із спостережень, то вибіркове середнє для спостережень, що залишились, визначається як

$$\sum_{\substack{k=1 \\ k\neq i}}^{N} \frac{x_i}{N-1} = \bar{x} - \frac{\tilde{x}}{N-1}. \qquad (8.2.7)$$

Якщо виділити декілька значень $x_1, x_2, \ldots, x_r$, то вибіркове середнє дорівнює

$$\bar{x} - \frac{\tilde{x}_1 + \tilde{x}_2 + \ldots + \tilde{x}_r}{N-r}. \qquad (8.2.8)$$

Позначимо максимальне відхилення через $\tilde{x}_m = x_m - \bar{x}$. Тепер правило визначення екстремального значення можна сформулювати так: *при заданому значенні c спостереження x_m відкидається, якщо* $|\tilde{x}_m| > c\,S_x$, де S_x – вибіркове стандартне відхилення змінної X.

Якщо вибірка має досить великий об'єм, то значення x_m видаляється і аналіз продовжується. Величина константи c може змінюватись зі зміною довжини вибірки; вона зв'язана неявно з t – статистикою:

$$\sqrt{\frac{N c^2 (\nu + \nu_0 - 1)}{\nu(\nu + \nu_0 - \frac{N c^2}{\nu})}} \approx t_{1-\alpha/2}^{\nu_0+\nu-1}, \qquad (8.2.9)$$

де $\nu = N - 1$; α – рівень значущості; ν_0 – будь-яке інше число додаткових ступенів свободи, яке зв'язане з оцінюванням σ_x^2 за вибіркою, об'єм якої не дорівнює N ($\nu_0 = 0$, якщо такої інформації немає). Також існує наближений вираз для c через розподіл F:

$$c \approx \left(\frac{\nu}{N}\right)^{1/2} \left(\frac{3F_{1-q}}{1 + [3F_{1-q} - 1)/(\nu + \nu_0)]}\right)^{1/2}, \qquad (8.2.10)$$

де $q = \Delta\hat{\sigma}_x^2 \frac{\nu}{N}$; $\Delta\hat{\sigma}_x^2$ – очікуваний приріст дисперсії внаслідок появи екстремальних значень.

При використанні (8.2.10) значення c визначається додатнім значенням квадратного кореня.

Виразом (8.2.10) можна скористатись наступним чином: якщо з ряду значення не видалялись, то допустимий (очікуваний) відносний приріст дисперсії ("премію") $\Delta\hat{\sigma}_x^2$ необхідно помножити на ν / N і таким чином отримаємо q. За його допомогою знайдемо відповідну верхню процентну точку

для відношення дисперсій F_{1-q} при трьох і $\nu+\nu_0-1$ ступенях свободи. За виразом (8.2.10) обчислимо значення c і застосуємо критерій до x_m. Очікуваний відносний приріст дисперсії (“премія”) залежить від того, наскільки ймовірною є поява екстремальних значень, наприклад, можна прийняти невеликий відносний приріст $\Delta\hat{\sigma}_x^2 = 0{,}01 \div 0{,}03$.

Наприклад, якщо $N=4,\ \nu=3$ і $\nu/N=0{,}75$, то при $\Delta\hat{\sigma}_x^2=0{,}02$ маємо: $q=0{,}02\cdot 0{,}75=0{,}05$. При 3-х ступенях свободи $F_{1-q}=F_{1-0{,}05}=9{,}28$. Тепер знайдемо значення c:

$$c=(0{,}75)^2\left(\frac{3F_{0{,}95}}{1+(3F_{0{,}95}-1)/3}\right)^{1/2}=1{,}449.$$

Спостереження x_m необхідно видалити, якщо $\left|\tilde{x}_m\right|>1{,}449S_x$. Можливі інші підходи до аналізу екстремальних значень.

Приклад 8.1. Обчислення критерію для виявлення екстремального значення. Є ряд значень: $X=\{23{,}2\ \ 23{,}4\ \ 23{,}5\ \ 24{,}1\ \ 25{,}5\}$. Встановити, чи можна вважати значення 25,5 екстремальним і чи необхідно його видалити з вибірки?

Розв’язок. Обчислимо: $\bar{x}=23{,}9;\ \tilde{x}_5=25{,}5-23{,}9=1{,}6;\ S_x=0{,}77$. Для $\alpha=0{,}05;\ \nu=4$ і $N=5$ за виразом (8.2.9) маємо:

$$\left(\frac{15c^2}{16-5c^2}\right)^{1/2}=2{,}776^3.$$

Звідси, за методом проб та помилок, $c=1{,}49$. Згідно з критерієм,

$$\left|1{,}6\right|>1{,}49\cdot 0{,}77=1{,}05,$$

тобто спостереження x_5 видаляється.

Видалення (фільтрація) шумів з вимірів

При необхідності застосовують *фільтрацію даних від шумів*. Для розв’язання цієї важливої задачі використовують *цифрові* або *оптимальні* фільтри. Обидва методи фільтрації широко застосовуються у технічних системах, а також у системах обробки економічних, фінансових та інших типів даних. Розглянемо схематично ці можливості попередньої обробки даних.

Цифровий фільтр (ЦФ) можна представити, наприклад, рівнянням типу АР(р):

$$y(k)=a_1y(k-1)+a_2y(k-2)+\ldots+a_py(k-p).$$

Такий фільтр має амплітудно-частотну характеристику (АЧХ), яка визначається значеннями коефіцієнтів рівняння. Мета застосування ЦФ: пропустити корисну частину спектру і затримати шумову або просто не

потрібну для аналізу складову. Сьогодні існують високорозвинені методи оптимізаційного проектування ЦФ, які дозволяють спроектувати ефективні структури фільтрів з частотними характеристиками заданої форми. Зокрема, корисний інструментарій для проектування ЦФ містить система Matlab.

Оптимальний фільтр потребує, як правило, модель процесу, представлену у просторі станів. Так, фільтр Калмана ґрунтується саме на моделі такого типу.

Нехай нестаціонарна лінійна система описується у дискретному часі рівняннями з непостійними в часі коефіцієнтами (непостійність означає залежність від дискретного часу $k = 0, 1, 2, ...$):

$$\mathbf{x}(k) = \mathbf{A}(k, k-1)\mathbf{x}(k-1) + \mathbf{B}(k, k-1)\mathbf{u}(k-1) + \mathbf{w}(k),$$

де $\mathbf{x}(k)$ – n-вимірний вектор станів системи; $\mathbf{u}(k-1)$ – m-вимірний вектор детермінованих вхідних величин (сигнали керування); $\mathbf{w}(k-1)$ – n- вимірний вектор випадкових зовнішніх збурень; $\mathbf{A}(k, k-1)$ – $(n \times n)$ матриця динаміки системи (вона містить коефіцієнти, що характеризують динаміку, тобто швидкість зміни станів у часі); $\mathbf{B}(k, k-1)$ – $(n \times m)$ матриця коефіцієнтів керування. Подвійний часовий аргумент у вигляді $(k, k-1)$ означає, що величина з цим аргументом використовується в момент k, але її значення ґрунтується на попередніх даних, які відомі на момент $k-1$ включно. Далі будемо записувати для простоти матриці **A** і **B** з одним аргументом, тобто $\mathbf{A}(k)$ та $\mathbf{B}(k)$. Очевидно, що стаціонарна система описується матрицями з постійними коефіцієнтами, які записують просто **A** і **B**. Оскільки матриця **A** зв'язує поточний стан із попереднім, то її називають ще перехідною матрицею станів. Нагадаємо, що дискретний час k зв'язаний з неперервним часом t періодом дискретизації вимірів T_s: $t = kT_s$.

У класичній постановці задачі оптимальної фільтрації послідовність зовнішніх збурень $\mathbf{w}(k)$ задовольняє властивостям білого гаусового шуму з нульовим середнім значенням і коваріаційною матрицею **Q**, тобто статистики шуму мають вигляд:

$$E[\mathbf{w}(k)] = 0, \quad \forall k,$$
$$E[\mathbf{w}(k)\mathbf{w}^T(j)] = \mathbf{Q}(k)\,\delta_{kj},$$

де δ_{kj} – дельта-функція Кронекера, що визначається так: $\delta_{kj} = \begin{cases} 0 & для\ k \neq j \\ 1 & для\ k = j \end{cases}$; $\mathbf{Q}(k)$ – додатно визначена коваріаційна матриця зовнішніх збурень стану розмірності $(n \times n)$. Діагональні елементи матриці представляють собою дисперсії компонент вектора збурень $\mathbf{w}(k)$.

Початковим станом системи $\mathbf{x}_0$ будемо вважати випадкові змінні з відомими статистиками:

$$E[\mathbf{x}_0] = \bar{\mathbf{x}}_0; \quad E[\mathbf{x}_0\mathbf{x}_0^T] = \mathbf{M}; \quad E[\mathbf{w}(k)\mathbf{x}_0^T] = 0, \ \forall k.$$

Нехай, вектор вимірів $\mathbf{z}(k)$ вихідних змінних доступний у будь-який момент часу t_k, а його компоненти лінійно зв'язані з вектором стану і на них впливає шум вимірів, тобто

$$\mathbf{z}(k) = \mathbf{H}(k)\,\mathbf{x}(k) + \mathbf{v}(k),$$

де $\mathbf{H}(k)$ – матриця спостережень вимірності $(r \times n)$, $\mathbf{v}(k)$ – r-вимірний вектор випадкових величин шуму вимірів з відомими статистиками:

$$E[\mathbf{v}(k)] = 0, \quad E[\mathbf{v}(k)\mathbf{v}^T(j)] = \mathbf{R}(k)\delta_{kj},$$

де $\mathbf{R}(k)$ – додатно визначена коваріаційна матриця шумів вимірів вимірності $(r \times r)$, діагональні елементи якої є дисперсіями адитивного шуму у кожному каналі вимірів. Шум вимірів також задовольняє властивостям білого гаусового шуму. Він вважається не корельованим із зовнішнім збуренням $\mathbf{w}(k)$ і початковим станом системи, тобто

$$E[\mathbf{v}(k)\mathbf{w}^T(j)] = 0 \quad \forall\ k, j;$$
$$E[\mathbf{v}(k)\mathbf{x}_0^T] = 0 \quad \forall\ k.$$

Для визначеної вище системи з вектором стану $\mathbf{x}(k)$ необхідно знайти оцінку стану $\hat{\mathbf{x}}(k)$ в момент t_k як лінійну комбінацію оцінки $\hat{\mathbf{x}}(k-1)$ в момент t_{k-1} і самого останнього виміру (статистичних даних) $\mathbf{z}(k)$.

Оцінка $\hat{\mathbf{x}}(k)$ повинна обчислюватися як найкраща за мінімумом середнього значення суми квадратів похибок оцінок. Іншими словами, оцінка повинна бути такою, щоб

$$E[(\hat{\mathbf{x}}(k) - \mathbf{x}(k))^T(\hat{\mathbf{x}}(k) - \mathbf{x}(k))] = \min_K,$$

де $\mathbf{x}(k)$ – точне значення вектора стану, яке можна обчислити за допомогою детермінованої складової математичної моделі процесу (тобто без врахування випадкових складових – збурень стану і шумів вимірів); $\mathbf{K}$ – оптимальний матричний коефіцієнт фільтра, який необхідно обчислити в результаті розв'язання оптимізаційної задачі.

Таким чином, фільтр слід будувати та використовувати для уточнення оцінок стану процесу в умовах впливу випадкових зовнішніх збурень та наявності шумів (похибок) вимірів. Сьогодні оптимальні фільтри – це практично обов'язкова складова комп'ютерних систем обробки експериментальних і статистичних даних. Докладніше оптимальний фільтр буде розглянуто нижче.

8.4 Аналіз наявності нелінійностей

Однією з проблем при визначенні структури моделі є встановлення факту наявності нелінійностей у досліджуваному процесі та їх типу. Для розв'язання

цієї проблеми обов'язково використовують *візуальний аналіз* даних та формальні тести на наявність нелінійностей. Досвідченому фахівцю з моделювання візуальний аналіз дозволяє оперативно виявити наявність ділянок з лінійним або нелінійним трендом, наявність гетероскедастичності та значних імпульсних викидів (екстремальних значень), які можуть суттєво впливати на якість моделі. Необхідно зазначити, що існують окремі навчальні курси з візуального аналізу даних. Це свідчить про те, що не варто нехтувати такою доступною, але ефективною можливістю дослідження даних, адже, на думку психологів, один інформативний рисунок може замінити до двох тисяч слів.

Існує також ряд формальних тестів на наявність нелінійності. Розглянемо простий тест для визначення наявності нелінійності. Цей тест можна застосувати у випадку, коли можна набрати декілька груп (реалізацій) спостережень для одного і того ж процесу:

$$\widehat{F} = \frac{\frac{1}{m-2}\sum_{i=1}^{m} n_i (\bar{y}_i - \hat{y}_i)^2}{\frac{1}{n-m}\sum_{i=1}^{m}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2},$$

де $\bar{y}_i$ – середнє значення для i –ї реалізації (вибірки або групи) даних; $\hat{y}_i$ – середнє для лінійної апроксимації даних; m – кількість груп даних; n_i – кількість вимірів в i –й групі; n – загальна кількість вимірів. Фактично, наведена статистика представляє собою таке відношення:

$$\widehat{F} = \frac{Відхилення\ середніх\ значень\ від\ прямої\ регресії}{Відхилення\ значень\ y(k)\ від\ групових\ середніх}.$$

Якщо статистика $\widehat{F}$ з $v_1 = m-2$ та $v_2 = n-m$ ступенями свободи досягає або перевищує рівень значущості, то гіпотезу стосовно лінійності процесу необхідно відхилити. Недоліком даного підходу є те, що для його застосування необхідно мати декілька (не менше трьох) груп (реалізацій) даних для одного і того ж процесу, які можна отримати в результаті виконання повторних експериментів. Очевидно, що це не завжди можливо.

Наявність нелінійності можна встановити також за допомогою вибіркових *нелінійних кореляційних функцій* (НКФ), тобто кореляційних функцій, розрахованих за вибірками експериментальних (статистичних) даних. Наприклад, якщо дискретна НКФ

$$r_{yx^2}(s) = r_{y(k)x^2(k-s)} = \frac{1}{N}\frac{\sum_{k=s+1}^{N}\left\{[y(k)-\bar{y}][x(k-s)-\bar{x}]^2\right\}}{\sigma_y \sigma_x^2}, \quad s = 0,1,2,3,\ldots,$$

містить значення, які суттєво відрізняються від нуля в статистичному смислі, то процес містить квадратичну нелінійність відносно регресора x.

Наявність нелінійного детермінованого тренду у процесі можна визначити шляхом оцінювання рівняння

$$y(k) = a_0 + c_1 k + c_2 k^2 + \ldots + c_m k^m + \varepsilon(k),$$

яке являє собою поліном порядку m відносно часу; $\varepsilon(k)$ – випадковий процес, причини появи якого будуть розглянуті нижче у даному розділі. Якщо хоча б один із коефіцієнтів моделі, c_i, $i = 1,\ldots,m$, є статистично значущим, то гіпотеза щодо відсутності тренду відхиляється. Якщо тренд відносно швидко змінює свій напрям руху і для нього складно знайти адекватний функціональний опис, то застосовують моделі випадкових трендів, які ґрунтуються на комбінаціях випадкових величин. Наявність нелінійності стосовно регресора $x(k)$ можна встановити за допомогою відповідного полінома:

$$y(k) = a_0 + c_1 x(k) + c_2 x^2(k) + \ldots + c_m x^m(k) + \varepsilon(k).$$

Автоматично оцінює структуру математичної моделі *метод групового врахування аргументів* (МГВА), запропонований академіком О. Г. Івахненком (Київський інститут кібернетики) [18]. Цей метод вже багаторазово застосовували до опису широкого класу процесів з метою оцінювання прогнозів та створення систем керування. Його успішно застосовують і сьогодні до моделювання процесів різної природи з нелінійностями та нестаціонарностями. Подальшим розвитком даного методу є нечіткий МГВА, який ґрунтується на нечіткому представленні параметрів оцінюваної моделі. Загалом задача встановлення наявності та визначення типу нелінійності залишається предметом дослідження.

8.5 Формування інших елементів структури моделі

На третьому етапі необхідно вибрати структури моделей - кандидатів. Поняття структури моделі було розглянуто вище, але нагадаємо, що воно включає *порядок моделі* (найвищий порядок рівнянь, що його утворюють); *вимірність* (кількість рівнянь моделі); час *запізнення* на вході (лаг) та його оцінка; *можливі нелінійності* та їх тип; *зовнішні збурення* та їх тип (детерміновані або випадкові; адитивні та мультиплікативні; імпульсні та неперервні); *обмеження* на параметри і змінні моделі.

Для того, щоб визначити які незалежні змінні (регресори) необхідно включити в праву частину рівняння, обчислюють коефіцієнт кореляції між залежною та відповідною незалежною змінною.

Коефіцієнт кореляції, а в загальному випадку кореляційна функція, дає змогу встановити факт існування зв’язку між змінними. Кореляція може бути лінійною або нелінійною, в залежності від типу функціональної залежності, яка фактично має місце між змінними. У більшості практичних випадків розглядають лінійну кореляцію (взаємозв'язок) між змінними, але більш глибокий аналіз потребує використання нелінійних залежностей. Складну нелінійну залежність можна спростити, але знати про її існування необхідно

для того, щоб побудувати при необхідності (складнішу за структурою) модель процесу з вищим ступенем адекватності.

Кореляційна матриця дає можливість встановити існування зв'язку між залежною (ендогенною) змінною та незалежними (екзогенними) змінними у правій частині. Розглянемо кореляційну матрицю **R** розмірності 3×3, яка будується для трьох змінних $x,\ y,\ z$:

$$\mathbf{R}=\begin{bmatrix} r_{yy} & r_{xy} & r_{zy} \\ r_{yx} & r_{xx} & r_{zx} \\ r_{yz} & r_{xz} & r_{zz} \end{bmatrix}, \quad \text{де} \quad r_{yx}=r_{xy},\ \ r_{yz}=r_{zy},\ \ r_{xz}=r_{zy}.$$

Нехай $y-$ показник якості технологічного процесу; $x,z-$ технологічні параметри, які, за припущенням, впливають на показник якості. Тобто ставиться задача встановлення існування залежності вигляду:

$$y=f(x,z),$$

яка може бути представлена у формі регресії змінної y на незалежні змінні $x,\ z$:

$$y(k)=a_0+a_1x(k)+a_2z(k)+\varepsilon(k),$$

де $k-$ дискретний час (наприклад, у долях секунди, секундах, хвилинах, годинах, днях, тижнях, місяцях і т.ін.); $\varepsilon(k)-$ випадкова змінна, введення якої у модель пояснюється наступними причинами:

– часто буває неможливо встановити всі незалежні змінні, які впливають на залежну змінну, а тому наведене рівняння описує процес з похибкою;

– можуть існувати такі незалежні змінні, які неможливо виміряти і включити в модель, а тому їх розглядають як збурення і вважають, що їх спільний вплив на залежну змінну описується випадковою змінною $\varepsilon(k)$;

– у наведене вище регресійне рівняння можна ввести пояснювальні змінні, які формально корельовані із залежною змінною, але фактично не впливають на неї;

– для будь-якого методу оцінювання параметрів рівнянь властиві методичні обчислювальні похибки, які слід по-можливості врахувати в моделі.

Вважається, що сукупний вплив усіх вказаних факторів можна описати деякою мірою за допомогою випадкової змінної $\varepsilon(k)$. Оскільки вона не вимірюється, то оцінити її значення (*похибку* моделі або *залишок*) можна тільки після оцінювання коефіцієнтів моделі, тобто

$$\hat{\varepsilon}(k)=e(k)=y(k)-\hat{y}(k),$$

де $\hat{y}(k)-$ оцінка змінної $y(k)$, отримана за допомогою моделі; $y(k)-$

фактичний вимір.

Для обчислення елементів матриці **R** необхідно мати синхронізовані у часі вибірки значень всіх трьох змінних y, x, z. Формула для розрахунку вибіркових коефіцієнтів кореляції має вигляд

$$r_{yx} = \frac{\frac{1}{N-1}\sum_{k=1}^{N}\{[x(k)-\bar{x}][y(k)-\bar{y}]\}}{\sigma_x \sigma_y},$$

де $\bar{x}, \bar{y}-$ вибіркові середні значення змінних x, y; σ_x, σ_y – стандартні відхилення цих змінних, тобто корені квадратні з їх дисперсії. Наприклад,

$$\sigma_y = \sqrt{\sigma_y^2} = \left[\frac{1}{N-1}\sum_{k=1}^{N}[y(k)-\bar{y}]^2\right]^{1/2},$$

де $N-$ число вимірів змінної y; $\bar{y}-$вибіркове середнє значення ряду $\{y(k)\}$, яке обчислюється за відомою формулою

$$\bar{y} = \frac{1}{N}\sum_{k=1}^{N} y(k).$$

Очевидно, що перед формальним обчисленням коефіцієнтів кореляції, необхідно зробити аналіз процесу і визначити наявність (або відсутність) логічних зв'язків між змінними. Це дозволяє обмежитись розглядом тільки тих змінних, які дійсно впливають на залежну змінну, наприклад, на показник якості продукції. Очевидно, що для правильного вибору незалежних (екзогенних або пояснювальних) змінних необхідно глибоко знати технологічний або інший процес, що моделюється.

На основі значень коефіцієнтів кореляції ухвалюється рішення про включення їх у рівняння регресії:

$$y(k) = a_0 + b_1 x(k) + b_2 z(k) + \varepsilon(k),$$

яке може бути представлено в загальному вигляді так:

$$y(k) = a_0 + a_1 x_1(k) + a_2 x_2(k) + a_3 x_3(k) + \ldots + a_{p-1} x_{p-1}(k) + \varepsilon(k).$$

Можна показати, що між коефіцієнтами регресії b_1, b_2 і коефіцієнтами кореляції r_{yx}, r_{yz} існує однозначний аналітичний взаємозв'язок. Емпіричні дослідження свідчать, що незалежні змінні необхідно включати в регресію, якщо $|r_{yx}| \geq 0{,}4$. Однак необхідно пам'ятати, що ця рекомендація має досить наближений характер. Існують випадки, коли у модель слід включати регресори, які мають менші значення коефіцієнта кореляції із залежною змінною.

Останнє рівняння представляє собою *лінійну регресію*, яка містить p параметрів (коефіцієнтів), але досить часто необхідно застосовувати складніші

нелінійні моделі. Характерним представником нелінійної стосовно змінних регресії є поліноміальна регресія порядку $p-1$:

$$y(k) = a_0 + a_1 x(k) + a_2 x^2(k) + a_3 x^3(k) + \ldots + a_{p-1} x^{p-1}(k) + \varepsilon(k).$$

Хоча в це рівняння включено тільки одну незалежну змінну $x(k)$, очевидно, що воно може бути розширене будь-якими іншими змінними.

Регресори, які включають у модель з фактичними значеннями їх лагів, називають *провідними індикаторами*. Наприклад, якщо результат (прибуток) від інвестицій з'являється через три квартали, то при використанні квартальних даних регресійне рівняння буде мати такий вигляд:

$$y(k) = a_0 + a_1 x(k-3) + \varepsilon(k).$$

Таке рівняння дає можливість коректно прогнозувати залежну змінну на три кроки вперед.

Для визначення необхідності введення в рівняння регресії авторегресійної складової необхідно обчислити і дослідити вибіркову *автокореляційну функцію* змінної $y(k)$. Рівняння з авторегресійною складовою має вигляд:

$$y(k) = a_0 + a_1 y(k-1) + a_2 y(k-2) + b_1 x(k) + b_2 z(k) + \varepsilon(k),$$

тобто в рівняння регресії введено авторегресійну (АР) складову другого порядку. Порядок авторегресії визначається за допомогою автокореляційної функції (АКФ) і часткової АКФ (ЧАКФ). Кількість значень автокореляційної функції, які відмінні від нуля в статистичному смислі й визначаються параметром зсуву s (що використовується при обчисленні кореляційних функцій), і буде становити оцінку порядку авторегресії.

Природа авторегресії пояснюється існуванням так званої "пам'яті" процесу, яка проявляється у тому, що його поточний стан може значною мірою визначатись попередніми станами. Наприклад, стан людини вранці залежить від того, яким було самопочуття ввечері та в попередні дні. На формування ринкових цін суттєво впливають значення цін у попередні періоди часу. Очевидно, що поточний рівень валового внутрішнього продукту (ВВП) залежить від його попередніх значень, а поточний стан технічної системи або технологічного процесу також залежить від їх стану у попередні моменти часу.

АКФ та ЧАКФ використовують для визначення попередньої оцінки порядку авторегресійної частини моделі, тобто скільки затриманих в часі значень слід брати для опису процесу. При цьому необхідно врахувати, що ЧАКФ дає «чіткішу» оцінку порядку процесу, ніж АКФ. Наприклад, для процесу АР(1) значення основної змінної $y(k)$ та $y(k-2)$ будуть корельованими, незважаючи на те, що $y(k-2)$ не присутнє у моделі. Кореляція між $y(k)$ і $y(k-2)$, тобто ρ_2, дорівнює коефіцієнту кореляції між значеннями $y(k)$ і $y(k-1)$, помноженому на коефіцієнт кореляції між $y(k-1)$ і $y(k-2)$

або $\rho_2 = \rho_1 \rho_1 = \rho_1^2$. Подібні «непрямі» кореляції наявні в АКФ будь-якого процесу авторегресії.

Вибіркова АКФ обчислюється за виразом:

$$r_y(s) = r_{y(k)y(k-s)} = \frac{1}{N-1} \frac{\sum_{k=s+1}^{N} \{[y(k) - \bar{y}][y(k-s) - \bar{y}]\}}{\sigma_y^2}, \quad s = 1,2,3,\ldots ,$$

де $\sigma_y^2 -$ вибіркова дисперсія змінної $y(k)$; $\bar{y} -$ середнє значення вибірки даних.

Кількість значень АКФ, відмінних від нуля в статистичному смислі, вказує на порядок авторегресійної частини моделі. Для стаціонарного процесу (це процес із постійними середнім, дисперсією та коваріацією) коефіцієнти $r_y(s)$ мають нормальний розподіл та нульове середнє.

На відміну від АКФ, часткова АКФ між значеннями $y(k)$ та $y(k-s)$ виключає вплив величин $y(k-1)\ldots y(k-s+1)$, а це означає, що коефіцієнти ЧАКФ чіткіше відображають зв'язок між окремими значеннями основної змінної. Так, для процесу АР(1) ЧАКФ між $y(k)$ та $y(k-2)$ дорівнює нулю за визначенням, що підтверджується обчисленими значеннями ЧАКФ. *Для того, щоб знайти попередню оцінку порядку моделі, вибіркові коефіцієнти ЧАКФ (тобто коефіцієнти, знайдені за вибіркою даних) можна обчислити також за допомогою простого методу, який полягає в наступному*:

1) Обчислюють додатковий часовий ряд з відхилень основної змінної:

$$\{y'(k)\} = \{y(k)\} - \mu,$$

де $\mu -$ середнє значення ряду.

2) Формують рівняння першого порядку:

$$y'(k) = \Phi_{11} y'(k-1) + e(k),$$

де $e(k) -$ похибка моделі. У такому рівнянні Φ_{11} відіграє роль коефіцієнта АКФ та ЧАКФ між $y(k)$ та $y(k-1)$. Для оцінювання двох коефіцієнтів можна сформувати рівняння другого порядку:

$$y'(k) = \Phi_{11} y'(k-1) + \Phi_{22} y'(k-2) + e(k),$$

де $\Phi_{22} -$ коефіцієнт ЧАКФ між $y(k)$ та $y(k-2)$.

Таким чином, дискретні значення ЧАКФ можна обчислити за допомогою значень АКФ, використовуючи такі вирази [17]:

$$\Phi_{11} = r(1), \qquad \Phi_{22} = \frac{r_2 - r_1^2}{1 - r_1^2};$$

$$\Phi_{ss} = \frac{r_s - \sum_{j=1}^{s-1} \Phi_{s-1,j} r_{s-j}}{1 - \sum_{j=1}^{s-1} \Phi_{s-1,j} r_j}.$$

У загальному випадку коефіцієнти ЧАКФ стаціонарного процесу АРКС(p,q) повинні збігатися до нуля, починаючи із p-го значення. АКФ процесу АРКС(p,q) починає збігатися до нуля при значеннях зміщення $s \geq q$.

Твердження, що значення автокореляційної функції повинні бути відмінними від нуля в статистичному смислі, означає, що існує вираз (формула), який дає змогу підтвердити або спростувати цей факт. Одним із загальноприйнятих підходів до встановлення того факту, що значення АКФ суттєво відмінні від нуля в статистичному смислі, є обчислення та аналіз значущості статистичного параметра (або просто статистики) Люнга-Бокса $Q(r_k)$ за формулою:

$$Q(r_k) = N(N+2) \sum_{k=1}^{s} r_k^2 /(N-k),$$

де $N-$ довжина вибірки даних змінної, для якої обчислено значення автокореляційної функції r_k; $s-$ кількість відліків АКФ, які досліджуються на суттєву відмінність від нуля. Якщо дані згенеровані процесом АР чи АРКС, то значення $Q(r_k)$ асимптотично мають розподіл χ^2 з s ступенями свободи, а тому для перевірки їх значущості необхідно користуватись відповідними статистичними таблицями. Очевидно, що більші значення вибіркової автокореляційної функції призводять до більших значень $Q(r_k)$.

Так, для ідеального процесу білого шуму $Q(r_k) = 0$. Якщо значення $Q(r_k)$, обчислене за наведеним виразом, перевищує критичне значення з розподілу χ^2 з s ступенями свободи, то існує щонайменше одне значення $r(k)$, яке є відмінним від нуля в статистичному смислі. Статистику Люнга-Бокса можна застосовувати також для встановлення ступеня близькості залишків моделі до білого шуму. Однак необхідно пам'ятати, що при обчисленні s значень кореляційної функції кількість ступенів свободи зменшується на кількість коефіцієнтів моделі. Таким чином, при аналізі залишків моделі АРКС(p,q) статистика $Q(r_k)$ має розподіл χ^2 з $s-p-q$ ступенями свободи, а із врахуванням константи $s-p-q-1$.

Цей етап закінчується формуванням структур декількох моделей-кандидатів з векторами параметрів $\theta_1,...,\theta_m$, де $m-$ число кандидатів. Кандидатів може бути декілька, оскільки встановити структуру точно за один раз, як правило, неможливо. Загалом оцінювання структури моделі високого ступеня адекватності – це трудомісткий ітераційний процес, який вимагає

поглиблених знань, досвіду моделювання та значних зусиль. На наступному етапі оцінюють числові значення параметрів знайдених моделей-кандидатів.

8.6 Оцінювання параметрів моделей-кандидатів

На *четвертому етапі* оцінюють параметри (коефіцієнти) рівнянь-кандидатів за умови відомої структури цих рівнянь, використовуючи принцип економії або збереження. Цей принцип означає, що *кількість коефіцієнтів, що оцінюються, не повинна перевищувати їх необхідну кількість* ("необхідність" можна визначити, наприклад, як необхідність збереження в моделі основних статистичних характеристик процесу – математичного сподівання, дисперсії та коваріації).

При моделюванні процесів будь-якої природи необхідно пам'ятати, що поведінку процесу слід коректно *апроксимувати* за допомогою рівнянь, а не намагатися описати її до найменших дрібниць. Необхідно враховувати також, що різні за структурою моделі можуть мати дуже схожі властивості. Наприклад, рівняння авторегресії першого порядку

$$y(k) = 0{,}5y(k-1) + \varepsilon(k)$$

еквівалентне ковзному середньому у вигляді

$$y(k) = \varepsilon(k) + 0{,}5\varepsilon(k-1) + 0{,}25\varepsilon(k-2) + 0{,}125\varepsilon(k-3) + \ldots \ .$$

Модель, що оцінюється, повинна задовольняти принципу інверсії, тобто щоб за допомогою отриманого рівняння можна було б згенерувати початковий ряд, на основі якого оцінювались коефіцієнти. Це означає, що хоча модель і спрощена, вона повинна співпадати з досліджуваним процесом за такими основними характеристиками як середнє, дисперсія та коваріація.

У процедурі оцінювання часто використовують не абсолютні значення змінних, а їх відхилення від середнього, тобто

$$y(k) = Y(k) - \mu_y,$$

де $Y(k)$ – значення виміру, μ_y – середнє значення ряду. Якщо для оцінювання параметрів використовується рекурсивна процедура, то поточне середнє можна обчислювати за формулою

$$\mu_y(k) = \mu_y(k-1) + \frac{1}{k}[y(k) - \mu_y(k-1)].$$

Найбільш поширеними методами оцінювання параметрів моделі є такі:

- *метод найменших квадратів* (МНК),
- *метод максимальної правдоподібності* (ММП),
- *метод допоміжної (інструментальної) змінної* (МДП),
- *нелінійний метод найменших квадратів* (НМНК),

– *метод Монте- Карло для марковських ланцюгів* (МКМЛ)

та їх рекурсивні версії (РМНК, РММП, РМДП). Деякі методи розглянуті в розділі, присвяченому методам оцінювання. Оцінки (звичайного) МНК обчислюють за допомогою такого виразу:

$$\hat{\theta} = \left[\mathbf{X}^T \mathbf{X}\right]^{-1} \mathbf{X}^T \mathbf{y},$$

де $\theta\,[p]$ – вектор оцінок параметрів вимірності p; $\mathbf{X}[N \times p]$ – матриця вимірів; $\mathbf{y}[N]$ – вектор вимірів залежної змінної. У квадратних дужках вказана вимірність векторів і матриці. Елементи матриці вимірів формують по-своєму для кожної конкретної моделі. Так, для моделі

$$y(k) = a_0 + a_1 x_1(k) + a_2 x_2(k) + a_3 x_3(k) + \varepsilon(k)$$

матриця вимірів має вигляд

$$\mathbf{X} = \begin{bmatrix} 1 & x_1(1) & x_2(1) & x_3(1) \\ 1 & x_1(2) & x_2(2) & x_3(2) \\ . & . & . & . \\ 1 & x_1(N) & x_2(N) & x_3(N) \end{bmatrix}.$$

Одиниці у першому стовпчику матриці $\mathbf{X}$ означають, що вимір при коефіцієнті a_0 завжди дорівнює одиниці.

Елементи матриці вимірів дещо ускладнюються у випадку використання поліноміальної моделі, але її також можна оцінювати за допомогою лінійних методів. Безпосереднє застосування методу мінімізації суми квадратів похибок до поліноміальної моделі порядку p призводить до формування такої матриці вимірів:

$$\mathbf{X}' = \begin{bmatrix} N & \sum_{k=1}^{N} x(k) & \sum_{k=1}^{N} x^2(k) & \dots & \sum_{k=1}^{N} x^p(k) \\ \sum_{k=1}^{N} x(k) & \sum_{k=1}^{N} x^2(k) & \sum_{k=1}^{N} x^3(k) & \dots & \sum_{k=1}^{N} x^{p+1}(k) \\ . & . & . & \dots & \dots \\ \sum_{k=1}^{N} x^p(k) & \sum_{k=1}^{N} x^{p+1}(k) & \sum_{k=1}^{N} x^{p+2}(k) & \dots & \sum_{k=1}^{N} x^{2p}(k) \end{bmatrix}.$$

При такому представлені векторно-матричне рівняння для N вимірів залежної та незалежної змінних можна записати так: $\mathbf{y}' = \mathbf{X}'\,\theta$; звідси $\hat{\theta} = [\mathbf{X}']^{-1}\mathbf{y}'$, де $\mathbf{y}' = \left[\sum_{1}^{N} y(k) \quad \sum_{1}^{N} x(k)\,y(k) \quad \cdots \quad \sum_{1}^{N} x^p(k)\,y(k)\right]^T$; $\hat{\theta}$ – вектор оцінок параметрів моделі. Тобто, оцінку вектора параметрів можна знайти шляхом знаходження розв’язку системи лінійних (нормальних) рівнянь.

Для отримання незміщених, консистентних та ефективних оцінок вектора параметрів θ лінійної регресійної математичної моделі, наприклад, моделі змішаної регресії:

$$y(k) = a_0 + a_1 y(k-1) + a_2 y(k-2) + b_1 x(k) + b_2 z(k) + \varepsilon(k),$$

за допомогою методу найменших квадратів необхідно задовольнити наступні умови:

а) $\varepsilon(k)-$ некорельована послідовність однаково (нормально) розподілених випадкових чисел з нульовим середнім, тобто $E[\varepsilon(k)] = 0$,

$$\operatorname{cov}[\varepsilon(k)] = E[\varepsilon(k)\varepsilon(j)] = \begin{cases} \sigma_\varepsilon^2, & k = j; \\ 0, & k \neq j. \end{cases}$$

б) послідовності $\varepsilon(k)$ і $y(k)$ не повинні бути корельовані між собою.

Зазначимо, що перевірити виконання наведених умов ми можемо тільки після оцінювання коефіцієнтів моделі (апостеріорно), а до оцінювання (апріорно) можна тільки постулювати їх виконання. Тобто, після оцінювання параметрів моделі оцінка значень випадкового процесу визначається похибками моделі:

$$\hat{\varepsilon}(k) = e(k) = y(k) - \hat{y}(k),$$

що дає можливість виконати аналіз характеристик випадкового процесу $\{\varepsilon(k)\}$.

8.7 Діагностика моделей – вибір кращої з множини оцінених кандидатів

На п'ятому етапі аналізується якість моделі, тобто виконується перевірка оцінених кандидатів на адекватність процесу. Діагностика побудованих моделей складається з кроків, наведених нижче.

а) *Візуальне дослідження графіка похибок* моделі $e(k) = y(k) - \hat{y}(k)$, де $\hat{y}(k)-$ оцінка змінної, отримана за допомогою побудованої моделі. На графіку не повинно бути значних викидів та довгих інтервалів, на яких похибка набуває великих значень (тобто довгих інтервалів суттєвої неадекватності). У випадку застосування рекурсивних методів оцінювання, найбільші похибки будуть у перехідному процесі, коли інформаційна матриця ще не містить достатньо інформації про процес. Однак, для моделей низьких порядків (1-2) перехідний процес повинен закінчуватись через 20-30 кроків (не більше), а оцінки параметрів мають прямувати до точних значень.

Для аналізу характеристик похибок необхідно застосовувати статистичні характеристики (статистики), які свідчать про близькість цього випадкового процесу до білого шуму. Це коефіцієнт асиметрії, ексцес, статистика Жак-Бера.

б) *Похибки моделі не повинні бути корельовані* між собою. Для аналізу наявності кореляції між значеннями похибок необхідно обчислити АКФ та ЧАКФ для ряду $\{e(k)\}$ і за допомогою $Q-$статистики визначити ступінь корельованості (наприклад, $Q-$статистика вважається несуттєвою до рівня 10%).

Крім того, корельованість похибок визначають за допомогою статистики Дарбіна-Уотсона (DW), яка розраховується за формулою

$$DW = 2 - 2\rho,$$

де $\rho = E[e(k)e(k-1)]/\sigma_e^2$ – коефіцієнт кореляції між сусідніми значеннями похибки; σ_e^2- дисперсія послідовності похибок $\{e(k)\}$. Таким чином, при повній відсутності кореляції між похибками $DW = 2$ – це ідеальне значення. Граничними значеннями для DW є 0 (при $\rho = 1$) та +4 (при $\rho = -1$).

Отримати формулу $DW = 2 - 2\rho$ можна досить просто. Автори цієї статистики (Durbin і Watson) запропонували скористатись для перевірки корельованості похибок моделі таким виразом [17]:

$$DW = \frac{\sum_{k=2}^{N}\left[e(k) - e(k-1)\right]^2}{\sum_{k=1}^{N} e^2(k)} = \frac{\sum_{k=2}^{N}\left[e(k) - e(k-1)\right]\left[e(k) - e(k-1)\right]}{\sum_{k=1}^{N} e^2(k)},$$

Тобто, DW можна, в деякій мірі, трактувати як коефіцієнт автокореляції для (перших різниць) приростів похибок.

Розкриваючи квадрат різниці у чисельнику, отримаємо:

$$DW = \frac{\sum_{k=2}^{N} e^2(k)}{\sum_{k=1}^{N} e^2(k)} + \frac{\sum_{k=2}^{N} e^2(k-1)}{\sum_{k=1}^{N} e^2(k)} - 2\,\frac{\sum_{k=2}^{N} e(k)e(k-1)}{\sum_{k=1}^{N} e^2(k)},$$

де $\frac{\sum_{k=2}^{N} e^2(k)}{\sum_{k=1}^{N} e^2(k)} \approx 1$; $\frac{\sum_{k=2}^{N} e^2(k-1)}{\sum_{k=1}^{N} e^2(k-1)} \approx 1$; а $\frac{\sum_{k=2}^{N} e(k)e(k-1)}{\sum_{k=1}^{N} e^2(k-1)} = \rho$.

Тому можна записати, що $DW = 2 - 2\rho$.

в) Для лінійної моделі 2-3-го порядку *оцінки параметрів повинні збігатися до усталених значень* після 30-40 (не більше) ітерацій алгоритму оцінювання. Якщо кількість ітерацій набагато перевищує вказані числа, то це свідчить про те, що процес може бути нестаціонарним.

г) Перевірка статистичної значущості оцінок параметрів моделі. *Статистика Стьюдента або* $t-$*статистика* (випадкова величина, що має $t-$розподіл), яка використовується для визначення значущості оцінки кожного коефіцієнта в статистичному смислі, визначається за виразом:

$$t = \frac{\hat{a} - a^0}{SE_{\hat{a}}},$$

де $\hat{a}-$ оцінка коефіцієнта моделі; a^0- нуль-гіпотеза (початкова гіпотеза) щодо цієї оцінки; $SE_{\hat{a}}-$ стандартна похибка оцінки. В якості нуль-гіпотези щодо значущості оцінки можна висувати будь-яку: що коефіцієнт значущий, тобто $H_0 : a^0 \neq 0$, або незначущий ($H_0 : a^0 = 0$). Статистична теорія перевірки гіпотез пропонує висувати нуль-гіпотезу, яка є протилежною бажаному результату. У даному випадку бажаним результатом є статистична значущість коефіцієнтів математичної моделі. Таким чином, необхідно висувати нульову гіпотезу, що коефіцієнт незначущий. Це дає можливість коректно підійти до визначення значущості оцінок коефіцієнтів та дещо спростити розрахунки.

Для того, щоб встановити, чи оцінка коефіцієнта значуща, необхідно знати довжину вибірки даних N (потужність вибірки); кількість ступенів свободи $f = N - n$, де $n-$ число коефіцієнтів моделі, які оцінюються на основі ряду даних, і вибрати рівень значущості $\alpha = 1\%$, або $\alpha = 5\%$, або $\alpha = 10\%$ (для цих значень існують розраховані таблиці критичних значень $t-$статистики). Фактично, рівень значущості означає ймовірність припуститись *помилки першого роду* при перевірці гіпотези. Згадаємо, що

$$\alpha = p\{X \in G/\omega \,|\, H_0\} = \int\limits_{n-m(G/\omega)} L_{H_o}(X)\,dx,$$

де $X = [x_1,...,x_n] \in R^n-$ вся вибірка, яка розбивається на дві множини, що перетинаються: ω і G/ω ($\omega-$ область прийняття нуль-гіпотези); $G/\omega-$ критична область: якщо $X \in G/\omega$, то H_0 відхиляється; $L_{H_0}(X)-$ закон розподілу X. Помилка першого роду означає відхилення вірної гіпотези.

Користуючись значеннями N, f і α, з таблиць для $t-$розподілу знаходять критичне значення $t-$статистики, тобто $t_{кр}$. Для перевірки правильності висунутої гіпотези розраховане значення t порівнюють з критичним $t_{кр}$. Якщо

$$-t_{кр} < t < t_{кр} \quad \text{або} \quad |t| < |t_{кр}|,$$

то нуль-гіпотеза стосовно незначущості коефіцієнта приймається (його можна не враховувати в регресії). Звідси випливає, що чим більшим є значення $t-$статистики для оцінки коефіцієнта, тим імовірніше, що цей коефіцієнт є значущим.

Загалом послідовність дій при перевірці значущості оцінок коефіцієнтів побудованої моделі можна сформулювати так:

- сформулювати нуль-гіпотезу стосовно значущості коефіцієнта;
- обчислити значення $t-$статистики для кожного коефіцієнта регресії (це робить кожний пакет для математичного моделювання);
- за допомогою значень N, f і α знайти із таблиць для $t-$статистики її критичне значення;
- перевірити нуль-гіпотезу за наведеним вище простим правилом (аналіз виконання нерівності $-t_{кр} < t < t_{кр}$).

д) *Коефіцієнт множинної детермінації* R^2, який обчислюється так:

$$R^2 = \frac{\mathrm{var}(\hat{y})}{\mathrm{var}(y)} = 1 - \frac{SSE}{SST},$$

де $\mathrm{var}(\hat{y})-$ дисперсія залежної змінної, оціненої за допомогою побудованої моделі; $\mathrm{var}(y)-$ дисперсія вимірів залежної змінної; $SSE = \sum_{k=1}^{N}[y(k) - \hat{y}(k)]^2 -$ сума квадратів похибок (залишків) моделі (*sum of squared errors*); $SST = \sum_{k=1}^{N}[y(k) - \bar{y}]^2 -$ загальна сума квадратів (*total sum of squares*); $\bar{y}-$ середнє значення; $SST = SSE + SSR$, де $SSR = \sum_{k=1}^{N}[\hat{y}(k) - \bar{y}]^2 -$ загальна сума квадратів для регресії (*sum of squares for regression*).

Очевидно, що найкращим значенням є $R^2 = 1$, коли дисперсії вимірів змінної та оцінок цієї ж змінної, отриманих за моделлю, збігаються. Цей параметр можна трактувати також, як міру інформативності моделі, якщо за міру інформативності вибрати дисперсію. Таким чином, R^2 показує рівень інформативності моделі по відношенню до інформативності вибірки даних, за допомогою якої вона була оцінена.

е) *Сума квадратів похибок* для вибраної моделі повинна бути мінімальною, тобто

$$\sum_{k=1}^{N} e^2(k) = \sum_{k=1}^{N}[\hat{y}(k) - y(k)]^2 \to \min_{\hat{\theta}}$$

у порівнянні з усіма іншими моделями.

є) Для оцінки адекватності моделі також використовують *інформаційний критерій Акайке*:

$$AIC = N \ln\left(\sum_{k=1}^{N} e^2(k)\right) + 2n,$$

та *критерій Байєса-Шварца*:

$$BSC = N \ln\left(\sum_{k=1}^{N} e^2(k)\right) + n \ln(N),$$

де $n = p + q + 1-$ кількість параметрів моделі, які оцінюються за допомогою статистичних даних (p – число параметрів авторегресійної частини моделі; q – кількість параметрів ковзного середнього; „1" з'являється тоді, коли оцінюється зміщення (або *перетин*, тобто a_0).

Критерії Акайке і Байєса-Шварца містять у правій частині суму квадратів похибок, а тому за цими критеріями вибирають ту модель, для якої ці критерії набувають найменших значень. Введення нового регресора призводить до збільшення критерію (при цьому збільшується n), але одночасно зменшується сума квадратів похибок і критерій в цілому зменшується. Якщо регресор не покращує модель, то критерій збільшується. Необхідно також зазначити, що асимптотичні властивості для довгих вибірок кращі у критерію Байєса-Шварца, тобто його рекомендують застосовувати при відносно великих значеннях N ($N > 100$).

ж) Окрім згаданих параметрів, для визначення адекватності моделі в цілому, використовують $F-$*статистику Фішера*, яка пропорційна відношенню:

$$F \sim \frac{R^2}{1 - R^2},$$

а для множинної (багатофакторної) регресії вона визначається за виразом

$$F = \frac{R^2}{1 - R^2} \cdot \frac{(N - p - 1)}{p},$$

де, як і раніше, $N -$ число значень ряду; $p -$ число параметрів моделі без врахування перетину (константи).

Таким чином, якщо $R^2 \to 1$, то $F \to \infty$. Порядок застосування $F-$статистики такий же, як і $t-$статистики. Нуль-гіпотезою в даному випадку є припущення про те, що модель не адекватна в цілому, тобто

$$H_0 : a_1 = a_2 = \ldots = a_p = 0,$$

проти альтернативної гіпотези:

H_1 : *хоча б одне значення* a_i *відмінне від нуля в статистичному смислі*.

Значення $F_{крит}$ знаходять із таблиць для $F-$розподілу. Послідовність застосування цієї статистики можна представити так:

1. Сформулювати нуль-гіпотезу стосовно адекватності моделі в цілому.

Наприклад, H_0: *модель не адекватна в цілому* (або

$$H_0 : a_1 = a_2 = \ldots = a_p = 0).$$

2. Розрахувати значення F для оціненої моделі (як правило, воно розраховується у всіх пакетах статистичної обробки даних).
3. Вибрати рівень значущості $\alpha = 1\%$, або $\alpha = 5\%$, або $\alpha = 10\%$.
4. Користуючись значеннями N, f і α, знайти критичне значення $F_{крит}$ з таблиць для $F-$ розподілу при $(p, N-p-1)$ ступенях свободи.
5. Перевірити нуль-гіпотезу: якщо $F > F_{крит}$,

то нуль-гіпотеза щодо неадекватності моделі в цілому відкидається на вибраному рівні значущості.

Коректне застосування запропонованої методики забезпечує побудову адекватної математичної моделі процесу, якщо статистичні (експериментальні) дані відповідають *вимогам представництва* та *інформативності*. Перша вимога означає, що вибірка даних повинна охоплювати досить довгий проміжок часу, щоб повністю відображати поведінку того режиму функціонування процесу, для якого будується модель. Вимога *інформативності* означає, що вибірка повинна містити в собі об'єм інформації, достатній для оцінювання коефіцієнтів моделі. Наприклад, якщо моделюється процес другого порядку, то вибірка повинна забезпечувати коректне обчислення першої та другої похідної. Іноді інформативність формально оцінюють за допомогою величини дисперсії процесу, а також за кількістю гармонійних складових, які містяться у процесі. Чим більше гармонік містить вибірка, тим вищою є її інформативність.

Умову інформативності даних пов'язують з *умовою достатнього збудження* процесу. Достатнє збудження означає, що вхідний сигнал повинен охоплювати всю смугу частот, які може пропускати на вихід процес (об'єкт). Тобто, вхідний вплив (сигнал) повинен охоплювати всю амплітудно-частотну характеристику процесу. Ця вимога залишається справедливою для процесів будь-якої природи.

8.8 Приклади побудови моделей за статистичними даними

Приклад 8.2. Розглянемо можливості побудови математичних моделей динаміки для таких макроекономічних процесів України: формування внутрішнього валового продукту (ВВП), індекс споживчих цін (ІСЦ) і грошовий агрегат М3.

Для побудови використано фактичні щомісячні дані з січня 1996 по січень 2005 року, всього 109 значень.

Значення кореляції між показниками:

- кореляція між ІСЦ та ВВП дорівнює -0.3043;

- кореляція між ІСЦ та М3 дорівнює -0.2491;
- кореляція між ВВП та М3 дорівнює 0.9317.

Корельованість ІСЦ з ВВП та агрегатом М3 незначна; в подальшому ця інформація буде використана при побудові альтернативних варіантів математичних моделей процесів, що розглядаються. Корельованість між ВВП і агрегатом М3 становить приблизно 0,932 – це велике значення, яке може негативно вплинути на якість оцінок моделі при використанні цих змінних у правій частині рівняння.

Модель індексу споживчих цін

Авторегресійна модель індексу споживчих цін

Спочатку розглянемо можливість опису індексу споживчих цін моделлю авторегресії з ковзним середнім. Моделі парної регресії та авторегресійні моделі найпростіші за своєю структурою, але досить часто вони дають можливість досягти високого ступеня адекватності досліджуваному процесу, прийнятного для подальшого застосування моделі.

Автокореляційна функція процесу наведена в таблиці 8.1. При побудові моделі, індекс споживчих цін позначимо через *isc* . У таблиці 8.1 прийнято такі скорочення: АКФ – автокореляційна функція, а ЧАКФ – часткова АКФ.

Таблиця 8.1. Автокореляційна функція процесу формування індексу споживчих цін (ІСЦ). Результаті використання системи EViews 3.1

Часові дані: 1996:01 2005:01						
Всього спостережень: 109						
АКФ	Часткова АКФ		АКФ	ЧАКФ	Q-стат.	Ймовірність
. |**** |	. |**** |	1	0.570	0.570	36.373	0.000
. |** |	.*|. |	2	0.230	-0.140	42.378	0.000
. |* |	. |* |	3	0.129	0.089	44.284	0.000
. |* |	. |. |	4	0.076	-0.022	44.957	0.000
. |. |	. |. |	5	0.064	0.043	45.433	0.000
. |* |	. |* |	6	0.120	0.101	47.137	0.000
. |* |	. |* |	7	0.180	0.088	50.960	0.000
. |* |	.*|. |	8	0.068	-0.136	51.515	0.000
. |. |	. |* |	9	0.052	0.105	51.840	0.000
. |* |	. |. |	10	0.072	-0.001	52.468	0.000

Як видно із наведених значень та графіку АКФ, при побудові моделей необхідно починати з моделей нижчих порядків, які часто мають прийнятну адекватність процесу і забезпечують високу якість прогнозу.

Результати оцінювання моделі авторегресії першого порядку за методом найменших квадратів наведені в таблиці 8.2.

Таблиця 8.2. Результати оцінювання моделі АР(1) для ІСЦ.

Метод оцінювання: метод найменших квадратів				
Скорегована часова вибірка даних: 1996:02 2005:01				
Використано спостережень: 108 після корегування				
Модель АР(1): I=C(1)+C(2)*I(-1)				
	Оцінки коефіцієнтів	Стандартна похибка	t-статистика	Ймовірність
C(1)	43.29386	6.624742	6.535178	0.0000
C(2)	0.569826	0.065703	8.672790	0.0000
R-квадрат	0.415067	Середнє залежної змінної		100.7407
Скорегований R-квадрат	0.409549	Станд. відхил. зал. змінної		1.506189
Станд. похибка регресії	1.157368	Інформ. критерій Акайке		3.148519
Сума квадратів похибок	141.9871	Критерій Шварца		3.198188
		Стат. Дарбіна-Уотсона		1.931805

Отримана модель АР(1):

$$isc(k) = a_0 + a_1\, isc(k-1) + \varepsilon(k) =$$

$$= 43{,}294 + 0{,}57\, isc(k-1) + e(k),$$

де $e(k)$ – залишки (похибки) моделі, значення яких можна знайти у відповідному файлі пакету програм, що застосовується для побудови моделі. Основні статистичні характеристики якості моделі:

$$R^2 = 0{,}415; \quad J = СКП = 141{,}99; \quad DW = 1{,}931.$$

Коефіцієнт детермінації має низьке значення (0,415), сума квадратів похибок досить високе (141,99), а статистика Дарбіна-Уотсона (1,931) наближається до найкращого значення. Таким чином, загалом адекватність моделі АР(1) досить низька, а тому структура моделі потребує уточнення. Характеристики якості (історичного, на навчальній вибірці) однокрокового прогнозу:

$$СеКП = 1{,}36; \quad САП = 1{,}02; \quad САПП = 1{,}008; \quad U = 0{,}0067,$$

тобто середньоквадратична похибка (*СеКП*), середня абсолютна похибка (*САП*), середня абсолютна похибка у процентах (*САПП*) і коефіцієнт Тейла U, який свідчить про загальну придатність моделі для прогнозування (його ідеальне значення – нуль).

У таблиці 8.3 наведено результати оцінювання авторегресійної моделі АР(3). Усі коефіцієнти моделі статистично значущі.

Таким чином, можна записати таку модель:

$$isc(k) = a_0 + a_1\, isc(k-1) + a_2\, isc(k-2) + a_3\, isc(k-3) + \varepsilon(k) =$$
$$= 46{,}96 + 0{,}61\, isc(k-1) - 0{,}15\, isc(k-2) + 0{,}08\, isc(k-3) + e(k).$$

Таблиця 8.3. Результати оцінювання моделі АР(3).

Метод оцінювання: метод найменших квадратів				
Скорегована часова вибірка даних: 1996:04 2005:01				
Всього спостережень після корегування крайніх значень: 106				
Модель: I=C(1)+C(2)*I(-1)+C(3)*I(-2)+C(4)*I(-3)				
	Оцінки коеф-в	Станд. похиб.	t-статистика	Ймовірність
C(1)	46.96556	9.330171	5.033729	0.0000
C(2)	0.613458	0.098563	6.223995	0.0000
C(3)	-0.157670	0.113613	-1.387787	0.1682
C(4)	0.077517	0.086493	0.896229	0.3722
R-квадрат	0.316655	Середнє залежної змінної		100.6604
Скорегований R-квадрат	0.296556	Станд. відхил. залеж. змінної		1.372428
Станд. похибка регресії	1.151076	Інформ. критерій Акайке		3.156277
Сума квадратів похибок	135.1476	Критерій Шварца		3.256785
		Стат. Дарбіна-Уотсона		1.992389

Для цієї моделі спостерігається зменшення коефіцієнта детермінації від 0,415 до 0,317 (деяке погіршення); зменшення суми квадратів похибок від 141,99 до 135,148 і покращення значення статистики Дарбіна-Уотсона від 1,931 до 1,992 (похибки моделі можна вважати практично некорельованими). Тобто, отримані такі значення характеристик моделі:

$$R^2 = 0{,}317; \quad J = СКП = 135{,}148; \quad DW = 1{,}992.$$

Характеристики однокрокового прогнозу для даної моделі:

$$СеКП = 1{,}36; \quad САП = 1{,}02; \quad САПП = 1{,}01; \quad U = 0{,}0068,$$

тобто, модель загалом придатна для прогнозування (за коефіцієнтом Тейла, який наближається до ідеального нульового значення), а три інших показники свідчать про високу точність прогнозу. Необхідно зазначити, що показники якості прогнозу для моделей АР(1) і АР(3) є практично однаковими.

Розглянемо характеристики моделі вищого порядку. У таблиці 8.4 наведені результати оцінювання моделі АР(7).

Таблиця 8.4. Результати оцінювання моделі АР(7) для ІСЦ

Метод оцінювання: метод найменших квадратів				
Скорегована часова вибірка даних: 1996:08 2005:01				
Всього використано спостережень після корегування крайніх значень: 102				
Модель: I=C(1)+C(2)*I(-1)+C(3)*I(-2)+C(4)*I(-3)+C(5)*I(-5)+C(6)*I(-6)+C(7)*I(-7)				
	Оцінки коеф.	Станд. похибка	t-статистика	Ймовірність
C(1)	33.13446	12.79240	2.590168	0.0111
C(2)	0.603211	0.101625	5.935625	0.0000
C(3)	-0.184907	0.117804	-1.569623	0.1198
C(4)	0.125878	0.101891	1.235415	0.2197
C(5)	-0.031530	0.102954	-0.306254	0.7601
C(6)	0.093246	0.117601	0.792904	0.4298
C(7)	0.064850	0.089527	0.724365	0.4706
R-квадрат	0.346348	Середнє залежної змінної		100.6667
Скорегований R-квадрат	0.305064	Станд. відхил. залеж. змінної		1.388306
Станд. похибка регресії	1.157330	Інформ. критерій Акайке		3.196268
Сума квадратів похибок	127.2443	Критерій Шварца		3.376413
		Стат. Дарбіна-Уотсона		1.811821

Порівнюючи моделі АР(3) і АР(7), можна сказати, що коефіцієнт детермінації збільшився від 0,317 до 0,346; сума квадратів похибок моделі зменшилась від 135,15 до 127,24, а статистика Дарбіна-Уотсона зменшилась від 1,992 до 1,811. Характеристики однокрокового прогнозу для АР(7):

$$СеКП = 1{,}36; \quad САП = 1{,}012; \quad САПП = 1{,}002; \quad U = 0{,}0067,$$

тобто, на 0,008 зменшились середня абсолютна похибка і середня абсолютна похибка в процентах.

Результати оцінювання моделі 12-го порядку наведено у таблиці 8.5. Модель АР(12) має кращі характеристики, ніж попередні моделі:

$$R^2 = 0{,}435; \quad J = СКП = 97{,}80; \quad DW = 1{,}94.$$

Значно зменшилась сума квадратів похибок, підвищилось значення R^2, а значення статистики Дарбіна-Уотсона майже таке ж, як для моделі АР(3).

Таблиця 8.5. Результати оцінювання моделі АР(12) для ІСЦ

Метод оцінювання: метод найменших квадратів				
Скорегована часова вибірка даних: 1997:01 2005:01				
Всього використано спостережень після корегування крайніх значень: 97				
Модель: I=C(1)+C(2)*I(-1)++C(3)*I(-2)+C(4)*I(-3)+C(5)*I(-6)+C(6)*I(-7)+C(7)*I(				
-8)+C(8)*I(-9)+C(9)*I(-11)+C(10)*I(-12)				
	Оцінки коеф-в	Станд. похибка	t-статистика	Ймовірність
C(1)	26.23923	14.28763	1.836500	0.0697
C(2)	0.666410	0.102198	6.520741	0.0000
C(3)	-0.247996	0.122488	-2.024654	0.0460
C(4)	0.179378	0.103248	1.737345	0.0859
C(5)	-0.038240	0.096715	-0.395392	0.6935
C(6)	0.016650	0.112406	0.148120	0.8826
C(7)	-0.037042	0.111025	-0.333639	0.7395
C(8)	0.030055	0.095744	0.313915	0.7543
C(9)	0.020445	0.093392	0.218911	0.8272
C(10)	0.149194	0.081977	1.819964	0.0722
R-квадрат	0.435002	Середнє залежної змінної		100.6082
Скорегований R-квадрат	0.376553	Станд. відхил. залеж. змінної		1.342857
Станд. похибка регресії	1.060301	Інформ. критерій Акайке		3.052366
Сума квадратів похибок	97.80880	Критерій Шварца		3.317800
		Статистика Дарбіна-Уотсона		1.940130

Характеристики однокрокового прогнозу для цієї моделі такі:

$$СеКП = 1{,}337; \quad САП = 1{,}02; \quad САПП = 1{,}013; \quad U = 0{,}0066.$$

Таким чином, характеристики однокрокового прогнозу також найкращі для моделі АР(7). Можна зробити висновок, що процес формування індексу оптових цін може бути описаний моделлю авторегресії АР(7) з високим ступенем адекватності. Ця модель забезпечує також отримання кращого однокрокового прогнозу.

- *Модель авторегресії для відхилень ІСЦ від середнього*

Якщо, з вихідних (фактичних) значень ряду ІСЦ відняти середнє $(isc(k) - 100{,}66)$, $k = 1,\ldots,109$, то отримаємо ряд відхилень від середнього. Автокореляційна функція залишається фактично незмінною. Результати оцінювання моделі АР(1) наведені в таблиці 8.6.

Таблиця 8.6. Результати оцінювання моделі АР(1) для відхилень ІСЦ

Метод оцінювання: метод найменших квадратів				
Скорегована часова вибірка даних: 1996:02 2005:01				
Всього використано спостережень після корегування крайніх значень: 109				
Модель: IV=C(1)+C(2)*IV(-1)				
	Оцінки коеф-в	Станд. похибка	t-статистика	Ймовірність
C(1)	-0.007477	0.111831	-0.066857	0.9468
C(2)	0.569826	0.065703	8.672790	0.0000
R-квадрат	0.415067	Середнє залежної змінної		0.080741
Скорегований R-квадрат	0.409549	Станд. відхил. залеж. змінної		1.506189
Станд. похибка регресії	1.157368	Інформ. критерій Акайке		3.148519
Сума квадратів похибок	141.9871	Критерій Шварца		3.198188
		Стат. Дарбіна-Уотсона		1.931805

Три вибрані статистичні характеристики адекватності цієї моделі:

$$R^2 = 0{,}415; \quad J = СКП = 141{,}99; \quad DW = 1{,}931.$$

Характеристики якості однокрокового прогнозу:

$$СеКП = 1{,}36; \quad САП = 1{,}02; \quad САПП = 100{,}40; \quad U = 0{,}662.$$

У порівнянні з моделлю АР(1) для повних значень (без віднімання середнього) середньоквадратична похибка і середня абсолютна похибка не змінились, але в 100 разів збільшилась *САПП* і коефіцієнт Тейла: від 0,0067 до 0,662. Таким чином, модель АР(1) для відхилень не придатна для прогнозування. Це можна пояснити тим, що відхилення мають різні знаки, що ускладнює побудову моделі.

Авторегресія 13-го порядку для відхилень ІСЦ від середнього наведена у табл. 8.7.

Таблиця 8.7. Результати оцінювання моделі AP(13) для відхилень ІСЦ

Метод оцінювання: метод найменших квадратів				
Скорегована часова вибірка даних: 1997:02 2005:01				
Всього використано спостережень після корегування крайніх значень: 96				
Модель: IV=C(1)+C(2)*IV(-1)+C(3)*IV(-2)+C(4)*IV(-3)+C(5)*IV(-5)+C(6)*IV(-6)				
+C(7)*IV(-7)+C(8)*IV(-8)+C(9)*IV(-9)+C(10)*IV(-11)+C(11)*IV(				
-12)+C(12)*IV(-13)				
	Оцінки коеф-в	Станд. похибка	t-статистика	Ймовірність
C(1)	-0.045818	0.110205	-0.415756	0.6786
C(2)	0.706976	0.107583	6.571429	0.0000
C(3)	-0.261301	0.123875	-2.109395	0.0379
C(4)	0.194879	0.105046	1.855172	0.0671
C(5)	-0.110121	0.107924	-1.020352	0.3105
C(6)	0.043907	0.120182	0.365334	0.7158
C(7)	0.002351	0.115561	0.020347	0.9838
C(8)	-0.023657	0.114078	-0.207375	0.8362
C(9)	0.033030	0.096947	0.340705	0.7342
C(10)	0.002612	0.096331	0.027112	0.9784
C(11)	0.187169	0.109248	1.713242	0.0904
C(12)	-0.074932	0.085883	-0.872494	0.3854
R-квадрат	0.440922	Середнє зал. змінної		-0.066250
Скорегований R-квадрат	0.367709	Станд. відхил. залеж. змінної		1.342254
Станд. похибка регресії	1.067316	Інформ. крит. Акайке		3.084639
Сума квадратів похибок	95.68969	Критерій Шварца		3.405183
Логарифм правдоподібн.	-136.0627	Статистика Дарбіна-Уотсона		1.965193

Три вибрані статистичні характеристики адекватності цієї моделі:

$$R^2 = 0{,}440; \quad J = СКП = 95{,}69; \quad DW = 1{,}965.$$

Характеристики якості однокрокового прогнозу:

$$СеКП = 1{,}346; \quad САП = 1{,}02; \quad САПП = 109{,}09; \quad U = 0{,}848.$$

Спостерігається погіршення характеристик прогнозу, особливо погіршилось значення *САПП* і коефіцієнта Тейла: від 0,662 для моделі AP(1) до 0,848 для моделі AP(13). За цим параметром модель не придатна для прогнозування. Коефіцієнти C(6) – C(10) є статистично незначущими, а тому їх можна вилучити з моделі.

Після вилучення цих коефіцієнтів із моделі та відповідних їм складових процесу отримаємо модель, параметри якої наведені в таблиці 8.8.

Таблиця 8.8. Результати оцінювання моделі АР(13) для відхилень

Метод оцінювання: метод найменших квадратів				
Скорегована часова вибірка даних: 1997:02 2005:01				
Всього використано спостережень після корегування крайніх значень: 96				
Модель: IV=C(1)+C(2)*IV(-1)+C(3)*IV(-2)+C(4)*IV(-3)+C(5)*IV(-5)+C(6)*IV(				
-12)+C(7)*IV(-13)				
	Оцінки коеф-в	Станд. похибка	t-статистика	Ймовірність.
C(1)	-0.046129	0.107257	-0.430077	0.6682
C(2)	0.700062	0.103640	6.754728	0.0000
C(3)	-0.256660	0.120078	-2.137440	0.0353
C(4)	0.195599	0.101703	1.923247	0.0576
C(5)	-0.081877	0.083177	-0.984368	0.3276
C(6)	0.190417	0.089745	2.121760	0.0366
C(7)	-0.067455	0.081058	-0.832180	0.4075
R-квадрат	0.438765	Середнє залеж. змінної		-0.066250
Скорегований R-квадрат	0.400929	Станд. відхил. залеж. змінної		1.342254
Станд. похибка регресії	1.038900	Інформ. критерій Акайке		2.984323
Сума квадратів похибок	96.05881	Критерій Шварца		3.171307
Логарифм правдоподібн.	-136.2475	Статистика Дарбіна-Уотсона		1.959533

Вибрані статистичні характеристики адекватності цієї моделі:

$$R^2 = 0{,}439; \quad J = СКП = 96{,}05; \quad DW = 1{,}96.$$

Характеристики якості однокрокового прогнозу:

$$СеКП = 1{,}346; \quad САП = 1{,}017; \quad САПП = 108{,}2; \quad U = 0{,}854.$$

Таким чином, після видалення з моделі незначимих коефіцієнтів її характеристики залишились практично незмінними. В цілому можна зробити висновок, що побудовані авторегресійні моделі для відхилень ІСЦ від середнього не придатні для прогнозування. Це свідчить про те, що побудувати модель для процесу, який має різні знаки вимірів у різні моменти часу (тобто розвиток відбувається в двох квадрантах), складніше, ніж для процесу, який змінюється в межах одного квадранту.

Авторегресія з ковзним середнім для індексу споживчих цін

Розглянемо можливість описання ІСЦ за допомогою моделі АРКС. Характеристики моделі АРКС(1,1) наведені в таблиці 8.9.

Таблиця 8.9. Результати оцінювання моделі АРКС(1, 1) для ІСЦ

Метод оцінювання: метод найменших квадратів				
Скорегована часова вибірка даних: 1996:02 2005:01				
Всього використано спостережень після корегування крайніх значень: 108				
Змінна	Оцінки коеф-в	Станд. похибка	t-статистика	Ймовірність
C	100.6529	0.263645	381.7742	0.0000
AR(1)	0.546826	0.093191	5.867793	0.0000
КС(1)	0.060696	0.133153	0.455838	0.6494
R-квадрат	0.415932	Середнє залеж. змінної		100.7407
Скорегований R-квадрат	0.404807	Станд. відхил. залеж. змінної		1.506189
Станд. похибка регресії	1.162006	Інформ. критерій Акайке		3.165558
Сума квадратів похибок	141.7772	Критерій Шварца		3.240062
Логарифм правдоподібн.	-167.9401	F-статистика		37.38676
Статист. Дарбіна-Уотсона	1.996202	Ймовірність (F-стат.)		0.000000
Інвертовані АР корені	.55			
Інвертовані КС корені	-.06			

Вибрані статистичні характеристики адекватності цієї моделі:

$$R^2 = 0{,}416; \quad J = СКП = 141{,}78; \quad DW = 1{,}996.$$

Характеристики якості однокрокового прогнозу:

$$СеКП = 1{,}362; \quad САП = 1{,}016; \quad САПП = 1{,}005; \quad U = 0{,}0067.$$

Моделі АР(1) і АРКС(1,1) мають практично однакові характеристики адекватності та якості однокрокового прогнозу, а тому перевагу (при виборі з цих двох моделей) можна надати моделі АР(1), яка є простішою. Нижче наведена порівняльна таблиця для всіх побудованих моделей (табл.8.12).

Врахування впливу на індекс споживчих цін агрегату М3

Коефіцієнт кореляції між ІСЦ та агрегатом М3 від'ємний: –0,249, тобто формальний взаємозв'язок між цими змінними незначний, але цікаво розглянути вплив М3 на ІСЦ за допомогою моделі. Врахування регресора може покращити деякі характеристики моделі, а також врахувати причинний зв'язок між вибраними змінними.

Характеристики змішаної моделі: авторегресія АР(1) + парна регресія наведені у таблиці 8.10.

Спостерігається незначне покращення характеристик моделі та якості прогнозу у порівнянні з АР(1), але коефіцієнт при М3 дуже малий (-2,82E-06). Однак формально він є значимим. Таким чином, на розглянутому часовому інтервалі вплив агрегату М3 на індекс споживчих цін незначний і ним можна знехтувати.

Таблиця 8.10. Результати оцінювання моделі ІСЦ: АР(1) + регресор М3

Метод оцінювання: метод найменших квадратів				
Скорегована часова вибірка даних: 1996:02 2005:01				
Всього використано спостережень після корегування крайніх значень: 108				
Модель: I=C(1)+C(2)*I(-1)+C(3)*M				
	Оцінки коеф-в	Станд. похибка	t-статистика	Ймовірність
C(1)	44.97869	6.920057	6.499757	0.0000
C(2)	0.554264	0.068261	8.119764	0.0000
C(3)	-2.82E-06	3.30E-06	-0.854463	0.3948
R-квадрат	0.419106	Середнє залежної змінної		100.7407
Скорегований R-квадрат	0.408041	Станд. відхил. залеж. змінної		1.506189
Станд. похибка регресії	1.158844	Інформ. критерій Акайке		3.160108
Сума квадратів похибок	141.0066	Критерій Шварца		3.234612
Логарифм правдоподібн.	-167.6458	Стат. Дарбіна-Уотсона		1.919823

Характеристики однокрокового прогнозу:

$$СеКП = 1{,}340; \quad САП = 1{,}004; \quad САПП = 0{,}994; \quad U = 0{,}0066.$$

Можна припустити, що в обігу було недостатньо грошової маси для того, щоб її вплив на ІСЦ був істотним. З іншого боку, недостатній об'єм грошової маси в національній валюті компенсувався (і продовжує компенсуватись) "твердою" іноземною валютою, зокрема доларами США і, деякою мірою, євро. Таким чином, слабка українська економіка позитивно впливає на курс долара США, завдяки фактичному введенню його в частковий обіг на відносно великій території України.

Встановити фактичне співвідношення між об'ємами національної та іноземної валют в обігу можливо, але для цього необхідно отримати додаткові статистичні дані і виконати спеціальне дослідження. Зокрема, необхідно мати дані стосовно об'ємів реалізації торгівельних операцій у валюті підприємствами усіх форм власності. Очевидно, що отримати такі дані для тіньового обігу досить складно.

Визначення впливу ВВП на ІСЦ

Як було показано на початку цього параграфа, коефіцієнт кореляції між ІСЦ та ВВП складає –0,304, тобто формально це невелике значення. Також логічно припустити, що зростання ВВП має призводити до зменшення ІСЦ (про це свідчить також знак коефіцієнта кореляції між цими змінними).

Характеристики змішаної моделі: авторегресія АР(1) + парна регресія наведені в таблиці 8.11.

Таблиця 8.11. Результати оцінювання змішаної регресії для ІСЦ і ВВП

Метод оцінювання: метод найменших квадратів				
Скорегована часова вибірка даних: 1996:02 2005:01				
Всього використано спостережень після корегування крайніх значень: 108				
Модель: I=C(1)+C(2)*I(-1)+C(3)*V				
	Оцінки коеф-в	Станд. похибка	t-статистика	Ймовірність .
C(1)	45.53125	7.156626	6.362112	0.0000
C(2)	0.549611	0.070124	7.837741	0.0000
C(3)	-1.37E-05	1.64E-05	-0.833627	0.4064
R-квадрат	0.418913	Середнє залежної змінної		100.7407
Скорегований R-квадрат	0.407844	Станд. відхил. залеж. змінної		1.506189
Станд. похибка регресії	1.159037	Інформ. критерій Акайке		3.160441
Сума квадратів похибок	141.0536	Критерій Шварца		3.234945
Логарифм правдоподібн.	-167.6638	Стат. Дарбіна-Уотсона		1.916289

Отримано таке рівняння:

$$isc(k) = 45{,}53 + 0{,}55\ isc(k-1) - (1{,}37E-05)\,vvp(k) + e(k)$$

з характеристиками:

$$R^2 = 0{,}419; \quad J = СКП = 141{,}05; \quad DW = 1{,}92.$$

Характеристики якості однокрокового прогнозу:

$$СеКП = 1{,}335; \quad САП = 1{,}004; \quad САПП = 0{,}993; \quad U = 0{,}0066.$$

Коефіцієнт при змінній *ВВП* невеликий і знаходиться на межі статистичної значимості, тобто вплив *ВВП* на *ІСЦ* незначний. Причиною такого незначного впливу може бути некоректний розподіл *ВВП*, який призводить до того, що більшість населення проживає на межі або нижче межі бідності. Одночасно інша (менша) частина привласнює більшість благ і користується ними, але очевидно, що це користування не призводить до позитивного впливу на *ІСЦ*. Звідси маємо, що *ВВП* зростає значними темпами, а добробут населення помітно від нього відстає. Таким чином, можна зробити висновок, що на часовому періоді, який розглядається в даному прикладі, відбувався несправедливий розподіл суспільних благ, що проявилось у незначному впливі *ВВП* на споживчі ціни для більшості населення України. Побудована математична модель є формальним (об'єктивним) підтвердженням відомого факту нерівномірного розподілу матеріальних благ, вироблених в Україні.

У таблиці 8.12 зведені характеристики математичних моделей, побудованих для індексу споживчих цін, і характеристики однокрокових прогнозів, обчислених на основі цих моделей. Ця таблиця дає можливість оперативно порівняти результати моделювання та прогнозування, отримані за

допомогою методики Бокса-Дженкінса, а також встановити можливість практичного використання цих результатів.

Таблиця 8.12. Результати моделювання і однокрокового прогнозування індексу споживчих цін

Тип моделі	Характеристики моделі			Характеристики однокрокового прогнозу			
	R^2	$\sum e^2(k)$	*DW*	*СеКП*	*САП*	*САПП*	Коеф-т Тейла
АР(1)	0,415	141,99	1,931	1,360	1,020	1,008	0,0067
АР(3)	0,317	135,148	1,992	1,360	1,020	1,011	0,0068
АР(7)	0,346	127,244	1,811	1,360	1,012	1,002	0,0067
АР(12)	0,435	97,80	1,941	1,337	1,020	1,013	0,0066
АРКС(1,1)	0,416	141,78	1,996	1,362	1,016	1,005	0,0067
АР(1)+М3	0,419	141,007	1,919	1,340	1,004	0,994	0,0066
АР(1)+ВВП	0,419	141,054	1,916	1,335	1,004	0,993	0,0066
Моделі для відхилень ІСЦ від середнього							
АР(1)	0,415	141,99	1,931	1,360	1,020	100,40	0,662
АР(13)	0,440	95,69	1,965	1,346	1,020	109,09	0,848

Результати моделювання, наведені в таблиці 8.12, свідчать про те, що практично всі моделі, побудовані для ІСЦ (перші сім моделей), є придатними для прогнозування, оскільки коефіцієнт Тейла вимірюється тисячними частками. Характеристики однокрокового прогнозу для моделей різного порядку відрізняються несуттєво. Найкращі характеристики щодо прогнозування має модель АР(1)+ВВП. Для моделі АР(12) отримано найменше значення суми квадратів похибок моделі, але характеристики прогнозів, отриманих за цією моделлю, не кращі від інших.

Дві моделі, які побудовані для відхилень ІСЦ від середнього, мало-придатні для прогнозування. Про це свідчать високі значення коефіцієнта Тейла та середньої абсолютної похибки у процентах. Цей факт можна пояснити тим, що описати за допомогою АРКС процес, динаміка якого спостерігається у двох квадрантах, складніше, ніж процес, який спостерігається в одному квадранті. Тому, для прогнозування відхилень від середнього необхідно знайти іншу структуру математичної моделі, придатну для опису різнознакових величин часового ряду.

8.9 Контрольні запитання і вправи

1. Назвіть етапи побудови математичних моделей за методикою Бокса-Дженкінса. Що забезпечує коректне використання цієї методики на практиці?
2. Яка мета аналізу функціонування процесу? Які елементи структури моделі можна встановити за допомогою попереднього аналізу процесу на основі відомої інформації?

3. Яка мета попередньої обробки даних? Назвіть основні операції, які виконують у процесі попередньої обробки даних. До чого призводить визначення значень змінних у великому числовому діапазоні?
4. Які два основних типи нелінійностей зустрічаються в аналізі часових рядів? Який з них ускладнює процедуру оцінювання параметрів моделі? Поясніть на прикладах.
5. Яким чином можна встановити наявність нелінійностей у процесі? У яких випадках нелінійний процес можна описати лінійною моделлю?
6. Який метод дає можливість автоматизувати процес визначення та врахування нелінійностей процесу?
7. Яку інформацію можна отримати на основі візуального аналізу даних? Як можна нею скористатись?
8. Яким чином можна знайти оцінку порядку авторегресійної частини моделі?
9. У чому полягає відмінність між автокореляційною та частковою автокореляційною функціями процесу? Чи існує необхідність розрахунку обох функцій у процесі аналізу даних?
10. На чому ґрунтується відбір незалежних змінних (регресорів, екзогенних змінних) для включення у праву частину математичної моделі?
11. Назвіть три умови коректного застосування методу найменших квадратів до оцінювання параметрів математичної моделі. Як можна перевірити виконання цих умов?
12. Для чого призначена статистика Люнга-Бокса і як вона обчислюється? Скористайтесь наявним пакетом програм для статистичного аналізу даних для обчислення цього статистичного параметра і поясніть його значущість (чи незначущість) на прикладі.
13. У чому полягає принципова різниця між методом найменших квадратів (МНК), призначеним для оцінювання лінійних моделей, та МНК для оцінювання моделей, нелінійних відносно параметрів? Які критерії якості мінімізують ці методи?
14. Виведіть формулу МНК для поліноміальної моделі.
15. Сформуйте матрицю вимірів для математичної моделі такого вигляду:

$$y(k) = a_0 + a_1\, y(k-1) + b_1\, x_1(k) + b_2\, x_2(k) + \varepsilon(k).$$

16. Про що свідчить корельованість похибок моделі між собою? За допомогою якого статистичного параметра якості моделі можна визначити ступінь корельованості похибок? Яке значення приймає ця статистика в ідеальному випадку?
17. Що означає значущість оцінки параметра (коефіцієнта) моделі в статистичному смислі? За допомогою якої статистики можна встановити значущість оцінки параметра моделі?
18. Що означають помилки першого і другого роду при перевірці статистичних гіпотез?
19. Поясніть фізичну суть коефіцієнта (множинної) детермінації, яке його ідеальне значення? Чому R^2 корегують із врахуванням кількості ступенів свободи?

20. У чому полягає різниця між критеріями Акайке та Байєса-Шварца? Чи можна обмежитись використанням тільки одного з цих критеріїв?
21. Яка мета розрахунку статистики Фішера? Наведіть послідовність розрахунку та аналізу цієї статистики.
22. Сформулюйте правила перевірки гіпотез для $t-$статистики Стьюдента і $F-$статистики Фішера. Яким чином можна визначити критичні (порогові) значення цих статистик?
23. Поясніть фізичну суть умови достатнього збудження процесу. До чого призводить невиконання цієї умови?
24. Яким чином досягають достатнього збудження процесів (об'єктів) на практиці? Які сигнали використовують для досягнення цієї мети? Які властивості цих сигналів використовують у даному випадку?
25. Нарисуйте наближений графік перетворення Фур'є для білого шуму і поясніть його.
26. Які інші ортогональні перетворення даних ви знаєте, крім перетворення Фур'є? Чому на практиці використовують декілька різних ортогональних перетворень, а не одне?
27. Нарисуйте графік перетворення Фур'є для гармонічної функції: $y(k)=a_1\sin(\omega_1 k+\varphi_1)+a_2\cos(\omega_2 k+\varphi_2)$.

РОЗДІЛ 9

STATISTICAL ANALYSIS SYSTEM – СУЧАСНИЙ ІНСТРУМЕНТ СТАТИСТИЧНОЇ ОБРОБКИ ДАНИХ

9.1 Історичні відомості щодо системи SAS

SAS Institute – міжнародна приватна компанія, що займається розробкою програмного забезпечення і засобів бізнес-аналітики, заснована в 1976 році Ентоні Барром (Anthony Barr), Джеймсом Гуднайтом (James Goodnight), Джоном Саллом (John Sall) та Джейном Хальвінгом (Jane Helving). Свою назву SAS Institute отримала від скорочення повної назви – Statistical Analysis System, хоча сьогодні перелік функцій, що реалізовані у продуктах компанії вийшов далеко за межі статистичного аналізу.

Голова компанії Джеймс Гуднайт народився 6 січня 1943 року в Солсбері, штат Північна Кароліна (Salisbury, North Carolina). Коли Джеймсу було дванадцять років, сім'я переїхала в Уілмінгтон (Wilmington, North Carolina), де він закінчив середню школу New Hanover. У шкільні роки, завдяки вчителю хімії, захопився її вивченням і досяг не абияких успіхів з хімії та математики. Після закінчення школи вступив до Університету штату Північної Кароліни в Ролі (англ. North Carolina State University at Raleigh, NCSU), де почав серйозно займатися комп'ютерною технікою та програмуванням. Першим професійним досвідом програмування було написання програм для економічного відділу департаменту сільського господарства під час літніх канікул.

Працюючи над магістерською дисертацією поглиблено вивчав задачі, пов'язаними з Луною та програмою Apollo, завдяки чому, після закінчення університету отримав посаду програміста в технологічній компанії, що розробляла електронне обладнання для наземних станцій, які забезпечували зв'язок з капсулою Apollo. Пізніше Джеймс Гуднайт здобув ступінь доктора в галузі статистики Університету штату Північна Кароліна, де працював з 1972 по 1976 роки. В тому ж університеті професорами-викладачами працювали і співзасновники компанії SAS Institute Джон Солл та Ентоні Барр.

Відправною точкою створення компанія SAS Institute став науково-дослідний проект з аналізу даних для сільського господарства в Університеті штату Північної Кароліни, що розпочався у 1966 році, коли виникла необхідність розробки спеціалізованого програмного забезпечення для автоматизації обробки великої кількості сільськогосподарських даних, зібраних в рамках гранту Міністерства сільського господарства США (USDA). Проект було реалізовано спеціально створеним Університетом Статистики Південних Експериментальних Станцій (The University Statisticians of the Southern Experiment Stations, USSES) - консорціумом з восьми південних університетів

США. Лідером серед них був Державний університет Північної Кароліни, завдяки тому, що більшість програмістів були його співробітниками і він добре фінансувався за рахунок гранту від Національного інституту здоров'я. Основною задачею консорціуму була розробка статистичного програмного забезпечення. Не останню роль в розвитку проекту відіграла компанія IBM, для якої цей проект був вкрай вигідним, завдяки застосуванню в ньому IBMSystem/360, яка в свою чергу, потребувала перепрограмування всього програмного забезпечення створеного для комп'ютерів попереднього покоління.

В подальшому всі учасники проекту взяли участь у розробці універсального статистичного пакету прикладних програм, призначеного для аналізу сільськогосподарських даних в рамках гранту від Національного Інституту Здоров'я (англ. National Institute of Health, NIH).

В 1972 році Національний Інститут Здоров'я припинив фінансування проекту. Але члени консорціуму продовжували фінансування з метою підтримки розвитку проекту, який ставав дедалі популярнішим. В результаті – вже у 1976 році більше 100 постійних клієнтів були споживачами їх програмного забезпечення. Саме в 1976 році команда, що працювала під керівництвом професорів-викладачів Ентоні Барра, Джеймса Гуднайта, Джона Солла, а також їхнього студента-випускника програміста – Джейна Хальвіга вирішила покинути університет з метою заснування власної компанії – SAS Institute. Свій перший офіс вони відкрили через дорогу від будівлі Університету штату Північної Кароліни в Ролі. Дохід компанії в перший рік існування склав 138 тис. доларів, впродовж цього ж 1976 року було прийнято на роботу близько 300 працівників, з яких біля 100 – вихідці з академічного середовища. Зростання SAS у наступне десятиліття було феноменальним. Новий кампус штаб-квартири виріс з однієї будівлі на Хіллсборо стрит, де працювали 50 співробітників, до 18 будинків, де крім офісних приміщень були навчальний центр, архів, відео студія.

Починаючи свій розвиток як науково-дослідний проект з аналізу даних для сільського господарства в Університеті штату Північної Кароліни, компанія SAS Institute, сьогодні є найбільшою у світі приватною IT-компанією. У її 400 офісах, розташованих по всьому світу працюють більше 13,6 тис. працівників [19].

Клієнтами SAS є більше 60 тисяч організацій у 135 країнах світу. Серед них – 90 компаній з першої сотні лідерів, що увійшли до списку "2012 FORTUNE Global 500®" [19] .

Частка компанії на ринку бізнес-аналітики складає більше 35%, що дозволяє їй бути безумовним світовим лідером в даній галузі. Також SAS посідає лідируючі позиції в різних технологічних рейтингах незалежних дослідників. Коло рішень та послуг, що пропонує компанія постійно розширюється, значною мірою завдяки цілеспрямованій інноваційній політиці, адже близько 25% доходу компанії щороку реінвестується в дослідження та

розробки (R&D) [19]. SAS належить до числа стало прибуткових компаній: з моменту заснування її дохід щороку зростає і за підсумками 2012 року досяг 2,87 млрд. доларів [19].

У 1996 році у Москві був відкритий перший в СНД офіс компанії, а в Україні – у 2007 році. З загального об'єму продажів в країнах Центральної та Східної Європи 34% припадає на Росію. За країнами СНД вони розподілені наступним чином: 85% – Росія, 12% – Казахстан, 2% – Україна. По об'єму продажів лідерами є телекомунікаційний сектор (44%), банки та страхові компанії (29%), транспорт (15%) і 12% – всі інші сектори економіки.

Базові статистичні програмні продукти SAS є конкурентами для таких рішень як DAP, GenStat, Angoss, Mathematica, Matlab, R, SPSS, Stata, STATISTICA, WPS та інших.

В цьому розділі буде коротко описано майже два десятки продуктів, але найбільш докладно – базовий продукт, призначений для статистичної обробки даних – SAS Base. Також будуть розглянуті приклади використання системи для вирішення типових задач прикладної статистики. Всі наведені приклади можуть бути повторені самостійно в будь-якому середовищі чи програмному забезпеченні, що підтримують мову програмування SAS Base.

9.2 Класифікація світових брендів бізнес інтелектуальних платформ за рейтингом Gartner і місце SAS у ньому

Магічні квадранти для бізнес інтелектуальних платформ [1, 19], що розробляються щорічно міжнародною аналітичною компанією Gartner, презентують глобальний огляд головних розробників програмного забезпечення. Наведена в них інформація зазвичай розглядається організаціями, установами та підприємствами при пошуку постачальників бізнес інтелектуальних рішень та програмного забезпечення.

За визначенням Gartner, "Бізнес інтелектуальні платформи повинні задовольняти всім типам користувачів – від IT фахівців-консультантів до бізнес користувачів – в створенні програмного забезпечення, що допомагає організаціям краще вивчати та розуміти власний бізнес" [19].

Gartner описує запропоновані категорії таким чином.

Leaders. Лідери – це досить потужні вендори, що забезпечують широку функціональність своїх бізнес аналітичних платформ та можуть забезпечити масштабованість рішення на рівні підприємств для підтримки загальної бізнес інтелектуальної стратегії. Завдяки чому керівництво здатне пропонувати покупцям бізнес пропозицію, що заснована на глобальному фундаменті, який підтримується життєздатністю та операційною потужності при реалізації.

Challengers. Претенденти на лідерство пропонують цікаві з точки зору функціональності бізнес інтелектуальні рішення та мають хороші можливості для досягнення успіху на ринку. При цьому. вони можуть бути обмежені певними умовами використання, технічними можливостями та сферою

застосування. Їх глобальне бачення може бути обмежене із-за відсутності координації. Стратегія для різних продуктів в їхньому портфелі бізнес інтелектуальних платформ може мати недоліки з точки зору маркетингових зусиль, каналів продажу, географічної присутності, галузевого змісту та розуміння, у порівнянні з квадрантом лідерів.

Visionaries. Провидці – це постачальники, що мають потужне бачення перспектив постачання бізнес інтелектуальних платформ. Вони відрізняються інноваційністю, відкритістю та гнучкістю архітектури рішень і пропонують поглиблену функціональність додатків в галузях, для яких призначені, але можуть мати суттєві недоліки, що пов'язані з реалізацією широкого кола функціональних вимог. Провидці – це ринкові лідери-новатори, але вони як правило. ще не досягли достатнього масштабу або зацікавлені в послідовному рості.

Nicheplayers. Нішові гравці – це ті хто має специфічний сегмент ринку, для якого розроблене бізнес інтелектуальне рішення, наприклад, звітність або панелі пристроїв. При цьому, вони дещо обмежені у інноваціях та переході на рішення інших постачальників. Вони можуть зосередитися на конкретній галузі або якомусь аспекті бізнес інтелектуальної платформи, але їм не вистачає поглибленої функціональності в інших місцях. Нішові гравці також мають обмежене використання та підтримку клієнтів, наприклад, географічні та індустріальні обмеження або не досягли необхідного масштабу, щоб закріпити свою позицію на ринку.

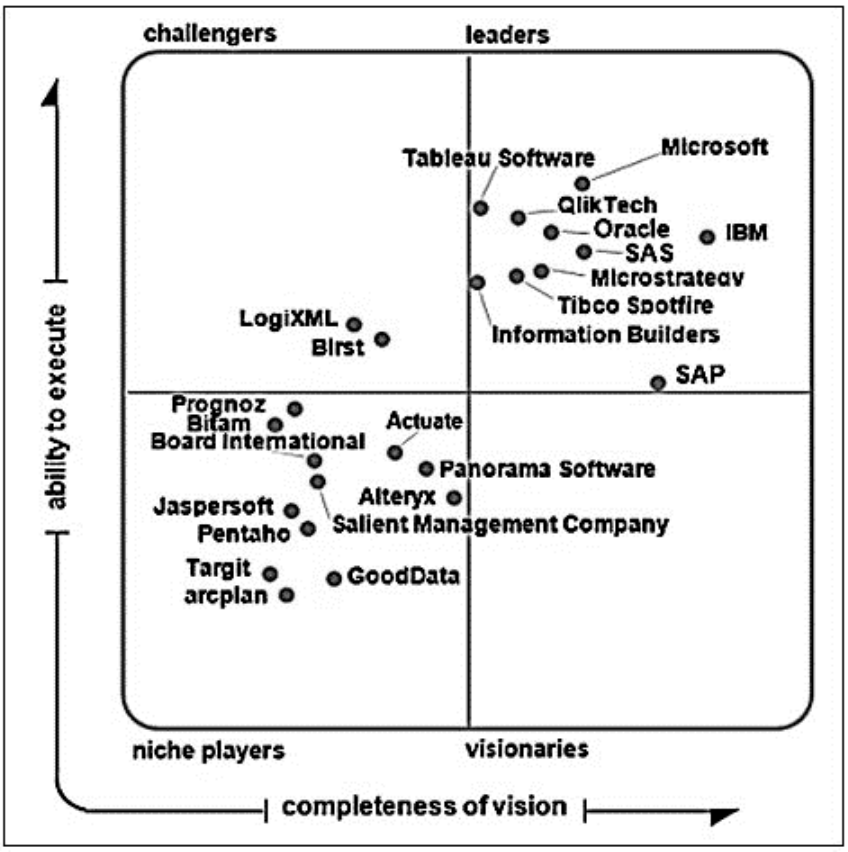


Рис. 9.*1* Магічний квадрант; порівняння Бізнес Інтелектуальних Платформ за лютий 2013 року[19]

Як видно рис. 9.1, згідно з наведеною у аналітичному звіті компанії Gartner за лютий 2013 року класифікації бізнес інтелектуальних платформ, компанія SAS належить до глобальних світових Лідерів і посідає перше місце серед усіх постачальників аналітичних рішень, задовольняє майже всі потреби клієнтів у якісній бізнес-аналітиці.

9.3 Короткий огляд продуктів системи SAS

SAS Base призначена для керування даними та виклику спеціалізованих статистичних процедур, також реалізовано можливість використання процедур SQL (Structured Query Language). Реалізовані у **SAS Base** механізми доступу до даних, надають можливість роботи з різноманітними структурами даних, в тому числі і Oracle. Є можливість створення звітів в файлах, форматів RTF, PDF, XML, HTML та звичайних текстових файлах. Середовище SAS реалізовано у вигляді вікон інтерактивного графічного інтерфейсу користувача, призначеного для запуску і тестування sas-програм. SAS Base є основою розроблених компанією продуктів та пов'язана з іншими модулями, як показано на рис. 9.2.

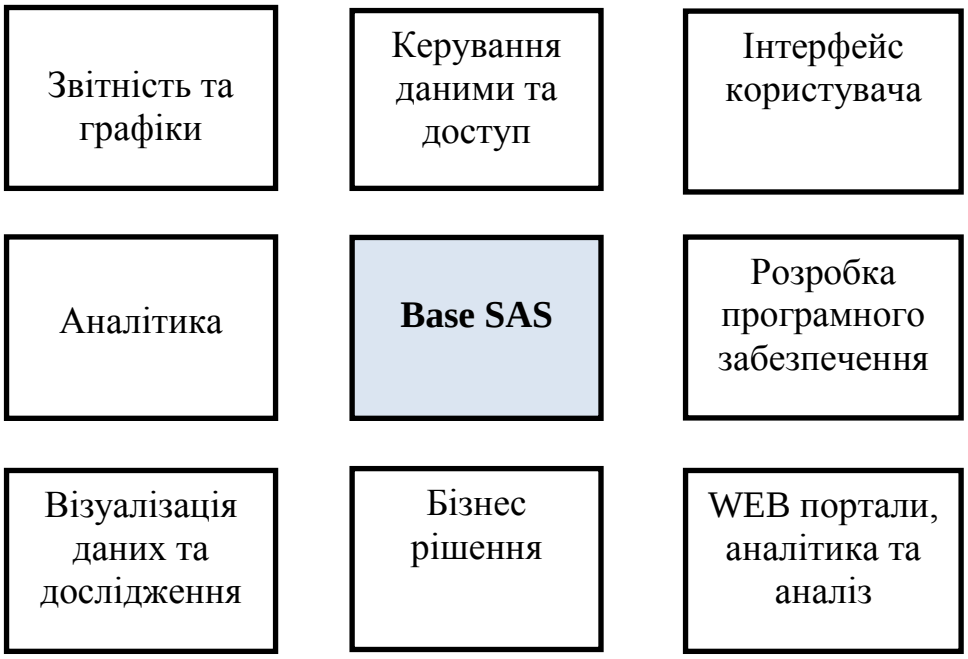


Рис. 9.*2* Місце SAS Base серед продуктів компанії

Окрім SAS Base, компанія SAS Institute пропонує багато інших програмних продуктів. Загалом, компанія пропонує більше 200 продуктів для різних галузей, призначених для зберігання, обробки та аналізу даних. Нижче наведено короткий опис близько двох десятків рішень та програм, розроблених компанією, обраних за суб'єктивним поглядом авторів з точки зору поточної популярності або перспектив використання в країнах СНД [19].

SAS App Dev Studio – програмний продукт для створення бізнес-аналітичних додатків на мові Java для різних платформ із використанням реалізованих в sas процедур та функцій.

SAS Enterprise Guide – програма, в якій реалізована технологія Drag and Drop розміщення різноманітних sas-процедур і функцій для побудови аналітичного процесу за принципами технологічного. Ця програма містить в собі процедури та функції SAS Base, які можна використовувати, як у вигляді sas-програм, так і технологічного процесу.

SAS/GRAPH – програмне забезпечення для візуалізації даних, використовується аналітиками та IT-керівниками для кращого розуміння процесів. Входить до складу багатьох індустріальних рішень, а також SAS Enterprise Guide.

SAS/ETS (Econometrics and Time-Series analysis) пропонує широкий спектр методів економетрики та обробки часових рядів, які дозволяють моделювати, аналізувати, прогнозувати та імітувати бізнес-процеси для поліпшення стратегічного і тактичного планування. Цей інструмент допомагає краще зрозуміти ефекти від впливу різних факторів, таких як економічні і ринкові умови, демографія, цінова політика і маркетингові заходи в бізнесі.

SAS/GIS (Geographic Information System) – інструмент для інтерактивного дослідження географічної інформації, що дозволяє систематизувати і аналізувати дані, які можна пов'язувати з фізичним місцем розташування. Використовується для демографічних, маркетингових та епідеміологічних досліджень.

SAS/IML (Interactive Matrix Language) – програмне забезпечення, в якому реалізована потужна та гнучка спеціалізована мова матричного програмування в динамічному середовищі (IML). Розраховане на програмістів, статистиків, дослідників і аналітиків. Підтримується робота з R-кодом, дозволяючи розширити стандартний функціонал. Має достатньо простий синтаксис, що дозволяє формалізувати математичні обчислення. Реалізовано велику кількість операцій над символами та числами. Використовується в навчальних закладах, поряд з такими системами як Matlab, MathCad та R.

SAS/OR (Operation Research) – програмне забезпечення для вирішення задач оптимізації, моделювання та календарного планування, визначення оптимальних стратегій дій в умовах обмежених ресурсів та часу. Його застосування дозволяє підприємствам та організаціям розглядати широкий спектр альтернативних дій і сценаріїв, оптимізувати розподіл ресурсів і визначати найкращі плани дій для досягнення мети.

SAS/STAT (Statistical Analysis) призначений як для звичайних, так і корпоративних аналітичних служб. Реалізує, як традиційні методи статистичного аналізу даних – дисперсійного аналізу і прогнозного моделювання, так і спеціалізовані – байєсівський, категоріальний, багатовимірний, живучості, психометричний, кластерний, непараметричний, планування та інші типи аналізу.

SAS Data Quality Solution – корпоративне рішення для профілювання, очищення, збагачення та інтеграції даних при створенні ланцюгів послідовної

надійної інформації.

SAS Clinical Data Integration – програмний продукт, призначений для організації процесу збору, зберігання, стандартизації та менеджменту медичних даних клінічних досліджень та метаданих, що забезпечує надійність клінічних даних при крос-дослідженні та поглибленому аналізі.

SAS Data Management Advanced призначений для керування практично всіма джерелами даних (у тому числі великих даних), а саме: завантажувати, очищувати, перетворювати, агрегувати, підтримувати сховища даних, робити міграцію та синхронізацію. Підтримує можливість роботи, як в реальному часі, так і batch-режимі (відкладений автоматичний режим роботи на сервері даних). Використовується в організаціях, що потребують обробки великих розподілених обсягів даних для задоволення потреб користувачів у достовірній, повній інформації та своєчасному доступі до неї.

SAS Forecast Server призначений для аналізу, моделювання та прогнозування великої кількості економетричних або часових рядів даних в автоматичному режимі без залучення аналітика-людини, є можливість виконання сценарного аналізу "what-if". Програмне забезпечення автоматично вибирає найбільш адекватну модель прогнозування, оптимізує параметри моделі і видає прогнози, також є можливість попередньої обробки транзакційних даних, перетворюючи їх у формат часових рядів. Ідея застосування інструмента заснована на тому, що для близько 80% всіх існуючих часових рядів даних, математичні моделі можуть бути побудовані в автоматичному режимі.

SAS Enterprise Miner – інтегрований компонент системи SAS, створений для виявлення закономірностей у дуже великих масивах даних інформації, необхідної для прийняття рішень. Це спеціалізований інструмент для спрощення процесу аналізу даних при створенні високоточних інтелектуальних та описових моделей, розроблений спеціально для пошуку та аналізу прихованих закономірностей в даних. Інструмент містить широкий набір методів інтелектуального аналізу даних: регресійний аналіз, дерева рішень, нейронні мережі, машина опорних векторів, кластерний аналіз та багато інших. Інструмент використовується для вирішення задач виявлення шахрайства, мінімізації ризиків, прогнозування потреби в ресурсах, скорочення часу простою активів, збільшення ефективності маркетингових кампаній, запобігання відтоку клієнтів, розробки скорінгових карт, збільшення продажів товарі та інше.

SAS Text Miner – це спеціалізований інструмент для виявлення і вилучення знань з текстових документів. Дозволяє аналізувати інформацію, приховану у великих обсягах неструктурованого тексту. Він перетворює текстові дані в працездатний, зрозумілий формат, який полегшує класифікацію документів, пошук явних зв'язків або зв'язків між документами і кластеризацію документів на категорії. При проведені аналізу підтримує наступні формати даних PDF, ASCII, HTML, Microsoft Word та інші. Станом на 2013 рік

реалізовано 31 мову, в тому числі російську та українську. Серверний варіант продукту має власний траулер – інструмент для завантаження інформації з Інтернету по заданих адресах. Це перше рішення інтелектуального аналізу даних, що щільно інтегрує текстову інформацію зі структурованими даними для поліпшення якості аналітичного процесу і прийняття рішень.

SAS Enterprise Content Categorization використовується для обробки природної мови (NLP) із використанням передових лінгвістичних методів для визначення ключових тем і фраз в електронному тексті з метою автоматичної класифікації великих обсягів багатомовного контенту. Інструмент дозволяє скоротити час ручного маркування текстів та індексації посилань, що загалом покращує спільний розвиток, збереження і використання знань.

SAS Sentiment Analysis призначений для аналізу думок і досвіду, що розміщені на веб-сайтах та комунікаційних центрах, тобто інформації, що прихована в повідомленнях електронної пошти, звітах, результатах опитування та внутрішніх файлах організації. Об'єднання статистичного навчання зі спеціалізованими лінгвістичними методами дозволяє програмному забезпеченню знаходити і аналізувати цифровий контент в режимі реального часу з метою визначення настрою клієнтів та користувачів Інтернету.

Варто зауважити, що описані вище інструменти для вирішення задач текстової аналітики використовуються для потреб бізнес-орієнтованих організацій з метою більш чіткого розуміння потреб клієнтів та сегментації ринку, а також для вирішення окремих задач безпеки.

SAS Information Delivery Portal – забезпечує єдину точку доступу до агрегованої інформації за допомогою веб-інтерфейсу. Використовується для розміщення контенту (результатів, звітів, графіків, OLAP-кубів), створеного в будь-яких SAS продуктах на Інтернет-порталі організації для загального або персоніфікованого доступу.

Окрім представлених вище програмних продуктів компанії SAS Institute, існує багато інших. Крім того, розроблені спеціалізовані індустріальні рішення для різних сфер людської діяльності, зокрема, грального бізнесу (казино), телекомунікацій, освіти, фінансових послуг, державного управління, медицини, страхування (в тому числі медичного), готельного бізнесу, промисловості, медіа та реклами, нафтогазової промисловості, роздрібної торгівлі та маркетингу, туристичної сфери та транспорту, комунального обслуговування, тощо [19].

9.4 Базові процедури статистичної обробки даних в системі SAS Base

У мові програмування **SAS BASE** реалізовані наступні базові процедури статистичної обробки даних.

CORR – попарне обчислення коефіцієнтів кореляції між даними.

MEANS та **SUMMARY** – базові статистики в наборах даних.

TABULATE – створення комплексних таблиць базових статистик.

UNIVARIATE – обчислення одновимірних статистик та виведення діаграм розподілу випадкових величин.

CHART – побудова гістограм, графіків, сегментних діаграм.

FREQ – обчислення частотних розподілів для категоріальних змінних, побудова багатонаправленої крос-табуляції, описові та дедуктивні статистики.

Таблиця 9.1 Базові статистики, що доступні в наведених вище процедурах та перелік додаткових операцій з ними

Статистика	MEANS	UNVARIATE	SUMMARY	TABULAT	CORR
Кількість “пропусків” в даних	X	X	X	X	
Кількість “не порожніх” значень	X	X	X	X	X
Кількість значень	X		X		
Сума ваг	X	X	X	X	X
Середнє	X	X	X	X	X
Сума	X	X	X	X	X
Мінімум	X	X	X	X	X
Максимум	X	X	X	X	X
Діапазон	X	X	X	X	X
Некорегована сума квадратів	X	X	X	X	X
Скорегована сума квадратів	X	X	X	X	X
Дисперсія	X	X	X	X	X
Стандартне відхилення	X	X	X	X	X
Стандартна похибка	X	X	X	X	
Коефіцієнт асиметрії	X	X	X		
Коефіцієнт ексцесу	X	X	X		
Т-тест Стьюдента	X	X	X	X	
Prob>|t|	X	X	X	X	
Медіана		X			X
Основні перцентилі		X			
Мода		X			
Кореляція Пірсона					X
Друк результатів	X	X	X	X	X
Збереження результатів в окремому файлі	X	X	X		X

В подальшому, при викладенні матеріалу будуть використовуватися наступні математичні позначення:

x_i	значення *i*-ї змінної (не порожнє)
w_i	вага, що відповідає значенню x_i, якщо є значення WEIGHT, в іншому випадку вага дорівнює одиниці
w	кількість не порожніх значень змінної
$\bar{x}$	математичне сподівання, що обчислюється за формулою $\frac{\sum w_i x_i}{\sum w_i}$
$d = n$	якщо $VARDEF = DF$
$d = n - 1$	якщо $VARDEF = DF$
$d = \sum w_i$	якщо $VARDEF = WEIGHT$
$d = (\sum w_i) - 1$	якщо $VARDEF = WDF$
s^2	обчислюється за формулою $\frac{\sum w_i \bullet (x_i - \bar{x})^2}{d}$
z_i	централізація, що обчислюється за формулою $\frac{(x_i - \bar{x})}{s}$

Наведені математичні позначення використовуються при обчислені статистик по відношенню до не порожніх значень.

При використані статистичних процедур можуть використовуватися наступні оператори:

N	– кількість не порожніх спостережень
NMISS	– кількість порожніх спостережень
MIN	– мінімальне значення в наборі даних
MAX	– максимальне значення в наборі даних
RANGE	– найбільший розкид (ранг), що обчислюється за формулою $MAX - MIN$
SUM	– зважена сума, обчислюється за формулою $\sum w_i - x_i$
SUMWGT	– сума ваг, обчислюється за формулою $\sum w_i$
MEAN	– середнє арифметичне, $\bar{x}$
USS	– не корегована сума квадратів, що обчислюється за формулою $\sum w_i \cdot x_i^{\,2}$
CSS	– корегована сума квадратів, що обчислюється за формулою $\sum w_i \cdot (x_i - \bar{x})^2$
VAR	– дисперсія (варіація)

COVARIANCE	– коваріація, що обчислюється за формулою $COV_{XY} = \frac{\sum w_i \cdot (x_i - \bar{x}) \cdot (y_t - \bar{y})}{d}$
STD	– стандартне середньоквадратичне відхилення
STDERR	– стандартна похибка, що обчислюється за формулою $\frac{s}{\sqrt{n}}$.
CV	– процентний коефіцієнт дисперсії, що обчислюється за формулою $\frac{s}{x} \cdot 100\%$
SKEWNESS	– коефіцієнт асиметрії, що обчислюється за формулою $\sum z_i^3 \cdot \left[\frac{n}{(n-1)(n-2)} \right]$
KURTOSIS	- коефіцієнт ексцесу, що обчислюється за формулою $\sum z_i^3 \cdot \left[\frac{n \cdot (n+1)}{(n-1)(n-2)(n-3)} \right] - \frac{3 \cdot (n-1)^2}{(n-2)(n-3)}$
Т–критерій	Т–критерій Стьюдента для H_0 (середнє популяції дорівнює нулю), що обчислюється за формулою $t = \frac{\bar{x} \cdot \sqrt{n}}{s}$
PRT	– двосторонній критерій Стьюдента з $n-1$ ступенем свободи. Ймовірність при H_0 отримання абсолютного значення $t-value$ більша ніж $t-value$ значення, що спостерігається, в даному випадку $\Pr ob > \lvert t \rvert$
MEDIAN	– середнє значення x_i, якщо n непарне та середнє значення двох середніх, якщо n парне
QUARTILE	– значення для якого чверть найбільших x_i знаходиться вище, і значення, для якого чверть найменших знаходиться нижче
MODE	– значення, що найчастіше зустрічається в наборі даних

9.5 Вимоги до даних для обчислення базових статистик

У випадку, коли з якихось причин дані містять пропуски, статистики обчислюються не залежно від кількості порожніх та не порожніх значень. Статистики SUM, MEAN, MAX, MIN, ДІАПАЗОН:, USS і CSS будуть обчислюватися за наявності хоча б одного не порожнього значення.

Вимоги для решти статистик наступні:

- VAR, STD, STDERR, CV, T і PRT потребують мінімум двох не порожніх значень.
- SKEWNESS вимагають мінімум трьох не порожніх спостережень.
- KURTOSIS вимагає чотирьох не порожніх спостережень.
- SKEWNESS, KURTOSIS, T і PRT вимагають, щоб STD було більше за нуль.
- CV вимагає, щоб MEAN було відмінно від нуля.

Популяція і параметри

У статистиці використовується термін ***популяція (population)***, який має свій власний, відмінний від загальновживаного, сенс. Цей термін не асоціюється в статистиці з якою-небудь групою людей. Термін популяція означає просто певний набір будь-яких значень – генеральну сукупність значень, тобто весь набір можливих випадків одного роду.

Популяція характеризується функціями, а саме:

- функцією розподілу випадкової величини;
- функцією щільності розподілу випадкової величини.

Зазвичай, при вживанні терміну ***розподіл (distribution)*** розуміється функція щільності або кумулятивна функція випадкової величини. Для визначення (опису) розподілу випадкової величини використовують ряд вимірів, які узагальнюють властивості розподілу, що нас цікавить.

Кожний такий вимір, обчислений за значеннями випадкової величини, зазвичай називають параметром. Одним з параметрів, що частіше за все використовують, є середнє значення даної випадкової величини.

Вибірки і статистики

Зазвичай в реальному житті не можливо виміряти всі значення популяції. Тому, використовують вибіркові набори значень (sample), для яких обчислюють характеристики. Математичну функцію від вибірки зазвичай називають ***статистикою.*** У випадку, коли вибірка з популяції формується випадковим чином і має необхідну репрезентативність і величину, то статистики, отримані з такої вибірки, називають ***оцінками (estimates)*** параметрів популяції. Якість (точність) цих оцінок залежить від виду випадкової величини, методів оцінки та ступеня задання точності.

Методи локалізації або місцеположення розподілу

Методами, що вказують на розташування розподілу прийнято вважати середнє, медіану та моду (вони вказують на центральне значення розподілу). Окрім цього, в якості додаткових методів використовують процентилі та квантилі – 25% та 75%.

- ***Середнє популяції*** оцінюється як вибіркове середнє $\bar{x} = \frac{x_i}{n}$.
- ***Медіана популяції*** – це центральне значення популяції, його оцінкою є медіана вибірки. Для непарного числа спостережень воно збігається з центральним значенням вибірки, а для парного – знаходиться як середнє двох центральних значень.
- ***Мода*** – це значення, при якому щільність популяції досягає максимуму. Вибірковою оцінкою його є значення, що найбільш часто зустрічається у вибірці.
- ***Процентилі*** **і** ***квантилі*** характеризують розподіл більш детально. Для набору даних, що впорядкований за зростанням, значення процентиля показує, який відсоток випадків знаходиться під ним, і відповідно, його доповнення до ста – який над ним. Найпопулярніші квантилі – 25% та 75%.

Міри варіабельності

Іншими важливими характеристиками при вивченні розподілів групи статистик є це міри варіабельності або розкиду:

- ***Розкид*** - оцінюється як різниця між найбільшим і найменшим значенням у вибірці.
- ***Міжквантильний розкид*** оцінюється як різниця між 25% та 75% квантилями.
- ***Дисперсія*** визначається за формулами $\sigma^2 = E(x-\mu)^2$ або $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$. Різницю між значенням і середнім називають відхиленням від середнього, тобто дисперсія є середнім квадрата відхилення.
- ***Стандартне відхилення для популяції*** – обчислюється як квадратний корінь з дисперсії.
- ***Коефіцієнт дисперсії*** – безрозмірна відсоткова величина, що показує пропорцію дисперсії. У випадку будь-якої лінійної зміни значень у вибірці, даний коефіцієнт залишиться незмінним.

Міри форми

Дуже часто важливу інформацію про розподіл надають його вид або форма, які характеризуються наступними коефіцієнтами.

- ***Коефіцієнт асиметрії для популяції*** обчислюється за формулою $S = \frac{E(x-\mu)^3}{\sigma^3}$, а для вибірки як $S = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \frac{(x_i - \bar{x})^3}{s^3}$.
- ***Коефіцієнт ексцесу для популяції*** обчислюється за формулою

$K = \frac{E(x-\mu)^4}{\sigma^4} - 3$, а для вибірки – $K = \frac{n(n-1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \frac{(x_i - \bar{x})^4}{s^4} - 3\frac{(n-1)^2}{(n-2)(n-3)}$.

Нормальний розподіл

Найбільш важливим сімейством теоретичних розподілів є так званий нормальний або Гаусовський розподіл. Він характеризується тим, що:

- графік функції має форму дзвону;
- повністю характеризується двома параметрами – середнім μ і стандартним середньоквадратичним відхиленням σ;
- приблизно 68% всіх значень лежать в межах одного стандартного відхилення від середнього, 95% всіх значень лежать в межах двох стандартних відхилень, а 99% значень в межах трьох стандартних відхилень.

Вибірковий розподіл середніх

При виборі з популяції вибірки розміру n, декілька разів та обчислені для кожного випадку середнього по вибірці, отримують ***вибірковий розподіл середніх (sampling distribution).***

Даний розподіл має такі властивості:

– вибіркове середнє середніх співпадає із середнім популяції μ;

– стандартне відхилення середніх (стандартна похибка)дорівнює $\frac{\sigma}{n}$ або $\frac{s}{\sqrt{n}}$ для оцінки.

Тестування гіпотез

Зазвичай гіпотези формулюються в сенсі ймовірності відповідності або невідповідності параметру певному значенню чи інтервалу. Наприклад, гіпотезу відповідності називати нульовою, а невідповідності – альтернативою.

Нехай: $H_0 : \mu = 0$, $H_1 : \mu < 0$, $H_2 : \mu > 0$. При цьому, можливість прийняття і неприйняття гіпотез характеризується ймовірнісними значеннями:

- ймовірність відхилення нульової гіпотези в разі її вірності називають ***значимістю або рівнем значущості (significance level)*** α**;**
- ймовірність відхилення нульової гіпотези в разі її невірності називають ***надійністю*** **(power).**

Для нормальних розподілів є оцінки через 2σ – 95% та 3σ – 99%. Таким чином, для $H_0 : \mu = 0$, $H_1 : \mu < 0$, $H_2 : \mu > 0$, якщо $0 - 2\sigma \le \bar{x} \le 0 + 2\sigma$, то з ймовірністю 95% приймаємо гіпотезу $H_0 : \mu = 0$, а значення меж інтервалу зазвичай називають ***критичними значеннями***.

Однак, в реальній ситуації, рідко буває достовірно відомо значення σ, тому можна використовувати тільки оцінку *s*. В зв'язку з цим, для перевірки гіпотез використовують спеціальну статистику $t = \frac{(\bar{x} - \mu_0)}{(s/\sqrt{n})}$, що є характеристикою відхилення середнього вибірки $\bar{x}$ від гіпотетичного μ_0.

Для випадкової величини проводять тест *t* на рівність її середнього нулю. Розподіл *t* називають також розподілом Стьюдента і для не прийняття або прийняття нульової гіпотези про рівність її середнього нулю з заданим наперед рівнем значущості, використовують таблиці розподілу.

Наприклад, для $\alpha = 5\%$ і ступенів свободи $n - 1 = 8$ відповідні критичні значення дорівнюють -2,3 і 2,3.

Іншим засобом відображення результату статистичного тесту гіпотез є обчислення $p - value$. Це значення виступає як ймовірність підтвердження того, що абсолютне для двостороннього тесту або звичайне значення t більше, ніж спостережуване у вибірці.

Таким чином, для визначення вірності нульової гіпотези із заданим рівнем значущості порівнюють цей рівень з $p - value$, і якщо $[(p - value) > |r|] < \alpha$, то нульова гіпотеза не приймається.

Двовимірні міри

Двовимірні статистики відповідають за вимір ступеня залежності між двома величинами, що змінюються. Двовимірні статистики:

- кореляції;
- коефіцієнта кореляції;
- ступеня узгодженості.

9.6. Створення набору даних в системі SAS

В якості прикладу розглянемо набір даних, що описує залежність вартості квадратного метру житла від його характеристик. Змінними аналізу використано наступні:

– Location – місце розташування квартири. Спеціальний коефіцієнт, що в кількісній формі показує ступінь популярності та престижності району міста (чим більше значення, тим краще).

– Quality_NUM – стан житла (чисельна змінна). Спеціально розроблений коефіцієнт (чим більше його значення, тим вища якість житла).

– Quality_NOMINAL – стан житла (категоріальна змінна). Приймає наступні значення:

Low – не задовільний стан.
Medium – задовільний стан.
Good – якісний стан.

Normal – відмінний стан.
Western – євроремонт (western style).

- Storey – поверховість (поверх). Спеціально розроблений коефіцієнт, на основі інформації про поверх на якому розміщене житло (чим більше значення коефіцієнта, тим краще). Коефіцієнт обчислюється з урахуванням кількості поверхів в будівлі та престижністю відповідного поверху житла.
- Material_NUM – material of house's walls – матеріали стін будівлі (чисельна змінна). Два стани: 1–залізо-бетонні панелі та 2 – цегла.
- Material_NOMINAL – матеріали стін будівлі (категоріальна змінна). Два стани: concrete – залізобетоні панелі та brick– цегла.
- TARGET_PRICE – вартість 1 $м^2$ квартири. Ця змінна буде використовуватися в якості цільової, тобто для прогнозування значень на основі інформації про попередні показники.

В табл. 9.2 наведено набір даних з 20 вимірів для задачі.

Таблиця 9.2. Значення змінних аналізу, що описує залежність вартості квадратного метру житла від його характеристик

Num	Location	Quality NUM	Quality NOMINAL	Storey	Material NUM	Material NOMINAL	TARGET PRICE
1	1,05	1,76	medium	1,01	1	concrete	885
2	1,9	1,14	low	1,98	1	concrete	940
3	1,9	3,03	normal	1,01	1	concrete	982
4	1,9	1,14	low	1,01	1	concrete	915
5	1,05	4,11	western_	1,98	1	concrete	1019
6	1,9	1,76	medium	1,01	2	brick	1023
7	1,9	1,76	medium	1,01	2	brick	1031
8	1,9	1,76	medium	1,01	2	brick	1039
9	3,05	1,76	medium	1,01	1	concrete	1064
10	1,9	1,76	medium	4	1	concrete	1078
11	3,05	1,76	medium	1,01	2	brick	1106
12	1,9	3,03	normal	3,02	2	brick	1163
13	1,9	4,96	western	1,98	1	concrete	1189
14	3,05	3,03	normal	1,98	1	concrete	1212
15	1,9	4,96	western	3,02	1	concrete	1226
16	1,9	3,03	normal	3,02	2	brick	1240
17	1,9	4,11	good	4	1	concrete	1240
18	1,9	4,11	good	3,02	2	brick	1250
19	1,9	3,03	normal	4	2	brick	1275
20	1,9	4,11	good	3,02	2	brick	1316

Будь-яка SAS-програма складається з двох етапів: етапу даних (DATA step) та етапу процедур (PROC step). Ці два етапи поодинці або разом створюють SAS-програму. Для створення набору даних, наведеного в табл. 9.2 використовується процедура DATA:

```
DATA PRICE;
INPUTNum Location Quality_NUMQuality_NOMINAL $
StoreyMaterial_NUMMaterial_NOMINAL $ TARGET_PRICE;
DATALINES;
1 1.05 1.76 medium 1.01 1 concrete 885
2 1.9 1.14 low 1.98 1 concrete 940
3 1.9 3.03 normal 1.01 1 concrete 982
4 1.9 1.14 low 1.01 1 concrete 915
5 1.05 4.11 western 1.98 1 concrete 1019
6 1.9 1.76 medium 1.01 2 brick 1023
7 1.9 1.76 medium 1.01 2 brick 1031
8 1.9 1.76 medium 1.01 2 brick 1039
9 3.05 1.76 medium 1.01 1 concrete 1064
10 1.9 1.76 medium 4 1 concrete 1078
11 3.05 1.76 medium 1.01 2 brick 1106
12 1.9 3.03 normal 3.02 2 brick 1163
13 1.9 4.96 western 1.98 1 concrete 1189
14 3.05 3.03 normal 1.98 1 concrete 1212
15 1.9 4.96 western 3.02 1 concrete 1226
16 1.9 3.03 normal 3.02 2 brick 1240
17 1.9 4.11 good 4 1 concrete 1240
18 1.9 4.11 good 3.02 2 brick 1250
19 1.9 3.03 normal 4 2 brick 1275
20 1.9 4.11 good 3.02 2 brick 1316
RUN;
```

9.7 Дослідження розподілів даних за допомогою PROC UNIVARIATE

При проведені статистичного аналізу даних, перше ніж перейти до статистичних тестів бажано спочатку витратити час на попереднє дослідження даних процедурою UNIVARIATE, щоб отримати перше уявлення про дані, що формують досліджувану вибірку. Ця процедура є частиною SASBASE, та являє собою набір інструментів для отримання інформації про дані, з якими працює користувач:

- визначення типу розподілу;
- розрахунок статистичних характеристик;
- побудова різних графіків і діаграм;
- розрахунок квантилів.

Статистики включають середнє, медіану, стандартне відхилення, коефіцієнт асиметрії, ексцесу і т.д.

Використання PROC UNIVARIATE досить просте. Після оператора PROC, можна вказати одну або кілька змінних числового типу, які необхідно досліджувати, в операторі **VAR**:

PROC UNIVARIATE;
VAR variable-list;

За відсутності оператора VAR, процедура обчислить статистики для всіх числових змінних, присутніх у вхідному наборі даних. Вхідний набір задається опцією DATA в операторі PROC.

PROC UNIVARIATE DATA = dataset;

Відзначимо, що в операторі PROC також можна використовувати й інші опції, наприклад, такі як, PLOT і NORMAL.

PROC UNIVARIATE PLOT NORMAL;

Опція **NORMAL** виконує тест на нормальність розподілу, а опція **PLOT** будує графіки даних – "текстовий графік", "ящик-з-вусами" та "нормально-ймовірнісного розподілу".

Можливості даної процедури достатньо різноманітні. Наприклад, використовуючи оператор **BY,** можна групувати дані для аналізу. Також можна задати порядок виводу результатів: (1) Ascending– по зростанню значень підгруп або (2) Descending– за зменшенням значень підгруп.

Синтаксис PROC UNIVARIATE

Повний синтаксис процедури **UNIVARIATE** має вигляд.

PROCUNIVARIATE<*option(s)*>;
BY<DESCENDING>*variable-1* <...<DESCENDING>*variable-n*><NOTSORTED>;
CLASS *variable-1*<(*variable-option(s)*)><*variable-2*<(*variableoption(s)*)>><**KEYLEVEL=** '*value1*'|('*value1*' '*value2*')>;
FREQ *variable*;
HISTOGRAM <*variable(s)*>< *option(s)*>;
ID *variable(s)*;
INSET <*keyword(s)*>< *option(s)*>;
OUTPUT <OUT=*SAS-data-set*>*statistic-keyword-1=name(s)* <... *statistickeyword-n=name(s)*><*percentiles-specification*>;
PROBPLOT <*variable(s)*>< *option(s)*>;
QQPLOT <*variable(s)*>< *option(s)*>;
VAR *variable(s)*;

WEIGHT *variable*;

Операнд **VAR** містить змінну або набір змінних, по відношенню до якої буде виконаний аналіз.

Операнд **FREQ** визначає змінну, що містить частоту появи спостереження. Можна задати тільки одну змінну для FREQ. Розглянемо ситуацію, коли є певний ряд даних (табл. 9.3).

Таблиця 9.3 Приклад ряду даних

1	3	2	1	1	4	2	3	3	1

Його можна представити і у вигляді варіаційного ряду (табл. 9.4).

Таблиця 9.4 Приклад варіаційного ряду даних, побудованого на основі ряду з табл. 9.3.

Значення	1	2	3	4
Частота	4	2	3	1

В цьому випадку, для аналізу змінної "Значення" задається операнд VAR, а змінної "Частота" – операнд FREQ.

Операнд **WEIGHT** містить вагові коефіцієнти для змінної аналізу, що відповідає операнду VAR. В операнді WEIGHT можна задати тільки одну змінну.

Операнд **CLASS,** аналогічно з BY дозволяє виконувати аналіз всередині підгруп, але, в даному випадку, можна задати дві змінні, значення яких будуть використовуватися, як змінні класифікації.

Типи розподілів, що підтримує PROC UNIVARIATE

Процедура UNIVARIATE дозволяє в автоматичному режимі виконувати порівняння емпіричного розподілу (ряду даних, що є в наявності у користувача) з теоретичними розподілами, що наведені нижче.

1. *Нормальний розподіл, з параметрами*:

- Mu – математичне сподівання.
- Sigma – середньоквадратичне відхилення.

2. *Логнормальний розподіл, з параметрами*:

- Zeta – коефіцієнт масштабу.
- Theta – значення порогу.
- Sigma – коефіцієнт згладжування.

3. Експоненційний розподіл, з параметрами:

- Sigma – коефіцієнт масштабу.

- Theta – значення порогу.

4. Розподіл Вейбулла, з параметрами:

- Sigma – коефіцієнт масштабу.
- Theta – значення порогу.
- C – коефіцієнт згладжування.

5. Бета розподіл, з параметрами:

- Sigma – коефіцієнт масштабу.
- Theta – порогове значення.
- Alpha та beta – коефіцієнти згладжування.

6. Гамма розподіл, з параметрами:

- Sigma – коефіцієнт масштабу.
- Theta – порогове значення.
- Alpha – коефіцієнти згладжування.

7. Розподіл Ядра;. можна задати наступні типи функцій ядра:

- Normal (використовується за замовчуванням).
- Quadratic.
- Triangular.

Значення Value задає смугу пропускання, яку можна вказати відразу задавши декілька значень (але не більше п'яти) розділених пропуском (наприклад, 0.1 0.2 0.5).

Графіки PROC UNIVARIATE

При побудові графіків, за допомогою процедури **UNIVARIATE**, користувач зазвичай може обрати тип візуалізації емпіричного і теоретичного розподілів. Всього передбачено побудову п'яти типів графіків:

1. гістограми;

2. графік ймовірностей (probability plot);

3. графік квантилів (quantiles plot);

4. "ящик з вусами" (box-and-whiskers plot);

5. текстовий графік (text based plots).

Для кожного графіка, окрім текстового. передбачена можливість налаштування кольору і товщини координатних осей, а також фону заливки графіка.

Гістограми в PROC UNIVARIATE

Створення гістограм з можливістю накладання кривих функцій щільності теоретичного розподілу, а також функцій щільності функціоналу ядра (kerneldensity function). Тип розподілу задається користувачем. Якщо задані

змінні класифікації (операнди BY та CLASS), то для кожної групи створюється своя гістограма.

На рис. 9.3 наведений приклад побудованої гістограми. Прямокутними стовпчиками представлено емпіричний розподіл, а суцільною лінією теоретичний розподіл нормального.

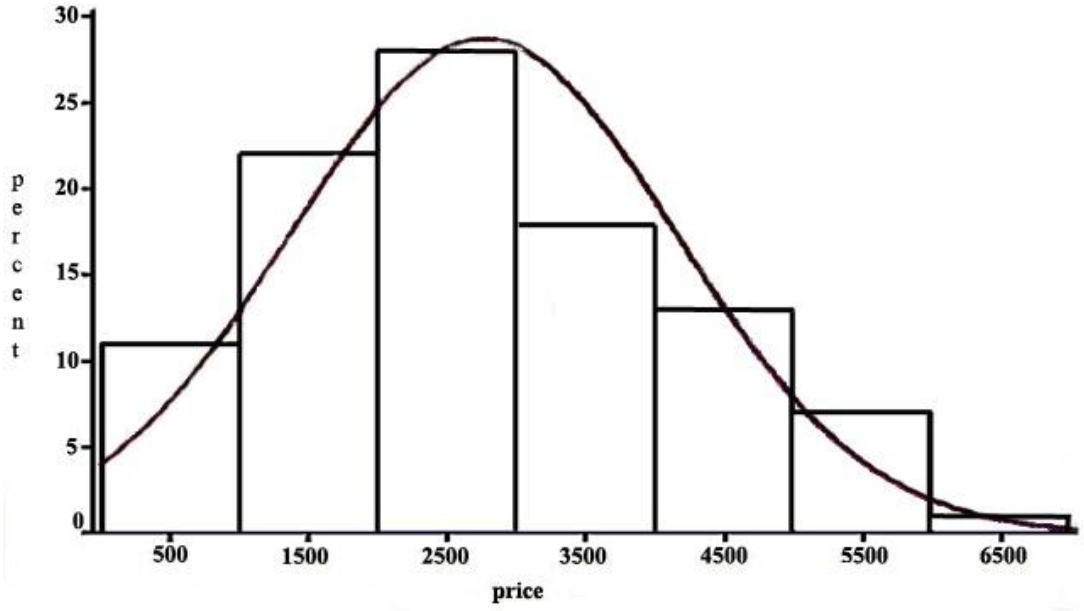


Рис. 9.3 Приклад теоретичного та емпіричного розподілів

Графіки ймовірностей в PROC UNIVARIATE

Використовується для візуального порівняння емпіричного і теоретичних функцій розподілу випадкової величини.

У прикладній статистиці використовується два типи графіків ймовірностей (probability plot):

1. PP графік, скорочення від "Probability-Probability" і "Percent-Percent" графік.

2. QQ графік, скорочення від "Quantile-Quantile" графік.

Обидва типи графіків PP і QQ реалізовані в системі SAS.

Для PP графіка по осі абсцис (горизонтальній) відкладаються процентилі розподілу, а по осі ординат (вертикальній) впорядковані відповідні значення розподілу.

На рис. 9.4 наведений приклад побудованого PP графіка. Суцільна (червона лінія) – це процентилі нормального теоретичного розподілу, а сині точки – процентилі теоретичного. Розподіли вважаються близькими, якщо точки знаходяться щільно навколо суцільної прямої лінії (теоретичних процентилів).

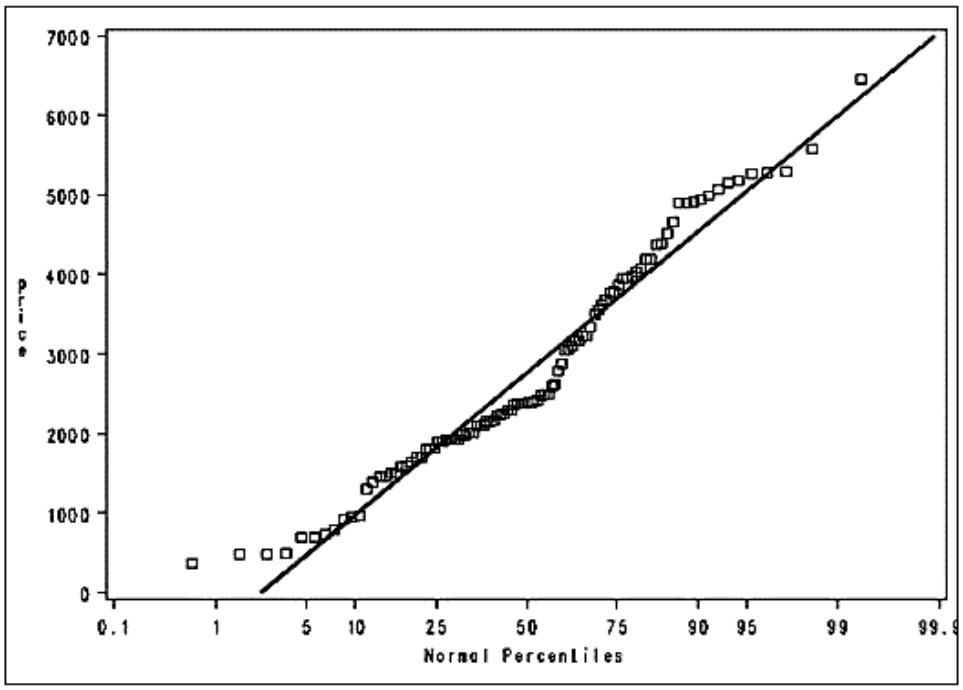


Рис. 9.4 Приклад графіка ймовірностей теоретичного та емпіричного розподілів

Графіки квантилів в PROC UNIVARIATE

Для графіка квантилів (QQ графіка) по осі абсцис (горизонтальній) відкладаються квантилі розподілу, а по осі ординат (вертикальної) впорядковані відповідні значення розподілу.

На рис. 9.5 наведений приклад побудованого графіка квантилів. Суцільна пряма лінія – це квантилі нормального теоретичного розподілу, а сині точки – квантилі теоретичного. Розподіли вважаються близькими, якщо точки лягають щільно навколо суцільної прямої лінії (теоретичних квантилів).

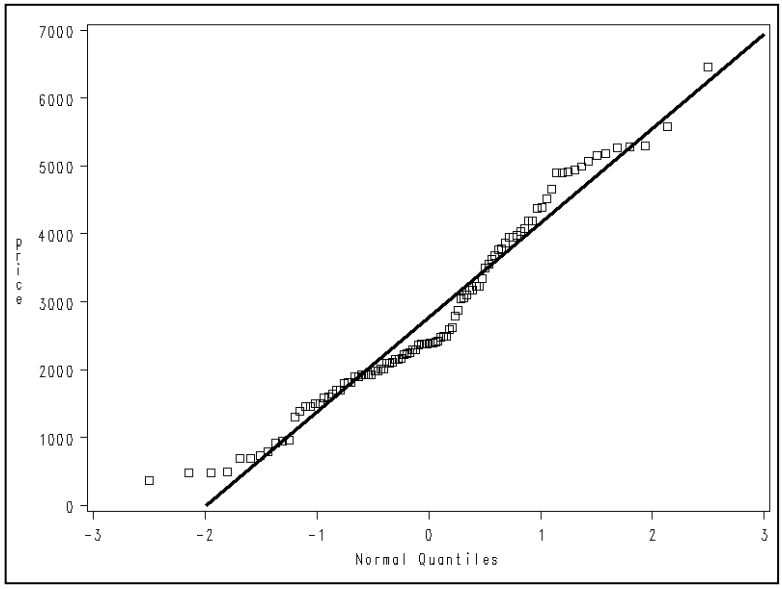


Рис. 9.5 Приклад графіка квантилів теоретичного та емпіричного розподілів

Графік у вигляді "коробки з вусами" в PROC UNIVARIATE

"Коробка з вусами" (box-and-whiskers plot або просто boxplot) – це графік, що компактно відображає розподіл ймовірностей. Кілька таких "ящиків" можна намалювати пліч-о-пліч, щоб візуально порівнювати один розподіл з іншим. Ні в якому разі не слід плутати з "японськими свічками", що використовуються на Forex.

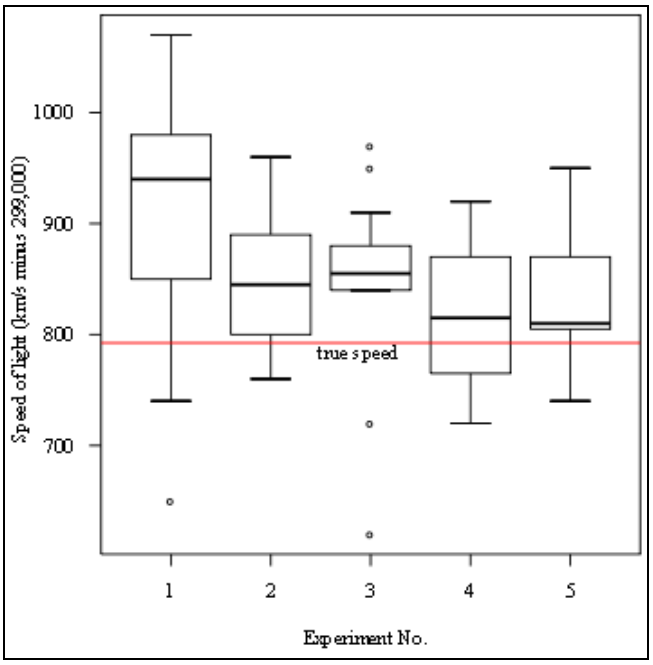


Рис. 9.6 Приклад "коробок з вусами" розміщених пліч-о-пліч

Побудова "коробки з вусами"

Для побудови графіка обчислюють медіану x_{50}, квартилі (x_{25}, x_{75}) і статистично значимий діапазон. Наприклад,

$$x_1 = \max(x_{\min}, x_{25} - 1{,}5 \cdot (x_{75} - x_{25}))$$
$$x_2 = \max(x_{\min}, x_{25} + 1{,}5 \cdot (x_{75} - x_{25}))$$

“Коробка” будується від квартиля до квартиля, на ній відкладається відмітка – медіана.

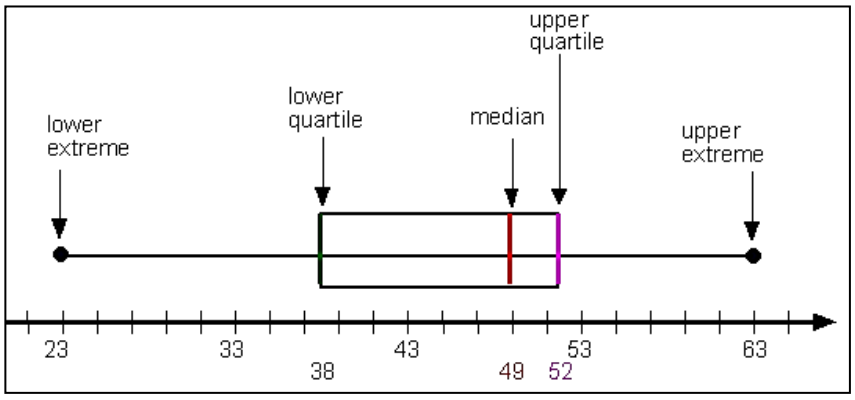


Рис. 9.7 Детальний приклад "коробки з вусами"

“Вуса” йдуть від квартилів до статистично значимих точок x_1 і x_2. Точки, що не входять в статистично значимий діапазон – це викиди, які зображують окремо. Іноді на коробці з вусами ще роблять додаткову мітку, що позначає довірчий інтервал математичного сподівання.

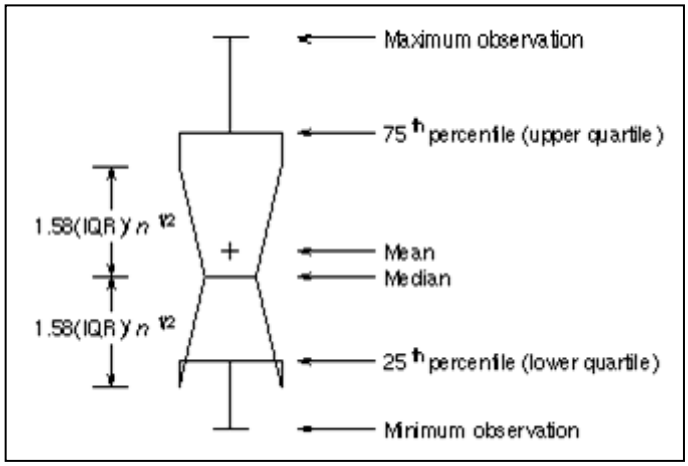


Рис. 9.8 Детальний опис "коробки з вусами"

Текстовий графік в PROC UNIVARIATE

В англомовній літературі по прикладній статистиці даний тип графіків зустрічається під назвами templot або stem-and-leafplot. Цей графік використовується для представлення кількісних даних у графічному форматі. Вперше був запропонований сером Артуром Лайон Боулі в 1900-х. Цей тип графіків широко застосовувався в 70-ті за використання псевдографіки на перших комп'ютерах, як дуже швидкий засіб для графічного представлення даних. Сьогодні, завдяки розвитку техніки і появи нових графічних інструментів, використовується рідко.

Принцип побудови текстового графіка.

Спостереження сортуються за зростанням. Кожне спостереження можна представити у вигляді пари чисел –stem (стовбур) і leaf (лист). Наприклад, число 6468 можна представити у вигляді:

Стовбур = 6
Лист = 5

В даному випадку468 заокруглили до 500.

На рис. 9.9 наведений приклад побудованих значень графіку, де

- Stem – стовбур спостереження.
- Leaf – лист спостереження.
- # – частота спостереження (сума дорівнює 100).
- BoxPlot – коробка з вусами (що була розглянута вище).

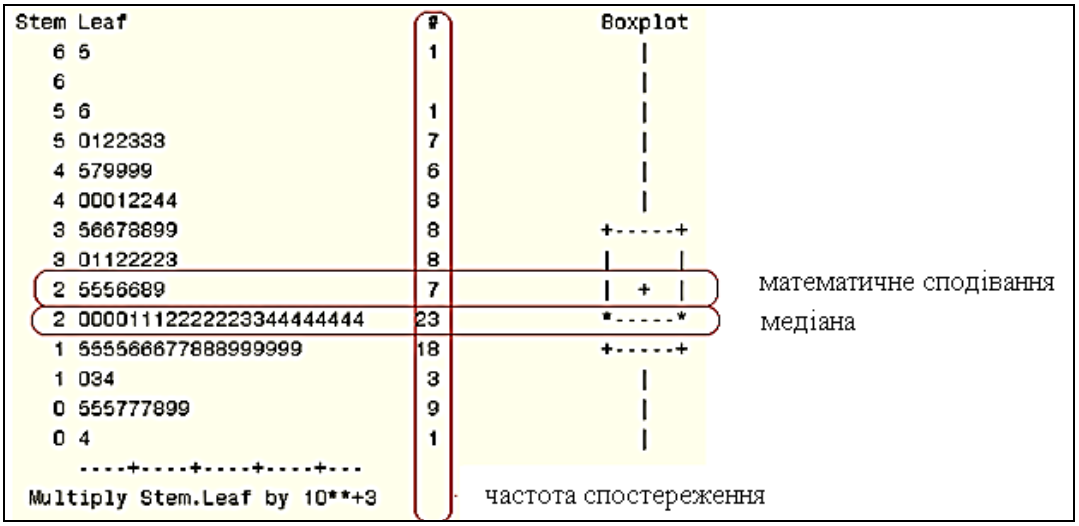


Рис. 9.9 Детальний опис текстового графіка

У стовпці листа може бути не одне, а кілька чисел. Наприклад, розглянемо випадок коли стовбур = 1, а лист = 034. Це означає, що записано три числа.

Графік отриманий на основі отриманих значень має вигляд наведений на рис. 9.10.

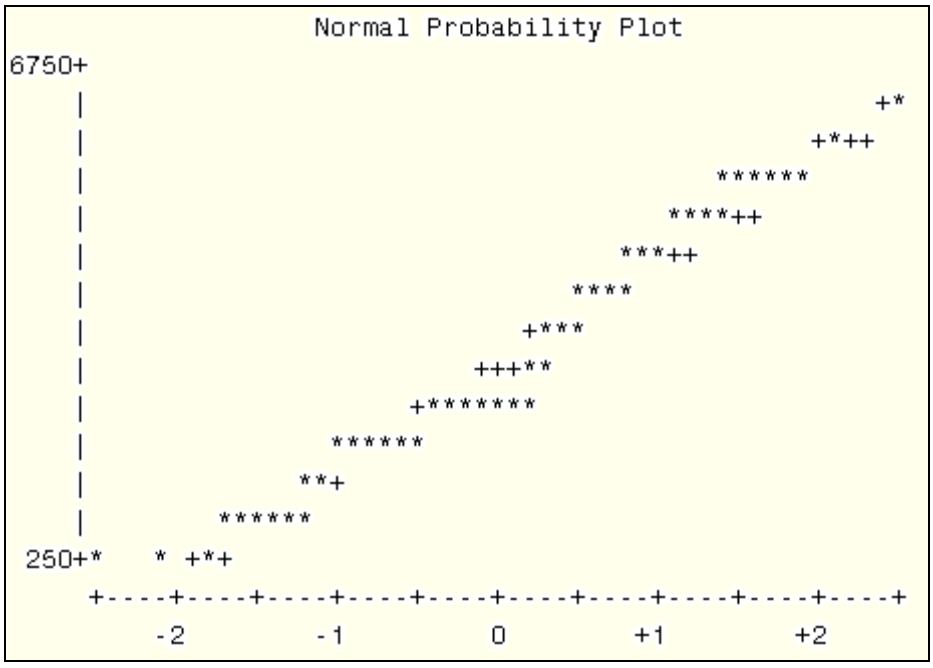


Рис. 9.10 Приклад текстового графіка

По осі абсцис відкладені квантилі емпіричного розподілу.

Критерії оцінювання емпіричної вибірки

- Критерій Колмогорова-Смірнова

Критерій Колмогорова-Смірнова в якості відстані між емпіричним та

теоретичним законом використовується значення обчислене за формулою:

$$D = \sup_x \left| F_n(x) - F(x) \right|$$

В наведеній формулі sup – це точна верхня грань множини $\left| F_n(x) - F(x) \right|$. Тобто, шукається найбільша вертикальна різниця між емпіричним $F_n(x)$ та теоретичним $F(x)$ розподілами. Зазвичай використовується статистика виду:

$$S_k = \frac{6nD_n + 1}{6\sqrt{n}},$$

де

$$D_n = \max(D_n^+, D_n^-),$$

$$D_n^+ = \max_{1 \le i \le n}\left(\frac{i}{n} - F(x_{i,}\Theta)\right),$$

$$D_n^+ = \max_{1 \le i \le n}\left(F(x_{i,}\Theta) - \frac{i-1}{n}\right).$$

n – об'єм вибірки, x_1, x_2, ..., x_n – впорядковані за зростанням вибіркові значення, $F(x_{i,}\Theta)$ – функція закону розподілу, узгодженість з яким перевіряється. Розподіл величини S_k при простій гіпотезі відповідає закону Колмогорова $K(S)$:

$$\forall S > 0: \lim_{n \to \infty} P(\sqrt{n} \cdot D_n \le S) = K(S) = \sum_{j=-\infty}^{+\infty} (-1)^j \cdot \exp(-2 \cdot j^2 \cdot S^2)$$

Якщо статистика $\sqrt{n} \cdot D_n$ перебільшує квантиль розподілу Колмогорова $K(S)$ заданого рівня значимості S, то тоді нуль-гіпотеза H_0 (про відповідність закону $F(x)$ відкидається. В іншому випадку, гіпотеза приймається на рівні S.

- Чисельне значення статистики Колмогорова-Смірнова.

Чим менше чисельне значення приймає статистика Колмогорова-Смірнова, тим краще.

Наприклад, є вибірка $F_n(x)$, яку перевірюють на відповідність логнормальному та експоненціальним законам розподілу, результати наведені на рис. 9.11.

Goodness-of-Fit Tests for Lognormal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.04268096	Pr > D	>0.500

Goodness-of-Fit Tests for Exponential Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.3464670	Pr > D	<0.001

Рис. 9.11 Приклад статистики Колмогорова-Смірнова

Отримані результати показують, що для логнормального розподілу статистика Колмогорова-Смірнова приймає менші значення, ніж за експоненційного. Тобто вибірка, що аналізується, більше відповідає логнормальному розподілу, ніж експоненціальному.

Значення *p-Value* – це значення ймовірності прийняття гіпотези.

На рис. 9.11 значення *p-Value*> 0,500, тобто з імовірністю більше 50% можна прийняти гіпотезу про те, що вибірка відповідає логнормальному розподілу.

Значення *p-Value* <0,001 означає, що з імовірністю менше 0,1% можна прийняти гіпотезу про те, що вибірка відповідає експоненціальному розподілу.

- Критерій Андерсона-Дарлінга

Даний критерій належить до класу квадратичних метрик $(F_n(x)-F(x))^2$. В загальному вигляді квадратичний статистичний критерій відповідності емпіричної вибірки заданому теоретичному закону розподілу записується як:

$$Q = n \cdot \int_{-\infty}^{+\infty} (F_n(x)-F(x))^2 \cdot \psi(x) \cdot dF(x)$$

Функція $\psi(x)$ – ваговий квадратичний коефіцієнт метрики $(F_n(x)-F(x))^2$.

$$\psi(x) = \frac{1}{F(x)\cdot(1-F(x))}$$

Критерій Андерсона-Дарлінга має вигляд:

$$A^2 = n \cdot \int_{-\infty}^{+\infty} (F_n(x)-F(x))^2 \cdot (1-F(x))^{-1} \cdot dF(x)$$

В кінцевому варіанті для обчислення статистики Андерсона-Дарлінга використовується формула:

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}\left[(2\cdot i-1)\cdot \log(F(x_i,\theta)) + (2\cdot n+1-2\cdot i)\cdot \log(1-F(x_i,\theta))\right]$$

- Критерій Крамера-Мізеса

Як і критерій Андерсона-Дарлінга, критерій Крамера-Мізеса належить до класу квадратичних метрик $(F_n(x)-F(x))^2$. Якщо в загальній формулі квадратичного статистичного критерію:

$$Q = n \cdot \int_{-\infty}^{+\infty} ((F_n(x)-F(x))^2 \cdot \psi(x) \cdot dF(x)$$

вагова квадратична функція $\psi(x)=1$, одержимо критерій Крамера-Мізеса:

$$Q = n \cdot \int_{-\infty}^{+\infty} ((F_n(x)-F(x))^2 \cdot dF(x)$$

Для обчислення статистики Крамера-Мізеса використовується формула:

$$W^2 = \sum_{i=1}^{n} (F(x_i, \theta) - \frac{2 \cdot i - 1}{2 \cdot n})^2 + \frac{1}{12 \cdot n}$$

- Критерій Шапіро-Уілка

Критерій Шапіро-Уілка (Shapiro-Wilk), був запропонований в 1965 році, базується на аналізі лінійної комбінації різниць порядкових статистик. Область допустимих значень критерію $0 < W \leq 1$. Значення критерію, що близькі до нуля означають, що емпіричний розподіл не відповідає нормальному (нуль- гіпотеза відхиляється).

Для обчислення значення критерію Шапіро-Уілкі використовується формула:

$$W = \frac{1}{S^2}\left[\sum_{i=1}^{k} a_{n-i+1} \cdot (x_{n-i+1} - x_i)\right]^2$$

де $S^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2$, за математичним змістом S^2 – коваріація, а $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ – математичне сподівання.

Коефіцієнт a_{n-i+1} беруть зі спеціальних таблиць.

При великих розмірах вибірки емпіричних значень, таблиці коефіцієнтів стають дуже громіздкими. Тому, в системі SAS даний критерій обчислюється для невеликих вибірок, що містять не більше ніж 2000 значень. Також існує модифікація критерію Шапіро-Уілка, яка дозволяє уникнути проблеми більшої розмірності вибірки – критерій Шапіро-Франчіа (в системі SAS не реалізований).

Важливе зауваження. Тому, що функціонал критерію Шапіро-Уілка має високий ступінь асиметричності, навіть високі значення $W > 0{,}9$ можна трактувати, як відхилення нуль-гіпотези.

Тому, при прийнятті рішення про відповідність емпіричного розподілу нормальному, потрібно в першу чергу орієнтуватися не на чисельне значення критерію *W*, а на *p-Value*, яке відповідає цьому значенню.

Приклад використання PROC UNIVARIATE

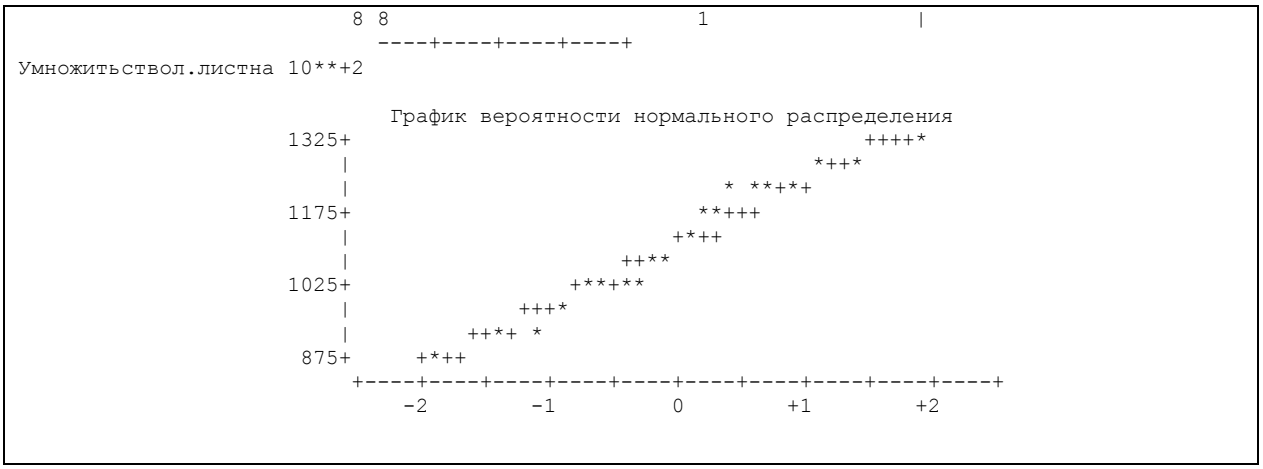
Наступний приклад описує роботу процедур аналізу вартості квадратного метра житла. Дані для аналізу взяті з файлу PRICE, що був створений раніше, аналізується тільки стовпець цін – показник TARGET_PRICE.

Код програми виглядає наступним чином:

```
procunivariatedata=PRICE plotnormal;
var TARGET_PRICE;
run;
```

Результати роботи процедури UNIVARIATE для наведеного коду sas-програми мають вигляд:

```
                                Процедура UNIVARIATE
                              Переменная:  TARGET_PRICE

                                      Моменты

              Кол-во                     20    Сумма весов               20
              Среднее                1109.65    Сумма набл-й           22193
              Ст. отклонение       129.81577    Дисперсия          16852.1342
              Асимметрия          -0.1132524    Эксцесс            -1.2154097
Неиспр.сумма кв.      24946653    Испр. сумма кв.     320190.55
Коэфф. вариации     11.6988033    Станд.ош.сред.     29.0276887

Basic Statistical Measures

                       Location                     Variability

Среднее   1109.650     Станд. отклонение       129.81577
Медиана   1092.000     Дисперсия                  16852
Мода      1240.000     Размах                  431.00000
Межеркварт-ый размах    212.00000

Tests for Location: Mu0=0

                       Test            -Statistic-    -----p Value------

                       t Стьюдента     t  38.22729    Pr> |t|    <.0001
Знаков          M        10    Pr>= |M|   <.0001
                       Знак.рангов     S       105    Pr>= |S|   <.0001


Tests for Normality

                Test                    --Statistic---    -----p Value------

Шапиро-Уилка            W     0.947762    Pr< W      0.3344
Колмогорова-Смирнова    D     0.134776    Pr> D     >0.1500
Крамера фон Мизеса      W-Sq  0.071025    Pr> W-Sq>0.2500
                Андерсона-Дарлинга      A-Sq  0.415958    Pr> A-Sq>0.2500

Quantile       Estimate

                                    100% Макс        1316.0
                                    99%              1316.0
                                    95%              1295.5
                                    90%              1262.5
                                    75% Q3           1233.0
                                    50% Медиана      1092.0
                                    25% Q1           1021.0
                                    10%               927.5
                                    5%                900.0
                                    1%                885.0
                                    0% Мин            885.0

                                    Extreme Observations

                                ----Lowest----        ----Highest---

                                Value      Obs        Value      Obs

                                  885        1         1240       16
                                  915        4         1240       17
                                  940        2         1250       18
                                  982        3         1275       19
                                 1019        5         1316       20


СтволЛист                     #                Box
13 2                        1                |
                           12 58                       2               |
                           12 1344                     4             +-----+
                           11 69                       2             |     |
                           11 1                        1             |  +  |
                           10 68                       2             *-----*
                           10 2234                     4             +-----+
                            9 8                        1                |
                            9 24                       2                |
```



Вивід результатів роботи процедури UNIVARIATE починається з числової інформації про розподіл: кількість спостережень, середнє, стандартне відхилення, коефіцієнти ексцесу та асиметрії, середнє, медіана, мода, міжквартильний розмах.

Тест на приналежність емпіричного ряду даних включає перевірку по критеріям:

- Шапіро-Уілка
- Колмогорова-Смірнова
- Крамера-Мізеса
- Андерсона-Дарлінга

Згідно отриманих результатів за тестом Шапіро-Уілка з ймовірністю 33%, можна прийняти гіпотезу про те, що ряд даних TARGET_PRICE описується нормальним розподілом. За всіма іншими тестами значення ймовірності прийняття гіпотези ще менше. Можна зробити висновок, що даний ряд треба порівнювати з іншими типами розподілів.

В кінці чисельної частини знаходиться інформація про процентилі та екстремальні значення.

Графічна частина виводить такі графіки:

1. текстовий графік (text based plots)
2. "ящик з вусами" (box-and-whiskers plot)
3. графік ймовірностей нормального розподілу

9.8 Обчислення статистик за допомогою PROC MEANS

Більшість статистик, які можна обчислювати за допомогою процедури UNIVARIATE також можна отримати процедурою MEANS, що є частиною SASBASE. Однак, в даному випадку їх необхідно вказувати явно. Процедура UNIVARIATE зазвичай застосовується для обчислення всіх статистик змінної, що аналізується, разом з тим, при обчислені характеристик для декількох змінних виходить дещо громіздкою. Якщо є необхідність отримати тільки

вибіркові статистики, і не по одній, а по декільком змінним, то використовують процедуру MEANS. У даній процедурі є можливість безпосередньо вказати тільки необхідні статистики, що цікавлять аналітика.

У простому випадку, PROC MEANS вимагає тільки одного оператора:

PROC MEANS *statistic-keywords*;

Приклад sas-програми із використанням PROCMEANS має вигляд:

```
procMEANSdata=PRICE;
var TARGET_PRICE;
run;
```

Результат роботи процедури має вигляд:

```
Процедура MEANS

                          Analysis Variable : TARGET_PRICE

          N            Mean         StdDev         Minimum         Maximum
          ---------------------------------------------------------------------
          20         1109.65     129.8157703     885.0000000         1316.00
          ---------------------------------------------------------------------
```

Зауважимо, що за відсутності будь-яких опцій-статистик, процедура MEANS автоматично обчислить для всіх числових змінних наступні статистики:

- кількість не порожніх спостережень;
- математичне сподівання;
- стандартне відхилення;
- мінімальне значення;
- максимальне значення.

Для отримання решти можливих статистик необхідно вказати відповідні опції в процедурі.

За замовчуванням, ***рівень значущості для довірчого інтервалу*** задається як 0,05 або 95%. У разі необхідності ***задання іншого рівня значущості***, слід використовувати в PROC MEANS опцію ALPHA. Наприклад, якщо необхідний рівень 0,10, то слід вказати:

PROC MEANS ALPHA = .10 CLM;

За замовчуванням, процедура MEANS обчислить статистики для всіх числових змінних. Це може бути незручно. ***Для задання тільки необхідних змінних,*** використовують оператор VAR.

PROC MEANS options;
VAR variable-list;

Статистики, що можна обчислити процедурою MEANS, наведені в табл. 9.5.

Таблиця 9.5 Перелік операндів та статистик, що використовуються в процедурі MEANS

Операнд	Опис статистики
CLM	Двосторонній довірчий інтервал
CSS	Скоригована сума квадратів
CV	Коефіцієнт дисперсії (варіації)
KURTOSIS	Коефіцієнт ексцесу
LCLM	Нижня межа довіри
MAX	Максимум
MEAN	Середнє
MEDIAN (P50)	Медіана
MIN	Мінімум
N	Кількість непустих спостережень
NMISS	Кількість порожніх спостережень
PROBIT	Імовірність для оцінки Стьюдента
RANGE	Діапазон
SKEWNESS	Коефіцієнт асиметрії
STDDEV	Стандартне відхилення
STDERR	Стандартна помилка середнього
SUM	Сума значень
SUMWGT	Сума зважених значень
UCLM	Верхня межа довіри
USS	Скоригована сума квадратів
VAR	Дисперсія (варіація)
T	Т-критерій Стьюдента
Q1 (P25)	Нижній квантиль
Q3 (P75)	Верхній квантиль
P1	1% процентиль
P5	5% процентиль
P10	10% процентиль
P90	90% процентиль
P95	95% процентиль
P99	99% процентиль

У разі використання операнду **CLASS**, PROC MEANS слід обчислити окремо статистики по групам спостережень, відповідно зі значеннями групової змінної. Якщо аналітик бажає ***отримати результати для груп***, що складають набір даних, використовують операнд **CLASS**. Окрім цього також може бути використаний операнд **BY** ***для розбиття даних на групи***. Це ефективно при

роботі з великими наборами даних. Однак, для використання оператора **BY**, необхідно попередньо відсортувати набір за змінною, яку аналітик бажає використовувати для розбиття на групи. Якщо ж необхідно залишити дані в їх первісному (природному) порядку, то використовується операнд **CLASS.**

PROC MEANS може, за бажанням аналітика створювати один або більше вихідних наборів даних, що містять обчислені статистики. Якщо немає потреби друкувати обчислені статистики, наприклад, їх зберігають для подальшої роботи або для виконання обчислень, то для ***відміни друку*** використовується опція **NOPRINT**.

Інші SAS-процедури, також обчислюють базові статистики. Однак, PROC MEANS є найбільш зручною і простою процедурою тому, що безпосередньо обчислює базові статистики, як результат своєї роботи. Решта процедур мають інші прямі призначення. Наведемо нижче ***синтаксис процедури MEANS.***

Синтаксис PROC MEANS

PROCMEANS<*option(s)*><*statistic-keyword(s)*>;
BY <DESCENDING>*variable-1* <... <DESCENDING>*variable-n*><NOTSORTED>;
CLASS *variable(s)* </ *option(s)*>;
FREQ *variable*;
ID *variable(s)*;
OUTPUT <OUT=*SAS-data-set*><*output-statistic-specification(s)*>
<*id-group-specification(s)*><*maximum-id-specification(s)*>
<*minimum-id-specification(s)*></ *option(s)*>;
TYPES *request(s)*;
VAR *variable(s)* < / WEIGHT=*weight-variable*>;
WAYS *list*;
WEIGHT *variable*;

Приклад використання PROC MEANS

Наступний приклад полягає в аналізі вартості квадратного метра житла. Дані для аналізу взяті з файлу PRICE, що був створений раніше. Аналізується тільки стовпець цін – показник TARGET_PRICE. Зауважимо, що цього разу добавлена назва для звіту за допомогою оператора **TITLE**.

Код програми виглядає наступним чином:

```
Proc MEANS data=PRICE nmeanmedianclmalpha=.10;
var TARGET_PRICE;
title "Приклади використання PROC MEANS";
run;
```

Результати роботи процедури MEANS для наведеного коду sas-програми мають вигляд:

```
                          Пример использования PROCMEANS

                                  Процедура MEANS

Analysis Variable : TARGET_PRICE

                                                  Lower 90%       Upper 90%
 N            Mean          Median     CL for Mean     CL for Mean
-------------------------------------------------------------------------
20         1109.65         1092.00         1059.46         1159.84
-------------------------------------------------------------------------
```

Отримавши даний звіт, аналітик може зробити наступні висновки. Середня ціна квадратного метра житла складає 1109,65, однак медіана дорівнює 1092,00, тобто половина всіх квартир має вартість до 1092 умовних одиниць. Довірчий інтервал вказує, що з ймовірністю 90% математичне сподівання для всієї популяції квартир буде в межах від 1059,46 до 1159,84 умовних одиниць. Таким чином, за результатами даного аналізу, аналітик може зробити висновок, що для збільшення ймовірності придбання житла, вартість квадратного метра повинна бути в межах від 1059 до 1160 умовних одиниць. Зрозуміло, що місце розташування, якість та тип житла також мають значення.

9.9 Дослідження категоріальних даних з використанням PROC FREQ

PROC FREQ – ще одна процедура SAS BASE, призначена для отримання статистик по категоріальним даним. Найбільш відома статистика χ-квадрат, проте є й інші – всі вони призначені для аналізу нульової гіпотези про наявність взаємозв'язків між змінними в даних. Всі міри таких взаємозв'язків є показниками сили або ступеня залежності між змінними. Найпростіший приклад використання PROC FREQ:

PROC FREQ;
TABLES variable-combinations / options;

Для отримання необхідних статистик треба вказати відповідні опції в процедурі.

Синтаксис PROC FREQ

PROC FREQ <*option(s)*>;
BY <DESCENDING>*variable-1* <...<DESCENDING>*variable-n*><NOTSORTED>;
EXACT *statistic-keyword(s)* </ *option(s)*>;
OUTPUT *statistic-keyword(s)* <OUT=*SAS-data-set*>;
TABLES *request(s)* </ *option(s)*>;
TEST *statistic-keyword(s)*;

WEIGHT *variable*;

Статистики, що можна обчислити процедурою FREQ, наведені в табл. 9.6.

Таблиця 9.6 Перелік операндів та статистик, що використовуються в процедурі FREQ

Операнд	**Опис статистики**
AGREE	Тести і міри МакНемара, Боукера, Q-Кокрен, к-тест
CHISQ	χ-квадрат тест на однорідність і міри взаємозв'язку
CL	Довірчі інтервал для заходів взаємозв'язку
CMN	Статистика Кокрен-Мантеля-Хаєнзеля
EXACT	Тест Фішера для матриць розміром більше, ніж 2x2
MEASURES	Коефіцієнти кореляції Пірсона і Спірмана, γ-статистика, статистика Кендалла τ-b, τ-c Стюарта, D-Соммерса, λ, ймовірність успіху, ймовірність ризику, довірчі інтервали для статистичних мір
RELRISK	Міра ризиків для матриці 2x2
TREND	Тест Кокрен-Ермітеджа для тренда (тенденції)

Приклад використання PROC FREQ

Наступний приклад полягає в аналізі категоріальних змінних Quality_NOMINAL та Material_NOMINAL. Дані для аналізу взяті з файлу PRICE, що був створений раніше (у підрозділі 9.6). Нагадаємо, що змінні мають наступний сенс та значення:

- Material_NOMINAL – матеріали стін будівлі (категориальна змінна). Два стани concrete – залізобетоні панелі та brick – цегла.
- Quality_NOMINAL – стан житла (категоріальна змінна). Приймає наступні значення:

Low – не задовільний стан.
Medium – задовільний стан.
Good – якісний стан.
Normal – відмінний стан.
Western – євроремонт (western style).

Код програми виглядає наступним чином:

```
Proc freq data=PRICE;
tablesquality_NOMINAL*material_NOMINAL / chisq;
title;
run;
```

Результати роботи процедури FREQ для наведеного коду sas-програми мають вигляд:

```
The FREQ Procedure

           Table of Quality_NOMINAL by Material_NOMINAL

Quality_NOMINAL
Material_NOMINAL

               Frequency|
               Percent  |
               Row Pct  |
               Col Pct  |brick   |concrete|  Total
               ---------+--------+--------+
               good     |      2 |      1 |      3
                        |  10.00 |   5.00 |  15.00
                        |  66.67 |  33.33 |
                        |  22.22 |   9.09 |
               ---------+--------+--------+
               low      |      0 |      2 |      2
                        |   0.00 |  10.00 |  10.00
                        |   0.00 | 100.00 |
                        |   0.00 |  18.18 |
               ---------+--------+--------+
               medium   |      4 |      3 |      7
                        |  20.00 |  15.00 |  35.00
                        |  57.14 |  42.86 |
                        |  44.44 |  27.27 |
               ---------+--------+--------+
               normal   |      3 |      2 |      5
                        |  15.00 |  10.00 |  25.00
                        |  60.00 |  40.00 |
                        |  33.33 |  18.18 |
               ---------+--------+--------+
               western  |      0 |      3 |      3
                        |   0.00 |  15.00 |  15.00
                        |   0.00 | 100.00 |
                        |   0.00 |  27.27 |
               ---------+--------+--------+
               Total           9       11       20
                            45.00    55.00   100.00

   Statistics for Table of Quality_NOMINAL by Material_NOMINAL

      Statistic                     DF       Value      Prob
      ------------------------------------------------------
      Chi-Square                     4      5.5315    0.2370
      Likelihood Ratio Chi-Square    4      7.4156    0.1155
      Mantel-Haenszel Chi-Square     1      0.6939    0.4049
   Phi Coefficient                      0.5259
      Contingency Coefficient              0.4655
      Cramer's V                           0.5259

       WARNING: 100% of the cells have expected counts less
than 5. Chi-Square may not be a valid test.

SampleSize = 20
```

Значення тесту χ-квадрат дорівнює 5,5315, а *p-value* дорівнює 0,2370, що створює передумови для прийняття нульової гіпотези про відсутність взаємозв'язку. Це підтверджує можливу відсутність залежності між матеріалом стін будівля та якістю стану житла.

9.10 Дослідження кореляції PROC CORR

Процедура CORR є ще однією процедурою SAS BASE, що відповідає за обчислення парної кореляції. Коефіцієнт кореляції є мірою взаємозв'язку між двома змінними. Якщо дві змінні повністю незалежні одна від одної, то коефіцієнт кореляції для такої пари дорівнює нулю. У разі ж їх повної залежності (кореляції) – коефіцієнт буде 1,0 або -1,0. В реальному житті цей коефіцієнт розподіляється усередині вказаного відрізку. Обчислювати даний коефіцієнт досить зручно і просто за допомогою процедури CORR:

PROC CORR;

Ці два слова вказують SAS необхідність обчислити кореляції між усіма числовими змінними у всіх наборах даних, що були створені за останній час (нещодавно).

Також, є можливість використовувати оператори **VAR** і **WITH** для визначення змінних, що необхідно аналізувати.

VAR variable-list;

WITH variable-list;

В цьому випадку змінні, визначені в **VAR**, розташуються в заголовках стовпців кореляційної матриці, а змінні, визначені в **WITH** – в мітках рядків. При використанні тільки оператора **VAR**, матриця буде містити обчислені змінні і в заголовках стовпчиків і в заголовках рядків.

За замовчуванням, процедура CORR обчислює коефіцієнт кореляції Пірсона. Однак, є можливість, за допомогою опцій, задавати обчислення непараметричних коефіцієнтів. Наприклад, опція **SPEARMAN** задає обчислення кореляційної пропорції Спірмана замість коефіцієнта Пірсона:

PROC CORR SPEARMAN;

Можливі й інші статистики, наприклад **HOEFFDING** обчислює D-статистику Хоефдінга, а опція **KENDALL**– тау-б коефіцієнт Кендалла. Так само, в процедурі єможливість збереження результатів в окремому наборі даних.

Синтаксис PROC CORR

PROC CORR <*option(s)*>;

BY <DESCENDING>*variable-1*<...<DESCENDING>*variable-n*><NOTSORTED>;

FREQ *frequency-variable*;

PARTIAL *variable(s)*;
VAR *variable(s)*;
WEIGHT *weight-variable*;
WITH *variable(s)*;

Кореляція Пірсона

Для метричних величин використовується коефіцієнт кореляції Пірсона, точна формула якого була виведена Френсісом Гальтоном (Francis Galton).

Нехай *X*, *Y* – дві випадкові величини, що визначені на одному ймовірнісному просторі. В цьому випадку їх коефіцієнт задається формулою:

$$R_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \cdot \sigma_Y},$$

де *cov* – це коваріація між *X* та *Y*, а σ – середньоквадратичне відхилення відповідної величини.

$$\sigma_X = \sqrt{D[X]},$$

де *D* – дисперсія.

$$R_{X,Y} = \frac{M[X,Y] - MX \cdot MY}{\sqrt{\left(M[X^2] - (MX)^2\right)} \cdot \sqrt{\left(M[Y^2] - (MY)^2\right)}},$$

де *M* – математичне сподівання.

Для графічного представлення подібного зв'язку можна використовувати прямокутну систему координат з осями, що відповідають обом змінним. Кожна пара значень маркується за допомогою певного символу. Такий графік називається "діаграмою розсіювання" (аналог такого графіка можна подивитися на рис. 9.12).

Статистична перевірка наявності кореляції Пірсона

Гіпотеза H_0 відсутності лінійного зв'язку між вибірками *X* та *Y* $(R_{XY} = 0)$. Статистика критерію:

$T = \frac{R_{XY} \cdot \sqrt{n-2}}{\sqrt{1 - R_{XY}^2}} \approx t_{n-2}$ – розподіл Стьюдента з *n-2* ступенями свободи.

Критерій $T \in [t_\alpha, t_{1-\alpha}]$, де α – квантиль розподілу Стьюдента.

Недоліки кореляції Пірсона:

- нестійкість до викидів;
- за допомогою коефіцієнта кореляції можна визначити лінійну залежність

між величинами, інші взаємозв'язки виявляються методами регресійного аналізу;

- необхідно розуміти відмінність понять "незалежність" і “некоррельованість”. З першого витікає друге, але не навпаки.

Інтерпретація результатів обчислення кореляції Пірсона в SAS Base

Область допустимих значень: $\rho \in [-1;\ +1]$

$\rho = +1$ – повний позитивний зв'язок між змінними.

$\rho = 0$ – змінні незалежні одна від одної.

$\rho = -1$ – повний негативний зв'язок між змінними.

Класифікація видів кореляційного зв’язку наведена в таблиці 9.7.

Таблиця 9.7 Класифікація видів кореляційного звязку

Кореляція	Позитивна	Негативна
Слабка	[0,1 ; 0,3)	(-0,3 ; -0,1]
Середня	[0,3 ; 0,5)	(-0,5 ; -0,3]
Сильна	[0,5 ; 1]	[-1 ; -0,1]

Дані, що використані для прикладу обчислення кореляції Пірсона представлены в таблиці 9.8.

Таблиця 9.8 Набір даних, що використовується для прикладу обчислення кореляції Пірсона

N	X_1	X_2
1	0	1
2	1	0
3	0	1
4	1	0
5	0	1
6	0	1
7	1	0
8	0	1
9	1	0
10	1	1

Код програми має наступний вигляд:

```
DATA DATA_COR;
INPUT N     X1     X2;
DATALINES;
1 0 1
2 1 0
3 0 1
4 1 0
5 0 1
6 0 1
7 1 0
8 0 1
9 1 0
10 1 1
RUN;
proccorrdata=DATA_COR PEARSON;
        var X2;
        with X1;
run;
```

Прилад результатів обчислення кореляції Пірсона, для значень змінних з табл. 9.8:

```
      Pearson Correlation Coefficients, N = 10
Prob> |r| under H0: Rho=0

X2

              X1     -0.81650
                      0.0039
```

Отримані результати мають наступне пояснення:

- 0,8165 – значення кореляції Пірсона;
- 0,0039 – ймовірність відхилення нульової гіпотези.

В даному випадку з ймовірністю 0,39% між змінними немає зв'язку.

Приклад використання PROC CORR для обчислення кореляції Пірсона

Наступний приклад полягає в аналізі наявності кореляції Пірсона між цільовою змінною "Вартість квадратного метра житла" (TARGET_PRICE) та регресорами "Місцеположення житла" (Location), "Поверх житла" (Storey). Дані для аналізу взяті з файлу PRICE, що був створений раніше (у підрозділі 9.6).

Код програми виглядає наступним чином:

```
proccorrdata=PRICE;
var Location Storey;
```

```
with TARGET_PRICE;
title 'Кореляція ціни квадратного метра житла';
title2 'з місцеположенням';
title3 'і поверхом';
run;
```

Необхідно зауважити, що в даному прикладі в явному вигляді не вказується, що обчислюється саме кореляція Пірсона тому, що за саме вона обчислюється за замовчуванням. Для того, щоб вказати в явному вигляді, необхідно підправити перший рядок коду програми, як показано нижче.

```
Proccorr data=PRICE PEARSON;
```

Результати роботи процедури CORR для наведеного коду sas-програми мають вигляд:

Корреляция цены квадратного метра жилья
с местоположением
иэтажностью

The CORR Procedure

1 With Variables: TARGET_PRICE
2 Variables: Location Storey

Simple Statistics

Variable	N	Mean	StdDev	Sum	Minimum	Maximum
TARGET_PRICE	20	1110	129.81577	22193	885.00000	1316
Location	20	1.98750	0.52613	39.75000	1.05000	3.05000
Storey	20	2.15500	1.13703	43.10000	1.01000	4.00000

Pearson Correlation Coefficients, N = 20
Prob> |r| under H0: Rho=0

	Location	Storey
TARGET_PRICE	0.25354 0.2808	0.71786 0.0004

Звіт про результати роботи процедури CORR починається з описових статистик по кожній змінній. Після чого виводиться кореляційна матриця з коефіцієнтами кореляції Пірсона і ймовірністю отримання більшого абсолютного значення для кожного з них. В наведеному прикладі зміна Storey корелює з цільовою змінною, при чому зв'язок додатній. Тобто поверховість житла позитивно впливає на вартість квадратного метра. Змінна Location (розташування житла), як не дивно має слабкий зв'язок з цільовою змінною.

Непараметричний кореляційний аналіз

Кореляція Пірсона використовується для нормально розподілених величин. Для інших розподілів більш ефективні методи вивчення зв'язку між випадковими величинами, що засновані на застосуванні порядкових статистик,

або на заміні спостережуваних величин їх рангами. Такі методи, володіючи підвищеною стійкістю до відхилень розподілу від нормального, в більшості випадків дозволяють спростити обчислення, залишаючи на прийнятному рівні статистичні характеристики висновків, що отримують за гіпотезами.

У загальному випадку, непараметричний кореляційний аналіз можна розділити на дві групи:

1. На основі порядкових статистик.

   1.1.Оцінка кореляції за допомогою тренду.

2. Рангова кореляція.

Кореляційний аналіз на основі порядкових статистик ґрунтується на попередньо упорядкованих за зростанням вибірках (ранжируваних сукупностях).

- Оцінка кореляції за допомогою тренду

Оцінка кореляції за допомогою тренду заснована на наступній ідеї. Якщо, значення однієї змінної (наприклад, *X*) попередньо упорядкувати (наприклад, за зростанням), то поведінка другої змінної (наприклад, *Y*) може бути використана в якості індикатора шуканої кореляції. Тобто, наявність кореляції повинна призводити до впорядкування значень другої змінної (тобто до їх тренду), відсутність кореляції не повинна змінювати випадковий характер поведінки значень *Y* при їх розміщенні вздовж впорядкованої послідовності значень *X*.

- Рангова кореляція

Розуміючи під ***рангом вибіркового значення випадкової величини*** його номер, у впорядкованій за зростанням вибірці, можна розглядати для оцінки сили зв'язку випадкових величин не їх чисельні значення, а відповідні їм ранги. В SASBase реалізовані наступні методи рангової кореляції:

- Гефдінга;

- Кендалла;

- Спірмена.

Існує ряд ситуацій, коли обчислення коефіцієнтів рангової кореляції доцільно при визначенні щільності зв'язку саме двох кількісних ознак. Наприклад, якщо зв'язок двох кількісних ознак має нелінійний, але монотонний характер. Якщо кількість об'єктів у вибірці невелика або якщо для дослідника суттєва інформація про знак зв'язку, то використання кореляційного відношення Пірсона може виявитися неадекватним. Обчислення ж коефіцієнта рангової кореляції дозволяє обійти зазначені труднощі.

- Кореляція Гефдінга

Критерій Гефдінга (Hoeffding) є ранговим аналогом критерію Блума-Кіфера-Розенблатта, тому спочатку буде розглянуто саме критерій Блума-

Кіфера-Розенблатта, для того щоб збільшити ступінь розуміння кореляції Гефдінга.

- Кореляція Блума-Кіфера-Розенблатта

Статистика критерію, запропонованого Блумом, Кіфером і Розенблаттом, будується наступним чином. Є сукупність точок (x_i, y_i), $i = 1,...,n$. Через точку з координатами (x_i, y_i) проводяться прямі, паралельні осям координат і підраховується кількість точок $m_1(i)$, що знаходяться в першому квадранті (для якого $x_j > x_i$ та $y_j > y_i$), $m_2(i)$ – перебувають в другому квадранті (для якого $x_j < x_i$ та $y_j > y_i$), $m_3(i)$ – знаходяться в третьому квадранті (для якого $x_j < x_i$ та $y_j < y_i$), $m_4(i)$ – знаходяться в четвертому квадранті (для якого $x_j > x_i$ та $y_j < y_i$).

Статистикою критерію є величина:

$$B = \frac{1}{n^3} \cdot \sum_{i=1}^{n} \left[m_1(i) \cdot m_4(i) - m_2(i) \cdot m_3(i) \right]^2$$

Критичні значення $B(\alpha)$ при $\alpha \to \infty$ ($n > 30$) дорівнюють: $B(0{,}90) = 0{,}0469$; $B(0{,}95) = 0{,}0584$; $B(0{,}99) = 0{,}868$. При $B \geq B(\alpha)$ кореляція вважається значимою з ймовірністю α.

Приклад обчислення значення кореляції Блума-Кіффера-Розенблатта

Для набору даних, наведеного в таблиці 9.9, обчислимо значення кореляції Блума-Кіффера-Розенблатта при значенні довіри ймовірності $\alpha = 0{,}95$

Таблиця 9.9 Набір даних для прикладу обчислення кореляції Блума-Кіффера-Розенблатта

n	*x(i)*	*y(i)*	$m_1(i)$	$m_2(i)$	$m_3(i)$	$m_4(i)$
1	1	12	0	2	0	6
2	2	17	0	0	2	6
3	4	8	3	3	0	2
4	12	14	1	0	8	1
5	7	1	6	3	0	0
6	6	2	2	4	0	2
7	7	1	0	2	0	6
8	8	13	6	0	3	0
9	9	4	4	3	0	0
10	10	10	3	1	4	1

На рис. 9.12 представлена графічна ілюстрація розміщення значень по відповідним квадрантам.

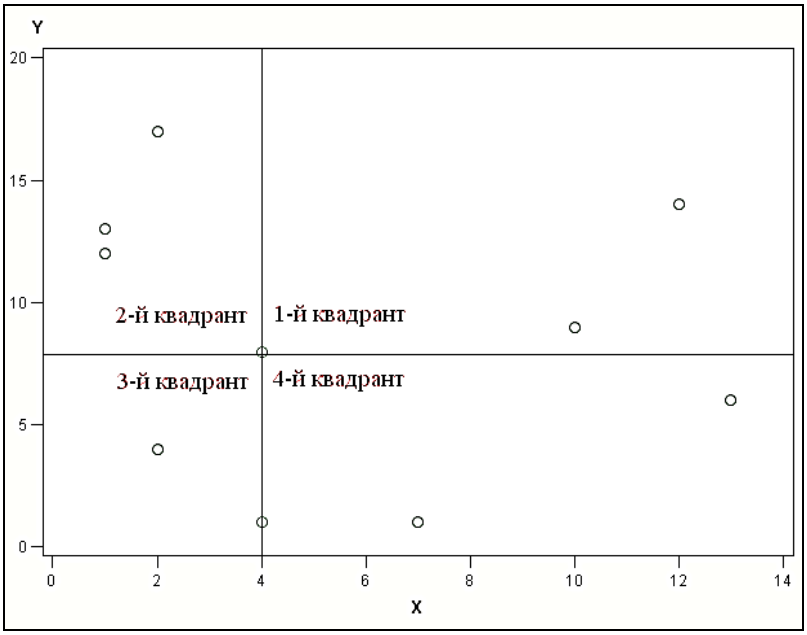


Рис. 9.12 Діаграма розсіювання для набору даних з табл. 9.9 при обчислення кореляції Блума-Кіффера-Розенблатта

Статистика критерію дорівнює

$$B = \frac{1}{10^3} \cdot \sum_{i=1}^{10} \left[m_1(i) \cdot m_4(i) - m_2(i) \cdot m_3(i) \right]^2 =$$

$$= 10^{-3} \cdot \left[(0 \cdot 6 - 2 \cdot 0)^2 + (0 \cdot 6 - 0 \cdot 2)^2 + ... + (3 \cdot 1 - 1 \cdot 4)^2 \right] = 0{,}054$$

оскільки $B = 0{,}054 < B(0{,}95) = 0{,}0584$, з імовірністю $\alpha = 0{,}95$ гіпотеза про наявність кореляції між змінними відхиляється.

Обчислення кореляції Гефдінга

Статистика критерію Гефдінга будується наступним чином: значення *X* та *Y* спочатку розміщають за рангом, а потім замінюються їх рангами.

Для обчислення використовуються наступні формули:

$$D = 30 \cdot \frac{(n-2) \cdot (n-3) \cdot D_1 + D_2 - 2 \cdot (n-2) \cdot D_3}{n \cdot (n-1) \cdot (n-2) \cdot (n-3) \cdot (n-4)}$$

де

$$D_1 = \sum_i \left((Q_i - 1) \cdot (Q_i - 2) \right)$$

$$D_2 = \sum_i \left((R_i - 1) \cdot (R_i - 2) \cdot (S_i - 1) \cdot (S_i - 2) \right)$$

$$D_3 = \sum_i \left((R_i - 2) \cdot (S_i - 2) \cdot (Q_i - 1) \right).$$

$$Q_i = 1 + \sum_{\substack{v=1 \\ v \neq i}}^{n} \left(\varphi(x_v, x_i) \cdot \varphi(y_v, y_i) \right),$$

де

$\varphi(x_v, x_i) = 1$, якщо $x_v < x_i$

$\varphi(x_v, x_i) = \frac{1}{2}$, якщо $x_v = x_i$

$\varphi(x_v, x_i) = 0$, якщо $x_v > x_i$

$\varphi(y_v, y_i) = 1$, якщо $y_v < y_i$

$\varphi(y_v, y_i) = \frac{1}{2}$, якщо $y_v = y_i$

$\varphi(y_v, y_i) = 0$, якщо $y_v > y_i$

Інтерпретація результатів обчислення кореляції Гефдінга в SAS Base

Область допустимих значень: $D \in [-0{,}5;\ +1]$

$D = +1$ – повна залежність між змінними *X* та*Y*.

При цьому кореляція Гефдінга не вказує, яка саме залежність: позитивна чи негативна.

$D = -0{,}5$ – повна незалежність між змінними *X* та*Y*.

$D < -0{,}5$.

При невеликому обсязі вибірки (малих *n*) і сильного зв'язку між *X* та *Y* , коефіцієнт Гефдінга може приймати значення менше -0,5.

Приклад обчислення кореляції Гефдінга в SAS Base

В якості прикладу обчислення використаємо дані з табл. 9.8.

Код програми має наступний вигляд:

```
proccorrdata=DATA_COR HOEFFDING;
        var X2;
        with X1;
run;
```

Результати роботи мають наступний вигляд:

```
Hoeffding Dependence Coefficients, N = 10
Prob> D under H0: D=0

X2
              X1     -0.13393
                      1.0000
```

Отримані результати мають наступне пояснення:

- -0,13393 - значення коефіцієнта кореляції Гефдінга;
- друге значення в таблиці результатів – ймовірність відхилення нульової гіпотези про існування зв'язку між змінними.

В даному прикладі кореляція Гефдінга показує, що між змінними відсутній зв'язок з імовірністю в 100% - це може бути пов’язано з малим розміром популяції.

Коефіцієнт кореляції рангів Спірмена

- класичний підхід щодо обчислення кореляції Спірмена

В 1904 році Спірмен запропонував близькість двох рядів *X* та *Y* характеризувати статистикою:

$$S = \sum_{i=1}^{n} (R_i - S_i)^2 ,$$

де R_i – ранг і-го значенні в *X* , а S_i – ранг і-го значенні в *Y*.

$S = 0$ тоді і тільки тоді, коли послідовності повністю збігаються.

$S \to \max$ коли послідовності повністю протилежні.

При цьому шукана сума – це сума квадратів послідовних непарних чисел:

$$S_{\max} = (n-1)^2 + [(n-1)-2]^2 + ... + 1^2 = \frac{(n^3 - n)}{3}$$

Коефіцієнт Спірмена:

$$P = 1 - \frac{2 \cdot S}{S_{\max}} = 1 - \frac{2 \cdot \sum_{i=1}^{n} (R_i - S_i)^2}{(n^3 - n)/3} = 1 - \frac{6 \cdot \sum_{i=1}^{n} (R_i - S_i)^2}{n \cdot (n^2 - 1)}$$

Область допустимих значень: $P \in [0;\ +1]$

$P = 0$ – змінний не залежні.

$P < 0{,}3$ – слабкий зв'язок або зв'язку немає.

$P \in [0{,}3;\ 0{,}7)$ – помірний зв'язок.

$P \geq 0{,}7$ – сильний зв'язок.

Різницю рангів кожної пари значень позначають, як $D_i = R_i - S_i$, *n*–число спостережень.

$$P == 1 - \frac{6 \cdot \sum_{i=1}^{n} D_i^2}{n \cdot (n^2 - 1)}$$

Приклад обчислення коефіцієнта Спірмена за класичним підходом

Для набору даних з табл. 9.8, виконаємо обчислення кореляції Спірмена за класичним підходом.

Таблиця 9.10 Таблиця проміжних результатів обчислення кореляції Спірмена

N	X_1	X_2	*Ранг* X_1	*Ранг* X_2	*Різниця рангів D*
1	0	1	0	1	-1
2	1	0	1	0	1
3	0	1	0	1	-1
4	1	0	1	0	1
5	0	1	0	1	-1
6	0	1	0	1	-1
7	1	0	1	0	1
8	0	1	0	1	-1
9	1	0	1	0	1
10	1	1	1	1	0

$$P == 1 - \frac{6 \cdot \sum_{i=1}^{n} D_i^2}{n \cdot \left(n^2 - 1\right)} = 1 - \frac{6 \cdot 9}{10 \cdot \left(100 - 1\right)} = 1 - \frac{54}{990} = 0{,}9455$$

Обчислення коефіцієнта Спірмена в SAS Base

В SAS Base реалізований наступний підхід до обчислення коефіцієнта Спірмена.

$$P = \frac{\sum \left(R_i - \bar{R}\right) \cdot \left(S_i - \bar{S}\right)}{\sqrt{\sum \left(R_i - \bar{R}\right)^2 \cdot \sum \left(S_i - \bar{S}\right)^2}},$$

де

R_i – ранг і-го значенні вX.

S_i – ранг і-го значенні вY.

$\bar{R}$ – математичне сподівання для рангів R_i.

$\bar{S}$ – математичне сподівання для рангів S_i.

Для набору даних з таблиці 9.8 обчислимо кореляцію Спірмена, користуючись наведеними формулами. В табл. 9.11 наведені проміжні результати у вигляді розставлених рангів.

Таблиця 9.11 Таблиця проміжних результатів обчислення кореляції Спірмена

N	X_1	X_2	*Ранг* X_1	*Ранг* X_2
1	0	1	0	1
2	1	0	1	0
3	0	1	0	1
4	1	0	1	0
5	0	1	0	1
6	0	1	0	1
7	1	0	1	0
8	0	1	0	1
9	1	0	1	0
10	1	1	1	1

$\bar{R} = 0{,}5$

$\bar{S} = 0{,}6$

$$P = \frac{-2}{\sqrt{2{,}5 \cdot 2{,}4}} = \frac{-2}{\sqrt{6}} = -0{,}81649$$

Приклад обчислення коефіцієнта Спірмена в SAS Base

Для набору даних з табл. 9.8 (теж саме в табл. 9.11) обчислимо коефіцієнт Спірмена за допомогою наступного коду sas-програми:

```
proccorrdata=DATA_COR SPEARMAN;
        var X2;
        with X1;
run;
```

Результати роботи мають наступний вигляд:

```
Spearman Correlation Coefficients, N = 10
Prob> |r| under H0: Rho=0

X2
              X1     -0.81650
                      0.0039
```

Отримані результати мають наступне пояснення:

- -0,8165 - значення кореляції Спірмена.
- 0,0039 - ймовірність відхилення нуль-гіпотези. У даному випадку з імовірністю 0,39% між змінними немає зв'язку.

Чисельний результат класичної формули обчислення коефіцієнта Спірмена не збігається з результатом, що видає система SASBase, в силу відмінності формул, що використовуються для обчислення значення кореляції. Використання формули для системи SAS задає область допустимих значень коефіцієнта $-1 \le P \le +1$, в класичному ж варіанті $0 \le P \le +1$.

Область допустимих значень коефіцієнта Спірмена в SAS Base $-1 \le P \le +1$, а сила зв'язку визначається як:

$P = 1$– строга лінійна кореляція.
$P = 0$ – відсутність кореляції між перемінними.
$P = -1$ – строга негативна лінійна кореляція.

Коефіцієнт кореляції тау-б Кендалла

Класичні варіант коефіцієнта Кендалла

В 1938 році Кендалл запропонував свій коефіцієнт, який позначається як τ , а вимовляється як тау-б.Застосовується для виявлення взаємозв'язку між кількісними або якісними показниками, якщо їх можна ранжувати. Значення показника *X* впорядковують за зростанням і присвоюють ранги. Ранжирують значення показника *Y* і розраховують коефіцієнт кореляції Кендалла. В літературі можна зустріти різні формули для обчислення. Найчастіше зустрічається наступна формула:

$$\tau = \frac{n_c - n_d}{\frac{1}{2} \cdot n \cdot (n-1)} \quad ,$$

де

n_c – кількість збігів пар.

n_d – кількість розбіжностей пар.

n – розмір навчальної вибірки.

Про те, які пари маються на увазі, буде більш зрозуміло з практичного прикладу, що наведений нижче.

Наведемо приклад обчислення кореляції Кендалла за класичним підходом. Розглянемо набір даних з табл. 9.12. Для підвищення якості розуміння процесу обчислення, кожний запис позначаємо літерою.

Таблиця 9.12 Дані для прикладу обчислення кореляції Кендалла.

	n	Рік	X_1 (кількість адміністративних порушень на підприємствах, тис.)	X_2 (кількість кримінальних порушень на підприємствах, тис.)
A	1	1992	79	10
B	2	1993	75	7
C	3	1994	45	5
D	4	1995	38	6
E	5	1996	59	4
F	6	1997	93	12
G	7	1998	68	8

Складемо таблицю пар, як наведено в табл. 9.13.

Таблиця 9.13 Таблиця пар, побудована по даним з табл. 9.12

	Пара	*X(1,i) ? X(1,j)*	*X(2,i) ? X(2,j)*	tick / cross
1	AB	6>5	6>4	
2	AC	6>2	6>2	
3	AD	6>1	6>3	
4	AE	6>3	6>1	
5	AF	6<7	6<7	
6	A G	6>4	6>5	
7	B C	5>2	4>2	
8	B D	5>1	4>3	
9	B E	5>3	4>1	
10	B F	5<7	4<7	
11	B G	5>4	4<5	X
12	C D	2>1	2<3	X
13	C E	2<3	2>1	X
14	C F	2<7	2<7	
15	C G	2<4	2<5	
16	D E	1<3	3>1	X
17	D F	1<7	3<7	
18	D G	1<4	3<5	
19	E F	3<7	1<7	
20	E G	3<4	1<5	
21	F G	7>4	7>5	

$$n_c = 21 - n_d = 21 - 4 = 17$$

$$n_d = 4$$

$$\tau = \frac{n_c - n_d}{\frac{1}{2} \cdot n \cdot (n-1)} = \frac{17 - 4}{\frac{1}{2} \cdot 7 \cdot 6} = 0{,}619$$

Обчислення коефіцієнта Кендалла в SAS Base

В SAS Base реалізований наступний підхід до обчислення коефіцієнта Кендалла:

$$\tau = \frac{\sum_{i<j} \operatorname{sgn}(x_i - x_j) \cdot \operatorname{sgn}(y_i - y_j)}{\sqrt{(T_0 - T_1) \cdot (T_0 - T_2)}},$$

де

$$T_0 = \frac{n \cdot (n-1)}{2}$$

$$T_1 = \frac{\sum t_i \cdot (t_i - 1)}{2}$$

$$T_2 = \frac{\sum u_i \cdot (u_i - 1)}{2}$$

t_k – кількість зв'язків змінної x в k – й групі змінної x.

u_k – кількість зв'язків змінної y в k – й групі змінної y.

n – кількість спостережень.

$\operatorname{sgn}(z) = 1$, якщо $z > 0$.

$\operatorname{sgn}(z) = 0$, якщо $z = 0$.

$\operatorname{sgn}(z) = -1$, якщо $z < 0$

Приклад обчислення коефіцієнта Кендала в SAS Base

Для набору даних з табл. 9.12 обчислимо коефіцієнт Кендала за допомогою наступного коду sas-програми:

```
data DATA_KENDALL;
input n X1 X2;
datalines;
1 79 10
2 75 7
3 45 5
4 38 6
5 59 4
6 93 12
7 68 8
RUN;
```

```
PROCCORRDATA=DATA_KENDALL KENDALL;
        VAR X2;
        WITH X1;
RUN;
```

Результати роботи мають наступний вигляд:

```
Kendall Tau b Correlation Coefficients, N = 7
Prob> |tau| under H0: Tau=0

X2
                X1      0.61905
                        0.0509
```

Отримані результати мають наступне пояснення:

- 0,61905 – значення кореляції тау-б Кендалла.
- 0,0509 – ймовірність відхилення нуль-гіпотези. У даному випадку з імовірністю 5% між змінними немає зв'язку.

Чисельний результат класичної формули обчислення коефіцієнта Кендалла збігається з результатом, розрахованим системою SASBase.

Область допустимих значень коефіцієнта Кендалла: $-1 \le \tau \le +1$:

$\tau = 1$ – сувора лінійна кореляція.

$\tau = 0$ – відсутність кореляції між змінними.

$\tau = -1$ – сувора негативна лінійна кореляція.

9.11 Використання PROC REG для однофакторного регресійного аналізу

Процедура **REG** є процедурою побудови лінійної регресійної моделі методом наближення найменшими квадратами для проведення регресійного аналізу. Процедура є однією з багатьох процедур продукту **SAS / STAT**. Нижче буде розглянутий приклад однофакторної регресійної моделі. Процедура REG здатна будувати моделі з множиною регресійних змінних та за різними способами побудови, включаючи покрокову регресію (stepwise), пряму вибірку (forward), виключення на зворотному проході (backward).

Продукт **SAS/STAT** пропонує багато інших статистичних процедур для побудови нелінійних, комбінованих і логістичних регресій, методів дисперсійного аналізу та інших аналітичних статистичних методів. Окрім цього, в продукті **SAS ETS** реалізовані методи аналізу часових рядів. Для більш докладного вивчення цих продуктів, рекомендується звернутися до спеціалізованої літератури або пройти відповідні навчальні курси SAS.

Процедура **REG** має тільки два обов'язкових оператора. Вона повинна починатися з оператора **PROC REG** і мати в тілі процедури оператор **MODEL**, що визначає вид бажаної моделі:

PROC REG;
MODEL dependent = independent;

В операторі **MODEL** – залежна змінна (для опису якої будується модель) вказується зліва, а незалежна (або декілька незалежних) – вказується праворуч.

Оператор **PLOT** – це один з багатьох операторів процедури **REG**. Даний оператор використовують для створення діаграм розсіювання за наявними даними, включаючи деякі статистики, обчислені за регресійним методом. Для використання оператора **PLOT** необхідно мати додатково продукт **SAS GRAPH**, оскільки цей оператор є його частиною.

PLOT dependent = independent;

При відсутності **SAS GRAPH**, можна використовувати опцію **LINEPRINTER** в операторі **PROC** для псевдографічного виводу:

PROC REG LINEPRINTER;
MODEL *dependent=independent*;
PLOT *dependent*independent = 'symbol'* P. *
***independent = 'symbol'* / OVERLAY;**

Параметр **symbol** задає вид символу псевдографіки для відображення ліній на графіку. При відображенні декількох ліній на одному графіку, бажано використовувати різні символи. Опція **P** – це ключове слово для позначення значень прогнозу.

В даному підрозділі розглянуто спосіб отримання графіків ліній регресії за допомогою процедури **REG**. Взагалі ця процедура містить набагато більше можливостей та відповідних опцій, що дозволяють отримувати інші види графіків, зокрема, діаграму залишків, діаграму залишків Стьюдента, довірчий інтервал для регресії та інші.

Синтаксис PROCREG

Синтаксис процедуры REG
PROC REG < options >**;**
<label: >**MODEL** dependents=<regressors>< / options >**;**
BY variables **;**
FREQ variable **;**
ID variables **;**
VAR variables **;**
WEIGHT variable **;**
ADD variables **;**
DELETE variables **;**
<label: >**MTEST** <equation, ... ,equation>< / options >**;**
OUTPUT <**OUT=**SAS-data-set > keyword=names < ... keyword=names >**;**

PAINT <condition | **ALLOBS**>< / options > | <**STATUS | UNDO**>**;**
PLOT <yvariable*xvariable><=symbol>< ...yvariable*xvariable><=symbol>< / options>**;**
PRINT < options ><**ANOVA** ><**MODELDATA** >**;**
REFIT;
RESTRICT equation, ... ,equation**;**
REWEIGHT <condition | **ALLOBS**>< / options > | <**STATUS | UNDO**>**;**
<label: >**TEST** equation,<, ...,equation>< / option >**;**

Приклад використання PROCREG

На основі зробленого аналізу привабливості нерухомості (дані для аналізу взяті з файлу PRICE, що був створений у підрозділі 9.6), було з'ясовано, що між вартістю квадратного метра житла (TARGET_PRICE) та поверхом, на якому воно розташоване (STOREY), існує значимий кореляційний зв'язок. На основі цієї інформації сформулюємо нульову і альтернативну гіпотези. H_0– вартість квадратного метра житла не залежить від його поверховості, H_1– існує залежність.

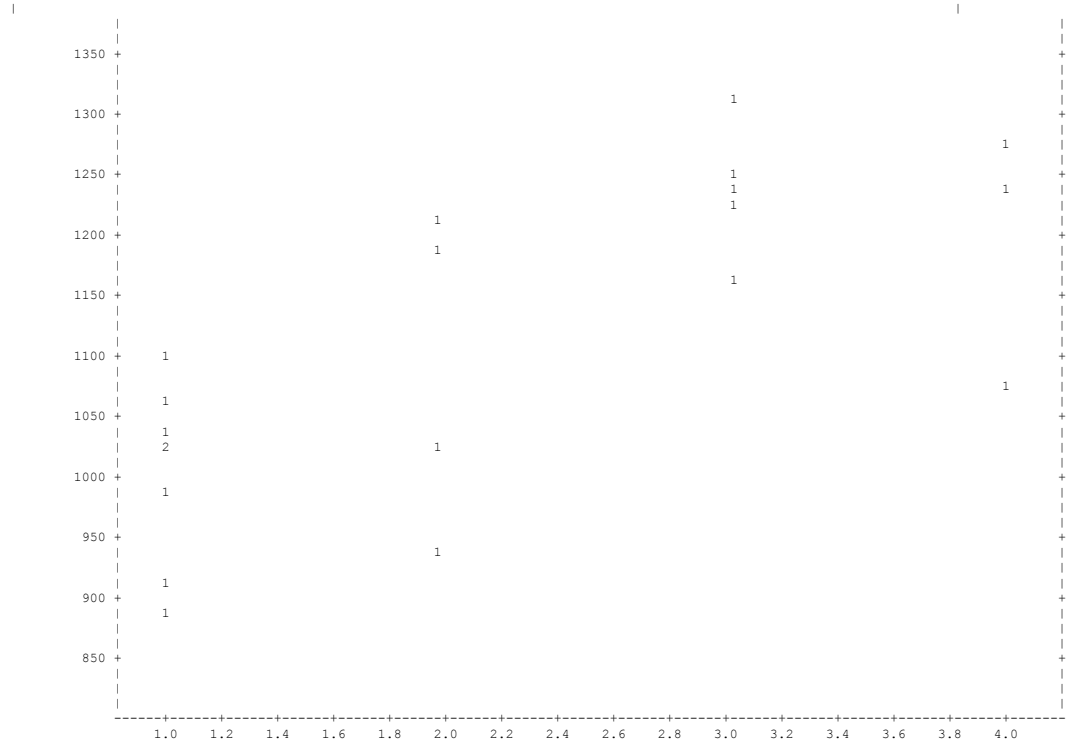
Код програми виглядає наступним чином:

```
procregdata=PRICE LINEPRINTER;
        model TARGET_PRICE=Storey;
        plot TARGET_PRICE*Storey;
        title'Результатрегрессионногоанализа';
run;
```

Результати роботи процедури **REG** для наведеного коду sas-програми мають вигляд:

```
Результатрегрессионногоанализа

TheREGProcedure
Model: MODEL1
                                  Dependent Variable: TARGET_PRICE

Number of Observations Read          20
Number of Observations Used          20

          Analysis of Variance

                                  Sum of           Mean
Source                    DF      Squares         Square    F Value    Pr> F

Model                      1       165003         165003      19.14    0.0004
Error                     18       155188     8621.53514
Corrected Total           19       320191


Root MSE             92.85222    R-Square     0.5153
Dependent Mean     1109.65000    Adj R-Sq     0.4884
CoeffVar             8.36770


                          Parameter Estimates

                       Parameter       Standard
Variable        DF      Estimate          Error    t Value    Pr> |t|
Intercept        1     933.02815       45.39884      20.55      <.0001
```

```
Storey           1       81.95909       18.73457       4.37      0.0004

The REG Procedure
          Model: MODEL1
Dependent Variable: TARGET_PRICE
TARGET_PRICE
```



Результат роботи процедури можна розділити на три частини – графічну і дві текстові.

У першій текстовій частині представлені характеристики даних (табл. 9.14).

Таблиця 9.14 Перелік статистик, що виводяться в першій частині звіту процедури REG

Статистика	**Опис статистики**
DF	Число ступенів свободи
Mean Square	Середнє квадратичне (сума квадратів поділена на число ступенів свободи)
F-value	F-статистика, тест нульової гіпотези
Pr > F	Достовірність (імовірнісна)
Root MSE	Корінь з середньоквадратичної похибки середнього
CoeffVar	Коефіцієнт дисперсії
Adj R-sq	R-квадрат з урахуванням ступенів свободи.

В другій частині звіту виводяться параметри моделі (табл.9.15).

Таблиця 9.15 Перелік статистик, які описують отримані математичну модель, що виводяться в другій частині звіту процедури REG

Статистика	Опис статистики
DF	Число ступенів свободи змінної
t value	Т-тест на рівність параметра нулю
Pr > | t |	Достовірність (імовірнісна) параметра

Таким чином, за результатами, представленими у другій частині, можна побудувати наступну модель залежності ціни квадратного метра від поверховості:

$$TARGET_{PRICE} = 933{,}02815 + 81{,}95909 \times STOREY$$

Саме цю лінійну залежність відображає графічна частина результатів, яка містить лінію регресії і точки розташування реальних даних. В наведеному прикладі вартість дійсно залежить від поверху. Модель є статистично значущою з оцінкою вірогідності моделі 0,0004 (статистика *Pr>F*), однак залежність є середньою, тому що R-квадрат дорівнює 0,5153 (статистика R-Square), що відповідає залежності вартості житла від поверховості приблизно на 51,53%, можливо для однофакторного аналізу кращими оцінками буде матеріали стін будівлі або стан житла.

Тест на рівність параметра нулю виключає нульову гіпотезу з достовірністю 0,0004 (статистика *Pr* > |*t*|), таким чином, оцінка параметра також є значущою.

9.12 Використання PROC ANOVA для однофакторного дисперсійного аналізу

Процедура **ANOVA** є однією з процедур **SAS/STAT**, призначена для проведення дисперсійного аналізу. Дана процедура спеціально розроблена для рівноважних збалансованих даних, тобто таких, де число спостережень однаково в кожній класифікаційній групі. Якщо дані не задовольняють цій умові, то слід використовувати інші процедури **SAS/STAT**, наприклад, **GLM**, яка за своїм синтаксису схожа на процедуру **ANOVA**.

Процедура **ANOVA** має два обов'язкових оператора – **CLASS** і **MODEL**:

```
PROC ANOVA;
    CLASS variable-list;
    MODEL dependent = effects;
```

Оператор **CLASS** повинен бути визначений раніше оператора **MODEL** для попереднього задання змінних класифікації. Для однофакторного дисперсійного аналізу вказують тільки одну змінну класу. **MODEL** – визначає залежну змінну і ефекти, а для однофакторного аналізу ефектом є змінна класу.

Як і попередніх процедури, **ANOVA** містить велику кількість операторів і опцій налаштування. Один з них, оператор **MEANS**, що обчислює середнє

залежної змінної для всіх ефектів, вказаних в **MODEL**. Відзначимо, що оператор **MEANS** реалізує ряд порівняльних тестів, наприклад тест Бонферроні – опція **BON**, багатовимірний тест Дункана – опція **DUNCAN**, *T*-тест, порівняльний тест Шеффе – **SCHEFF**, тест Туке – **TUKEY** та багато інших.

MEANS effects / options;

Ефектами в операторі можуть бути будь-які із перерахованих в моделі.

Синтаксис PROC ANOVA

PROC ANOVA < options >**;**
CLASS variables **;**
MODEL dependents=effects < / options >**;**
ABSORB variables **;**
BY variables **;**
FREQ variable **;**
MANOVA < test-options >< / detail-options >**;**
MEANS effects < / options >**;**
REPEATED factor-specification < / options >**;**
TEST <**H=**effects >**E=**effect **;**

Приклад використання PROC ANOVA

Аналіз результатів виступів юніорських команд з волейболу показав, що команди, які складаються з підлітків вищого зросту, ніж їх суперники, частіше перемагають.

В рамках наступного прикладу перевіряється гіпотеза – чи дійсно є значуща відмінність між командами.

Sas-код програми для створення набору даних має вигляд:

```
data teams (keep=team height);
format team $7.;
input t $1. +1 height 2. +1 @@;
If t='r'then team='red';
If t='b'then team='blue';
If t='g'then team='gray';
If t='v'then team='violet';
If t='m'then team='magenta';
if t ^= "thenoutput;
cards;
r 55 r 48 r 53 r 47 r 51 r 43 r 45 r 46 r 55 r 54 r 45 r 52
b 46 b 56 b 48 b 47 b 54 b 52 b 49 b 51 b 45 b 48 b 55 b 47
g 55 g 45 g 47 g 56 g 49 g 53 g 48 g 53 g 51 g 52 g 48 g 47
v 53 v 53 v 58 v 56 v 50 v 55 v 59 v 57 v 49 v 55 v 56 v 57
m 53 m 55 m 48 m 45 m 47 m 56 m 55 m 46 m 47 m 53 m 51 m 50;
run;
```

Зауважимо, що в наведеному sas-коді використаний оператор **KEEP**, призначений для відбору для подальшого аналізу тільки змінних teamheight,.

Код для перевірки гіпотези виглядає наступним чином:

```
procanovadata=teams;
class team;
model height=team;
means team / scheffe;
title'Рост спортсменов';
run;
```

Результати роботи процедури **ANOVA** для наведеного коду sas-програми мають вигляд:

```
Ростспортсменов
                The ANOVA Procedure

Dependent Variable: height

                                    Sum of
Source                      DF         Squares     Mean Square    F Value    Pr> F
Model                        4     228.0000000      57.0000000       4.14    0.0053
Error                       55     758.0000000      13.7818182
Corrected Total             59     986.0000000


           R-Square     CoeffVar      Root MSE    height Mean
                            0.231237      7.279190      3.712387       51.00000


Source                      DF        Anova SS     Mean Square    F Value    Pr> F
team                         4     228.0000000      57.0000000       4.14    0.0053

                         The ANOVA Procedure
КритерийШеффедля height

  NOTE: This test controls the Type I experimentwise error rate.


Альфа                              0.05
Степени свободы ошибки               55
Средний квадрат ошибки          13.78182
CriticalValueof F               2.53969
Минимальная значимая разность    4.8306


Means with the same letter are not significantly different.

ScheffeGrouping          Mean      N   team
          A         54.833     12    violet
          A
     B    A         50.500     12    magenta
    B     A
    B     A        50.333     12    gray
     B
    B              49.833     12    blue
B
     B              49.500     12    red
```

Кожна команда складається з 12 юніорів, тобто дані збалансовані і отже, можна використовувати процедуру ANOVA. Для з'ясування, яка команда значимо "вище інших", застосуємо множинне порівняння Шеффе – оператор SCHEFFE. Оператор MEANS обчислить середнє значення для кожної команди.

Результати роботи процедури ANOVA також складаються з декількох частин. Перша частина містить:

– інформацію про змінну класу – TEAM, що вона має 5 значень і всі вони виведені – blue, gray, magenta, red, violet.

– кількість спостережень дорівнює 60.

Друга частина звіту містить безпосередньо характеристики дисперсійного аналізу. В таблиці 9.16 наведений опис відповідних статистик.

Таблиця 9.16 Перелік статистик, що виводяться в другій частині звіту процедури ANOVA

Статистика	**Опис статистики**
Source	Джерело варіації
DF	Число ступенів свободи для моделі, помилки і загальне
Sum of Square	Суму квадратів для моделі, помилки і загальна
Mean Square	Середнє квадратичне (сума квадратів поділена на число ступенів свободи)
F value	F статистика, тест нульової гіпотези (середнє квадратичне моделі діленої на середнє квадратичне помилки середнього)
Pr > F	Достовірність (імовірнісна) - F-статистика
R-square	R-квадрат
Coeff Var	Коефіцієнт дисперсії
Root MSE	Корінь з середньоквадратичної похибки середнього
Height Mean	Середнє значення залежної змінної

З отриманих результатів видно, що модель є значущою, тому що достовірність дорівнює 0,0053 (статистика Pr > F). Отже, гіпотеза про рівність росту в командах відкидається.

Третя частина звіту містить порівняльну інформацію про команди. Команди, не мають достовірної значущої різниці у зрості. Об'єднані в групи. Групи мають мітки A, B, C, і так далі за абеткою. Таким чином, видно, що для рівня значущості альфа рівного 0,05, найбільшу різницю мають команди RED і VIOLET, середній зріст 49,500 та 54,833 відповідно.

9.13 Досвід використання програмного забезпечення SAS у ННК ІПСА НТУУ КПІ

Існує досить багато інструментів статистичної обробки даних, як комерційних, так і безкоштовних. Найбільш популярні в академічному середовищі – це Matlab, MathCad, Mathematica, Statistica, R, Python та інші. На вибір Statistical Analysis System в якості практичного інструмента для цієї книги вплинув той факт, що автори мають доступ до системи з 2007 року, коли в Україні було відкрито офіс компанії SAS Institute.

В 2010 році Науково-навчальний комплекс "Інститут Прикладного Системного Аналізу" (ННК ІПСА), який входить до складу Національного Технічного Університету України "Київський Політехнічний Інститут" (НТУУ КПІ) приєднався до Глобальної Академічної Програми SAS, що призначена для професорів, студентів, дослідників та керівників ВУЗів. Ця програма підтримує використання програмного забезпечення SAS у навчанні, викладанні та наукових дослідженнях.

Доступ до програмного забезпечення SAS на пільгових умовах дозволив доповнити вже існуючі курси навчальних дисциплін практично значущими лабораторними роботами, а також розробити нові навчальні курси, як-то "Прикладна комп'ютерна лінгвістика із використанням SAS Text Miner", "Прикладна статистика із використанням SAS", "Аналіз, моделювання та прогнозування із використанням SAS Enterprise Miner". Отримані знання надають можливість випускникам працювати в організаціях, де використовується система, а це майже половина українських банків зі списку TOP 30, телекомунікаційна компанія МТС, страхові компанії, тощо.

В лютому 2012 року на базі ННК ІПСА НТУУ КПІ, було створено Центр Компетенції SAS, задачами якого є:

- надання консультаційних послуг при впроваджені системи SAS в державних та комерційних установах;
- підготовка методичних навчальних матеріалів;
- навчання студентів та фахівців роботі з системою та галузевими продуктами SAS;
- проведення досліджень в різноманітних предметних галузях на основі використання аналітичних і статистичних підходів, з метою виявлення закономірностей та побудови математичних моделей, їх впровадження та подальший моніторинг;
- розробка нових методів статистичного та інтелектуального аналізу даних.

Співробітники Центру Компетенції SAS при ННК ІПСА брали участь в тестуванні пакету українського мови для продуктів компанії, призначених для вирішення задач текстової аналітики. Окрім цього, вони є активними учасниками проектів по налаштуванню та впровадженню системи в різних організаціях.

9.14 Контрольні запитання і вправи

Контрольні запитання

1. Для чого використовуються магічні квадранти аналітичної компанії Gartner? Дайте класифікацію для лідерів, претендентів, провидців та нішевих компаній.
2. В яких форматах дозволяє створювати звіти система SAS Base? Який формат використовується за замовчуванням при первинному налаштуванні системи?
3. В якому році та ким була створена компанія SAS Institute? Для якої сфери людської діяльності використовувалися самі перші розробки компанії?
4. Які індустріальні та спеціалізовані рішення окрім SAS Base ще пропонує компанія? Скільки всього та для яких сфер існує рішень на базі платформи SAS Base?
5. Які базові процедури статистичної обробки даних реалізовані в SAS Base? Наведіть приклад.
6. Що означає термін “популяція” в статистиці? Якими функціями характеризується?
7. Для чого використовуються вибіркові набори значень в статистиці? Від чого залежить точність або якість оцінок параметрів популяції?
8. Які методи локалізації розподілу можете навести? Надайте практичний приклад.
9. З якими мірами варіабельності знайомі? Наведіть формули для обчислення та наведіть практичний приклад.
10. Які міри форми використовуються в статистиці? Який між ними зв'язок з центральними моментами випадкових величин?
11. Що таке гаусівський нормальний розподіл? Поясніть правило однієї, двох та трьох сігм для нормального розподілу.
12. В чому полягає принцип тестування гіпотез для нормальних розподілів? Які статистики використовуються для перевірки гіпотез? Наведіть формули для обчислення та приклад використання.
13. Які оператори в SAS Base використовуються для створення наборів даних? Наведіть приклад відповідної sas-програми.
14. Для чого використовується PROC UNIVARIATE? Які типи розподілів та побудови графіків використовує PROC UNIVARIATE? Наведіть синтаксис та приклад використання.
15. Що представляє собою графік типу “коробка з вусами” та де використовується?
16. Які критерії оцінювання емпіричної вибірки знаєте? Наведіть відповідні формули для обчислення та дайте пояснення значенню p-value?
17. Для чого використовується PROC MEANS? Наведіть синтаксис та приклад використання.
18. Яким чином налаштовується довірчий інтервал для PROC MEANS? Для чого використовується операнд VAR? Яким чином проводяться групові обчислення,

19. Для чого використовується PROC FREQ? Наведіть синтаксис та приклад використання.
20. Для чого використовується PROC CORR? Наведіть синтаксис та приклад використання.
21. Поясніть суть кореляції Пірсона та наведіть формулу для обчислення? Які ще типи та види кореляції існують? Які з них підтримує PROC CORR? Яким чином виконується статистична перевірка наявності кореляції Пірсона?
22. Якими недоліками вона характеризується кореляція Пірсона? Наведіть класифікацію сили зв'язку та поясніть результати, що видаються в SAS Base для неї.
23. Що таке непараметричний кореляційний аналіз та чим він відрізняється від кореляції Пірсона? Наведіть приклад використання кореляції Блума-Кіфера-Розенблатта. Які види непараметричної кореляції реалізовані в PROC CORR?
24. Наведіть формулу та приклад sas програми для обчислення кореляції Гефдінга. Надайте класифікацію інтерпретації результатів кореляції Гефдінга в SAS Base.
25. Що таке кореляція рангів Спірмена? В чому різниця між класичним підходом та реалізованим в SAS Base? Наведіть приклад використання в SAS Base у вигляді sas програми.
26. Наведіть формули для обчислення коефіцієнта тау-б Кендалла за класичним підходом та реалізованим в PROC CORR. Наведіть приклад sas програми та дайте інтерпретацію результатів отриманих в SAS Base.
27. Для чого використовується PROC REG? Наведіть синтаксис та приклад використання.
28. Для чого використовуються операнди PLOT та MODEL в PROC REG? Яким чином можна побудувати графіки в PROC REG без використання SAS GRAPH?
29. Для чого використовується PROC ANOVA? Наведіть синтаксис та приклад використання.
30. Для чого використовуються операнди KEEP та DROP при роботі з наборами даних в SAS Base?

Вправи

1. В таблиці наведений лист з оцінками п'яти студентів.

ID	AGE	GENDER	Grade Point Average (GPA)	College Entrance Exam Score (CSCORE)
1	18	M	3.7	650
2	18	F	2.0	490
3	19	F	3.2	580
4	23	M	2.8	530
5	21	M	3.5	640

(a) Створить sas програму, що здатна створити відповідний data-set/

(b) Додайте до програми необхідні функції для обчислення середніх значень GPA та CSCORE.

(c) Для обчислення індексу успішності студентів пропонується використовувати наступну формулу:

$$INDEX = GPA + 3 \times CSCORE/500$$

Змініть sas програму відповідним чином, щоб для кожного студента можна було обчислити INDEX та надрукувати список студентів впорядкованим за значенням INDEX. Список повинен містити ID, GPA, CSCORE та INDEX.

2. В таблиці наведена інформація про учасників програми зі схуднення за спеціальною дієтою.

Variable Name	Description	Type	Starting Column	Ending Column
SUBJ	Subject number	Character	1	3
HEIGHT	Height in inches	Numeric	4	5
WT_INIT	Initial weight (in pounds)	Numeric	6	8
WT_FINAL	Final weight	Numeric	9	11

(a) Створіть файл даних, що буде міст наведену в таблиці інформацію. Окрім цього файл даних повинен містити додаткові змінні BMI_INIT (початковий індекс маси тіла), BMI_FINAL (індекс маси тіла після закінчення програми) та BMI_DIFF (зміна індексу маси тіла впродовж програми зі схуднення). BMI обчислюється як вага учасника програми в кілограмах поділена на його зріст в метрах в квадраті. Для конвертації ваги в фунтах в кілограми необхідно фунти ділити на коефіцієнт 2,2, для переводу дюймів в метри необхідно множити на 0,0254. BMI_DIFF обчислюється як різниця між BMI_FINAL та BMI_INIT. Зауваження: Великі значення BMI свідчать про збиткову вагу учасника програми. Якщо програма зі схуднення працює, то BMI_DIFF повинно приймати від'ємне значення.

(b) Сформуйте звіт, що містить значення набору даних за зменшенням значень змінної SUBJ. Звіт повинен містити змінні SUBJ, HEIGHT, BMI_INIT, BMI_FINAL та BMI_DIFF.

Для роботи можна використовувати наступні вхідні дані учасників програми:

00768155150
00272250240
00563240200
00170345298

3. В наявності є наступний файл даних:

Social Security Number	Annual Salary	Age	Race
123874414	28,000	35	W
646239182	29,500	37	B
012437652	35,100	40	W
018451357	26,500	31	W

(a) Створіть файл даних використовуючи дані з таблиці. Обчисліть середньорічну зарплатню та середній вік. Зауваження: у зв'язку з тим, що в розділі не розглядався випадок роботи з даними, що містять кому в якості розділювача розрядів та використовуйте замість неї крапку.

(b) Всі працівники сплачують 30% податок на дохід. Обчисліть розмір податку на основі інформації про річний дохід та надрукуйте звіт, що містить Social Security Number, річну зарплатню та податок. Інформація повинна бути впорядкована за Social Security Number.

4. Вам досталася sas-програма від колеги. Необхідно в цю програму внести необхідні зміни:

(a) Додайте дві нові змінні. OVERALL обчислюється, як середнє значення IQ, MATH та SCIENCE розділене на 500. GROUP задається рівним 1 якщо IQ приймає значення між 0 та 100 (включно), 2 якщо IQ приймає значення між 101 та 140 (включно), 3 якщо IQ приймає значення більше 140.

(b) Надрукуйте звіт, що містить всі дані, впорядкований за IQ.

(c) Обчисліть частоту значень змінної GROUP.

```
*** Program to read in IQ and other test scores;
DATA IQ_AND_TEST_SCORES;
        INPUT ID 1-3
            IQ 4-6
                MATH 7-9
                SCIENCE 10-12;
DATALINES;
001128550590
002102490501
003140670690
004115510510
;
```

5.

(a) Що неправильного в нижченаведеній sas програмі? (Надайте детальну відповідь у письмовому вигляді).

(b) Створить коректну sas-програму.

DATA MISTAKE;
INPUT ID **1-3** TOWN **4-6** REGION **7-9** YEAR **11-12** BUDGET **13-14**
VOTER-TURNOUT **16-20**
DATALINES;
00104005 0422 12345
(MORE DATA LINES)
;
PROC MEANS DATA=MISTAKE;
VAR ID REGION VOTER-TURNOUT;
N, STD, MEAN;
RUN;

6. При проведенні опитування респонденти відповідали на п’ять запитань. Відповідь на кожне з п’яти запитань відзначається літерою A, B, C, D або E. Нижче наведені результати опитувань восьми респондентів. Створить власну sas програму, що здатна зчитати ці відповіді у вигляді п’яти змінних QUES1, QUES2, QUES3, QUES4, QUES5. Зауваження: всі відповіді – це символи, а не цифри, між якими немає вільного місця. Також програма повинна генерувати частотний розподіл для кожного з п’яти запитань, при цьому звіт не повинен містити кумулятивні значення статистики.

Набір даних:

ABCDE
AACCE
BBBBB
CABDA
DDAAC
CABBB
EEBBB
ACACA

7. Велика промислова компанія заінтересована в розумінні свого типового споживача, що купує їх продукцію. CEO бажає щоб був створений профіль “типового покупця”. Набір даних про покупців складається з наступної інформації: age, gender, race, income, marital status, and homeowner/renter. Створить власний набір даних та напишіть sas програму, що будує профіль “типового покупця”.

Зауваження: можливі труднощі при створені власного набору даних

Рекомендації:

1. Для таких змінних як “homeowner”, краще всього використовувати бінарний тип, коли 0 – негативний стан / відповідь (відсутність власного

житла), 1 - позитивний стан / відповідь (наявність власного житла).

2. При групуванні чисельних значень в категорії, перевірте що створені групи задовольнять цілям аналізу. Наприклад, якщо товар який виробляє компанія це косметика для шкіри, то наступні вікові категорії покупців:

1 =< 21 2 = 21 – 35 3 = 36 – 50 4 => 50

дещо не задовольняють цілям аналізу, тому що головний покупець – це люди старшого покоління. Цілям аналізу більш відповідають наступні групи:

1 =< 50 2 = 50 – 59 3 = 60 – 69 4 => 69

8. В наявності є інформація про зарплатню та категорію працівника. Зробить sas програму, що створює набір даних з наступною інформацією: EMPID (ідентифікатор працівника), SALARY (зарплатня), JCLASS (категорія працівника) та BONUS. Змінна BONUS обчислюється наступним чином:

10% від зарплатні для працівників класу 1;

15% від зарплатні для працівників класу 2;

20% від зарплатні для працівників класу 3 (CEO та президент компанії).

На основі інформації про зарплатню та бонус обчисліть нову змінну NEW_SALARY, що є сумою зарплатні та бонусу. Використовуючи PROC PRINT створить звіт, що містить відповідний набір даних.

Список даних для роботи:

137 28000 1
214 98000 3
199 150000 3
355 57000 2

Змінна JCLASS може бути як чисельного так і рядкового типу.

10.В наявності є наступний набір даних:

ID	RACE	SBP	DBP	HR
001	W	130	80	60
002	B	140	90	70
003	W	120	70	64
004	W	150	90	76
005	B	124	86	72

Зауваження: SBP – це систолічний тиск крові, DBP – це діастолічний

тиск крові, а HR – це частота серцевих скорочень.

Створіть sas-програму, що виводить наступний звіт:

ID	RACE	SBP	DBP
003	W	120	70
005	B	124	86
001	W	130	80
002	B	140	90
004	W	150	90

Зауваження:

1. Звіт не повинен мати стовбець “OBS”, для цього треба задати операнд NOOBS в PROC PRINT.

2. Дані повинні бути виведені за збільшенням значення SBP.

3. Значення HR не повинні бути включені в звіт.

4. Звіт повинен мати назву.

9.15 Відповіді до контрольних запитань і вправ

1 (a)

```
DATA COLLEGE;
        INPUT ID AGE GENDER $ GPA CSCORE;
DATALINES;
1 18 M 3.7 650
2 18 F 2.0 490
3 19 F 3.3 580
4 23 M 2.8 530
5 21 M 3.5 640
;
```

1 (b)

```
PROC MEANS DATA=COLLEGE;
        VAR GPA CSCORE;
RUN;
```

1 (c)

```
DATA COLLEGE;
        INPUT ID AGE GENDER $ GPA CSCORE;
        INDEX = GPA + 3*CSCORE/500;
DATALINES;
```

```
1 18 M 3.7 650
2 18 F 2.0 490
3 19 F 3.3 580
4 23 M 2.8 530
5 21 M 3.5 640
;
PROC SORT DATA=COLLEGE;
        BY INDEX;
RUN;
PROC PRINT DATA=COLLEGE;
        TITLE "Students in Index Order";
        ID ID;
        VAR GPA CSCORE INDEX;
RUN;
```

2.

```
DATA sas_data;
INPUT SUBJ $1-3
          HEIGHT 4-5
          WT_INIT 6-8
          WT_FINAL 9-11;
BMI_INIT=(WT_INIT/2.2)/(HEIGHT*0.254);
BMI_FINAL=(WT_FINAL/2.2)/(HEIGHT*0.254);
BMI_DIFF=BMI_FINAL-BMI_INIT;
DATALINES;
00768155150
00272250240
00563240200
00170345298
;
PROC SORT data=sas_data out=sort_sas_data;
by SUBJ;
RUN;

PROC PRINT data=sort_sas_data;
VAR SUBJ HEIGHT BMI_INIT BMI_FINAL BMI_DIFF;
        TITLE 'TASK 2';
RUN;
```

3.

```
DATA sas_data;
INPUT SOCIAL_SECURITY_NUMBER ANNUAL_SALARY AGE RACE $;
TAX = 0.3*ANNUAL_SALARY;
DATALINES;
123874414 28000 35 W
```

```
646239182 29500 37 B
012437652 35100 40 W
018451357 26500 31 W
;
RUN;

PROC MEANS data= sas_data NOPRINT NWAY;
VAR ANNUAL_SALARY AGE;
OUTPUT OUT=AVERAGE
        MEAN=Average_Salary Average_Age;
RUN;

PROC PRINT DATA=AVERAGE;
VAR Average_Salary Average_Age;
TITLE 'Average Means';
RUN;

PROC SORT DATA=sas_data OUT=sort_sas_data;
BY SOCIAL_SECURITY_NUMBER;
RUN;

PROC PRINT DATA=sort_sas_data;
VAR ANNUAL_SALARY TAX;
TITLE 'Annual Salary and Tax';
RUN;

4.
DATA datas;
        INPUT ID 1-3
            IQ 4-6
                MATH 7-9
                SCIENCE 10-12;
if 0 LE IQ LE 100 THEN GROUP = 1;
ELSE IF 101 LE IQ LE 140 THEN GROUP = 2;
ELSE IF IQ GT 140 THEN GROUP = 3;
DATALINES;
001128550590
002102490501
003140670690
004115510510
;
PROC SORT data=datas out=sort_datas;
BY IQ;
RUN;
DATA GETOVERALL;
```

```
SET datas;
OVERALL=(IQ+MATH+SCIENCE)/500;
RUN;

DATA output;
MERGE sort_datas
        GETOVERALL(keep=OVERALL);
RUN;

PROC PRINT data=output;
      TITLE ";
RUN;

PROC FREQ data=output;
tables GROUP;
RUN;
```

5.

(a)

Строчка 2. Немає помилки, але можливо користувач бажає щоб ID було строкового типу.
Строчка 3. Ім'я змінної не може містити "-". Пропущений ";" в кінці строчки.
Строчка 7. ID – це ідентифікаційний код клієнта, для якого в загальному випадку середнє значення не обчислюють. Біль коректним є використання PROC FREQ по відношенню до категоріальних змінних, наприклад, REGION.
Строчка 8. Операнди PROC MEANS повинні знаходитися між службовим словом MEANS та ";". Також операнди повинні мати пробіл між собою, замість коми.

(b)

```
DATA MISTAKE;
INPUT ID 1-3 TOWN 4-6 REGION 7-9 YEAR 11-12 BUDGET 13-14
VOTER_TURNOUT 16-20;
DATALINES;
00104005  0422 12345
;
PROC MEANS DATA=MISTAKE;
VAR ID REGION VOTER_TURNOUT;
RUN;
```

6.

```
DATA datas;
      INPUT QUES1 $1
```

```
                QUES2 $2
                QUES3 $3
                QUES4 $4
                QUES5 $5;
DATALINES;
ABCDE
AACCE
BBBBB
CABDA
DDAAC
CABBB
EEBBB
ACACA
;

PROC FREQ data=datas;
tables QUES1 /NOCUM;
tables QUES2 /NOCUM;
tables QUES3 /NOCUM;
tables QUES4 /NOCUM;
tables QUES5 /NOCUM;
TITLE 'TASK 6';
RUN;
```

7. Набір даних повинен включати змінні: AGE, GENDER, RACE, INCOME, MARITAL та HOME (homeowner або renter).

Variable Name	Col(s)	Description and Formats
AGE	1	Вікові групи клієнтів: 1 = 10 to 19 2 = 20-29 3 = 30-39 4 = 40-49 5 = 50-59 6 = 60+
GENDER	2	Стать: 1 = male 2 = female
RACE	3	Раса: 1 = white 2 = African Am. 3 = Hispanic 4 = other
INCOME	4	Дохід: 1 = 0 to $9,999 2 = 10,000 to 19,999 3 = 20,000 to 39,000 4 = 40,000 to 59,000 5 = 60,000 to 79,000 6 = 80,000 and over

MARITAL	5	Сімейний стан: 1 = single 2 = married 3 = separated 4 = divorced 5 = widowed
HOME	6	Власник житла або орендатор: 1 = homeowner 0 = renter

```
DATA CEO;
    INPUT
 AGE     $ 1
        GENDER   $ 2
        RACE     $ 3
        INCOME   $ 4
        MARITAL  $ 5
        HOME     $ 6;
DATALINES;
311411
122310
411221
;
PROC FREQ DATA=CEO ORDER=FREQ;
  TITLE 'Frequencies and Contingency Tables for CEO Report';
  TABLES AGE GENDER RACE INCOME MARITAL HOME
  AGE*GENDER*RACE INCOME*AGE*GENDER MARITAL*HOME;
RUN;

PROC GCHART DATA=CEO;
  TITLE "Histograms";
  VBAR AGE GENDER RACE INCOME MARITAL HOME / DISCRETE;
RUN;
```

8.

```
data sas_data;
input EMPID SALARY JCLASS;
if JCLASS = 1 then BONUS = 0.1*SALARY;
if JCLASS = 2 then BONUS = 0.15*SALARY;
if JCLASS = 3 then BONUS = 0.2*SALARY;
NEW_SALARY=SALARY+BONUS;
Datalines;
137 28000 1
214 98000 3
199 150000 3
355 57000 2
Run;
```

```
PROC PRINT data=sas_data;
TITLE 'Employee salaries and job class';
Run;
```

9.

```
DATA PROB1_9;
  INPUT ID RACE $ SBP DBP HR;
DATALINES;
001   W   130   80   60
002   B   140   90   70
003   W   120   70   64
004   W   150   90   76
005   B   124   86   72
;

PROC SORT DATA=PROB1_9;
  BY SBP;
RUN;

PROC PRINT DATA=PROB1_9;
  TITLE "Race and Hemodynamic Variables";
  ID ID;
  VAR RACE SBP DBP;
RUN;
```

Перелік посилань

1. Бідюк, П. І. Прикладна статистика : (конспект лекцій) / П. І. Бідюк. – К. : НТУУ «КПІ», 2009. – 220 с.
2. Гаркавий, В. К. Математична статистика / В. К. Гаркавий, В. В. Ярова. – К. : Професіонал, 2004. – 384 с.
3. Бобик, О. І. Теорія ймовірностей і математична статистика / О. І. Бобик, Г. І. Берегова, Б. І. Копитко. – К. : Професіонал, 2007. – 560 с.
4. Кобзарь, А. И. Прикладная м атематическая статистика / А. И. Кобзарь. – М. : Физматлит, 2006. – 815 с.
5. Гмурман, В. Е. Теория вероятностей и прикладная статистика / В. Е. Гмурман. – М. : Высшая школа, 2002. – 479 с.
6. Хастингс, Н. Справочник по статистическим распределениям / Н. Хастингс, Дж. Пикок. – М. : Статистика, 1980. – 95 с.
7. Эфрон, Б. Нетрадиционные методы многомерного статистического анализа / Б. Эфрон. – М. : Финансы и статистика, 1988. – 263 с.
8. Harvey, A. C. Forecasting, structural time series models and the Kalman filter / A. C. Harvey. – Cambridge : Cambridge University Press, 1989. – 554 p.
9. Franses, Ph. H. Time series models for business and economic forecasting / Ph. H. Franses. – Cambridge : Cambridge University Press, 1998. – 280 p.
10. Tsay, R. S. Analysis of financial time series / R. S. Tsay. – Chicago : Wiley & Sons, Ltd. – 2010. – 715 p.
11. Bayesian Data Analysis / A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin. – New York : Chapman and Hall, CRC Press, 2000. – 670 p.
12. Rossi, P., Bayesian Statistics and Marketing / P. Rossi, G. M. Allenby. – Hoboken (New Jersey) : Wiley & Sons, Ltd., 2005. – 350 p.
13. Magic Quadrant for Business Intelligence Platforms // Technical Report G00225500, 6 Feb 2012. – NY : Gartner Inc. – 27 p.
14. Magic Quadrant for Business Intelligence Platforms / Kurt Schlegel, Rita L. Sallam, Daniel Yuen, Joao Tapadinhas // Technical Report G00239854, 5 Feb 2013. – NY : Gartner Inc. – 64 p.
15. Constable, N. SAS programming for Enterprise Guide Users / N. Constable. – Cary, NC : SAS Institute Inc, 2010. – 292 p.
16. Cody, R. Learning SAS by Example : A Programmer's Guide / R. Cody. – Cary, NC : SAS Institute Inc, 2007. – 665 p.
17. Анализ временных рядов: прогноз и управление [Текст] / Дж. Бокс, Г. Дженкинс; Пер. с англ. А. Л. Левшин, Ред. В. Ф. Писаренко. - М. : Мир, 1974 - 1974. Вып. 1. - 1974. - 406 с.
18. Ивахненко, А. Г. Долгосрочное прогнозирование и управление сложными системами [Текст] / А. Г. Ивахненко. - К. : Техніка, 1975. - 311 с.
19. О компании SAS [Електронний ресурс] : офіційний web-сайт. – Режим доступу http://www.sas.com/offices/europe/russia/company/index.html

Примітка: біля 90% всіх наукових публікацій у світі – англомовні, сучасна література з прикладної статистики також англомовна.

Навчальне видання

Навчальний посібник

П. І. Бідюк, О. М. Терентьєв, Т. І. Просянкіна-Жарова

ПРИКЛАДНА СТАТИСТИКА

Здано в набір 07.10.2013. Підписано до друку 14.11.2013.
Формат 60х84/16. Папір офсетний. Друк різографічний
Ум. друк. арк. 17.67. Гарнітура Times New Roman.
Замовлення № 67304
Наклад 300 екземплярів.

Віддруковано ПП «ТД «Едельвейс і К»
М.Вінниця, вул. 600-річчя, 17
Тел.: (0432) 550-333
Свідоцтво про внесення до Державного реєстру
ДК №3736