МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Харківський національний педагогічний університет

імені Г.С.Сковороди

Пономарьова Н.О., Голиш І. Є.

ОСНОВИ ОПРАЦЮВАННЯ СТАТИСТИЧНИХ ДАНИХ

Навчальний посібник
для здобувачів першого (бакалаврського) рівня вищої освіти
за спеціальністю 014 Середня освіта (Інформатика)

Харків-2025

УДК 006.91:519.22 (075.8)

Рецензенти:

Олефіренко Н.В. – доктор педагогічних наук, професор, завідувач кафедри інформатики Харківського національного педагогічного університету імені Г. С. Сковороди.

Котова О.М. – викладач вищої категорії, голова циклової комісії природничо-математичних дисциплін Харківського політехнічного фахового коледжу.

Затверджено вченою радою Харківського національного педагогічного університету імені Г. С. Сковороди

протокол № 12 від 05.12.2024 р.

Пономарьова Н.О., Голиш І.Є. Основи опрацювання статистичних даних. Навчальний посібник. – Харків, 2025. – 171 с. рис. 56, табл. 22.

Навчальний посібник присвячений розкриттю особливостей навчання основ статистики майбутніми учителями інформатики в закладах вищої педагогічної освіти. У посібнику наведено навчально-методичні матеріали для лабораторного практикуму з опрацювання статистичних даних для майбутніх учителів інформатики. Посібник призначений для здобувачів першого (бакалаврського) рівня вищої освіти за спеціальністю 014 Середня освіта (Інформатика), а також для науково-педагогічних працівників закладів вищої педагогічної освіти, учителів інформатики.

Видано за рахунок авторів

ЗМІСТ

ВСТУП

Професійна діяльність вимагає від сучасного вчителя здійснення наукових розвідок та володіння методами математичної статистики й активного їх прикладного застосування. Учитель має вирішувати важливі завдання педагогічної практики, як-от отримання науково обґрунтованих висновків, демонстрація достовірності фактичного матеріалу та забезпечення об'єктивності експериментальних даних.

Учитель повинен бути здатним виконувати педагогічні дослідження, що включають експериментальну перевірку гіпотез та підтвердження або спростування теоретичних припущень.

Математична статистика забезпечує науковців інструментами для аналізу даних і валідації гіпотез, що особливо важливо при проведенні педагогічних експериментів. Це дозволяє здійснювати аналіз результатів досліджень, виявляти закономірності та формулювати обґрунтовані висновки.

Підготовка майбутніх учителів у цьому контексті є важливим завданням вищої педагогічної освіти, особливо для учителів інформатики. Вони вже мають достатній рівень цифрової компетентності, але потребують розвитку практичних навичок застосування програмного забезпечення для вирішення прикладних задач, зокрема, обробки експериментальних даних.

Навчальний посібник присвячений розкриттю особливостей навчання основ статистики майбутніми учителями інформатики в закладах вищої педагогічної освіти. У посібнику наведено навчально-методичні матеріали для лабораторного практикуму з опрацювання статистичних даних для майбутніх учителів інформатики.

Систематичне виконання завдань лабораторного практикуму допоможе майбутньому вчителю інформатики отримати належні знання, уміння та навички роботи з методами математичної статистики, а також сформує мотиваційні засади до їх застосування у професійній діяльності.

РОЗДІЛ 1. ОСОБЛИВОСТІ ВИВЧЕННЯ ОПРАЦЮВАННЯ СТАТИСТИЧНИХ ДАНИХ МАЙБУТНІМИ ВЧИТЕЛЯМИ ІНФОРМАТИКИ

Підготовка майбутніх вчителів інформатики в системі вищої педагогічної освіти України забезпечують класичні та педагогічні університети. Освітні програми як для бакалаврського, так і для магістерського рівнів вищої освіти, які розроблені в закладах вищої педагогічної освіти, містять низку складових серед яких дисципліни загальної підготовки, дисципліни професійної підготовки, різні види практик, різноманітні форми атестацій.

До переліку обов’язкових освітніх компонентів підготовки майбутніх учителів інформатики, які мають професійну інформатичну спрямованість, у переважної більшості закладів вищої освіти віднесено низку дисциплін з основ вищої математики (алгебра, геометрія, математичний аналіз, дискретна математика тощо) та з основ комп’ютерних наук (теоретичні основи інформатики та інформаційні технології, архітектура комп’ютера, комп’ютерні мережі та веб-дизайн, основи програмування, інформаційні системи, комп’ютерна графіка тощо).

Особливе місце у цьому переліку належить дисциплінам, пов’язаним з опанування майбутніми вчителями інформатики основ комп’ютерних технології статистичного опрацювання експериментальних даних.

Сучасний учитель повинен уміти виконувати педагогічні дослідження, що нерозривно пов'язане з експериментальною перевіркою гіпотез шляхом доведення або спростування теоретичних припущень. Це вимагає від учителя вирішення ряду важливих завдань, таких як отримання науково обґрунтованих висновків, демонстрація достовірності фактичного матеріалу та забезпечення об'єктивності експериментальних даних. З метою ефективного вирішення цих завдань учитель повинен вміти доцільно використовувати математичні методи обробки експериментальних даних у поєднанні з сучасними цифровими інструментами, що сприятиме швидкій обробці і аналізу даних, автоматизації обчислень та уникненню потенційних помилок, а також сприятиме формуванню його цифрової компетентності [36]

Наведемо результати систематизації досвіду підготовки майбутніх учителів інформатики основам статистичної обробки у вітчизняних закладах вищої освіти, а саме:

- Бердянський державний педагогічний університет;
- Донбаський державний педагогічний університет;
- Дрогобицький державний педагогічний університет імені Івана Франка;
- Житомирський державний університет імені Івана Франка;
- Закарпатський угорський інституті ім. Ференца Ракоці ІІ;
- Запорізький національний університеті;
- Кременецька обласна гуманітарно-педагогічна академія ім. Тараса Шевченка;
- Кременчуцький національний університет імені Михайла Остроградського;
- Харківський національний педагогічний університет імені Г.С. Сковороди;
- Черкаський національний університет імені Богдана Хмельницького.

Так, у Бердянському державному педагогічному університеті для здобувачів магістерського рівня професійної освіти запропонована дисципліна «Комп’ютерні технології статистичного опрацювання експериментальних даних». Ця дисципліна входить до навчального плану як обов’язкова компонента професійної підготовки, має обсяг 3 кредити, форма контролю – екзамен [38].

У Донбаському державному педагогічному університеті для здобувачів бакалаврського рівня вищої освіти запропонована дисципліна «Теорія ймовірностей та математична статистика». Ця дисципліна є обов’язковою компонентою професійної підготовки, складає 5 кредитів, форма контролю – екзамен [43].

У Дрогобицькому державному педагогічному університеті імені Івана Франка запропонована дисципліна «Теорія ймовірностей та математична статистика» для здобувачів бакалаврського рівня вищої освіти. Ця дисципліна входить до навчального плану як обов'язкова компонента професійної підготовки, складає 4 кредити, форма контролю – залік. Також для здобувачів бакалаврського рівня вищої освіти тут пропонується споріднена за тематикою навчальна дисципліна «Інформаційні технології». Ця дисципліна - обов'язкова компонента професійної підготовки, складає 6 кредити, форма контролю – екзамен [45].

В Закарпатському угорському інституті ім. Ференца Ракоці ІІ запропонована дисципліна «Теорія ймовірностей та математична статистика» для здобувачів бакалаврського рівня вищої освіти. Ця дисципліна входить до навчального плану як обов'язкова компонента професійної підготовки, має обсяг 5 кредити, форма контролю – екзамен [50].

В Запорізькому національному університеті запропонована дисципліна «Теорія ймовірностей та математична статистика» для здобувачів бакалаврського рівня вищої освіти. Ця дисципліна є обов'язковою компонентою професійної підготовки, складає 4 кредити, форма контролю – екзамен [20].

В Кременецькій обласній гуманітарно-педагогічній академії ім. Тараса Шевченка запропонована дисципліна «Математична статистика» для здобувачів бакалаврського рівня вищої освіти. Ця дисципліна входить до навчального плану як обов'язкова компонента професійної підготовки, складає 3 кредити, форма контролю – екзамен [32].

В Кременчуцькому національному університеті імені Михайла Остроградського запропонована дисципліна «Ймовірнісно-статистичні методи інформаційних технологій» для здобувачів бакалаврського рівня вищої освіти. Ця дисципліна входить до навчального плану як обов'язкова компонента професійної підготовки, складає 5 кредити, форма контролю – залік. Також для здобувачів бакалаврського рівня вищої освіти передбачена і навчальна дисципліна «Теорія ймовірностей та математична статистика». Ця дисципліна є обов'язковою компонентою професійної підготовки, складає 5 кредити, форма контролю – залік [29, 41].

В Харківському національному педагогічному університеті імені Г.С. Сковороди запропонована дисципліна «Теорія ймовірностей та математична статистика» для здобувачів бакалаврського рівня вищої освіти. Ця дисципліна до навчального плану входить як обов'язкова компонента професійної підготовки, складає 3 кредити, форма контролю – залік [26].

В Міжнародному економіко-гуманітарному університеті імені академіка Степана Дем'янчука для здобувачів бакалаврського рівня вищої освіти запропонована дисципліна «Теорія ймовірностей та математична статистика». Ця дисципліна входить до навчального плану як обов'язкова компонента професійної підготовки, складає 5 кредити, форма контролю – екзамен [42].

В Черкаський національний університет імені Богдана Хмельницького для здобувачів бакалаврського рівня вищої освіти передбачена дисципліна «Теорія ймовірностей та математична статистика». Ця дисципліна входить до навчального плану як обов'язкова компонента професійної підготовки, складає 3 кредити, форма контролю – залік [27].

Для здобувачів бакалаврського рівня вищої освіти у Житомирському державному університеті імені Івана Франка підготовлена дисципліна «Теоретичні основи інформатики та інформаційно-комунікаційні технології», яка є обов'язковою компонентою професійної підготовки, складає 9 кредити, форма контролю – екзамен [44].

Таким чином, в усіх закладах вищої освіти основи математичної статистики - обов'язкова складова бакалаврської підготовки майбутніх учителів інформатики. Це засвідчує вагому роль цієї освітньої компоненти для формування професійної компетентності учителя інформатики.

У таблиці 1.1. наведено підсумки проведеного аналізу освітніх програм з підготовки майбутніх учителів інформатики у закладах вищої освіти України.

Таблиця 1.1

Аналіз освітніх програм підготовки майбутніх учителів інформатики щодо вивчення основ опрацювання статистичних даних

Назва ЗВО	Назва спеціальності. Освітня програма. Кваліфікація	Рік	Назва дисципліни	К-ть кредитів, вид кон-тролю	Блок
1	2	3	4	5	6
Бердянський державний педагогічний університет	Професійна освіта. Комп'ютерні технології. Магістр професійної освіти.	2023	Комп'ютерні технології статистичного опрацювання експериментальних даних	3 Екзамен	Професійна підготовка. Обов'язкова компонента
Донбаський державний педагогічний університет	Середня освіта (Інформатика). Бакалавр. Середня освіта «Інформатика»	2023	Теорія ймовірностей та математична статистика	5 Екзамен	Професійна підготовка. Обов'язкова компонента
Дрогобицький державний педагогічний університет імені Івана Франка	Середня освіта (Інформатика). Бакалавр. Середня освіта (Інформатика)	2023	Теорія ймовірностей та математична статистика	4 Залік	Професійна підготовка. Обов'язкова компонента
Дрогобицький державний педагогічний університет імені Івана Франка	Середня освіта (Інформатика). перший (бакалаврський) рівень вищої освіти	2023	Інформаційні технології	6 Екзамен	Професійна підготовка. Обов'язкова компонента
Житомирський державний університет імені Івана Франка	Середня освіта (Інформатика). Бакалавр. Середня освіта (Інформатика)	2023	Теоретичні основи інформатики та інформаційно-комунікаційні технології	9 Екзамен	Професійна підготовка. Обов'язкова компонента
Закарпатський угорський інститут ім. Ференца Ракоці II	Середня освіта (Інформатика). Бакалавр. Середня освіта (Інформатика)	2023	Теорія ймовірностей та математична статистика	5 Екзамен	Професійна підготовка. Обов'язкова компонента

Продовження табл. 1.1.

1	2	3	4	5	6
Запорізький національний університет	Середня освіта (Інформатика). Бакалавр. Середня освіта (Інформатика)	2023	Теорія ймовірності та математична статистика	4 Екзамен	Професійна підготовка. Обов'язкова компонента
Кременецька обласна гуманітарно-педагогічна академія ім. Тараса Шевченка	Середня освіта (Інформатика) Бакалавр середньої освіти. Вчитель інформатики	2023	Математична статистика	3 Екзамен	Професійна підготовка. Обов'язкова компонента
Кременчуцький національний університет імені Михайла Остроградського	Інформаційні технології. Комп'ютерні науки. Бакалавр з комп'ютерних наук	2023	Ймовірнісно-статистичні методи інформаційних технологій	5 Залік	Професійна підготовка. Обов'язкова компонента
Кременчуцький національний університет імені Михайла Остроградського	Середня освіта (Інформатика) Бакалавр середньої освіти. Вчитель інформатики	2023	Теорія ймовірностей та математична статистика	5 Залік	Загальна підготовка. Обов'язкова компонента
Харківський національний педагогічний університет імені Г.С. Сковороди	Середня освіта (Інформатика). Інформатика в закладах освіти	2023	Теорія ймовірностей та математична статистика	3 Залік	Професійна підготовка. Обов'язкова компонента
"Міжнародний економіко-гуманітарний університет імені академіка Степана Дем'янчука"	Середня освіта (Інформатика та математика) Бакалавр середньої освіти. (Інформатика)	2023	Теорія ймовірностей та математична статистика	5 Екзамен	Професійна підготовка. Обов'язкова компонента
Черкаський національний університет імені Богдана Хмельницького	Середня освіта (Інформатика) Бакалавр освіти. Вчитель інформатики	2022	Теорія ймовірностей та математична статистика	3 Залік	Професійна підготовка. Обов'язкова компонента

На особливу увагу заслуговує аналіз навчальних програм дисциплін з основ математичної статистики в таких закладах освіти, як Закарпатський угорський інститут ім. Ференца Ракоці II, Дрогобицький державний педагогічний університет імені Івана Франка, Житомирський державний університет імені Івана Франка, Харківський національний педагогічний університет імені Г.С. Сковороди.

Так, наприклад у Закарпатському угорському інституті ім. Ференца Ракоці II навчальна програма дисципліни «Теорія ймовірностей та математична статистика» серед двадцяти однієї теми курсу включає вісім тем, пов'язаних із статистичною обробкою експериментальних даних, а саме:

- Вибірка. Емпіричні початкові і центральні моменти, асиметрія та ексцес.
- Точкові статистичні оцінки. Інтервальні статистичні оцінки.
- Статистична гіпотеза. Перевірка правдивості статистичних гіпотез про рівність двох генеральних середніх та двох дисперсій, ознаки яких мають нормальні закони розподілу.
- Перевірка правдивості нульової гіпотези нормального закону розподілу ознаки генеральної сукупності.
- Емпіричні та теоретичні частоти.
- Функціональна, статистична і кореляційна залежно
- Множинна регресія, визначення статистичних оцінок для параметрів лінійної множинної функції регресії. Нелінійна регресія.

Водночас, освітня програма не передбачає наявності навчальної дисципліни з цілеспрямованої підготовки майбутніх учителів до застосування комп'ютерних технологій статистичної обробки експериментальних даних.

У Дрогобицькому державному педагогічному університеті імені Івана Франка навчальна програма дисципліни «Теорія ймовірностей та математична статистика» включає три розділи «Випадкові події», «Випадкові величини», «Основи математичної статистики». Вивчення основ математичної статистики включає теми, пов'язані із статистичними розподілами вибірок та їх числовими характеристиками, статистичними оцінками параметрів генеральної сукупності, статистичними гіпотезами [49]. Одночасно, у цьому ж університеті навчальна дисципліна «Інформаційні технології» включає розділ, дотичний до вивчення присвячений застосуванню табличного процесора Microsoft Excel для обробки статистичних даних. Вивчення цього розділу передбачає, у тому числі, опанування основ візуалізації результатів навчання та опрацювання даних статистичного аналізу засобами табличного процесора Microsoft Excel [41].

У Житомирському державному університеті імені Івана Франка навчальна програма дисципліна «Теоретичні основи інформатики та інформаційно-комунікаційні технології» хоча і включає такий модуль як «Обробка інформації засобами програми Microsoft Excel», але зміст модулів не спрямований на опанування технологій статистичної обробки експериментальних даних.

В Харківському національному педагогічному університеті імені Г.С. Сковороди бакалаврська підготовка майбутніх учителів інформатики включає навчальну дисципліну «Теорія ймовірностей та математична статистика». Програма навчальної дисципліни включає розділ «Елементи математичної статистики», який містить наступні теми:

- Вибірковий метод.

Предмет математичної статистики. Історія розвитку, сучасний стан дисципліни. Застосування. Генеральна сукупність і вибірка. Статистичні ознаки. Вимоги, яким повинна задовольняти вибірка Табличне подання експериментальних даних. Варіаційні ряди. Гістограма, полігон частот. Характеристики положення (середнє арифметичне, мода, медіана). Дисперсія та середнє квадратичне відхилення.

- Статистичні оцінки параметрів розподілу.

Оцінки параметрів генеральної сукупності за даними вибірки. Точкові та інтервальні оцінки. Визначення необхідного об’єму вибірки для отримання оцінок заданої точності.

- Статистична перевірка статистичних гіпотез.

Статистичні критерії і перевірка статистичних гіпотез. Нуль-гіпотеза і альтернативна. Помилки першого та другого роду при перевірці гіпотез. Односторонні та двосторонні критерії Критерії, що ґрунтуються на нормальному розподілі: порівняння вибіркового середнього з генеральним середнім;

порівняння вибіркових середніх для незалежних вибірок. Перевірка відповідності нормальному розподілу. Критерій «хі–квадрат». Непараметричні методи. [25]

Майбутнім учителям інформатики (бакалаврський рівень) тут пропонується також і вибіркова навчальна дисципліна «Поглиблений MS Excel. Основи статистики в Excel», програма якої передбачає окремий модуль «Статистична обробка даних». Зміст цього модулю передбачає вивчення тем:

- Статистична обробка експериментальних даних (Основні статистичні параметри. Діаграма розподілення. Нормальний закон розподілу. Перевірка на істотність змін. Перевірка на сталість/наявність тенденцій. Кореляційний аналіз. Кластерний аналіз).

- Регресійний аналіз (Математичний апарат кореляційного і регресійного аналізу; Аналіз можливостей Excel для проведення кореляційного і регресійного аналізу. Надійність та похибки регресійного аналізу) [34].

Таким чином можна зробити висновок, що хоча основи статистичної обробки експериментальних даних і входять до змісту професійної підготовки майбутніх учителів, сучасним комп’ютерним технологіям їх реалізації у інваріантній складовій професійної підготовки приділяється недостатня увага.

Як засвідчив аналіз методичних та спеціальних джерел, навчальний матеріал з опрацювання статистичних даних для майбутніх учителів інформатики для першого (бакалаврського) рівня вищої освіти має включати наступні теми та їх складові:

- *основні поняття описової статистики* (Завдання описової статистики. Методи описової статистики. Типи шкал і змінних в описовій статистиці. Кількісні та категоріальні дані. Безперервні і дискретні змінні. Інтервальна шкала. Шкала співвідношення);

- *випадкові числа, генерація випадкових чисел* (Загальне поняття випадкового числа і розподілу випадкових чисел. Сутність методу Монте-Карло. Інструменти для генерації випадкових чисел);

- *вибірка, статистичні оцінки параметрів вибірки (*Поняття генеральної сукупності. Вибірка та особливості її формування. Побудова варіаційного ряду. Емпірична функція розподілу. Багатокутник частот);

- *обробка вибірок даних* (Параметри вибірки та особливості їх розрахунку. Середнє арифметичне. Дисперсія. Стандартне відхилення. Медіана. Мода. Емпіричний момент. Коефіцієнт асиметрії. Ексцес);

- *основи кореляційного аналізу* (Поняття про кореляцію. Характеристика кореляційних зв’язків. Напрямок. Потужність. Ймовірність. Форма. Параметри кореляції та їх розрахунок);

- *метод вибірки* (Поняття похибки вибірки. Помилка параметра. Помилка середнього арифметичного. Похибка середньоквадратичного відхилення Помилка коефіцієнту варіації. Помилка індикатора асиметрії. Помилка коефіцієнта ексцесу. Критерії надійності. Довірчі інтервали);

- *статистичні гіпотези та їх перевірка* (Поняття статистичної гіпотези. Алгоритм перевірки статистичної гіпотези. Помилки перевірки гіпотези. Статистичні критерії та їх види. Критерій Стьюдента. Критерій Пірсона);

- *побудова регресійних моделей* (Поняття регресійної моделі. Алгоритм побудови та інтерпретації регресійної моделі).

РОЗДІЛ 2. ТЕОРЕТИЧНІ ЗАСАДИ ОПРАЦЮВАННЯ СТАТИСТИЧНИХ ДАНИХ

ОСНОВНІ ПОНЯТТЯ ОПИСОВОЇ СТАТИСТИКИ

Завдання описової статистики.

Описова статистика або дескриптивна статистика (англ. descriptive statistics) – розділ статистики, який займається обробкою емпіричних даних, їх систематизацією, наочним представленням у вигляді графіків та таблиць, а також їх кількісним описом через основні статистичні показники [33].

Описова статистика забезпечує короткий аналітичний підсумок про вибірку та про результати спостереження, які були зроблені.

У психолого-педагогічних дослідженнях описова статистика використовується для аналізу результатів експериментальної діяльності - для відбору, класифікації, упорядкування, порівняння різноманітних кількісних показників, планування та достовірного прогнозування експериментальних досліджень [30,38].

Методи описової статистики.

Описова статистика використовує три основних методи представлення даних:

1. Табличне представлення даних.
2. Графічне представлення даних.
3. Розрахунок статистичних показників

Основними способами представлення емпіричних розподілів є *табличний, та графічний.*

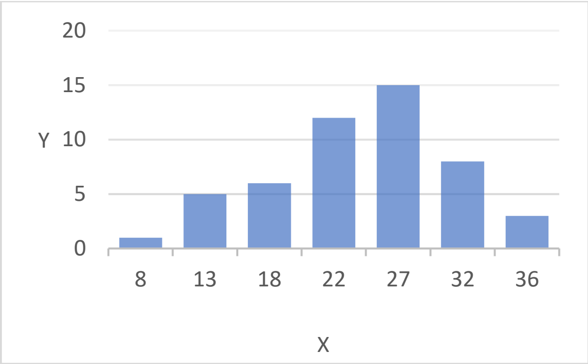
1. Табличний спосіб представлення розподілів продемонстровано на рис. 1.1.

	A	B	C	D	E	F	G	H
1	Результати тестування *IQ* (80 осіб)							
2	120	104	102	96	121	97	106	93
3	83	115	109	119	96	114	91	92
4	95	112	104	116	85	106	89	102
5	111	85	113	97	115	105	90	94
6	92	95	118	104	94	97	109	99
7	117	97	80	99	86	96	112	102
8	93	124	98	106	137	93	100	113
9	120	112	89	78	83	92	72	97
10	79	80	80	83	87	93	84	87
11	103	100	107	90	88	105	93	105
12	Мінімальне IQ_{min}		72		Максимальне IQ_{max}		137	
13	K =	8	λ =	8,125	Частоти			
14	Діапазон значень *IQ*				диференціальні		інтегральні	
15					абсолютн	відносна	абсолютн	відносна
16	i	$IQ_{поч}$	$< IQ_i \le$	$IQ_{кін}$	m_i	f_i	m_j	F_j
17	1	70	$< IQ_1 \le$	80	6	0,0750	6	0,0750
18	2	80	$< IQ_2 \le$	90	14	0,1750	20	0,2500
19	3	90	$< IQ_3 \le$	100	26	0,3250	46	0,5750
20	4	100	$< IQ_4 \le$	110	16	0,2000	62	0,7750
21	5	110	$< IQ_5 \le$	120	15	0,1875	77	0,9625
22	6	120	$< IQ_6 \le$	130	2	0,0250	79	0,9875
23	7	130	$< IQ_7 \le$	140	1	0,0125	80	1,0000
24	Суми:				80	1		

Рис. 1.1. Результати розрахунку розподілу результатів тестування *IQ*

По-різному називають такі таблиці: *таблицею емпіричних частот* або *табличною формою* представлення *розподілу.* Табличний спосіб є основним розрахунковим методом і передумовою його графічної форми. Разом вони дають цілісне уявлення щодо властивостей вибірки.

2. Графічний спосіб представлення – це відображення розподілу графічними засобами, серед яких найпоширенішими є *гістограма і полігон.* На рис. 1.2 показано диференціальний відносний розподіл незгрупованих частот у двох варіантах.



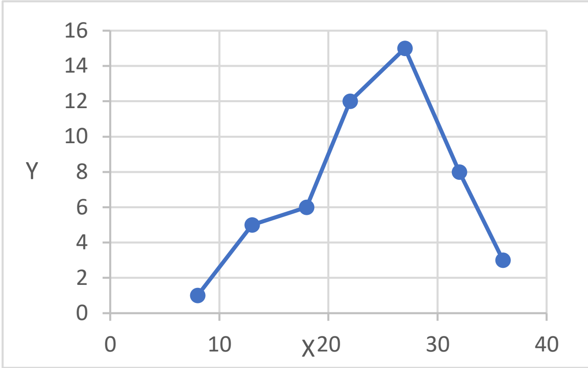


Рис. 1.2. Гістограма і полігон частот

Комбіновані способи представлення, які об'єднують у межах однієї графічної форми різні типи розподілу (диференціальний та інтегральний), можна побачити на рис. 1.3.

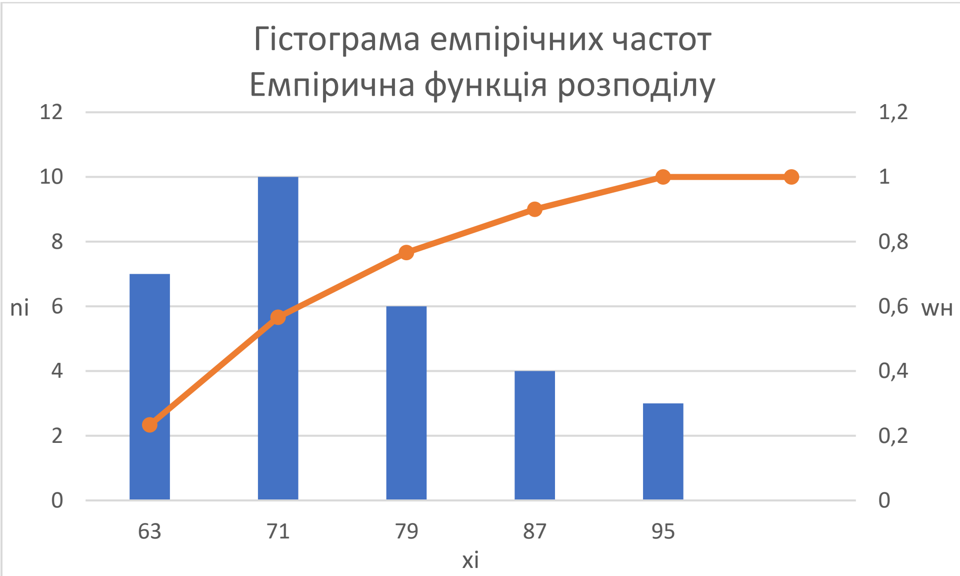


Рис. 1.3. Графік розподілу [38]

Статистичні показники, що використовуються в описовій статистиці:

Середнє арифметичне (mean) – середнє значення набору даних.

Медіана (median) – середнє значення, яке ділить набір даних на дві рівні частини.

Мода (mode) – найбільш часто зустрічається значення в наборі даних.

Дисперсія (variance) – міра розсіювання даних навколо середнього значення.

Стандартне відхилення (standard deviation) – корінь квадратний з дисперсії, що показує середнє відхилення значень від середнього.

Вимірювання варіації (measure of variation) полягає в знаходженні чисел, які характеризують ступінь розкиду даних щодо центру розподілу.

Розмах (range) – різниця між найбільшим і найменшим значенням у наборі даних.

Квартилі (quartile) – значення, що ділять набір даних на чотири рівні частини.

Квартильний розмах (snter quartile range – IQR) – різниця між третім і першим квартилями.

Інтерквартильний розмах (snterquartile range, IQR) – різниця між третім (Q3) і першим (Q1) квартилями.

Асиметрія (skewness) – міра асиметрії розподілу даних.

Ексцес (kurtosis) – міра "гостроти" або "плоскості" розподілу даних у порівнянні з нормальним розподілом [24].

Типи шкал і змінних в описовій статистиці.

Змінна — будь-яка характеристика об’єкта, що вимірюється чи досліджується. Змінна може набувати різних значень для різних об’єктів. Для змінних виділяють такі типи шкал: номінальна та порядкова, а також метричні.

Номінальна шкала, також відома як *шкала найменувань* (лат. nomen — ім’я), — передбачає фіксацію якісних (не кількісних) ознак об’єкта. Основним інформативним параметром таких об’єктів є їх кількість, яку визначають за допомогою лічби. Значення не можна порівнювати між собою. Наприклад, не можна вважати один колір або професію кращий чи гірший за інший. Шкали найменувань використовують, вимірюючи такі об’єктивні ознаки, як кольори, сімейний стан, вік, стать, стаж роботи, кваліфікацію тощо, а також суб’єктивне ставлення респондентів до певних аспектів будь-якого явища, процесу: мотиви поведінки, судження і твердження з різних питань, ціннісні орієнтації, інтереси тощо. Для статистичного аналізу даних використовують лише дослідження частоти повторення певного значення вимірюваної величини.

Порядкова (рангова) шкала (англ. — ordinal quantity) — застосовується, коли вимірювані значення можна порівнювати між собою, наприклад, розташувати за зростанням, але не можна застосувати одиницю вимірювання. Розглядаються поняття більше, дорівнює, менше. Наприклад, розподіл місць на змаганнях, класифікація учнів за рівнями досягнень (низький, середній, достатній, високий). Для статистичного аналізу даних використовують не лише дослідження, які застосовуються для шкали найменувань, але й є можливість визначати міру зв'язку між різними властивостями (кореляція рангів) та здійснювати порівняння за рівнем прояву певної ознаки.

Шкала інтервалів, або *інтервальна шкала* (англ. — interval scales), — характеризується наявністю одиниці вимірювання, але точка відліку є умовною, тобто немає природно обумовленого нульового значення. Розглядаються поняття «на скільки менше», «дорівнює», «на скільки більше». Віднесення даних до метричної шкали дає можливість обчислювати числові параметри, такі як середнє значення, дисперсія, визначати близькість емпіричного розподілу до розподілів, що відомі у математичній теорії, та застосовувати найбільш потужні методи статистичного аналізу.

Шкала відношень (англ. — ratio scales), або *пропорцій,* — передбачає наявність одиниці вимірювання та природного початку відліку. Наприклад, час виконання завдання, кількість виконаних завдань. З даними припустимі всі математичні операції [32].

<u>Кількісні та категоріальні дані.</u>

Кількісні (числові) дані – це показники, які беруть числові значення, які виходять шляхом деяких вимірювань або підрахунків.

Категоріальні (іменовані) дані – представляють собою категорії або назви, які використовуються для класифікації об'єктів або ознак [1].

<u>Безперервні і дискретні змінні.</u>

Кількісні дані поділяються на дискретні (кількість дітей у сім'ї, кількість медалей у команди) та безперервні (ріст, вага, заробітна плата)

Безперервна – коли змінна може приймати будь-яке значення в деякому діапазоні значень.

Дискретні змінні - коли змінна приймає на себе зліченну кількість значень [17].

Номінальні та порядкові дані.

Категоріальні дані поділяються на номінальні (невпорядковані) (імена, назви міст, група крові), порядкові (ступені освіти, рейтинг популярності) та бінарні (так, ні; 0,1) [33].

Номінальні дані – різновид категоріальних даних, коли немає природного впорядкування серед категорій.

Порядкові дані – категоріальні дані, для яких існує природний порядок серед категорій. Приклад - шкали ранжирування або літерні оцінки [17].

ВИПАДКОВІ ЧИСЛА, ГЕНЕРАЦІЯ ВИПАДКОВИХ ЧИСЕЛ

Загальне поняття випадкового числа.

Якщо результат досліду описується не однією, а декількома випадковими величинами, вони утворюють систему випадкових величин. Системою n випадкових величин (n-вимірною випадковою величиною, або випадковим вектором) називають впорядкований набір n випадкових величин (X_1, X_2, …, X_n) [46].

Дискретною випадковою величиною називають величину, можливі значення якої можна занумеровати. Кількість можливих значень дискретної випадкової величини скінчена або нескінчена.

Розподіл випадкових чисел

Законом розподілу будь-якої дискретної величини називається співвідношення, що визначає залежність між значеннями випадкової величини та ймовірностями, з якими ці значення набуваються. Закон розподілу можна задати у вигляді таблиці, аналітично (у вигляді формули) та графічно.

Найчастіше закон розподілу дискретної випадкової величини подають у вигляді таблиці розподілу або ряда розподілу. У першому рядку таблиці записуються можливі значення x_1 ,x_2 ,x_3 ,..., x_n, а у другому рядку відповідні ймовірності p_1 , p_2 , p_3 ,..., p_n .

Сутність методу Монте-Карло.

Метод статистичного моделювання, відомий під назвою «метод Монте-Карло», — це числовий метод розв'язування математичних задач за допомогою моделювання випадкових величин.

Перша особливість цього методу — простота алгоритму. Як правило, складають програму для одного випробування. Потім цей дослід повторюють N разів, при цьому кожен дослід не залежить від інших. За результатами дослідів визначають середні величини, тому цей метод одержав іншу назву — «метод статистичних випробувань» [22].

Не існує єдиного методу Монте-Карло, цей термін описує великий і широко використовуваний клас підходів. Проте ці підходи використовують в своїй основі єдиний шаблон:

1. Визначити область можливих вхідних даних.

2. Випадковим чином згенерувати вхідні дані із визначеної вище області за допомогою деякого заданого розподілу ймовірностей.

3. Виконати детерміновані обчислення над вхідними даними.

4. Проміжні результати окремих розрахунків звести у кінцевий результат [20].

Таблиця 1.2

Інструменти для генерації випадкових чисел в Excel [16].

Формула	Опис
= RAND () = СЛЧИС ()	Генерація випадкових десяткових чисел від 0 до 1.
=RAND()*N = СЛЧИС ()*N	Створення випадкових десяткових чисел від 0 до N.
=RAND()*(BA)+A = СЛЧИС ()*(BA)+A	Генерація випадкових десяткових чисел між будь-якими двома вказаними числами. (A є нижньою межею та B це верхня межа значення.)
=INT(RAND()*N) =ЦЕЛОЕ (СЛЧИС ()*N)	Генерація випадкових цілих чисел від 0 до N.
=INT(RAND()*(BA)+A) =ЦЕЛОЕ (СЛЧИС()*BA)+A)	Генерація випадкових цілих чисел між будь-якими двома вказаними числами. (A є нижньою межею та B це верхня межа значення.)
=RANDBETWEEN(bottom, top)	Отримання найменшого та найбільшого числа діапазону випадкових чисел
=RandomNumbers(50,200,0)	Генерація випадкових цілих чисел (X, Y, Z)

ВИБІРКА, СТАТИСТИЧНІ ОЦІНКИ ПАРАМЕТРІВ ВИБІРКИ

<u>Поняття генеральної сукупності.</u>

Генеральна сукупність — сукупність об’єктів, підданих статистичному аналізу. Обсяг генеральної сукупності — це кількість об’єктів у ній. Об’єкти генеральної сукупності можуть мати одну (декілька) кількісних або якісних ознак [27].

Термін «генеральна сукупність» використовується тоді, коли йдеться про велику, але кінцеву сукупність досліджуваних об’єктів. Генеральні сукупності можуть стягати значних обсягів, бути скінченими і нескінченими. На практиці мають справу зі скінченими сукупностями. Якщо відношення обсягу генеральної сукупності до обсягу вибірки складає більш, ніж 100, то, за словами Гласса і Стенлі методи оцінювання для скінчених і нескінчених сукупностей дають у сутності однакові результати. Генеральною сукупністю можна назвати і повну сукупність значень якоїсь ознаки. Приналежність вибірки до генеральної сукупності є головною підставою для оцінки характеристик генеральної сукупності за характеристиками вибірки [38].

Вибірка та особливості її формування.

Вибіркова сукупність (вибірка) — сукупність випадково відібраних об'єктів. Репрезентативна (представницька) вибірка — вибірка, до якої з однаковою ймовірністю може потрапити будь-який елемент генеральної сукупності. Результати репрезентативної вибірки переносять на весь обсяг генеральної сукупності [27].

Згідно з вибірковим підходом використання математико-статистичних методів може проводитись у такій послідовності:

- із генеральної сукупності, властивості якої підлягають дослідженню, певними методами формують вибірку – типову але обмежену кількість об'єктів, до яких застосовують дослідницькі методи;
- в результаті спостережень, експериментальних дій і вимірювань над об'єктами вибірки отримують емпіричні дані;
- обробка емпіричних даних за допомогою методів описової статистики дає показники вибірки, які називаються статистиками – як і назва, до речи;
- застосовуючи методи статистичних висновків до статистик, отримують параметри, які характеризують властивості генеральної сукупності [38].

Побудова варіаційного ряду.

Якщо ряд розподілу побудований за кількісною ознакою, то такий ряд називають *варіаційним*. Побудувати варіаційний ряд - означає упорядкувати кількісний розподіл одиниць сукупності по значеннях ознаки, а потім підрахувати кількість одиниць сукупності з цими значеннями (побудувати групову таблицю). Виділяють три форми варіаційного ряду: ранжований ряд, дискретний і інтервальний ряд.

Ранжований ряд – це розподіл окремих одиниць сукупності в порядку зростання або убування досліджуваної ознаки, що дозволяє легко розділити кількісні дані по групах, відразу виявити найменше і найбільше значення ознаки, виділити значення, які найчастіше повторюються. В такому ряді кожному об’єкту присвоюється певний ранг – місце в цьому ряд. У тому випадку, якщо кілька елементів мають той самий ранг, то кожному з них присвоюється середнє від займаних ними місць.

Дискретний ряд – це такий варіаційний ряд, в основу побудови якого покладені ознаки з переривчастою зміною (дискретні ознаки). До останніх можна віднести тарифний розряд, кількість дітей в сім'ї, число працівників на підприємстві… Ці ознаки можуть мати лише певну кількість визначених значень.

Якщо ознака неперервно змінюється (розмір доходу, стаж роботи, вартість основних фондів підприємства і так далі, які в певних межах можуть набувати будь-яких значень), то для цієї ознаки потрібно будувати *інтервальний варіаційний ряд*. Групова таблиця такого ряду містить дві колонки. У першій вказується значення ознаки в інтервалі «від – до» (варіанти) у другій - число одиниць, що входять в інтервал (частота). *Частота* – це число повторень окремого варіанту значень ознаки, позначається f_i, іноді вказується сума накопичених частот, рівна об'єму досліджуваної сукупності [3].

Емпірична функція розподілу. Багатокутник частот.

Емпіричні дані, які отримані шляхом вимірювання властивостей вибіркових об’єктів, повинні пройти обробку і систематизацію: внесення у табличні форми (етап табуляції), упорядкування у варіаційні послідовності (ряди), представлення у вигляді емпіричних або статистичних розподілів [38].

Емпірична функція розподілу $F^*(x)$. Ця функція визначається як відношення: $F^*(x) = \frac{n_x}{n}$, де n_x - кількість варіант СТРОГО МЕНШИХ, ніж x, при цьому «ікс» «пробігає» всі значення від «мінус» до «плюс» нескінченності: (див. рис. 1.4.)

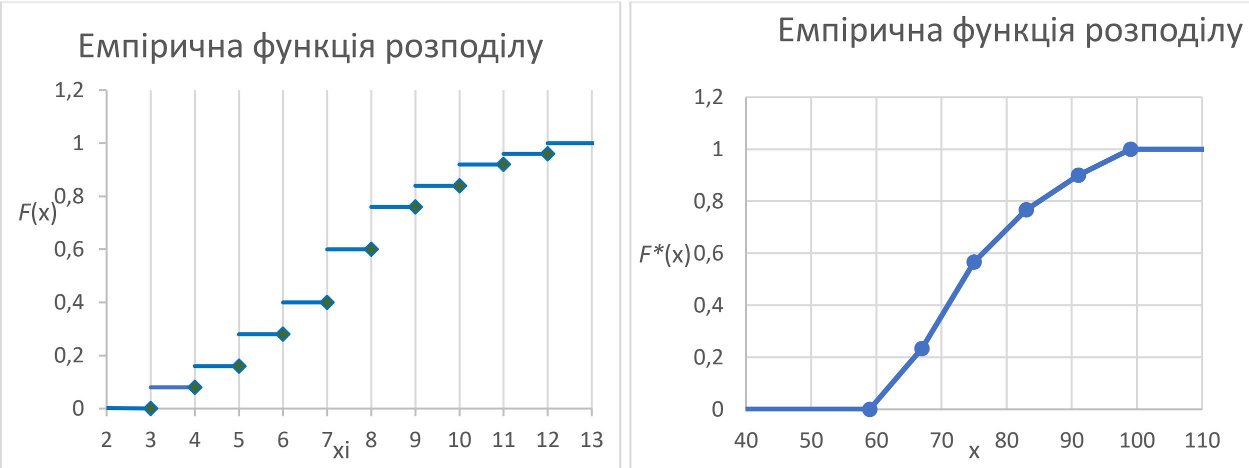


Рис. 1.4. Емпірична функція розподілу для дискретного і інтервального ряду

Ознака X є випадковою величиною, а статистичний ряд – емпіричним (тобто отриманим у результаті експерименту або спостережень) законом її розподілу.

Статистичний ряд називається дискретним, якщо він є законом розподілу дискретної випадкової величини, та інтервальним, якщо він є законом розподілу неперервної випадкової величини.

Таблиця 1.3

Дискретний статистичний ряд

Варіанти x_i	x_1	x_2	…	x_k
Частоти n_i (відносні частоти w_i)	n_1 (w_1)	n_2 (w_2)	…	n_k (w_k)

де k – кількість варіант.

Таблиця 1.4

Неперервний статистичний ряд

Інтервали $[a_i ;a_{i+1})$	$[a_1\, a_2)$	$[a_2 ;a_3)$	…	$[a_{k\text{-}1} ;a_k)$
Частоти n_i (відносні частоти w_i)	n_1 (w_1)	n_2 (w_2)	…	n_k (w_k)

де k – кількість інтервалів [36].

Багатокутник частот (полігон частот) – це графічний пристрій для розуміння форм розподілів. Вони служать тій же меті, що і гістограми, але особливо корисні для порівняння наборів даних. Частотні полігони також є хорошим вибором для відображення кумулятивних розподілів частоти. [18]

Полігоном частот (відносних частот) називається ламана лінія, що сполучає точки з координатами (див. рис 1.5.):

(x_i; n_i) або (x_i; w_i) для дискретного статистичного ряду;

(x_i; n_i) або (x_i: w_i) для інтервального ряду, де x_i – середина і-того інтервалу,

$$x_i = \frac{a_i + a_{i+1}}{2}$$

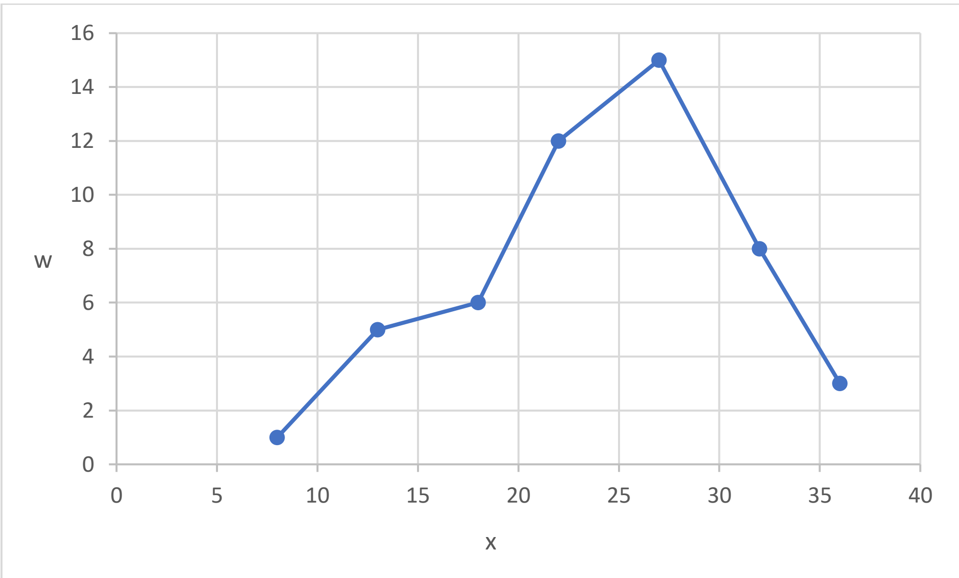


Рис. 1.5. Полігон частот

Гістограмою називається ступінчаста фігура, яка складається з прямокутників з основами, що дорівнюють довжинам інтервалів, та висотами, які пропорційні частотам n_i (відносним частотам w_i) і обчислюються як відношення частот n_i (відносних частот w_i) до довжини відповідних інтервалів (див. рис 1.6.)

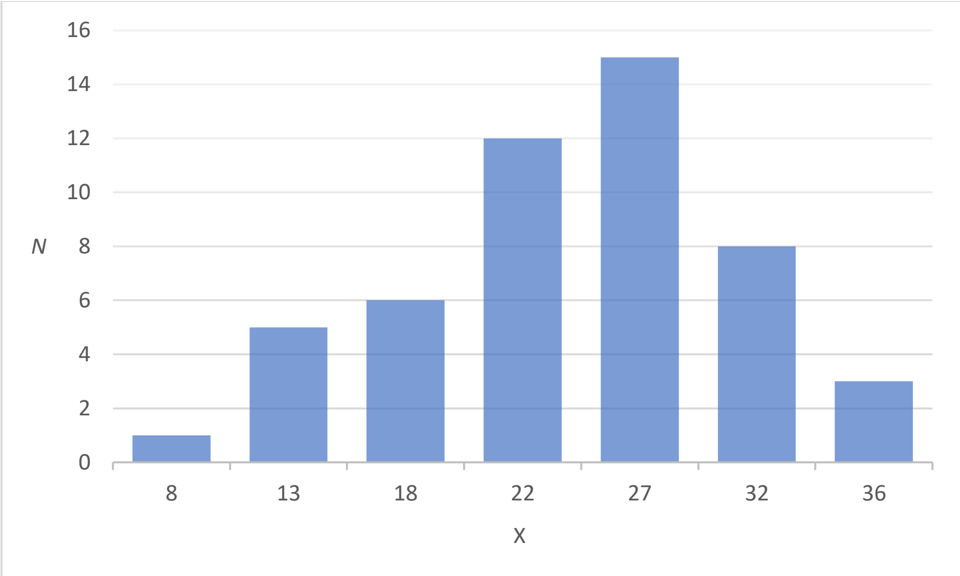


Рис. 1.6. Гістограма

Площа гістограми частот дорівнює об’єму вибірки n.

Площа гістограми відновних частот дорівнює одиниці [36].

ОБРОБКА ВИБІРОК ДАНИХ

Параметри вибірки та особливості їх розрахунку.

Середнє арифметичне.

Середнє арифметичне вибірки визначається як сума всіх значень вибірки, поділена на її обсяг.

$$\bar{x} = \frac{\sum x}{n}, \tag{1.1}$$

де $\sum x$= сума всіх значень вибірки, n = обсяг вибірки.

Середнє значення для згрупованих даних можна знаходити за більш зручною формулою:

$$\bar{x} = \frac{\sum(f * x)}{n}, \tag{1.2}$$

де $\sum (f * x)$ позначає суму добутків кожного з значень ознаки на його частоту, n є загальною кількістю спостережень, що дорівнює також сумі всіх частот [24].

Зважене середнє – середнє значення, що отримується при об'єднанні кількох груп спостережень.

Якщо групи мають однаковий обсяг, то групове середнє можна обчислити як середнє арифметичне наявних середніх значень по кожній групі. Якщо ж групи мають різний обсяг, то групове середнє можна знайти за такою формулою:

$$\bar{X} = \frac{\sum(\bar{x} * n)}{N}, \tag{1.3}$$

де $\sum(\bar{x} * n)$ = сума добутків середніх в групі на кількість елементів в цій групі, N = загальна кількість спостережень у всіх групах [24].

Дисперсія – це показник, який описує, наскільки розкидані дані від середнього значення. Для обчислення дисперсії необхідно взяти кожне значення вибірки, відняти від нього середнє значення, піднести результат до квадрату, підсумувати всі такі значення та розділити їх на розмірність вибірки мінус один [15].

Основна формула (за визначенням) для знаходження дисперсії:

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}, \tag{1.4}$$

Формула для знаходження дисперсії вибірки:

$$s^2 = \frac{n * \sum x^2 - \left(\sum x\right)^2}{n * (n - 1)}, \tag{1.5}$$

[24].

Стандартне відхилення – це квадратний корінь з дисперсії. Воно показує, наскільки дані розкидані навколо середнього значення [15].

Стандартне відхилення позначається s і обчислюється за формулою:

$$s^2 = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}, \qquad (1.6)$$

[24].

Медіана визначається як серединне значення вибірки, або значення, вище і нижче якого розташовується однакове кількість спостережень. Для знаходження медіани обов'язково впорядкувати дані [24].

Медіану визначають за формулою:

$$m_e = \frac{n_{max} - n_{min}}{2} + n_{min}, \qquad (1.7)$$

[27].

Мода – значення, що найбільш часто зустрічається у вибірці або наборі даних. У разі, якщо дані згруповані і побудовано розподіл частот, модою є значення, що має найбільшу частоту.

Мода *Мо* цілком придатна для вимірювання центральної тенденції хоча б тому, що це єдиний спосіб описувати номінальний розподіл не гірше порядкового або інтервального. Обмеження в застосуванні пов’язані з тим, що мода розглядає лише одну особливість розподілу, а саме, розташування найбільш частого значення. Інші важливі особливості, такі як число спостережень вище або нижче моди, відстань між модальними значеннями і інші характеристики, залишаються поза увагою [24].

<u>Емпіричний момент.</u>

Середнє зважене значення варіант у степені k (k = 1, 2, 3,…) називають *початковим емпіричним моментом k-го порядку* ν^*_k який обчислюється за формулою:

$$v_k^* = \frac{\sum x_k n_i}{n} \qquad (1.8)$$

При k = 1 дістанемо початковий момент першого порядку:

$$v_1^* = \frac{\sum x_i n_i}{n} = \bar{x}_B \tag{1.9}$$

При k = 2 обчислимо початковий момент другого порядку:

$$v_2^* = \frac{\sum x_i^2 n_i}{n} \tag{1.10}$$

Отже, дисперсію вибірки можна подати через початкові моменти першого та другого порядків, а саме:

$$D_B = v_2^* - (v_1^*)^2, \tag{1.11}$$

Середнє зважене відхилення варіант у степені k (k = 1, 2, 3,…) називають *центральним емпіричним моментом k-го порядку*

$$\mu_k^* = \frac{\sum (x_i - \bar{x}_B)^k n_i}{n}, \tag{1.12}$$

При k = 1 дістанемо:

$$\mu_1^* = \frac{\sum (x_i - \bar{x}_B) n_i}{n} = \frac{\sum x_i n_i}{n} - \bar{x}_B \cdot \frac{\sum n_i}{n} = \bar{x}_B - \bar{x}_B = 0$$

При k = 2 маємо:

$$\mu_2^* = \frac{\sum (x_i - \bar{x}_B)^2 n_i}{n} = D_B$$

[24].

Коефіцієнт асиметрії (скошеності) – розраховується через момент третього порядку і використовується для характеристики ступеня асиметричності ряду або щільності розподілу ймовірності випадкової величини *X* відносно її математичного сподівання.

Коефіцієнт асиметрії – безрозмірна величина:

$$A_s = \frac{m_3}{\sigma_{\text{в}}^3}, \tag{1.13}$$

де $\sigma_{\text{в}}^3$ – куб стандартного вибіркового відхилення, а m_3 – так званий центральний емпіричний момент третього порядку. Для несгрупованої статичної сукупності він розраховується так:

$$m_3 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_{\text{в}})^3}{n},$$

де $\bar{x}_в$ – вибіркова середня,

а сформованого варіаційного ряду – так:

$$m_3 = \frac{\sum_{i=1}^{k}(x_i - \bar{x}_в)^3 \cdot n_i}{n},$$

де x_i – варіанти дискретного ряду чи середини часткових інтервалів інтервального ряду, а – відповідні частоти n_i.

Якщо $A_s > \pm\, 0.5$ – значне зміщення;

якщо $A_s < \pm\, 0.25$ – незначне зміщення;

$0.25 < A_s < 0.5$ – помірне зміщення [8].

Якщо $A_s = 0$, то розподіл має симетричну форму;

при $A_s > 0$ розподіл має додатню (правосторонню) асиметрію;

при $A_s < 0$ — від’ємну (лівосторонню) (див. рис. 1.6) [19].

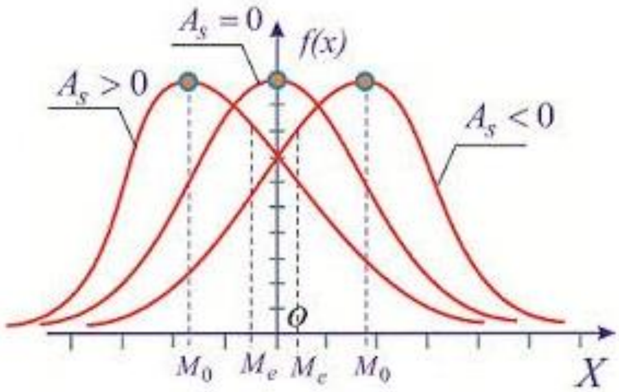


Рис. 1.6. Коефіцієнт асиметрії

Коефіцієнт гостровершинності (ексцесу) характеризує гострість вершини ряду, тобто відхилення від нормального розподілу варіант з виступанням або падінням вершини кривої розподілу.

Він визначається через четвертий центральний момент і є безрозмірною величиною:

$$E_k = \frac{m_4}{\sigma_в^4} - 3, \tag{1.14}$$

де m_4 – центральний емпіричний момент четвертого порядку:

$$m_4 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_\text{в})^4}{n}$$

– для несгрупованих даних, та

$$m_4 = \frac{\sum_{i=1}^{k}(x_i - \bar{x}_\text{в})^4 \cdot n_i}{n}$$

– для сформованого варіаційного ряду.

Властивості ексцесу:

Ексцес дорівнює нулю для нормального розподілу $E_k = 0$.

Ексцес додатний для кривих більш гострої форми, ніж нормальна щільність розподілу $E_k > 0$ – гостровершинність.

Ексцес від'ємний для кривих менш гострої форми порівняно із нормальною щільністю розподілу $E_k < 0$ – плосковершинність (див. рис. 1.7).

Для нормованого розподілу ($\bar{x}$ = 0, $\sigma = 1$) коефіцієнт гостровершинності [8].

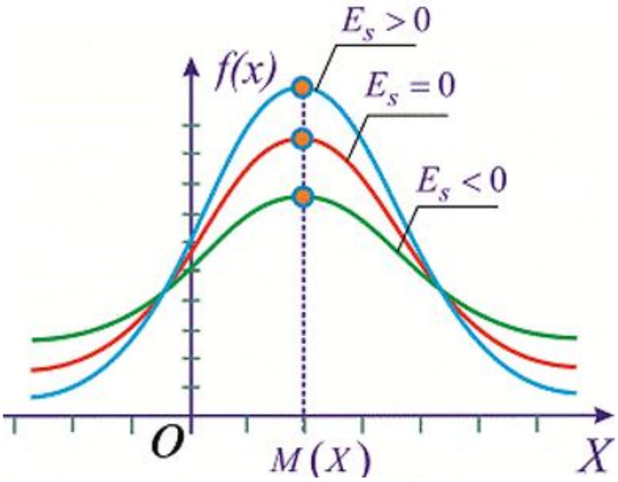


Рис. 1.7. Ексцес

ОСНОВИ КОРЕЛЯЦІЙНОГО АНАЛІЗУ

<u>Поняття про кореляцію.</u>

Кореляція (від лат. correlatio – співвідношення) – реально встановлений факт взаємозв'язку певного стану однієї змінної з певними значеннями іншої, коли зміна однієї з них супроводжується зміною другої [13].

Якщо дві деякі характеристики, отримані для одного і того ж «об’єкта», мають тенденцію змінюватися сумісно так, що створюється можливість передбачити одну з них за значенням іншої, то кажуть, що ці характеристики корелюють одна з одною. Відповідно в статистиці кореляція виражає ступінь взаємозв’язку між такими характеристиками. Кількісно ця ступінь взаємозв’язку виражається за допомогою коефіцієнта кореляції [6].

Характеристика кореляційних зв'язків.

Кореляційний зв'язок – це узгоджені зміни двох ознак або більшої кількості ознак (множинна кореляційний зв'язок).

Кореляційний зв'язок відображає той факт, що мінливість однієї ознаки знаходиться в деякій відповідності до мінливості іншої. З іншого боку, кореляційний зв'язок може говорити не про залежність ознак між собою, а про залежність цих ознак від іншої (інших). Кореляційні зв'язки не може розглядатися як свідчення причинно-наслідкового зв'язку, вони свідчать лише про те, що зміни однієї ознаки, як правило, супроводжують певними змінами іншої, але чи знаходиться причина змін в одній з ознак або вона виявляється за межами досліджуваної пари ознак, нам невідомо [6].

Кореляційні зв’язки розрізняються за:

- напрямком,
- потужністю (силою зв’язку),
- ймовірністю,
- формою.

Напрямок.

Прямий зв'язок (позитивна кореляція “+”) динаміка значень обох випадкових величин є односпрямованою — збільшення одного значення зумовлює збільшення іншого (див. рис. 1.8).

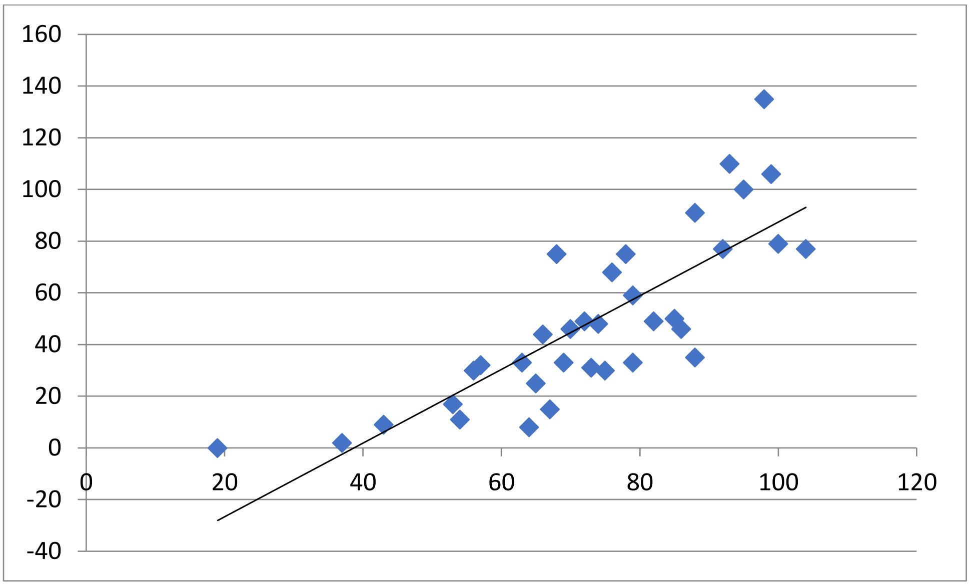


Рис. 1.8. Прямий зв’язок (позитивна кореляція)

Зворотний зв'язок (негативна кореляція “–”) — динаміка значень обох випадкових величин є різноспрямованою — збільшення одного значення зумовлює зменшення іншого (див. рис. 1.9).

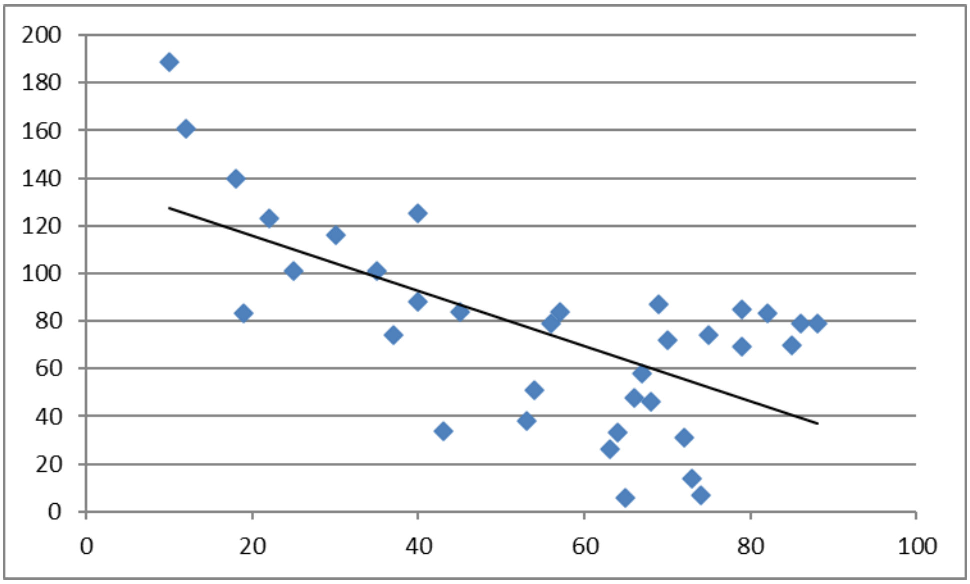


Рис. 1.9. Зворотний зв'язок (негативна кореляція)

Нульове значення коефіцієнту кореляції свідчить про повну незалежність двох випадкових величин (див. рис. 1.10) [29].

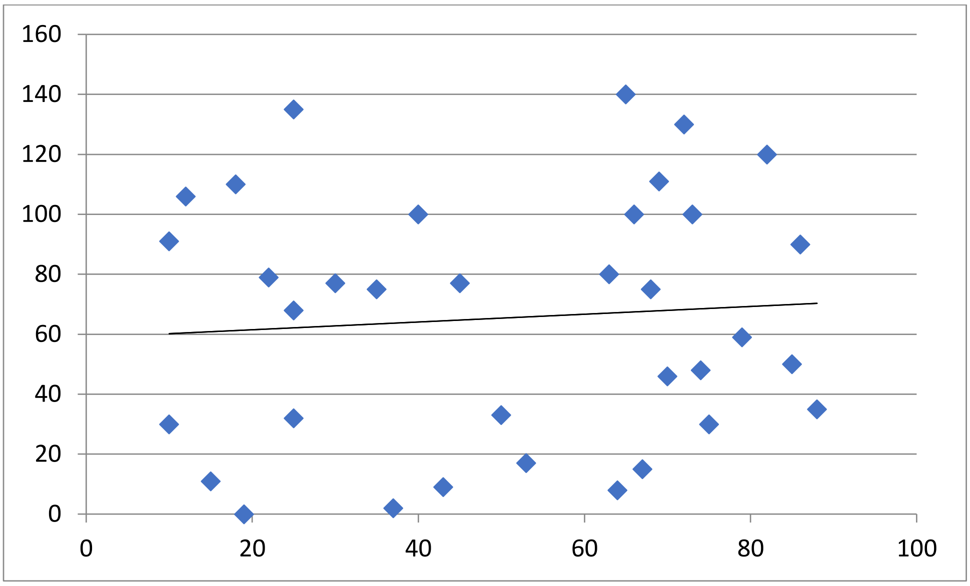


Рис. 1.10. Нульова кореляція

Потужність.

Використовуються кілька систем класифікації потужності (сили) кореляції. Загальна класифікація засвідчує, що кореляція:

- сильна, або тісна при $|r| \geq 0{,}7$;
- середня при $0{,}5 \leq |r| < 0{,}7$;
- помірна при $0{,}3 \leq |r| < 0{,}5$;
- слабка при $0{,}2 \leq |r| < 0{,}3$;
- дуже слабка при $|r| < 0{,}2$. [6]

За значенням коефіцієнта кореляції можна зробити такі висновки:

- якщо значення близьке до – 1, то між факторами існує щільний зворотній зв'язок;
- якщо дорівнює 0, то зв'язок відсутній
- якщо близьке до +1, то між факторами існує щільний прямий зв'язок [32].

Ймовірність.

Важливим елементом кореляційного аналізу є перевірка статистичної гіпотези про існування статистичного зв’язку між досліджуваними випадковими величинами проти нульової гіпотези про відсутність зв’язку. Для зручності проведення такого аналізу розраховані таблиці критичних значень коефіцієнта кореляції. Якщо коефіцієнт кореляції перебільшує критичне значення, то кореляція вважається вірогідною. Критичне значення визначається кількістю пар результатів спостереження випадкових величин у вибірках і заданим рівнем значущості, а також, особливостями обчислення коефіцієнта кореляції та умовами його застосування [32].

Форма.

За формою кореляційний зв'язок може бути лінійної і криволінійної:

прямолінійна – рівномірна зміна одного значення випадкової величини відповідає рівномірним змінам іншої (при незначних коливаннях);

криволінійна – рівномірна зміна одного значення випадкової величини відповідає нерівномірним змінам іншого параметра (нерівномірність має певну закономірність).

Якщо зв’язок між досліджуваними явищами суттєво відрізняється від лінійної, то коефіцієнт кореляції непридатний для визначення міри зв’язку.

Він може вказати на відсутність взаємозв’язку, там де простежується сильна криволінійна залежність.

При нелінійному кореляційному зв’язку рівномірним змінам однієї ознаки відповідають в середньому нерівномірні, які підлягають відповідній закономірності змін другої ознаки [29].

Параметри кореляції та їх розрахунок.

Найбільш поширеними способами розрахунку є:

- метод квадратів (Пірсона);
- метод рангів (Спірмена).

Умови для застосування коефіцієнта кореляції Пірсона

1. Змінні, що порівнюються повинні бути отримані в інтервальній шкалі або шкалі відносин.

2. Розподілу змінних X і Y повинні бути близькі до нормального.

3. Число ознак, що варіюють, у порівнянні зі іншими X і Y має бути однаковим.

4. Таблиці рівнів значущості для коефіцієнта кореляції Пірсона розраховані від n = 5 до n = 1000. Оцінка рівня значущості за таблицями здійснюється за числом ступенів свободи v=n – 2.

Формула для підрахунку коефіцієнта кореляції така:

$$r_{xy} = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{\left(n\sum x_i^2 - (\sum x_i)^2\right)\left(n\sum y_i^2 - (\sum y_i)^2\right)}}, \quad (1.15)$$

де x_i – значення, що приймаються змінною X;

y_i – значення, що приймаються змінною Y.

Коефіцієнт рангової кореляції Спірмена використовують для визначення щільності зв'язків між ознаками, якщо їх значення упорядковані або проранжовані за ступенем спадання або зростання ознаки. Коефіцієнт кореляції рангів розраховують за формулою:

$$R = 1 - \frac{6 \cdot \sum_{i-1}^{n} d^2}{n \cdot (n^2 - 1)}, \quad (1.16)$$

де n – обсяг вибірки;

d – різниця між рангами показників одних і тих самих реципієнтів

Умови для застосування коефіцієнта кореляції Спірмена

1. Змінні, що порівнюються повинні бути отримані в порядковій, інтервальній шкалах або в шкалі відносин.

2. Характер розподілу величин, корелюються не має значення.

3. Число ознак, що варіюють, повинно бути однаковим і перебувати в межах від 5 до 40.

4. У разі якщо є багато повторюваних рангів, то необхідно вносити поправку на однакові ранги [6].

МЕТОД ВИБІРКИ

Поняття похибки вибірки.

У загальному випадку розбіжності між даними будь-якого статистичного спостереження і дійсним значенням показників вважають помилками спостереження. Існують різні підходи до класифікації таких помилок. За причинами виникнення розрізняють помилки реєстрації та похибки вибірки (похибки репрезентативності).

Похибки вибірки (репрезентативності) виникають через те, що у вибіркових обстеженнях здійснюється відбір для обстеження лише частини сукупності. Вибіркова похибка відображає той факт, що оцінка може бути іншою, коли відбирається інша вибірка, за інших незмінних умов [7].

Похибки вибірки можуть бути:

- систематичними – виникають унаслідок порушення принципів проведення вибіркового спостереження; мають тенденційний характер викривлення величини досліджуваної ознаки в бік її збільшення або зменшення;
- випадкові – зумовлені тим, що вибіркова сукупність не відтворює точно середні і відносні показники генеральної сукупності.

Для узагальнюючої характеристики похибки вибірки розраховують середню похибку репрезентативності μ, її називають в деяких випадках стандартом [14].

Для визначення середньої помилки репрезентативності вибірки застосовують чотири формули для повторного і безповторного відбору (табл. 1.5).

Таблиця 1.5

Формули для визначення середньої помилки вибірки

Способи відбору	Помилка вибірки	
	при визначенні середньої	при визначенні частки
Повторний	$\mu = \sqrt{\frac{\sigma^2}{n}}$	$\mu = \sqrt{\frac{w \cdot (1-w)}{n}}$
Безповторний	$\mu = \sqrt{\frac{\sigma^2}{n} \cdot (1 - \frac{n}{N})}$	$\mu = \sqrt{\frac{w \cdot (1-w)}{n} \cdot (1 - \frac{n}{N})}$

де μ – середня похибка вибірки;

σ^2 – середній квадрат відхилень у вибірці;

n – чисельність вибіркової сукупності;

N – чисельність генеральної сукупності;

$1 - \frac{n}{N}$ – необстежена частка генеральної сукупності;

$\frac{n}{N}$ – частка обстеженої частини вибіркової сукупності;

w – частка даної ознаки у вибірці;

$(1 - w)$ – частка протилежної ознаки у вибірці [12].

Безповторний відбір гарантує більш точні результати, тому що він включає можливість обстеження одних і тих самих одиниць при відборі з генеральної сукупності.

Таким чином, *стандартна похибка вибірки* μ – це середнє квадратичне відхилення вибіркових оцінок від значення параметра в генеральній сукупності [14].

Помилка параметра.

Відмінності значень вибіркових параметрів від генеральних називаються *помилкою репрезентативності даного параметра* або просто (статистичної) помилкою. При збільшенні обсягу вибірки помилки репрезентативності прагнуть нуля (наслідок закону великих чисел). Чисельно виражені статистичні помилки є мірою тих меж, у яких вибіркові параметри можуть відхилятися від значень генеральних параметрів [9].

Помилка середнього арифметичного.

У статистиці під «помилкою» слід розуміти не помилку дослідження, а міру представництва даної величини, тобто наскільки середня арифметична величина, одержана із вибіркової сукупності (10—20 випадків) відрізняється від істинної, яка була б одержана на генеральній сукупності (100, 200, 300 і більше випадків). Вона визначається за формулою

$$m_M = \pm \frac{S}{\sqrt{n}} \tag{1.17}$$

У виразі під коренем ставиться n—1, якщо менше 30 випадків, якщо ж більше, то n.

Таким чином, достовірність середньої арифметичної залежить від дисперсії варіаційного ряду. Чим більш компактним є ряд, тим достовірнішою буде середня арифметична, і навпаки, чим він більше розсіяний, тим середня арифметична менш достовірна, міра її точності залежить також від кількості спостережень. Помилка середньої арифметичної ($\pm m_M$) застосовується при обчисленні довірливого коефіцієнта [9].

Похибка середньоквадратичного відхилення.

$$m_S = \frac{S}{\sqrt{2 \cdot n}}, \tag{1.18}$$

Помилка коефіцієнту варіації.

Коефіцієнт варіації – це відносна стандартна похибка, яка визначається шляхом ділення стандартної похибки в абсолютному значенні на очікуване значення оцінки показника [7].

помилка коефіцієнта варіації:

$$m_{CV} = \frac{CV}{\sqrt{2 \cdot n}}, \tag{1.19}$$

[9].

Похибка індикатора асиметрії

$$m_{As} = \sqrt{\frac{6 \cdot n(n-1)}{(n-2) \cdot (n+1) \cdot (n+3)}}$$

Похибка коефіцієнта ексцесу

$$m_{Ek} = \sqrt{\frac{24}{n}}$$

Критерії надійності. Довірчі інтервали.

Довірчим інтервалом називають інтервал, розрахований з даних таким чином, що існує відома вірогідність включення параметра генеральної сукупності, що цікавить нас (невідомого), в інтервал, і ця вірогідність інтерпретується з точки зору випадкового експерименту що починається з витягання випадкової вибірки. Межі довірчого інтервалу визначаються на основі точкової оцінки і граничної помилки вибірки, яка дорівнює добутку стандартної помилки і t–критерію Стьюдента. Гранична помилка вибірки показує максимально можливу помилку для прийнятої вірогідності, а довірче число t – як співвідносяться гранична і стандартна помилки.

$$\Delta = t \cdot \mu, \tag{1.20}$$

де t – коефіцієнт довір'я, який залежить від ймовірності визначення граничної помилки.

Вірогідність того, що параметр сукупності належатиме довірчому інтервалу називають *критерієм надійності*, який зазвичай встановлюють рівним 95%, хоча часто використовують і інші рівні – 90; 99; 99,9%. Чим вище рівень довірчості, тим ширше (а значить, і менш корисний) довірчий інтервал. Приблизне узагальнене формулювання твердження про довірчий інтервал має наступний вигляд:

ми упевнені на 95%, що значення параметра генеральної сукупності знаходиться між значенням оцінки мінус дві стандартні помилки оцінки і значенням оцінки плюс дві стандартні помилки оцінки.

Це твердження засноване на тому факті, що при нормальному розподілі з вірогідністю 0,95 слід чекати значення на відстані $t = 1.960$, тобто приблизно два стандартні відхилення від середнього (див. рис. 1.11).

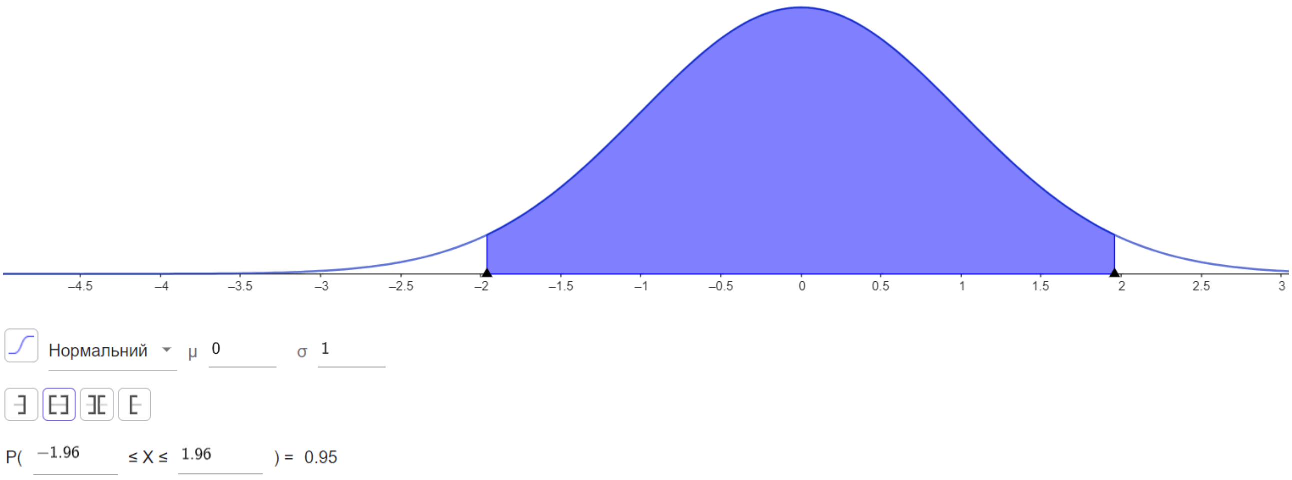


Рис. 1.11. Випадкова величина, розподілена по нормальному закону, з імовірністю 95% потрапляє в інтервал +/– 1,960 стандартних відхилень

Для того, щоб використання довірчого інтервалу було коректним, необхідне виконання двох наступних умов:

(1) дані мають бути випадковою вибіркою з даної генеральної сукупності;

(2) виміряні значення повинні підкорятися нормальному розподілу.

Перша умова гарантує, що дані правильно представляють невідомий параметр, а друге дає підставу використовувати t-таблицю для обчислення вірогідності [10].

СТАТИСТИЧНІ ГІПОТЕЗИ ТА ЇХ ПЕРЕВІРКА

Поняття статистичної гіпотези.

Статистична гіпотеза – це твердження про характер розподілу параметрів генеральної сукупності, яке перевіряється математичними методами.

Гіпотези є: нульова і альтернативна.

Нульову гіпотезу прийнято позначати H_0, а альтернативну гіпотезу – H_A. Прийнято нульову гіпотезу вважати початковою.

Нульова гіпотеза передбачає, що між параметрами генеральних сукупностей немає очікуваних розбіжностей. Запис нульової гіпотези має, наприклад, вигляд H_0: ($\bar{x} = \bar{y}$) що означає: між сукупностями X і $У$ статистично немає відмінностей варіацій.

Альтернативна гіпотеза передбачає протилежне твердження до нульової гіпотези, наприклад, H_A: ($\bar{x} \neq \bar{y}$), тобто варіації генеральних сукупностей статистично неоднакові [48].

<u>Алгоритм перевірки статистичної гіпотези.</u>

1. Формулювання нульової та альтернативної гіпотез.

2. Розрахунок необхідних статистичних характеристик.

3. Вибір рівня значущості α.

4. Вибір та обчислення розрахункового значення критерію для перевірки гіпотез. Критичне значення t-критерію Стьюдента визначається за таблицею.

5. Порівняння розрахункового і табличного значень критерію, на основі чого робиться висновок про прийняття чи відхилення гіпотези [48].

<u>Помилки перевірки гіпотези.</u>

Точність перевірки гіпотези вибирається з практичних міркувань. Відповідна похибка визначається рівнем значущості α.

Імовірність прийняття правильної гіпотези називають вірогідністю β і визначають за формулою $\beta = (1 - \alpha) \cdot 100$ %.

При перевірці гіпотез похибки є двох родів.

Похибка першого роду – відхилення істинного припущення.

Похибка другого роду – прийняття помилкового твердження [48].

Можливі чотири ситуації, які наведено в таблиці 1.6:

Таблиця 1.6

Помилки перевірки гіпотези

	H_0	H_A
H_0	правильне рішення, ймовірність 1-α	помилка другого роду
H_A	помилка першого роду, ймовірність α	правильне рішення

Помилка першого роду фактично відповідає ситуації, коли дослідник бачить відмінності, якийсь вплив або зв'язок там, де їх немає. Якщо за результатами аналізу спостережених даних виявляється, що ймовірність помилки першого роду менша за вибране критичне значення (в педагогіці це критичне значення звичайно вибирають рівним 0,05), то приймається гіпотеза H_A. Інакше вважають, що немає підстав відхилити гіпотезу H_0. Критичне значення ймовірності помилки першого роду називають рівнем значущості [5].

Статистичні критерії та їх види.

Процедура перевірки гіпотез зводиться до обчислення за вибірковими даними величини, яку називають статистичним критерієм.

Критерій – це математичний метод, який дозволяє перевірити статистичну гіпотезу і забезпечує прийняття істинної і відхилення невірної гіпотези з наперед заданою імовірністю [48].

Відповідно до статистичних гіпотез статистичні критерії діляться на параметричні й непараметричні.

Параметричні критерії використовуються в завданнях перевірки параметрів гіпотез і включають у свій розрахунок показники розподілу, наприклад, середні, дисперсію тощо. Це такі відомі класичні критерії, як z-критерій, t-критерій Стьюдента, F-критерій Фішера та ін. *Непараметричні критерії* перевірки гіпотез засновані на операціях з іншими даними, зокрема, частотами, рангами тощо. Це λ-критерій Колмогорова-Смірнова, U-критерій Уілкоксона-Манна-Уітні та багато інших.

Параметричні критерії дозволяють прямо оцінити рівень основних параметрів генеральних сукупностей, різниці середніх і відмінності в дисперсіях. Критерії спроможні виявити тенденції зміни ознаки при переході від умови до умови, оцінити взаємодію двох і більш факторів у впливі на зміни ознаки. Параметричні критерії вважаються дещо більш потужними, ніж непараметричні, за умов, якщо ознака виміряна за інтервальною шкалою і нормально розподілена. Проте з інтервальною шкалою можуть виникнути певні проблеми, якщо дані, представлено не в стандартизованих оцінках. До того ж перевірка розподілу «на нормальність» вимагає досить складних розрахунків, результат яких заздалегідь невідомий. Найчастіше розподіли ознак відрізняються від нормального, тоді доводиться звертатися до непараметричних критеріїв.

Непараметричні критерії позбавлені перерахованих вище обмежень. Проте вони не дозволяють здійснити пряму оцінку рівня таких важливих параметрів, як середнє або дисперсія, з їхньою допомогою неможливо оцінити взаємодію двох і більше умов або факторів, що впливають на зміну ознаки. Непараметричні критерії дозволяють вирішити деякі важливі завдання, які супроводжують дослідження в психології і педагогіці: виявлення відмінностей у рівні досліджуваної ознаки, оцінка зсуву значень досліджуваної ознаки, виявлення відмінностей у розподілах ознак.

Застосування критеріїв для прийняття (відхилення) статистичних гіпотез завжди здійснюються з довірчою ймовірністю, інакше кажучи, на певному *рівні значущості* [38].

Параметричний критерій Стьюдента призначено для перевірки гіпотези про рівність математичних сподівань в генеральних сукупностях, з яких отримано вибірки. Для застосування критерію потрібно виконання певних умов:

• метрична шкала вимірювання (інтервальна або відносна);

• розподіл імовірностей у кожній з генеральних сукупностей відповідає нормальному закону;

• вибірки є випадковими вибірками з генеральних сукупностей, які є незалежними та мають однакові дисперсії.

В практиці застосування індуктивної статистики в педагогіці математичні сподівання та дисперсії невідомі, тоді обчислення критерію будується на їх оцінках:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2 + \sum_{i=1}^{m}(y_i - \bar{Y})^2}{n + m - 2}}} \cdot \sqrt{\frac{m \cdot n}{m + n}} \qquad (1.21)$$

де x_i , y_i – значення випадкових величин X та Y у вибірці;

n, m – кількість елементів у вибірках відповідно для величин X та Y.

Величина t має розподіл імовірностей Стьюдента з ($n + m - 2$) ступенями вільності, які можна знайти в спеціальних таблицях.

Незважаючи на те, що оцінки дисперсії за вибірками можуть розрізнятися, таблиці критичних значень критерію Стьюдента розраховані за припущенням, що дисперсії однакові. Порушення цього припущення вважається несуттєвим, якщо обсяги обох вибірок однакові, тобто $m = n$. Якщо обсяги вибірок не однакові та не має підстав вважати, що дисперсії генеральних сукупностей рівні, то задача перевірки гіпотези про рівність математичних сподівань вирішується іншими методами.

Деякі табличні процесори надають користувачеві можливість обчислювати імовірність помилки першого роду на підставі застосування критерію Стьюдента. Це позбавляє користувача від рутинної роботи щодо застосування таблиць критичних значень. Якщо отримана оцінка помилки першого роду менше за 0,05, то приймається альтернативна гіпотеза, інакше немає підстав для відхилення нульової гіпотези. В україномовному табличному процесорі Microsoft Excel таку можливість надає функція ТТЕСТ [5].

Критерій Пірсона дає можливість порівняти розподіли двох випадкових величин за результатами спостереження. Для застосування цього критерію слід подати дані у вигляді таблиці частот, можливе здійснити групування окремих значень за частковими інтервалами. Обчислення критерію передбачає наявність двох таблиць – емпіричних частот і теоретичних частот. Таблиця емпіричних частот містить результати спостереження (вимірювання) двох випадкових величин (наприклад частоти оцінок за контрольну роботу в двох групах), які слід порівняти. Таблиця теоретичних частот будується таким чином, щоб забезпечити однаковий розподіл відносних частот для кожної випадкової величини, такий саме, як розподіл емпіричних частот, якщо додати результати вимірювання обох випадкових величин, які порівнюємо.

Критерій обчислюється за формулою:

$$X^2 = \sum_{i=1}^{m} \sum_{i=1}^{k} \frac{\left(E_{i,j} - T_{i,j}\right)^2}{T_{i,j}}, \tag{1.22}$$

де $E_{i,j}$,$T_{i,j}$ – емпіричні та теоретичні частоти відповідно;

i, j – номери рядків і стовпців таблиць частот відповідно;

m, k – кількість рядків і стовпців у таблиці частот відповідно.

Отримане емпіричне значення критерію χ2 слід порівняти з критичним значенням критерію, яке слід знайти в довідкових таблицях критичних значень для заданого рівня значущості (0,05 – в педагогіці) та наявної кількості ступенів вільності (обчислюється як добуток $(m-1)\cdot(k-1)$, де m, k – кількість рядків і стовпців у таблиці частот відповідно). Якщо $X^2_{\text{емпіричне}} > X^2_{\text{критичне}}$, то приймається альтернативна гіпотеза, інакше немає підстав для відхилення нульової гіпотези. В україномовному табличному процесорі Microsoft Excel таку можливість надає функція ХИ2ТЕСТ [5].

ПОБУДОВА РЕГРЕСІЙНИХ МОДЕЛЕЙ

Поняття регресійної моделі.

Регресійний аналіз полягає у визначенні аналітичного виразу зв'язку у вигляді рівняння регресії. Регресією називається залежність середнього значення випадкової величини результативної ознаки від факторної величини, а рівнянням регресії – рівняння, що описує кореляційну залежність між результативною ознакою і однією або декількома факторними. Якщо зв'язок між явищами може бути приблизно виражений рівнянням прямої лінії, то його називають лінійним зв'язком вигляду: $y = ax + b$. Якщо ж зв'язок може бути виражений рівнянням якої-небудь кривої лінії (параболи, гіперболи і ін.), то такий зв'язок називають нелінійним (криволінійним) зв'язком.

В залежності від кількості вхідних параметрів системи (факторних даних) розрізняють парну або однофакторну залежність, коли є один вхідний аргумент і, якщо ж аргументів більше, ніж один, то залежність називається множинною або багатофакторною [3].

Алгоритм побудови та інтерпретації регресійної моделі.

Регресійний аналіз проводиться за такими етапами:

Встановлення виду кореляційної залежності результативної ознаки Y від факторної ознаки X.

Побудова регресійної моделі.

Перевірка статистичної значущості побудованої моделі.

Перший етап регресійного аналізу є найважливішим, оскільки помилки у виборі виду залежності призводять до побудови регресійної моделі, що не відповідає емпіричним даним і не може використовуватися для прогнозування.

Вибіркові дані для вивчення кореляційного зв'язку між ознаками X та Y, зазвичай, мають вигляд пар їх значень: $(x_1; y_1)$, $(x_2; y_2)$, …,$(x_n; y_n)$,

x_i – значення величини X,

y_i - значення Y,

n – кількість пар значень, $i = \overline{1, n}$.

Якщо їх кількість достатньо велика, то для зручності розрахунків дані групуються і будується статистичний ряд, що містить значення X, відповідні середні значення Y та частоти.

Таблиця 1.7

Статистичний ряд

x_i	x_1	x_2	…	x_k
$\bar{y}_{x_i}$	$\bar{y}_{x_1}$	$\bar{y}_{x_2}$	…	$\bar{y}_{x_k}$
n_i	n_1	n_2	…	n_k

Згруповані дані зображуються графічно, що часто дозволяє визначити вид залежності Y від X.

Ламана лінія, що сполучає точки з координатами $(x_i, \bar{y}_{x_i})$ називається емпіричною лінією регресії.

Якщо емпірична лінія регресії значно наближається до прямої лінії, то висувається гіпотеза про наявність лінійного зв'язку між досліджуваними ознаками (див. рис 1.12) [36].

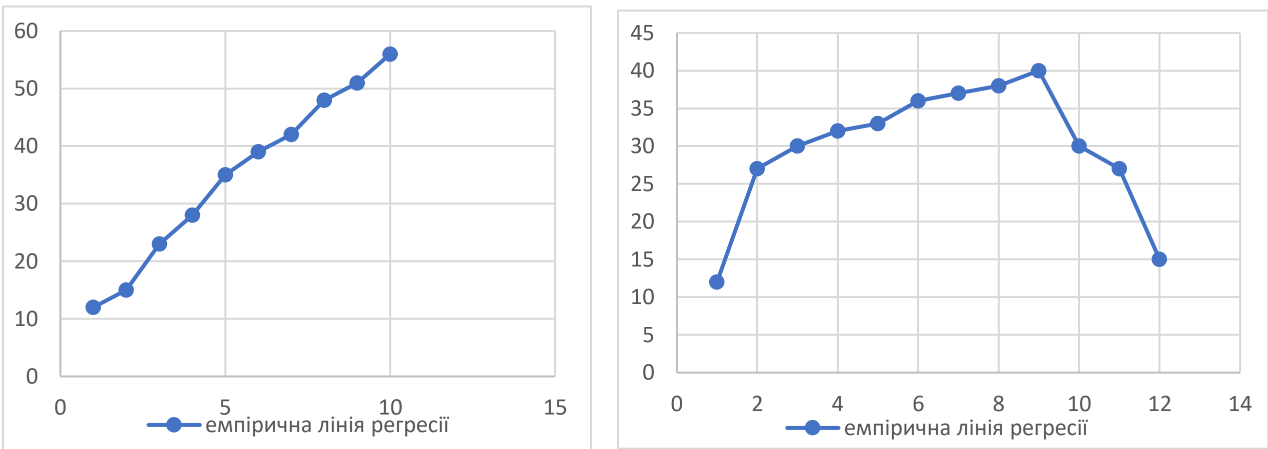


Рис 1.12. Гіпотетичні лінійна і нелінійна залежність

РОЗДІЛ 2. ЛАБОРАТОРНИЙ ПРАКТИКУМ З ОПРАЦЮВАННЯ СТАТИСТИЧНИХ ДАНИХ

ЛАБОРАТОРНА РОБОТА 1
ОСНОВНІ ПОНЯТТЯ ОПИСОВОЇ СТАТИСТИКИ

Мета:

- засвоїти основні поняття та терміни описової статистики;
- опанувати методи обробки та аналізу статистичних даних засобами табличного процесора Microsoft Excel;
- повторити основні принципи та методи представлення даних у вигляді таблиць, графіків та діаграм.

Хід роботи

1. Завдання 1.

Згідно з наданим дискретним статистичним розподілом вибірки успішності учнів , було отримано такі результати (див. табл. 2.1):

Таблиця 2.1

Дискретний статистичний розподіл вибірки

$X = x_i$	2	4	6	8	10	12
n_i	5	10	15	20	40	10

Виконайте наочне графічне представлення даних у вигляді полігону частот і відносних частот та емпіричній функції розподілу - фрагмент робочого листа Microsoft Excel, який потрібно створити в процесі виконання завдання, наведено на рис 2.1.

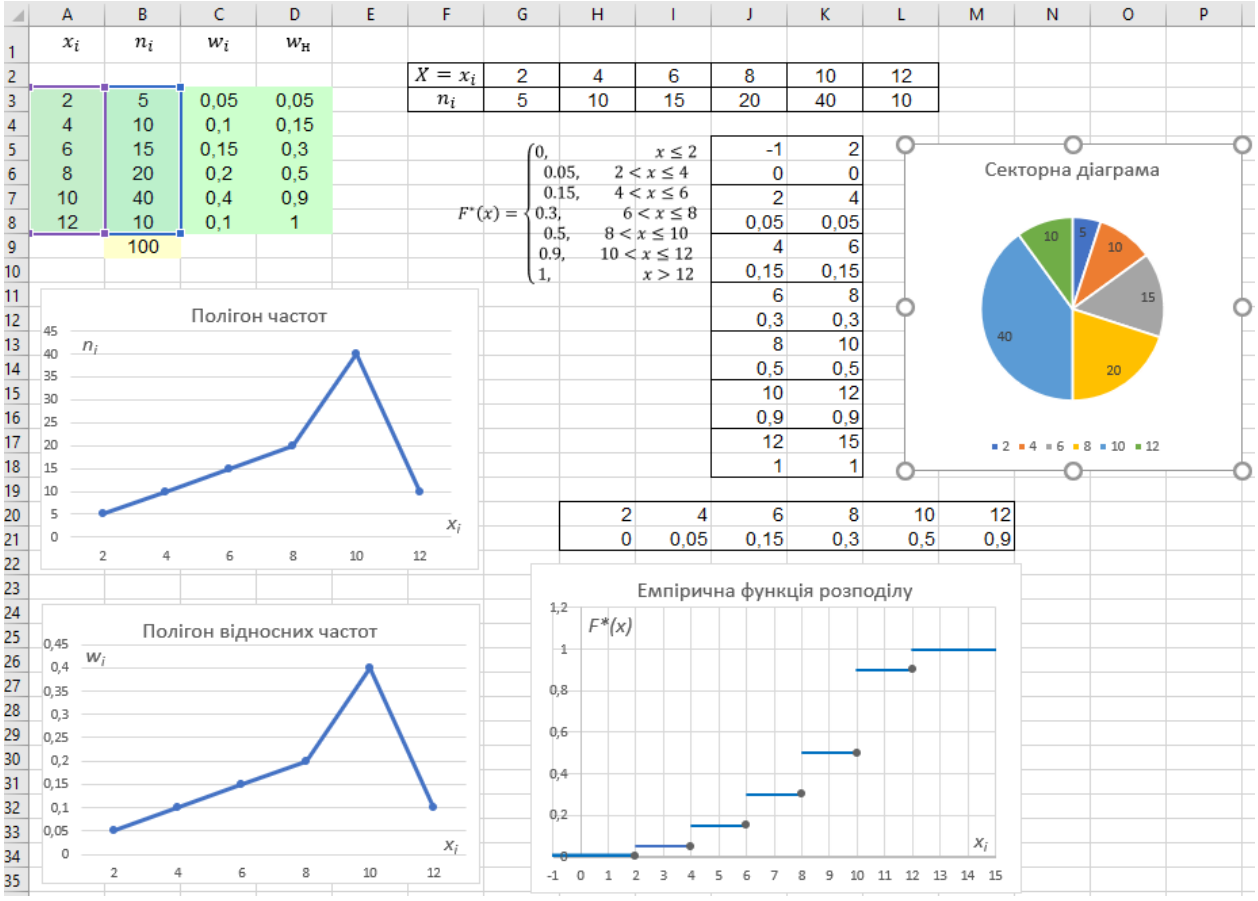


Рис. 2.1. Лабораторна робота 1, завдання 1.

2. Завдання 2.

Згідно з наданим інтервальним статистичним розподілом вибірки успішності учнів , було отримано такі результати (див. табл. 2.2):

Таблиця 2.2

Інтервальний статистичний розподіл вибірки

$h=10$	40-50	50-60	60-70	70-80	80-90	90-100
n_i	5	10	20	25	30	10

Виконати наочне графічне представлення даних у вигляді гістограми, полігону частот і відносних частот та емпіричній функції розподілу.

Фрагмент робочого листа Microsoft Excel, який потрібно створити в процесі виконання завдання, наведено на рис 2.2.

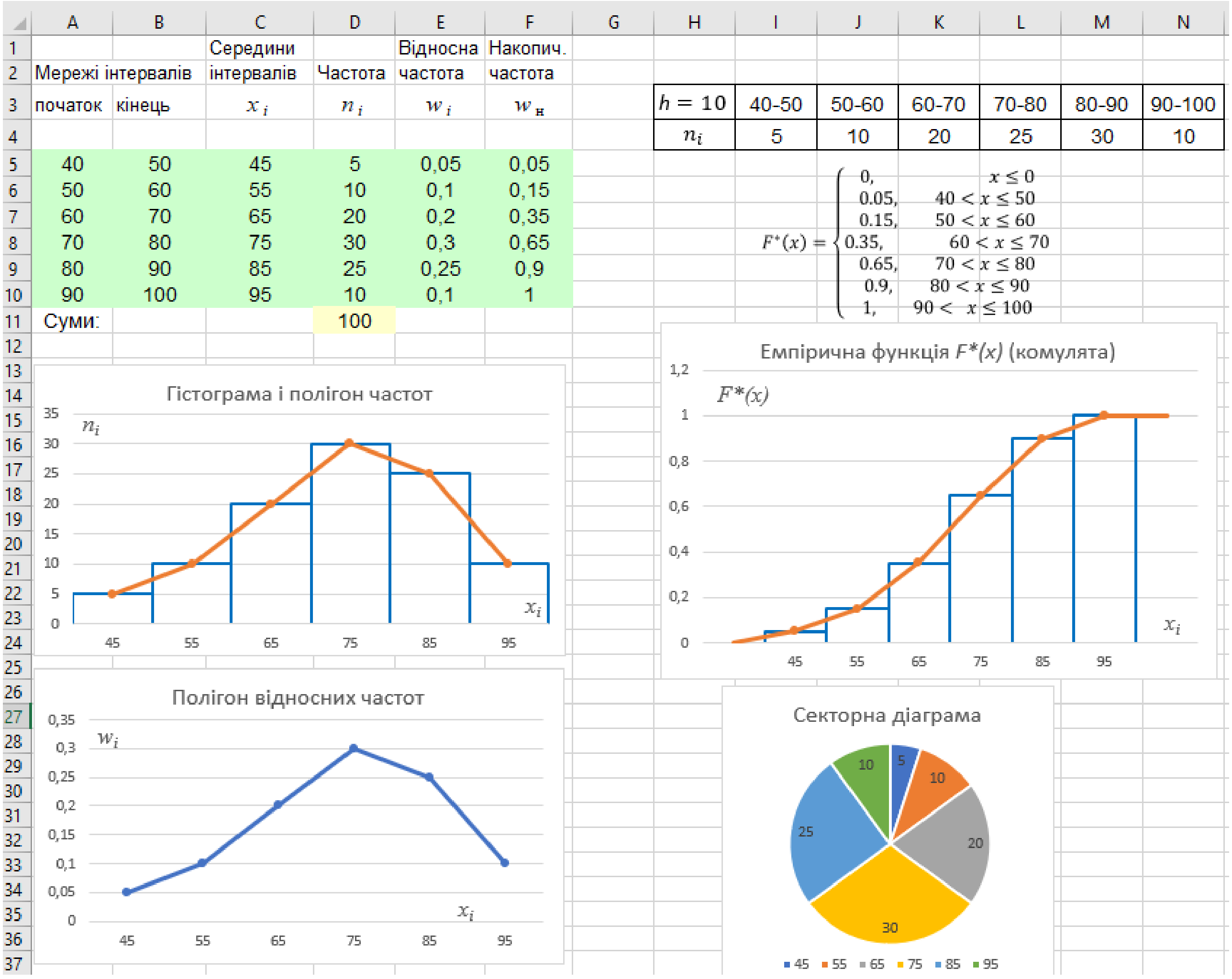


Рис. 2.2. Лабораторна робота 1, завдання 2

ЛАБОРАТОРНА РОБОТА 2

ВИПАДКОВІ ЧИСЛА, ГЕНЕРАЦІЯ ВИПАДКОВИХ ЧИСЕЛ

Мета:

- засвоїти основні поняття про випадкові числа та їх властивості;
- опанувати методи генерації випадкових чисел за допомогою табличного процесора Microsoft Excel;
- повторити принципи та застосування випадкових чисел в описовій у статистиці.

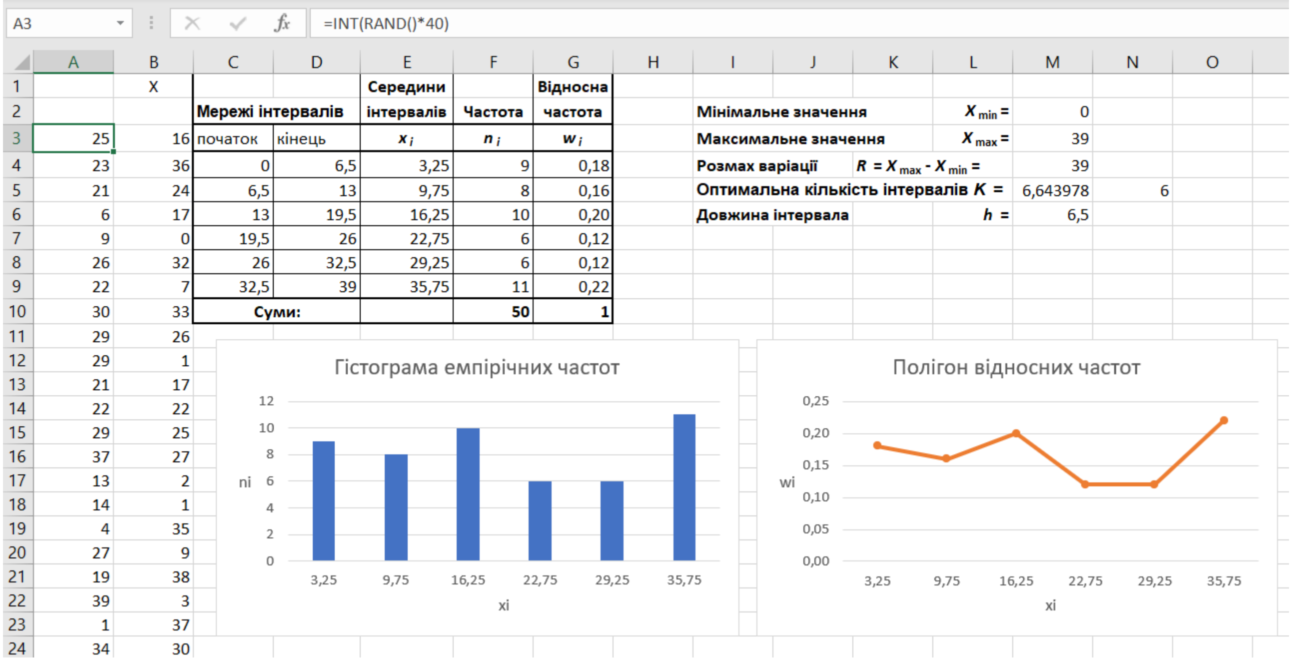
Хід роботи

1. Ознайомитись з матеріалом щодо способів генерації випадкових чисел за допомогою табличного процесора Microsoft Excel за посиланням: https://uk.extendoffice.com/documents/excel/6624-excel-generate-random-data.html#a3

2. Завдання 1.

Використовуючи вбудовані функції, згенеруйте випадкові бали (від 0 до 40) для 50 студентів за підсумками тесту з предмету «Інформатика». Необхідно представити результати у вигляді варіаційного ряду та таблиці емпіричних частот, дати наочне графічне представлення даних у вигляді гістограми емпіричних частот, полігону відносних частот.

Фрагмент робочого листа Microsoft Excel, який потрібно створити в процесі виконання завдання, наведено на рис. 2.3.



	A	B	C	D	E	F	G
1		X			Середини		Відносна
2			Мережі інтервалів		інтервалів	Частота	частота
3	25	16	початок	кінець	x_i	n_i	w_i
4	23	36	0	6,5	3,25	9	0,18
5	21	24	6,5	13	9,75	8	0,16
6	6	17	13	19,5	16,25	10	0,20
7	9	0	19,5	26	22,75	6	0,12
8	26	32	26	32,5	29,25	6	0,12
9	22	7	32,5	39	35,75	11	0,22
10	30	33	Суми:			50	1
11	29	26					
12	29	1					
13	21	17					
14	22	22					
15	29	25					
16	37	27					
17	13	2					
18	14	1					
19	4	35					
20	27	9					
21	19	38					
22	39	3					
23	1	37					
24	34	30					

	I	L	M	N
2	Мінімальне значення	X_{min} =	0	
3	Максимальне значення	X_{max} =	39	
4	Розмах варіації	$R = X_{max} - X_{min}$ =	39	
5	Оптимальна кількість інтервалів K =		6,643978	6
6	Довжина інтервала	h =	6,5	

Рис. 2.3. Лабораторна робота 2, завдання 1

3. Завдання 2.

Виконайте вправи з використанням вбудованих функцій:

- А) згенеруйте 100 випадкових чисел у діапазоні від 60 до 100 з двома знаками після коми, які представлятимуть оцінки студентів за підсумковий іспит;
- створіть таблицю з іменами студентів та їхніми оцінками;
- побудуйте гістограму розподілу оцінок.
- Б) згенеруйте випадкові оцінки (від 2 до 5) для 20 студентів за виконання групових проектів з предмету "Проектний менеджмент".
- створіть таблицю з іменами студентів та їхніми оцінками.
- побудуйте діаграму, яка показує відсоткове співвідношення студентів з різними оцінками.
- згенеруйте 50 випадкових цілих чисел у діапазоні від 1 до 50, які представлятимуть бали студентів за тест з предмету "Інформатика". Кожне число має бути унікальним.

Фрагмент робочого листа Microsoft Excel, який потрібно створити в процесі виконання завдання, наведено на рис. 2.4а, б.

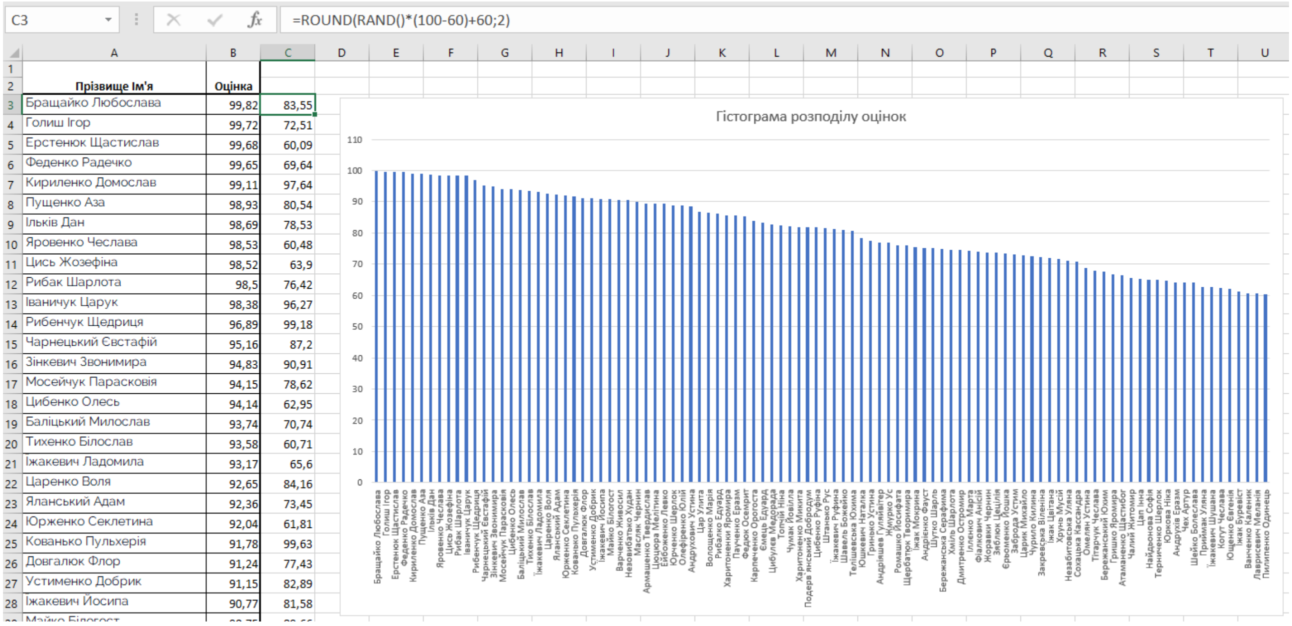


Рис. 2.4а. Лабораторна робота 2, завдання 2а

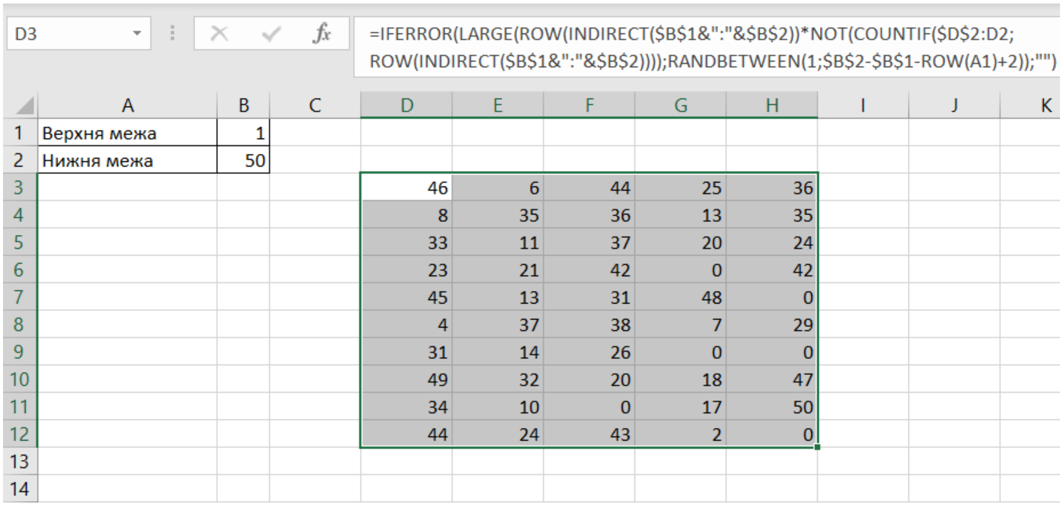


Рис. 2.4б. Лабораторна робота 2, завдання 2б

4. Завдання 3.

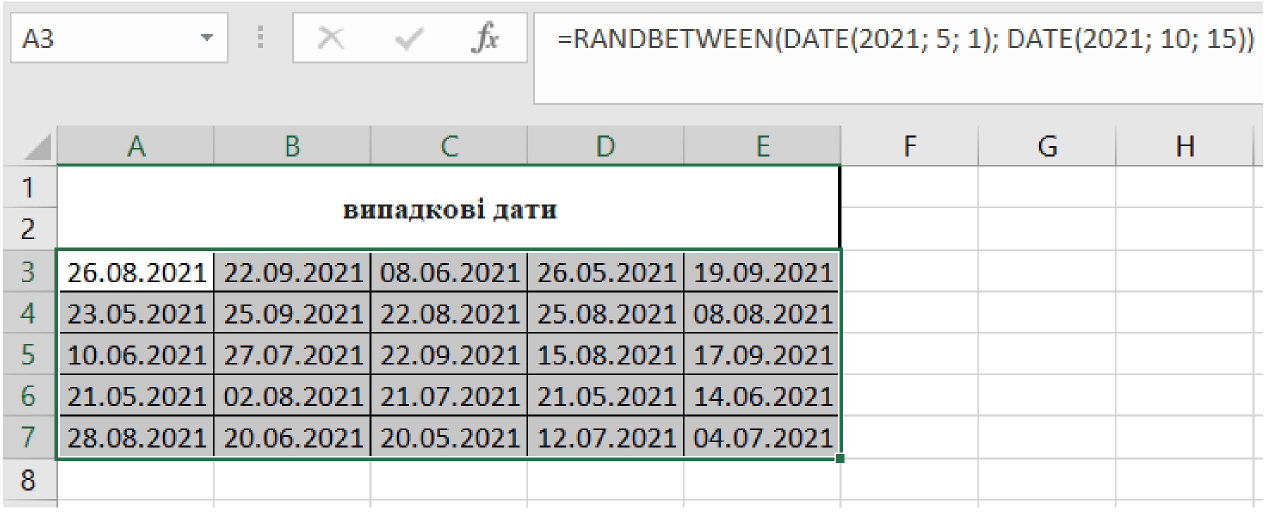
З використанням вбудованих функцій згенеруйте:

a) 25 випадкових дат між двома даними датами;

b) 25 випадкових показань годинника в зазначеному діапазоні;

c) 50 випадкових символів.

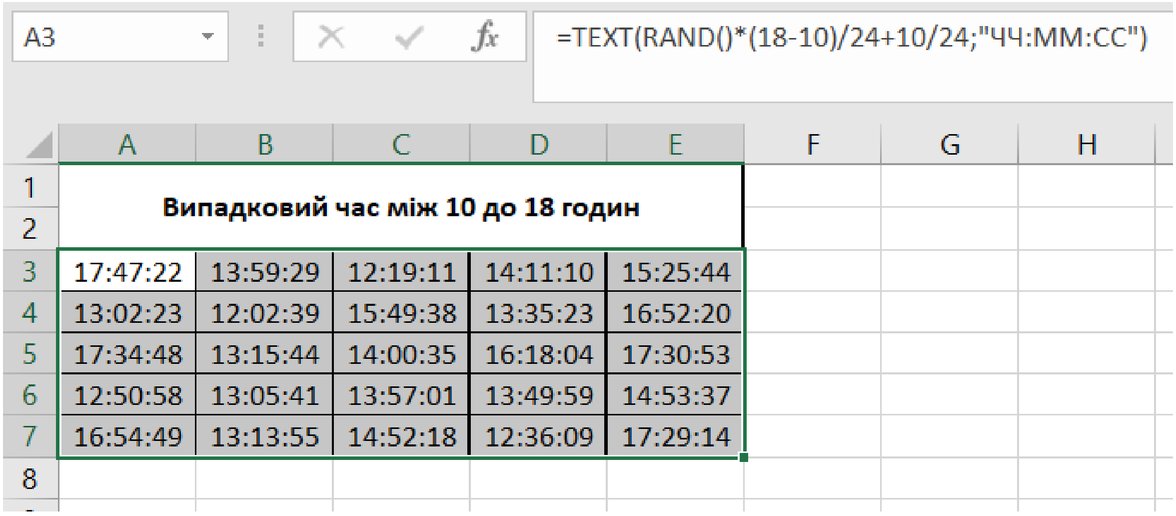
Фрагмент робочого листа Microsoft Excel, який потрібно створити в процесі виконання завдання, наведено на рис. 2.5a, b, c.



A3 =RANDBETWEEN(DATE(2021; 5; 1); DATE(2021; 10; 15))

	A	B	C	D	E
1–2	випадкові дати				
3	26.08.2021	22.09.2021	08.06.2021	26.05.2021	19.09.2021
4	23.05.2021	25.09.2021	22.08.2021	25.08.2021	08.08.2021
5	10.06.2021	27.07.2021	22.09.2021	15.08.2021	17.09.2021
6	21.05.2021	02.08.2021	21.07.2021	21.05.2021	14.06.2021
7	28.08.2021	20.06.2021	20.05.2021	12.07.2021	04.07.2021

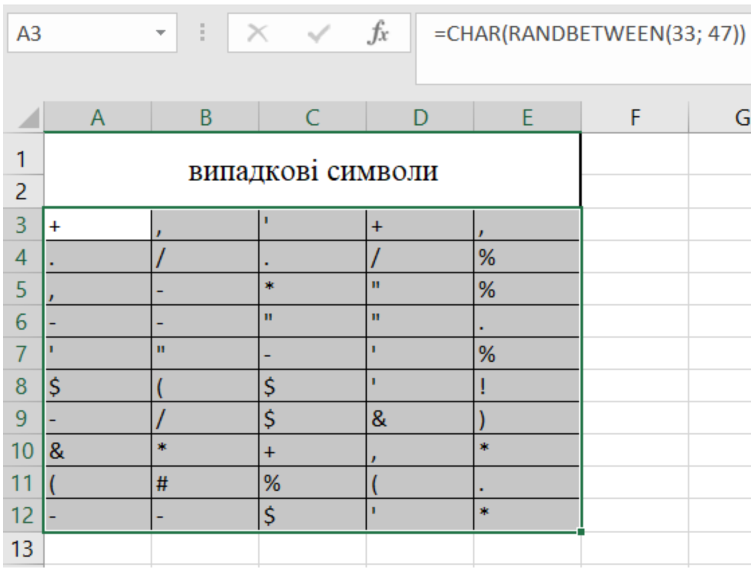
Рис. 2.5a. Лабораторна робота 2, завдання 3a



A3 =TEXT(RAND()*(18-10)/24+10/24;"ЧЧ:ММ:СС")

	A	B	C	D	E
1–2	**Випадковий час між 10 до 18 годин**				
3	17:47:22	13:59:29	12:19:11	14:11:10	15:25:44
4	13:02:23	12:02:39	15:49:38	13:35:23	16:52:20
5	17:34:48	13:15:44	14:00:35	16:18:04	17:30:53
6	12:50:58	13:05:41	13:57:01	13:49:59	14:53:37
7	16:54:49	13:13:55	14:52:18	12:36:09	17:29:14

Рис. 2.5b. Лабораторна робота 2, завдання 3b



A3 =CHAR(RANDBETWEEN(33; 47))

	A	B	C	D	E
1–2	випадкові символи				
3	+	,	'	+	,
4	.	/	.	/	%
5	,	-	*	"	%
6	-	-	"	"	.
7	'	"	-	'	%
8	$	(	$	'	!
9	-	/	$	&	)
10	&	*	+	,	*
11	(	#	%	(	.
12	-	-	$	'	*

Рис. 2.5c. Лабораторна робота 2, завдання 3c

ЛАБОРАТОРНА РОБОТА 3
ВИБІРКА, СТАТИСТИЧНІ ОЦІНКИ ПАРАМЕТРІВ ВИБІРКИ

Мета:

- засвоїти поняття генеральної сукупності та вибірки, розуміння їхніх ролей у статистичному аналізі;
- опанувати методи формування вибірки та її обробки, зокрема побудову варіаційного ряду за допомогою табличного процесора Microsoft Excel;
- научитися розраховувати та інтерпретувати статистичні оцінки параметрів вибірки, включаючи емпіричну функцію розподілу та багатокутник частот за допомогою табличного процесора Microsoft Excel;
- повторити основи візуалізації та аналізу даних для глибшого розуміння розподілу та характеристик вибірки.

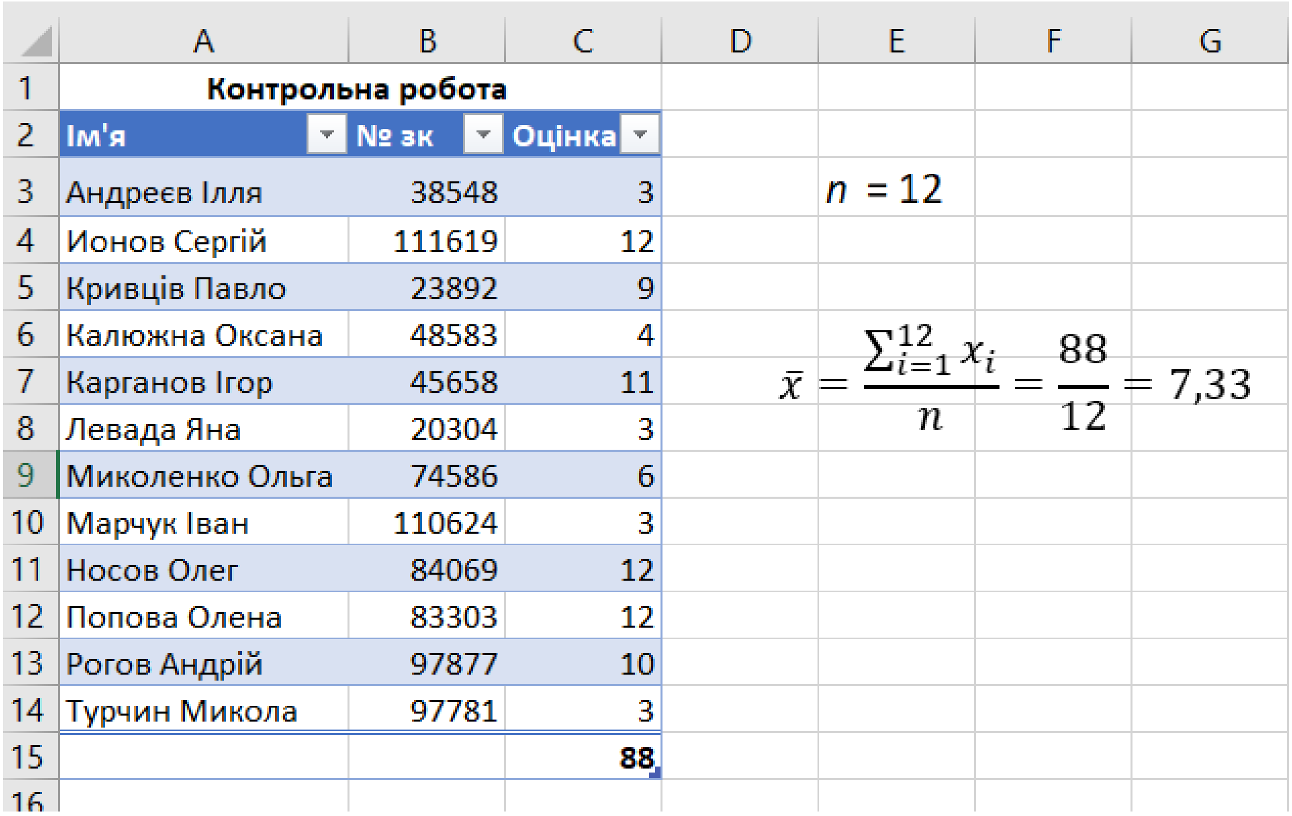
Хід роботи

1. Вивчити теоретичні поняття (див.п.1.2.)

2. Завдання 1.

Створити таблицю, яка містить оцінки 12 здобувачів освіти з предмету "Інформатика" за контрольну роботу за останній семестр. Вказати такі дані: прізвище та ім'я студента, номер індивідуального плану (залікової книжки), оцінка за роботу.

Фрагмент робочого листа Microsoft Excel, який потрібно створити в процесі виконання завдання, наведено на рис 2.6.



	A	B	C	D	E	F	G
1	**Контрольна робота**						
2	**Ім'я**	**№ зк**	**Оцінка**				
3	Андреєв Ілля	38548	3		n = 12		
4	Ионов Сергій	111619	12				
5	Кривців Павло	23892	9				
6	Калюжна Оксана	48583	4				
7	Карганов Ігор	45658	11				
8	Левада Яна	20304	3				
9	Миколенко Ольга	74586	6				
10	Марчук Іван	110624	3				
11	Носов Олег	84069	12				
12	Попова Олена	83303	12				
13	Рогов Андрій	97877	10				
14	Турчин Микола	97781	3				
15			**88**				
16							

$$\bar{x} = \frac{\sum_{i=1}^{12} x_i}{n} = \frac{88}{12} = 7{,}33$$

Рис. 2.6. Лабораторна робота 3, завдання 1

3. Завдання 2.

Студентська група здала колоквіум з математичного аналізу з наступними результатами (див. табл. 2.3):

Таблиця 2.3

Результати групи

x_i	3	4	5	6	7	8	9	10	11	12
n_i	2	2	3	3	5	4	2	2	1	1

Визначити середню успішність групи. Результати представити у вигляді варіаційного ряду та таблиці емпіричних частот, дати наочне графічне представлення даних у вигляді гістограми емпіричних частот, полігону відносних частот та емпіричній функції розподілу.

Фрагмент робочого листа Microsoft Excel, який потрібно створити в процесі виконання завдання, наведено на рис 2.7.

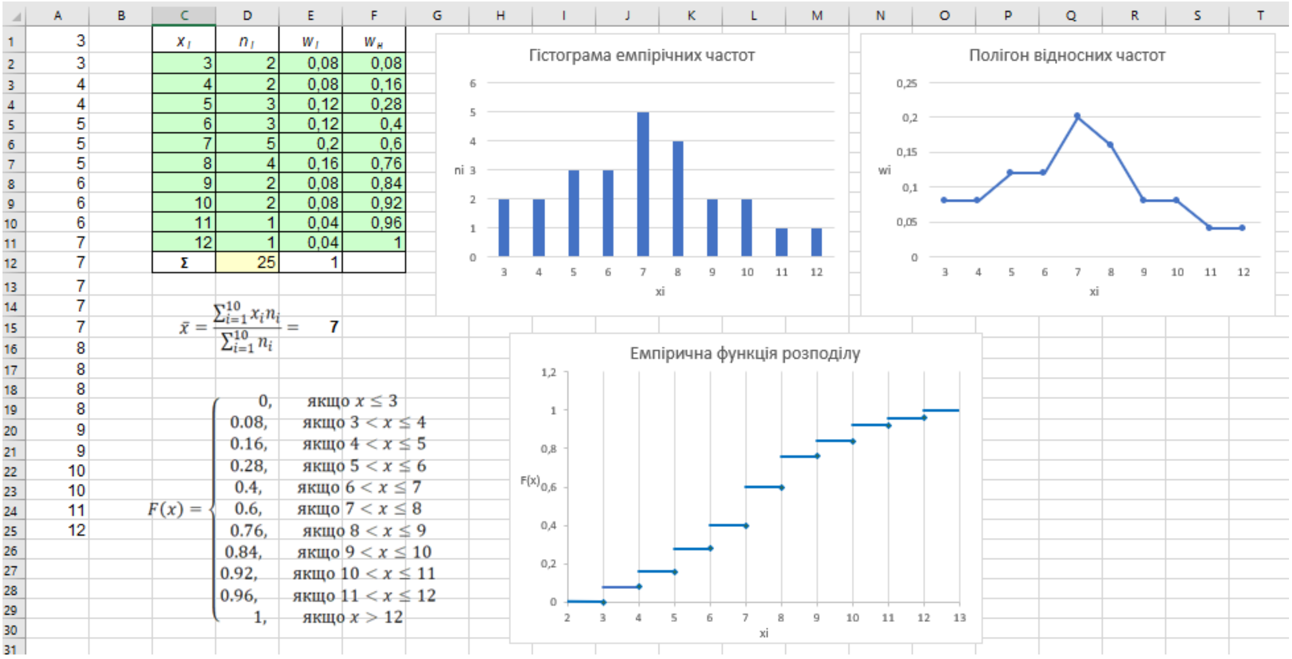


Рис. 2.7. Лабораторна робота 3, завдання 2

4. Завдання 3.

За результатами підсумкових оцінок за семестр здобувачів другого (магістерського) рівня вищої освіти для оцінки успішності був використаний вибірковий метод і отримані такі дані (див. табл. 2.4):

Таблиця 2.4

Підсумкові оцінки за семестр

75	76	87
61	99	98
70	60	83
60	82	83
74	71	91
68	96	63
63	85	58
75	89	72
70	80	75
75	80	67

Скласти варіаційний ряд розподілу, побудувати гістограму частот, гістограму та полігон відносних частот, та емпіричну функцію розподілу.

Фрагмент робочого листа Microsoft Excel, який потрібно створити в процесі виконання завдання, наведено на рис. 2.8.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	59		59	68	76	85	96						
2	60		60	70	80	87	98		75	76	87		
3	60		60	70	80	89	99		61	99	98		
4	61		61	71	82	91			70	60	83		
5	63		63	72	83				60	82	83		
6	63		63	74	83				74	71	91		
7	67		67	75					68	96	63		
8	68			75					63	85	58		
9	70			75					75	89	72		
10	70			75					70	80	75		
11	71								75	80	67		
12	72												
13	74												
14	75												
15	75		**Мінімальне значення**			X_{min} =	59						
16	75		**Максимальне значення**			X_{max} =	99						
17	75		**Розмах варіації**	$R = X_{max} - X_{min} =$			40						
18	76		**Оптимальна кількість інтервалів**				K =	5,906997	5				
19	80		**Довжина інтервала**			h =	8						
20	80				**Середини**		**Щільність**	**Відносна**	**Щільність**	**Накопич.**			
21	82		**Мережі інтервалів**		**інтервалів**	**Частота**	**частот**	**частота**	**від.частоти**	**від.частоти**		-10	0
22	83		початок	кінець	x_i	n_i	n_i/h	w_i	w_i/h	$w_н$		59	0
23	83		59	67	63	7	0,88	0,23	0,03	0,23		67	0,23
24	85		67	75	71	10	1,25	0,33	0,04	0,57		75	0,57
25	87		75	83	79	6	0,75	0,20	0,03	0,77		83	0,77
26	89		83	91	87	4	0,50	0,13	0,02	0,90		91	0,90
27	91		91	99	95	3	0,38	0,10	0,01	1,00		99	1,00
28	96		**Суми:**			**30**		**1**				120	1
29	98												
30	99												

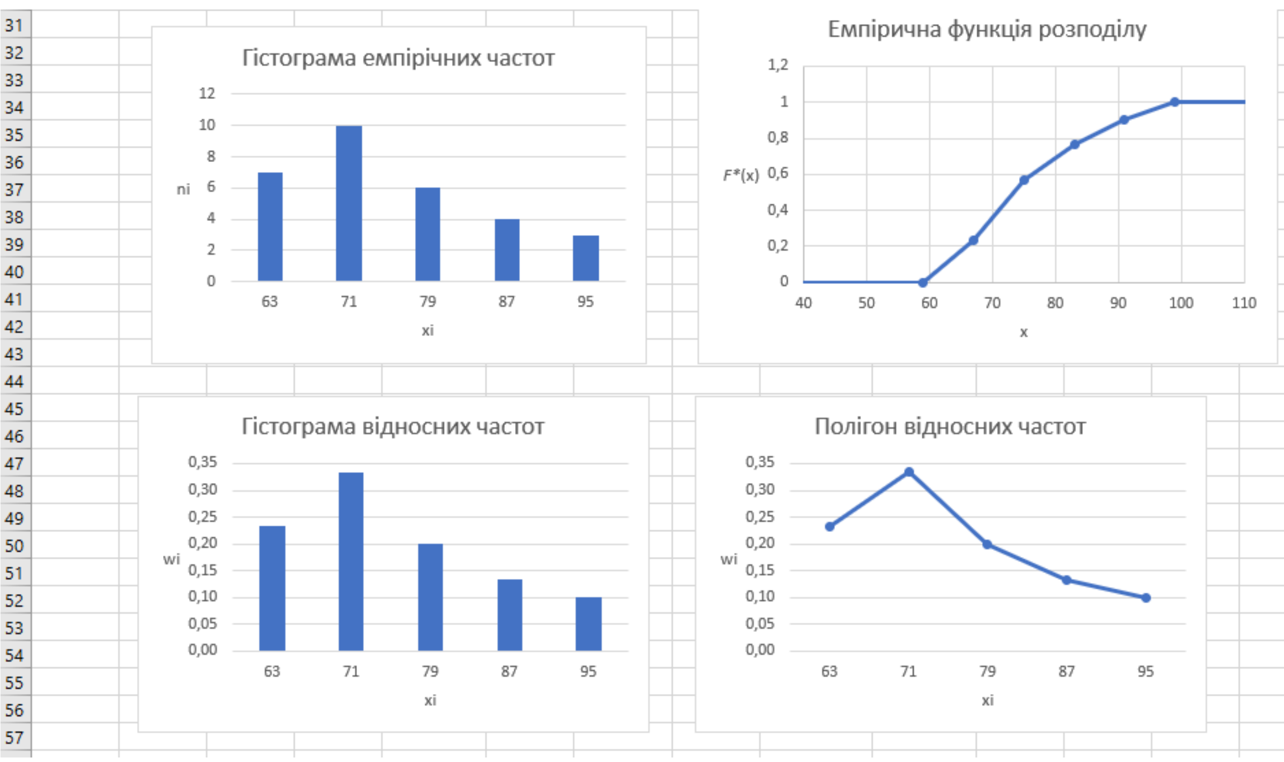


Рис. 2.8. Лабораторна робота 3, завдання 3

ЛАБОРАТОРНА РОБОТА 4
ОБРОБКА ВИБІРОК ДАНИХ

Мета:

- засвоїти методи обробки вибіркових даних, включаючи розрахунок основних статистичних параметрів;
- опанувати техніки визначення середнього арифметичного, дисперсії, стандартного відхилення, медіани, моди, а також емпіричного моменту за допомогою табличного процесора Microsoft Excel;
- научитися розраховувати та інтерпретувати коефіцієнти асиметрії і ексцесу для оцінки розподілу даних за допомогою табличного процесора Microsoft Excel;
- повторити принципи аналізу та інтерпретації вибіркових даних для виявлення закономірностей та аномалій.

Хід роботи

1. Вивчити теоретичні поняття (див.п.1.2.)

2. Завдання 1.

За результатами підсумкових оцінок за семестр здобувачів вищої освіти другого магістерського рівня одного із вузів для оцінки успішності був використаний вибірковий метод і отримані такі дані (див. табл. 2.4).

Знайти середнє значення, моду та медіану.

Фрагмент робочого листа Microsoft Excel, який потрібно створити в процесі виконання завдання, наведено на рис. 2.9.

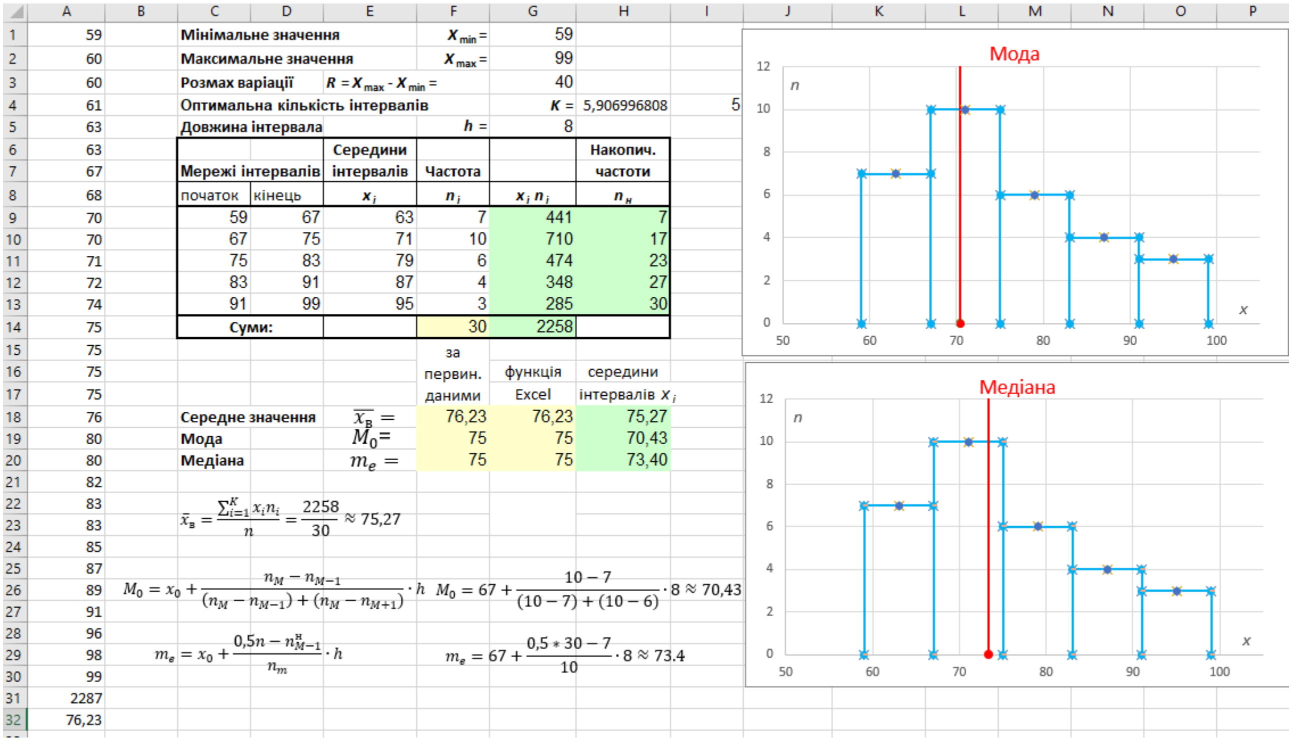


Рис. 2.9. Лабораторна робота 4, завдання 1

3. Завдання 2.

В результаті 10 незалежних вимірів деякої величини, виконаних з однаковою точністю, отримані дослідні дані, які представлені в таблиці 2.5:

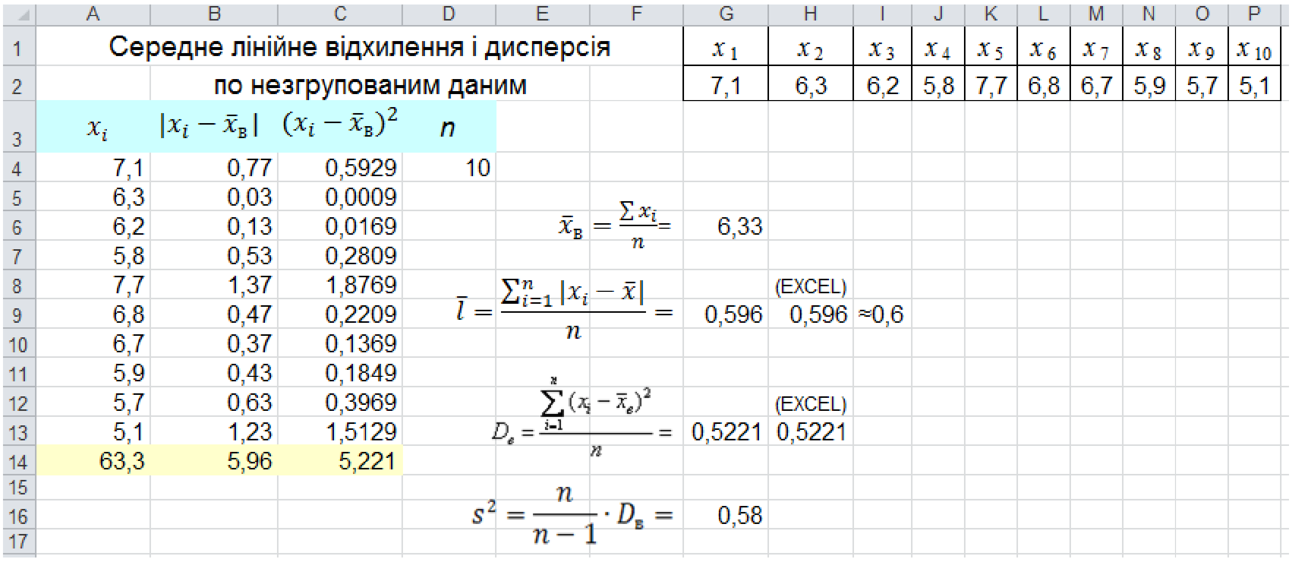
Таблиця 2.5

Виміри величини

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
7,1	6,3	6,2	5,8	7,7	6,8	6,7	5,9	5,7	5,1

Обчислити середнє лінійне відхилення, вибіркову дисперсію, виправлену вибіркову дисперсію.

Фрагмент робочого листа Microsoft Excel, який потрібно створити в процесі виконання завдання, наведено на рис. 2.10.



Середнє лінійне відхилення і дисперсія по незгрупованим даним

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
7,1	6,3	6,2	5,8	7,7	6,8	6,7	5,9	5,7	5,1

x_i	$\lvert x_i - \bar{x}_в \rvert$	$(x_i - \bar{x}_в)^2$	n				
7,1	0,77	0,5929	10				
6,3	0,03	0,0009					
6,2	0,13	0,0169		$\bar{x}_в = \frac{\sum x_i}{n} =$	6,33		
5,8	0,53	0,2809					
7,7	1,37	1,8769				(EXCEL)	
6,8	0,47	0,2209		$\bar{l} = \frac{\sum_{i=1}^{n} \lvert x_i - \bar{x} \rvert}{n} =$	0,596	0,596	≈0,6
6,7	0,37	0,1369					
5,9	0,43	0,1849					
5,7	0,63	0,3969				(EXCEL)	
5,1	1,23	1,5129		$D_в = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_в)^2}{n} =$	0,5221	0,5221	
63,3	5,96	5,221					
				$s^2 = \frac{n}{n-1} \cdot D_в =$	0,58		

Рис. 2.10. Лабораторна робота 4, завдання 2

4. Завдання 3.

З метою вивчення успішності студентів у навчальних закладах проведено вибіркове дослідження, в результаті якого отримано такі дані про оцінки студентів у межах одного предмету «Математика» (див. табл. 2.6):

Таблиця 2.6

Результати дослідження

Бали	**Кількість студентів**
до 40	32
от 40 до 60	56
от 60 до 80	120
от 80 до 100	104
понад 100	88
Разом:	**400**

Обчислити вибіркову дисперсію та середнє відхилення, оцінити відповідні показники генеральної сукупності. Визначити коефіцієнт варіації.

Фрагмент робочого листа Microsoft Excel, який потрібно створити в процесі виконання завдання, наведено на рис. 2.11.

Интервалы		x_i	n_i	$x_i n_i$	$x_i^2 n_i$
20	40	0,3	32	9,6	2,88
40	60	0,5	56	28	14
60	80	0,7	120	84	58,8
80	100	0,9	104	93,6	84,24
100	120	1,1	88	96,8	106,48
Суммы:			400	312	266,4

Бали	Кількість студентів
до 40	32
от 40 до 60	56
от 60 до 80	120
от 80 до 100	104
понад 100	88
Разом:	**400**

$$\bar{x}_в = \frac{\sum x_i n_i}{n} = \frac{312}{400} = 0{,}78 \text{ або 78 балів} \qquad (\bar{x}_в)^2 = 0{,}6084$$

$$D_в = \frac{\sum x_i^2 n_i}{n} - (\bar{x}_в)^2 = 0{,}0576 \qquad s^2 = \frac{n}{n-1} \cdot D_в = 0{,}057744$$

$$\sigma_в = 0{,}24 \qquad s = 0{,}2403$$

$$V = \frac{\sigma_в}{\bar{x}_в} \cdot 100\% = 30{,}77\ \%$$

Рис. 2.11. Лабораторна робота 4, завдання 3

5. Завдання 4.

Вибіркова перевірка результатів підсумкової контрольної роботи учнів 9 класів з математики дала наступні результати (див. табл. 2.7):

Таблиця 2.7

Результатів підсумкової контрольної роботи.

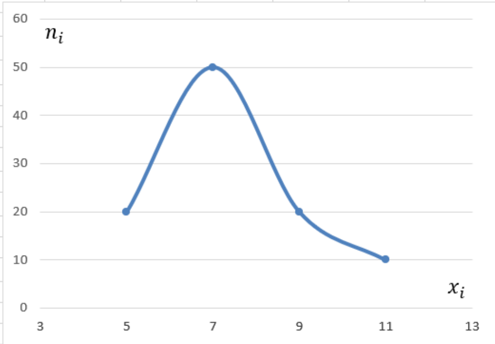
Оцінка x_i n_i	4 - 6	6 - 8	8 - 10	10 - 12
Кількість учнів,	20	50	20	10

Обчислити коефіцієнти асиметрії та ексцесу.

Фрагмент робочого листа Microsoft Excel, який потрібно створити в процесі виконання завдання, наведено на рис. 2.12.

Интервали		x_i	n_i	$x_i n_i$	$(x_i-\bar{x}_в)^2 n_i$	$(x_i-\bar{x}_в)^3 n_i$	$(x_i-\bar{x}_в)^4 n_i$
4	6	5	20	100	115,2	-276,48	663,552
6	8	7	50	350	8	-3,2	1,28
8	10	9	20	180	51,2	81,92	131,072
10	12	11	10	110	129,6	466,56	1679,616
Сумми:			100	740	304	268,8	2475,52

Оцінка x_i	4 - 6	6 - 8	8 - 10	10 - 12
Кількість учнів n_i	20	50	20	10



$\bar{x}_в = \frac{\sum x_i n_i}{n} =$ 7,4

$m_3 = \frac{\sum (x_i - \bar{x}_в)^3 \cdot n_i}{n} =$ 2,688

$\sigma_в = \sqrt{D_в} =$ 1,7436

$D_в = \frac{\sum (x_i - \bar{x}_в)^2 \cdot n_i}{n} =$ 3,04

$m_4 = \frac{\sum (x_i - \bar{x}_в)^4 \cdot n_i}{n} =$ 24,7552

$A_s = \frac{m_3}{\sigma_в^3} =$ 0,5071

$E_k = \frac{m_4}{\sigma_в^4} - 3 =$ -0,3213

$A_s > 0{,}5$ розподіл скошено вправо, асиметрія істотна.

$E_k < 0$ емпіричний розподіл є нижчим та пологішим щодо «еталонного» нормального розподілу з параметрами $a = \bar{x}_в$, $\sigma = \sigma_в$

Рис. 2.12. Лабораторна робота 4, завдання 4

ЛАБОРАТОРНА РОБОТА 5
ОСНОВИ КОРЕЛЯЦІЙНОГО АНАЛІЗУ

Мета:

- засвоїти поняття кореляції та основні характеристики кореляційних зв'язків, включаючи напрямок, потужність, ймовірність та форму взаємозв'язків між змінними;
- опанувати методи розрахунку параметрів кореляції, такі як коефіцієнт кореляції, що дозволяють кількісно оцінити силу і напрямок зв'язку між змінними;
- навчитися визначати типи кореляційних зв'язків та інтерпретувати результати аналізу для прийняття обґрунтованих рішень у дослідницькій та професійній діяльності;
- розвинути практичні навички використання табличного процесора Microsoft Excel для виконання кореляційного аналізу та обробки результатів.

Хід роботи

1. Вивчити теоретичні поняття (див.п.1.2.)

2. Завдання 1.

Наявні вибіркові дані про успішність 30-ти студентів (підсумковий бал) та кількість відвідувань ними занять (кількість відвіданих лекцій та практичних занять) за семестр (див. табл. 2.8):

Таблиця 2.8

Успішність та кількість відвідувань занять

ПІП	Лекції та ПР	Бали	ПІП	Лекції та ПР	Бали
1	16	65	16	15	52
2	18	78	17	15	62
3	12	41	18	16	69
4	14	54	19	17	85
5	16	66	20	16	70
6	18	80	21	16	71
7	13	45	22	15	64
8	14	57	23	17	72
9	16	67	24	19	88
10	18	81	25	16	73
11	18	92	26	16	74
12	13	48	27	19	96
13	16	59	28	16	75
14	16	68	29	20	101
15	17	83	30	17	76

Для наведених даних:

1. Методом аналітичного угруповання встановіть наявність та характер кореляційного зв'язку між кількістю відвідувань занять та підсумковим балом студента. Оформіть результати у вигляді групової та аналітичної таблиці.

2. Виміряйте тісноту кореляційного зв'язку між кількістю відвідувань занять та підсумковим балом емпіричним кореляційним відношенням. Зробіть висновки.

Фрагмент робочого листа Microsoft Excel, який потрібно створити в процесі виконання завдання, наведено на рис. 2.13 і 2.14.

	A	B	C
1	ПІП	Лекції та ПР	Бали
2	3	12	41
3	7	13	45
4	12	13	48
5	4	14	54
6	8	14	57
7	16	15	52
8	17	15	62
9	22	15	64
10	1	16	65
11	5	16	66
12	9	16	67
13	13	16	59
14	14	16	68
15	18	16	69
16	20	16	70
17	21	16	71
18	25	16	73
19	26	16	74
20	28	16	75
21	15	17	83
22	19	17	85
23	23	17	72
24	30	17	76
25	2	18	78
26	6	18	80
27	10	18	81
28	11	18	92
29	24	19	88
30	27	19	96
31	29	20	101

	E	F	G	H	I	J	K	L	M	N	O	P
1		Перша група		Друга група		Третя група		Четверта група		П'ята група		
2		12	41	14	54	16	65	17	83	19	88	
3		13	45	14	57	16	66	17	85	19	96	
4		13	48	15	52	16	67	17	72	20	101	
5				15	62	16	59	17	76			
6				15	64	16	68	18	78			
7						16	69	18	80			
8						16	70	18	81			
9						16	71	18	92			
10						16	73					
11						16	74					
12						16	75					
14	$n_i =$	3		5		11		8		3		30
16	Разом по групам		134		289		757		647		285	
18	$\bar{y}_i =$		44,667		57,8		68,818		80,875		95	
19	*(середне значення по групам)*											
20						Аналітична таблиця						
21						№ групи	Інтервали факторної ознаки		n_i	$\bar{y}_i$		
23	Мінімальне значення:			12		1	12	13,6	3	44,667		
24	Максімальне значення:			20		2	13,6	15,2	5	57,8		
25	Розмах варіації:			8		3	15,2	16,8	11	68,818		
26	Кількість інтервалів:			5,907	5	4	16,8	18,4	8	80,875		
27	Довжина інтервалу:			1,6		5	18,4	20	3	95		

	Q	R	S	T	U	V
1	ПІП	Лекції та ПР	Бали	ПІП	Лекції та ПР	Бали
2	1	16	65	16	15	52
3	2	18	78	17	15	62
4	3	12	41	18	16	69
5	4	14	54	19	17	85
6	5	16	66	20	16	70
7	6	18	80	21	16	71
8	7	13	45	22	15	64
9	8	14	57	23	17	72
10	9	16	67	24	19	88
11	10	18	81	25	16	73
12	11	18	92	26	16	74
13	12	13	48	27	19	96
14	13	16	59	28	16	75
15	14	16	68	29	20	101
16	15	17	83	30	17	76
19	зі збільшенням кількості відвіданих занять спостерігається збільшення середнього підсумкового балу					
23	існує пряма кореляційна залежність					

Рис. 2.13. Лабораторна робота 5, завдання 1.1

	B	C	D
1	Лекції та ПР	Бали	$(y_i-\bar{y})^2$
2	12	41	864,36
3	13	45	645,16
4	13	48	501,76
5	14	54	268,96
6	14	57	179,56
7	15	52	338,56
8	15	62	70,56
9	15	64	40,96
10	16	65	29,16
11	16	66	19,36
12	16	67	11,56
13	16	59	129,96
14	16	68	5,76
15	16	69	1,96
16	16	70	0,16
17	16	71	0,36
18	16	73	6,76
19	16	74	12,96
20	16	75	21,16
21	17	83	158,76
22	17	85	213,16
23	17	72	2,56
24	17	76	31,36
25	18	78	57,76
26	18	80	92,16
27	18	81	112,36
28	18	92	466,56
29	19	88	309,76
30	19	96	655,36
31	20	101	936,36
32		2112	6185,2

	E	F	G	H	I	J	K
1			y_i	y_i	y_i	y_i	y_i
2			41	54	65	83	88
3			45	57	66	85	96
4			48	52	67	72	101
5				62	59	76	
6				64	68	78	
7					69	80	
8					70	81	
9					71	92	
10					73		
11					74		
12					75		
14	$n_i =$		3	5	11	8	3
16	Разом по групам		134	289	757	647	285
18	$\bar{y}_i =$		44,667	57,8	68,818	80,875	95
19	*(середне значення по групам)*						
21	$(\bar{y}_i-\bar{y})^2\cdot n_i$		1986,6	793,8	27,524	877,8	1815,5
26			$n =$	30			
28			$\bar{y}=\frac{\sum y}{n}=$	70,4			
29				*(спільна середня)*			$\eta^2=\frac{\delta_y^2}{\sigma_y^2}=$
31	$\sigma_y^2=\frac{\sum_{i=1}^{30}(y_i-\bar{y})^2}{n}=$			206,17			
32				*(спільна дисперсія)*			

	L	M	N	O	P	Q
1		$(y_i-\bar{y}_1)^2$	$(y_i-\bar{y}_2)^2$	$(y_i-\bar{y}_3)^2$	$(y_i-\bar{y}_4)^2$	$(y_i-\bar{y}_5)^2$
2		13,444444	14,44	14,578512	4,515625	49
3		0,1111111	0,64	7,9421488	17,015625	1
4		11,111111	33,64	3,3057851	78,765625	36
5			17,64	96,396694	23,765625	
6			38,44	0,6694215	8,265625	
7				0,0330579	0,765625	
8				1,3966942	0,015625	
9				4,7603306	123,76563	
10				17,487603		
11				26,85124		
12				38,214876		
16	$\sigma_i^2\cdot n_i =$	24,666667	104,8	211,63636	256,875	86
18	$\sigma_i^2 =$	8,2222222	20,96	19,239669	32,109375	28,666667
19		*(групові дисперсії)*				
22	$\overline{\sigma^2}=\frac{\sum\sigma_i^2\cdot n_i}{n}=$		22,799268	*(внутрішньогрупова дисперсія)*		
25	$\delta_y^2=\frac{\sum_{i=1}^{5}(\bar{y}_i-\bar{y})^2\cdot n_i}{n}=$		183,37407	*(міжгрупова дисперсія)*		
27			*Правило складання дисперсій*			
28			$\overline{\sigma^2}+\delta_y^2=$	206,17333	$=\sigma_y^2$	
29	0,889417					
30				підсумковий бал дуже сильно залежить від кількості відвіданих занять		
31	$\eta=\sqrt{\frac{\delta_y^2}{\sigma_y^2}}=\frac{\delta_y}{\sigma_y}\approx$		0,94			

Рис. 2.14. Лабораторна робота 5, завдання 1.2

3. Завдання 2.

Є вибіркові дані щодо прогулів занять за певний період та успішності навчання для 8-ми студентів: X – кількість прогулів за певний період часу та Y – сумарна успішність за період (див. табл. 2.9):

Таблиця 2.9

Прогули і успішність студентів

X	12	9	8	14	15	11	10	15
Y	42	107	100	60	78	79	90	54

На основі наявних даних:

1) висловити припущення про наявність та напрямок *кореляційної залежності ознаки-результату* Y від *ознаки-фактора* X;

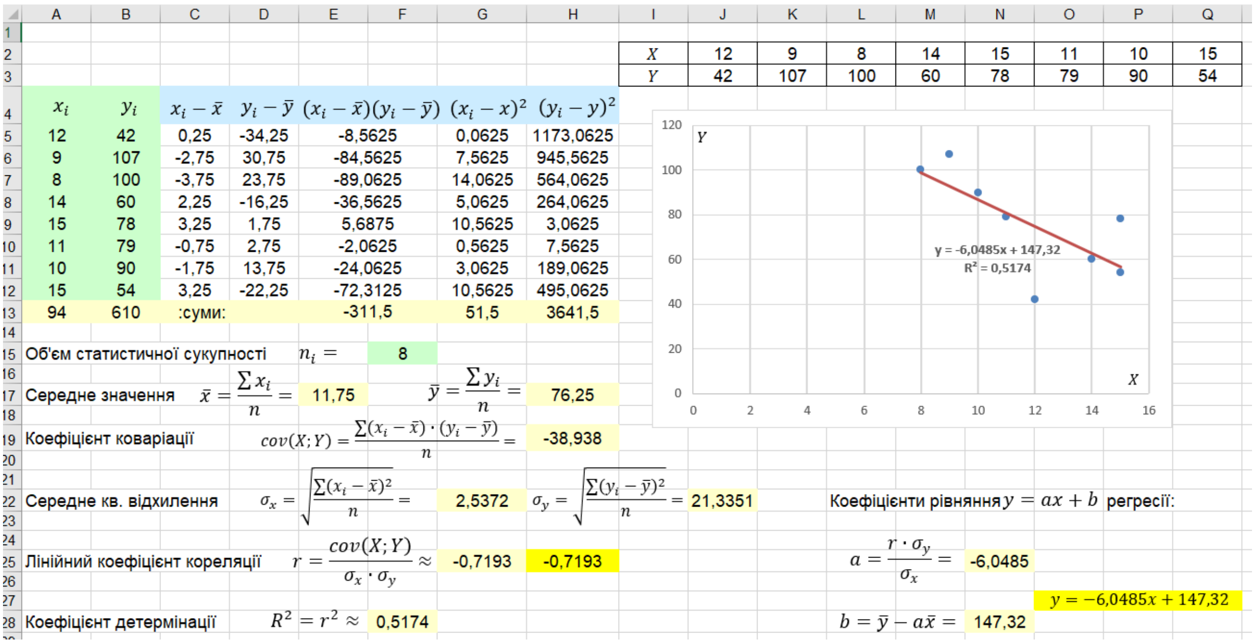
2) побудувати та проаналізувати *діаграму розсіювання*, зробити висновок про *форму* залежності;

3) обчислити *лінійний коефіцієнт кореляції*, зробити висновок;

4) обчислити *коефіцієнт детермінації*, зробити висновок;

5) визначити рівняння *лінійної регресії* Y на X, здійснити креслення.

Фрагмент робочого листа Microsoft Excel, який потрібно створити в процесі виконання завдання, наведено на рис. 2.15



x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-x)^2$	$(y_i-y)^2$
12	42	0,25	-34,25	-8,5625	0,0625	1173,0625
9	107	-2,75	30,75	-84,5625	7,5625	945,5625
8	100	-3,75	23,75	-89,0625	14,0625	564,0625
14	60	2,25	-16,25	-36,5625	5,0625	264,0625
15	78	3,25	1,75	5,6875	10,5625	3,0625
11	79	-0,75	2,75	-2,0625	0,5625	7,5625
10	90	-1,75	13,75	-24,0625	3,0625	189,0625
15	54	3,25	-22,25	-72,3125	10,5625	495,0625
94	610	:суми:		-311,5	51,5	3641,5

Об'єм статистичної сукупності $n_i=$ 8

Середнє значення $\bar{x}=\frac{\sum x_i}{n}=$ 11,75 $\bar{y}=\frac{\sum y_i}{n}=$ 76,25

Коефіцієнт коваріації $cov(X;Y)=\frac{\sum(x_i-\bar{x})\cdot(y_i-\bar{y})}{n}=$ -38,938

Середнє кв. відхилення $\sigma_x=\sqrt{\frac{\sum(x_i-\bar{x})^2}{n}}=$ 2,5372 $\sigma_y=\sqrt{\frac{\sum(y_i-\bar{y})^2}{n}}=$ 21,3351

Лінійний коефіцієнт кореляції $r=\frac{cov(X;Y)}{\sigma_x\cdot\sigma_y}\approx$ -0,7193 -0,7193

Коефіцієнт детермінації $R^2=r^2\approx$ 0,5174

Коефіцієнти рівняння $y=ax+b$ регресії:

$a=\frac{r\cdot\sigma_y}{\sigma_x}=$ -6,0485

$y=-6{,}0485x+147{,}32$

$b=\bar{y}-a\bar{x}=$ 147,32

Рис. 2.15. Лабораторна робота 5, завдання 2

ЛАБОРАТОРНА РОБОТА 6
МЕТОД ВИБІРКИ

Мета:

- засвоїти основні поняття вибіркового методу, зокрема визначення вибірки та її значення у статистичних дослідженнях;
- опанувати розрахунок похибок вибіркових параметрів, включаючи помилку середнього арифметичного, середньоквадратичного відхилення, коефіцієнта варіації, індикатора асиметрії та коефіцієнта ексцесу;
- навчитися визначати критерії надійності статистичних оцінок та будувати довірчі інтервали для параметрів вибірки, що дозволить оцінювати точність і надійність отриманих результатів;
- розвинути навички застосування методів вибірки для оцінки параметрів генеральної сукупності та інтерпретації результатів у контексті практичних задач і прийняття обґрунтованих рішень.

Хід роботи

1. Вивчити теоретичні поняття (див.п.1.2.)

2. Завдання 1.

Розрахувати модельні характеристики генеральної сукупності за вибіркою**:**

11,3; 11,4; 11,4; 11,5; 11,5; 11,6; 11,6; 11,7; 11,7; 11,8; 11,8; 11,9; 11,9; 11,9; 11,9; 12,00; 12,00;12,10; 12,00; 12,00; 12,00; 12,20; 12,20; 12,20; 12,20; 12,20; 12,20; 12,30; 12,30; 12,30; 12,40; 12,40; 12,40; 12,50; 12,50; 12,50; 12,50; 12,60; 12,70; 12,70; 12,80; 12,90; 12,90; 13,00; 13,00; 13,00; 13,10; 13,20; 13,20; 13,40; 13,80; 14,00; 14,00; 14,10; 14,20; 14,17; 15,00; 15,20; 15,20.

Для наведених даних розрахувати доступні функції:

- похибка середнього арифметичного;
- помилка стандартного відхилення;
- похибка коефіцієнта варіації;

- похибки показників асиметрії та ексцесу;
- довірчий інтервал.

Фрагмент робочого листа Microsoft Excel, який потрібно створити в процесі виконання завдання, наведено на рис 2.16.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	x			Мінімальне значення			$x_{min} =$	11,3					
2	11,3			Максимальне значення			$x_{max} =$	15,2					
3	11,4			Розмах варіації		$R = x_{max} - x_{min} =$		3,9					
4	11,4			Оптимальна кількість інтервалів				$K =$	6,88	6			
5	11,5			Довжина інтервала			$h =$	0,65					
6	11,5												
7	11,6					Середини							
8	11,6			Мережі інтервалів		інтервалів	Частота						
9	11,7			початок	кінець	x_i	n_i	$x_i n_i$	$(x_i - \bar{x}_в)^2 n_i$	$(x_i - \bar{x}_в)^3 n_i$	$(x_i - \bar{x}_в)^4 n_i$		
10	11,7			11,3	11,95	11,63	15	174,38	13,47	-12,76	12,09		
11	11,8			11,95	12,6	12,28	23	282,33	2,04	-0,61	0,18		
12	11,8			12,6	13,25	12,93	11	142,18	1,37	0,48	0,17		
13	11,9			13,25	13,9	13,58	2	27,15	2,01	2,02	2,02		
14	11,9			13,9	14,55	14,23	5	71,13	13,65	22,56	37,29		
15	11,9			14,55	15,2	14,88	3	44,63	15,91	36,62	84,32		
16	11,9			Суми:			59	741,78	48,44	48,32	136,07		
17	12												
18	12												
19	12				$\bar{x}_в = \frac{\sum x_i n_i}{n} =$	12,57		$m_3 = \frac{\sum (x_i - \bar{x}_в)^3 \cdot n_i}{n} =$		0,82		$\alpha =$	0,05
20	12												
21	12												
22	12,1			$D_в = \frac{\sum (x_i - \bar{x}_в)^2 \cdot n_i}{n} =$		0,82		$m_4 = \frac{\sum (x_i - \bar{x}_в)^4 \cdot n_i}{n} =$		2,31		$k = n - 1 =$	58
23	12,2												
24	12,2												
25	12,2				$\sigma_в = \sqrt{D_в} =$	0,91			$A_s = \frac{m_3}{\sigma_в^3} =$	1,10	$t(\alpha, k) = t(0{,}05; 58) \approx$		2
26	12,2												
27	12,2												
28	12,2				$s^2 = \frac{n}{n-1} \cdot D_в =$	0,84			$E_k = \frac{m_4}{\sigma_в^4} - 3 =$	0,42		$m = \frac{\sigma}{\sqrt{n}} =$	0,12
29	12,3												
30	12,3												
31	12,3				$s = \sqrt{s^2} =$	0,91			$\bar{x}_в - m \cdot t =$	12,34	$< \bar{X} <$	12,81	$= \bar{x}_в + m$
32	12,4												
33	12,4												
34	12,4				$V = \frac{\sigma_в}{\bar{x}_в} \cdot 100\% =$	7,21	%					$m_s = \frac{s}{\sqrt{2 \cdot n}} =$	0,08
35	12,5												
36	12,5												
37	12,5											$m_V = \frac{V}{\sqrt{2 \cdot n}} \cdot \sqrt{1 + 2\left(\frac{V}{100}\right)^2} =$	0,67
38	12,5												
39	12,6												
40	12,7											$m_{As} = \sqrt{\frac{6 \cdot n(n-1)}{(n-2) \cdot (n+1) \cdot (n+3)}} =$	0,31
41	12,7												
42	12,8												
43	12,9											$m_{Ek} = \sqrt{\frac{24}{n}} =$	0,64
44	12,9												

Рис. 2.16. Лабораторна робота 6, завдання 1

3. Завдання 2.

З метою визначення успішності отримання знань по предмету «Інформатика» була проведена загальна контрольна робота учнів 10 класу м. Харків. Проведено 5% вибіркове обстеження 100 учнів, відібраних у випадковому порядку, в результаті якого отримані такі дані (див. табл. 2.10):

Таблиця 2.10

Результати вибірки

Бали	**Кількість студентів**
до 4	10
от 4 до 6	25
от 6 до 8	40
от 8 до 10	20
понад 10	5
Разом:	**100**

З ймовірністю 0,9973 визначити *граничну помилку* вибірки та можливі межі, в яких очікується середня успішність в школах.

Фрагмент робочого листа Microsoft Excel, який потрібно створити в процесі виконання завдання, наведено на рис. 2.17.

Интервали		x_i	n_i	$x_i n_i$	$x_i^2 n_i$
2	4	3	10	30	90
4	6	5	25	125	625
6	8	7	40	280	1960
8	10	9	20	180	1620
10	12	11	5	55	605
Сумми:			100	670	4900

Бали	Кількість студентів
до 4	10
от 4 до 6	25
от 6 до 8	40
от 8 до 10	20
понад 10	5
Разом:	**100**

$$N = \frac{100\%}{5\%} \cdot n = 2000$$

$$\bar{x}_\text{в} = \frac{\sum x_i n_i}{n} = 6{,}7$$

$$\gamma = 0{,}9973$$

$$2\Phi(t) = \gamma \Longrightarrow \Phi(t) = \frac{\gamma}{2} = 0{,}49865$$

$$D_\text{в} = \frac{\sum x_i{}^2 n_i}{n} - (\bar{x}_\text{в})^2 = 4{,}11$$

$$t = 3$$

$$s^2 = \frac{n}{n-1} \cdot \sigma_\text{в}^2 = 4{,}15$$

$$\mu = \sqrt{\frac{s^2}{n}\left(1 - \frac{n}{N}\right)} = 0{,}19859354 \approx 0{,}20$$

$$\Delta = t \cdot \mu = 0{,}60$$

$$\bar{x}_\text{в} - \Delta < \bar{x}_\Gamma < \bar{x}_\text{в} + \Delta$$

$$6{,}1 < \bar{x}_\Gamma < 7{,}3$$

Рис. 2.17. Лабораторна робота 6, завдання 2

ЛАБОРАТОРНА РОБОТА 7
СТАТИСТИЧНІ ГІПОТЕЗИ ТА ЇХ ПЕРЕВІРКА

Мета:

- засвоїти поняття статистичної гіпотези, її роль у статистичних дослідженнях та значення в прийнятті рішень на основі даних. Засвоїти застосування основних статистичних критеріїв, таких як критерій Стьюдента та критерій Пірсона, для перевірки гіпотез і визначення значущості відмінностей або взаємозв'язків у даних;
- опанувати алгоритм перевірки статистичних гіпотез, включаючи етапи формулювання гіпотез, вибору відповідного статистичного критерію та інтерпретації результатів;
- навчитися розпізнавати та оцінювати помилки першого та другого роду при перевірці гіпотез, а також розуміти їх вплив на достовірність висновків;
- розвинути навички критичного аналізу та оцінки статистичних результатів.

Хід роботи

1. Вивчити теоретичні поняття (див.п.1.2.)

2. Завдання 1.

Наявні дані про оцінку знань з математики (у балах) для 9 студентів, які було отримано на початку навчання на 1-му курсі та після його закінчення (див. табл. 2.11):

Таблиця 2.11

Рівень знань студентів

x_i	76	71	57	49	70	69	26	65	59
y_i	81	85	52	52	70	63	33	83	62

(у 1-му рядку кількість балів на час вступу, у 2-му – після 1-го курсу)

На рівні значимості 0,05 встановити, значимо чи незначно поліпшилася підготовка студентів з математики, у припущенні, що кількість балів розподілено нормально.

Фрагмент робочого листа Microsoft Excel, який потрібно створити в процесі виконання завдання, наведено на рис. 2.18.

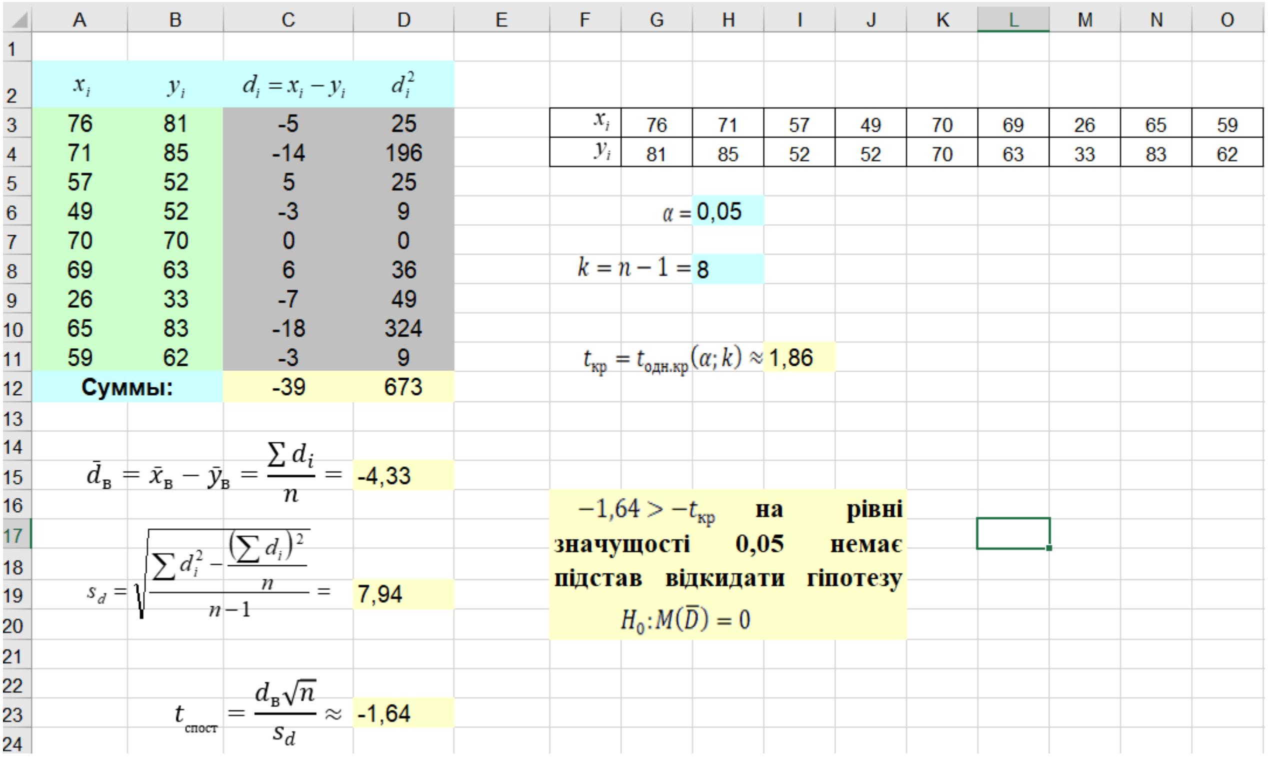


Рис. 2.18. Лабораторна робота 7, завдання 1

3. Завдання 2.

Наявні дані про те, що дві групи студентів-першокурсників написали контрольну роботу з математичного аналізу з наступними результатами (див. табл. 2.12):

Таблиця 2.12

Результати двох груп студентів

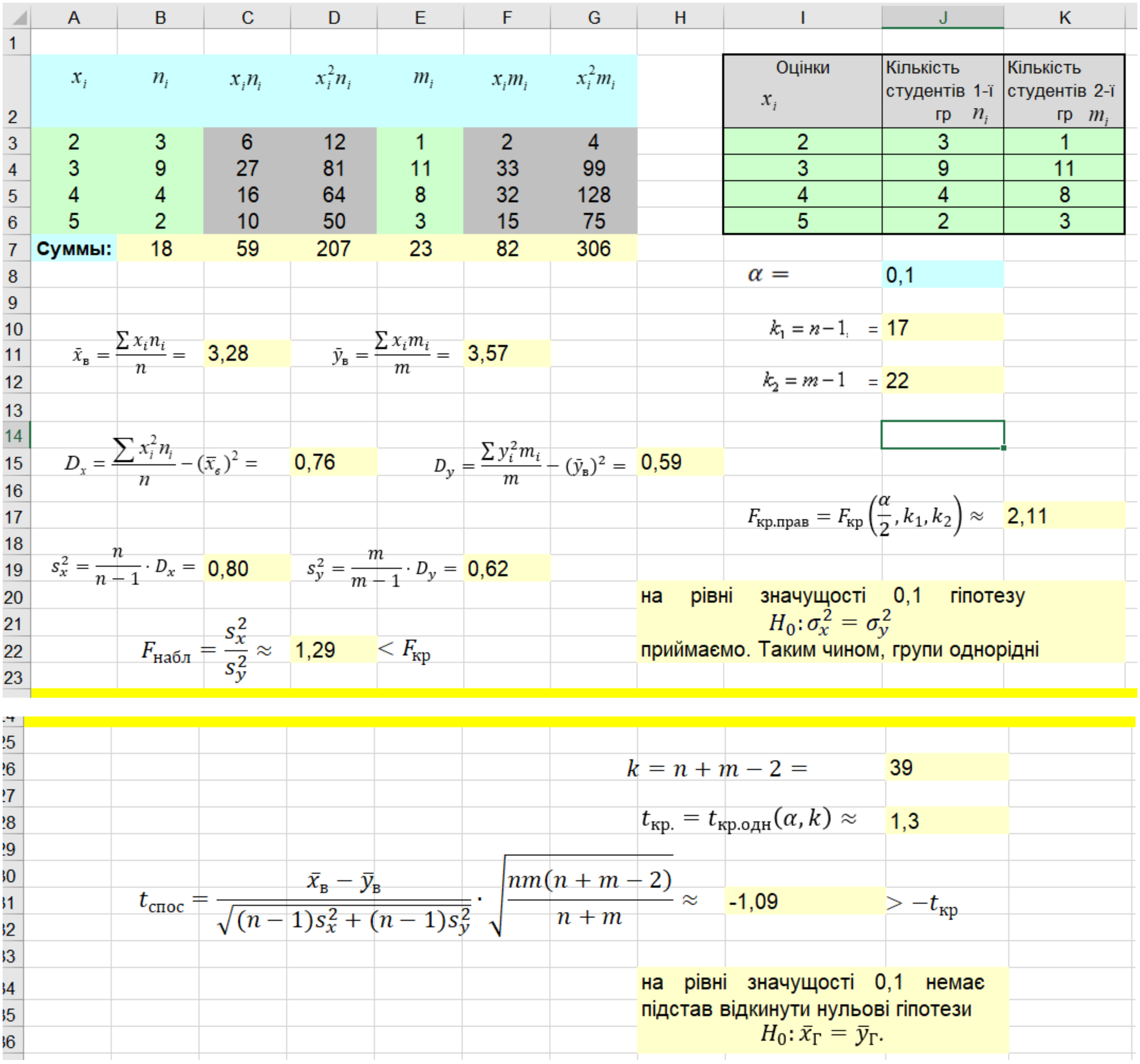
Оцінки x_i	Кількість студентів 1-ї гр n_i	Кількість студентів 2-ї гр m_i
2	3	1
3	9	11
4	4	8
5	2	3

Припускаючи, що успішність студентів розподілена нормально, на рівні значущості 0,1:

1) Перевірити гіпотезу $H_0: \sigma_x^2 = \sigma_y^2$ – у тому, що групи однорідні за складом (у плані співвідношення краще і гірше встигаючих студентів) проти конкуруючої гіпотези $H_1: \sigma_x^2 \neq \sigma_y^2$, і у разі однорідності груп.

2) Перевірити гіпотезу $H_0: \bar{x}_\Gamma = \bar{y}_\Gamma$ – про однакову успішність груп проти гіпотези у тому, що одна з груп слабша за успішністю.

Фрагмент робочого листа Microsoft Excel, який потрібно створити в процесі виконання завдання, наведено на рис. 2.19.



x_i	n_i	$x_i n_i$	$x_i^2 n_i$	m_i	$x_i m_i$	$x_i^2 m_i$
2	3	6	12	1	2	4
3	9	27	81	11	33	99
4	4	16	64	8	32	128
5	2	10	50	3	15	75
Суммы:	18	59	207	23	82	306

Оцінки x_i	Кількість студентів 1-ї гр n_i	Кількість студентів 2-ї гр m_i
2	3	1
3	9	11
4	4	8
5	2	3

$\alpha =$ 0,1

$k_1 = n-1$ = 17

$k_2 = m-1$ = 22

$\bar{x}_в = \dfrac{\sum x_i n_i}{n} =$ 3,28 $\quad \bar{y}_в = \dfrac{\sum x_i m_i}{m} =$ 3,57

$D_x = \dfrac{\sum x_i^2 n_i}{n} - (\bar{x}_в)^2 =$ 0,76 $\quad D_y = \dfrac{\sum y_i^2 m_i}{m} - (\bar{y}_в)^2 =$ 0,59

$F_{кр.прав} = F_{кр}\left(\dfrac{\alpha}{2}, k_1, k_2\right) \approx$ 2,11

$s_x^2 = \dfrac{n}{n-1} \cdot D_x =$ 0,80 $\quad s_y^2 = \dfrac{m}{m-1} \cdot D_y =$ 0,62

$F_{набл} = \dfrac{s_x^2}{s_y^2} \approx$ 1,29 $< F_{кр}$

на рівні значущості 0,1 гіпотезу $H_0: \sigma_x^2 = \sigma_y^2$ приймаємо. Таким чином, групи однорідні

$k = n + m - 2 =$ 39

$t_{кр.} = t_{кр.одн}(\alpha, k) \approx$ 1,3

$t_{спос} = \dfrac{\bar{x}_в - \bar{y}_в}{\sqrt{(n-1)s_x^2 + (n-1)s_y^2}} \cdot \sqrt{\dfrac{nm(n+m-2)}{n+m}} \approx$ -1,09 $> -t_{кр}$

на рівні значущості 0,1 немає підстав відкинути нульові гіпотези $H_0: \bar{x}_\Gamma = \bar{y}_\Gamma$.

Рис. 2.19. Лабораторна робота 7, завдання 2

4. Завдання 3.

За результатами вибіркового дослідження знайдено розподіл середніх оцінок за модульну контрольну роботу з математики студентів першого курсу закладу вищої освіти (див. табл. 2.13):

Таблиця 2.13

Середні оцінки студентів

Бали	**Кількість студентів**
7,5 - 10,5	2
10,5 - 13,5	6
13,5 - 16,5	10
16,5 - 19,5	17
19,5 - 22,5	33
22,5 - 25,5	11
25,5 - 28,5	9
28,5 - 31,5	7
31,5 - 34,5	5
Разом:	**100**

На рівні значимості 0,05 перевірити гіпотезу у тому, що генеральна сукупність (середній бал) розподілена нормально. Побудувати емпіричну гістограму та теоретичну криву.

Фрагмент робочого листа Microsoft Excel, який потрібно створити в процесі виконання завдання, наведено на рис. 2.20.

Як перевірити гіпотезу про нормальний розподіл генеральної сукупності за допомогою критерію погодження Пірсона

Знайдемо середини інтервалів x_i, вибіркову середню $\bar{x}_в = \frac{\sum x_i n_i}{n} =$ 21,21, серед. квадр. відхилення $\sigma_в = \sqrt{\frac{\sum x_i^2 n_i}{n} - (\bar{x}_в)^2} =$ 5,47

Знайдемо теоретичні частоти за формулою $n'_i = \frac{h \cdot n}{\sigma_в} \cdot f(z_i)$, де $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$, $z_i = \frac{x_i - \bar{x}_в}{\sigma_в}$

Інтервали		x_i	n_i	$x_i n_i$	$x_i^2 n_i$	z_i	$f(z_i)$	n'_i
7,5	10,5	9	2	18	162	-2,232	0,033	1,81
10,5	13,5	12	6	72	864	-1,6836	0,0967	5,30
13,5	16,5	15	10	150	2250	-1,1352	0,2095	11,49
16,5	19,5	18	17	306	5508	-0,5868	0,3358	18,42
19,5	22,5	21	33	693	14553	-0,0384	0,3986	21,86
22,5	25,5	24	11	264	6336	0,51	0,3503	19,21
25,5	28,5	27	9	243	6561	1,0584	0,2279	12,50
28,5	31,5	30	7	210	6300	1,6068	0,1097	6,02
31,5	34,5	33	5	165	5445	2,1552	0,0391	2,14
$h=$	3	$n=$	100	2121	47979			98,75
(довжина часткового		(об'єм вибірки)						

n_i	n'_i	$\frac{(n_i - n'_i)^2}{n'_i}$
8	7,12	0,11
10	11,49	0,1923
17	18,42	0,1092
33	21,86	5,6746
11	19,21	3,5087
9	12,50	0,9778
12	8,16	1,8053
	$\chi^2_{набл.} \approx$	12,378

Бали	Кількість
7,5 - 10,5	2
10,5 - 13,5	6
13,5 - 16,5	10
16,5 - 19,5	17
19,5 - 22,5	33
22,5 - 25,5	11
25,5 - 28,5	9
28,5 - 31,5	7
31,5 - 34,5	5
Разом:	**100**

На рівні значущості $\alpha = 0{,}05$

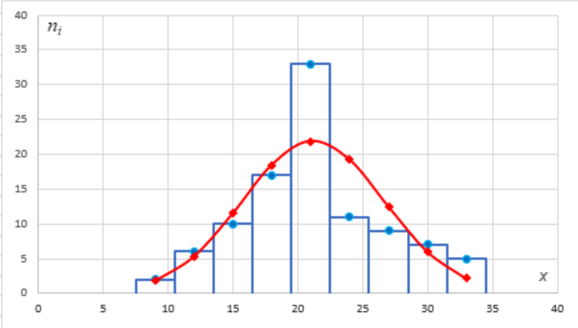
і кількісті ступенів свободи $k = m - r - 1$ = 4

знаходимо критичне значення $\chi^2_{кр.} = \chi^2_{кр.}(\alpha, k) \approx$ 9,4877

H_0 генеральна сукупність розподілена нормально.

H_1 генеральна сукупність **не** розподілена нормально.

$$\chi^2_{досл} \approx 12{,}378 > \chi^2_{кр}$$



$n_i(x_i, n'_i)$ $\sum n'_i \approx 98{,}75$ $\sum n_i = 100$

Рис. 2.20. Лабораторна робота 7, завдання 3

ЛАБОРАТОРНА РОБОТА 8
ПОБУДОВА РЕГРЕСІЙНИХ МОДЕЛЕЙ

Мета:

- засвоїти поняття регресійної моделі та її значення в аналізі залежностей між змінними;
- опанувати алгоритм побудови регресійних моделей за допомогою табличного процесора Microsoft Excel;
- навчитися інтерпретувати параметри регресійної моделі, оцінювати її адекватність та робити висновки щодо взаємозв'язків між досліджуваними змінними;
- розвинути навички застосування регресійного аналізу для прогнозування та прийняття рішень на основі даних.

Хід роботи

1. Вивчити теоретичні поняття (див.п.1.2.)

2. Завдання 1.

Наявні вибіркові дані щодо 40 студентів факультету (див. табл. 2.14): кількість годин, витрачених студентами на самостійну підготовку до контрольної роботи та кількість їх правильних відповідей на контрольній роботі.

Таблиця 2.14

Комбінаційна таблиця

Кількість годин на підготовку до КР	Кількість правильних відповідей			
	4-6	6-8	8-10	10-12
2,5-3,5	2			
3,5-4,5	6	3		
4,5-5,5	2	5	7	
5,5-6,5		2	2	3
6,5-7,5			1	7

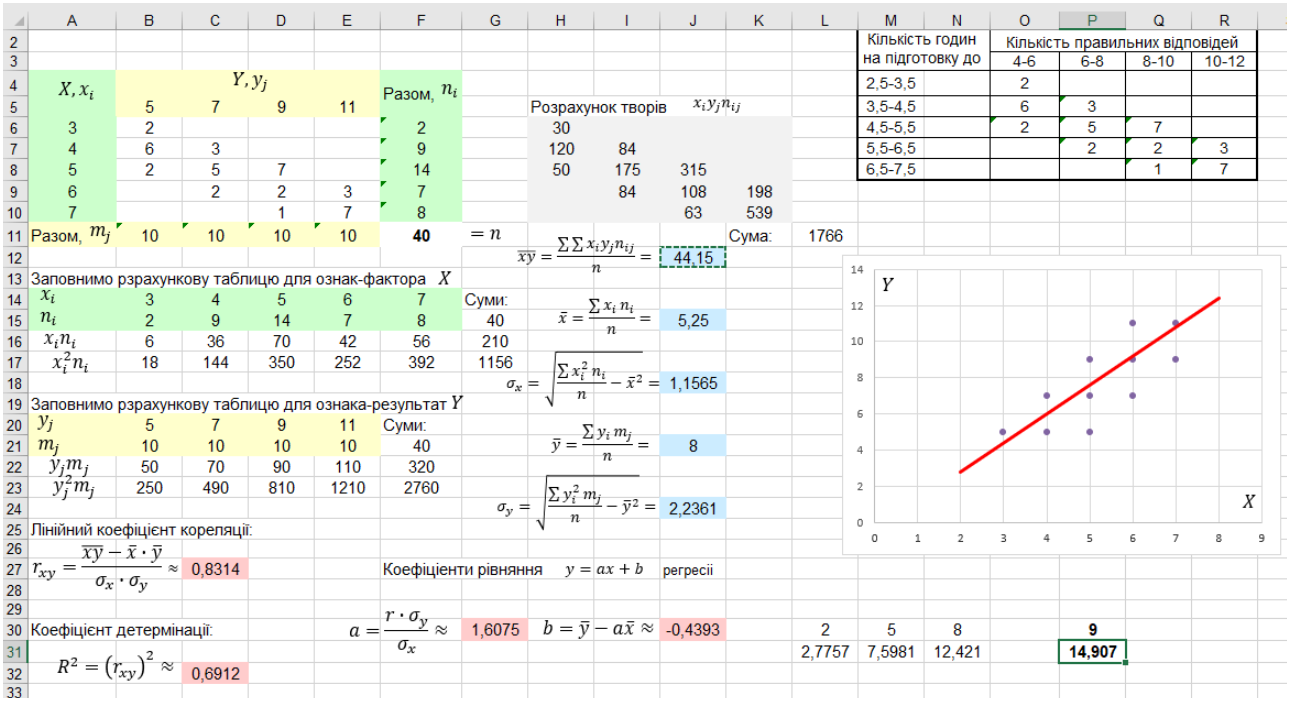
За даними таблиці:

1) Визначити *ознаку-фактор* X і *ознаку-результат* Y і висловити припущення про наявність та *напрямок кореляційної залежності* Y від X. Побудувати *кореляційне поле* та висунути гіпотезу про можливу форму залежності.

2) Обчислити лінійний коефіцієнт кореляції та детермінації, зробити висновки.

3) Знайти рівняння лінійної регресії Y і X зобразити відповідну пряму на кресленні. Спрогнозувати кількість правильних відповідей для студента, який готувався 9 годин.

Фрагмент робочого листа Microsoft Excel, який потрібно створити в процесі виконання завдання, наведено на рис. 2.21.



X, x_i	Y, y_j = 5	7	9	11	Разом, n_i
3	2				2
4	6	3			9
5	2	5	7		14
6		2	2	3	7
7			1	7	8
Разом, m_j	10	10	10	10	40

Кількість годин на підготовку до	Кількість правильних відповідей 4-6	6-8	8-10	10-12
2,5-3,5	2			
3,5-4,5	6	3		
4,5-5,5	2	5	7	
5,5-6,5		2	2	3
6,5-7,5			1	7

Рис. 2.21. Лабораторна робота 8, завдання 1

3. Завдання 2.

Дослідник вирішив встановити динаміку рівня знань студентів з програмування протягом першого року навчання. Він вибрав групу з 10 студентів, які ніколи раніше не займалися програмуванням, і відстежував їхні результати за регулярними тестами складності. Отримані дані представлені в таблиці 2.15:

Таблиця 2.15

Результати дослідження

X	1	2	3	4	5	6	7	8	9	10
Y	7	8	15	20	32	42	60	100	120	220

де X – кількість тижнів самостійної підготовки студентів;

Y – сума балів студентів на регулярних тестах.

За наявними даними:

1) Побудувати діаграму розсіювання та підібрати лінію, яка ефективно наближає емпіричні дані.

2) Методом найменших квадратів визначити *рівняння регресії* Y на X, порівняти з лінією на діаграмі.

3) Обчислити індекс детермінації та індекс кореляції.

4) Перевірити значимість отриманої моделі лише на рівні значимості $\alpha = 0{,}05$.

5) Знайти середню помилку апроксимації.

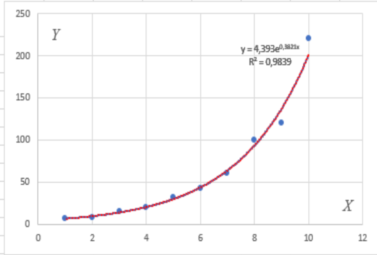
6) Спрогнозувати кількість балів, які зможуть отримати студенти за 12 та 24 тижні підготовки.

Фрагмент робочого листа Microsoft Excel, який потрібно створити в процесі виконання завдання, наведено на рис. 2.22.

X	1	2	3	4	5	6	7	8	9	10
Y	7	8	15	20	32	42	60	100	120	220

Фактор X	Резул-т Y							
x_i	y_i	x_i^2	$\ln y_i$	$x_i \ln y_i$	$(y_i-\bar{y})^2$	$y(x_i)=4{,}393e^{0{,}382x_i}$	$(y_i-y(x_i))^2$	$\left\lvert\frac{y_i-y(x_i)}{y_i}\right\rvert$
1	7	1	1,94591	1,94591	3069,16	6,4370	0,3169	0,0804
2	8	4	2,07944	4,15888	2959,36	9,4321	2,0510	0,1790
3	15	9	2,70805	8,12415	2246,76	13,8208	1,3904	0,0786
4	20	16	2,99573	11,9829	1797,76	20,2516	0,0633	0,0126
5	32	25	3,46574	17,3287	924,16	29,6745	5,4077	0,0727
6	42	36	3,73767	22,426	416,16	43,4819	2,1961	0,0353
7	60	49	4,09434	28,6604	5,76	63,7138	13,7921	0,0619
8	100	64	4,60517	36,8414	1413,76	93,3594	44,0977	0,0664
9	120	81	4,78749	43,0874	3317,76	136,7989	282,2035	0,1400
10	220	100	5,39363	53,9363	24837,76	200,4506	382,1801	0,0889
55	624	385	35,8132	228,492	40988,4		733,6990	0,8157

$n = 10$



$\Delta = \begin{vmatrix} 385 & 55 \\ 55 & 10 \end{vmatrix} =$ 825 $\neq 0$

$\Delta_a = \begin{vmatrix} 228{,}492 & 55 \\ 35{,}8132 & 10 \end{vmatrix} =$ 315,196

$a = \frac{\Delta_a}{\Delta} =$ 0,382

$\Delta_B = \begin{vmatrix} 385 & 228{,}492 \\ 55 & 35{,}8132 \end{vmatrix} =$ 1221

$B = \frac{\Delta_B}{\Delta} =$ 1,48

$b = e^B \approx e^{1{,}48} \approx$ 4,393

$\bar{y} = \frac{\sum y_i}{n} =$ 62,4

$Q = \sum (y_i - \bar{y})^2 =$ 40988,4

$Q_e = \sum (y_i - y(x_i))^2 \approx$ 733,6990

$R_I^2 = 1 - \frac{Q_e}{Q} =$ 0,9821

$R_I = \sqrt{R_I^2} \approx$ 0,991

$\bar{A} = \frac{1}{n}\sum \left\lvert\frac{y_i - y(x_i)}{y_i}\right\rvert \cdot 100\% \approx$ 8,157 %

$\alpha =$ 0,05

$k_1 =$ 1

$k_2 =$ 8

$F_{кр}(\alpha;\ k_1;\ k_2) \approx$ 5,3177

$F_{спос} = \frac{R_I^2(n-2)}{1-R_I^2} \approx$ 438,92

$F_{спос} = 438{,}92 > F_{кр} = 5{,}3177$

$H_0: R_\Gamma^2 = 0$ відкидаємо

$H_0: R_\Gamma^2 > 0$ приймаємо

$y(12) = 4{,}393e^{0{,}382\cdot 12} \approx$ 430

$y(24) = 4{,}393e^{0{,}382\cdot 24} \approx$ 42165

Рис. 2.22. Лабораторна робота 8, завдання 2

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Класифікація даних по типу шкали вимірювання ознаки. URL: https://stud.com.ua/93305/statistika/klasifikatsiya_danih_tipu_shkali_vimiryuvannya_oznaki (дата звернення: 07.08.2024).
2. Подання даних вимірювань дискретних величин. URL: https://drive.google.com/file/d/1v1UowXgiov4Cd4dA-XFdA3olyFtgv0Ds/view?usp=sharing (дата звернення: 07.08.2024).
3. Типи рядів розподілу даних. URL: https://elib.lntu.edu.ua/sites/default/files/elib_upload/%D0%95%D0%9D%D0%9F_%D0%AF%D0%BA%D0%B8%D0%BC%D1%87%D1%83%D0%BA_%D0%A1%D0%B5%D0%BB%D0%B5%D0%BF%D0%B8%D0%BD%D0%B0/page9.html#:~:text=%D0%9F%D0%BE%D0%B1%D1%83%D0%B4%D1%83%D0%B2%D0%B0%D1%82%D0%B8%20%D0%B2%D0%B0%D1%80%D1%96%D0%B0%D1%86%D1%96%D0%B9%D0%BD%D0%B8%D0%B9%20%D1%80%D1%8F%D0%B4%20%2D%20%D0%BE%D0%B7%D0%BD%D0%B0%D1%87%D0%B0%D1%94%20%D1%83%D0%BF%D0%BE%D1%80%D1%8F%D0%B4%D0%BA%D0%BA% (дата звернення: 07.08.2024).
4. Побудова моделі системи. URL: https://elib.lntu.edu.ua/sites/default/files/elib_upload/%D0%95%D0%9D%D0%9F_%D0%AF%D0%BA%D0%B8%D0%BC%D1%87%D1%83%D0%BA_%D0%A1%D0%B5%D0%BB%D0%B5%D0%BF%D0%B8%D0%BD%D0%B0/page13.html (дата звернення: 07.08.2024).
5. Індуктивна статистика в педагогічних дослідженнях. URL: https://lms.hnpu.edu.ua/pluginfile.php/5102/mod_resource/content/4/st2014_lit_indukt.pdf (дата звернення: 07.08.2024).
6. Математична статистика та математичні методи в психології. URL: https://moodle.znu.edu.ua/pluginfile.php/425789/mod_resource/content/1/%D0%9B%D0%B5%D0%BA%D1%86%D1%96%D1%8F%207.pdf (дата звернення: 07.08.2024).

7. Похибки вибіркових обстежень підприємств. URL: https://pidru4niki.com/1580011957634/statistika/pohibki_vibirkovih_obstezhen_pidpriyemstv (дата звернення: 07.08.2024).

8. Характеристики асиметрії та ексцесу. URL: https://studfile.net/preview/10098693/ (дата звернення: 07.08.2024).

9. Помилка середнього арифметичного. URL: https://studfile.net/preview/2299085/page:8/ (дата звернення: 07.08.2024).

10. Поняття про помилку вибірки. URL: https://studfile.net/preview/2502080/page:2/ (дата звернення: 07.08.2024).

11. Шкала Чеддока. URL: https://studfile.net/preview/3009913/page:39/ (дата звернення: 07.08.2024).

12. Похибки вибірки. URL: https://studfile.net/preview/5226519/page:20/ (дата звернення: 07.08.2024).

13. Кореляційні дослідження у психології. URL: https://studfile.net/preview/5484781/ (дата звернення: 07.08.2024).

14. Обчислення похибок вибірки та необхідної чисельності вибірки. URL: https://studfile.net/preview/5775952/page:16/ (дата звернення: 07.08.2024).

15. Обчислення основних статистичних характеристик вибірки. URL: https://ua5.org/model/1726-obchyslennya-osnovnyh-statystychnyh-harakterystyk-vybirky.html (дата звернення: 07.08.2024).

16. Випадкові дані Excel. URL: https://uk.extendoffice.com/documents/excel/6624-excel-generate-random-data.html#a3 (дата звернення: 07.08.2024).

17. Типи змінних. URL: https://ukrayinska.libretexts.org/%D0%A1%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0/%D0%9F%D1%80%D0%B8%D0%BA%D0%BB%D0%B0%D0%B4%D0%BD%D0%B0_%D1%81%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0/%D0%91%D1%96%D0%BE%D1%81%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0_-

_%D0%9F%D1%96%D0%B4%D1%80%D1%83%D1%87%D0%BD%D0%B8%D0%BA_%D0%B7_%D0%B2%D1%96%D0%B4%D0%BA%D1%80%D0%B8%D1%82%D0%BE%D0%B3%D0%BE_%D0%BD%D0%B0%D0%B2%D1%87%D0%B0%D0%BD%D0%BD%D1%8F/%D0%9F%D0%BE%D0%BF%D0%B5%D1%80%D0%B5%D0%B4%D0%BD%D1%96_%D0%B5%D1%82%D0%B0%D0%BF%D0%B8/%D0%A2%D0%B8%D0%BF%D0%B8_%D0%B7%D0%BC%D1%96%D0%BD%D0%BD%D0%B8%D1%85 (дата звернення: 07.08.2024).

18. Багатокутники частоти. URL: https://ukrayinska.libretexts.org/%D0%A1%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0/%D0%92%D1%81%D1%82%D1%83%D0%BF%D0%BD%D0%B0_%D1%81%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0/%D0%9A%D0%BD%D0%B8%D0%B3%D0%B0%3A_%D0%92%D1%81%D1%82%D1%83%D0%BF%D0%BD%D0%B0_%D1%81%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0_(Lane)/02%3A_%D0%93%D1%80%D0%B0%D1%84%D1%96%D1%87%D0%BD%D1%96_%D0%B4%D0%B8%D1%81%D1%82%D1%80%D0%B8%D0%B1%D1%83%D1%82%D0%B8%D0%B2%D0%B8/2.05%3A_%D0%91%D0%B0%D0%B3%D0%B0%D1%82%D0%BE%D0%BA%D1%83%D1%82%D0%BD%D0%B8%D0%BA%D0%B8_%D1%87%D0%B0%D1%81%D1%82%D0%BE%D1%82%D0%B8 (дата звернення: 07.08.2024).

19. Асиметрії коефіцієнт. URL: https://vue.gov.ua/%D0%90%D1%81%D0%B8%D0%BC%D0%B5%D1%82%D1%80%D1%96%D1%97_%D0%BA%D0%BE%D0%B5%D1%84%D1%96%D1%86%D1%96%D1%94%D0%BD%D1%82 (дата звернення: 07.08.2024).

20. Метод статистичних випробовувань. URL: https://wiki.cusu.edu.ua/index.php/%D0%9C%D0%B5%D1%82%D0%BE%D0%B4_%D1%81%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D1%87%D0%BD%D0%B8%D1%85_%D0%B2%D0%B8%D0%BF%D1%80%D0%BE%D0%B1%D0%BE%D0%B2%D1%83%D0%B2%D0%B0%D0%BD%D1%8C (дата звернення: 07.08.2024).

21. Вища математика в прикладах та задачах. URL: https://nmetau.edu.ua/file/kaplmath_19933.pdf (дата звернення: 07.08.2024).

22. Горонескуль М.М. Таблиці функцій та критичних точок розподілів. URL: http://repositsc.nuczu.edu.ua/bitstream/123456789/1530/1/Tablici.pdf (дата звернення: 07.08.2024).

23. Загальні відомості про метод статистичного моделювання. URL: https://elib.tsatu.edu.ua/dep/mtf/ophv_12/page6.html (дата звернення: 07.08.2024).

24. Жлуктенко В. І., Наконечний С. І., Савіна С. С. Теорія ймовірностей і математична статистика: Навч.- метод. посібник: К.: КНЕУ, 2001. 336 с. URL: https://kpdi.edu.ua/biblioteka/%D0%A2/%D0%A2%D0%B5%D0%BE%D1%80%D1%96%D1%8F%20%D0%B9%D0%BC%D0%BE%D0%B2%D1%96%D1%80%D0%BD%D0%BE%D1%81%D1%82%D0%B5%D0%B9%20%D1%96%20%D0%BC%D0%B0%D1%82%D0%B5%D0%BC%D0%B0%D1%82%D0%B8%D1%87%D0%BD%D0%B0%20%D1%81%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0%20%D0%96%D0%BB%D1%83%D0%BA%D1%82%D0%B5%D0%BD%D0%BA%D0%BE%20%D0%92.%D0%86.1.pdf (дата звернення: 07.08.2024).

25. Описова статистика. Лекція. URL: https://moodle.znu.edu.ua/pluginfile.php/648858/mod_resource/content/1/6%20%D0%9B%D0%B5%D0%BA%D1%86%D1%96%D1%8F%20-%20%D0%9E%D0%BF%D0%B8%D1%81%D0%BE%D0%B2%D0%B0%20%D1%81%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0.pdf (дата звернення: 07.08.2024).

26. Інформатика в закладах освіти. Освітньо-професійна програма. ХНПУ ім. Г.С. Сковороди. URL: http://smc.hnpu.edu.ua/files/Osv%D1%96tn%D1%96_programi/Osvitni_programu_bakalavr/2023_rik/Informatyka_v_zakladakh_osvity_bakalavr.pdf (дата звернення: 07.08.2024).

27. Інформатика. Освітньо-професійна програма. ЧНУ ім. Богдана Хмельницького. URL:

https://drive.google.com/file/d/18FgPGqo8E8lkijFf9sFqCJQRp5kll7Xl/view (дата звернення: 07.08.2024).

28. Ковальчук Лариса. Моделювання науково-педагогічних досліджень. Навчальний посібник. Львів: Видавничий центр ЛНУ імені Івана Франка, 2020. 520 с. URL: https://pedagogy.lnu.edu.ua/wp-content/uploads/2020/09/Koval-chuk-L.-Modeliuvannia-naukovo-pedahohichnykh-doslidzhen-.-Posibnyk.pdf (дата звернення: 07.08.2024).

29. Комп'ютерні науки. Освітньо-професійна програма. КНУ ім. Михайла Остроградського. URL: https://www.kdu.edu.ua/new/OPP/OPP_122_bac_2023.pdf (дата звернення: 07.08.2024).

30. Кореляційний аналіз. Лекція. URL: https://docs.google.com/presentation/d/1VDJCzVPQjrFGaibUYY8fe7mUb9nlOz5A/edit?usp=sharing&ouid=111101370347815307544&rtpof=true&sd=true (дата звернення: 07.08.2024).

31. Літнарович Р. М. Основи математичної статистики у психології. Частина 3. Навчальний посібник. – МЕГУ, Рівне, 2006, – 49 с. URL: https://essuir.sumdu.edu.ua/bitstream-download/123456789/3040/1/%d0%9e%d1%81%d0%bd%d0%be%d0%b2%d0%b8%20%d0%bc%d0%b0%d1%82%d0%b5%d0%bc%d0%b0%d1%82%d0%b8%d1%87%d0%bd%d0%be%d1%97%20%d1%81%d1%82%d0%b0%d1%82%d0%b8%d1%81%d1%82%d0%b8%d0%ba%d0%b8%20%d1%83%20%d0%bf%d1%81%d0%b8%d1%85%d0%be%d0%bb%d0%be%d0%b3%d1%96%d1%97.pdf;jsessionid=F5AEFA6548CBA36A60B7F5C9FE1DF5A6 (дата звернення: 07.08.2024).

32. Математична статистика. Освітньо-професійна програма. КОГПА ім. Тараса Шевченка. URL: http://kogpi.edu.te.ua/images/stories/Henrikh/OPP/Projects/2023-2024/INF.pdf ((дата звернення: 07.08.2024).

33. Науково-дослідна практика у підготовці майбутніх учителів інформатики: навч. посібник / Олефіренко Н.В., Андрієвська В.М., Колгатіна

Л.С., Остапенко Л.П., Пономарьова Н.О. – Харків: ХНПУ імені Г.С.Сковороди, 2022. – 60 с. URL: https://dspace.hnpu.edu.ua/server/api/core/bitstreams/c7f461fc-919a-49fd-a932-66610945a9c9/content (дата звернення: 07.08.2024).

34. Несnúmпляк Д. М. Застосування інноваційних технологій у психології: конспект лекцій. URL: https://dspace.lvduvs.edu.ua/bitstream/1234567890/1164/1/%D0%BD%D0%B5%D1%81%D0%BF%D0%BB%D1%8F%D0%BA%20%D0%B7%D0%B0%D1%81%D1%82%D0%BE%D1%81%D1%83%D0%B2%D0%B0%D0%BD%D0%BD%D1%8F%20%D1%96%D0%BD%D0%BD%D0%BE%D0%B2%D0%B0%D1%86%20%D1%82%D0%B5%D1%85%D0%BD%D0%BE%D0%BB..pdf (дата звернення: 07.08.2024).

35. Поглиблений MS Excel. Основи статистики в Excel. Навчальна програма. ХНПУ імені Г.С.Сковороди. URL: https://drive.google.com/file/d/14JGkBHM3VMUtvpkcaD4xYrJX_js2KnLW/view (дата звернення: 07.08.2024).

36. Поліщук Т.В. Використання цифрових інструментів під час обробки результатів педагогічного дослідження як ефективний засіб формування цифрової компетентності майбутніх учителів. *Збірник тез доповідей III Міжнародної науково-практичної конференції* URL: http://dspace.tnpu.edu.ua/bitstream/123456789/19224/1/32_Polishchuk.pdf (дата звернення: 07.08.2024).

37. Коцюбинський В.О. Прикладна статистика. Курс лекцій. URL: http://lib.pu.if.ua:8080/bitstream/123456789/2609/1/%D0%9F%D1%80%D0%B8%D0%BA%D0%BB%D0%B0%D0%B4%D0%BD%D0%B0%20%D1%81%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0%2B.pdf 14.04.2024 (дата звернення: 07.08.2024).

38. Професійна освіта. Комп’ютерні технології. Освітньо-професійна програма. БДПУ URL: https://bdpu.org.ua/wp-content/uploads/2023/06/015.039_PO_TSyfrovi-tekhnolohii.pdf (дата звернення: 07.08.2024).

39. Руденко В. М. Математична статистика. URL: https://shron1.chtyvo.org.ua/Rudenko_Volodymyr/Matematychna_statystyka.pdf (дата звернення: 07.08.2024).

40. Середня освіта (інформатика). Освітньо-професійна програма. ЗНУ URL: https://www.znu.edu.ua/opp/bak/math/opp_so-inform_23.pdf (дата звернення: 07.08.2024).

41. Середня освіта (Інформатика). Освітньо-професійна програма. КНУ ім. Михайла Остроградського. URL: https://www.kdu.edu.ua/new/OPP/OPP_014_bac_2023.pdf (дата звернення: 07.08.2024).

42. Середня освіта (Інформатика). Освітньо-професійна програма. РДГУ. URL: https://www.rshu.edu.ua/images/osvitni_programi/2023/osv_prog_bak_014_so_inf_2023.pdf (дата звернення: 07.08.2024).

43. Середня освіта (Інформатика). Освітня програма. ДДПУ. URL: https://ddpu.edu.ua/images/stories/news/2023/07_july/17/01/opp_bak_SO_informatika-2023.pdf (дата звернення: 07.08.2024).

44. Середня освіта (Інформатика). Освітня програма. ЖДУ ім. Івана Франка. URL: https://eportfolio.zu.edu.ua/media/TeachingProgram/199/uxbscvvm.pdf (дата звернення: 07.08.2024).

45. Середня освіта (Інформатика, математика). Освітньо-професійна програма. ДДПУ ім. Івана Франка. URL: https://dspu.edu.ua/wp-content/uploads/2023/07/serednya-osvita-informatyka-matematyka_bakalavr.pdf (дата звернення: 07.08.2024).

46. Системи випадкових величин. URL: http://ftplec.nlu.edu.ua/%D0%9B%D0%B5%D0%BA%D1%86%D1%96%D0%B9%D0%BD%D1%96%20%D0%BC%D0%B0%D1%82%D0%B5%D1%80%D1%96%D0%B0%D0%BB%D0%B8/%D0%9A%D0%B0%D1%84%D0%B5%D0%B4%D1%80%D0%B0%20%D0%B5%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D1%96%D1%87%D0%BD%D0%BE%D1%97%20%D1%82%D0%B5%D0%BE%D1%80%D1%96%D1%97/%D0%9D%D0%B0%D0%B2%D1%87%D0%B0%D0%BB%D1%8C%D0%BD%D0%B0%20%D0%B4%D0%B8%D1%81%D1%86%D0%B8%D0%BF%D0%BB%D1%96%D0%BD%D0%B0%20%22%D0%92%D0%B8%D1%89%D0%B0%20%D1%82%D0%B0%20%D0%BF%D1%80%D0%B8%D0%BA%D0%BB%D0%B0%D0%B4%D0%BD%D0%B0%20%D0%BC%D0%B0%D1%82%D0%B5%D0%BC%D0%B0%D1%82%D0%B8%D0%BA%D0%B0%22/%D0%A1%D0%B8%D1%81%D1%82%D0%B5%D0%BC%D0%B8%20%D0%B2%D0%B8%D0%BF%D0%B0%D0%B4%D0%BA%D0%BE%D0%B2%D0%B8%D1%85%20%D0%B2%D0%B5%D0%BB%D0%B8%D1%87%D0%B8%D0%BD.%20%D0%A2%D0%B5%D0%BC%D0%B0%205.pdf (дата звернення: 07.08.2024).

47. Статистичне вивчення кореляційного зв'язку випадкових величин. Лекція. URL: https://e-tk.lntu.edu.ua/pluginfile.php/25218/mod_resource/content/0/%D0%A2%D0%B5%D0%BC%D0%B0%2020.1.%20%D0%A1%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D1%87%D0%BD%D0%B5%20%D0%B2%D0%B8%D0%B2%D1%87%D0%B5%D0%BD%D0%BD%D1%8F%20%D0%BA%D0%BE%D1%80%D0%B5%D0%BB%D1%8F%D1%86%D1%96%D0%B9%D0%BD%D0%BE%D0%B3%D0%BE%20%D0%B7%D0%B2%E2%80%99%D1%8F%D0%B7%D0%BA%D1%83%20%D0%B2%D0%B8%D0%BF%D0%B0%D0%B4%D0%BA%D0%BE%D0%B2%D0%B8%D1%85%20%D0%B2%D0%B5%D0%BB%D0%B8%D1%87%D0%B8%D0%BD.pdf (дата звернення: 07.08.2024).

48. Статистичні гіпотези. Лекція. URL: https://repository.ldufk.edu.ua/bitstream/34606048/10763/1/%D0%9B%D0%B5%D0%BA%D1%86%20%E2%84%965.pdf (дата звернення: 07.08.2024).

49. Теорія ймовірностей та математична статистика. Навчальна програма. ДДПУ ім. Івана Франка. URL: https://dspu.edu.ua/ifmeit/wp-content/uploads/sites/2/2019/11/59.pdf (дата звернення: 07.08.2024).

50. Теорія ймовірностей та математична статистика. Силабус. ЗУІ ім. Ференца Ракоці II.URL: https://okt.kmf.uz.ua/mit/oktat-mit/2023-2024/Informatika_BSc/II_Teoriya_ymovirnostey_ta_matematychna_statystyka_4.pdf (дата звернення: 07.08.2024).

ДОДАТОК А.
ІНСТРУКЦІЇ ДО ВИКОНАННЯ ЛАБОРАТОРНИХ РОБІТ

ДОДАТОК А.1.

Лабораторна робота 1. Завдання 1

Інструкція до виконання:

Створити таблицю А.1:

Таблиця А.1

Розрахунок частот

	A	B	C	D
1	x_i	n_i	w_i	$w_н$
2				
3	2	5	0,05	0,05
4	4	10	0,1	0,15
5	6	15	0,15	0,3
6	8	20	0,2	0,5
7	10	40	0,4	0,9
8	12	10	0,1	1
9		100		

До комірок А3:А8 занести значення x_i – оцінки (варіант), а в В3:В8 n_i – частоти (кількість студентів, які отримали ту чи іншу оцінку).

Підрахувати загальну кількість групи $N = \sum_{i=1}^{6} n_i$. До комірки В9 занести формулу =SUM(B3:B8).

Розрахувати відносні частоти $w_i = \frac{n_i}{N}$, де N – обсяг усіх студентів. До комірки С3 занести формулу =B3/$B$9 і скопіювати її до комірок С4:С8.

Розрахувати накопичені частоти. Комірка D3 =C3. До комірки D4 занести формулу =D3+C4 і скопіювати її до комірок D5:D8.

У табличному процесорі Microsoft Excel створити діаграму «Лінійчата діаграма з маркерами» [23].

Зауважимо, що полігони частот та відносних частот зображено на рис.А.1, А.2.

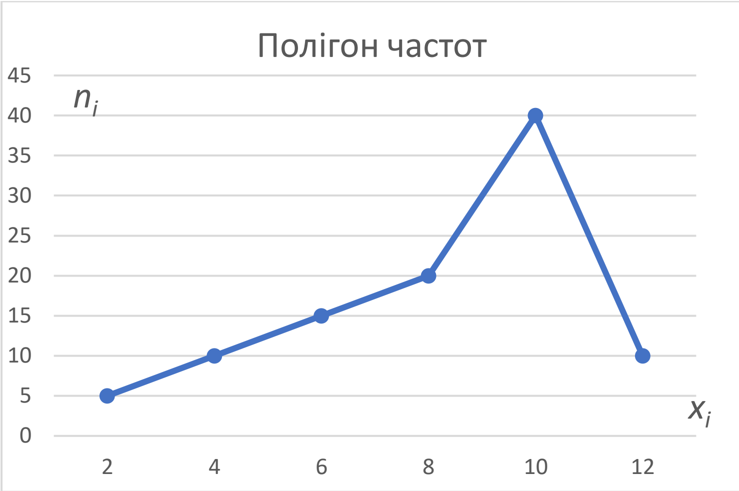


Рис. А.1. Полігони частот

Рис. А.2. Полігони відносних частот

Згідно з означенням та властивостями $F^*(x)$ має такий вигляд:

$$F^*(x) = \begin{cases} 0, & x \leq 2 \\ 0{,}05, & 2 < x \leq 4 \\ 0{,}15, & 4 < x \leq 6 \\ 0{,}3, & 6 < x \leq 8 \\ 0{,}5, & 8 < x \leq 10 \\ 0{,}9, & 10 < x \leq 12 \\ 1, & x > 12 \end{cases}$$

Для побудови графіка необхідно створити допоміжну таблицю А.2 для відповідних і точок:

Таблиця А.2

Допоміжна таблиця для сходинок і точок

-1	2
0	0
2	4
0,05	0,05
4	6
0,15	0,15
6	8
0,3	0,3
8	10
0,5	0,5
10	12
0,9	0,9
12	15
1	1

2	4	6	8	10	12
0	0,05	0,15	0,3	0,5	0,9

Тут доцільно використовувати вкладку «Точкова діаграма з прямими лініями».

Графічне зображення $F^*(x)$ подано на рис. А.3.

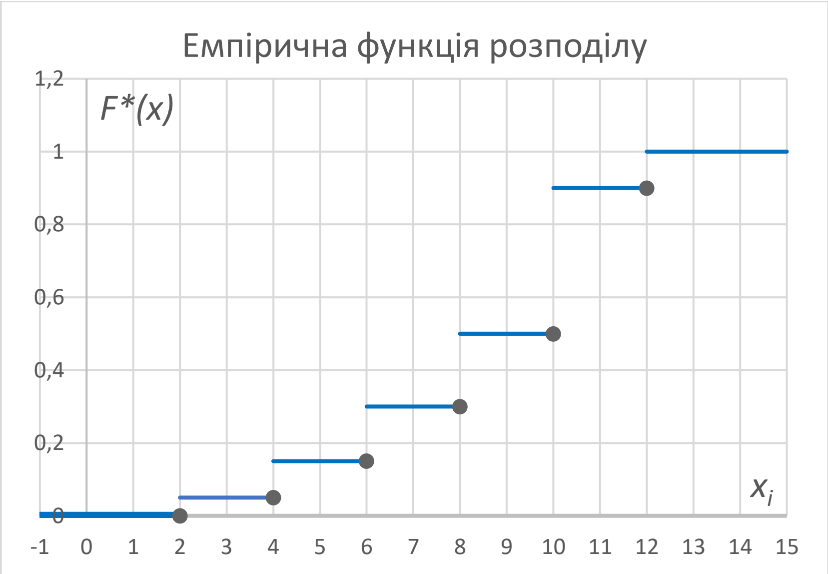


Рис. А.3. Емпірична функція розподілу

Створити «Секторну діаграму» (див. рис А.4).

Рис А.4. Секторна діаграма

Лабораторна робота 1. Завдання 2

Інструкція до виконання:

Створити таблицю А.3:

Таблицю А.3

Розрахунок частот

	A	B	C	D	E	F
1			Середини		Відносна	Накопич.
2	Мережі інтервалів		інтервалів	Частота	частота	частота
3	початок	кінець	x_i	n_i	w_i	$w_н$
4						
5	40	50	45	5	0,05	0,05
6	50	60	55	10	0,1	0,15
7	60	70	65	20	0,2	0,35
8	70	80	75	30	0,3	0,65
9	80	90	85	25	0,25	0,9
10	90	100	95	10	0,1	1
11	Суми:			100		

Розраховуємо середини інтервалів x_i: До комірки С5 вставити формулу =(A5+B5)/2; і скопіювати до комірок С6:С10.

Підрахувати загальну кількість групи $N = \sum_{i=1}^{6} n_i$. До комірки D11 занести формулу =SUM(D5:D10).

Розрахувати відносні частоти $w_i = \frac{n_i}{N}$, де N – обсяг усіх студентів. До комірки Е5 занести формулу =D5/$D$11 і скопіювати її до комірок Е6:Е10.

Розрахувати накопичені частоти. Комірка F5 =E5. До комірки F6 занести формулу =F5+E6 і скопіювати її до комірок F7:F10.

У табличному процесорі Microsoft Excel за допомогою вкладок «Звичайна стовпчаста діаграма» і «Лінійчата діаграма з маркерами» створити гістограму і полігон частот (див. рис. А.5) [23].

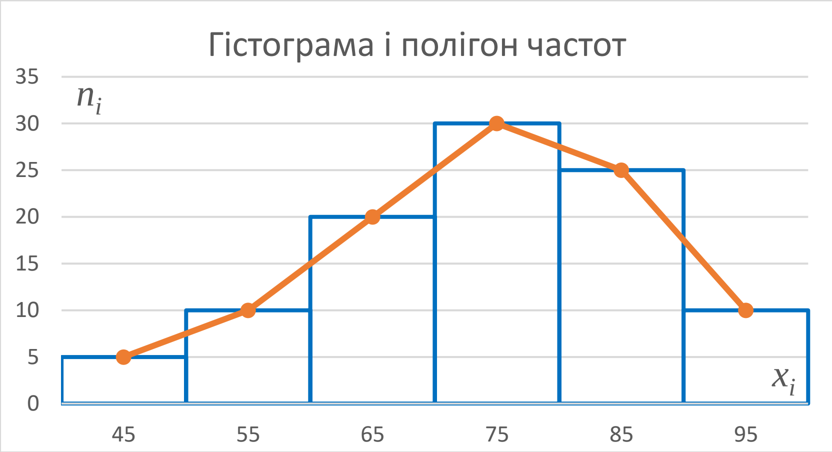


Рис. А.5. Гістограма і полігон частот

Полігони відносних частот зображено на рис.А.6:

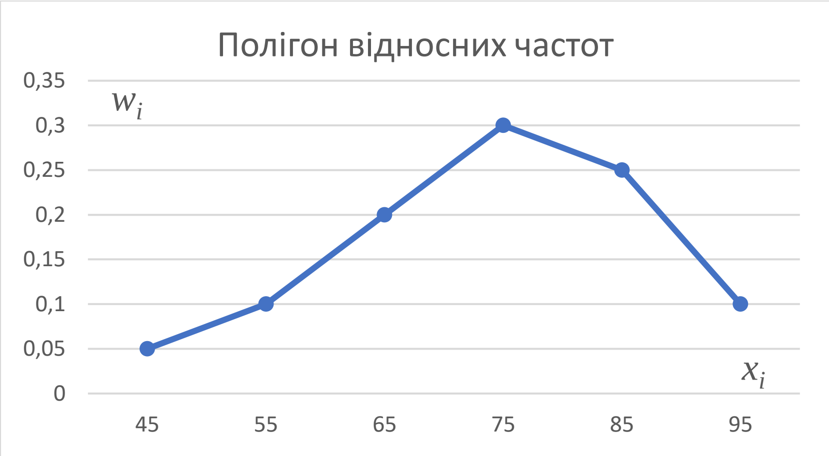


Рис.А.6. Полігони відносних частот

Згідно з означенням та властивостями $F^*(x)$ має такий вигляд:

$$F^*(x) = \begin{cases} 0, & x \le 0 \\ 0{,}05, & 40 < x \le 50 \\ 0{,}15, & 50 < x \le 60 \\ 0{,}35, & 60 < x \le 70 \\ 0{,}65, & 70 < x \le 80 \\ 0{,}9, & 80 < x \le 90 \\ 1, & 90 < x \le 100 \end{cases}$$

Графік $F^*(x)$ зображено на рис. А.7 :

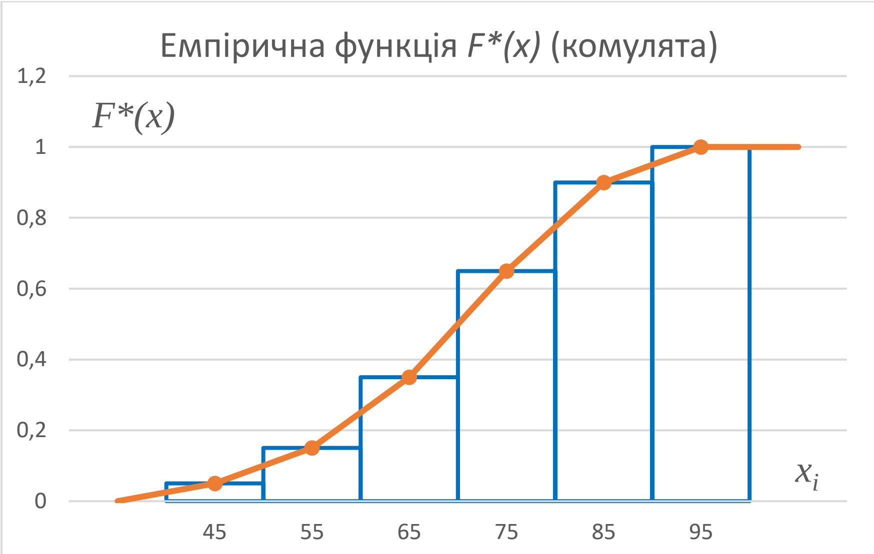


Рис. А.7. Емпірична функція $F^*(x)$ (комулята)

Створити «Секторну діаграму» (див. рис А.8).

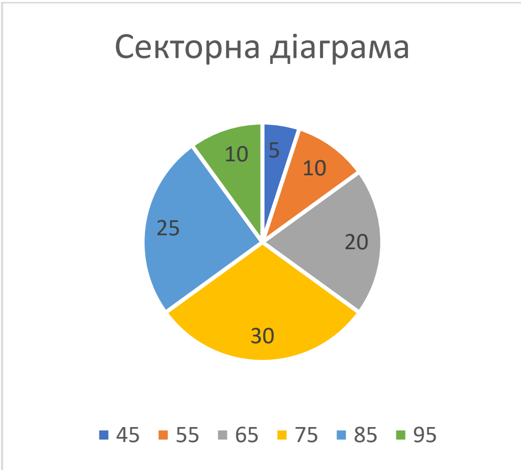


Рис А.8. Секторна діаграма

ДОДАТОК А.2.

Лабораторна робота 2. Завдання 1

Інструкція до виконання:

У табличному процесорі Microsoft Excel створити таблицю інтервального варіаційного ряду.

За допомогою формули =INT(RAND()*40) в комірки A3;A52 згенеруйте оцінки 50 студентів, та скопіюйте їх як значення до комірок B3:B52 [16].

Потрібно розмістити оцінки в одну колонку і відсортувати їх за зростанням.

Знайти мінімальне та максимальне значення з використанням функцій МІН та MAX та записати їх до комірок M2 та M3 відповідно.

Обчислимо розмах варіації: $R = x_{max} - x_{min}$ =M3-M2 – довжина загального інтервалу, в межах якого варіюються оцінки.

Тепер його потрібно розбити на часткові інтервали по формулі Стерджеса: k = 1 + 3.322lg n, де lg n – десятковий логарифм від обсягу вибірки та k – оптимальна кількість інтервалів, при цьому результат округляють до найближчого лівого цілого значення.

У нашому випадку отримуємо: =LOG10(50)*3,322+1≈6,6439≈6 інтервалів.

Довжини часткових інтервалів можуть бути різні, але в більшості випадків використовує рівноінтервальне угруповання: $h = \frac{x_{max} - x_{min}}{k}$ =M4/N5.

Встановити межі інтервалів:

- скопіюємо в комірку C4 x_{min} – початок першого інтервалу;
- кінець першого інтервалу – $x_{min} + h$ =C4+$M$6;
- початок другого інтервалу – кінець першого – =D4;
- перетягнемо D4 на одне ділення вниз – це кінець другого інтервалу;
- виділяємо дві комірки C5 і D5 і перетягнемо на чотири комірки вниз;
- ми отримали наші шість інтервалів. Переконуємося в тому, щоб кінець п’ятого інтервалу збігався з x_{max}.

Розраховуємо середини інтервалів x_i: До комірки E4 вставимо формулу =(C4+D4)/2; і скопіювати до комірок E5:E9.

Розрахуємо частоти за даними оцінок для кожного з інтервалів:

- виділити комірки F4:F9;
- вставити формулу =FREQUENCY(B3:B52;D4:D9);
- не натискати кнопку «ОК», а закінчити введення формули натисненням клавіш Ctrl+Shift+Enter (таблична формула);
- обов'язково переконуємося у тому, що нічого не втрачено: = SUM(F23:F27) $N = 50$ – обсяг вибірки.

Обчислимо *відносні частоти*: $w_i = \frac{n_i}{N}$ до комірки G4 занесимо формулу =F4/$F$10 і коиіюємо до комірок G5:G9 (округлюємо їх до 2 знаків після коми). У комірці G28 обчислити суму відносних частот і впевнитись, що вона дорівнює 1 [2].

Побудувати гістограму частот та полігон відносних частот.

Лабораторна робота 2. Завдання 2

Інструкція до виконання:

А) Для генерації імен та прізвищ скористайтесь сервісом: https://generator-online.com/uk/names/

До табличного процесору Microsoft Excel у комірки A3:A102 скопіюйте отримані прізвища. На жаль сервіс генерує максимум 99 прізвищ.

До комірок C3:C102 згенеруйте числа у діапазоні від 60 до 100 з двома знаками після коми за допомогою формули =ROUND(RAND()*(100-60)+60;2), та скопіюйте їх як значення до комірок B3:B102.

Отримати варіаційний ряд шляхом упорядкування даних за зростанням (на справді здійснюється впорядкування за не спаданням, оскільки серед спостережених значень є однакові).

Побудуйте гістограму розподілу оцінок.

Б) У табличному процесорі Microsoft Excel вкажіть нижню та верхню межі у двох комірках B1 та B2.

Скопіюйте у комірку D3 формулу =IFERROR(LARGE(ROW(INDIRECT($B$1&":"&$B$2))*NOT(COUNTIF($D$2:D2;ROW(INDIRECT($B$1&":"&$B$2))));RANDBETWEEN(1;$B$2-$B$1-ROW(A1)+2));"") і протягніть її до комірки D52 [16].

Лабораторна робота 2. Завдання 3

Інструкція до виконання:

а) у табличному процесорі Microsoft Excel до комірки А3 введіть формулу: =RANDBETWEEN(DATE(2021; 5; 1); DATE(2021; 10; 15)) [16].

Перетягніть і скопіюйте цю формулу в інші клітинки.

Зауваження: формат числа потрібно відформатувати як: «Короткий формат дати» (див. рис А.9).

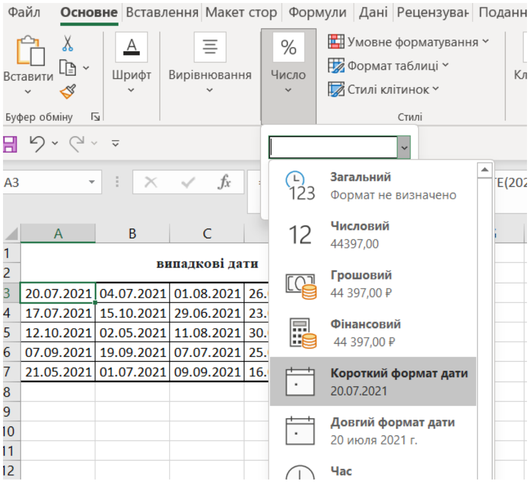


Рис. А.9. Короткий формат дати

b) У табличному процесорі Microsoft Excel до комірки А3 введіть формулу: =TEXT(RAND()*(18-10)/24+10/24;"ЧЧ:ММ:СС"). У наведеній вище формулі число 18 – час закінчення, і 10 означає час початку.

Перетягніть і скопіюйте цю формулу в інші клітинки.

c) У табличному процесорі Microsoft Excel до комірки А3 введіть формулу: = CHAR(RANDBETWEEN(33; 47)). Ця формула генерує випадковий символ діапазону ASCII-кодів від 33 до 47 [16].

Перетягніть і скопіюйте цю формулу в інші клітинки.

ДОДАТОК А.3.

Лабораторна робота 3. Завдання 1

Інструкція до виконання:

У табличному процесорі Microsoft Excel створити таблицю з іменами, та номером залікової.

За допомогою формули = RANDBETWEEN(1;12) слід згенерувати оцінки 12 учнів, та скопіювати їх як значення до комірок C3:C14. Отримані значення називаються *варіантами*, як і сукупність – *варіаційним рядом*. Варіанта – у статистиці означає окремо взяте *емпіричне значення*.

Далі слід обчислити суму усіх оцінок = SUM(C3:C14) та розрахувати *середнє арифметичне* $\bar{x} = \frac{\sum_{i=1}^{12} x_i}{n}$.

Лабораторна робота 3. Завдання 2

Інструкція до виконання:

Створити таблицю (в статистиці дані частіше мають не в рядках, а в стовпцях) До комірок С2:С11 занести значення x_i – оцінок (варіант), а в D2:D11 n_i – частоти (кількість студентів, які отримали ту чи іншу оцінку). Приклад *дискретного варіаційного ряду*.

Підрахувати загальну кількість групи $N = \sum_{i=1}^{10} n_i$. До комірки D12 занести формулу = SUM(D2:D11).

Визначити середню успішність групи за допомогою так званої середньозваженої середньої: $\bar{x} = \frac{\sum_{i=1}^{10} x_i n N_i}{\sum_{i=1}^{10} N_i} = \frac{x_1 N_1 + x_2 N_2 + \cdots + x_{10} N_{10}}{N} = 7$. Формула Microsoft Excel = AVERAGE () – оцінка математичного сподівання через вибіркове середнє арифметичне. У даному випадку =AVERAGE(A1:A25)

Розрахувати відносні частоти $w_i = \frac{n_i}{N}$, де N – обсяг усіх студентів. Результати занести у комірки Е2:Е11.

Застосувати таблицю спостережених частот для побудови гістограми (значення для кожної категорії подані стовпчиками відповідної висоти) та полігону (графік, де точки з'єднані відрізками) спостережених частот. Використати можливі значення оцінок (комірки С2:С11) як підписи по осі X (ось категорій), а частоти (комірки D2:D11 для гістограми) і відносні частоти (комірки Е2:Е11 для полігону) як значення.

Побудувати емпіричну функцію розподілу $F^*(x)$, яка визначається як відношення: $F^*(x) = \frac{n_x}{n}$, де n_x - кількість варіант **строго менших**, ніж x, при цьому «ікс» набуває всіх можливих значень від «мінус» до «плюс» нескінченності.

В інтервалі $x \in (-\infty; 3]$ функцію розподілу $F^*(x) = 0$, і, крім того, функція дорівнює нулю ще й у точці $F^*(3)$. Тому що значення $F^*(3)$ визначає кількість варіантів, які **строго** менше трійки, а ця кількість дорівнює нулю.

На проміжку $x \in (3;4]$ $F^*(x) = \frac{n_x}{n} = \frac{n_1}{n} = w_1 = 0.08$ – і знову значення F*(4) не враховує студентів, які отримали оцінку 4, т.я. мова йде про варіанти, які **строго** менше чотирьох.

На проміжку $x \in (4;5]$ $F^*(x) = \frac{n_x}{n} = \frac{n_1+n_2}{n} = w_1 + w_2 = 0.08 + 0.08 = 0.16$ і далі процес триває за принципом накопичення частот:

Якщо $5 < x \leq 6$ то $F^*(x) = \frac{n_x}{n} = \frac{n_1+n_2+n_3}{n} = w_1 + w_2 + w_3 = 0.08 + 0.08 + 0.12 = 0.28 \dots$

І, якщо 12 < x < +∞, то $F^*(x) = w_1 + w_2 + w_3 + \cdots + w_{11} = 0.08 + 0.08 + 0.12 + 0.12 + 0.2 + 0.16 + 0.08 + 0.08 + 0.04 + 0.04 = 1$ – для БУДЬ-ЯКОГО «х» з інтервалу (12, +∞) ВСІ частоти розташовані **строго** ліворуч від цього «х».

Результати накопичених відносних частот занести у комірки F2:F11.

Далі слід скласти емпіричну функцію розподілу:

$$F(x) = \begin{cases} 0, & \text{якщо } x \leq 3 \\ 0.08, & \text{якщо } 3 < x \leq 4 \\ 0.16, & \text{якщо } 4 < x \leq 5 \\ 0.28, & \text{якщо } 5 < x \leq 6 \\ 0.4, & \text{якщо } 6 < x \leq 7 \\ 0.6, & \text{якщо } 7 < x \leq 8 \\ 0.76, & \text{якщо } 8 < x \leq 9 \\ 0.84, & \text{якщо } 9 < x \leq 10 \\ 0.92, & \text{якщо } 10 < x \leq 11 \\ 0.96, & \text{якщо } 11 < x \leq 12 \\ 1, & \text{якщо } x > 12 \end{cases}$$

Побудувати графік емпіричної функції розподілу.

Лабораторна робота 3. Завдання 3

Інструкція до виконання:

Маємо вибіркову сукупність обсягом n = 30 спостережень. Серед запропонованих оцінок є однакові, але їхній розкид досить великий, і тому тут доцільно провести інтервальне розбиття. Для цього потрібно розмістити оцінки в одну колонку і відсортувати їх за зростанням, далі знайти мінімальне та максимальне значення з використанням функцій МІН та МАХ та записати їх до комірок G15 та G16 відповідно.

Обчислити розмах варіації: $R = x_{max} - x_{min}$ =G16-G15 – довжина загального інтервалу, в межах якого варіюються оцінки.

Далі інтервал потрібно розбити на часткові інтервали по формулі Стерджеса: k = 1 + 3.322lg n, де lg n – десятковий логарифм від обсягу вибірки та k – оптимальна кількість інтервалів, при цьому результат округляють до найближчого лівого цілого значення.

У нашому випадку отримуємо: =LOG10(30)*3,322+1≈5,9≈5 інтервали.

Довжини часткових інтервалів можуть бути різні, але в більшості випадків використовує рівноінтервальне угруповання: $h = \frac{x_{max} - x_{min}}{k}$ =G17/I18.

Встановити межі інтервалів:

- скопіюємо в комірку C23 x_{min} – початок першого інтервалу;
- кінець першого інтервалу – $x_{min} + h$ =C23+$G$19;
- початок другого інтервалу – кінець першого – D23;
- перетягнемо D23 на одне ділення вниз – це кінець другого інтервалу;
- виділяємо дві комірки C23 і D23 і перетягнемо на три комірки вниз;
- Ми отримали наші п’ять інтервалів. Переконуємося в тому, щоб кінець п’ятого інтервалу збігався з x_{max}.

Для кращої візуалізації можна скопіювати інтервали до стовпців C, D, E, F і G. Як що значення попадає на стик інтервалу, то в таких випадках його зазвичай відносять до поточного інтервалу.

Розрахувати середини інтервалів x_i: до комірки E23 вставити формулу =(C23+D23)/2; і скопіювати до комірок E24:E27.

Розрахувати частоти за даними оцінок для кожного з інтервалів:

- виділити комірки F23:F27;
- вставити формулу =FREQUENCY(A1:A30;D23:D27);
- не натискати кнопку «ОК», а закінчити введення формули натисненням клавіш Ctrl+Shift+Enter (таблична формула);
- обов'язково переконуємося у тому, що нічого не втрачено: = SUM(F23:F27) N = 30 – обсяг вибірки.

Обчислити щільність частот: $\frac{n_i}{h}$. До комірки G23 вставити формулу =F23/$G$19; і скопіювати до комірок G24:G27.

Обчислити *відносні частоти*: $w_i = \frac{n_i}{N}$ до комірки G23 занести формулу =F23/$F$28 і скопіювати до комірок G24:G27 (округлюємо їх до 2 знаків після коми). В комірці G28 обчислити суму відносних частот і впевнитись, що вона дорівнює 1.

Обчислити *щільність відносних частот*: $\frac{w_i}{h}$. До комірки I23 вставити формулу =H23/$G$19; і скопіювати до комірок I24:I27.

Побудувати гістограму частот, гістограму та полігон відносних частот, та емпіричну функцію розподілу.

Побудувати емпіричну функцію розподілу:

Для цього потрібно розрахувати накопичені відносні частоти. Розраховуються вони так: $w_{н1} = w_1$ – до комірки J23 занести формулу =H23 і кожна наступна одержується як сума попередньої накопиченої частоти і поточної відносної частоти. До комірки J23 слід занести формулу =J23+H24 і скопіювати до комірок J25:J27.

Емпірична функція розподілу визначається так само, як у дискретному випадку: $F^*(x) = \frac{n_x}{n}$, де n_x - кількість варіант строго менших, ніж x, який набуває всі значення від «мінус» до «плюс» нескінченності.

Далі слід побудувати шматково-ламану лінію, з проміжними точками ($x_{\text{прав}}$, $w_н$), де $x_{\text{прав}}$ – праві кінці інтервалів, а $w_н$ – відносна частота, яка встигла накопичитися на всіх пройдених інтервалах:

При цьому $F^*(x) = 0$ якщо $x \leq 59$ і $F^*(x) = 1$ якщо x >99.

ДОДАТОК А.4.

Лабораторна робота 4. Завдання 1

Інструкція до виконання:

Обчислити середнє значення можна у три способи:

1) щоб знайти середнє значення за первинними даними, потрібно підсумувати всі варіанти та розділити отриманий результат на обсяг сукупності:

=SUM(A1:A30)/30 = 76,23;

2) за допомогою функції Excel =AVERAGE(A1:A30) = 76,23;

3) за наявності готового інтервального ряду для обчислення середнього значення застосовуються середини x_i інтервалів (див. табл. А.4):

Таблиця А.4

Інтервальний ряд і середини інтервалів

	C	D	E	F	G
6			Середини		
7	Мережі інтервалі		інтервалів	Частота	
8	початок	кінець	x_i	n_i	$x_i\, n_i$
9	59	67	63	7	441
10	67	75	71	10	710
11	75	83	79	6	474
12	83	91	87	4	348
13	91	99	95	3	285
14	Суми:			30	2258

$$\bar{x}_{\text{в}} = \frac{\sum_{i=1}^{K} x_i n_i}{n} = \frac{2258}{30} \approx 75{,}27$$

Розбіжність із більш точним значенням (76,23), обчисленим за первинними даними, становить ≈ 1.

Зауважимо, що при виконання завдання пропонується прийом наближення ряду до дискретного. Разом із тим слід відзначити, що такий прийом не завджи постає продуктивним через сучасні можливості обчислень, які дозволяють працювати із значними обсягами даних [24].

Визначити моду (значення, що найбільш часто зустрічається у вибірці або наборі даних) можна одним із способів:

1) візуально (у завданні – 75) або на полігоні частот – це абсциса найвищої точки;

2) за допомогою функції Excel =MODE(A1:A30) = 75;

3) за наявності готового інтервального ряду необхідно визначити *модальний інтервал з максимальною частотою (у завданні* це інтервал (67;75) із частотою 10), і скористатися наступною формулою:

$$M_0 = x_0 + \frac{n_M - n_{M-1}}{(n_M - n_{M-1}) + (n_M - n_{M+1})} \cdot h$$

де: $x_0 = 67$ – нижня межа модального інтервалу;

h = 75 – 67 = 8 – довжина модального інтервалу;

$n_M = 10$ – частота модального інтервалу;

$n_{M-1} = 7$ – частота попереднього інтервалу;

$n_{M+1} = 6$ – частота наступного інтервалу.

Отже:

$$M_0 = 67 + \frac{10 - 7}{(10 - 7) + (10 - 6)} \cdot 8 \approx 70{,}43$$

З рис. А.10 видно, що мода помітно відрізняється від середнього арифметичного - мода зміщена щодо центру модального інтервалу у бік лівого інтервалу з більшою частотою [23].

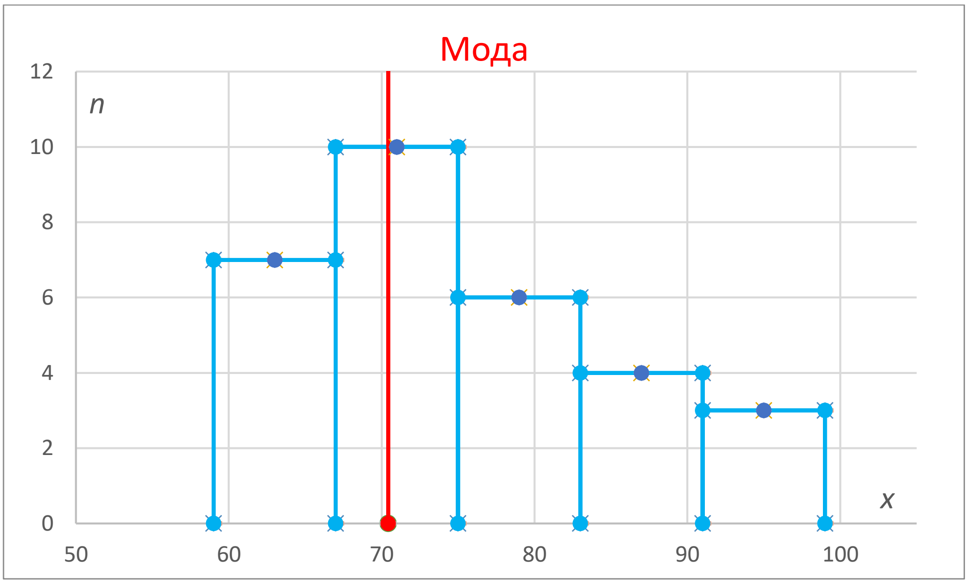


Рис. А.10. Мода

Медіана m_e варіаційного ряду (дискретного чи інтервального, генеральної сукупності чи вибіркової) – це значення, яка поділяє його на дві рівні частини (за кількістю варіантів). Медіану можна знайти кількома способами.

1) якщо дані первинні дані, то їх сортують за зростанням або спаданням і ділять обсяг вибірки навпіл, і оскільки вона складається з парної кількості варіант, то медіана дорівнює середньому арифметичному 15-й та 16-й варіанти *впорядкованого* варіаційного ряду: $m_e = \frac{x_{15}+x_{16}}{2} = \frac{75+75}{2} = 75$;

2) за допомогою стандартної функції Excel: =MEDIAN(A1:A30) = 75;

3) за умов наявності готового інтервального ряду медіана розраховується за формулою медаіни, але спочатку знаходять *медіанний інтервал* – це інтервал, що містить варіант (або 2 варіанти), яка ділить варіаційний ряд на дві рівні частини.

Розраховують «звичайні» накопичені частоти $n_н$ Обчислювальний алгоритм – перше значення зносимо ліворуч, і кожне наступне виходить як сума попереднього з поточною частотою з лівого стовпця (див. табл. А.5):

Таблиця А.5

Інтервальний ряд і накопичені частоти

	C	D	E	F	G	H
6			**Середини**			**Накопич.**
7	**Мережі інтервалі**		**інтервалів**	**Частота**		**частоти**
8	початок	кінець	x_i	n_i	$x_i\ n_i$	$n_н$
9	59	67	63	7	441	7
10	67	75	71	10	710	17
11	75	83	79	6	474	23
12	83	91	87	4	348	27
13	91	99	95	3	285	30
14	**Суми:**			30	2258	

Числа у правому стовпці – це кількість варіантів, які встигли "накопичитися" на всіх "пройдених" інтервалах, включаючи поточний.

Оскільки в завданні парна кількість варіант (30 штук), то медіанним буде той інтервал, який містить 30/2 = 15-ю та 16-ту варіанту. І орієнтуючись на накопичені частоти, можна дійти висновку, що ці варіанти містяться в інтервалі (67; 75).

Формула медіани:

$$m_e = x_0 + \frac{0{,}5n - n_{M-1}^{\text{н}}}{n_m} \cdot h$$

де:

n = 30 – обсяг статистичної сукупності;

$x_0 = 67$ – нижня межа медіанного інтервалу;

h = 75 – 67 = 8 – довжина медіанного інтервалу;

$n_m = 10$ – частота медіанного інтервалу;

$n_{M-1}^{\text{н}} = 7$ – *Накопичена частота* **попереднього** інтервалу [23].

Отже:

$$m_e = 67 + \frac{0{,}5 * 30 - 7}{10} \cdot 8 \approx 73.4$$

Зауважимо, що медіанне значення, навпаки, виявилося зміщене правіше, т.я. праворуч знаходиться значна кількість варіантів (див. рис А.11):

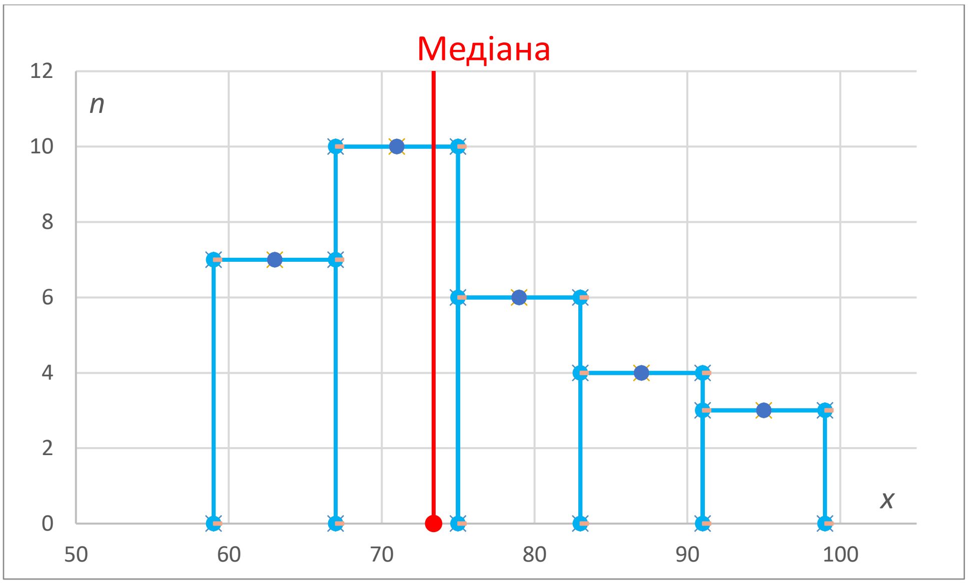


Рис. А.11. Медіана

Відповідь: $\bar{x}_{\text{в}} \approx 75{,}27$, $M_0 \approx 70{,}43$, $m_e \approx 73{,}4$ [24]

Лабораторна робота 4. Завдання 2

Інструкція до виконання:

Щоб знайти *середню* за первинними даними, потрібно підсумувати всі варіанти та розділити отриманий результат на обсяг сукупності:

$$\bar{x}_в = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- до комірки A14 занести формулу =SUM(A4:A13).
- до комірки G6 занести формулу =A14/D4.

Тепер слід знайти модулі відхилень від середньої:

$$|x_1 - \bar{x}_в| = |7{,}1 - 6{,}33| = |0{,}77| = 0{,}77$$

$$|x_2 - \bar{x}_в| = |6{,}3 - 6{,}33| = |-0{,}03| = 0{,}03$$

...

$$|x_{10} - \bar{x}_в| = |5{,}1 - 6{,}33| = |-1{,}23| = 1{,}23$$

- до комірки B4 занести формулу =ABS(A4-$G$6) та і скопіювати до комірок B5:B13.

На завершальному етапі слід розрахувати суму:

$$\sum_{i=1}^{n} |x_i - \bar{x}_в| = \sum_{i=1}^{10} |x_i - \bar{x}_в| = 0{,}77 + 0{,}03 + \cdots + 1{,}23 = 5{,}96$$

- до комірки B14 занести формулу =SUM(B4:B13), або комірку A14 скопіювати в B14;

Далі - середнє лінійне відхилення:

$$\bar{l} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}_в|}{n} = \frac{5{,}96}{10} = 0{,}596 \approx 0{,}6$$

- до комірки G9 занести формулу =B14/D4;

У табличному процесорі Microsoft Excel існує вбудована формула, занести її до комірки G10: =AVEDEV(A4:A13).

$\bar{l} \approx 0{,}6$ – означає, що виміряні значення x_i в середньому відрізняються від $\bar{x}_в = 6{,}33$ приблизно від 0,6 од.

Зауважимо, що для оцінки розсіювання варіант щодо середньої існує досконаліший і поширеніший підхід, який полягає в тому, щоб використовувати не модулі, а зведення відхилень у квадрат: $(x_1 - \bar{x}_в)^2$ (щоб ліквідувати негативні значення, що зустрічаються).

Вибіркова дисперсія – це *середнє арифметичне* квадратів відхилень всіх *варіантів* вибірки від її *середньої*:

$$D_в = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_в)^2}{n}$$

– для негрупованих даних, і:

$$D_в = \frac{\sum_{i=1}^{k}(x_i - \bar{x}_в)^2 \cdot n_i}{n}$$

– для сформованого варіаційного ряду, де x_i – кратні (однакові за значенням) *варіанти* у дискретному випадку або середини часткових інтервалів – *в* інтервальному, і – відповідні частоти n_i.

Замість модулів розраховуємо квадрати відхилень:

$$(x_1 - \bar{x}_в)^2 = (7{,}1 - 6{,}33)^2 = (0{,}77)^2 = 0{,}5929$$

$$(x_2 - \bar{x}_в)^2 = (6{,}3 - 6{,}33)^2 = (-0{,}03)^2 = 0{,}0009$$

...

$$(x_{10} - \bar{x}_в)^2 = (5{,}1 - 6{,}33)^2 = (-1{,}23)^2 = 1{,}5129$$

- до комірки C4 занести формулу =POWER(A4-$G$6;2) та і скопіювати до комірок C5:C13.

Далі слід розрахувати суму квадратів:

- до комірки C14 занести формулу =SUM(C4:C13), або комірку B14 скопіювати в C14.

$$D_в = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_в)^2}{n} = \frac{5{,}221}{10} = 0{,}5221$$

квадратних одиниць.

- до комірки G13 занести формулу =C14/D4.

У табличному процесорі Microsoft Excel існує вбудована формула, її слід занести до комірки H13: =VARP(A4:A13).

Для повернення до розмірності завдання з дисперсії слід взяти корінь.

У випадку з отриманими лінійними відхиленнями $\bar{l}$ – вони без будь-якої закономірності варіюватимуться навколо генерального значення $\bar{l}_{\text{Г}}$. Отримані значення вибіркової дисперсії $D_{\text{в}}$ даватимуть *систематично* занижену оцінку генеральної дисперсії $D_{\text{Г}}$. І тому вибіркову дисперсію слід обчислити за «виправленою» формулою:

$$s^2 = \frac{n}{n-1} \cdot D_{\text{в}}$$

s^2 – *виправлена вибіркова дисперсія*, яка є *незміщеною* оцінкою генеральної дисперсії.

Таким чином,

$$s^2 = \frac{n}{n-1} \cdot D_{\text{в}} = \frac{10}{9} \cdot 0{,}5221 \approx 0{,}58$$

- до комірки G16 заносимо: =(D4/(D4-1))*G13.

Слід зазначити, що для великої вибірки (від 100 і навіть від 30 варіант) цією поправкою можна знехтувати, так як при $n \to \infty$ дріб $\frac{n}{n-1}$ прагне одиниці і $D_{\text{в}} \to s^2$.

Лабораторна робота 4. Завдання 3

Інструкція до виконання:

За умовою завдання обсяг вибірки n = 400, але тут не закриті крайні інтервали. Оскільки довжини внутрішніх інтервалів становлять h = 200, то слід розглянути таку ж довжину і по краях, тобто інтервали від 20 до 40 і від 100 до 120 балів.

Для розрахунку числових характеристик слід перейти до інтервального варіаційного ряду, вибравши як *варіант* x_i середини інтервалів (окрім того, варіанти доцільно зменшити у 100 разів, оскільки в ході подальших обчислень виходитимуть гігантські числа):

- до комірки С3 занести формулу =((A3+B3)/2)/100 і скопіювати її до С4;С7.

Далі слід спочатку обчислити вибіркову середню, для чого знайти добуток твори $x_i n_i$ та їх суму:

- до комірки E3 занести формулу =C3*D3 і скопіювати до E4;E7;
- до комірки E8 занести формулу =SUM(E3:E7);

і за відповідною формулою далі слід розрахувати середній бал:

$$\bar{x}_{\text{в}} = \frac{\sum x_i n_i}{n} = \frac{312}{400} = 0{,}78 \text{ або } 78 \text{ балів}$$

- до комірки E12 занести формулу =E8/D8.

Дисперсія за визначенням може розраховуватися так:

$$D_{\text{в}} = \frac{\sum (x_i - \bar{x}_{\text{в}})^2 \cdot n_i}{n}$$

але простішою є формула:

$$D_{\text{в}} = \overline{x^2} - (\bar{x}_{\text{в}})^2 = \frac{\sum {x_i}^2 n_i}{n} - (\bar{x}_{\text{в}})^2$$

Для її застосування слід розрахувати добутки ${x_i}^2 n_i$ та їх суму $\sum {x_i}^2 n_i$:

- до комірки F2 занести формулу =C3*C3*D3 і скопіювàти до F4;F7;
- до комірки F8 занести формулу =SUM(F3:F7).

Розрахувати $(\bar{x}_{\text{в}})^2$:

- до комірки I12 занести формулу =E12*E12.

Отже, за формулою обчислення дисперсії отримуємо:

$$D_{\text{в}} = \frac{\sum x_i{}^2 n_i}{n} - (\bar{x}_{\text{в}})^2 = \frac{266{,}4}{400} - 0{,}78^2 = 0{,}666 - 0{,}6084 = 0{,}0576$$

(урахувуючи другий степінь, оскільки дисперсія є величина квадратична).

- до комірки F16 занести формулу =F8/D8-I12.

Для повернення до розмірності завдання з дисперсії слід взяти квадратний корінь:

$$\sqrt{D_{\text{в}}} = \sqrt{0{,}0576} = 0{,}24,$$

або 24 балів. Отриманий показник - *середнє квадратичне відхилення*, або *стандартне відхилення*.

- до комірки F19 занести формулу =SQRT(F16).

$$\sigma_{\text{в}} = 0{,}24$$

Зауважимо, що чим менше стандартне відхилення (і дисперсія), тим менша варіація – тим більша кількість варіантів знаходиться поблизу вибіркової середньої. Оскільки у завданні розкид досить великий, то значна кількість балів розташована далеко від $\bar{x}_{\text{в}} = 0{,}78$, і тому значення $\sigma_{\text{в}}$ вийшло чималим.

Коректна оцінка генеральну дисперсію D_{Γ} та генеральне середнє квадратичне відхилення σ_{Γ} виходить з наступних міркувань.

Вибіркова дисперсія є **зміщеною** оцінкою генеральної дисперсії. Це означає, що якщо проводити неодноразові вибірки з тієї ж генеральної сукупності, то отримані значення $D_{\text{в}}$ будуть *систематично занижено* оцінювати D_{Γ}, хоча це не означає, що $D_{\text{в}}$ завжди буде менше, ніж D_{Γ}.

Тому вибіркову дисперсію потрібно виправити:

$s^2 = \frac{n}{n-1} \cdot D_{\text{в}} = \frac{400}{399} \cdot 0.0576 \approx 0{,}057744$ – *виправлена вибіркова дисперсія.*

- до комірки I16 *занести* формулу =(D8/(D8-1))*F16.

і, відповідно:

$s = \sqrt{s^2} \approx \sqrt{0{,}057744} \approx 0{,}2403$ або $24{,}03$ бала – *виправлене середнє квадратичне відхилення*.

- до комірки I19 *занести* формулу =SQRT(I16).

Тут s^2 і s це вже незміщені оцінки генеральної дисперсії $D_Г$ та генерального стандартного відхилення $\sigma_Г$ відповідно.

Виходячи з викладеного $D_в = 0{,}0576, \sigma_в = 0{,}24$, а як оцінку відповідних генеральних показників приймаємо $s^2 \approx 0{,}057744$ та $s \approx 0{,}2403$.

В завданні, якщо не зменшувати варіанти в 100 разів, то дисперсія вийде занадто велика і тут слід привести результати до єдиного стандарту.

Для цього існують показники *відносні*, і найвідомішим з них є *коефіцієнт варіації*. Це відношення *стандартного відхилення* до *середньої*, виражене у відсотках:

$$V = \frac{\sigma}{\bar{x}} \cdot 100\%$$

Тоді у вираженні в балах:

$$V = \frac{\sigma_в}{\bar{x}_в} \cdot 100\% = \frac{24}{78} * 100\% \approx 30{,}77\%$$

У вираженні в сотнях балів:

$$V = \frac{\sigma_в}{\bar{x}_в} \cdot 100\% = \frac{0{,}24}{0{,}78} * 100\% \approx 30{,}77\%$$

- до *комірки* H22 занести формулу =F19/E12*100.

Зауважимо, що у статистиці існує наступний емпіричний орієнтир:

- якщо показник варіації становить приблизно 30% і менше, то статистична сукупність вважається ***однорідною***. Це означає, що більшість *варіантів* знаходиться недалеко від *середньої* і знайдене значення $\bar{x}$ добре характеризує центральну тенденцію сукупності.

– якщо показник варіації становить значно більше 30%, то сукупність ***неоднорідна***, тобто, значну кількість *варіант* перебувають далеко від $\bar{x}$, і середня вибіркова погано характеризує типову варіанту [23].

Лабораторна робота 4. Завдання 4

Інструкція до виконання:

Для розрахунку числових характеристик слід перейти до інтервального варіаційного ряду, вибравши як *варіант* x_i середини інтервалів.

Оскільки у формулах асиметрії та ексцесу фігурує стандартне відхилення, то спочатку потрібно розрахувати вибіркову середню та дисперсію.

Обчислити добутки твори $x_i n_i$, їх суму та середнє значення:

$$\bar{x}_{\text{в}} = \frac{\sum x_i n_i}{n} = \frac{740}{100} = 7{,}4 \text{ балів}$$

- до комірки E4 занести формулу =C4*D4 та скопіювати її до комірок E5:E7;
- до комірки D8 занести формулу =SUM(D4:D7) та скопіювати її до комірки E8 (об'єм вибірки і сума творів $x_i n_i$):
- до комірки C12 занести формулу =E8/D8 (вибіркова середня).

Дисперсію тут зручніше знайти не за формулою, а за визначенням:

$$D_{\text{в}} = \frac{\sum (x_i - \bar{x}_{\text{в}})^2 \cdot n_i}{n}$$

Для цього слід розрахувати добутки $(x_i - \bar{x}_{\text{в}})^2 n_i$ і потім $(x_i - \bar{x}_{\text{в}})^3 n_i$, $(x_i - \bar{x}_{\text{в}})^4 n_i$ (див. табл. А.6):

Таблиця А.6

Розрахунок творів

	A	B	C	D	E	F	G	H
3	Интервалы		x_i	n_i	$x_i n_i$	$(x_i - \bar{x}_в)^2 n_i$	$(x_i - \bar{x}_в)^3 n_i$,	$(x_i - \bar{x}_в)^4 n_i$
4	48	49	48,5	20	970	28,8	-34,56	41,472
5	49	50	49,5	50	2475	2	-0,4	0,08
6	50	51	50,5	20	1010	12,8	10,24	8,192
7	51	52	51,5	10	515	32,4	58,32	104,976
8	Суммы:			100	4970	76	33,6	154,72

- до комірки F4 занести формулу =POWER($C4-$C$12;2)*$D4;

- скопіювати формулу до комірок G4:H4, і в цих комірках поставити відповідну ступень (3 і 4);
- виділити комірки F4:G4:H4 і скопіювати до рядка F7:G7:H7;
- виділити комірку E8 і скопіювати її до комірки H8.

Фінальні розрахунки:

$$D_{\text{в}} = \frac{\sum(x_i - \bar{x}_{\text{в}})^2 \cdot n_i}{n} = \frac{304}{100} = 3{,}04$$

- до комірки B17 занести формулу =F8/D8 (дисперсія);

$$\sigma_{\text{в}} = \sqrt{D_{\text{в}} =} \sqrt{3{,}04} = 1{,}7436$$

- до комірки F14 занести формулу =SQRT(D17) (середнє квадратичне відхилення).

Центральні моменти 3-го та 4-го порядків:

$$m_3 = \frac{\sum(x_i - \bar{x}_{\text{в}})^3 \cdot n_i}{n} = \frac{268{,}8}{100} = 2{,}688$$

- до комірки H12 занести формулу =G8/D8;

$$m_4 = \frac{\sum(x_i - \bar{x}_{\text{в}})^4 \cdot n_i}{n} = \frac{2475{,}52}{100} = 24{,}7552$$

- до комірки H17 занести формулу =H8/D8.

Обчислення коефіцієнту асиметрії виконують за формулою:

$$A_s = \frac{m_3}{\sigma_{\text{в}}^3} = \frac{2{,}688}{(1{,}7436)^3} = 0{,}5071 \approx 0{,}51$$

- до комірки E21 занести формулу =H12/POWER(F14;3).

Зауважимо, що якщо $A_s > 0{,}5$, то розподіл скошено вправо, якщо $A_s < 0{,}5$ – то вліво. При цьому прийнята наступна умовна градація: якщо отримане значення по модулю менше, ніж 0,25, асиметрія незначна, якщо $0{,}25 < |A_s| < 0{,}5$, то помірна, і якщо $|A_s| > 0{,}5$, то істотна.

Чим **менше** по модулю A_s**,** тим емпіричний розподіл, що розглядається, тим **ближче** до нормального розподілу з параметрами $a = \bar{x}_{\text{в}}$, $\sigma = \sigma_{\text{в}}$.

У випадку завдання $A_s \approx 0{,}51 > 0$, тобто розподіл має істотну правосторонню асиметрію.

Обчислення коефіцієнту виконують за формулою:

$$E_k = \frac{m_4}{\sigma_{\text{в}}^4} - 3 = \frac{24{,}7552}{(1{,}7436)^4} - 3 = -0{,}3213 \approx -0{,}32$$

- до комірки E21 заносимо формулу =-3+H17/POWER(F14;4).

Якщо $E_k > 0$, то емпіричний розподіл є вищим («гостровершинним») – щодо «еталонного» нормального розподілу з параметрами $a = \bar{x}_{\text{в}}$, $\sigma = \sigma_{\text{в}}$. Якщо ж $E_k < 0$ – то нижчим та пологішим. І чим більше E_k за модулем, тим «аномальніша» висота в той чи інший бік.

У випадку завдання $E_k \approx -0{,}32 < 0$, розподіл помітно нижче, ніж нормальний розподіл з параметрами $a = \bar{x}_{\text{в}} = 7{,}4$, $\sigma = \sigma_{\text{в}} \approx 1{,}74$.

Відповідь: $A_s \approx 0{,}51$, $E_k \approx -0{,}32$ [23].

ДОДАТОК А.5.

Лабораторна робота 5. Завдання 1

Інструкція до виконання:

За умовою завдання дано два варіаційні ряди: *X* – кількість відвіданих лекцій та практичних занять і *Y* – підсумкові бали. Кількість відвіданих лекцій та практичних занять не завжди впливає на підсумковий бал, однак *загальна тенденція* полягає в тому, що при збільшенні відвіданих лекцій збільшується і *середній бал*. Така *нежорстка* залежність називається *кореляційною.*

Показник *X* (кількість відвіданих занять) називається *факторним* (причинним) або *ознакою-фактором*. Показник *Y* (підсумковий бал) називається *результативним* (наслідковим) або *ознакою-результатом*.

За умовою завдання потрібно встановити наявність та характер кореляційного зв'язку між кількістю відвіданих занять та підсумковими балами. Під *характером* зв'язку розуміється її *напрямок*, при цьому тут можливі такі варіанти:

– *прямий зв'язок* («чим більше, тим більше» – наш випадок);
– *зворотній зв'язок* («чим більше, тим менше»);
– відсутність зв'язку («чим більше, тим так само хаотично»).

Метод *аналітичного угруповання* дозволяє з'ясувати наявність та напрямок *кореляційного зв'язку* між *ознакою-фактором X* та *ознакою-результатом* Y.

Загальний алгоритм розв’язання може бути наступний:

1. Виконуємо припущення, що середній бал по групі студентів залежить від кількості відвіданих занять, причому чим більше відвіданих занять, тим вище може бути бал. Таким чином, кількісті відвіданих занять X є ознакою-фактором, а підсумкові бал Y – ознакою-результатом; передбачувана кореляційна залежність – пряма.

Для перевірки гіпотези (припущення) методом *аналітичного угруповання* спочатку потрібно впорядкувати сукупність за ознакою-фактором – за зростанням відвіданих занять:

Далі слід виконати угруповання сукупності *за ознакою-фактором* (кількістю відвіданих занять) стандартним рівноінтервальним угрупованням. Розмах варіації складає:

$$R = x_{max} - x_{min} = 20 - 12 = 8$$

Оптимальну кількість інтервалів слід визначити за формулою Стерджеса, для обсягу сукупності $n = 30$ воно складає:

$$k = 1 + 3{,}322 lg 30 \approx 5{,}9 \rightarrow 5 \text{ інтервалів.}$$

Таким чином, довжина кожного інтервалу:

$$h = \frac{R}{k} = \frac{8}{5} = 1{,}6,$$

в результаті чого отримано інтервали: 12-13,6, 13,6-15,2, 15,2-16,8, 16,8-18,4 та 18,4-20 занять. Далі слід організувати групи та заповнити групову таблицю (див. рис А.12) .

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	ПІП	Лекції та ПР	Бали		Перша група		Друга група		Третя група		Четверта група		П'ята група	
2	3	12	41		12	41	14	54	16	65	17	83	19	88
3	7	13	45		13	45	14	57	16	66	17	85	19	96
4	12	13	48		13	48	15	52	16	67	17	72	20	101
5	4	14	54				15	62	16	59	17	76		
6	8	14	57				15	64	16	68	18	78		
7	16	15	52						16	69	18	80		
8	17	15	62						16	70	18	81		
9	22	15	64						16	71	18	92		
10	1	16	65						16	73				
11	5	16	66						16	74				
12	9	16	67						16	75				
13	13	16	59											
14	14	16	68	$n_i =$										
15	18	16	69											
16	20	16	70	Разом по групам										
17	21	16	71											
18	25	16	73	$\bar{y}_i =$										
19	26	16	74	*(середне значення по групам)*										
20	28	16	75						Аналітична таблиця					
21	15	17	83						№ групи	Інтервали		n_i	$\bar{y}_i$	
22	19	17	85							факторної ознаки				
23	23	17	72	Мінімальне значення:			12		1					
24	30	17	76	Максимальне значення:			20		2					
25	2	18	78	Розмах варіації:			8		3					
26	6	18	80	Кількість інтервалів:			5,907	5	4					
27	10	18	81	Довжина інтервалу:			1,6		5					
28	11	18	92											
29	24	19	88											
30	27	19	96											
31	29	20	101											
32														

Рис А.12. Групова таблиця

У нижній частині таблиці підраховують кількість n_i за кожною групою, суму значень *результативної* ознаки (бали), і *середні значення* $\bar{y}_i$ *результативної* ознаки за групами.

Основні результати слід представити у вигляді *аналітичної таблиці* (див. табл. А.7):

Таблиця А.7

Аналітична таблиця

	J	K	L	M	N
20	Аналітична таблиця				
21	№ групи	Інтервали факторної ознаки		n_i	$\bar{y}_i$
22					
23	1	12	13,6	3	44,667
24	2	13,6	15,2	5	57,8
25	3	15,2	16,8	11	68,818
26	4	16,8	18,4	8	80,875
27	5	18,4	20	3	95

Зауважимо, що аналітичне угруповання показало, що зі збільшенням кількості відвіданих занять спостерігається збільшення середнього підсумкового балу. Таким чином, між розглянутими факторами справді існує *пряма кореляційна залежність*.

2. Для оцінки сили впливу факторної ознаки на результат (оцінки, наскільки тісний зв'язок між факторами) застосовують *емпіричний коефіцієнт детермінації*, та пов'язане з ним *емпіричне кореляційне ставлення*.

Для цього слід виміряти *тісноту* кореляційного зв'язку між кількістю відвіданих занять та підсумковим балом за допомогою *емпіричного кореляційного відношення*.

Спочатку обчислюють емпіричний коефіцієнт детермінації. Підсумкові бали (ознака-результат) *варіюється* під впливом різних чинників, їх може бути досить багато. І загальна дисперсія σ_y^2 враховує **всі** чинники, які впливають на зміну балів.

Далі розбивають сукупність на групи, де основою угруповання має бути відвідування занять (один із чинників). Суттєвість впливу цього фактора вимірюється міжгруповою дисперсією δ_y^2. І емпіричний коефіцієнт детермінації $\eta^2 = \frac{\delta_y^2}{\sigma_y^2}$ показує **долю** впливу групувальної ознаки – чим ближче цей коефіцієнт до одиниці, тим сильніший вплив причини на слідство.

Загальну дисперсію підсумкових балів обчислюють за відсортованими даними. Спочатку слід розрахувати середнє значення балів по всій вибірці:

$$\bar{y} = \frac{y_1 + y_2 + y_3 + \cdots + y_{30}}{n} = \frac{41 + 45 + 48 + \cdots + 101}{30} = 70{,}4$$

- до комірки C32 занести формулу =SUM(C2:C31);
- до комірки H28 занести формулу =C32/H26.

Дисперсію слід обчислити за визначенням:

$$\sigma_y^2 = \frac{\sum_{i=1}^{30}(y_i - \bar{y})^2}{n} = \frac{6185{,}2}{30} = 206{,}17$$

- до комірки D2 занести формулу =POWER(C2-$R$22;2) і копіювати її до комірок D3:D31;
- до комірки D32 занести формулу =SUM(D2:D31);
- до комірки H31 занести формулу =D32/H26.

Далі обчислюють *міжгрупову дисперсію*:

Для цього потрібно розрахувати квадрат відхилень

$$(\overline{y_i} - \bar{y})^2 \cdot n_i$$

- до комірки G21 занести формулу =POWER(G18-$H$28;2)*G14 і скопіювати її до комірок H21:K21.

Міжгрупова дисперсія:

$$\delta_y^2 = \frac{\sum_{i=1}^{5}(\overline{y_i} - \bar{y})^2 \cdot n_i}{n} = \frac{1986{,}6 + 793{,}8 + 27{,}52 + 877{,}8 + 1815{,}5}{30} \approx 183{,}37$$

- до комірки N25 занести формулу =SUM(G21:K21)/H26.

Для перевірки можна ще обчислити *внутрішньогрупову дисперсію*, щоб перевірити *правило складання дисперсій*. Для цього слід заповнити таблицю А.8.

Таблиця А.8

Обчислення групових дисперсій

	L	M	N	O	P	Q
1		$(y_i-\bar{y}_1)^2$	$(y_i-\bar{y}_2)^2$	$(y_i-\bar{y}_3)^2$	$(y_i-\bar{y}_4)^2$	$(y_i-\bar{y}_5)^2$
2		13,44444444	14,44	14,5785124	4,515625	49
3		0,111111111	0,64	7,94214876	17,015625	1
4		11,11111111	33,64	3,305785124	78,765625	36
5			17,64	96,39669421	23,765625	
6			38,44	0,669421488	8,265625	
7				0,033057851	0,765625	
8				1,396694215	0,015625	
9				4,760330579	123,765625	
10				17,48760331		
11				26,85123967		
12				38,21487603		
13						
14						
15						
16	$\sigma_i^2 \cdot n_i =$	24,66666667	104,8	211,6363636	256,875	86
17						
18	$\sigma_i^2 =$	8,222222222	20,96	19,23966942	32,109375	28,66666667
19		*(групові дисперсії)*				

- до комірки M2 занести формулу =POWER(G2-G$18;2) і скопіювати її до комірок N2:Q2;
- комірку M2 скопіювати до комірок M3:M4;
- комірку N2 скопіювати до комірок N3:N6;
- комірку O2 скопіювати до комірок O3:O12;
- комірку P2 скопіювати до комірок P3:P9;
- комірку Q2 скопіювати до комірок Q3:Q4;

Обчислити добутки $\sigma_i^2 \cdot n_i$:

- до комірки M16 занести формулу =SUM(M2:M12) і скопіювати її до комірок N16:Q16.

Внутрішньогрупова дисперсія:

$$\overline{\sigma^2} = \frac{\sum \sigma_i^2 \cdot n_i}{n} = \frac{24{,}66 + 104{,}8 + 211{,}64 + 256{,}88 + 86}{30} \approx 22{,}8$$

- до комірки N22 занести формулу =SUM(M16:Q16)/H26.

По правилу складання дисперсій слід переконатися, що виконані розрахунки вірні:

$$\overline{\sigma^2} + \delta_y^2 = 22{,}8 + 183{,}37 = 206{,}17 = \sigma_y^2$$

- до комірки O28 занести формулу =N22+N25.

У результаті:

$$\eta^2 = \frac{\delta_y^2}{\sigma_y^2} = \frac{183{,}37}{206{,}17} \approx 0{,}8894$$

- до комірки L29 занести формулу =N25/H31.

Таким чином 88,94% варіації підсумкових балів обумовлено кількістю відвіданих занять. Решта варіації $(100 - 88{,}94 = 11{,}06\%)$ обумовлена іншими чинниками.

Слід зважити, що за коефіцієнтом η^2 можна міркувати про силу впливу причини на слідство, однак для оцінки цієї сили існує спеціалізований показник, а саме – *емпіричне кореляційне відношення*, що дорівнює квадратному кореню η^2. Фактично це ставлення відповідних стандартних відхилень:

$$\eta = \sqrt{\frac{\delta_y^2}{\sigma_y^2}} = \frac{\delta_y}{\sigma_y} \approx \sqrt{0{,}8894} \approx 0{,}94$$

- до комірки N31 заносим формулу =SQRT(N25/H31).

Це відношення теж змінюється в межах $0 \leq \eta \leq 1$ і характеризує *тісноту (силу) зв'язку* між ознаками. Для якісної оцінки цієї сили часто використовують так звану *шкалу Чеддока* (див. табл. А.9) [11]:

Таблиця А.9

Шкала Чеддока

Діапазон значень η	Кореляційна залежність Y від X
0 – 0,1	практично відсутня
0,1 – 0,3	слаба
0,3 – 0,5	помірна
0,5 – 0,7	помітна
0,7 – 0,9	сильна
0,9 – 0,99	дуже сильна
0,99 - 1	практично функціональна

У випадку завдання $\eta \approx 0{,}94$, таким чином, підсумковий бал дуже сильно залежить від кількості відвіданих занять.

Лабораторна робота 5. Завдання 2

Інструкція до виконання:

При виконанні завдання слід виходити з наступних міркувань. Очевидно, що чим більше студент прогулює заняття, тим ймовірніше, що в нього погана успішність. Однак, це може бути не завжди, оскільки успішність залежить багатьох чинників. Проте дійсно існує загальна тенденція у тому, що зі збільшенням кількості прогулів середня успішність студентів падатиме, яка є кореляційною.

За своїм *напрямом* залежність буває *прямою* («чим більше, тим більше») і *зворотною* («чим більше, тим менше»). У завданні висловлюємо припущення про наявність *зворотної* кореляційної залежності *Y* – успішності студентів від *X* кількості їх прогулів. Для перевірки слід:

1) Побудувати *діаграму розсіювання* (див. рис. А.13):

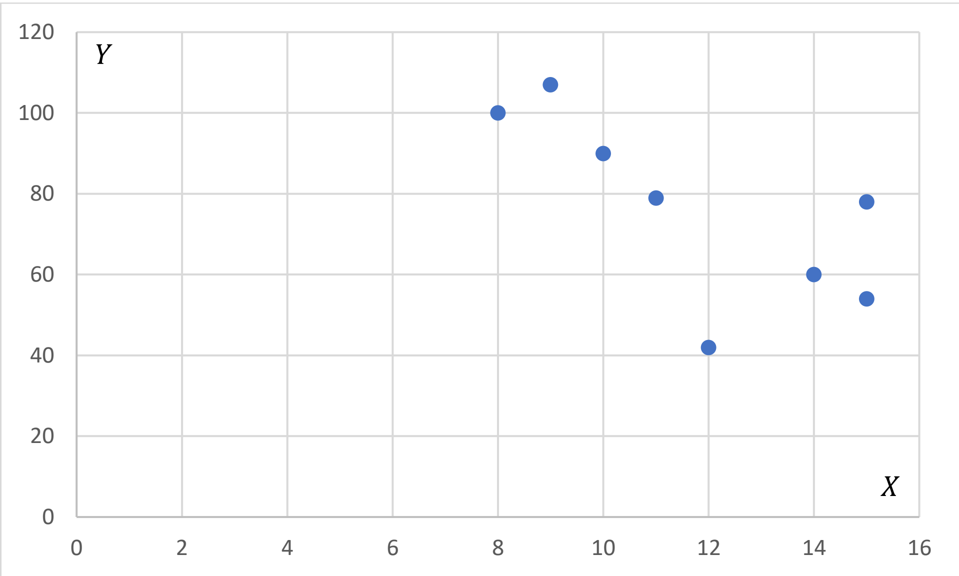


Рис. А.13. Діаграма розсіювання

За *діаграмою розсіювання* добре видно, що зі збільшенням числа прогулів успішність *переважно* падає, що підтверджує наявність зворотної кореляційної залежності успішності кількості прогулів. Більше того, майже всі точки розміщуються приблизно по прямій, що дає підставу припустити, що ця залежність близька до лінійної. Потрібно знайти рівняння прямої, яка проходить *максимально близько* до емпіричних точок, а також оцінити *тісноту* лінійної кореляційної залежності - наскільки близько розташовані ці точки до побудованої прямої.

2) Лінійний коефіцієнт кореляції слід обчислити за такою формулою:

$$r = \frac{cov(X;Y)}{\sigma_x \cdot \sigma_y},$$

де $\sigma_x \cdot \sigma_y$ – стандартні відхилення ознак X, Y;

$cov(X;Y)$ – *кореляційним моментом* чи *коефіцієнтом коваріації* (спільної варіації) ознак, він розраховується так:

$$cov(X;Y) = \frac{\sum(x_i - \bar{x}) \cdot (y_i - \bar{y})}{n} =,$$

де n – обсяг статистичної сукупності,

а $\bar{x}$, $\bar{y}$ – середні значення ознак.

Даний коефіцієнт показує, наскільки *узгоджено* відхиляються парні значення $(x_i;\ y_i)$ від своїх середніх у той чи інший бік.

Для виконання завдання слід заповнити розрахункову таблицю А.10:

Таблиця А.10

Обчислення добутків і сум добутків

	A	B	C	D	E–F	G	H
4	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - x)^2$	$(y_i - y)^2$
5	12	42	0,25	-34,25	-8,5625	0,0625	1173,0625
6	9	107	-2,75	30,75	-84,5625	7,5625	945,5625
7	8	100	-3,75	23,75	-89,0625	14,0625	564,0625
8	14	60	2,25	-16,25	-36,5625	5,0625	264,0625
9	15	78	3,25	1,75	5,6875	10,5625	3,0625
10	11	79	-0,75	2,75	-2,0625	0,5625	7,5625
11	10	90	-1,75	13,75	-24,0625	3,0625	189,0625
12	15	54	3,25	-22,25	-72,3125	10,5625	495,0625
13	94	610	:суми:		-311,5	51,5	3641,5

Спочатку слід розрахувати ліві нижні суми і середні значення ознак:

- до комірки А13 занести формулу =SUM(A5:A12) і скопіювати її до комірки В13;

$$\bar{x} = \frac{\sum x_i}{n} = \frac{94}{8} = 11{,}75,$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{610}{8} = 76{,}25,$$

- до комірок E17 і H17 занести формули =A13/F15 і =B13/F15.

Далі слід заповнити стовпці таблиці, що залишилися:

- до комірки C5 занести формулу =A5-$E$17 і скопіювати її до комірок C6:C12;
- до комірки D5 занести формулу =B5-$H$17 і скопіювати її до комірок D6:D12;
- до комірки E5 занести формулу =C5*D5 і скопіювати її до комірок E6:E12;
- до комірки G5 занести формулу =POWER(C5;2) і скопіювати її до комірки H5 і комірки G5:H5 скопіювати до G6:H6; G12:H12;
- до комірки E13 занести формулу =SUM(E5:F12) і до комірок G13:H13 занести формули =SUM(G5:G12) і =SUM(H5:H12).

Обчислити коефіцієнт коваріації:

$$cov(X;Y) = \frac{\sum(x_i - \bar{x}) \cdot (y_i - \bar{y})}{n} = \frac{-311{,}5}{8} = -38{,}938.$$

- до комірки H19 занести формулу =E13/F15.

Стандартні відхилення обчислити як квадратний корінь з дисперсій:

$$\sigma_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}} = \sqrt{\frac{51{,}5}{8}} \approx 2{,}5372;$$

$$\sigma_y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n}} = \sqrt{\frac{3641{,}5}{8}} \approx 21{,}3351.$$

- до комірки G22 занести формулу =SQRT(G13/F15);
- до комірки J22 занести формулу =SQRT(H13/F15).

Таким чином, коефіцієнт кореляції:

$$r = \frac{cov(X;Y)}{\sigma_x \cdot \sigma_y} \approx \frac{-38,938}{2,5372 \cdot 21,3351} \approx -0,7193.$$

- до комірки G25 занести формулу =H19/G22/J22.

У Microsoft Excel можна відразу обчислити цей коефіцієнт за допомогою вбудованої функції:

- до комірки H25 занести формулу =CORREL(A5:A12;B5:B12).

Коефіцієнт кореляції може змінюватися в межах $-1 \le r \le 1$ і чим він ближче по модулю до одиниці, тим тісніше лінійна кореляційна залежність – тим ближче розташовані точки до прямої, тим якісніша і достовірніша лінійна модель. Якщо $r = -1$ або $r = 1$, то йдеться про строгу лінійну залежність, при якій всі емпіричні точки виявляться на побудованій прямій. Навпаки, що ближче r до нуля, то точки розсіяні далі, тим лінійна залежність виражена менше.

Для оцінки тісноти зв'язку використовують *шкалу Чеддока* (див. табл. А.9) [11].

При цьому якщо $r < 0$ кореляційний зв'язок *зворотний*, а якщо $r > 0$, то *прямий*.

У випадку завдання $r \approx -0,72$, і таким чином, **існує сильна зворотна лінійна кореляційна залежність Y сумарної успішності від X кількості прогулів.**

Лінійний коефіцієнт кореляції показує не тільки тісноту, але ще й *напрямок* залежності, а й її *форму* (лінійну).

3) Коефіцієнт детермінації показує *частку варіації* ознаки-результату Y, яка обумовлена впливом ознаки-фактора X. Лінійний коефіцієнт детермінації може змінюватися в межах $0 \le R^2 \le 1$, і чим він ближче до одиниці, тим краще лінійна модель наближає емпіричні дані.

$$R^2 = r^2 \approx (-0,7193)^2 \approx 0,5174.$$

- до комірки F28 занести формулу =POWER(G25;2).

Таким чином, у рамках побудованої моделі успішність на 51,74% залежить від кількості прогулів. Решта варіації успішності (48,26%) обумовлена іншими причинами.

4) Рівняння лінійної регресії Y на X це оптимальна пряма $y = ax + b$, яка проходить максимально близько до емпіричних точок.

Якщо відомі значення r, σ_x, σ_y, то коефіцієнти рівняння $y = ax + b$ регресії розраховуються за такими формулами:

$$a = \frac{r \cdot \sigma_y}{\sigma_x} \approx \frac{-0{,}7193 \cdot 21{,}3351}{2{,}5372} \approx -6{,}0485;$$

$$b = \bar{y} - a\bar{x} \approx 76{,}25 - (-6{,}0485) \cdot 11{,}75 \approx 147{,}32.$$

- до комірки N25 занести формулу =G25*J22/G22;
- до комірки N25 заносимо формулу =H17-N25*E17.

Таким чином:

$$y = -6{,}0485x + 147{,}32.$$

Далі слід побудувати лінію регресії. Для цього на діаграмі слід додати лінію тренду, параметри лінії тренду: лінійна (див. рис А.14).

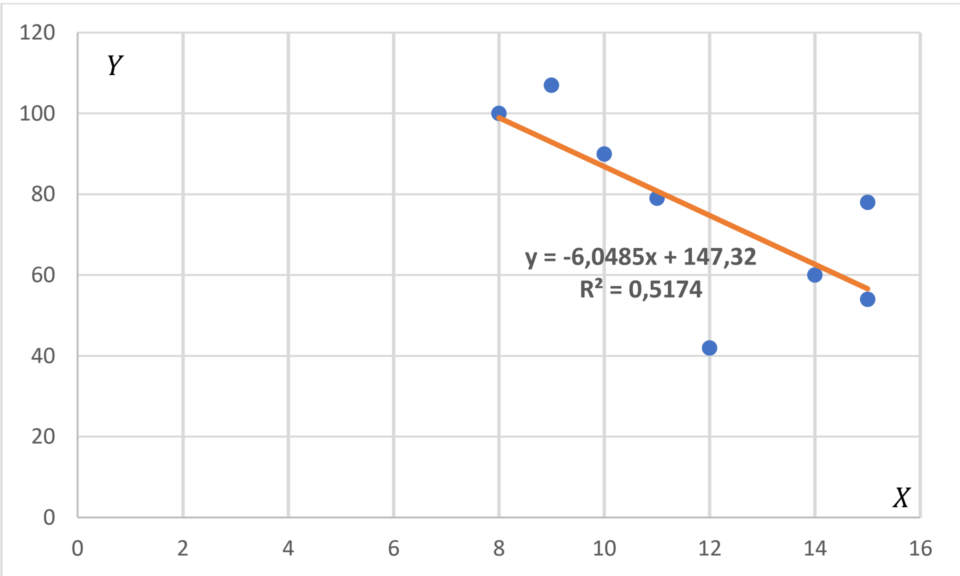


Рис А.14. Діаграма розсіювання і лінія регресії

ДОДАТОК А.6.

Лабораторна робота 6. Завдання 1

Інструкція до виконання:

В завданні дано вибіркову сукупність обсягом n = 59 спостережень. Для розвязання доцільно провести інтервальне розбиття. Для цього потрібно розмістити варіанти в одну колонку і відсортувати їх за зростанням. Далі слід знайти мінімальне та максимальне значення з використанням функцій МІН та МАХ та записати їх до комірок Н1 та Н2 відповідно. Обчислення розмаху варіації виконують за формулою: $R = x_{max} - x_{min}$ =H2-H1 (як довжину загального інтервалу, в межах якого варіюються варіанти). Далі його потрібно його розбити на часткові інтервали по формулі Стерджеса: k = 1 + 3.322lg n, де lg n – десятковий логарифм від обсягу вибірки та k – оптимальна кількість інтервалів, при цьому результат округляють до найближчого лівого цілого значення.

У випадку завдання: =LOG10(59)*3,322+1≈6,88≈6 інтервалів.

Довжини часткових інтервалів можуть бути різні, але в більшості випадків використовує рівноінтервальне угруповання: $h = \frac{x_{max}-x_{min}}{k}$ =H3/J4.

Для виконання завдання слід заповнити таблицю інтервального варіаційного ряду (див. табл. А.11).

Таблиця А.11

Інтервальний варіаційний ряд і розрахунки допоміжних значень

	D	E	F	G	H	I	J	K
1	Мінімальне значення			$x_{\min}=$	11,3			
2	Максимальне значення			$x_{\max}=$	15,2			
3	Розмах варіації		$R=x_{\max}-x_{\min}=$		3,9			
4	Оптимальна кількість інтервалів				$K=$	6,88	6	
5	Довжина інтервала			$h=$	0,65			
6								
7			Середини					
8	**Мережі інтервалів**		**інтервалів**	**Частота**				
9	початок	кінець	x_i	n_i	$x_i n_i$	$(x_i-\bar{x}_в)^2 n_i$	$(x_i-\bar{x}_в)^3 n_i$	$(x_i-\bar{x}_в)^4 n_i$
10	11,3	11,95	11,63	15	174,38	13,47	-12,76	12,09
11	11,95	12,6	12,28	23	282,33	2,04	-0,61	0,18
12	12,6	13,25	12,93	11	142,18	1,37	0,48	0,17
13	13,25	13,9	13,58	2	27,15	2,01	2,02	2,02
14	13,9	14,55	14,23	5	71,13	13,65	22,56	37,29
15	14,55	15,2	14,88	3	44,63	15,91	36,62	84,32
16	**Суми:**			59	741,78	48,44	48,32	136,07

Для розв’язання слід встановити межі інтервалів:

- скопіювати в комірку D10 x_{min} – початок першого інтервалу;
- комірка E10 – кінець першого інтервалу – $x_{min}+h$ = D10+$H$5;
- комірка D11 – початок другого інтервалу – кінець першого – E10;
- перетягнути E10 на одну комірку вниз – кінець другого інтервалу;
- виділити дві комірки D11 і E11 і перетягнути на чотири комірки вниз;

Завуважимо, що тут слід перевірити, щоб кінець п’ятого інтервалу збігався з x_{max}.

Для розрахунку середини інтервалів x_i до комірки F10 слід вставити формулу =(D10+E10)/2; і скопіювати до комірок F11:F15.

Розрахунок частоти за даними оцінок для кожного з інтервалів виконують так:

- виділити комірки G10:G15;
- вставити формулу =FREQUENCY(A2:A60;E10:E15);
- не натискати кнопку «ОК», а закінчити введення формули натисненням клавіш Ctrl+Shift+Enter (таблична формула);

- обов'язково переконатися втому, що нічого не втрачено: =SUM(G10:G15) $N = 59$ – обсяг вибірки.

Для розрахунку вибіркової середньої та дисперсії слід обчислити добутки $x_i n_i$, їх суму та середню:

- до комірки H10 занести формулу =F10*G10 і скопіювати її до комірок H11:H15;
- до комірки H16 занести формулу =SUM(H10:H15) (сума творів $x_i n_i$):
- до комірки F19 занести формулу =H16/G16 (вибіркова середня).

$$\bar{x}_{\text{в}} = \frac{\sum x_i n_i}{n} = \frac{741{,}78}{59} = 12{,}57$$

Дисперсію тут слід знайти за визначенням:

$$D_{\text{в}} = \frac{\sum (x_i - \bar{x}_{\text{в}})^2 \cdot n_i}{n}$$

Для цього розрахуємо добутки $(x_i - \bar{x}_{\text{в}})^2 n_i$ і $(x_i - \bar{x}_{\text{в}})^3 n_i$, $(x_i - \bar{x}_{\text{в}})^4 n_i$ (див. табл. 2.18).

- до комірки I10 занести формулу =POWER($F0-$F$19;2)*$G10;
- скопіювати формулу до комірок J10:K10, і в цих комірках поставити відповідну ступень (3 і 4);
- виділити комірки I10: J10:K10 і скопіювати їх до рядка I15: J15:K15;
- виділити комірку H16 і скопіювати її до комірок I16: J16:K16.

Розрахунки значень:

$$D_{\text{в}} = \frac{\sum (x_i - \bar{x}_{\text{в}})^2 \cdot n_i}{n} = \frac{48{,}44}{59} = 0{,}82$$

- до комірки F22 занести формулу =I16/G16 (дисперсія);

$$\sigma_{\text{в}} = \sqrt{D_{\text{в}} =} \sqrt{0{,}82} = 0{,}91$$

- до комірки F25 занести формулу =SQRT(F22) (середнє квадратичне відхилення).

Виправлена вибіркова дисперсія:

$$s^2 = \frac{n}{n-1} \cdot D_{\text{в}} = \frac{59}{59-1} \cdot 0{,}82 = 0{,}84$$

- до комірки F28 занести формулу =(G16/(G16-1))*F22,

і, відповідно *виправлене середнє квадратичне відхилення*:

$$s = \sqrt{s^2} = \sqrt{0{,}84} = 0{,}91$$

- до комірки F31 занести формулу =SQRT(F28).

Коефіцієнт варіації:

$$V = \frac{\sigma_{\text{в}}}{\bar{x}_{\text{в}}} \cdot 100\% = \frac{0{,}91}{12{,}57} * 100\% = 30{,}77\%$$

- до комірки F34 занести формулу =(F25/F19)*100.

Центральні моменти 3-го та 4-го порядків:

$$m_3 = \frac{\sum (x_i - \bar{x}_{\text{в}})^3 \cdot n_i}{n} = \frac{48{,}32}{59} = 0{,}82$$

- до комірки J19 занести формулу =J16/G16;

$$m_4 = \frac{\sum (x_i - \bar{x}_{\text{в}})^4 \cdot n_i}{n} = \frac{136{,}07}{59} = 2{,}31$$

- до комірки J22 занести формулу =K16/G16.

Обчислення коефіцієнту асиметрії:

$$A_s = \frac{m_3}{\sigma_{\text{в}}^3} = \frac{0{,}82}{(0{,}91)^3} = 1{,}1$$

- до комірки J25 занести формулу =J19/POWER(F25;3).

Обчислення коефіцієнту ексцесу:

$$E_k = \frac{m_4}{\sigma_{\text{в}}^4} - 3 = \frac{2{,}31}{(0{,}91)^4} - 3 = 0{,}42$$

- до комірки J28 занести формулу =-3+J22/POWER(F25;4).

Для рівня значущості $\alpha = 0{,}05$ та кількості ступенів свободи $k = n - 1 = 59 - 1 = 58$ за таблицею критичних точок розподілу Стьюдента слід знайти критичне значення:

$$t(\alpha, k) = t(0{,}05; 58) \approx 2$$

Похибка середнього арифметичного, як що $n > 30$:

$$m = \frac{\sigma}{\sqrt{n}} = \frac{0{,}91}{\sqrt{59}} = 0{,}12$$

- до комірки M28 занести формулу =F25/SQRT(G16).

Складання довірчогоінтервалу:

$$\bar{x}_{\text{в}} - m \cdot t < \bar{X} < \bar{x}_{\text{в}} + m \cdot t$$

$$12{,}34 < \bar{X} < 12{,}81$$

- до комірки J31 занести формулу =F19-M28*M25;
- до комірки L31 занести формулу =F19+M28*M25.

Похибка середнього квадратичного відхилення:

$$m_s = \frac{s}{\sqrt{2 \cdot n}} = \frac{0{,}91}{\sqrt{2 \cdot 59}} = 0{,}08$$

- до комірки M34 занести формулу =F31/SQRT(2*G16).

Похибка коефіцієнта варіації:

$$m_V = \frac{V}{\sqrt{2 \cdot n}} \cdot \sqrt{1 + 2\left(\frac{V}{100}\right)^2} = \frac{7{,}21}{\sqrt{2 \cdot 59}} \cdot \sqrt{1 + 2\left(\frac{7{,}21}{100}\right)^2} = 0{,}67$$

- до комірки M37 занести формулу =(F34/SQRT(2*G16))*(SQRT(1+2*(POWER((F34/100);2)))).

Похибка індикатора асиметрії:

$$m_{As} = \sqrt{\frac{6 \cdot n(n-1)}{(n-2) \cdot (n+1) \cdot (n+3)}} = \sqrt{\frac{6 \cdot 59(59-1)}{(59-2) \cdot (59+1) \cdot (59+3)}} = 0{,}31$$

- до комірки M40 занести формулу =SQRT((6*G16*(G16-1))/((G16-2)*(G16+1)*(G16+3))).

Похибка коефіцієнта ексцесу

$$m_{Ek} = \sqrt{\frac{24}{n}} = \sqrt{\frac{24}{59}} = 0{,}64$$

- до комірки M43 занести формулу =SQRT(24/G16).

Лабораторна робота 6. Завдання 2

Інструкція до виконання:

Оскільки вибірка 5%-на, вона становить 1/20-ю частина генеральної сукупності, отже, загальна кількість учнів становить:

$N = \frac{100\%}{5\%} \cdot n = 20 \cdot 100 = 2000$ учнів.

До комірки G11 занести формулу =100/5*D9.

Потрібно знайти *граничну помилку* вибірки $\Delta = t \cdot \mu$, де t – коефіцієнт довіри, відповідний довірчої ймовірності $\gamma = 0{,}9973$, і якщо вибірка *безповторна* і генеральна дисперсія не відома, то *середня помилка* розраховується по формулі $\mu = \sqrt{\frac{s^2}{n}\left(1 - \frac{n}{N}\right)}$. Далі потрібно знайти інтервал $\bar{x}_в - \Delta < \bar{x}_Г < \bar{x}_в + \Delta$, який із ймовірністю 99,73% накриє *генеральну середню* $\bar{x}_Г$ успішності учнів в м. Харків.

У завданні запропоновано інтервальний варіаційний ряд із відкритими крайніми інтервалами. Оскільки довжина часткового інтервалу становить h = 2 бали, то тут слід закрити інтервали: 2-4 та 10-12 балів.

Далі слід знайти середини x_i інтервалів (виконати перехід до дискретного ряду), добутки $x_i n_i$, $x_i^2 n_i$ та їх суми (див. табл. А.12).

Таблиця А.12

Розрахунок середин інтервалів, добутків та їх сум

	A	B	C	D	E	F
2	**Интервали**		x_i	n_i	$x_i n_i$	$x_i^2 n_i$
3	2	4	3	10	30	90
4	4	6	5	25	125	625
5	6	8	7	40	280	1960
6	8	10	9	20	180	1620
7	10	12	11	5	55	605
8	Сумми:			100	670	4900

- до комірки E4 занести формулу =C4*D4 і скопіювати її до комірок E5:E7;

- до комірки F4 занести формулу =C4*C4*D4 і скопіювати її до комірок F5:F7;
- до комірки D8 занести формулу =SUM(D4:D7) і скопіювати її до комірки E8: F8 (об'єм вибірки і сума творів $x_i n_i$ і $x_i^2 n_i$);
- до комірки C12 занести формулу =E8/D8 (вибіркова середня).

Обчислення вибіркової середньої:

$$\bar{x}_{\text{в}} = \frac{\sum x_i n_i}{n} = \frac{670}{100} = 6{,}7 \text{ балів}$$

Вибіркову дисперсію обчислюють за формулою:

$$D_{\text{в}} = \sigma_{\text{в}}^2 = \frac{\sum {x_i}^2 n_i}{n} - (\bar{x}_{\text{в}})^2 = \frac{4900}{100} - 6{,}7^2 = 4{,}11,$$

- до комірки D16 занести формулу =F9/D9-C12^2.

Виправлена дисперсія:

$$s^2 = \frac{n}{n-1} \cdot \sigma_{\text{в}}^2 = \frac{100}{99} \cdot 4{,}11 \approx 4{,}15,$$

- до комірки D19 занести формулу =D9/(D9-1)*D16.

Далі слід скласти довірчий інтервал $\bar{x}_{\text{в}} - \Delta < \bar{x}_{\Gamma} < \bar{x}_{\text{в}} + \Delta$ для оцінки генеральної середньої успішності учнів в м. Харків.

Обчислення граничної помилки $\Delta = t \cdot \mu$ виконують так: оскільки обсяг вибірки $n > 30$, то коефіцієнт довіри шукають із співвідношення $2\Phi(t) = \gamma$. Оскільки $\gamma = 0{,}9973$, то:

$$2\Phi(t) = \gamma \Rightarrow \Phi(t) = \frac{\gamma}{2} = \frac{0{,}9973}{2} = 0{,}49865,$$

- до комірки I15 занести формулу =H13/2.

За таблицею значень функції Лапласа далі можна визначити, що цьому значенню функції відповідає аргумент $t = 3$ [23].

Обчислення *середньої помилку* без повторної вибірки:

$$\mu = \sqrt{\frac{s^2}{n}\left(1 - \frac{n}{N}\right)} = \sqrt{\frac{4{,}15}{100}\left(1 - \frac{100}{2000}\right)} \approx 0{,}20,$$

- до комірки H20 занести формулу =SQRT(D19/D9*(1-D9/G11)).

Гранична помилка становить:

$$\Delta = t \cdot \mu = 3 \cdot 0{,}20 = 0{,}6 \text{ балів,}$$

- до комірки F23 занести формулу =I18*H20.

Довірчий інтервал:

$$\bar{x}_{\text{в}} - \Delta < \bar{x}_{\Gamma} < \bar{x}_{\text{в}} + \Delta$$

$$6{,}7 - 0{,}6 < \bar{x}_{\Gamma} < 6{,}7 + 0{,}6$$

$$6{,}1 < \bar{x}_{\Gamma} < 7{,}3 \text{ (балів)}$$

– межі, в яких очікується середня успішність учнів в м. Харків з ймовірністю $\gamma = 0{,}9973$.

Відповідь: $\Delta = 3 \cdot 0{,}20 = 0{,}6$ балів, $6{,}1 < \bar{x}_{\Gamma} < 7{,}3$ балів.

ДОДАТОК А.7.

Лабораторна робота 7. Завдання 1

Інструкція до виконання:

Слід врахувати, що процес перевірки статистичної гіпотези складається з наступних етапів:

1) Обробка вибіркових даних та висування основної H_0 та конкуруючої H_1 гіпотез.

2) Вибір статистичного критерію K (безперервної випадкової.

3) Вибір рівня значущості α.

4) Знаходження *критичного значення* $k_{кр}$ – значення випадкової величини K, яке залежить від обраного рівня значущості α та опціонально з інших параметрів. Критичне значення визначає *критичну область (*лівосторонню, правосторонню чи двосторонню) (див. рис. А.15):

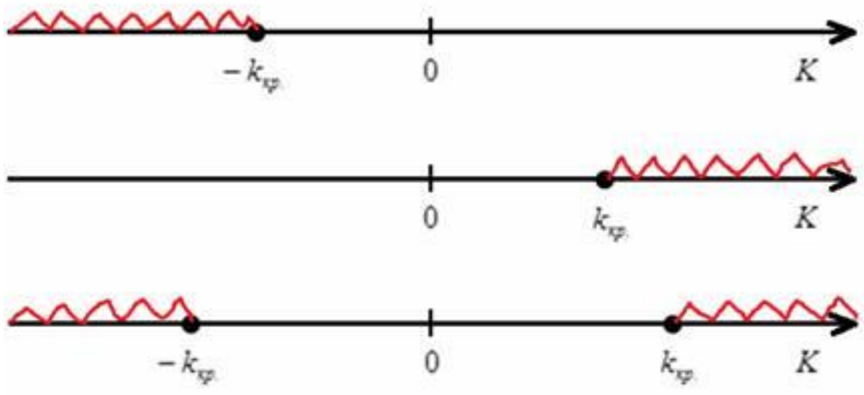


Рис. А.15. Критична область

Критична область – це область заперечення нульової гіпотези. Незаштриховану область на рис.2.33 називають областю прийняття гіпотези.

5) Виходячи з вибіркових даних розраховують *значення критерію*: $k_{спост}$.

У висновку, якщо $k_{спост}$ до критичної області не потрапляє, то гіпотеза H_0 на рівні значущості α приймається. При цьому існує ймовірність β того, що припущена помилка 2-го роду (прийно невірну гіпотезу).Якщо $k_{спост}$ потрапляє в критичну область, то гіпотеза H_0 на рівні значущості α відкидається (при цьому, якщо, наприклад, $\alpha = 0{,}05$ то в середньому в 5 випадках зі 100 відкитаться правильна гіпотеза, тобто зроблено помилку 1-го роду) [23].

У завданні слід перевірти гіпотезу про те, що математичне очікування випадкової величини $\overline{D} = \overline{X}_{\text{в}} - \overline{Y}_{\text{в}}$ (різниці між випадковими середніми) дорівнює нулю $H_0: M(\overline{D}) = a_0 = 0$ проти конкуруючої гіпотези $H_1: M = (\overline{D}) < 0$ (оскільки поліпшення підготовки студентів виражається більшим значенням Y і негативною різницею). Тобто критична область тут – лівостороння.

Так як генеральна дисперсія цієї випадкової величини не відома, то слід використати критерій $T = \frac{(\overline{D} - a_0)\sqrt{n}}{S} = \frac{\overline{D}\sqrt{n}}{S}$, де $\overline{D}$ випадкова різниця між вибірковими середніми і S відповідне виправлене стандартне відхилення, цей критерій має розподіл Стьюдента з кількістю ступенів свободи $k = n - 1$.

Для рівня значущості $\alpha = 0{,}05$ та $k = 9 -$ слід знайти критичне значення лівосторонньої критичної області (по нижньому рядку таблиці «Критичні точки розподілу Стьюдента» [21]:

$$t_{\text{кр}} = t_{\text{одн.кр}}(\alpha; k) = t_{\text{одн.кр}}(0{,}05; 8) \approx 1{,}86$$

При $t > -t_{\text{кр}}$ нульову гіпотезу приймають, а при $t < -t_{\text{кр}}$ – відкидають (див. рис. А.16):

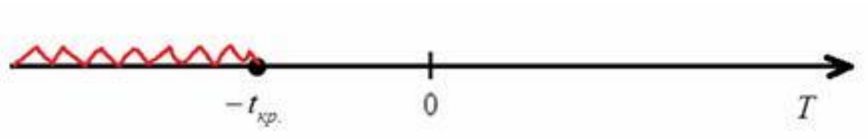


Рис. А.16. Лівостороння критична область

Для знаходження критерію, що спостерігається, $t_{\text{спост}} = \frac{\overline{D}\sqrt{n}}{S}$ потрібно розрахувати вибіркові характеристики – обчислити різниці між варіантами $d_i = x_i - y_i$, їх квадрати d_i^2 та суми (див. табл. А.13):

Таблиця А.13

Розрахунок різниці між варіантами

	A	B	C	D
1				
2	x_i	y_i	$d_i = x_i - y_i$	d_i^2
3	76	81	-5	25
4	71	85	-14	196
5	57	52	5	25
6	49	52	-3	9
7	70	70	0	0
8	69	63	6	36
9	26	33	-7	49
10	65	83	-18	324
11	59	62	-3	9
12	**Суммы:**		-39	673

- до комірки C3 занести формулу =A3-B3 і скопіювати її до комірок A4:A11;
- до комірки D3 занести формулу =POWER(C3;2) і скопіювати її до комірок D4:D11;
- до комірки C12 занести формулу =SUM(C3:C11) і скопіювати її до комірки D12.

Далі слід обчислити вибіркову середню різницю:

$$\bar{d}_{\text{в}} = \bar{x}_{\text{в}} - \bar{y}_{\text{в}} = \frac{\sum d_i}{n} = -\frac{39}{9} \approx -4{,}33,$$

- до комірки D15 занести формулу =C12/9.

Обчислити виправлене стандартне відхилення:

$$s_d = \sqrt{\frac{\sum d_i^2 - \frac{\left(\sum d_i\right)^2}{n}}{n-1}} = \sqrt{\frac{673 - \frac{(-39)^2}{9}}{8}} = \sqrt{\frac{508}{8}} = \sqrt{63},$$

- до комірки D19 занести формулу =SQRT((D12-((C12^2)/9))/H8).

Таким чином:

$t_{\text{спост}} = \frac{d_{\text{в}}\sqrt{n}}{s_d} = \frac{-4{,}33\cdot\sqrt{9}}{\sqrt{63}} \approx -1{,}64 > -t_{\text{кр}}$ тому **на рівні значущості 0,05 немає підстав відкидати гіпотезу** $H_0: M(\bar{D}) = 0$**.**

Таким чином, середня різниця $\bar{d}_{\text{в}} \approx -4{,}33$ між варіантами x_i *(*під час вступу*)* та відповідними варіантами y_i (після 1-го курсу) статистично незначна.

У відповіді отримаємо : на рівні значущості 0,05 немає підстав стверджувати, що після 2-го курсу рівень знання значно покращився.

Лабораторна робота 7. Завдання 2

Інструкція до виконання:

Заповнити розрахункову таблицю.

Розрахувати добутки $x_i n_i$, $x_i^2 n_i$ та їх суми (див. табл. А.14):

Таблиця А.14

Розрахунок добутків та їх сум

	A	B	C	D	E	F	G
2	x_i	n_i	$x_i n_i$	$x_i^2 n_i$	m_i	$x_i m_i$	$x_i^2 m_i$
3	2	3	6	12	1	2	4
4	3	9	27	81	11	33	99
5	4	4	16	64	8	32	128
6	5	2	10	50	3	15	75
7	**Суммы:**	18	59	207	23	82	306

- до комірки C3 занести формулу =A3*B3 і скопіювати її до комірок C4:C6;
- до комірки D3 занести формулу =A3*A3*B3 і скопіювати її до комірок D4:D6;
- до комірки F3 занести формулу =A3*E3 і скопіювати її до комірки F4:F6;
- до комірки G3 занести формулу =A3*A3*E3 і скопіювати її до комірок G4:G6;
- до комірки B7 занести формулу =SUM(B3:B6) і скопіювати її до комірок C7:G7.

.

Середній бал:

$$\bar{x}_{\text{в}} = \frac{\sum x_i n_i}{n} = \frac{59}{18} \approx 3{,}27,$$

- до комірки C11 занести формулу =C7/B7.

$$\bar{y}_{\text{в}} = \frac{\sum x_i m_i}{m} = \frac{82}{23} \approx 3{,}56,$$

- до комірки F11 занести формулу =F7/E7.

Вибіркові дисперсії:

$$D_x = \frac{\sum {x_i}^2 n_i}{n} - (\bar{x}_{\text{в}})^2 = \frac{207}{18} - 3{,}27^2 \approx 0{,}76,$$

- до комірки D15 занести формулу =D7/B7-C11^2.

$$D_y = \frac{\sum {x_i}^2 m_i}{m} - (\bar{y}_{\text{в}})^2 = \frac{306}{23} - 3{,}56^2 \approx 0{,}59,$$

- до комірки H15 занести формулу =G7/E7-F11^2.

Виправлені дисперсії:

$$s_x^2 = \frac{n}{n-1} \cdot D_x = \frac{18}{17} \cdot 0{,}76 \approx 0{,}80,$$

- до комірки C19 занести формулу =(B7/(B7-1))*D15.

$$s_y^2 = \frac{m}{m-1} \cdot D_y = \frac{23}{22} \cdot 0{,}59 \approx 0{,}62,$$

- до комірки F19 занести формулу =(E7/(E7-1))*H15.

1) На рівні значущості 0,1 слід перевірити гіпотезу $H_0{:}\, \sigma_x^2 = \sigma_y^2$ проти конкуруючої гіпотези $H_1{:}\, \sigma_x^2 \neq \sigma_y^2$. Для цього застосовують критерій $F = \frac{S_{\text{Б}}^2}{S_{\text{м}}^2}$, де $S_{\text{Б}}^2$ – більша виправлена дисперсія, а $S_{\text{м}}^2$ – менша.

Пошук правого критичного значення двосторонньої критичної галузі виконують так: для рівня значущості $\alpha = 0{,}1$ та числа ступенів свободи $\mathrm{k}_1 = \mathrm{n} - 1 = 18 - 1 = 17$, $\mathrm{k}_2 = \mathrm{m} - 1 = 23 - 1 = 22$ знаходять за допомогою таблиці критичні точки розподілу Фішера-Снедекора (F-розподіл) : $F_{\text{кр.прав}} = F_{\text{кр}}\left(\frac{\alpha}{2}, k_1, k_2\right) = F_{\text{кр}}(0{,}5; 17; 22) \approx 2{,}11$ [21].

Обчислення значення критерію спостереження:

$$F_{\text{спост}} = \frac{S_{\text{Б}}^2}{S_{\text{м}}^2} = \frac{s_x^2}{s_y^2} = \frac{0{,}80}{0{,}62} \approx 1{,}29 < F_{\text{кр}},$$

- до комірки D22 занести формулу =C19/F19.

У висновку до завдання на рівні значущості 0,1 гіпотезу $H_0: \sigma_x^2 = \sigma_y^2$ слід прийняти . Таким чином, групи є однорідними [23].

2) На рівні значимості 0,1 слід перевірити гіпотезу $H_0: \bar{x}_\Gamma = \bar{y}_\Gamma$ проти гіпотези $H_1: \bar{x}_\Gamma < \bar{y}_\Gamma$ у тому, що 1-я група навчається слабше. Досліджувані сукупності досить малі (n < 30, m < 30) та їхні генеральні дисперсії невідомі, але у попередньому пункті статистично обґрунтована незначна відмінність генеральних дисперсій. Тож для перевірки гіпотези можна використовувати критерій

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{(n-1)S_x^2 + (n-1)S_y^2}} \cdot \sqrt{\frac{nm(n+m-2)}{n+m}}$$

де $\bar{X},\ \bar{Y}$ – Випадкові значення вибіркових середніх, a S_x^2, S_y^2 – відповідні виправлені вибіркові дисперсії.

Оскільки конкуруюча гіпотеза має вигляд $H_1: \bar{x}_\Gamma < \bar{y}_\Gamma$, то критична область буде лівосторонньою. Для рівня значущості $\alpha = 0{,}1$ та числа ступенів свободи

$$k = n + m - 2 = 18 + 23 - 2 = 39$$

- до комірки J26 занести формулу =B7+E7-2.

Далі критичне значення односторонньої області: $t_{\text{кр.}} = t_{\text{кр.одн}}(\alpha, k) = t_{\text{кр.одн}}(0{,}1; 39) \approx 1{,}3$ [23].

При $t > -t_{\text{кр}}$ нульову гіпотезу приймають, а при $t < -t_{\text{кр}}$ – відкидають (див. рис. А.17):

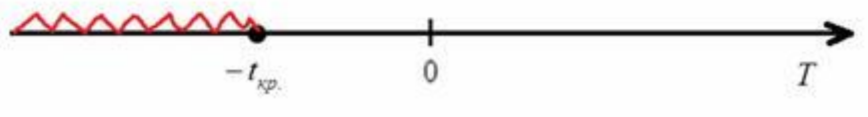


Рис. А.17. Критичне значення односторонньої області

Обчислення значення критерію показує:

$$t_{\text{спос}} = \frac{\bar{x}_{\text{в}} - \bar{y}_{\text{в}}}{\sqrt{(n-1)s_x^2 + (n-1)s_y^2}} \cdot \sqrt{\frac{nm(n+m-2)}{n+m}} = \frac{3{,}28 - 3{,}57}{\sqrt{17 \cdot 0{,}80 + 22 \cdot 0{,}62}} \cdot \sqrt{\frac{18 \cdot 23 \cdot 39}{41}} \approx -1{,}09 > -t_{\text{кр}},$$

тому на рівні значущості 0,1 немає підстав відкидати гіпотезу $H_0 \colon \bar{x}_{\Gamma} = \bar{y}_{\Gamma}$.

Таким чином, за результатами не можна стверджувати, що різниця між середніми оцінками $\bar{x}_{\text{в}} \approx 3{,}28$, $\bar{y}_{\text{в}} \approx 3{,}57$ обумовлена тим, що 1-а група слабкіша., - для перевірки цього припущення потрібен подальший моніторинг успішності.

Отже, на рівні значущості 0,1 немає підстав відкинути нульові гіпотези.

Лабораторна робота 7. Завдання 3

Інструкція до виконання:

На рівні значущості α слід перевірити гіпотезу H_0 про нормальний розподіл генеральної сукупності проти конкуруючої гіпотези H_1 (про те, що вона так не розподілена). Використовуємо критерій згоди Пірсона:

$$\chi^2 = \sum \frac{(n_i - n'_i)^2}{n'_i}$$

Емпіричні частоти відомі із запропонованого інтервального ряду і далі слід знайти теоретичні, для чого цього потрібно обчислити вибіркову середню $\bar{x}_\text{в}$ та вибіркове стандартне відхилення $\sigma_\text{в}$.

Слід обрати як *варіант* x_i середини часткових інтервалів (довжина кожного інтервалу h = 3) та заповнити розрахункову таблицю А.15 .

Таблицю А.15

Розрахунок добутків та їх сум

	A	B	C	D	E	F
8	*Інтервали*		x_i	n_i	$x_i n_i$	$x_i^2 n_i$
9	7,5	10,5	9	2	18	162
10	10,5	13,5	12	6	72	864
11	13,5	16,5	15	10	150	2250
12	16,5	19,5	18	17	306	5508
13	19,5	22,5	21	33	693	14553
14	22,5	25,5	24	11	264	6336
15	25,5	28,5	27	9	243	6561
16	28,5	31,5	30	7	210	6300
17	31,5	34,5	33	5	165	5445
18	h =	3	n =	100	2121	47979

Обчислення вибіркового середнього:

$$\bar{x}_\text{в} = \frac{\sum x_i n_i}{n} = \frac{2121}{100} = 21{,}21 \text{ бал},$$

- до комірки I4 занести формулу =E18/D18.

Вибіркове стандартне відхилення:

$$\sigma_\text{в} = \sqrt{D_\text{в}} = \sqrt{\frac{\sum x_i^2 n_i}{n} - (\bar{x}_\text{в})^2} = \sqrt{\frac{47979}{100} - (21{,}21)^2} = \sqrt{29{,}9259} \approx 5{,}47$$

балів,

- до комірки Р4 занести формулу =SQRT(F18/D18-I4^2).

Через великий обсяг вибірки його виправленням можна знехтувати.

Теоретичні частоти розраховуються за такою формулою:

$n'_i = \frac{h \cdot n}{\sigma_в} \cdot f(z_i)$, де $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$ – функція Гаусса, а $z_i = \frac{x_i - \bar{x}_в}{\sigma_в}$.

Вхідні дані відомі: $n = 100, h = 3, \bar{x}_в = 21{,}21, \sigma_в \approx 5{,}47$ тому можна продовжувати заповнювати таблицю за прикладом першого рядка (див. табл. А.16):

Таблиця А.16

Розрахунок теоретичних частот

	C	D	E	F	G	H	I
8	x_i	n_i	$x_i n_i$	$x_i^2 n_i$	z_i	$f(z_i)$	n'_i
9	9	2	18	162	-2,232	0,033	1,81
10	12	6	72	864	-1,6836	0,0967	5,3
11	15	10	150	2250	-1,1352	0,2095	11,49
12	18	17	306	5508	-0,5868	0,3358	18,42
13	21	33	693	14553	-0,0384	0,3986	21,86
14	24	11	264	6336	0,51	0,3503	19,21
15	27	9	243	6561	1,0584	0,2279	12,5
16	30	7	210	6300	1,6068	0,1097	6,02
17	33	5	165	5445	2,1552	0,0391	2,14
18	$n =$	100	2121	47979			98,75

Розрахунки для рядка:

$$z_1 = \frac{x_1 - \bar{x}_в}{\sigma_в} \approx \frac{9 - 21{,}21}{5{,}47} \approx -2{,}232$$

- до комірки G9 занести формулу =(C9-$I$4)/$P$4 і скопіювати її до комірок G10:G17.

$$f(z) \approx \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} = f(-2{,}232) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(-2{,}232)^2}{2}} \approx 0{,}033$$

– тут слід використовувати вбудовану функцію Excel: =NORMDIST(-2,232;0;1;0), перший аргумент якої дорівнює поточному значенню z_i.

- до комірки H9 занести формулу =NORMDIST(G9;0;1;0) і скопіювати її до комірок H10:H17.

Теоретична частота:

$$n'_i = \frac{h \cdot n}{\sigma_\text{в}} \cdot f(z_i) \approx \frac{3 \cdot 100}{5{,}47} \cdot 0{,}033 \approx 1{,}81,$$

- до комірки I9 занести формулу =$B$18*$D$18/$P$4*H9 і скопіювати її до комірок I10:I17, обчислити суму: до I18 занести формулу =SUM(I9:I17).

Побудова емпіричної гістограми та теоретичної кривої, яка проходить через крапки x_i, n'_i представлена нище (див. рис. А.18.):

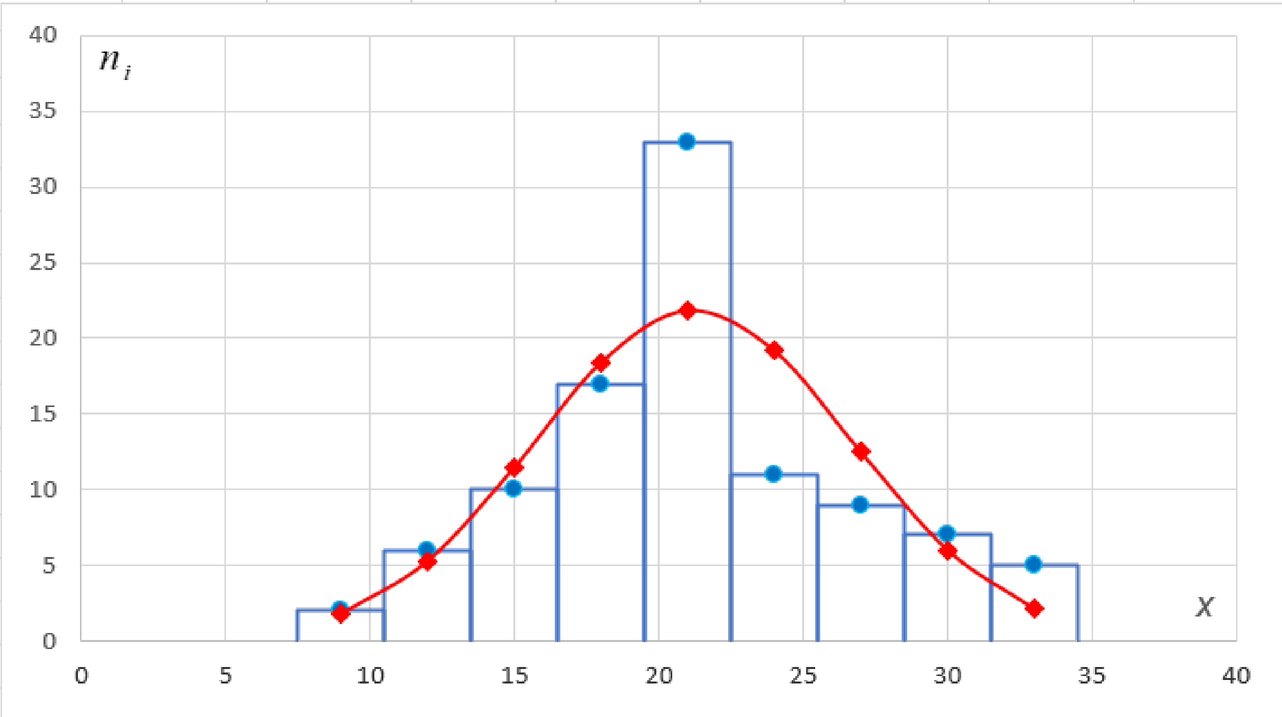


Рис. А.18. Емпірична гістограма та теоретична крива

Висота «сходів» дорівнює безпосередньо частотам. Нормальна крива побудована на основі вибіркових даних *(вибіркової середньої та стандартного відхилення)*, вона проходить через крапки і найкраще наближає гістограму. При цьому сума теоретичних частот виявилася трохи меншою за обсяг вибірки.

$$n_i(\, x_i, n'_i) \sum n'_i \approx 98{,}75$$

$$\sum n_i = 100$$

Тому, що емпірична гістограма *кінцева*, а нормальна крива – *нескінченна*, і невеликий «недобір» теоретичних частот посідає ділянки, що ліворуч і праворуч від гістограми.

Подальше завдання полягає в тому, щоб оцінити, наскільки **значимо** відрізняються емпіричні частоти *(сходинки гістограми)* від відповідних теоретичних частот *(рівень червоних точок)*.

На початку слід поєднати інтервали з малими (менше п'яти) частотами. В випадку завдання слід поєднати два перші та два останні інтервали (див. табл. А.17):

Таблиця А.17

Поєднання інтервалів

	D	I	J	K	L
8	n_i	n'_i		n_i	n'_i
9	2	1,81		8	7,12
10	6	5,3		10	11,49
11	10	11,49		17	18,42
12	17	18,42		33	21,86
13	33	21,86		11	19,21
14	11	19,21		9	12,5
15	9	12,5		12	8,16
16	7	6,02			
17	5	2,14			
18	100	98,75			

Це потрібно для того, щоб згладити невиправдано велику розбіжність між малими частотами по краях вибірки.

Далі слід знайти критичне значення $\chi^2_{\text{кр}} = \chi^2_{\text{кр}}(\alpha; k)$ критерію згоди Пірсона. Кількість ступенів свободи визначається за формулою свободи $k = m - r - 1$, де m – кількість інтервалів, а r – кількість параметрів закону аналізованого закону розподілу.

Після поєднання $m = 7$. У нормального закону оцінюємо $r = 2$ параметра.

Таким чином, $k = 7 - 2 - 1 = 4$ і для рівня значущості $\alpha = 0{,}05$:

$$\chi^2_{\text{кр}} = \chi^2_{\text{кр}}(0{,}05; 4) \approx 9{,}4877$$

Це значення можна знайти за таблицею критичних значень розподілу хі-квадрат [21], а також за допомогою вбудованої функції Excel CHIINV:

- до комірки Q21 занести формулу =CHIINV(0,05;4) [21].

При $\chi^2_{\text{досл}} > \chi^2_{\text{кр}}$ нульова гіпотеза відкидається, а $\chi^2_{\text{досл}} < \chi^2_{\text{кр}}$ таких підстав немає (див. рис. А.19):

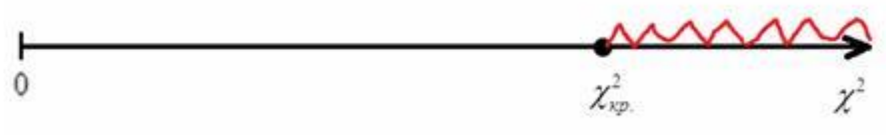


Рис. А.19. Критична область $\chi^2_{\text{кр}}$

Обчислення значення критерію

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - n'_i)^2}{n'_i}$$

– *розбіжностей між частотами* (див. табл. А.18):

Таблиця А.18

Розрахунок розбіжностей

	K	L	M
7			
8	n_i	n'_i	$\frac{(n_i - n'_i)^2}{n'_i}$
9	8	7,12	0,11
10	10	11,49	0,1923
11	17	18,42	0,1092
12	33	21,86	5,6746
13	11	19,21	3,5087
14	9	12,5	0,9778
15	12	8,16	1,8053
16		$\chi^2_{набл.} \approx$	12,378

Приклад розрахунку: $\frac{(n_1 - n'_1)^2}{n'_1} = \frac{(8-7,12)^2}{7,12} \approx 0,11$,

- до комірки M9 занести формулу =POWER(K9-L9;2)/L9 і скопіювати її до комірок M10:M15, обчислити суму у M16 за формулою =SUM(M9:M15).

У нижньому рядку таблиці готове значення $\chi^2_{\text{досл}} \approx 12,378 > \chi^2_{\text{кр}}$, тому на рівні значущості 0,05 гіпотезу H_0 про нормальний розподіл генеральної сукупності відкидають. При цьому з ймовірністю 5% наявна помилка 1-го роду.

Отже, на рівні значимості 0,05 гіпотезу про нормальний розподіл відкидають.

ДОДАТОК А.8.

Лабораторна робота 8. Завдання 1

Інструкція до виконання:

При виконанні завдання слід виходити з наступних міркувань:

1) Насамперед потрібно обґрунтувати причинно-наслідковий зв'язок між ознаками. Очевидно, що чим більша кількість годин, витрачених на самостійну підготовку до контрольної роботи тим більше кількість правильних відповідей. Однак це не є непорушним правилом. Тим не менш, *загальна тенденція* полягає в тому, що при збільшенні кількості годин на підготовку *середня* кількість правильних відповідей зростає. Така *нежорстка* залежність називається *кореляційною.*

Таким чином, можна припустити наявність *прямої кореляційної залежності* кількості правильних відповідей (ознака-результат Y) від кількісті годин, витрачених на самостійну підготовку (ознака-фактор X).

Якщо частоти мають тенденцію розташовуватися по діагоналі **від лівого верхнього до правого нижнього кута**, між ознаками існує *пряма кореляційна залежність*.

Для визначення *форми* залежності найпростіший спосіб – графічний, для чого потрібно модифікувати вихідну таблицю, а саме перейти від інтервальних варіаційних рядів *(лівий стовпець і 2-й зверху рядок)* до дискретних, вибравши як варіант x_i і y_j середину відповідних інтервалів (див. табл. А.19):

Таблиця А.19

Перехід до дискретних рядів і розрахунок частот

	A	B	C	D	E	F
4	X, x_i			Y, y_j		Разом, n_i
5		5	7	9	11	
6	3	2				2
7	4	6	3			9
8	5	2	5	7		14
9	6		2	2	3	7
10	7			1	7	8
11	Разом, m_j	10	10	10	10	**40**

Підрахунок суми частот:

$$\sum n_i = 2 + 9 + 14 + 7 + 8 = 40 = n$$

$$\sum m_j = 10 + 10 + 10 + 10 = 40 = n$$

- до комірки F6 занести формулу =SUM(B6:E6) і скопіювати її до комірок F7:F10;
- до комірки B11 занести формулу =SUM(B6:B10) і скопіювати її до комірок C11:E11;
- до комірки F11 занести формулу =SUM(F6:F10).

Для побудови кореляційного поля слід заповнити допоміжну таблицю, де верхній рядок – x_i, а наступні рядки – значення y_j, які відповідають не нульовим значенням частот (див. табл. А.20).

Таблиця А.20

Допоміжна таблиця

3	4	5	6	7
5	5	8		
	7	7	7	
		9	9	9
			11	11

За допомогою вставлення «Діаграма» слід обрати точкову діаграму (див. рис А.20).

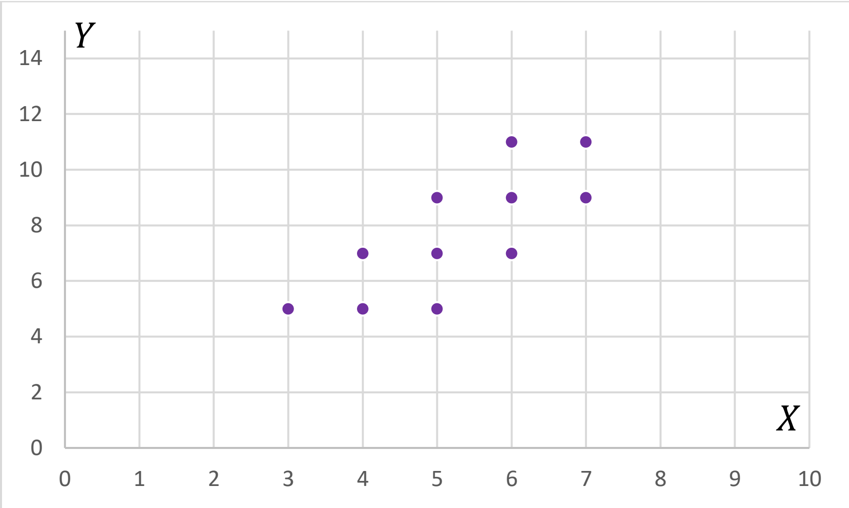


Рис А.20. Кореляційне поле

Аналізуючи креслення, можна зробити висновок, що емпіричні точки $(x_i;\ y_j)$ розмішуються приблизно по прямій, що дозволяє припустити наявність *лінійної* кореляційної залежності Y – кількість правильних відповідей від X кількісті годин, витрачених на підготовку.

Подальші дії полягають у тому, щоб знайти рівняння лінійної регресії $y = ax + b$, графік якої проходить максимально близько до емпіричних точок, а також оцінити тісноту лінійної кореляційної залежності – наскільки близько розташовані точки до побудованої прямої. Ця тіснота оцінюється за допомогою лінійного коефіцієнта кореляції.

2) Коефіцієнт кореляції обчислюють за формулою:

$$r_{xy} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y}.$$

Спочатку слід знайти середні $\bar{x},\ \bar{y}$ та стандартні відхилення $\sigma_x,\ \sigma_y$.

З комбінаційної таблиці слід виписати значення x_i, n_i і заповнити розрахункову таблицю А.21:

Таблиця А.21

Розрахункова таблиця ознаки-фактора X

	A	B	C	D	E	F	G
14	x_i	3	4	5	6	7	Суми:
15	n_i	2	9	14	7	8	40
16	$x_i n_i$	6	36	70	42	56	210
17	$x_i^2 n_i$	18	144	350	252	392	1156

- до комірки B16 занести формулу =B14*B15 і скопіювати її до комірок C16:F16;
- до комірки B17 занести формулу =B14*B14*B15 і скопіювати її до комірок C17:F17;
- до комірки G15 занести формулу =SUM(B15:F15) і скопіювати її до комірок G16:G17.

Далі слід обчислити середнє значення:

$$\bar{x} = \frac{\sum x_i n_i}{n} = \frac{210}{40} = 5{,}25,$$

- до комірки J15 занести формулу =G16/G15.

Середнє квадратичне відхилення, як корінь дисперсії, обчисленої за формулою:

$$\sigma_x = \sqrt{\frac{\sum x_i^2 n_i}{n} - \bar{x}^2} = \sqrt{\frac{1156}{40} - (5{,}25)^2} = \sqrt{1{,}3375} \approx 1{,}1565,$$

- до комірки J18 занести формулу =SQRT(G17/G15-J15*J15).

Так само заповнюють розрахункову таблицю для ознаки-результату Y (див. табл. А.22):

Таблиця А.22

Розрахункова таблиця ознака-результат Y

	A	B	C	D	E	F
20	y_j	5	7	9	11	Суми:
21	m_j	10	10	10	10	40
22	$y_j m_j$	50	70	90	110	320
23	$y_j^2 m_j$	250	490	810	1210	2760

- до комірки B22 занести формулу =B20*B21 і скопіювати її до комірок C22:E22;
- до комірки B23 занести формулу =B20*B20*B21 і скопіювати її до комірок C23:E23;
- до комірки F21 занести формулу =SUM(B21:E21) і скопіювати її до комірок F22:F23.

Далі слід обчислити середнє значення правильних відповідей:

$$\bar{y} = \frac{\sum y_i m_j}{n} = \frac{320}{40} = 8$$

- до комірки J21 занести формулу =F22/F21.

Середнє квадратичне відхилення:

$$\sigma_y = \sqrt{\frac{\sum y_i^2 m_j}{n} - \bar{y}^2} = \sqrt{\frac{2760}{40} - (8)^2} = \sqrt{5} \approx 2{,}2361,$$

- до комірки J18 занести формулу =SQRT(F23/F21-J21*J21).

Для пошуку середнього значення $\overline{xy}$ (добутку ознак) слід обчислити всі можливі добутки x_i і y_j на відповідні ненульові частоти n_{ij} (див. табл. А.23):

Таблиця А.23

Розрахунок добутків $x_i y_j n_{ij}$ (приклад розрахунку виділено рожевим)

	A	B	C	D	E	G	H	I	J	K
4	X, x_i	Y, y_j								
5		5	7	9	11		Розрахунок творів		$x_i y_j n_{ij}$	
6	3	2					30			
7	4	6	3				120	84		
8	5	2	5	7			50	175	315	
9	6		2	2	3			84	108	198
10	7			1	7				63	539

- До комірки H6 занести формулу =$A6*B$5*B6 і скопіювати її до комірок I6:K6;
- комірки H6:K6 скопіювати до комірок H7:K10.

Обчислимо суму цих добутків:

$\sum\sum x_i y_j n_{ij}$ =30+120+84+50+175+315+84+108+198+63+539=1766

- до комірки L11 занести формулу =SUM(H6:K10),

і середню шукану:

$$\overline{xy} = \frac{\sum\sum x_i y_j n_{ij}}{n} = \frac{1766}{40} = 44{,}15$$

- до комірки J12 занести формулу =L11/F11.

Таким чином, лінійний коефіцієнт кореляції:

$$r_{xy} = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\sigma_x\cdot\sigma_y} = \frac{44{,}15 - 5{,}25\cdot 8}{1{,}1565\cdot 2{,}2361} \approx 0{,}8314,$$

- до комірки C27 занести формулу =(J12-J15*J21)/J18/J24.

В результаті отримано позитивне число і, згідно з шкалою Чеддока (таблиця А.9), існує пряма лінійна кореляційна залежність Y кількість правильних відповідей від X кількості годин, витрачених на підготовку.

Обчислення коефіцієнту детермінації:

$$R^2 = \left(r_{xy}\right)^2 \approx (0{,}8314)^2 \approx 0{,}6912,$$

- до комірки C32 занести формулу =POWER(C27;2).

Таким чином, у рамках побудованої моделі кількість правильних відповідей на 69,12% залежить від кількості годин, витрачених на підготовку. Інші $100 - 69{,}12 = 30{,}88\%$ варіації обумовлено іншими факторами.

3) Пошук рівняння $y = ax + b$ *лінійної регресії* Y на X.

$$a = \frac{r\cdot\sigma_y}{\sigma_x} \approx \frac{0{,}8314\cdot 2{,}2361}{1{,}1565} \approx 1{,}6075;$$

$$b = \bar{y} - a\bar{x} \approx 8 - 1{,}6075\cdot 5{,}25 \approx -0{,}4393.$$

- До комірки G30 занести формулу =C27*J24/J18;
- до комірки J30 занести формулу =J21-G30*J15.

Таким чином:

$$y = -1{,}6075x + 0{,}4393 = 1{,}6075x - 0{,}4393.$$

Пошук трьох зручних точок для побудови графіка:

$$x = 2 \Rightarrow y \approx 1{,}61 \cdot 2 - 0{,}44 = 2{,}78;$$

$$x = 5 \Rightarrow y \approx 1{,}61 \cdot 2 - 0{,}44 = 7{,}6;$$

$$x = 8 \Rightarrow y \approx 1{,}61 \cdot 2 - 0{,}44 = 12{,}42.$$

Ці точки слід відзначити на кресленні *(червоний колір)* і провести лінію регресії (див. рис А.21):

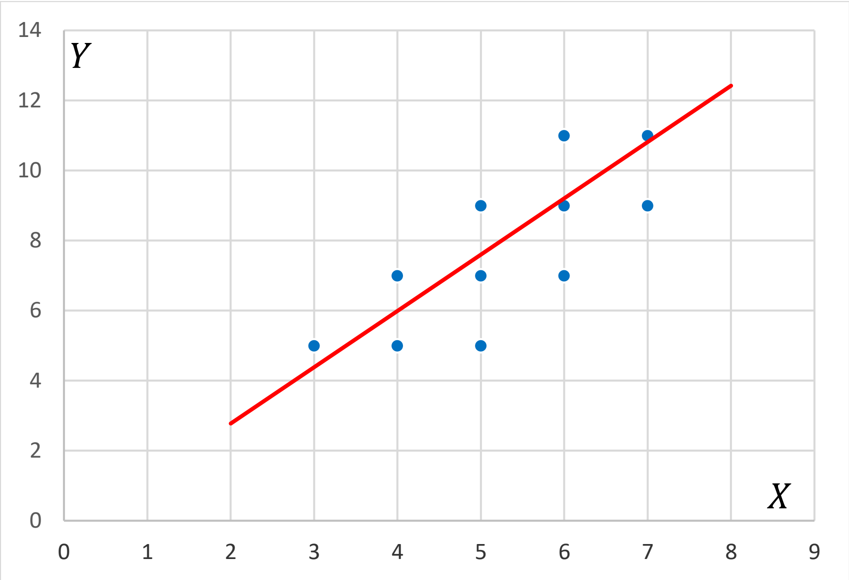


Рис А.21. Лінія регресії

Прогнозування середньої кількості правильних відповідей якщо на підготовку було витрачено 9 годин:

$$x = 9 \Rightarrow y \approx 1{,}61 \cdot 9 - 0{,}44 = 14{,}907$$

- до комірки P31 занести формулу =G30*P30-J30.

Отже, рівняння регресії повертає <u>середньоочікуване</u> значення $\tilde{y}(x)$ ознаки-результату при різних значеннях ознаки-фактора.

Лабораторна робота 8. Завдання 2

Інструкція до виконання:

Інструкція до виконання:

При виконанні завдання слід виходити з наступних міркувань:

1) Завдання побудови діаграми розсіювання полягає в тому, щоб підібрати лінію (її тип), яка вдало наблизить емпіричні точки. Відповідний тип лінії виявляють прямим перебором, за допомогою опції «лінія тренду», основних графіків – методом найменших квадратів будують оптимальну пряму, параболу, гіперболу, експоненту тощо, і аналізують, яка функція краще наближає емпіричні точки (див. рис. А.22). Якість наближення оцінюють за допомогою *індексу детермінації* R_I^2 *(що більше до одиниці, тим краще)* та середньої помилки апроксимації $\bar{A}$ *(чим ближче до нуля, тим краще)*.

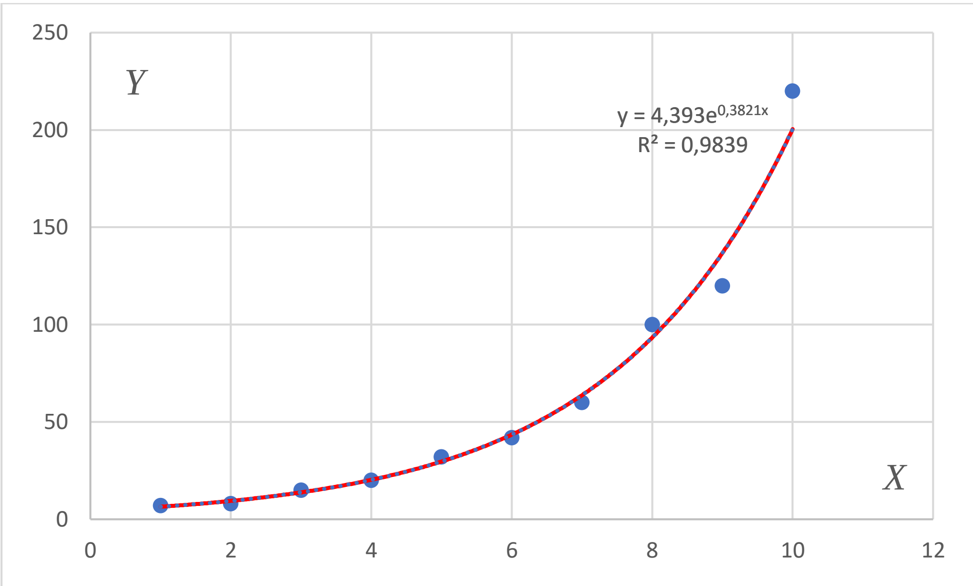


Рис А.22. Лінія регресії

У завданні найкращим вибором є експонента $y = be^{ax}$, а саме $y = 4{,}393e^{0{,}3821x}$.

2) Методом найменших квадратів слід знайти рівняння *нелінійної експоненційної регресії* $y = be^{ax}$.

Коефіцієнти a та b визначають з системи:

$$\begin{cases} a\sum x_i^2 + B\sum x_i = \sum x_i \ln y_i \\ a\sum x_i + Bn = \sum \ln y_i \end{cases},$$

де $b = e^B$.

Для вирішення слід заповнити розрахункову таблицю А.24:

Таблиця А.24

Розрахункова таблиця

	A	B	C	D	E
3	x_i	y_i	x_i^2	$\ln y_i$	$x_i \ln y_i$
4	1	7	1	1,9459	1,94591
5	2	8	4	2,0794	4,15888
6	3	15	9	2,7081	8,12415
7	4	20	16	2,9957	11,9829
8	5	32	25	3,4657	17,3287
9	6	42	36	3,7377	22,426
10	7	60	49	4,0943	28,6604
11	8	100	64	4,6052	36,8414
12	9	120	81	4,7875	43,0874
13	10	220	100	5,3936	53,9363
14	55	624	385	35,8132	228,492
15	$n =$ 10				

- До комірки С4 занести формулу =POWER(A4;2) і скопіювати її до комірок С5:С13;
- до комірки D4 занести формулу =LN(B4) і скопіювати її до комірок D5:D13;
- до комірки E4 занести формулу =A4*LN(B4) і скопіювати її до комірок E5:E13.

Обчислення суми:

- до комірки A14 занести формулу =SUM (A4:A13) і скопіювати її до комірок B14:E14.

Таким чином, отримано систему:

$$\begin{cases} a\sum x_i^2 + B\sum x_i = \sum x_i \ln y_i \\ a\sum x_i + Bn = \sum \ln y_i \end{cases} \Rightarrow \begin{cases} 385a + 55B = 228{,}492 \\ 55a + 10B = 35{,}8132 \end{cases}$$

Систему слід вирішити за формулами Крамера. Головний визначник:

$$\Delta = \begin{vmatrix} 385 & 55 \\ 55 & 10 \end{vmatrix} = 385 \cdot 10 - 55 \cdot 55 = 825 \neq 0,$$

отже, система має єдине рішення.

$$\Delta_a = \begin{vmatrix} 228{,}492 & 55 \\ 35{,}8132 & 10 \end{vmatrix} = 228{,}492 \cdot 10 - 35{,}8132 \cdot 55 = 315{,}196$$

$$a = \frac{\Delta_a}{\Delta} = \frac{315{,}194}{825} \approx 0{,}382$$

$$\Delta_B = \begin{vmatrix} 385 & 228{,}492 \\ 55 & 35{,}8132 \end{vmatrix} = 385 \cdot 35{,}8132 - 55 \cdot 228{,}492 = 1221$$

$$B = \frac{\Delta_B}{\Delta} = \frac{1221}{825} \approx 1{,}48.$$

Безпосередньо коефіцієнт:

$$b = e^B \approx e^{1{,}48} \approx 4{,}393.$$

- до комірки D17 занести формулу =C14*A15-A14*A14;
- до комірки D20 занести формулу =E14*A15-D14*A14;
- до комірки D23 занести формулу =D20/D17;
- до комірки D26 занести формулу =C14*D14-A14*E14;
- до комірки D29 занести формулу =D26/D17;
- до комірки D32 занести формулу =EXP(D29).

В результаті, шукана експонента: $y = 4{,}393e^{0{,}382x}$ [23].

3) Пошук *індекса детермінації* та *індекса кореляції*. Для цього слід обчислити середнє значення *ознаки-результату:*

$$\bar{y} = \frac{\sum y_i}{n} = 62{,}4$$

- до комірки H17 занести формулу =B14/A15.

Далі необхідно заповнити розрахункову таблицю А.25, одразу з додатковим стовпцем для розрахунку *СОА* у пункті 5:

Таблиця А.25

Розрахункова таблиця

	A	B	F	G	H	I
3	x_i	y_i	$(y_i-\bar{y})^2$	$y(x_i)=4{,}393e^{0{,}382x_i}$	$(y_i-y(x_i))^2$	$\left\lvert\frac{y_i-y(x_i)}{y_i}\right\rvert$
4	1	7	3069,16	6,4370	0,3169	0,0804
5	2	8	2959,36	9,4321	2,0510	0,1790
6	3	15	2246,76	13,8208	1,3904	0,0786
7	4	20	1797,76	20,2516	0,0633	0,0126
8	5	32	924,16	29,6745	5,4077	0,0727
9	6	42	416,16	43,4819	2,1961	0,0353
10	7	60	5,76	63,7138	13,7921	0,0619
11	8	100	1413,76	93,3594	44,0977	0,0664
12	9	120	3317,76	136,7989	282,2035	0,1400
13	10	220	24837,76	200,4506	382,1801	0,0889
14	55	624	40988,4		733,6990	0,8157
15	$n=$ 10					

- До комірки F4 занести формулу =POWER(B4-$H$17;2) і скопіювати її до комірок F5:F13;
- до комірки G4 занести формулу =$D$32*EXP($D$23*A4) і скопіювати її до комірок G5:G13;
- до комірки H4 занести формулу =POWER(B4-G4;2) і скопіювати її до комірок H5:H13;
- до комірки I4 занести формулу =ABS(B4-G4)/B4 і скопіювати її до комірок I5:I13.

Підрахуємо суми:

- До комірки F14 занести формулу =SUM(F4:F13) і скопіювати її до комірок H14:I14.

У результаті, загальна сума квадратів

$$Q=\sum(y_i-\bar{y})^2=40988{,}4,$$

залишкова сума квадратів

$$Q_e=\sum(y_i-y(x_i))^2\approx 733{,}6990$$

та *індекс детермінації*:

$$R_I^2 = 1 - \frac{Q_e}{Q} = 1 - \frac{733,6990}{40988,4} = 0,9821.$$

- комірка H20 – =F14, комірка H23 – =H14;
- до комірки H25 занести формулу =1-H23/H20.

Таким чином, в рамках побудованої моделі рівня знань (результат Y) на 98,21% залежить від тривалістю часу навчання (фактору X). Інші 1,79% варіації *ознаки-результату* обумовлені іншими, не врахованими в моделі факторами.

Обчислення *індексу кореляції*:

$$R_I = \sqrt{R_I^2} \approx \sqrt{0,9821} \approx 0,991,$$

- до комірки H29 занести формулу =SQRT(H26).

Таким чином, згідно з шкалою Чеддока (див. табл. А.9), існує практично функціональна залежність ознаки-результату Y від фактора X.

4) Перевірка *статистичної значимості* побудованої моделі.

На рівні значимості $\alpha = 0,05$ слід перевірити нульову гіпотезу $H_0: R_\Gamma^2 = 0$ – у тому, що генеральний індекс детермінації дорівнює нулю, проти конкуруючої гіпотези: $H_0: R_\Gamma^2 > 0$.

Статистичний критерій

$$F = \frac{R_{I(e)}^2 (n-2)}{1 - R_{I(e)}^2},$$

де $R_{I(e)}^2$ значення вибіркового індексу детермінації. У різних вибірках воно буде різним, а тому F є величина випадкова.

Для рівня значущості $\alpha = 0,05$ та кількості ступенів свободи $k_1 = 1, k_2 = 10 - 2$ за відповідною таблицею слід визначити критичне значення критерію [23]:

$$F_{\text{кр}} = F_{\text{кр}}(\alpha;\ k_1;\ k_2) = F_{\text{кр}}(0,05;\ 1;\ 8) \approx 5,3177$$

Обчислення *значення критерію*:

$$F_{\text{спос}} = \frac{R_I^2 (n-2)}{1 - R_I^2} \approx \frac{0,9821 \cdot 8}{1 - 0,9821} \approx 438,93$$

- до комірки K26 занести формулу =H26*(A15-2)/(1-H26).

Значення потрапило у критичну область $F_{\text{спос}} > F_{\text{кр}}$ (див. рис. А.23):

Рис А.23. Критична область

Тому на рівні значимості $\alpha = 0{,}05$ гіпотезу $H_0\!: R_\Gamma^2 = 0$ слід відкинути на користь гіпотези $H_0\!: R_\Gamma^2 > 0$.

Отже, отриманий результат $R_I^2 = 0{,}9821$ *статистично значущий*, отже, *статистично значуще* і вибіркове рівняння $y = 4{,}393e^{0{,}382x}$ експоненційної регресії.

5) Обчислення середньої помилки апроксимації:

$$\bar{A} = \frac{1}{n}\sum\left|\frac{y_i - y(x_i)}{y_i}\right| \cdot 100\% \approx \frac{1}{10} \cdot 0{,}8157 \cdot 100\% \approx 8{,}16\%,$$

- до комірки H33 занести формулу =(1/A15)*I14*100,

таким чином, регресійні значення $y(x_i)$ відрізняються від відповідних емпіричних значень y_i *у середньому* на 8,16%, що можна визнати добрим результатом.

6) Прогноз кількість балів за 12 та 24 тижні підготовки:

$$y(12) = 4{,}393e^{0{,}382\cdot 12} \approx 430 \text{ балів},$$

- до комірки P18 занести формулу =$D$32*EXP($D$23*12),

$$y(24) = 4{,}393e^{0{,}382\cdot 24} \approx 42165 \text{ балів}$$

- до комірки P20 занести формулу =$D$32*EXP($D$23*24).

Навчальне видання

Пономарьова Наталія Олександрівна

Голиш Ігор Євстафійович

ОСНОВИ ОПРАЦЮВАННЯ СТАТИСТИЧНИХ ДАНИХ

Навчальний посібник

Відповідальний за випуск: Олефіренко Н.В.

Комп’ютерна верстка: Пономарьова Н.О.

Коректор: Пономарьова Н.О.

Відповідальність за дотримання вимог академічної доброчесності несуть автори

Електронне видання