

МІНІСТЕРСТВО ОХОРОНИ ЗДОРОВ'Я УКРАЇНИ
ЗАПОРІЗЬКИЙ ДЕРЖАВНИЙ МЕДИКО-ФАРМАЦЕВТИЧНИЙ
УНІВЕРСИТЕТ
КАФЕДРА МЕДИЧНОЇ ТА ФАРМАЦЕВТИЧНОЇ ІНФОРМАТИКИ
І НОВІТНІХ ТЕХНОЛОГІЙ

Ю. М. Пишнограєв, О. А. Рижов, Н. І. Строїтелєва

ОСНОВИ БІОСТАТИСТИКИ
ТА МЕТОДИ СТАТИСТИЧНОГО
МОДЕЛЮВАННЯ

НАВЧАЛЬНИЙ ПОСІБНИК

для здобувачів третього освітньо-наукового ступеня доктора філософії

Запоріжжя

2026

*Затверджено на засіданні Вченої ради ЗДМФУ
та рекомендовано для використання в освітньому процесі
(протокол № від « » червня 2026 р.)*

Автори:

Ю. М. Пишнограєв – канд. фіз.-мат. наук, доцент кафедри медичної та фармацевтичної інформатики і НТ ЗДМФУ.

О. А. Рижов – д-р фарм. наук, проф., завідувач кафедри медичної та фармацевтичної інформатики і НТ ЗДМФУ;

Н. І. Строїтелева – канд. фіз.-мат. наук, доцент кафедри медичної та фармацевтичної інформатики і НТ ЗДМФУ;

Рецензенти:

Ю. О. Прокопчук д-р техн. наук, доцент, провідний науковий співробітник Інституту технічної механіки НАН;

А.В. Пожуєв - канд. фіз.-мат. наук, професор кафедри фундаментальної та прикладної математики ЗНУ.

Пишнограєв Ю.М.

ПЗ6 Основи біостатистики та методи статистичного моделювання: навч. посібник для здобувачів третього освітньо-наукового ступеня доктора філософії / Ю. М. Пишнограєв, О. А. Рижов, Н. І. Строїтелева. – Запоріжжя: ЗДМФУ, 2026. – 223 с.

У навчальному посібнику наведено теоретичні відомості та практичні завдання для формування практичних навичок використання сучасних методів статистичного моделювання, з якими майбутні науковці стикаються під час підготовки дисертаційних робіт.

ЗМІСТ

ПЕРЕДМОВА.....	5
СТАТИСТИКА ЯК СИСТЕМНИЙ МЕТОД ДОКАЗОВОЇ МЕДИЦИНИ.....	7
КЛАСИФІКАЦІЯ ОБ’ЄКТІВ МЕТОДАМИ КЛАСТЕРНОГО ТА ДИСКРИМІНАНТНОГО АНАЛІЗУ В МЕДИЦИНІ ТА ФАРМАЦІЇ.....	21
ПРАКТИЧНА РОБОТА 1	
Робота з даними в програмі Microsoft Excel.....	29
ПРАКТИЧНА РОБОТА 2	
Робота з даними в програмі Microsoft Excel.....	42
ПРАКТИЧНА РОБОТА 3	
Випадкова величина. Основні характеристики випадкової величини.	55
ПРАКТИЧНА РОБОТА 4	
Функція розподілу імовірностей випадкової величини. Деякі закони розподілу дискретної випадкової величини.....	67
ПРАКТИЧНА РОБОТА 5	
Щільність розподілу ймовірностей випадкової величини Нормальний закон розподілу випадкової величини.	80
ПРАКТИЧНА РОБОТА 6	
Перевірка статистичних гіпотез. Тест Шапіро–Уїлка. Критерій Пірсона.....	96
ПРАКТИЧНА РОБОТА 7	
Критерії Фішера і Стюдента.....	114
ПРАКТИЧНА РОБОТА 8	
Непараметричні методи перевірки гіпотез.....	131
ПРАКТИЧНА РОБОТА 9	
Дисперсійний аналіз. Критерій Краскела–Уолліса.....	145
ПРАКТИЧНА РОБОТА 10	
Довірчі інтервали.....	161
ПРАКТИЧНА РОБОТА 12	
Дискримінантний аналіз.....	201
РЕКОМЕНДОВАНА ЛІТЕРАТУРА.....	213
Додаток А.....	215
Додаток Б.....	216

Додаток В.....	218
Додаток Г	219
Додаток Д	220
Додаток Е.....	221
Додаток Є.....	222
Додаток Ж.....	223

ПЕРЕДМОВА

Сучасний етап розвитку медичної та фармацевтичної науки характеризується стрімким накопиченням емпіричних даних, що потребує від дослідника не лише глибоких професійних знань, а й високого рівня володіння методами статистичного аналізу. Біостатистика сьогодні є фундаментом доказової медицини, оскільки саме вона забезпечує об'єктивність висновків, дозволяє оцінити ефективність нових методів лікування та прогнозувати перебіг захворювань.

Даний навчальний посібник розроблений спеціально для здобувачів третього освітньо-наукового ступеня доктора філософії з медичних та фармацевтичних спеціальностей. Його метою є формування теоретичних знань і практичних навичок застосування сучасних методів статистичного аналізу та моделювання у медичних і фармацевтичних дослідженнях. Особливу увагу приділено використанню доступних інструментів, зокрема програмного середовища Microsoft Excel, що дозволяє виконувати статистичні розрахунки, візуалізацію даних і первинний аналіз без залучення спеціалізованого дороговартісного програмного забезпечення.

Структура посібника побудована таким чином, щоб забезпечити поступовий перехід від базових понять до складніших методів аналізу. Теоретичні розділи розкривають роль статистики як системного методу доказової медицини, принципи проектування клінічних досліджень, особливості інтерпретації результатів та оцінки їх достовірності, сучасні підходи до класифікації та моделювання даних. Візуальний контент теоретичних розділів виконаний із використанням сервісу штучного інтелекту Gemini.

Практична складова посібника представлена низкою практичних робіт, які охоплюють повний цикл обробки медичних даних: від їх структурування та візуалізації до перевірки статистичних гіпотез і побудови прогнозних моделей. Виконання цих завдань сприяє формуванню не лише технічних навичок, але й критичного мислення, необхідного для інтерпретації результатів досліджень.

Значна частина посібника присвячена перевірці статистичних гіпотез. Для полегшення навігації у різноманітні підходів у практикумі наведено алгоритм

вибору методу аналізу залежно від поставленого наукового завдання. Зокрема, якщо метою є аналіз відмінностей між групами, досліднику пропонується обрати відповідний статистичний критерій. Для виявлення та оцінки взаємозв'язків між параметрами застосовується кореляція, тоді як завдання прогнозування вирішуються за допомогою регресійних моделей. У випадках, коли необхідно класифікувати групи за сукупністю ознак, аспіранти можуть скористатися методами дискримінантного аналізу.

Ключовим елементом посібника є деталізована схема вибору конкретного статистичного критерію (Додаток А), яка базується на типі даних та характері їх розподілу:

- для кількісних даних із нормальним розподілом пропонується використання параметричних методів: t-критерію Стюдента (для порівняння двох незалежних або залежних груп) та дисперсійного аналізу (ANOVA), якщо кількість груп становить три і більше;
- у випадках ненормального розподілу або роботи з якісними (ранговими) даними, висвітлено алгоритми застосування непараметричних критеріїв, таких як U-критерій Манна-Уїтні, критерій Вілкоксона, а також методи Краскела-Уолліса та Фрідмана для порівняння декількох сукупностей;
- для номінальних та бінарних даних окрема увага приділена критерію Пірсона, що дозволяє коректно аналізувати частотні показники та якісні зміни у стані пацієнтів.

Методичне наповнення посібника базується на використанні пакету «Аналіз даних» (Data Analysis) та вбудованих статистичних функцій Excel, що робить складні розрахунки доступними без потреби у глибокому програмуванні.

Застосування методів, викладених у цьому виданні, допоможе аспірантам забезпечити високу достовірність наукових висновків, що є обов'язковою умовою для публікації статей у провідних міжнародних виданнях та успішного захисту дисертації.

СТАТИСТИКА ЯК СИСТЕМНИЙ МЕТОД ДОКАЗОВОЇ МЕДИЦИНИ

1. Системна методологія клінічного дослідження

У сучасній доказовій медицині статистика не може розглядатися лише як техніка обробки отриманих числових результатів. Її методологічне значення полягає у тому, що вона забезпечує формалізований перехід від клінічного спостереження до наукового доказу, від індивідуального досвіду до узагальненого знання, від невизначеності до кількісно оціненого ризику, ефекту та прогнозу. Однак для повноцінного планування медичного дослідження й інтерпретації його результатів статистичного апарату недостатньо, якщо він не включений у більш широку системну методологію. Такою методологічною рамкою виступає загальна теорія систем, яка дозволяє розглядати медичне дослідження як відкриту, цілеспрямовану, багаторівневу й ієрархічно організовану систему.

Поєднання Загальної теорії систем (ЗТС) із принципами доказової медицини (EBM- Evidence-Based Medicine) формує цілісний підхід до наукового дослідження в медицині. У межах цього підходу клінічна проблема розглядається як система, що має межі, структуру, елементи, зв'язки, входи, виходи, зворотні зв'язки, критерії ефективності та механізми регуляції. Дослідницький дизайн постає не як механічний вибір між когортним дослідженням, рандомізованим випробуванням або мета-аналізом, а як побудова моделі системи, у якій кожний етап — формулювання питання, добір вибірки, визначення змінних, контроль змішувальних факторів, статистичний аналіз і трансляція результатів у рекомендацію — має власну функцію в досягненні науково обґрунтованого висновку.

Доказова медицина сформувалася як методологія прийняття клінічних рішень, що поєднує найкращі доступні наукові докази, професійну експертизу лікаря та цінності пацієнта. У традиційному трактуванні ця тріада описує джерела інформації для клінічного рішення. Проте з позицій ЗТС її доцільно розглядати як взаємодію трьох підсистем: науково-довказової, експертно-клінічної та пацієнт-орієнтованої. Жодна з них не є самодостатньою. Докази без клінічного контексту

можуть бути непридатними для конкретного пацієнта; експертний досвід без критичної оцінки доказів може бути упередженим; цінності пацієнта без інформації про ризики й користь не забезпечують справді інформованого вибору.

Системний підхід дозволяє подолати спрощене уявлення про доказову медицину. ЕВМ є не набором статистичних критеріїв, а циклом керування клінічним знанням. Цей цикл включає постановку питання, пошук доказів, критичну оцінку, кількісну інтерпретацію, застосування до пацієнта та аналіз результату. У системній термінології це є контуром зворотного зв'язку: клінічна проблема створює запит на доказ; доказ змінює рішення; рішення змінює результат; результат повертається в систему як нова інформація для вдосконалення практики.

У медичній інформатиці та біостатистиці такий підхід має особливе значення. Дані пацієнтів, лабораторні показники, електронні медичні записи, результати рандомізованих випробувань, реєстри та систематичні огляди є інформаційними потоками відкритої системи охорони здоров'я. Статистика забезпечує формальні методи обробки цих потоків, а Загальна теорія систем — концептуальну рамку для розуміння, які саме елементи системи ми вимірюємо, які зв'язки аналізуємо, які межі дослідження встановлюємо і на який рівень узагальнення претендуємо.

Отже, доказова медицина у системному розумінні — це не лише медицина, що спирається на докази, а й медицина, що організовує докази у цілісну систему прийняття рішень. Статистика в цій системі виконує роль мови кількісного опису, інструменту оцінювання невизначеності та механізму контролю достовірності висновків.

Загальна теорія систем виходить з припущення, що складні об'єкти доцільно вивчати не лише через аналіз їх окремих частин, а через дослідження структури, функцій, зв'язків і поведінки цілого. Система — це сукупність взаємопов'язаних елементів, які утворюють певну цілісність і взаємодіють із середовищем. У медицині системами можуть бути організм, функціональна система регуляції артеріального тиску, клінічний маршрут пацієнта, лікувальна технологія, наукове дослідження, база медичних даних або система охорони здоров'я.



Будь-яке медичне дослідження також можна представити як систему. Його елементами є дослідницьке питання, гіпотеза, популяція, вибірка, змінні, методи вимірювання, втручання, порівняння, кінцеві точки, статистичні процедури, критерії інтерпретації та способи поширення результатів. Зв'язки між цими елементами визначають якість дослідження. Наприклад, неправильно сформульоване питання призводить до невідповідного дизайну; невідповідний дизайн не дозволяє коректно оцінити причинність; некоректно визначена кінцева точка знижує клінічну значущість результату; недооцінка змішувальних факторів викривлює статистичний висновок.

Системна методологія вимагає спочатку визначити межі досліджуваної системи. У контексті ЕВМ це відповідає визначенню популяції, втручання, порівняння, результату і часових меж у форматі PICO. Якщо межі не задані, дослідження втрачає фокус: результати можуть бути статистично оброблені, але їх неможливо змістовно інтерпретувати. Наприклад, твердження «препарат ефективний» не є системним описом. Системний опис потребує уточнення: для якої популяції, за яких умов, порівняно з якою альтернативою, за яким результатом, протягом якого часу та з яким допустимим рівнем ризику.

Другим принципом є ієрархічність. Медична система має рівні: молекулярний, клітинний, органний, організмний, поведінковий, соціальний, організаційний. Дослідження, яке оцінює біомаркер, належить до одного рівня; дослідження виживаності — до іншого; аналіз вартості лікування — до ще іншого. Помилка виникає тоді, коли висновки одного рівня безпосередньо переносять на інший. Зниження лабораторного маркера не завжди означає зниження смертності; покращення сурогатної кінцевої точки не гарантує покращення клінічно значущого результату. Саме системний підхід змушує перевіряти відповідність рівня вимірювання рівню клінічного рішення.

Третім принципом є емерджентність. Властивості цілого не завжди зводяться до властивостей його частин. У медицині це особливо очевидно: ефективність лікувальної програми залежить не лише від фармакологічної дії препарату, а й від прихильності пацієнта, комунікації з лікарем, доступності медичної допомоги, супутніх захворювань і соціального середовища. Тому статистична модель має враховувати не лише окремі фактори, а й можливі взаємодії, модифікацію ефекту та контекст.

2. Системне проектування плану медичного дослідження

Медичне дослідження є відкритою системою, оскільки воно взаємодіє із зовнішнім середовищем: клінічною практикою, етичними нормами, регуляторними вимогами, доступністю ресурсів, характеристиками популяції та інформаційною інфраструктурою. Відкритість означає, що дослідник не може повністю ізолювати об'єкт від контексту. Навіть у рандомізованому контрольованому випробуванні існують умови включення, критерії виключення, особливості центру, поведінка учасників, втрати спостереження і зовнішні події, які впливають на результат.



Цілеспрямованість дослідження проявляється у наявності мети та критеріїв її досягнення. У системній термінології мета визначає бажаний стан системи, а критерії — спосіб оцінки наближення до цього стану. У доказовій медицині такою метою може бути оцінка ефективності втручання, встановлення діагностичної точності тесту, визначення факторів ризику, прогнозування перебігу хвороби або синтез доказів для клінічної рекомендації. Критеріями можуть бути абсолютне зниження ризику, відносний ризик, відношення шансів, середня різниця та ін.

Планування дослідження можна представити як системне проектування. На вході системи знаходяться клінічна проблема, попередні знання, гіпотеза, ресурси, доступна популяція та методи вимірювання. Усередині системи відбуваються добір учасників, розподіл або спостереження, збір даних, контроль якості, статистичний аналіз і критична оцінка. На виході — оцінка ефекту, рівень невизначеності, висновок щодо валідності та потенційна рекомендація. Зворотний зв'язок виникає тоді, коли результати дослідження змінюють практику або породжують нові наукові питання.

Системна модель дослідження особливо корисна для запобігання методологічним помилкам. Якщо дослідник бачить дослідження як цілісну систему, він розуміє, що статистичний аналіз не може компенсувати помилки

формування вибірки, нечітке визначення кінцевої точки або систематичне зміщення вимірювання. Статистика є потужною лише тоді, коли вона застосована до правильно спроектованої системи даних.

Формат PICO у доказовій медицині можна трактувати як мінімальну системну модель клінічного питання. Компонент Р визначає межі популяційної підсистеми; І — активний вплив або втручання; С — альтернативний стан системи, з яким порівнюється вплив; О — вихідний результат, за яким оцінюється поведінка системи. У деяких випадках додають Т — час спостереження, що перетворює PICO на PICOT і робить модель динамічною.



Системне значення PICO полягає в тому, що воно не дозволяє змішувати різні рівні аналізу. Якщо популяція визначена нечітко, виникає проблема зовнішньої валідності; якщо втручання описане неповно, неможливо відтворити дослідження; якщо порівняння неадекватне, ефект не має клінічної інтерпретації; якщо результат не є клінічно значущим, статистичний висновок не перетворюється на практичну рекомендацію.

Наприклад, питання «чи ефективна антигіпертензивна терапія?» не є достатнім для системного дослідження. Коректна модель має вигляд: у дорослих

пацієнтів із артеріальною гіпертензією та високим серцево-судинним ризиком чи знижує комбінована терапія порівняно з монотерапією частоту інсульту, інфаркту міокарда або серцево-судинної смерті протягом трьох років? У такому формулюванні вже визначено популяцію, вплив, альтернативу, результат і часовий горизонт.

PICO також є основою інформаційного пошуку та статистичного планування. Воно визначає ключові слова для пошуку літератури, критерії включення досліджень у систематичний огляд, тип дизайну, первинну кінцеву точку, очікувану величину ефекту, необхідний обсяг вибірки та план аналізу. У системному сенсі PICO виконує функцію специфікації системи перед її моделюванням.



Дизайн дослідження визначає архітектуру доказу. Він задає, як елементи дослідницької системи пов'язані між собою, яким чином формується порівняння, як контролюється невизначеність, як мінімізуються систематичні помилки і який тип висновку можна зробити. З позицій Загальної теорії систем дизайн є не просто класифікаційною назвою, а способом організації інформаційних потоків у дослідженні.

Опис випадку та серія випадків мають відкриту, слабо контрольовану структуру. Вони корисні для виявлення нових явищ, формування гіпотез і опису рідкісних подій, але не забезпечують контроль альтернативних пояснень. У системній термінології це спостереження за окремими станами системи без керованого порівняння.

Дослідження «випадок–контроль» будуються ретроспективно: система формується від виходу до можливих входів. Вони ефективні для рідкісних захворювань, але чутливі до помилки добору, помилки пам'яті та неправильного визначення експозиції. Когортні дослідження, навпаки, рухаються від експозиції до результату і краще відображають часову послідовність, однак залишаються вразливими до змішування факторів. Рандомізоване контрольоване випробування є найбільш керованою системною архітектурою для оцінки причинного ефекту втручання. Рандомізація виконує роль механізму вирівнювання відомих і невідомих факторів між підсистемами порівняння.

Систематичний огляд і мета-аналіз становлять вищий рівень системної інтеграції доказів. Вони розглядають окремі дослідження як елементи метасистеми. При цьому якість метасистемного висновку залежить не лише від кількості досліджень, а й від їх гетерогенності, валідності, ризику публікаційного зміщення та сумісності популяцій, втручань і результатів.

Описова статистика є не лише попереднім етапом аналізу, а системною діагностикою якості даних. Вона дозволяє оцінити, чи відповідає фактична інформаційна структура дослідження запланованій моделі. На цьому етапі дослідник перевіряє розподіли, пропуски, викиди, баланс груп, одиниці вимірювання, логічну узгодженість змінних і можливі помилки введення.

Середнє, медіана, стандартне відхилення, міжквартильний інтервал і коефіцієнт варіації описують різні аспекти системної поведінки показника. Середнє характеризує центр симетричного розподілу; медіана є стійкою характеристикою для асиметричних даних; стандартне відхилення відображає масштаб варіабельності; міжквартильний інтервал показує розкид центральної частини

системи спостережень. Правильний вибір описових характеристик є передумовою коректної інтерпретації.



Графічний аналіз має особливе значення для системного мислення. Гістограма показує форму розподілу; boxplot виявляє викиди та асиметрію; діаграма розсіювання демонструє зв'язок між змінними; лінійний графік відображає динаміку системи в часі. Часто саме графік дозволяє побачити кластери, нелінійність, взаємодії або підгрупи, які неочевидні у таблиці середніх значень.

У доказовій медицині описова статистика також використовується для оцінки початкової порівнюваності груп. Якщо групи значно відрізняються за віком, тяжкістю хвороби або супутньою патологією, це може вплинути на результат. У рандомізованому дослідженні такі відмінності можуть бути випадковими, але їх усе одно потрібно враховувати змістовно. В обсерваційних дослідженнях вони часто є ознакою потенційного змішування факторів.

Довірчий інтервал є ключовим інструментом системного тлумачення невизначеності. Він показує не лише точкову оцінку ефекту, а й діапазон значень, сумісних із даними. Для клінічного рішення це принципово важливо, оскільки лікар

має оцінити не тільки факт статистичної значущості, а й межі можливої користі або шкоди.

Системний підхід особливо підкреслює значення абсолютних показників. Відносне зниження ризику може бути однаковим при великому і малому базовому ризику, але абсолютна користь буде різною. Тому для прийняття рішення необхідно переводити результати досліджень у клінічно зрозумілі показники: скільком пацієнтам допоможе втручання, скільки може зазнати шкоди, який баланс користі й ризику у конкретній системі пацієнта.

3. Системна інтерпретація та аналіз результатів медичного дослідження

Аналіз результатів досліджень не завершується статистичним розрахунком. Він є процесом системної інтерпретації, у якому числовий результат співвідноситься з дизайном, якістю даних, біологічною правдоподібністю, попередніми доказами та клінічним контекстом. Статистичний критерій відповідає на вузьке питання; системний аналіз відповідає на питання, що означає отриманий результат для науки, практики і пацієнта.

Першим кроком є перевірка якості даних: чи немає помилок, пропусків, викидів, порушень протоколу, дисбалансу груп. Другим — перевірка припущень статистичних методів. Третім — оцінка основного результату з величиною ефекту і довірчим інтервалом. Четвертим — аналіз чутливості, який перевіряє, чи зберігається висновок за альтернативних припущень або методів. П'ятим — оцінка застосовності результату до цільової клінічної системи.

Клінічна значущість має оцінюватися через абсолютні ризики, тяжкість результатів, небажані явища, вартість і цінності пацієнта. Наприклад, навіть статистично переконливе зниження лабораторного маркера може не мати значення, якщо воно не призводить до зменшення симптомів, госпіталізацій або смертності. Навпаки, невелике абсолютне зниження смертності може бути клінічно дуже важливим.

Системна інтерпретація також передбачає розгляд меж узагальнення. Результат, отриманий у вузькій групі пацієнтів, не можна автоматично переносити

на інші популяції. Потрібно враховувати вік, стать, коморбідність, тяжкість стану, доступність лікування, організаційні умови та культурний контекст. Саме тому ЕВМ не є механічним застосуванням протоколів, а потребує клінічного судження.

Сучасні медичні дослідження дедалі більше залежать від інформаційних технологій: електронних медичних записів, реєстрів, лабораторних інформаційних систем, статистичних пакетів, електронних таблиць, систем управління даними та програмного коду. З позицій системного підходу ці технології є інфраструктурою доказової системи. Їх якість безпосередньо впливає на якість наукового висновку.

Відтворюваність означає, що інший дослідник може зрозуміти і, за наявності даних, повторити аналітичний шлях від сирих даних до результату. Для цього потрібні попередньо зареєстрований протокол, план статистичного аналізу, документовані правила очищення даних, збереження вихідних даних, версіонування коду, прозоре повідомлення всіх запланованих результатів і чітке розмежування підтверджувального та пошукового аналізу.

Excel може бути корисним навчальним інструментом для базових розрахунків і первинної організації даних, але складні дослідження потребують контрольованого аналітичного середовища. Помилки формул, ручне копіювання, приховані рядки, некоректні діапазони або змішування сирих і оброблених даних можуть стати джерелом систематичних похибок. Тому будь-який інструмент має використовуватися з документуванням і контролем якості.

Статистика в доказовій медицині має етичний вимір, оскільки її результати впливають на лікування, профілактику, розподіл ресурсів і життя пацієнтів. Некоректно сплановане дослідження може не дати відповіді на важливе питання, а отже марно використати участь пацієнтів. Некоректний аналіз може створити ілюзію ефективності або безпеки. Неповне повідомлення небажаних явищ може завдати шкоди у майбутній практиці.

Системний підхід до етики означає, що відповідальність виникає на всіх етапах дослідницької системи: при формулюванні питання, включенні пацієнтів, виборі кінцевих точок, зборі даних, статистичному аналізі, інтерпретації, публікації та впровадженні. Етичність не обмежується отриманням інформованої згоди. Вона

включає методологічну доброчесність, прозорість, повагу до невизначеності та чесне повідомлення обмежень.

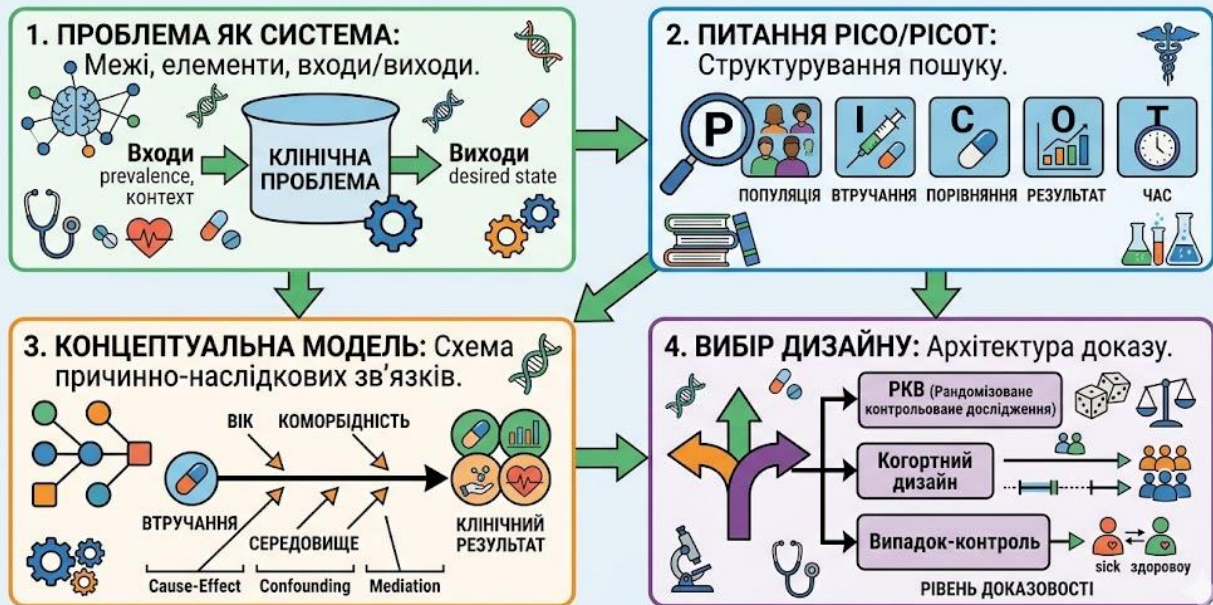
Для лікаря системно-доказове мислення означає здатність пояснити пацієнту не лише те, що втручання «ефективне», а наскільки воно ефективне, для кого саме, з яким ризиком, якою є альтернативна стратегія, наскільки впевненими є докази і як це співвідноситься з цінностями пацієнта. Такий підхід перетворює статистику на інструмент спільного прийняття рішення.

Етичним є також визнання меж знання. Якщо довірчий інтервал широкий, якщо дослідження має високий ризик bias, якщо популяція не відповідає конкретному пацієнту, якщо докази непрямі або суперечливі, це має бути відкрито зазначено. Доказова медицина не вимагає абсолютної певності; вона вимагає чесного керування невизначеністю.

Інтегрований алгоритм інтеграції ЗТС та ЕВМ у медичному дослідженні можна подати як послідовність системно-доказових кроків. Перший крок — визначити клінічну проблему як систему: її межі, рівень аналізу, ключові елементи, середовище, очікувані входи й виходи. Другий — сформулювати питання у форматі PICO/PICOT. Третій — побудувати концептуальну модель причинних або прогностичних зв'язків. Четвертий — обрати дизайн, який відповідає типу питання і рівню потрібного доказу.

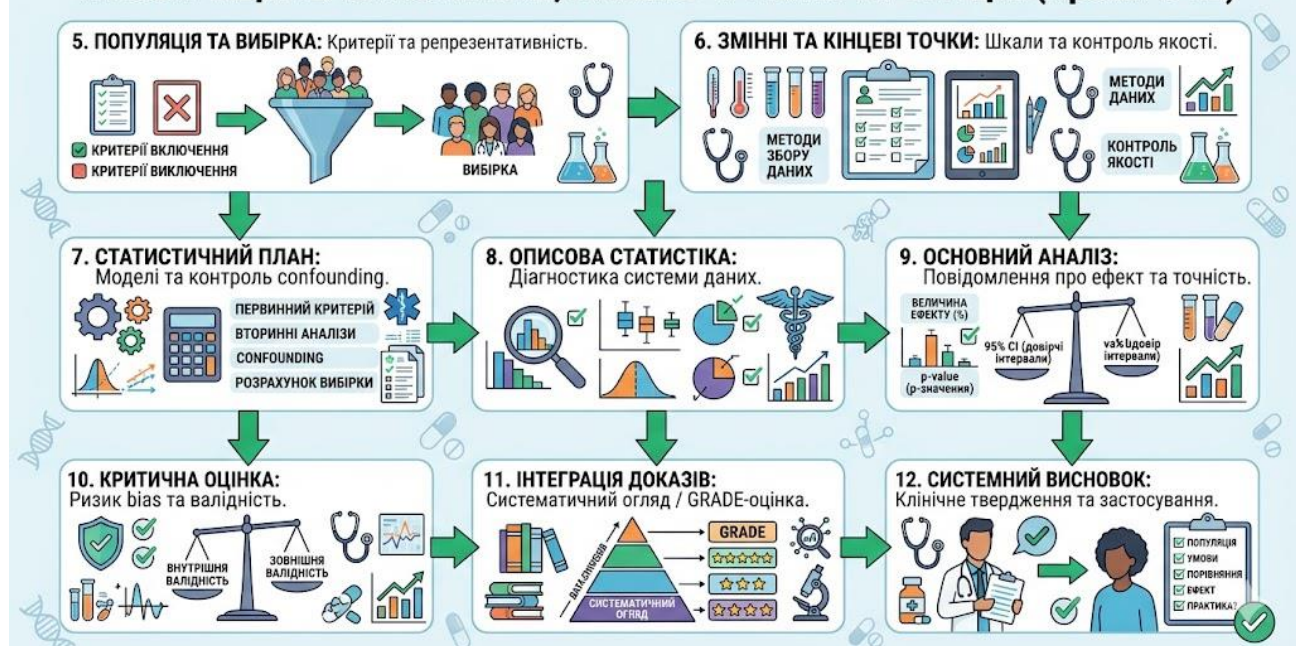
П'ятий крок — визначити популяцію, вибірку, критерії включення і виключення, методи добору та очікувану репрезентативність. Шостий — описати змінні, шкали вимірювання, кінцеві точки, методи збору даних і контроль якості. Сьомий — спланувати статистичний аналіз: первинний критерій або модель, вторинні аналізи, роботу з пропусками, контроль confounding, множинності, аналіз чутливості та розрахунок обсягу вибірки.

ІЛЮСТРАЦІЯ 1: ПЛАНУВАННЯ ТА ПОБУДОВА МОДЕЛІ (Кроки 1-4)



Восьмий крок — провести описову статистику як діагностику системи даних. Дев'ятий — виконати основний статистичний аналіз із повідомленням величини ефекту, довірчих інтервалів і р-значень. Десятий — оцінити ризик bias, внутрішню і зовнішню валідність, клінічну значущість і відповідність попереднім доказам. Одинадцятий — якщо потрібно, інтегрувати результат у систематичний огляд або GRADE-оцінку.

ІЛЮСТРАЦІЯ 2: ВИКОНАННЯ, АНАЛІЗ ТА ІНТЕРПРЕТАЦІЯ (Кроки 5-12)



Дванадцятий крок — сформулювати висновок як системне клінічне твердження: у якій популяції, за яких умов, порівняно з чим, який ефект спостерігається, наскільки він точний, наскільки клінічно значущий, які ризики та обмеження, який рівень впевненості і чи може результат бути застосований у практиці. Такий алгоритм дозволяє уникнути редукції доказової медицини до набору статистичних тестів і перетворює її на цілісну методологію наукового мислення.

Таким чином, для медичного дослідника системно-доказове мислення означає здатність бачити зв'язок між клінічною проблемою, дизайном, даними, статистичною моделлю, валідністю і практичним рішенням. Для лікаря воно означає здатність критично читати літературу і застосовувати докази до конкретного пацієнта. Для системи охорони здоров'я — здатність приймати рішення на основі прозорих, відтворюваних і клінічно значущих доказів.

Професійна діяльність сучасного науковця та лікаря вимагає переходу від інтуїтивного накопичення досвіду до суворої формалізації знань. Біостатистика, як фундамент доказової медицини, забезпечує цей перехід, дозволяючи за хаосом випадкових клінічних випадків побачити об'єктивні закономірності, що рятують життя. Опанування цих методів — від базової дескриптивної статистики до багатомірних моделей прогнозування — є обов'язковою умовою для формування достовірних наукових рішень та успішного захисту дисертаційних робіт у сучасному академічному просторі.

КЛАСИФІКАЦІЯ ОБ'ЄКТІВ МЕТОДАМИ КЛАСТЕРНОГО ТА ДИСКРИМІНАНТНОГО АНАЛІЗУ В МЕДИЦИНІ ТА ФАРМАЦІЇ

Проблема класифікації об'єктів є фундаментальною для розвитку медичної науки та фармацевтичної галузі, оскільки вона забезпечує перехід від описового накопичення фактів до систематизації знань та прогнозування явищ. У сучасних умовах цифровізації охорони здоров'я обсяг даних, що генеруються в результаті клінічних випробувань, генетичного секвенування та моніторингу якості лікарських засобів, зростає експоненціально. Це створює гостру потребу в застосуванні методів багатовимірної статистичного аналізу, серед яких центральне місце посідають кластерний та дискримінантний аналіз.

Класифікація в біомедицині може розглядатися як процес виявлення прихованих структур у даних або як віднесення нових об'єктів до вже відомих класів. У першому випадку дослідник має справу з навчанням без вчителя (unsupervised learning), де групи (кластери) не визначені заздалегідь, а формуються на основі внутрішньої подібності об'єктів. Типовим прикладом є пошук нових фенотипів захворювань або групування генів за профілями їхньої експресії. У другому випадку використовується навчання з вчителем (supervised learning), де на основі навчальної вибірки з відомими класами будується вирішальне правило для класифікації нових спостережень. Це критично важливо для діагностики, де на основі набору симптомів необхідно віднести пацієнта до групи “здоровий” або “хворий”.

Вибір між кластерним та дискримінантним аналізом залежить від наявності апріорної інформації про структуру об'єктів. Якщо діагнози пацієнтів відомі заздалегідь і потрібно знайти змінні, що найкраще їх розрізняють, застосовується дискримінантний аналіз. Якщо ж структура популяції пацієнтів невідома і метою є пошук природних підгруп, доцільно використовувати кластерний аналіз.

Кластерний аналіз — це багатовимірна статистична процедура, що збирає дані, які містять інформацію про вибірку об'єктів, і впорядковує їх у відносно гомогенні групи. Основне завдання полягає в тому, щоб об'єкти всередині одного

кластера були якомога подібнішими за своїми характеристиками, тоді як об'єкти з різних кластерів — максимально відмінними. В медицині це дозволяє ідентифікувати субпопуляції пацієнтів, які можуть по-різному реагувати на однакову терапію, що є основою персоналізованого підходу.

Однією з ключових особливостей кластерного аналізу є те, що він не є методом статистичного висновку у класичному розумінні. Це радше математичний інструмент для виявлення структури, який не вимагає суворих припущень щодо розподілу змінних (хоча нормальність може покращити результати). Результат кластеризації значною мірою детермінований вибором метрики відстані та алгоритму об'єднання, що вимагає від дослідника глибокого розуміння природи даних.

Математична репрезентація подібності об'єктів базується на понятті відстані в багатовимірному просторі ознак. Чим менша відстань між двома об'єктами, тим більше вони подібні. У медичних та фармацевтичних дослідженнях вибір метрики залежить від типу даних (кількісні, якісні) та наявності кореляцій між ними. Найбільш вживаною метрикою є Евклідова відстань, яка розраховується за формулою:

$$d(x, y) = \sqrt{\sum_{i=1}^p (x_i - y_i)^2}$$

де x_i та y_i — значення i -ї ознаки для об'єктів x та y , а p — кількість ознак. Ця метрика ефективна для некорельованих даних, але вона чутлива до масштабу вимірювання. Для нівелювання впливу одиниць вимірювання (наприклад, поєднання значень артеріального тиску в мм рт. ст. та концентрації холестерину в ммоль/л) необхідно проводити стандартизацію даних (Z-перетворення).

У випадках, коли змінні мають сильні кореляційні зв'язки, доцільно використовувати відстань Махаланобіса, яка враховує коваріаційну структуру даних:

$$D^2 = (x - \mu)^T \Sigma^{-1} (x - \mu)$$

де Σ^{-1} — зворотна коваріаційна матриця. Це дозволяє уникнути надмірної ваги дубльованої інформації, що міститься в корельованих біомаркерах.

Кластерний аналіз поділяється на дві великі категорії: ієрархічні та неієрархічні (partiційні) методи. Кожен з підходів має свої переваги та обмеження, що визначають їхнє застосування в конкретних дослідницьких сценаріях. Ієрархічні методи будують послідовність розбиття, що нагадує деревоподібну структуру — дендрограму. Агломеративний підхід (bottom-up) починається з того, що кожен об'єкт вважається окремим кластером, а потім вони послідовно об'єднуються. Дивізійний підхід (top-down) діє навпаки: вся вибірка розглядається як один кластер, який поступово розщеплюється.

Ключовим моментом в ієрархічному аналізі є метод зв'язку (linkage), який визначає відстань між кластерами:

1. Метод найближчого сусіда (Single Linkage): відстань між кластерами дорівнює мінімальній відстані між їхніми елементами. Це може призводити до утворення “ланцюгів” об'єктів.
2. Метод далекого сусіда (Complete Linkage): відстань базується на максимальному відхиленні елементів. Сприяє формуванню компактних, сферичних груп.
3. Метод Уорда (Ward's Method): мінімізує приріст внутрішньокластерної суми квадратів відхилень. Це один із найпотужніших методів для медичних даних, оскільки він створює кластери приблизно однакового розміру з високою щільністю.

Неієрархічною кластеризацією є метод k -середніх. Це ітераційний алгоритм, який вимагає попереднього визначення кількості кластерів k . Він спрямований на мінімізацію суми квадратів відстаней від об'єктів до центрів їхніх кластерів. Алгоритм особливо ефективний для великих масивів даних, де ієрархічні методи стають занадто ресурсоємними через необхідність зберігання матриці відстаней розміром $n \times n$. Процес включає чотири основні стадії: вибір початкових центрів, призначення об'єктів до найближчого центру, перерахунок центрів як середніх значень нових груп та повторення процесу до конвергенції. Основним критерієм завершення є відсутність переміщень об'єктів між групами або досягнення

мінімального порогу зміни суми квадратів помилок (SSE). Для аспірантів медичних та фармацевтичних спеціальностей оволодіння навичками ручної реалізації алгоритмів у MS Excel є критично важливим для розуміння внутрішньої логіки статистичних процедур.

Дискримінантний аналіз — це статистичний метод, що використовується для вивчення відмінностей між двома або більше групами об'єктів за сукупністю кількох ознак одночасно. На відміну від кластерного аналізу, тут групи відомі заздалегідь, і основною метою є побудова математичної моделі, яка дозволяє максимально точно передбачити належність об'єкта до групи.

У медицині дискримінантний аналіз виконує дві основні функції:

1. Описова: визначення, які саме біомаркери або симптоми роблять найбільший внесок у розрізнення хворих та здорових.
2. Прогностична: розробка алгоритму для автоматизованої діагностики нових пацієнтів.

Основою методу є лінійна дискримінантна функція (LDA) Фішера, яка має вигляд:

$$D = v_1X_1 + v_2X_2 + \dots + v_pX_p + a$$

де D — дискримінантний бал (score), v_i — дискримінантні коефіцієнти, X_i — незалежні змінні (предиктори), a — константа. Коефіцієнти підбираються так, щоб максимізувати співвідношення міжгрупової варіації до внутрішньогрупової.

Важливою умовою застосування LDA є припущення про рівність коваріаційних матриць груп (гомоскедастичність). Якщо це припущення порушується (що часто буває в медичних вибірках, де патологічна група має значно більшу варіативність), використовується квадратичний дискримінантний аналіз (QDA).

Для оцінки якості дискримінації використовують наступні критерії:

- Лямбда Вілкса (Λ): відображає частку загальної мінливості, яка не пояснюється груповими відмінностями. Значення Λ близькі до 0 свідчать про високу якість розділення груп, тоді як значення близькі до 1 — про неефективність моделі.

- F-критерій: використовується для тестування нульової гіпотези про рівність середніх значень дискримінантних функцій у групах.
- Канонічна кореляція: вказує на силу зв'язку між дискримінантною функцією та груповою належністю.

Фармацевтична галузь використовує кластерний та дискримінантний аналіз на всіх етапах життєвого циклу лікарського засобу — від розробки молекули до фармаконагляду. Дискримінантний аналіз на основі даних БІЧ-спектроскопії (NIR spectroscopy) є “золотим стандартом” для швидкої ідентифікації лікарської сировини. Модель навчається на спектрах автентичних зразків і здатна миттєво відрізнити оригінальний препарат від підробки, навіть якщо вони мають схожий зовнішній вигляд. Кожна серія виробництва повинна потрапляти в межі дискримінантного “хмари” еталона.

Методи класифікації застосовуються в експериментальній фармакології для вивчення суб'єктивних ефектів препаратів (drug discrimination procedures). Наприклад, тварини навчаються розрізняти введення певної дози препарату та фізіологічного розчину. Дискримінантний аналіз поведінкових реакцій дозволяє оцінити адиктивний потенціал нових сполук та їхню подібність до відомих наркотичних засобів.

Якщо кореляційний аналіз лише констатує факт наявності зв'язку та його силу, то регресійний аналіз дає кількісне математичне уявлення (модель) цього зв'язку. **Регресійний аналіз** — це статистичний метод дослідження залежності однієї величини (залежної змінної, Y) від однієї або кількох інших величин (незалежних змінних, предикторів, X), заданий у вигляді математичного рівняння.

Регресійний аналіз в медицині використовується з двох головних причин:

1. Пояснення (опис): опис залежності між змінними допомагає встановити наявність та характер можливого причинно-наслідкового механізму. Наприклад, як саме кожний міліграм препарату впливає на рівень ферменту.

2. Прогнозування (предикція): рівняння регресії дозволяє прогнозувати невідоме значення залежної змінної (Y) для нових пацієнтів за відомими значеннями незалежних змінних (X). Наприклад, передбачення ризику ускладнень

на основі біохімічних маркерів крові.

Основне завдання регресійного аналізу: встановити параметри (коефіцієнти), які найкращим чином описують експериментальну залежність, за умови, що емпіричні точки завжди мають певну випадкову похибку вимірювання (біологічну варіабельність). Регресія — це процес підбору невідомих коефіцієнтів аналітичної функції $f(x)$.

Класичний регресійний аналіз припускає, що фактори X вимірюються без помилок (вони детерміновані). На практиці в медицині будь-який предиктор (наприклад, рівень лейкоцитів, вага пацієнта) вимірюється з апаратною похибкою. Відмова від абсолютної детермінованості незалежних змінних зумовлює необхідність застосування компромісних методів — **конфлюентного аналізу**. Він враховує, що розкид значень факторів є нормально розподіленим та усіченим. Це дозволяє коригувати коефіцієнти регресії, уникаючи так званого “зміщення ослаблення” (attenuation bias), яке штучно занижує оцінку впливу фактора, виміряного з похибкою.

Будь-яка класифікаційна модель потребує перевірки надійності перед впровадженням у клінічну практику. Найбільш поширеним методом є побудова класифікаційної матриці (Confusion Matrix). Ось її ключові метрики ефективності:

- 1) точність: загальна частка правильно класифікованих випадків.
- 2) Чутливість: здатність моделі правильно виявляти хворих. Це критично для скринінгових тестів, де важливо не пропустити хворобу.
- 3) специфічність: здатність моделі правильно виявляти здорових людей, що мінімізує ризик зайвого лікування.
- 4) прогностична цінність позитивного результату: ймовірність того, що людина дійсно хвора при отриманні позитивного тесту.
- 5) F1-Score: гармонійне середнє між точністю та чутливістю, корисне при незбалансованих класах.
- 6) коефіцієнт кореляції Меттьюса: вважається одним із найкращих показників для оцінки бінарної класифікації.

У медичних дослідженнях також обов'язково використовується метод перехресної перевірки (cross-validation), зокрема “leave-one-out” (виключення по одному). Це дозволяє оцінити, наскільки стабільною буде модель при появі нових даних.

За допомогою кластерного аналізу виділяють Critical Quality Attributes (CQA) — критичні показники якості, які найбільше впливають на стабільність готової форми. Це дозволяє зосередити ресурси контролю якості на найбільш чутливих параметрах процесу.

При проведенні класифікації в медицині аспіранти мають враховувати кілька специфічних факторів:

- мультиколінеарність: багато медичних показників (наприклад, зріст та вага, систолічний та діастолічний тиск) сильно корелюють між собою. Це може призвести до нестабільності дискримінантних коефіцієнтів. Вирішенням є попереднє використання методу головних компонент (PCA) або використання відстані Махаланобіса.
- наявність викидів: патологічні стани часто супроводжуються екстремальними значеннями показників. Кластерний аналіз методом k -середніх дуже чутливий до таких точок, тому перед аналізом необхідно проводити скринінг даних.
- обсяг вибірки: для дискримінантного аналізу існує правило — кількість спостережень у найменшій групі повинна бути принаймні у 5-10 разів більшою за кількість незалежних змінних.
- суб'єктивність вибору k : у кластерному аналізі вибір кількості груп часто базується на евристиці. Рекомендується використовувати метод “ліктя” (elbow method), де будується графік залежності SSE від кількості кластерів і шукається точка зламу.

Методи кластерного та дискримінантного аналізу є потужними інструментами інтелектуального аналізу медичних та фармацевтичних даних. Вони дозволяють перейти від простого порівняння середніх значень до багатовимірного розуміння структури об'єктів.

Для отримання достовірних результатів аспірантам рекомендується дотримуватися наступної послідовності дій:

- 1) ретельна візуалізація та чистка даних, перевірка на наявність пропущених значень та аномалій.
- 2) обов'язкова стандартизація змінних, якщо вони мають різні одиниці вимірювання.
- 3) використання декількох методів кластеризації для перевірки стійкості отриманої структури (наприклад, порівняння результатів методу Уорда та k-середніх).
- 4) валідація дискримінантних моделей на незалежних вибірках або за допомогою крос-валідації.
- 5) інтерпретація результатів з огляду на клінічну значущість, а не лише статистичні показники.

Реалізація цих методів у середовищі MS Excel сприяє глибшому засвоєнню статистичної теорії та дозволяє проводити складні обчислення без необхідності використання високовартісного спеціалізованого програмного забезпечення на початкових етапах наукової роботи.

ПРАКТИЧНА РОБОТА 1

Робота з даними в програмі Microsoft Excel

Питання, що розглядаються:

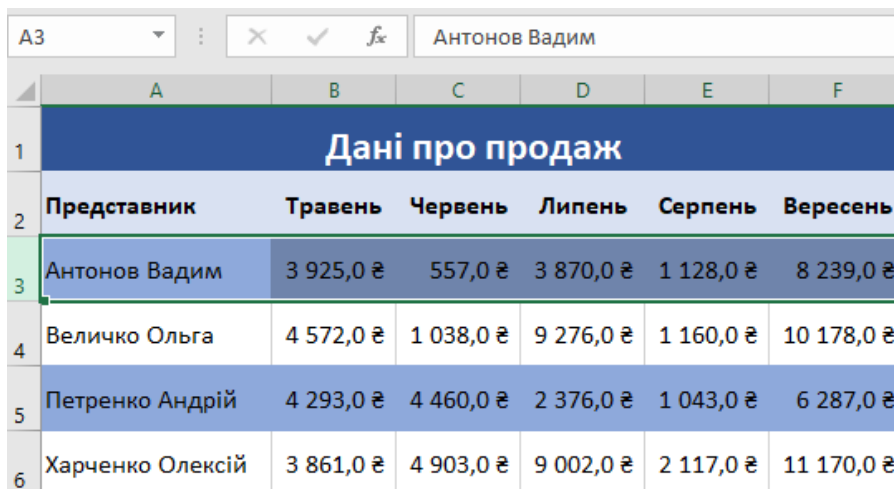
1. Закріплення областей.
2. Відкриття поточної книги в новому вікні.
3. Розділення листа на окремі області.
4. Сортювання даних.

При роботі з великими об'ємами даних можуть виникнути ускладнення при зіставленні інформації в книзі. Проте **Excel** має в своєму розпорядженні декілька інструментів, що спрощують одночасний перегляд вмісту різних розділів книги, наприклад, закріплення областей і розподіл вікон.

1. Закріплення областей

Іноді може виникати необхідність постійно бачити певні області на робочому аркуші **Excel**, зокрема заголовки. Закріплюючи рядки або стовпці, Ви зможете прокручувати вміст, при цьому закріплені клітинки залишаться в полі зору. Для цього необхідно виконати наступні кроки.

- 1.1 Виділіть рядок нижче за той, що потрібно закріпити. У нашому прикладі ми хочемо зафіксувати рядки 1 і 2, тому виділимо рядок 3.



A3						Антонов Вадим
	A	B	C	D	E	F
1	Дані про продаж					
2	Представник	Травень	Червень	Липень	Серпень	Вересень
3	Антонов Вадим	3 925,0 €	557,0 €	3 870,0 €	1 128,0 €	8 239,0 €
4	Величко Ольга	4 572,0 €	1 038,0 €	9 276,0 €	1 160,0 €	10 178,0 €
5	Петренко Андрій	4 293,0 €	4 460,0 €	2 376,0 €	1 043,0 €	6 287,0 €
6	Харченко Олексій	3 861,0 €	4 903,0 €	9 002,0 €	2 117,0 €	11 170,0 €

Рисунок 1.1

1.2 Відкрийте вкладку **Подання** на **Стрічці головного меню**.

1.3 Натисніть команду **Закріпити області** і з меню, що розкривається, виберіть однойменний пункт.

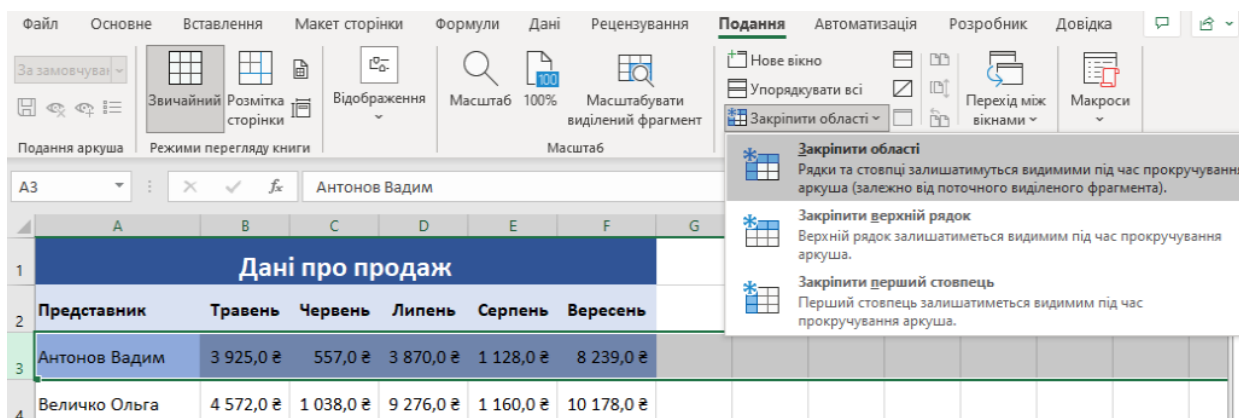


Рисунок 1.2

1.4 Рядки будуть закріплені, а область закріплення буде позначена сірою лінією. Тепер можна прокручувати робочий аркуш **Excel**, але закріплені рядки залишаються в полі зору у верхній частині аркуша. У нашому прикладі ми прокрутили аркуш до рядка 15 (рис.1.3).

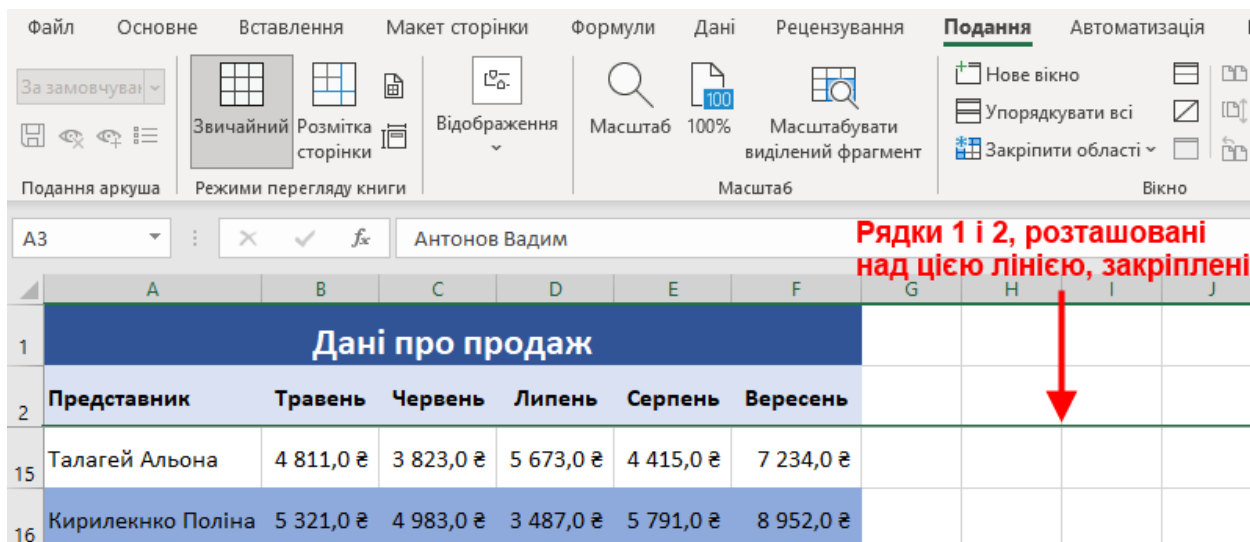


Рисунок 1.3

Зауваження. Для закріплення стовпців необхідно виконати дії аналогічні тим, що прописані вище для рядків. Щоб зняти закріплення рядків або стовпців,

натисніть **Закріпити області**, а потім з випадючого меню виберіть пункт **Звільнити області** (рис. 1.4).

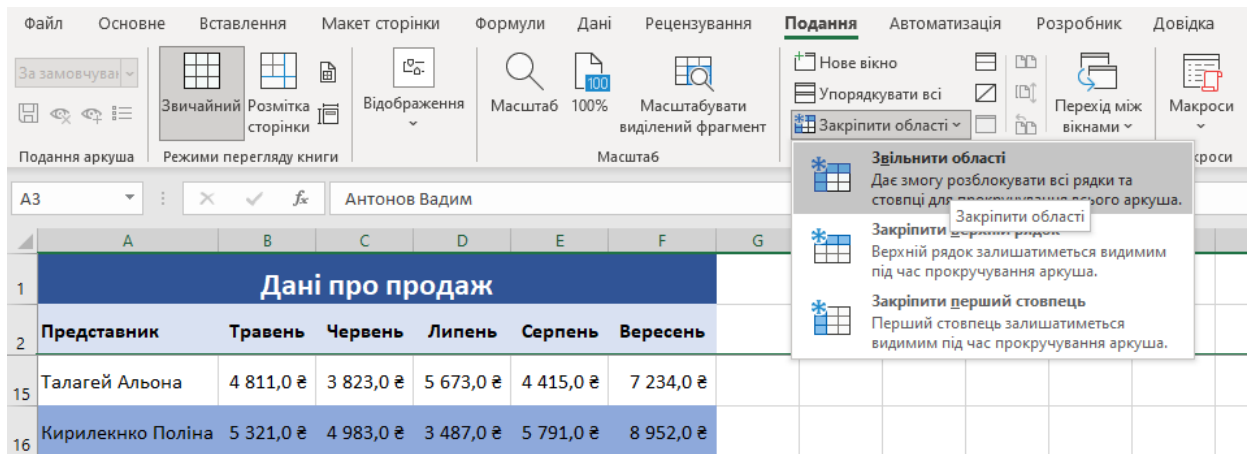


Рисунок 1.4

Зауваження. Якщо необхідно закріпити тільки верхній рядок (Рядок 1) або перший стовпець (Стовпець A), можете вибрати відповідну команду в меню, що розкривається.

2. Відкриття поточної книги в новому вікні

Excel дозволяє відкривати одну і ту ж книгу одночасно в декількох вікнах. У нашому прикладі ми скористаємося цією можливістю, щоб порівняти два різні аркуші однієї книги. Для цього:

2.1 Відкрийте вкладку **Подання** на **Стрічці головного меню**, а потім виберіть команду **Нове вікно** (рис.1.5).

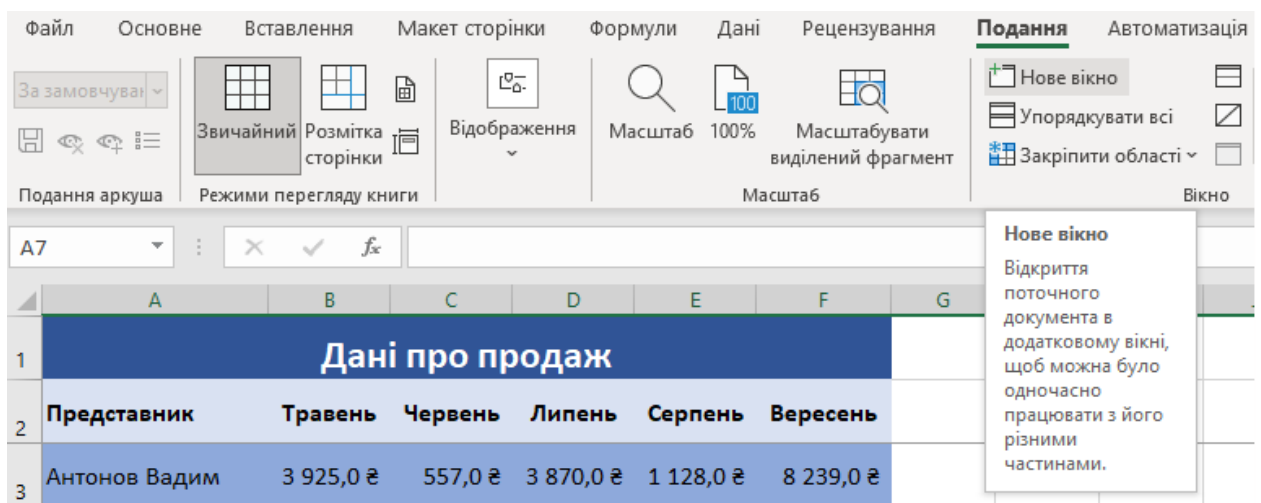


Рисунок 1.5

В результаті цих дій відкриється нове вікно для поточної книги (рис. 1.6).

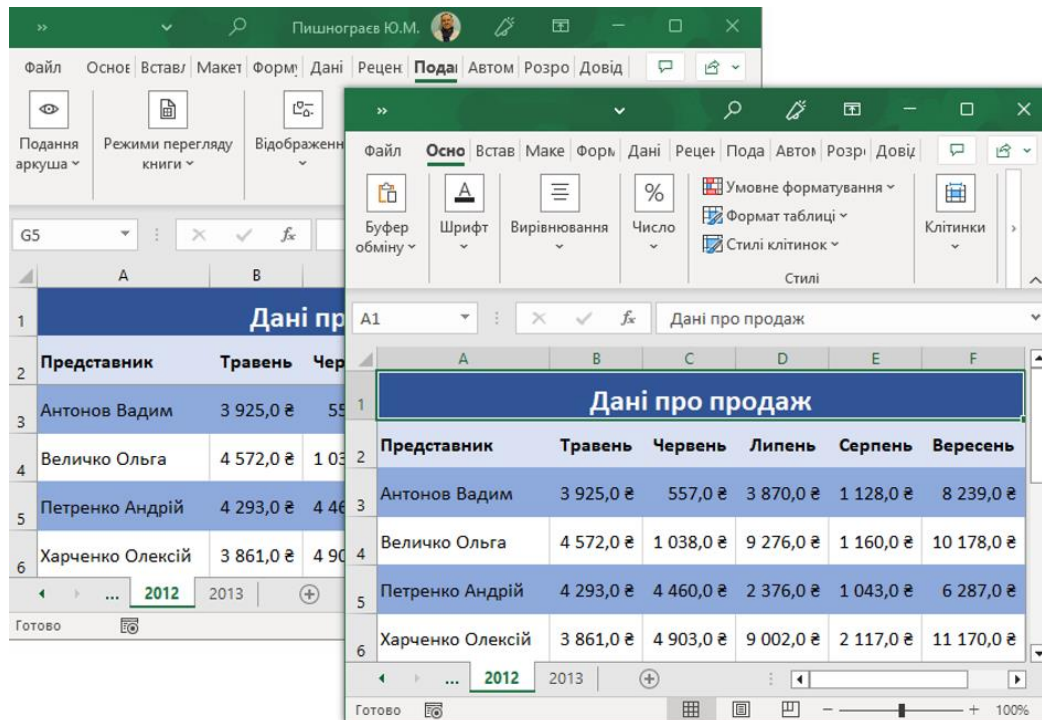


Рисунок 1.6

Тепер можна порівнювати листи однієї і тієї ж книги в різних вікнах. У нашому прикладі вибираємо звіт по продажах за 2013 рік, щоб порівняти продажі в 2012 і в 2013 роках (рис.1.7).

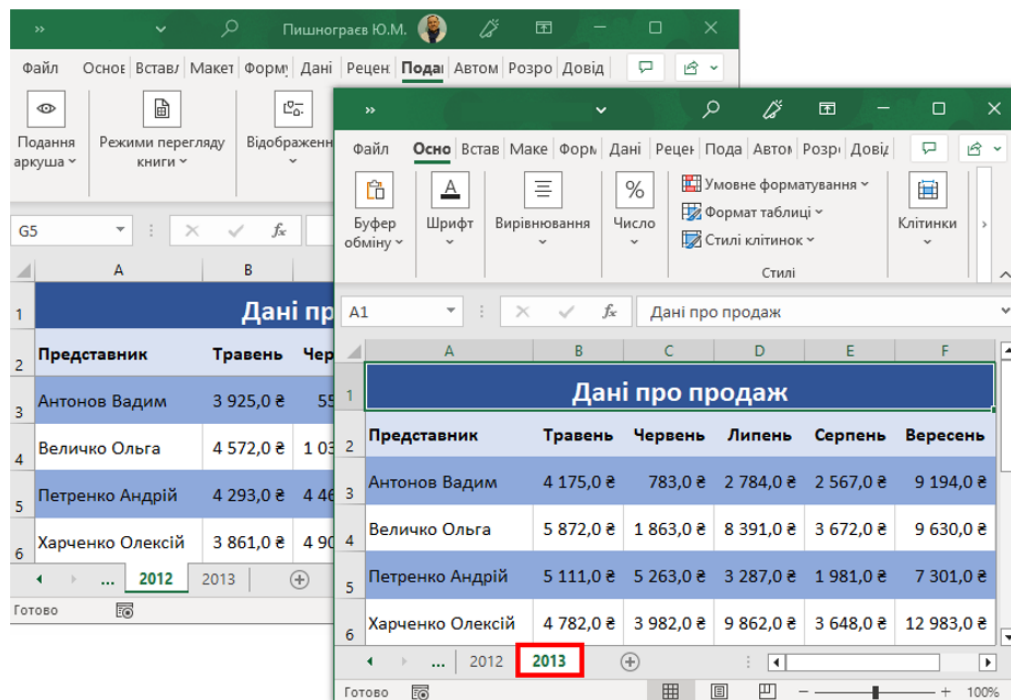


Рисунок 1.7

Зауваження. Якщо у Вас відкрито декілька вікон, можна скористатися командою **Упорядкувати всі** у вкладці **Подання** для швидкого угруповання вікон.

3. Розділення листа на окремі області

Excel дозволяє порівнювати розділи одного аркушу без створення додаткових вікон. Команда **Розділити** дозволяє розділити аркуш на окремі області, які можна прокручувати незалежно одна від одної.

3.1 Виділіть клітинку в тому місці, де необхідно розділити аркуш. Якщо виберете клітинку в першому стовпці або першому рядку, то аркуш розділиться на 2 частини, інакше на 4. У нашому прикладі вибираємо клітинку **C6** (рис.1.8).

C6						
	A	B	C	D	E	F
1	Дані про продаж					
2	Представник	Травень	Червень	Липень	Серпень	Вересень
4	Величко Ольга	5 872,0 €	1 863,0 €	8 391,0 €	3 672,0 €	9 630,0 €
5	Петренко Андрій	5 111,0 €	5 263,0 €	3 287,0 €	1 981,0 €	7 301,0 €
6	Харченко Олексій	4 782,0 €	3 982,0 €	9 862,0 €	3 648,0 €	12 983,0 €
7	Бойова Світлана	4 811,0 €	3 823,0 €	5 673,0 €	4 415,0 €	7 234,0 €
8	Лігун Єгор	5 321,0 €	4 983,0 €	3 487,0 €	5 791,0 €	8 952,0 €

Рисунок 1.8

3.2 Відкрийте вкладку **Подання** на **Стрічці головного меню**, а потім натисніть команду **Розділити** (рис.1.9). В результаті аркуш буде розділений на декілька областей і можна прокручувати кожен область окремо, використовуючи смуги прокрутки (рис.1.10). Це дозволяє порівнювати різні розділи одного і того ж аркуша.

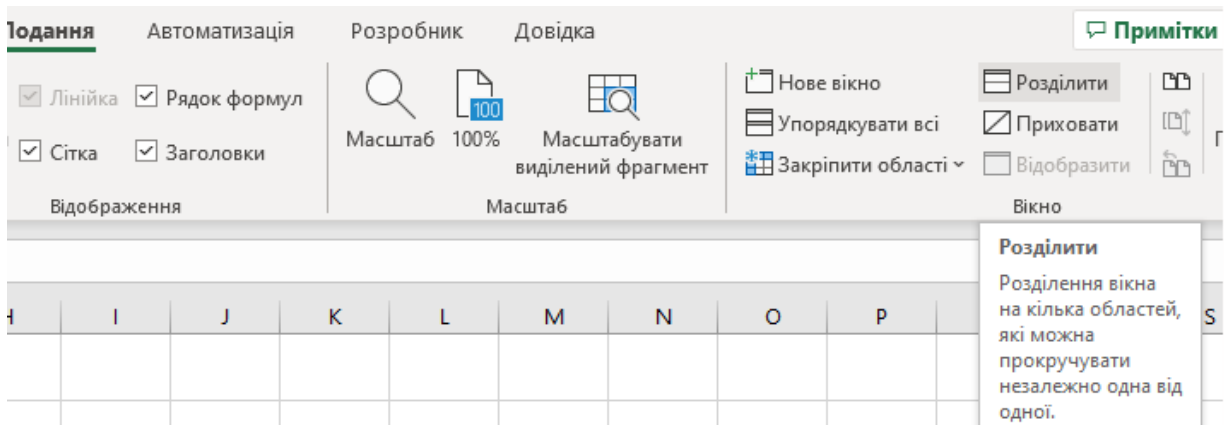


Рисунок 1.9

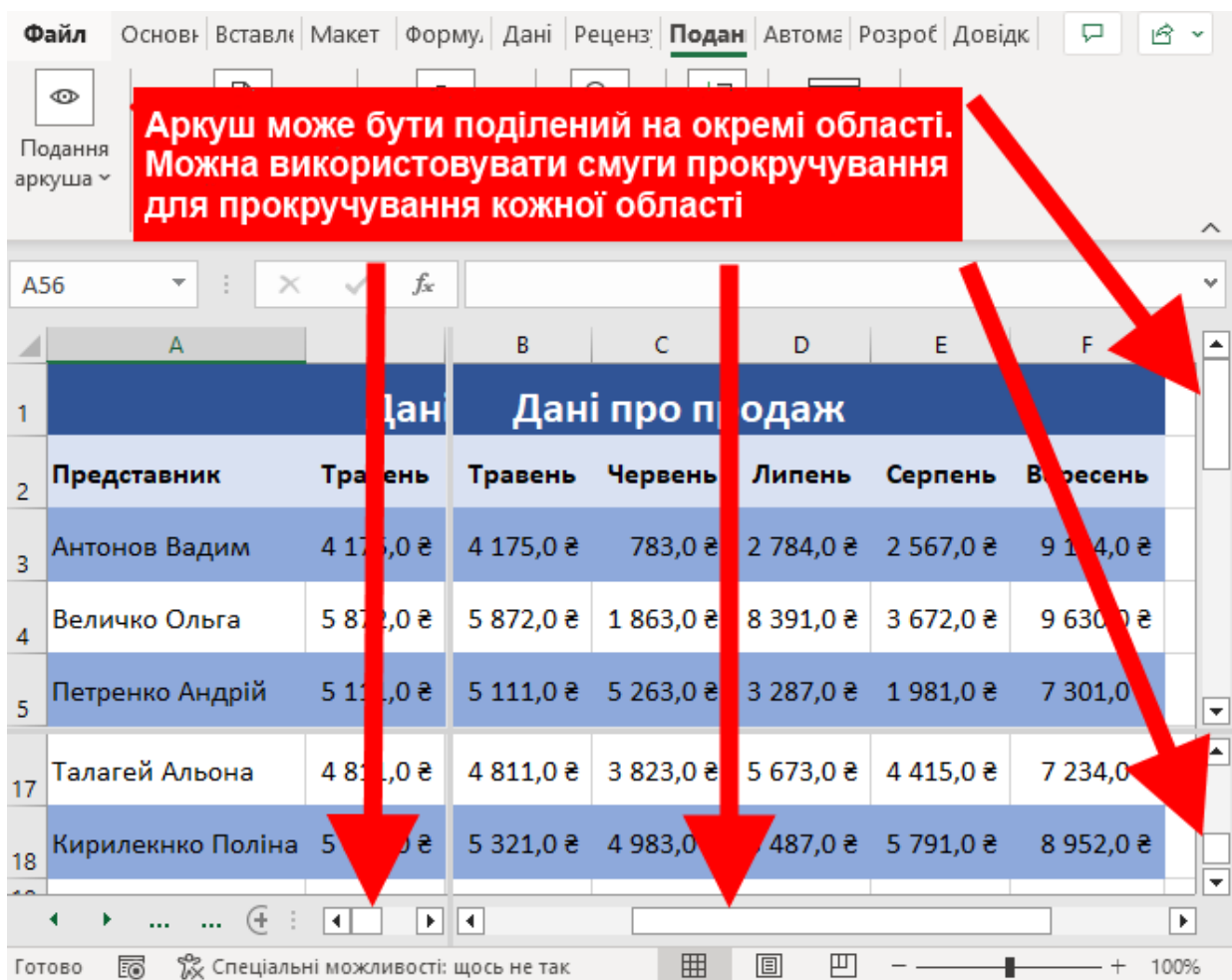


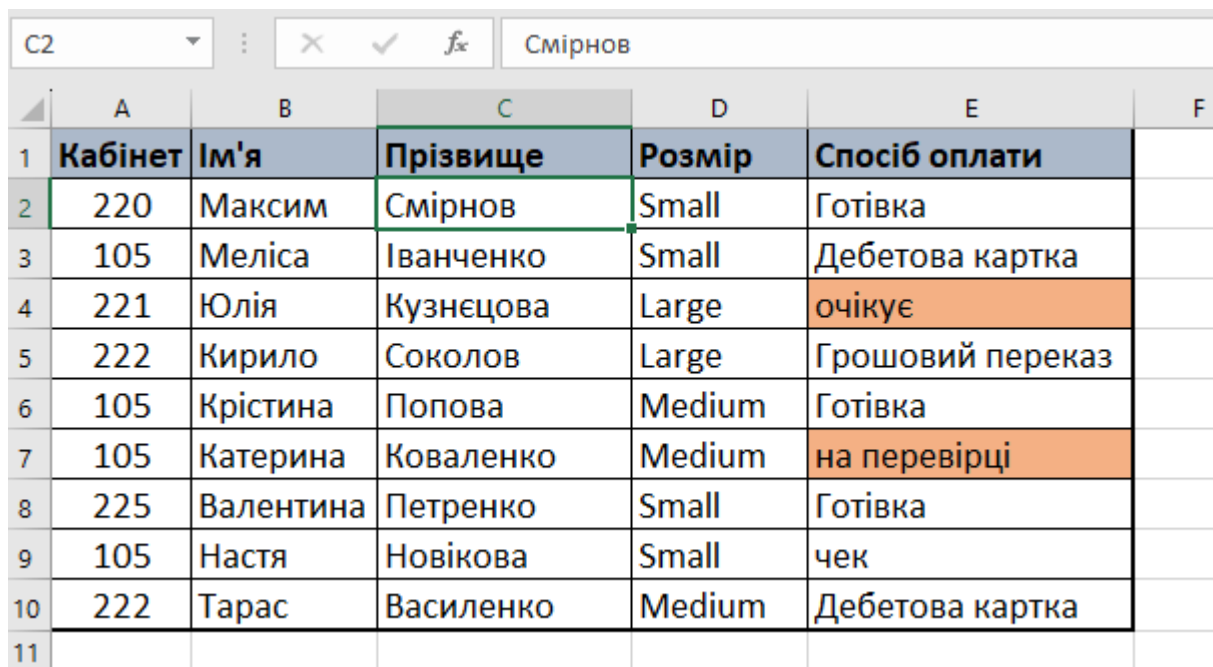
Рисунок 1.10

4. Сортвання даних

Сортвання даних в Excel – це дуже корисний інструмент, який дозволяє покращувати сприйняття інформації, особливо при її великих обсягах. В якості

прикладу відсортуюмо форму замовлення медичних халатів по Прізвищах (Стовпець 3) і розташуємо їх в алфавітному порядку.

4.1 Виділіть клітинку в стовпці, по якій необхідно виконати сортування. У нашому прикладі виділимо клітинку **C2**.



	A	B	C	D	E	F
1	Кабінет	Ім'я	Прізвище	Розмір	Спосіб оплати	
2	220	Максим	Смірнов	Small	Готівка	
3	105	Меліса	Іванченко	Small	Дебетова картка	
4	221	Юлія	Кузнєцова	Large	очікує	
5	222	Кирило	Соколов	Large	Грошовий переказ	
6	105	Крістіна	Попова	Medium	Готівка	
7	105	Катерина	Коваленко	Medium	на перевірці	
8	225	Валентина	Петренко	Small	Готівка	
9	105	Настя	Новікова	Small	чек	
10	222	Тарас	Василенко	Medium	Дебетова картка	
11						

Рисунок 1.11

4.2 Відкрийте вкладку **Дані** на **Стрічці головного меню**, потім натисніть команду **Сортування від А до Я**, щоб відсортувати за збільшенням, або команду **Сортування від Я до А**, щоб відсортувати за спаданням. У нашому прикладі ми виберемо команду **Сортування від А до Я**.

Зауваження. При виділенні діапазону сортування відбувається у межах цього діапазону.

Також **Excel** дозволяє зробити сортування що настроюється самим користувачем. В якості прикладу розглянемо сортування даних на аркуші **Excel** за розміром медичних халатів (стовпець **D**). Звичайне сортування розставить розміри в алфавітному порядку, що буде не зовсім правильне. Треба створити список, що настроюється, для сортування розмірів від меншого до більшого.

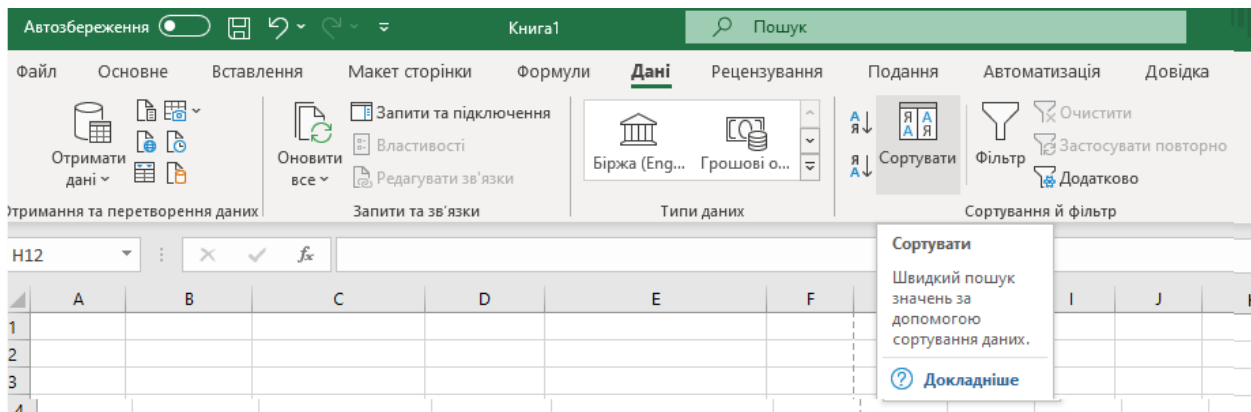


Рисунок 1.12

Таблиця буде відсортована по вибраному стовпцю, тобто по прізвищу (рисунок 1.13).

	A	B	C	D	E
1	Кабінет	Ім'я	Прізвище	Розмір	Спосіб оплати
2	222	Тарас	Василенко	Medium	Дебетова картка
3	105	Меліса	Іванченко	Small	Дебетова картка
4	105	Катерина	Коваленко	Medium	на перевірці
5	221	Юлія	Кузнєцова	Large	очікує
6	105	Настя	Новікова	Small	чек
7	225	Валентина	Петренко	Small	Готівка
8	105	Крістіна	Попова	Medium	Готівка
9	220	Максим	Смірнов	Small	Готівка
10	222	Кирило	Соколов	Large	Грошовий переказ
11					

Рисунок 1.13

4.4 Виділіть будь-яку клітинку в таблиці **Excel**, яку необхідно сортувати. У даному прикладі ми виділимо клітинку **D2**.

4.5 Відкрийте вкладку **Дані**, потім натисніть команду **Сортування**.

4.6 Відкриється діалогове вікно **Сортування**. Виберіть стовпець, за яким Ви хочете сортувати таблицю. В даному випадку ми виберемо сортування за **Розміром**

медичних халатів (рис.1.15). Потім в полі **Порядок** виберіть пункт **Настроюваний список...**

	A	B	C	D	E
1	Кабінет	Ім'я	Прізвище	Розмір	Спосіб оплати
2	222	Тарас	Василенко	Medium	Дебетова картка
3	105	Меліса	Іванченко	Small	Дебетова картка
4	105	Катерина	Коваленко	Medium	на перевірку
5	221	Юлія	Кузнєцова	Large	очікує
6	105	Настя	Новікова	Small	чек
7	225	Валентина	Петренко	Small	Готівка
8	105	Крістіна	Попова	Medium	Готівка
9	220	Максим	Смірнов	Small	Готівка
10	222	Кирило	Соколов	Large	Грошовий переказ

Рисунок 1.14

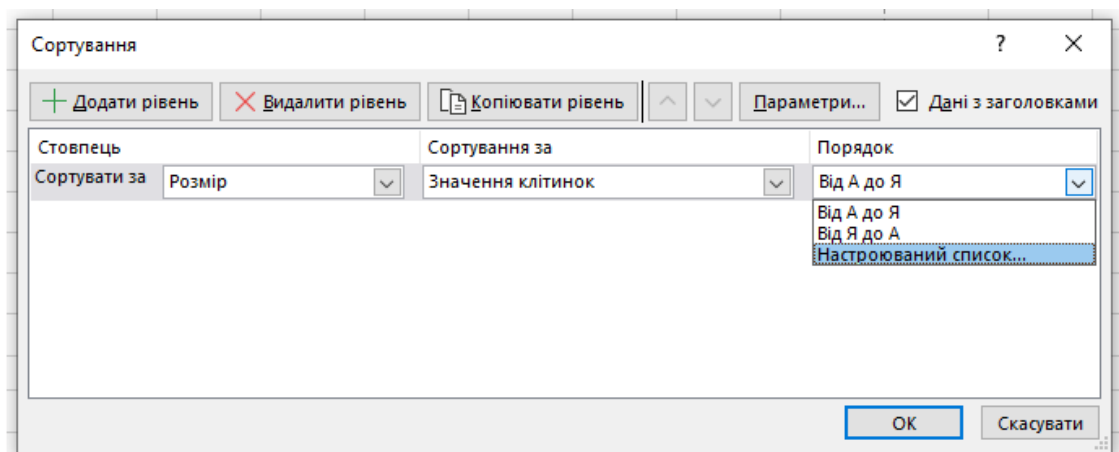


Рисунок 1.15

4.7 З'явиться діалогове вікно **Списки**. Виберіть **НОВИЙ СПИСОК** в розділі **Списки** (рис.1.15).

4.8 Введіть розміри халатів в поле **Елементи списку** в необхідному порядку. У нашому прикладі треба відсортувати розміри від меншого до більшого, тому введемо по черзі: Small, Medium, Large, натискаючи клавішу **Enter** після кожного елементу (рис.1.16).

4.9 Клацніть **Додати**, щоб зберегти новий порядок сортування. Список буде доданий в розділ **Списки**. Переконайтеся, що вибраний саме він, і натисніть **ОК**.

4.10 Діалогове вікно **Списки** закриється. Натисніть **ОК** в діалоговому вікні **Сортування** для того, щоб виконати призначене для користувача сортування (рис.1.17).

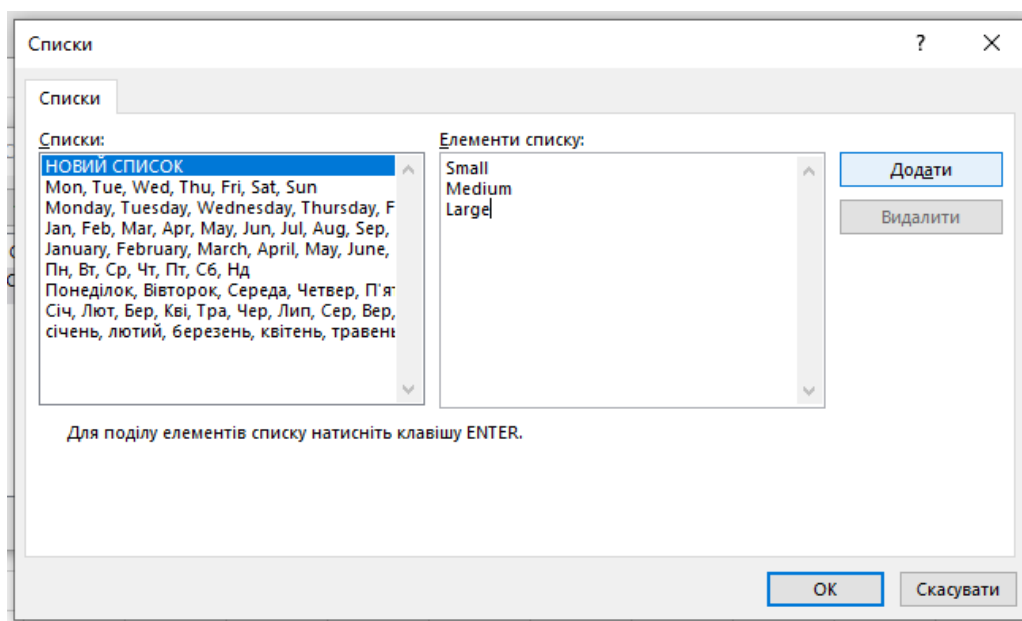


Рисунок 1.16

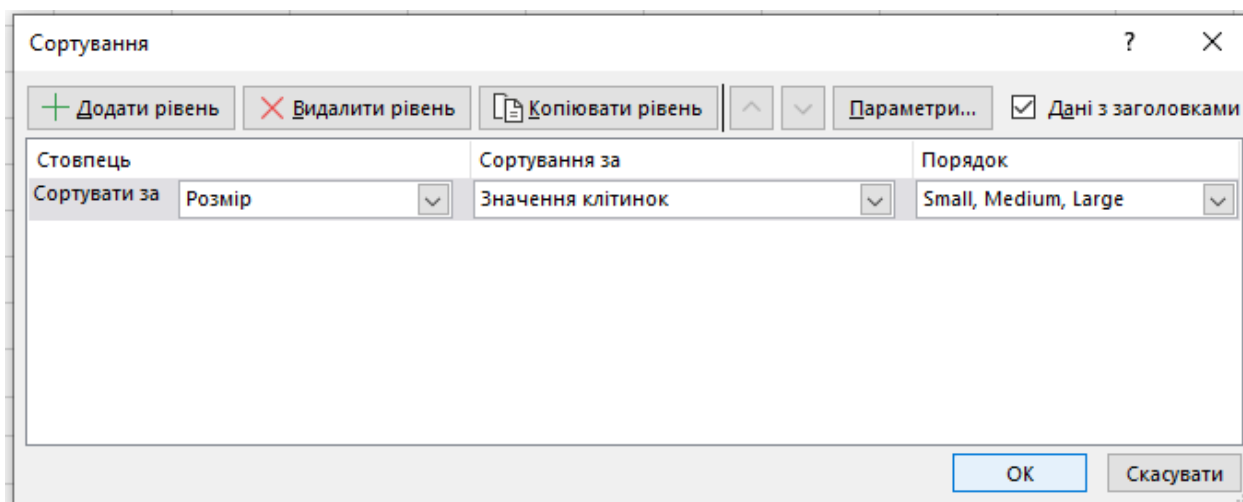


Рисунок 1.17

Таблиця **Excel** буде відсортована в необхідному порядку, в нашому випадку - за розміром медичних халатів від меншого до більшого (рис.1.18).

D2				Small	
	A	B	C	D	E
1	Кабінет	Ім'я	Прізвище	Розмір	Спосіб оплати
2	105	Меліса	Іванченко	Small	Дебетова картка
3	105	Настя	Новікова	Small	чек
4	225	Валентина	Петренко	Small	Готівка
5	220	Максим	Смірнов	Small	Готівка
6	222	Тарас	Василенко	Medium	Дебетова картка
7	105	Катерина	Коваленко	Medium	на перевірку
8	105	Крістіна	Попова	Medium	Готівка
9	221	Юлія	Кузнецова	Large	очікує
10	222	Кирило	Соколов	Large	Грошовий переказ
11					

Рисунок 1.18

Також **Excel** дозволяє виконувати множинне сортування, тобто таке сортування, що складається з декількох рівнів.

В якості прикладу розглянемо сортування таблиці за номером **Кабінету** (стовпець **A**) і за **Прізвищем** (стовпець **C**) (рис.1.19).

4.11 За допомогою миші виділіть усю таблицю.

A2				222	
	A	B	C	D	E
1	Кабінет	Ім'я	Прізвище	Розмір	Спосіб оплати
2	222	Тарас	Василенко	Medium	Дебетова картка
3	105	Меліса	Іванченко	Small	Дебетова картка
4	105	Катерина	Коваленко	Medium	на перевірку
5	221	Юлія	Кузнецова	Large	очікує
6	105	Настя	Новікова	Small	чек
7	225	Валентина	Петренко	Small	Готівка
8	105	Крістіна	Попова	Medium	Готівка
9	220	Максим	Смірнов	Small	Готівка
10	222	Кирило	Соколов	Large	Грошовий переказ
11					

Рисунок 1.19

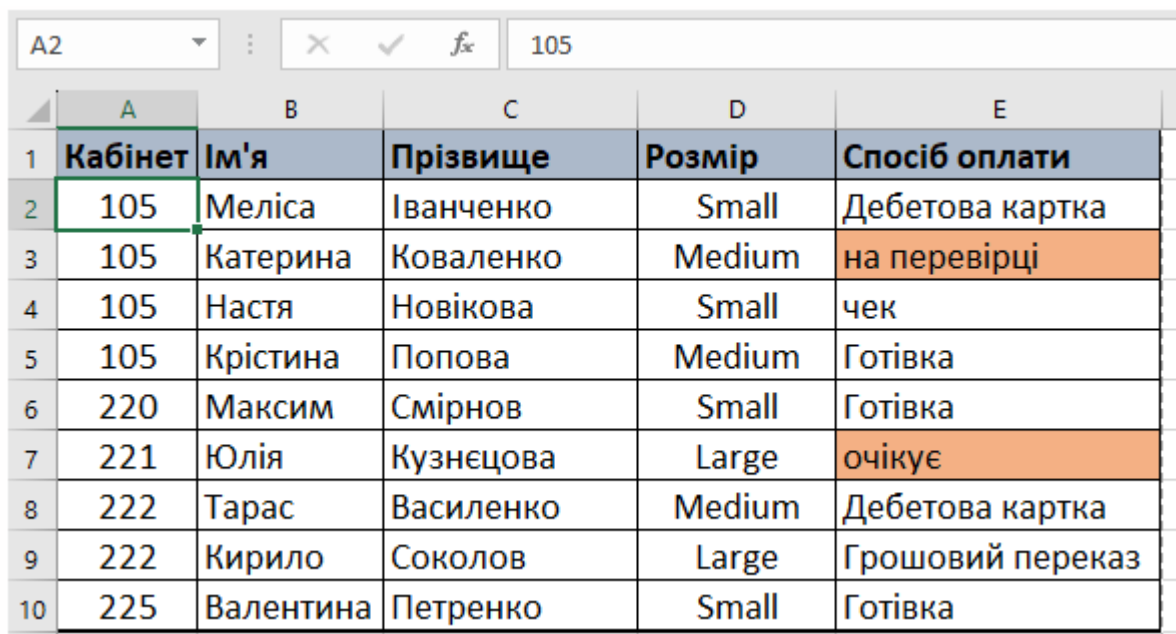
4.12 Відкрийте вкладку **Дані**, потім натисніть команду **Сортування**.

4.13 Відкриється діалогове вікно **Сортування**. З випадаючого списку виберіть стовпець, за яким треба виконати сортування. У даному прикладі ми сортуватимемо за стовпцем **Кабінет** (стовпець **A**).

4.14 Клацніть **Додати рівень**, щоб додати ще один стовпець для сортування.

4.15 З випадаючого списку виберіть другий стовпець, за яким треба сортувати таблицю, потім натисніть **ОК**. У нашому прикладі ми сортуватимемо за **Прізвищем** (стовпець **C**).

Таблиця буде відсортована відповідно до вибраного порядку (рис. 1.20). У нашому прикладі номери кабінетів сортуються за збільшенням, а для кожного **Кабінету** студенти розташовані за **Прізвищем** в алфавітному порядку.



	A	B	C	D	E
1	Кабінет	Ім'я	Прізвище	Розмір	Спосіб оплати
2	105	Меліса	Іванченко	Small	Дебетова картка
3	105	Катерина	Коваленко	Medium	на перевірці
4	105	Настя	Новікова	Small	чек
5	105	Крістіна	Попова	Medium	Готівка
6	220	Максим	Смірнов	Small	Готівка
7	221	Юлія	Кузнецова	Large	очікує
8	222	Тарас	Василенко	Medium	Дебетова картка
9	222	Кирило	Соколов	Large	Грошовий переказ
10	225	Валентина	Петренко	Small	Готівка

Рисунок 1.20

*Зауваження. Якщо необхідно змінити порядок багаторівневого сортування в **Excel**, то необхідно вказати стовпець, який сортується в першу чергу. Для цього виділіть необхідний стовпець, а потім натисніть стрілку вгору або вниз, щоб змінити пріоритет.*

Практичне завдання

1. Створіть та заповніть даними таблицю (див. рис.1.21). В якості діагнозу обрати: **хворий, нехворий**. Зробіть скріншот.

Ім'я	Прізвище	Стать	Вік	Температура тіла	Діагноз людини

Рисунок 1.21

2. Виконати сортування за кількома критеріями:

- діагноз пацієнта (настроюваний список: хворий, нехворий);
- прізвище (від А до Я);
- вік (за зростанням);
- температура (за спаданням).

Зробіть скріншот.

3. Змінити критерії сортування:

- прізвище (від А до Я);
- стать (настроюваний список: чоловіча, жіноча);
- температура (за зростанням);
- вік (за спаданням).

Зробіть скріншот.

4. Закріпіть 1,2 і 3 рядки і прокрутіть робочий аркуш, щоб було видно лише рядки 1,2,3,9,10,11. Зробіть скріншот.

5. Розділіть таблицю на 4 області так, щоб у ліву верхню область входило 8 рядків і 2 стовпці. Зробіть скріншот.

6. Відкрийте таблицю у новому вікні. Зробіть скріншот.

7. Усі скриншоти розмістіть у файлі Word, та прикріпіть файл до завдання.

ПРАКТИЧНА РОБОТА 2

Робота з даними в програмі Microsoft Excel

Питання, які розглядаються:

1. Фільтрація даних.
2. Графічне подання даних. Діаграми.

Програма **Excel** за допомогою фільтрації даних дозволяє відобразити ту частину інформації, яка потрібна, і також подати її у графічному вигляді.

1. Фільтрація даних

У наступному прикладі ми застосуємо фільтр до журналу обліку експлуатації устаткування, щоб відобразити тільки ноутбуки і планшети, доступні для перевірки.

Щоб фільтрація в **Excel** працювала правильно, аркуш повинен містити рядок заголовка, який використовується для завдання імені кожного стовпця. У наступному прикладі дані на аркуші організовані у вигляді стовпців із заголовками в рядку 1: ID #, Тип, Опис обладнання і так далі (рис.2.1).

1.1 Виділіть будь-яку клітинку в таблиці, наприклад, клітинку **A2**.

	A	B	C	D	E	F
1	ID #	Тип	Опис обладнання	Віддано на перевірку	Перевірено	Виконавець
2	3000	Камера	Saris Lumina Digital Camera	12.Тра.17	15.Тра.17	Геращенко Микола
3	3005	Камера	Saris Zoom Z-60 Camera	27.Чер.17	06.Сер.17	Зінченко Валерій
4	3070	Камера	Omega PixL Digital Camera	10.Жов.17	07.Жов.17	Клевцов Андрій
5	1021	Ноутбук	15" EDI SmartPad L200-3	15.Вер.17	01.Жов.17	Васильєв Іван
6	1022	Ноутбук	15" EDI SmartPad L200-3	14.Сер.17	16.Сер.17	Головко Юрій
7	1023	Ноутбук	15" EDI SmartPad L200-3	08.Сер.17	15.Сер.17	Густенко Вадим
8	1025	Ноутбук	17" Saris X-10 Laptop	26.Вер.17	04.Жов.17	Мащак Микита
9	1031	Ноутбук	17" Saris X-10 Laptop	04.Жов.17	08.Жов.17	Мігель Вадим
10	1032	Ноутбук	17" Saris X-10 Laptop	19.Вер.17	21.Вер.17	Мороз Максим
11	1033	Ноутбук	17" Saris X-10 Laptop	24.Вер.17	26.Вер.17	П'ятаченко Антон

Рисунок 2.1

1.2 Відкрийте вкладку **Дані**, потім натисніть команду **Фільтр** (рис.2.1).

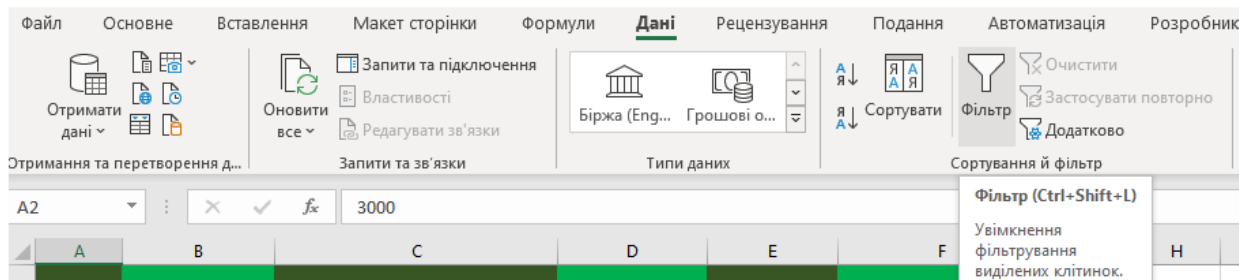


Рисунок 2.1

1.3 У заголовках кожного стовпця з'являться кнопки із стрілкою. Натисніть на таку кнопку в стовпці, який необхідно відфільтрувати. У нашому випадку ми застосуємо фільтр до стовпця **В**, щоб побачити тільки потрібні типи обладнання (рис.2.3).

	A	B	C	D	E	F
	ID #	Тип	Опис обладнання	Віддано на перевірку	Перевірено	Виконавець
1						
2	3000	Камера	Saris Lumina Digital Camera	12.Тра.17	15.Тра.17	Герашенко Микола
3	3005	Камера	Omega PIXL Digital Camera	27.Чер.17	06.Сеп.17	Зінченко Валерій
4	3070	Камера	Omega PIXL Digital Camera	10.Жов.17	07.Жов.17	Клевцов Андрій
5	1021	Ноутбук	15" EDI SmartPad L200-3	15.Вер.17	01.Жов.17	Васильєв Іван
6	1022	Ноутбук	15" EDI SmartPad L200-3	14.Сеп.17	16.Сеп.17	Головко Юрій
7	1023	Ноутбук	15" EDI SmartPad L200-3	08.Сеп.17	15.Сеп.17	Густенко Вадим
8	1025	Ноутбук	17" Saris X-10 Laptop	26.Вер.17	04.Жов.17	Мащак Микита
9	1031	Ноутбук	17" Saris X-10 Laptop	04.Жов.17	08.Жов.17	Мігель Вадим
10	1032	Ноутбук	17" Saris X-10 Laptop	19.Вер.17	21.Вер.17	Мороз Максим
11	1033	Ноутбук	17" Saris X-10 Laptop	24.Вер.17	26.Вер.17	П'ятаченко Антон

Рисунок 2.3

1.4 З'явиться меню фільтру. Зніміть прапорець **Виділити все**, щоб швидко зняти виділення зі всіх пунктів (рис.2.4).

1.5 Встановіть прапорці для тих типів обладнання, які необхідно залишити в таблиці, потім натисніть **ОК**. У нашому прикладі ми виберемо **Ноутбуки** і **Планшети**, щоб бачити тільки ці типи обладнання (рис.2.5).

Таблиця з даними буде відфільтрована, при цьому буде тимчасово прихований весь вміст, що не відповідає критерію. У нашому прикладі тільки ноутбуки і планшети залишилися у переліку опису обладнання (рис.2.6).

	A	B	C	D
	ID #	Тип	Опис обладнання	Віддано на перевірку
1				
я↓	Сортування від А до Я		на Digital Camera	12.Тра.17
я↓	Сортування від Я до А		n Z-60 Camera	27.Чер.17
	Сортування за кольором		L Digital Camera	10.Жов.17
	Подання аркуша		artPad L200-3	15.Вер.17
	Очистити фільтр у "Тип"		artPad L200-3	14.Сер.17
	Фільтрування за кольором		artPad L200-3	08.Сер.17
	Текстові фільтри		-10 Laptop	26.Вер.17
	Пошук		-10 Laptop	04.Жов.17
			-10 Laptop	19.Вер.17
			-10 Laptop	24.Вер.17
			-10 Laptop	25.Сер.17
			board L500-1	04.Жов.17
			board L500-1	29.Вер.17
			DigiCam Printer II	04.Сер.17
			Rolling Laptop Case	04.Жов.17
			50	11.Сер.17
			50	17.Лип.17
			50	01.Жов.17
			-80	28.Вер.17
			-81 Lite	26.Вер.17
			pro	04.Жов.17
			pro	29.Вер.17
24	1014	Планшет	Saris SlimPro	01.Жов.17
25	6200	Проектор	Saris Lux T-80	01.Вер.17

Рисунок 1.4

	A	B	C	D
	ID #	Тип	Опис обладнання	Віддано на перевірку
1				
я↓	Сортування від А до Я		на Digital Camera	12.Тра.17
я↓	Сортування від Я до А		n Z-60 Camera	27.Чер.17
	Сортування за кольором		L Digital Camera	10.Жов.17
	Подання аркуша		artPad L200-3	15.Вер.17
	Очистити фільтр у "Тип"		artPad L200-3	14.Сер.17
	Фільтрування за кольором		artPad L200-3	08.Сер.17
	Текстові фільтри		-10 Laptop	26.Вер.17
	Пошук		-10 Laptop	04.Жов.17
			-10 Laptop	19.Вер.17
			-10 Laptop	24.Вер.17
			-10 Laptop	25.Сер.17
			board L500-1	04.Жов.17
			50	17.Лип.17
			50	01.Жов.17
			-80	28.Вер.17
			-81 Lite	26.Вер.17
			pro	04.Жов.17
			pro	29.Вер.17
24	1014	Планшет	Saris SlimPro	01.Жов.17
25	6200	Проектор	Saris Lux T-80	01.Вер.17

Рисунок 1.5

	A	B	C	D	E	F
1	ID #	Тип	Опис обладнання	Віддано на перевірку	Перевірено	Виконавець
5	1021	Ноутбук	15" EDI SmartPad L200-3	15.Вер.17	01.Жов.17	Васильєв Іван
6	1022	Ноутбук	15" EDI SmartPad L200-3	14.Сер.17	16.Сер.17	Головко Юрій
7	1023	Ноутбук	15" EDI SmartPad L200-3	08.Сер.17	15.Сер.17	Густенко Вадим
8	1025	Ноутбук	17" Saris X-10 Laptop	26.Вер.17	04.Жов.17	Мащак Микита
9	1031	Ноутбук	17" Saris X-10 Laptop	04.Жов.17	08.Жов.17	Мігель Вадим
10	1032	Ноутбук	17" Saris X-10 Laptop	19.Вер.17	21.Вер.17	Мороз Максим
11	1033	Ноутбук	17" Saris X-10 Laptop	24.Вер.17	26.Вер.17	П'ятаченко Антон
12	1034	Ноутбук	17" Saris X-10 Laptop	25.Сер.17	27.Сер.17	Козулько Андрій
22	1011	Планшет	Saris SlimPro	04.Жов.17	07.Жов.17	Байдаченко Віктор
23	1012	Планшет	Saris SlimPro	29.Вер.17	30.Вер.17	Батаєв Юрій
24	1014	Планшет	Saris SlimPro	01.Жов.17	05.Жов.17	Видмиш Костянтин
31						
32						

Рисунок 2.6

Зауваження. Фільтрацію можна також застосувати, вибравши команду **Сортувати й фільтрувати** на вкладці **Основне** (рис.1.7).

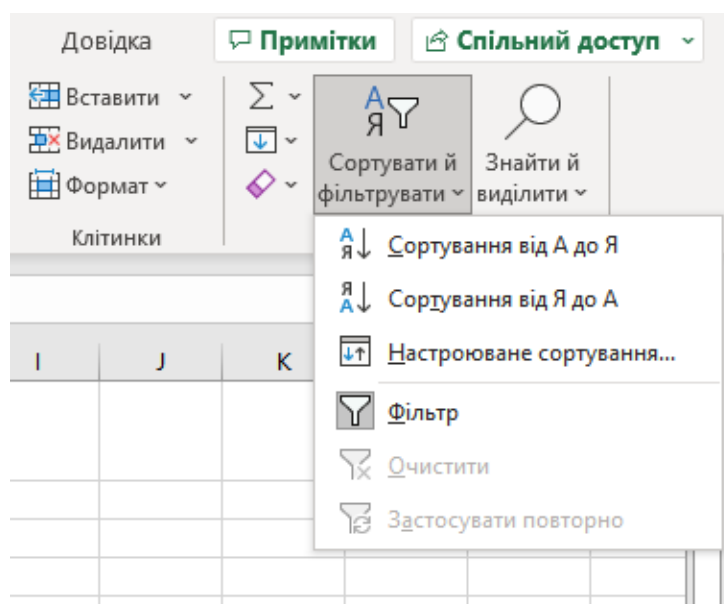


Рисунок 1.7

В **Excel** можна застосувати декілька фільтрів до однієї таблиці, щоб звузити результати фільтрації. У минулому прикладі ми вже відфільтрували таблицю, відобразив тільки ноутбуки і планшети. Тепер наше завдання звузити дані ще більше і показати тільки ноутбуки і планшети, віддані на перевірку в серпні.

1.6 Натисніть на кнопку із стрілкою в стовпці, який необхідно відфільтрувати. В даному випадку ми застосуємо додатковий фільтр до стовпця **D**, щоб проглянути інформацію по даті (рис.2.8).

	A	B	C	D	E	F
1	ID #	Тип	Опис обладнання	Віддано на перевірку	Перевірено	Виконавець
5	1021	Ноутбук	15" EDI SmartPad L200-3	15.Вер.17	01.Жов.17	Васильєв Іван
6	1022	Ноутбук	15" EDI SmartPad L200-3	14.Сер.17	16.Сер.17	Головко Юрій
7	1023	Ноутбук	15" EDI SmartPad L200-3	08.Сер.17	15.Сер.17	Густенко Вадим
8	1025	Ноутбук	17" Saris X-10 Laptop	26.Вер.17	04.Жов.17	Мащак Микита
9	1031	Ноутбук	17" Saris X-10 Laptop	04.Жов.17	08.Жов.17	Мігель Вадим
10	1032	Ноутбук	17" Saris X-10 Laptop	19.Вер.17	21.Вер.17	Мороз Максим
11	1033	Ноутбук	17" Saris X-10 Laptop	24.Вер.17	26.Вер.17	П`ятаченко Антон
12	1034	Ноутбук	17" Saris X-10 Laptop	25.Сер.17	27.Сер.17	Козулько Андрій
22	1011	Планшет	Saris SlimPro	04.Жов.17	07.Жов.17	Байдаченко Віктор
23	1012	Планшет	Saris SlimPro	29.Вер.17	30.Вер.17	Батаєв Юрій
24	1014	Планшет	Saris SlimPro	01.Жов.17	05.Жов.17	Видмиш Костянтин
31						

Рисунок 2.8

З'явиться меню фільтру. Встановіть або зніміть прапорці з пунктів залежно від даних, які необхідно відфільтрувати, потім натисніть **ОК**. Ми знімемо виділення зі всіх пунктів, окрім серпня (рис.2.9).

B2						Камера
	A	B	C	D	E	F
1	ID #	Тип	Опис обладнання	Віддано на перевірку	Перевірено	Виконавець
5	1021	Ноутбук	Сортування від найстаршого до наймолодшого		01.Жов.17	Васильєв Іван
6	1022	Ноутбук	Сортування від наймолодшого до найстаршого		16.Сер.17	Головко Юрій
7	1023	Ноутбук	Сортування за кольором		15.Сер.17	Густенко Вадим
8	1025	Ноутбук	Подання аркуша		04.Жов.17	Мащак Микита
9	1031	Ноутбук	Очистити фільтр у "Віддано на перевірку"		08.Жов.17	Мігель Вадим
10	1032	Ноутбук	Фільтрування за кольором		21.Вер.17	Мороз Максим
11	1033	Ноутбук	Фільтри дат		26.Вер.17	П`ятаченко Антон
12	1034	Ноутбук	Пошук (усі)		27.Сер.17	Козулько Андрій
22	1011	Планшет	(Виділити все)		07.Жов.17	Байдаченко Віктор
23	1012	Планшет	2017		30.Вер.17	Батаєв Юрій
24	1014	Планшет	серпень		05.Жов.17	Видмиш Костянтин
31			вересень			
32			жовтень			
33						
34						
35						
36						
37						
38						
39						
40						
41						
42						

Рисунок 1.9

Новий фільтр буде застосований, а в таблиці залишаться тільки ноутбуки і планшети, які були віддані на перевірку в серпні (рис.2.10).

B2						Камера
	A	B	C	D	E	F
1	ID #	Тип	Опис обладнання	Віддано на перевірку	Перевірено	Виконавець
6	1022	Ноутбук	15" EDI SmartPad L200-3	14.Сер.17	16.Сер.17	Головко Юрій
7	1023	Ноутбук	15" EDI SmartPad L200-3	08.Сер.17	15.Сер.17	Густенко Вадим
12	1034	Ноутбук	17" Saris X-10 Laptop	25.Сер.17	27.Сер.17	Козулько Андрій
31						
32						

Рисунок 2.10

*Зауваження. Для зняття фільтра в якомусь із стовпців потрібно після натискання стрілки вибрати пункт **Очистити фільтр**.*

Якщо раптом виникає необхідність виділити якісь специфічні дані, то, як правило, базові інструменти фільтрації з таким завданням вже не справляються. На щастя, **Excel** містить в залежності від типу даних стовпця багато **розширених фільтрів**, включаючи пошук тексту, дати і фільтрацію по числових значеннях. Це дозволяє звузити результати і допомогти знайти саме те, що потрібне. Таким чином **Excel** дозволяє шукати інформацію, яка містить точну фразу, число, дату і багато що іншого. Покажемо на прикладі, як фільтрувати дані за допомогою **числового фільтра**.

1.7 Відкрийте вкладку **Дані**, потім натисніть команду **Фільтр**. У кожному заголовку стовпця з'явиться кнопка із стрілкою. Якщо Ви вже застосовували фільтри в таблиці, можете пропустити цей крок. Натисніть на кнопку із стрілкою в стовпці, який необхідно відфільтрувати. В даному прикладі виберемо стовпець **A**, щоб побачити заданий ряд ідентифікаційних номерів (рис.2.11).

1.8 З'явиться меню фільтру. Наведіть покажчик миші на пункт **Фільтри чисел**, потім виберіть необхідний числовий фільтр в меню, що розкривається. У

даному прикладі ми виберемо **Між**, щоб побачити ідентифікаційні номери в певному діапазоні (рис.2.12).

	A	B	C	D	E	F
1	ID #	Тип	Опис обладнання	Віддано на перевірку	Перевірено	Виконавець
2	3000	Камера	Saris Lumina Digital Camera	12.Тра.17	15.Тра.17	Геращенко Микола
3	3005	Камера	Saris Zoom Z-60 Camera	27.Чер.17	06.Сер.17	Зінченко Валерій
4	3070	Камера	Omega PixL Digital Camera	10.Жов.17	07.Жов.17	Клевцов Андрій
5	1021	Ноутбук	15" EDI SmartPad L200-3	15.Вер.17	01.Жов.17	Васильєв Іван
6	1022	Ноутбук	15" EDI SmartPad L200-3	14.Сер.17	16.Сер.17	Головко Юрій
7	1023	Ноутбук	15" EDI SmartPad L200-3	08.Сер.17	15.Сер.17	Густенко Вадим
8	1025	Ноутбук	17" Saris X-10 Laptop	26.Вер.17	04.Жов.17	Мащак Микита
9	1031	Ноутбук	17" Saris X-10 Laptop	04.Жов.17	08.Жов.17	Мігель Вадим
10	1032	Ноутбук	17" Saris X-10 Laptop	19.Вер.17	21.Вер.17	Мороз Максим
11	1033	Ноутбук	17" Saris X-10 Laptop	24.Вер.17	26.Вер.17	П'ятаченко Антон
12	1034	Ноутбук	17" Saris X-10 Laptop	25.Сер.17	27.Сер.17	Козулько Андрій

Рисунок 2.11

	A	B	C	D	E	F
1	ID #	Тип	Опис обладнання	Віддано на перевірку	Перевірено	Виконавець
2	3000	Камера	Saris Lumina Digital Camera	12.Тра.17	15.Тра.17	Геращенко Микола
3	3005	Камера	Saris Zoom Z-60 Camera	27.Чер.17	06.Сер.17	Зінченко Валерій
4	3070	Камера	Omega PixL Digital Camera	10.Жов.17	07.Жов.17	Клевцов Андрій
5	1021	Ноутбук	15" EDI SmartPad L200-3	15.Вер.17	01.Жов.17	Васильєв Іван
6	1022	Ноутбук	15" EDI SmartPad L200-3	14.Сер.17	16.Сер.17	Головко Юрій
7	1023	Ноутбук	15" EDI SmartPad L200-3	08.Сер.17	15.Сер.17	Густенко Вадим
8	1025	Ноутбук	17" Saris X-10 Laptop	26.Вер.17	04.Жов.17	Мащак Микита
9	1031	Ноутбук	17" Saris X-10 Laptop	04.Жов.17	08.Жов.17	Мігель Вадим
10	1032	Ноутбук	17" Saris X-10 Laptop	19.Вер.17	21.Вер.17	Мороз Максим
11	1033	Ноутбук	17" Saris X-10 Laptop	24.Вер.17	26.Вер.17	П'ятаченко Антон
12	1034	Ноутбук	17" Saris X-10 Laptop	25.Сер.17	27.Сер.17	Козулько Андрій
13	1011	Планшет	Saris SlimPro	02.Жов.17	01.Жов.17	Калакай Олександр
14	1012	Планшет	Saris SlimPro	02.Жов.17	01.Жов.17	Кишко Олексій
15	1014	Планшет	Saris SlimPro	02.Жов.17	01.Жов.17	Красуляк Антон
16	1021	Ноутбук	15" EDI SmartPad L200-3	15.Вер.17	01.Жов.17	Красуляк Антон
17	1022	Ноутбук	15" EDI SmartPad L200-3	14.Сер.17	16.Сер.17	Байдаченко Віктор
18	1023	Ноутбук	15" EDI SmartPad L200-3	08.Сер.17	15.Сер.17	Батаєв Юрій
19	1025	Ноутбук	17" Saris X-10 Laptop	26.Вер.17	04.Жов.17	Видмиш Костянтин
20	1031	Ноутбук	17" Saris X-10 Laptop	04.Жов.17	08.Жов.17	Вітер Пилип
21	1032	Ноутбук	17" Saris X-10 Laptop	19.Вер.17	21.Вер.17	Волошко Михайло
22	1033	Ноутбук	17" Saris X-10 Laptop	24.Вер.17	26.Вер.17	Волошко Михайло
23	1034	Ноутбук	17" Saris X-10 Laptop	25.Сер.17	27.Сер.17	Волошко Михайло
24	1014	Планшет	Saris SlimPro	02.Жов.17	01.Жов.17	Калакай Олександр
25	6200	Проектор	Saris Lux T-80	04.Вер.17	04.Вер.17	Вітер Пилип
26	6301	Проектор	Saris Lux T-81 Lite	10.Вер.17	17.Вер.17	Волошко Михайло

Рисунок 2.12

1.9 У діалоговому вікні , що з'явилося, **Спеціальний автофільтр** введіть необхідні числа для кожної з умов, потім натисніть **ОК**. В даному прикладі ми хочемо отримати номери, які **більше або дорівнює 3000**, але **менше або дорівнює 4000** (рис.2.13).

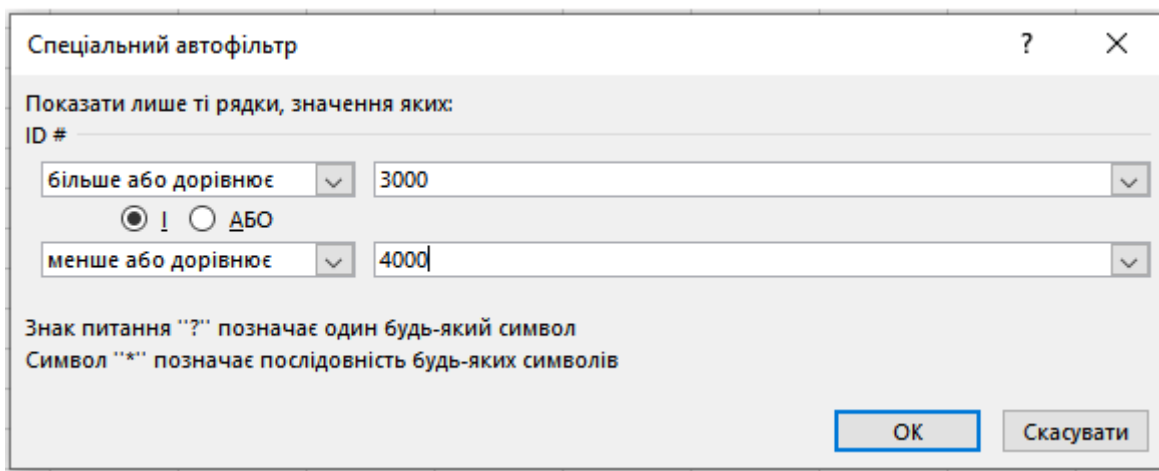


Рисунок 2.13

Дані будуть відфільтровані по заданому числовому фільтру. У нашому випадку відображаються тільки номери в діапазоні від **3000** до **4000** (рис.2.14).

	A	B	C	D	E	F
1	ID #	Тип	Опис обладнання	Віддано на перевірку	Перевірено	Виконавець
2	3000	Камера	Saris Lumina Digital Camera	12.Тра.17	15.Тра.17	Геращенко Микола
3	3005	Камера	Saris Zoom Z-60 Camera	27.Чер.17	06.Сер.17	Зінченко Валерій
4	3070	Камера	Omega PixL Digital Camera	10.Жов.17	07.Жов.17	Клевцов Андрій
15	3800	Інше	U-Go Saris DigiCam Printer II	04.Сер.17	05.Сер.17	Баранець Ігор
27	3900	Інше	U-Go Saris Label Maker	16.Чер.17	20.Чер.17	Засименко Євген
31						
32						

Рисунок 2.14

2. Графічне подання даних. Діаграми

Excel має в своєму розпорядженні прекрасний інструмент візуалізації даних – побудову діаграм. Діаграми в **Excel** здатні графічно представляти дані, що дозволяє легше сприймати великі об'єми інформації і оцінювати її. **Excel** пропонує

самі різні типи діаграм. Серед цього різноманіття можна підібрати саме ту діаграму, яка найточніше підходить під потреби. Щоб використовувати діаграми максимально ефективно, необхідно знати, як ними користуватися.



Рисунок 2.15

Діаграми в **Excel** містять 5 основних елементів, розглянемо їх (рис.2.15).

1. **Заголовок діаграми** повинен чітко описувати, що представлено на ній.
2. **Вертикальна вісь** (також відома як вісь Y) є вертикальною частиною діаграми. На вертикальній осі відображаються значення стовпців, тому її називають віссю значень. У поточному прикладі величиною вимірювання є кількість упаковок лікарського засобу, що була продана в аптеці.
3. **Ряд даних** складається із зв'язаних крапок (значень) на діаграмі. У поточному прикладі сині стовпці відображають, який самелікарський засіб був проданий у 1 кварталі. Ми розуміємо, що кількість упаковок відноситься саме до цього кварталу, завдяки легенді в правій частині діаграми.
4. **Легенда** вказує приналежність кожного ряду до кого-небудь або чого-небудь. У поточному прикладі легенда містить 4 кольори з відповідними кварталами. Дивлячись на легенду достатньо легко визначити кількість упаковок препарату, що були продані у відповідний квартал.

5. **Горизонтальна вісь** (також відома як вісь X) є горизонтальною частиною діаграми. Горизонтальна вісь представляє категорії. У даному прикладі категорії – це назви лікарських засобів.

2.1 Щоб побудувати діаграму виділіть клітинки, на основі яких бажаєте побудувати діаграму, включаючи заголовки стовпців і назви рядків. Ці осередки є джерелом даних для діаграми. У прикладі виберемо діапазон клітинок **A1:E9**.

	A	B	C	D	E
1	ПРОДАЖ ЛІКІВ В АПТЕЦІ ПО КВАРТАЛАХ				
2	НАЗВА	1 КВАРТАЛ	2 КВАРТАЛ	3 КВАРТАЛ	4 КВАРТАЛ
3	Мезим	123	139	100	56
4	Каптопрес	15	34	78	90
5	Німесил	120	80	56	23
6	Омнік	200	56	87	180
7	Парацетамол	200	300	350	190
8	Наклофен	50	70	65	100
9	Диклофенак	240	350	290	300

Рисунок 2.16

2.2 На вкладці **Вставлення** виберіть необхідну діаграму. У нашому прикладі ми виберемо **Гістограму** (рис.2.17).

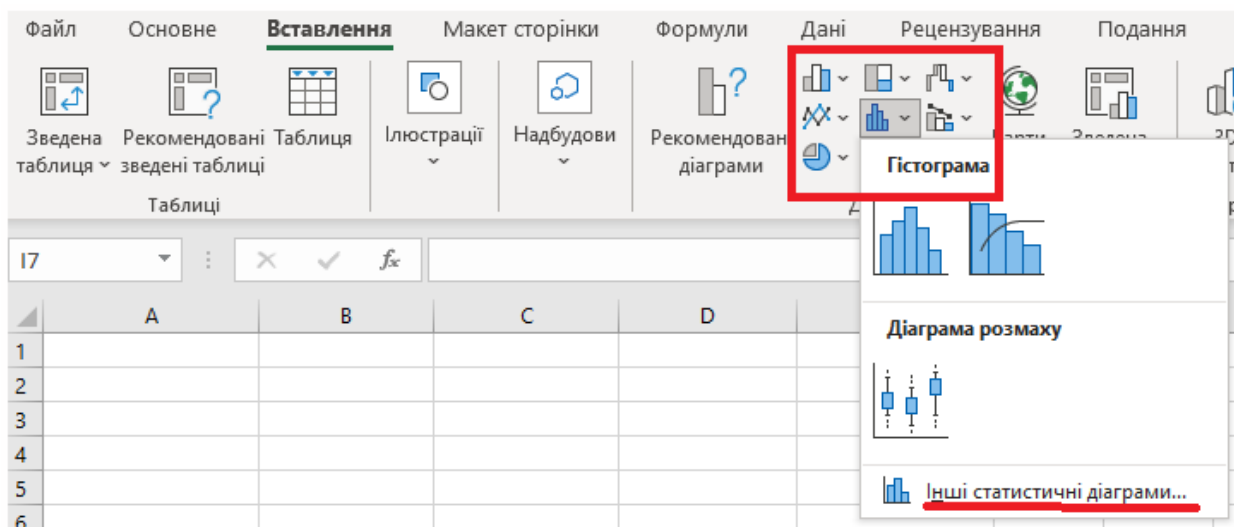


Рисунок 2.17

2.3 У меню, що розкривається, вкажіть відповідний тип гістограми **Стовпчаста** (рис.2.18).

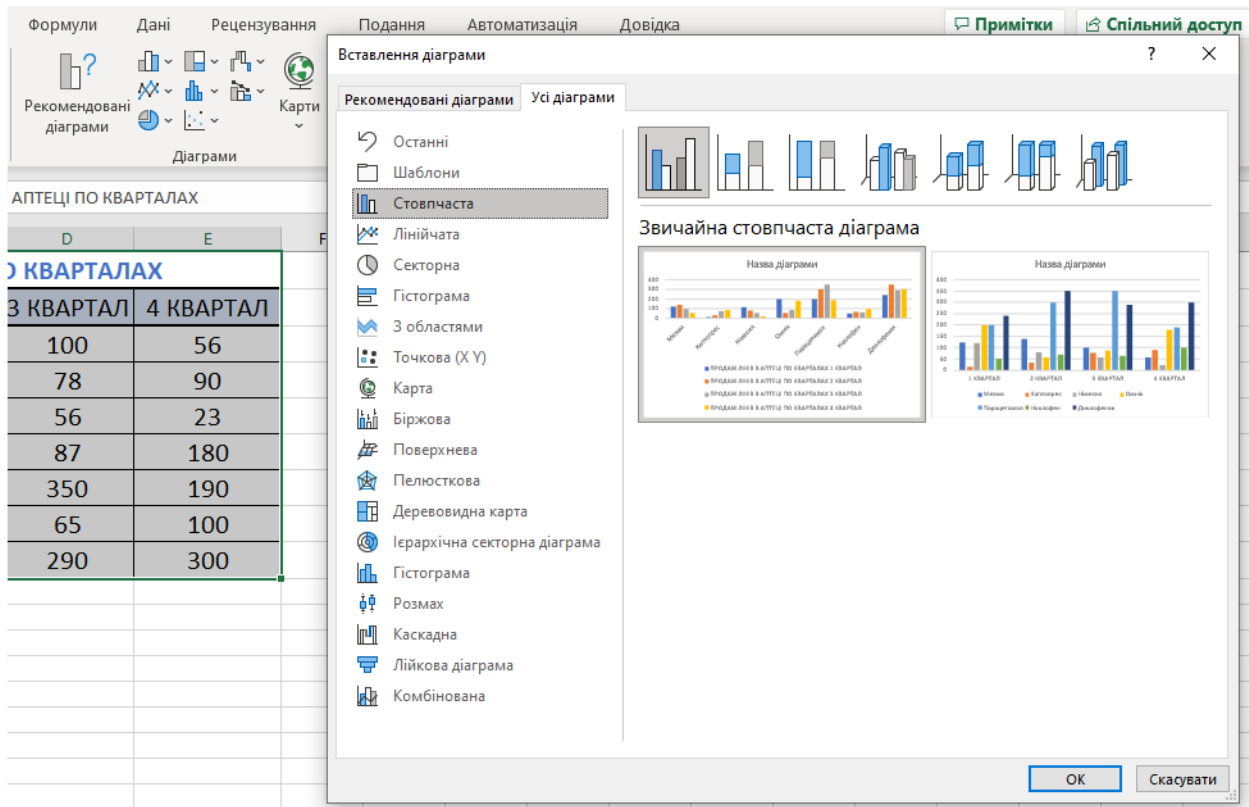


Рисунок 2.18

Вибрана діаграма з'явиться на листі Excel (рис.2.19).

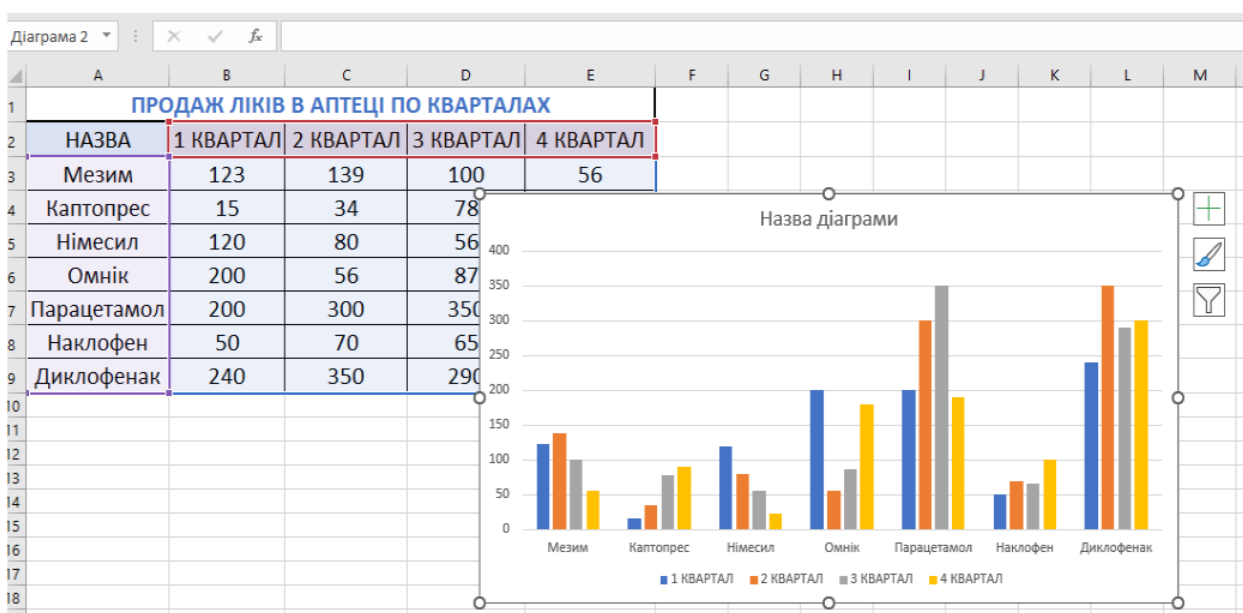


Рисунок 2.19

Якщо Ви не упевнені, який тип діаграми використовувати, команда **Інші статистичні діаграми** запропонує Вам різні варіанти на базі початкових даних.

Щоб змінити деякі параметри відображення даних, слід скористатися пунктом меню **Конструктор діаграм** (рис. 2.20).

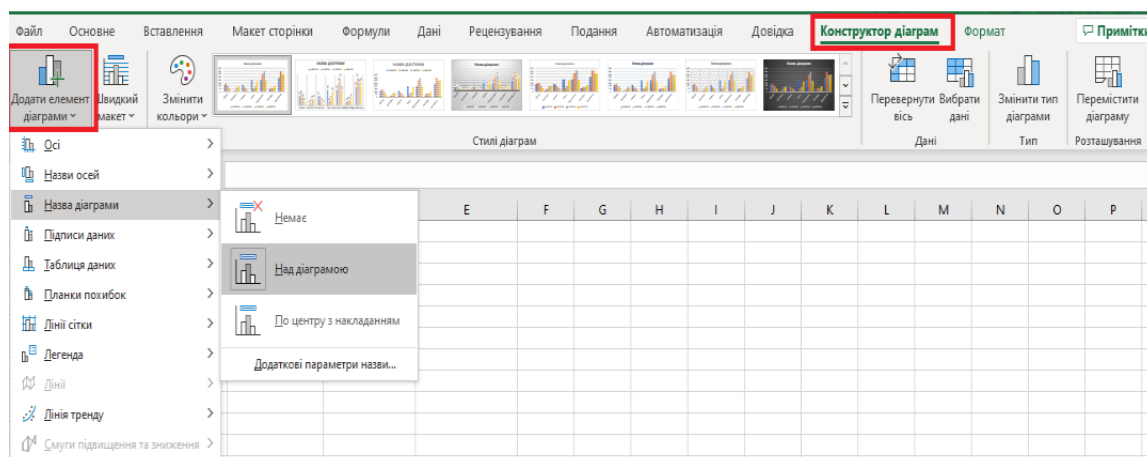


Рисунок 2.20

Цей пункт меню дає можливість змінити стиль діаграм, перевернути осі, змінити тип діаграм, перемістити діаграму та зробити інші дії.

Ви також можете скористатися кнопками швидкого доступу до форматування, щоб додати елементи на діаграму, змінити стиль або відфільтрувати дані (рисунок 2.21).

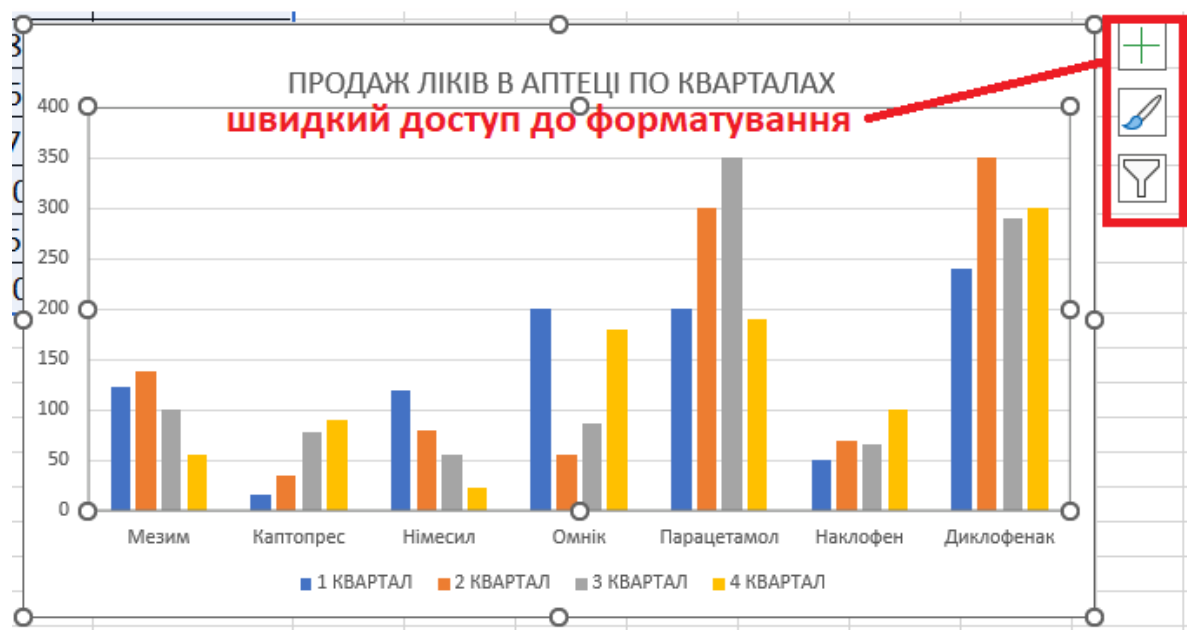


Рисунок 2.21

Практичне завдання

1. Створіть та заповніть даними таблицю (см 2.22).

Увага ! Додати до вартості ліків кількість гривень, яка дорівнює порядковому номеру в журналі. Зробіть скріншот.

Назва ліків	Категорія	Виробник	Форма випуску	Вартість, грн.	Кінцевий термін придатності
Цитрамон	знеболювальне	Дарниця	таблетки №6	13,00	27.09.2026
Ланейра	нервова система	ЗАТ Фармліга	таблетки №20	952,23	03.12.2025
Каптопрес	від високого тиску	Дарниця	таблетки №20	82,00	04.06.2027
Цитрамон	знеболювальне	АТ Лубнифарм	таблетки №10	13,90	17.07.2026
Ібупрофен	знеболювальне	Лекхим	таблетки №20	85,90	11.12.2025
Сономедин	нервова система	Дарниця	капсули №20	239,40	05.01.2027
Биспролол-КВ	від високого тиску	ЗАТ Фармліга	таблетки №30	18,30	15.08.2026
Ко-пренесса	від високого тиску	KRKA d.d.	таблетки №30	292,80	02.03.2026
Лоратадин	від алергії	Дарниця	таблетки №10	30,00	01.02.2027
Парацетамол	знеболювальне	ТОВ Здоров'я	капсули №10	24,00	02.05.2027

Рисунок 2.22

2. За допомогою фільтра відобразіть лише знеболювальні ліки. Зробіть скріншот.
3. За допомогою фільтра відобразіть тільки ліки з назвою «Цитрамон». Зробіть скріншот.
4. За допомогою фільтрів відобразіть ліки виробника "Дарниця", які коштують менше ніж 30 гривень. Зробіть скріншот.
5. За допомогою фільтрів відобразіть ліки у таблетках та капсулах по 20 штук, термін придатності яких перевищує 30 листопада 2025 року. Зробіть скріншот.
6. Створіть стовпчасту діаграму вартості ліків. Підпишіть осі. Зробіть скріншот.
7. Усі скріншоти розмістіть у файлі Word, та прикріпіть файл до завдання.

ПРАКТИЧНА РОБОТА 3

Випадкова величина. Основні характеристики випадкової величини.

Питання, які розглядаються:

1. Основні поняття.
2. Числові характеристики випадкової величини.
3. Побудова інтервального ряду випадкової величини.
4. Визначення числових характеристик засобами програми Excel.

Визначення параметрів набору значень довільної величини є важливим етапом статистичного аналізу даних.

1. Основні поняття

Подія називається *випадковою*, якщо в результаті випробування вона може відбутися, а може не відбутися. Наприклад, при покупці лотерейного квитка він може виявитися таким, що виграв або програв.

Відносною частотою події називають відношення числа випробувань m (частота), у яких подія з'явилася, до загальної кількості проведених випробувань n .

$$W = \frac{m}{n}.$$

Наприклад, підкинули монету 20 разів, при цьому “Решка” випала 15 разів. Отже, відносна частота випадання “Рішки”

$$W = \frac{15}{20} = 0,75.$$

Дуже важливо, що відносна частота має властивість *стійкості*. Тобто зі збільшенням числа випробувань її значення змінюється дедалі менше. Це значення коливається біля деякого постійного числа, яке і є ймовірністю появи події. Таким чином, відносну частоту вважатимуться *наближеним значенням ймовірності події*.

Ймовірністю $P(A)$ події A називають відношення числа сприятливих цій події наслідків m до загального числа всіх наслідків n , що утворюють повну групу рівноможливих несумісних наслідків.

$$P(A) = \frac{m}{n}.$$

Приклад 1. Монета підкидається двічі. Знайти ймовірність появи двох "Орлів".

Різними результатами є: "Орел, Орел", "Решка, Решка", "Орел, Решка", "Решка, Орел", тобто $n = 4$. Сприятливим є один результат: "Орел, Орел", отже $m = 1$. Тоді

$$P(A) = \frac{m}{n} = \frac{1}{4}.$$

Приклад 2. У ящику сім білих куль і три чорні. Випадковим чином витягується одна куля. Знайти ймовірність, що куля виявиться білою.

Усього десять куль, тобто $n = 10$. Сприятливий результат, коли витягується біла куля. Число таких результатів дорівнює кількості білих куль, тобто $m = 7$. Тоді

$$P(A) = \frac{m}{n} = \frac{7}{10} = 0,7.$$

Випадковою називається величина, яка набуває своїх значень з певною ймовірністю. Випадкові величини поділяють на дискретні та безперервні випадкові величини.

Дискретною називають випадкову величину, яка набуває окремих значень з певною ймовірністю. Наприклад, кількість хлопчиків, що народилися, серед загальної кількості новонароджених.

Безперервною називають випадкову величину, яка набуває своїх значень з деякого кінцевого або нескінченного проміжку. Наприклад, час безвідмовної роботи медичного обладнання. Правило, яке встановлює відповідність між можливими значеннями випадкової величини та його ймовірностями, називається **законом розподілу** випадкової величини.

2. Числові характеристики випадкової величини

Нехай випадкова величина набуває значення:

11, 9, 8, 9, 9, 10, 13, 13, 9, 10, 11, 9, 9, 12, 7, 12, 9, 13, 11, 10, 14.

Якщо ми впорядкуємо ці дані за зростанням, то отримаємо **варіаційний ряд**:

7, 8, 9, 9, 9, 9, 9, 9, 10, 10, 10, 11, 11, 11, 12, 12, 13, 13, 13, 14.

Його члени мають назву – **варіанти**, а закон розподілу дискретної випадкової величини задається як таблиця і називається **ряд розподілу**.

Для його побудови слід поєднати однакові значення X , вказавши їх число n (частоту). Також можна додати рядок відносної частоти W , яка є відношенням відповідної частоти до загального числа значень випадкової величини X . Відносна частота W відіграє роль ймовірності (див. табл. 2.1).

Таблиця 3.1

X	7	8	9	10	11	12	13	14
n	1	1	7	3	3	2	3	1
W	$\frac{1}{21}$	$\frac{1}{21}$	$\frac{7}{21}$	$\frac{3}{21}$	$\frac{3}{21}$	$\frac{2}{21}$	$\frac{3}{21}$	$\frac{1}{21}$

Зверніть увагу, що сума всіх відносних частот варіаційного ряду обов'язково повинна дорівнювати одиниці.

Використовуючи дані ряду розподілів у програмі **Excel**, легко побудувати **полігон частот**. Для цього переносимо значення випадкової величини X і відповідні частоти n на аркуш програми, потім виділяємо дані, вибираємо пункт меню **Вставка**, потім **Точкова діаграма з прямими лініями** (рис. 3.1).

Визначимо основні числові характеристики випадкової величини.

2.1. **Розмах ряду** - різниця між максимальним та мінімальним значеннями:

$$R = x_{\max} - x_{\min} = 14 - 7 = 7.$$

2.2 **Мода** - це значення випадкової величини, яке зустрічається найбільшу кількість разів. З визначення випливає, що значення моди дорівнює максимальному значенню частоти. Мода позначається M_0 . У нашому прикладі $M_0 = 9$.

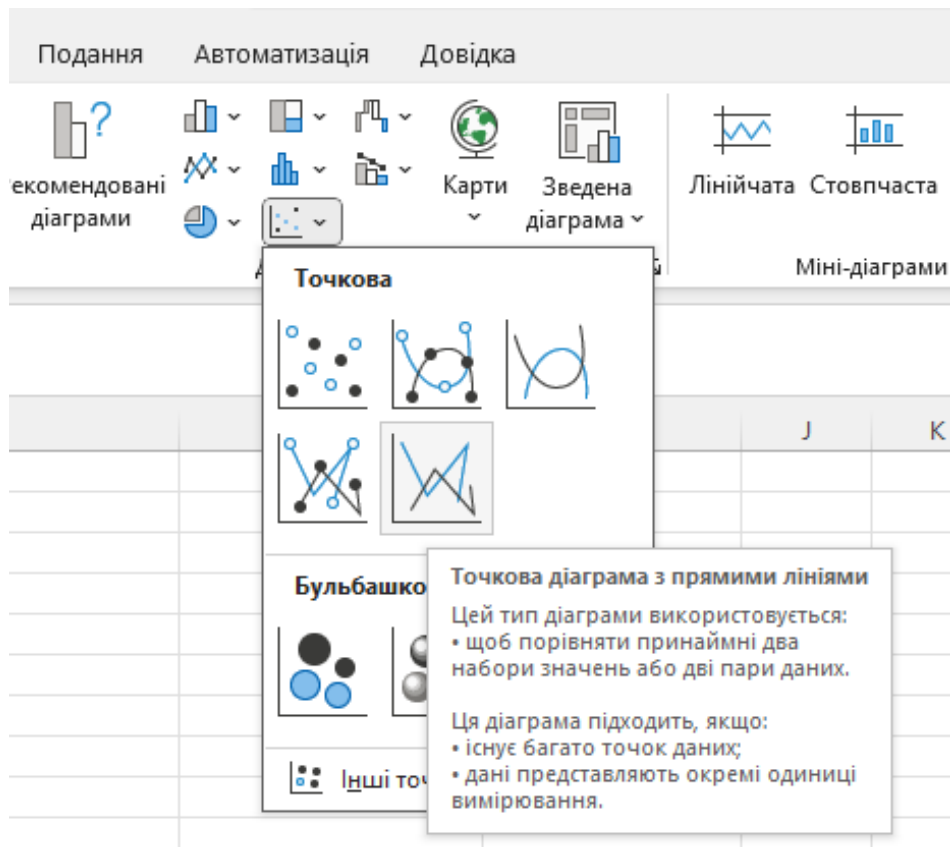


Рисунок 3.1

В результаті отримаємо полігон частот (рис.3.2).



Рисунок 3.2

2.3 Медіана – це середина, центральна варіанта, що ділить варіаційний ряд на дві рівні частини. Коли розподіл має непарну кількість спостережень, то медіана дорівнює середньому значенню; коли вона має парну кількість спостережень, то

центральне положення мають дві величини (тобто це середнє значення або півсума двох середніх варіант).

7, 8, 9, 9, 9, 9, 9, 9, 9, 10, **10**, 10, 11, 11, 11, 12, 12, 13, 13, 13, 14

У нашому прикладі це $M_e = 10$.

2.4 Вибіркове середнє - визначається як сума варіант варіаційного ряду, поділена на їх число. Якщо випадкова величина X представлена у вигляді ряду розподілу (див. табл. 2.1), вибіркове середнє розраховується за формулами

$$\bar{X} = \frac{\sum_{i=1}^k x_i n_i}{n} \text{ або } \bar{X} = \sum_{i=1}^k x_i w_i,$$

де n сума усіх частот, k їх кількість.

У нашому випадку

$$\bar{X} = \frac{\sum_{i=1}^k x_i n_i}{n} = \frac{7 \cdot 1 + 8 \cdot 1 + 9 \cdot 7 + 10 \cdot 3 + 11 \cdot 3 + 12 \cdot 2 + 13 \cdot 3 + 14 \cdot 1}{21} = \frac{218}{21} = 10,4.$$

2.5 Дисперсія – це середній квадрат відхилень значень X від вибіркового середнього. Якщо випадкова величина представлена у вигляді ряду розподілу (див. табл. 2.1), дисперсія (виправлена) розраховується за формулами

$$D = \frac{n}{n-1} \sum_{i=1}^k (x_i - \bar{X})^2 w_i, \text{ або } D = \frac{n}{n-1} (\sum_{i=1}^k x_i^2 w_i - \bar{X}^2).$$

У нашому випадку

$$D = \frac{n}{n-1} \sum_{i=1}^k (x_i - \bar{X})^2 w_i = \frac{70,96}{20} = 3,548.$$

Середньоквадратичне відхилення є коренем квадратним з дисперсії, тому воно вимірюється в тих самих одиницях, що й основні варіанти.

$$S = \sqrt{3,548} = 1,88.$$

Коефіцієнт варіації – середньоквадратичне відхилення, поділене на вибіркове середнє і виражене у відсотках.

$$V = \frac{S}{\bar{X}} = \frac{1,84}{10,4} \cdot 100\% = 17,7\%.$$

2. Побудова інтервального ряду випадкової величини

У випадку, коли кількість варіант велика, для аналізу та графічного подання вибірки замість варіаційного ряду раціонально використовувати інтервальний ряд. Для його побудови спочатку визначають кількість інтервалів k , а потім його довжину Δ за формулами:

$$k = 1 + 1,443 \ln n, \quad \Delta = \frac{R}{k},$$

де n – кількість варіант, R - розмах виборки, який дорівнює 7.

У нашому випадку

$$k = 5, \quad \Delta = 1,4.$$

Тоді інтервальний ряд має вигляд табл. 3.2:

Таблиця 3.2

Границі інтервалів	7 – 8,4	8,4 - 9,8	9,8 – 11,2	11,2 - 12,6	12,6 - 14
Частота	2	7	6	2	4

Дані інтервального ряду зручно використовувати для побудови частотної діаграми у програмі *Excel*.

Для цього переносимо таблицю 3.1 на аркуш *Excel*, виділяємо дані, вибираємо пункт меню **Вставка**, потім **Плоска стовпчаста діаграма** (рис. 3.3).

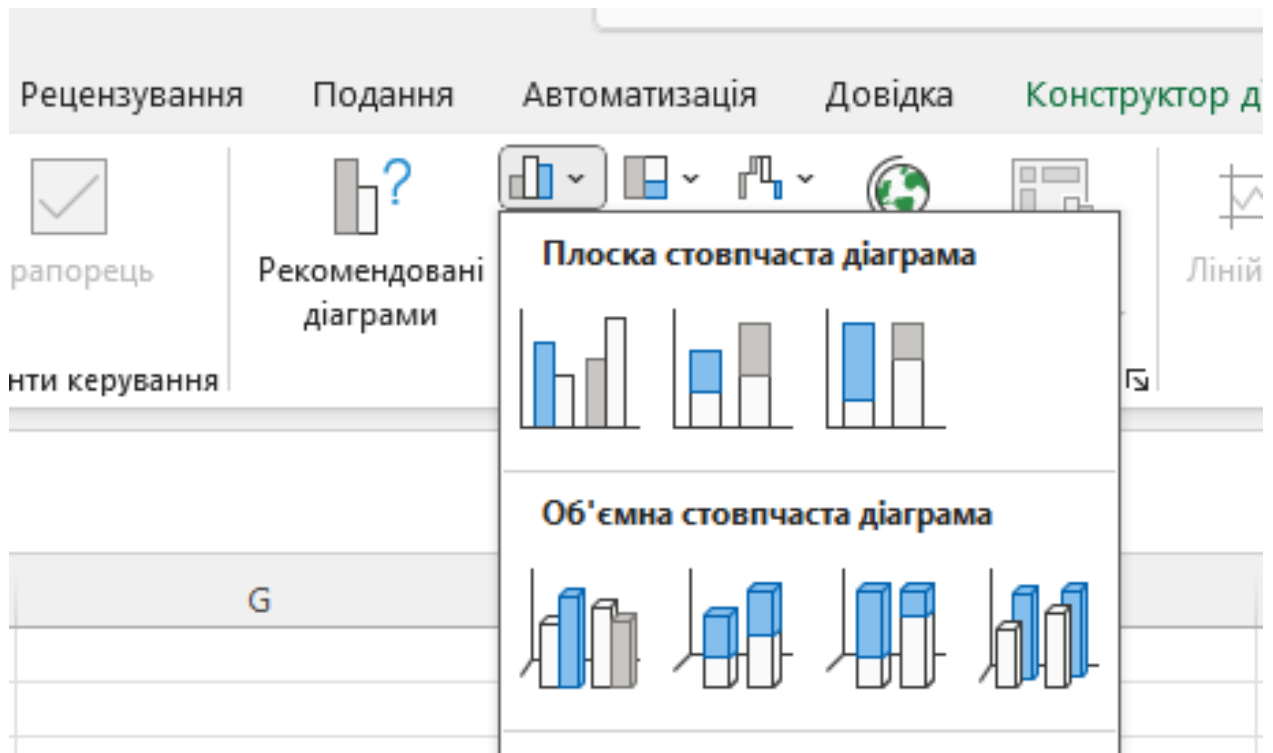


Рисунок 3.3

Отримуємо діаграму, де кожному інтервалу відповідає стовпчик, висота якого дорівнює частоті (рис.3.2).

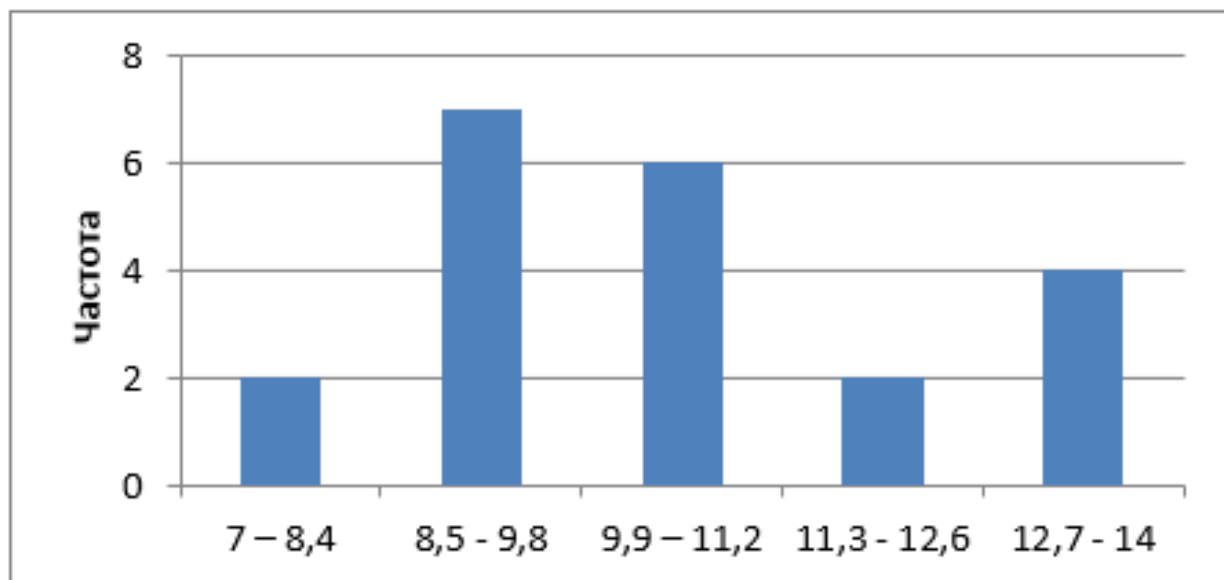
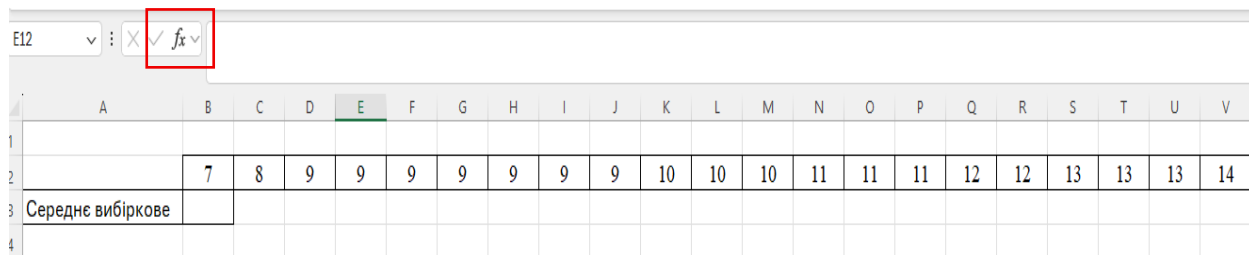


Рисунок 3.4

3. Визначення числових характеристик засобами програми Excel

4.1 Програма **Excel** має великий набір засобів для знаходження числових характеристик випадкової величини. В наступному прикладі обчислимо середнє вибіркове.

Для цього розмістимо на аркуші програми таблицю даних, виберемо клітинку **B3** і поруч з командним рядком натиснемо на значок f_x (рис. 3.5).



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1																						
2		7	8	9	9	9	9	9	9	9	10	10	10	11	11	11	12	12	13	13	13	14
3	Середнє вибіркове																					
4																						

Рисунок 3.5

Обираємо категорію **Статистичні**, потім функцію **AVERAGE**, натискаємо **OK**.

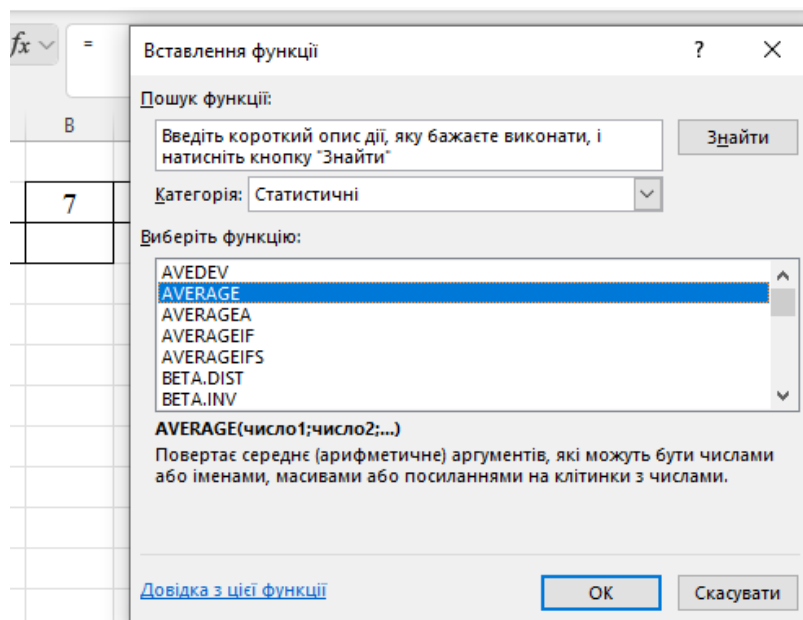


Рисунок 3.6

У вікні **Число1** вказуємо діапазон клітинок, в яких знаходяться дані та натискаємо **OK** (рис.3.7).

Зауваження. Якщо дані містять одне значення, що є найчастішим, слід використовувати функцію **MODA.SNGL**, якщо більше, ніж одне, - функцію **MODA.MULT**.

4.2 Статистичні числові характеристики можна обчислити іншим способом за допомогою пакета **Аналіз даних (Data Analysis)**. Якщо він відсутній, його слід встановити. Для цього обираємо **Файл, Параметри, Надбудови, Пакет аналізу** та натискаємо **ОК**.

Для запуску пакета обираємо **Дані**, потім **Data Analysis** (рис. 3.9).

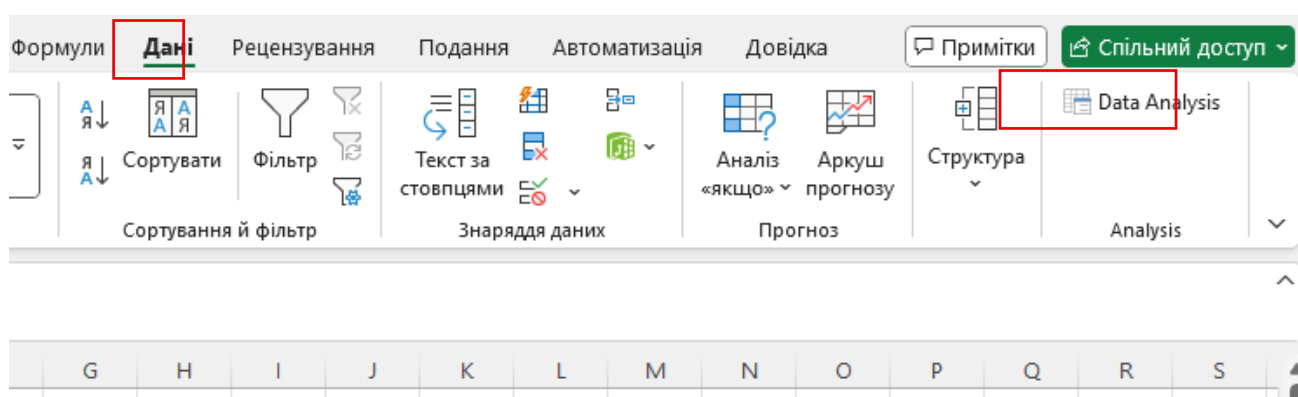


Рисунок 3.9

Відкривається вікно, в якому вибираємо **Descriptive Statistics** (рис.3.10).

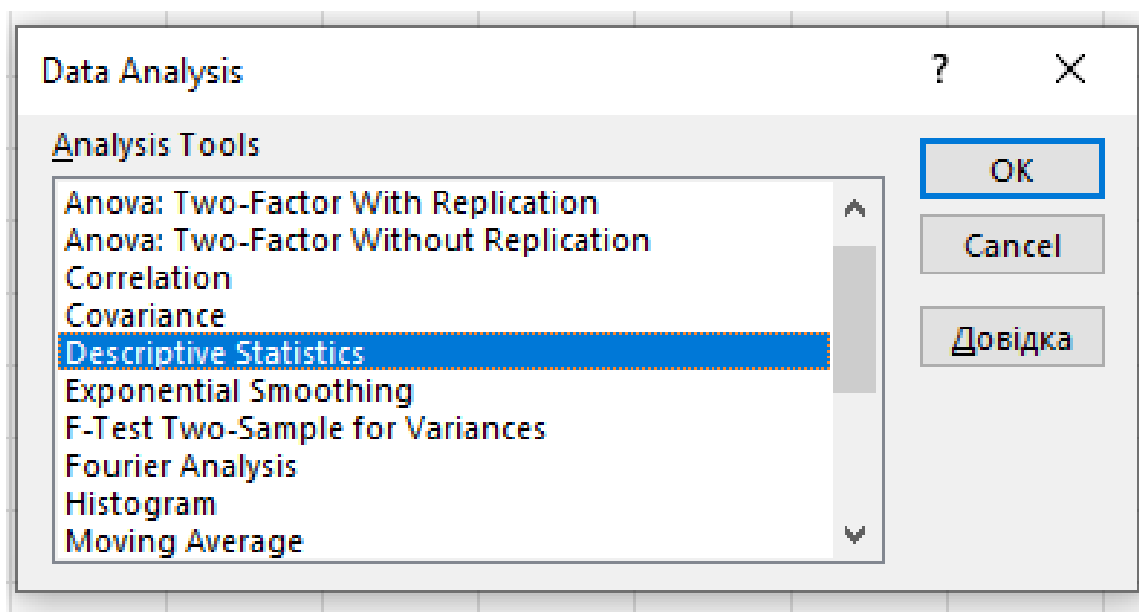


Рисунок 3.10

Після цього у вікні **Input Range** вказуємо діапазон даних, у вікні **Output Range** клітинку лівого верхнього краю вихідної таблиці та помічаємо всі пункти як показано на рис. 3.11.

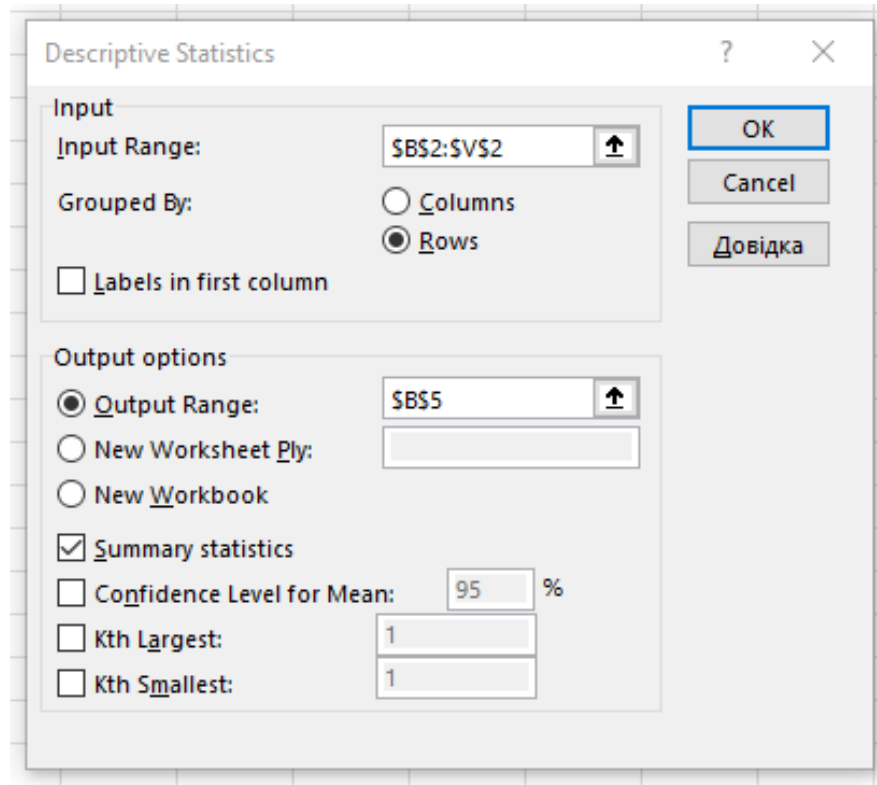


Рисунок 3.11

У результаті на аркуші з'явиться таблиця з набором обчислених числових характеристик (рис.3.12)..

Row1	
Mean	10,2
Standard Error	0,38797721
Median	10
Mode	9
Standard Deviation	1,735086832
Sample Variance	3,010526316
Kurtosis	-0,754107061
Skewness	0,263316192
Range	6
Minimum	7
Maximum	13
Sum	204
Count	20

Рисунок 3.12

Практичне завдання

Відповідно до номера свого варіанта, наведеного в табл.3.4, виконайте наступні дії:

1. Обчислити за допомогою статистичних функцій програми **Excel** середнє вибіркове, медіану, моду, дисперсію, стандартне відхилення.
2. Побудуйте ряд розподілів і на його основі полігон частот.
3. Побудуйте інтервальний ряд та на його основі діаграму частот.
4. За допомогою пакету **Аналіз даних** визначте числові характеристики випадкової величини.
5. Файл **Excel** із виконаними завданнями прикріпіть до завдання.

Таблиця 3.4

Варіант №	Дані
1	9, 13, 9, 6, 9, 9, 13, 10, 7, 10, 5, 11, 11, 11, 10, 12, 14, 11, 11, 9, 9
2	19, 19, 13, 19, 19, 9, 13, 12, 17, 15, 13, 17, 19, 14, 12, 16, 18, 15, 15, 12, 13
3	17, 14, 18, 17, 18, 17, 20, 15, 19, 22, 23, 21, 27, 20, 20, 11, 15, 21, 21, 16, 19
4	7, 6, 7, 9, 9, 14, 7, 9, 8, 9, 13, 12, 11, 10, 12, 8, 13, 10, 11, 8, 12
5	18, 12, 18, 15, 17, 6, 10, 16, 16, 20, 17, 15, 16, 10, 17, 14, 18, 16, 16, 21, 17
6	17, 11, 13, 21, 24, 15, 15, 17, 15, 17, 19, 22, 13, 23, 16, 19, 20, 15, 16, 24, 19
7	10, 10, 6, 9, 13, 12, 11, 11, 7, 14, 7, 11, 8, 8, 12, 7, 13, 8, 10, 8, 11
8	14, 16, 17, 15, 13, 15, 12, 13, 16, 15, 19, 18, 22, 17, 17, 15, 11, 19, 14, 15, 7
9	25, 18, 14, 20, 17, 15, 23, 18, 12, 20, 18, 20, 20, 15, 21, 19, 11, 16, 25, 20, 25
10	13, 12, 8, 10, 10, 15, 12, 9, 9, 7, 9, 7, 10, 14, 9, 11, 10, 6, 8, 9, 9
11	19, 16, 12, 13, 9, 18, 13, 16, 10, 14, 15, 14, 12, 23, 16, 17, 14, 10, 13, 18, 11
12	15, 16, 21, 20, 16, 16, 18, 20, 20, 11, 22, 17, 20, 17, 19, 17, 20, 13, 24, 21, 14
13	12, 8, 9, 8, 10, 9, 14, 12, 10, 9, 12, 8, 10, 11, 8, 11, 10, 12, 12, 9, 13
14	16, 14, 11, 16, 16, 13, 16, 13, 19, 19, 16, 17, 15, 15, 17, 11, 8, 14, 12, 15, 17
15	18, 20, 11, 14, 20, 13, 20, 15, 20, 19, 16, 25, 15, 25, 19, 21, 15, 17, 16, 20, 12

ПРАКТИЧНА РОБОТА 4

Функція розподілу імовірностей випадкової величини.

Деякі закони розподілу дискретної випадкової величини

Питання, які розглядаються:

1. Функція розподілу імовірностей випадкової величини.
2. Деякі закони розподілу дискретної випадкової величини.
3. Використання програми **Excel** для аналізу розподілів випадкової величини.

За допомогою функції програми **Excel** визначаються параметри деяких дискретних розподілів випадкової величини. Результати обчислень застосовуються для побудови законів розподілу.

1. Функція розподілу імовірностей випадкової величини

Функцією розподілу випадкової величини називається функція $F(x)$, що визначає ймовірність того, що випадкова величина набуде значення менше ніж x , тобто

$$F(x) = P(X < x).$$

Наприклад, випадкова величина задана наступним законом розподілу (табл. 4.1).

Таблиця 4.1

X	3	4	7	10
P	0,2	0,1	0,4	0,3

З визначення випливає, що функція розподілу має вигляд:

$$F(x) = \begin{cases} 0, & x \leq 3 \\ 0,2, & 3 < x \leq 4 \\ 0,3, & 4 < x \leq 7 \\ 0,7, & 7 < x \leq 10 \\ 1, & x > 10 \end{cases}$$

Графік функції розподілу показано на рис. 4.1.

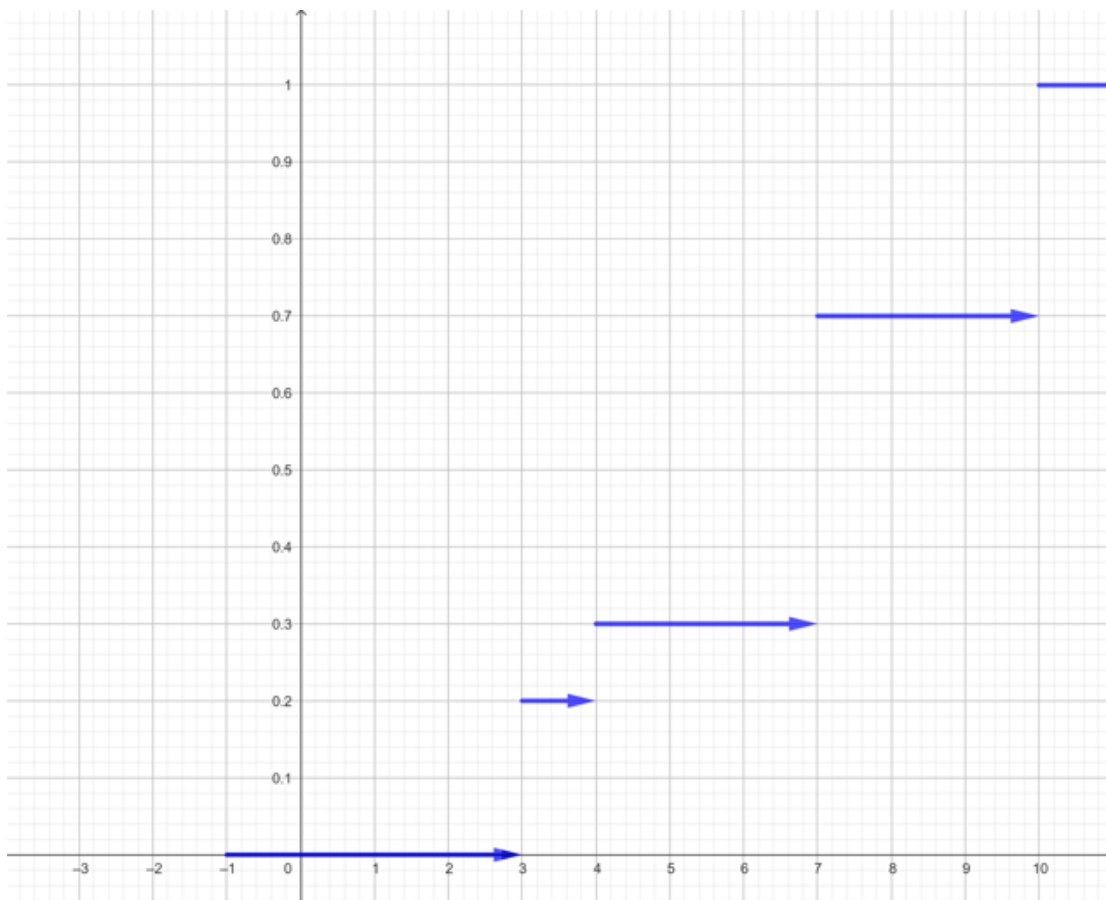


Рисунок 4.1

Функція розподілу є загальним способом завдання як дискретної, так і безперервної випадкової величини. Вона має наступні властивості:

1. $F(x) \in [0; 1]$,
2. $\lim_{x \rightarrow -\infty} F(x) = 0$,
3. $\lim_{x \rightarrow +\infty} F(x) = 1$,
4. $P(x_1 \leq X < x_2) = F(x_2) - F(x_1)$.

2. Деякі закони розподілу дискретної випадкової величини

2.1 Рівномірний розподіл

Рівномірний розподіл - розподіл, для якого ймовірність кожного із значень випадкової величини одна і та ж, тобто

$$P(X = x) = \frac{1}{n},$$

де n - кількість можливих значень випадкової величини.

Приклад. При киданні гральної кістки випадає цифра від одного до шести. Нехай значення випадкової величини - цифра, що випадає під час кидання гральної кістки. Імовірність кожної події дорівнює $\frac{1}{6}$ (табл. 4.2).

Таблиця 4.2

X	1	2	3	4	5	6
P	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

2.2 Біноміальний розподіл

Незалежними називаються випробування, у яких ймовірність події кожного випробування не залежить від результатів інших випробувань.

Нехай проводяться n незалежних випробувань, у кожному з яких подія настає з ймовірністю p . Тоді ймовірність того, що при n незалежних випробуваннях ця подія настане k раз, визначається формулою Бернуллі:

$$P_n(k) = C_n^k p^k (1 - p)^{n-k},$$

де $C_n^k = \frac{n!}{k!(n-k)!}$.

Якщо прийняти, що випадкова величина набуває всі можливі значення настання події, тобто $X = k$, то розподіл ймовірностей за формулою Бернуллі називається **біноміальним розподілом**. **Математичне сподівання** цього розподілу дорівнює np , а **дисперсія** - $np(1 - p)$.

Приклад. Нехай народилося п'ять дітей. Нехай випадкова величина - число новонароджених хлопчиків. Обчислимо ймовірності за формулою Бернуллі при

значеннях випадкової величини $k=0,1,2,3,4,5$, вважаючи, що ймовірність народження хлопчика дорівнює 0,51.

$$\begin{aligned} P_5(0) &= C_5^0 0,51^0 (1 - 0,51)^5 = 0,028, & P_5(1) &= C_5^1 0,51^1 (1 - 0,51)^4 = 0,147, \\ P_5(2) &= C_5^2 0,51^2 (1 - 0,51)^3 = 0,306, & P_5(3) &= C_5^3 0,51^3 (1 - 0,51)^2 = 0,318, \\ P_5(4) &= C_5^4 0,51^4 (1 - 0,51)^1 = 0,166, & P_5(5) &= C_5^5 0,51^5 (1 - 0,51)^0 = 0,035. \end{aligned}$$

Тоді закон розподілу має наступний вигляд (табл. 4.3).

Таблиця 4.3

X	0	1	2	3	4	5
P	0,028	0,141	0,306	0,318	0,166	0,035

Графічно закон розподілу можна зобразити, побудувавши точки (x_i, p_i) , та з'єднавши їх. Отримана фігура має назву **багатокутник розподілу** (рис.4.2).

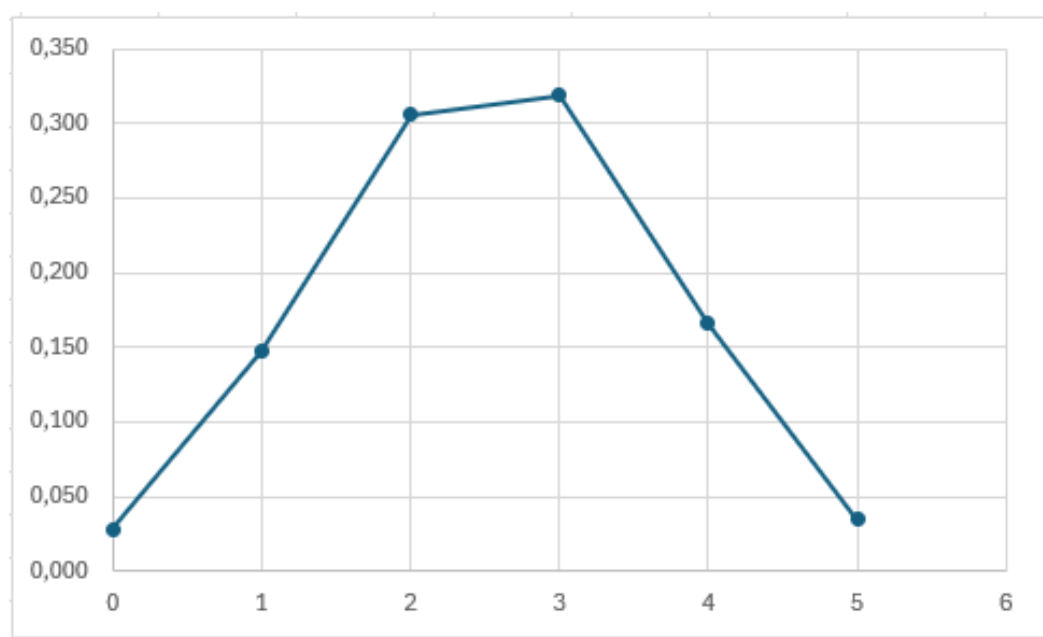


Рисунок 4.2

Математичне сподівання цього розподілу дорівнює

$$np = 5 \cdot 0,51 = 2,55,$$

а дисперсія -

$$np(1 - p) = 5 \cdot 0,51 \cdot 0,49 = 1,2495.$$

Зверніть увагу, що найбільше значення ймовірність приймає при значенні випадкової величини $k_0 = x_4 = 3$. Таке значення називають **найімовірнішим числом випробувань**. Його можна обчислити за такою формулою:

$$np - q \leq k_0 \leq np + p,$$

де $q = 1 - p$.

У нашому випадку маємо

$$5 \cdot 0,51 - 0,49 \leq k_0 \leq 5 \cdot 0,51 + 0,51p,$$

$$2,06 \leq k_0 \leq 3,06,$$

тобто $k_0 = 3$, таким чином ми отримали теж саме значення.

2.3 Розподіл Пуассона

У разі, коли p або q мало, при зростанні числа випробувань біноміальний розподіл прямує до **розподілу Пуассона**, який задається формулою.

$$P_k = \frac{\lambda^k e^{-\lambda}}{k!},$$

де $\lambda = np$ - інтенсивність (середня кількість подій за фіксований проміжок часу).

Математичне сподівання та **дисперсія розподілу Пуассона** дорівнюють $np = \lambda$. Таким чином, **розподіл Пуассон** є асимптотичним наближенням біноміального розподілу.

Приклад. Імовірність поставки в аптеку ліків, термін придатності яких закінчився, дорівнює 0,02. Ліки поставляються в упаковках по 100 штук в кожній. Знайти ймовірність того, що в упаковці кількість ліків, термін придатності яких закінчився, не перевищує двох.

З умови випливає: $n = 100$; $p = 0,02$; $k = 0,1,2$. Отже інтенсивність $\lambda = np = 100 \cdot 0,02 = 2$. Тоді за формулою Пуассона

$$P_0 = \frac{2^0 e^{-2}}{0!} \approx 0,135, \quad P_1 = \frac{2^1 e^{-2}}{1!} \approx 0,271, \quad P_2 = \frac{2^2 e^{-2}}{2!} \approx 0,271.$$

Оскільки потрібно визначити ймовірність здійснення однієї з трьох подій, ймовірності підсумовуємо і отримуємо:

$$P_{k \leq 2} = P_0 + P_1 + P_2 = 0,135 + 0,271 + 0,271 = 0,677,$$

математичне сподівання та дисперсія дорівнюють $\lambda = 2$.

Розподіл Пуассона широко використовується при моделюванні систем масового обслуговування, у тому числі в медичних установах.

2.4 Геометричний розподіл

Геометричний розподіл описує ймовірність настання сприятливого результату випробування після того, як перші $(k-1)$ результатів були несприятливі. Ймовірність такої події визначається за формулою:

$$P(X = k) = p(1 - p)^{k-1}.$$

Математичне сподівання цього розподілу дорівнює $\frac{1}{p}$, а **дисперсія** - $\frac{1-p}{p^2}$.

Приклад. У відділенні проводять поодинокі лабораторні тести на наявність патогену у матеріалі пацієнтів. Вважаємо, що кожен тест незалежний, і ймовірність того, що тест виявиться позитивним (тобто знайдено інфікований зразок) дорівнює $p = 0,03$. Знайти ймовірність того, що перший позитивний тест виявиться на 10-му досліді.

З умови завдання: $p = 0,03, k = 10$. Тоді за формулою ймовірності геометричного розподілу

$$P(X = 10) = 0,03(1 - 0,03)^9 \approx 0,023.$$

Математичне сподівання цього розподілу дорівнює $\frac{1}{p} = \frac{1}{0,03} \approx 33,3$, а дисперсія -

$$\frac{1-p}{p^2} = \frac{1-0,03}{0,03^2} \approx 32,84.$$

2.5 Гіпергеометричний розподіл

Нехай у скінченній популяції обсягу N є рівно K «успіхів» (об'єктів, що мають потрібну властивість) і $(N - K)$ «неуспіхів». Із популяції випадково вибирають n об'єктів без повернення (тобто вибірка без заміни). Позначимо випадкову величину X — кількість успіхів у вибірці розміру n .

Тоді X має *гіпергеометричний розподіл* з параметрами (N, K, n) . Його функція ймовірності обчислюється таким чином:

$$P(X = x) = \frac{C_K^x C_{N-K}^{n-x}}{C_N^n},$$

при цьому випадкова величина може приймати значення

$$X = 0, 1, 2, \dots, \min(K, n).$$

Математичне сподівання гіпергеометричного розподілу дорівнює

$$M(X) = n \cdot \frac{K}{N}.$$

Дисперсія гіпергеометричного розподілу дорівнює

$$D(X) = n \cdot \frac{K}{N} \cdot \left(1 - \frac{K}{N}\right) \cdot \frac{N - n}{N - 1}.$$

Приклад. У невеликому стаціонарі обстежено $N = 50$ пацієнтів — відомо, що серед них $K = 10$ інфіковані (позитивні випадки). Випадковим чином *без повернення* (тобто вибірка без заміни) вибрали $n = 5$ пацієнтів для експрес-тестування. Нехай X — кількість виявлених позитивних тестів у цій вибірці. Знайти ймовірність двох позитивних результатів у вибірці з 5 осіб.

За формулою гіпергеометричного розподілу обчислимо функцію ймовірності:

$$P(X = 2) = \frac{C_{10}^2 C_{40}^3}{C_{50}^5} = \frac{45 \cdot 9880}{2118760} \approx 0,21,$$

Математичне сподівання цього розподілу дорівнює

$$M(X) = n \cdot \frac{K}{N} = 5 \cdot \frac{10}{50} = 1.$$

Дисперсія цього розподілу дорівнює

$$D(X) = n \cdot \frac{K}{N} \cdot \left(1 - \frac{K}{N}\right) \cdot \frac{N - n}{N - 1} = 5 \cdot 0,2 \cdot 0,8 \cdot \frac{45}{49} \approx 0,735.$$

3. Використання програми Excel для аналізу розподілів випадкової величини

3.1 Біномінальний розподіл

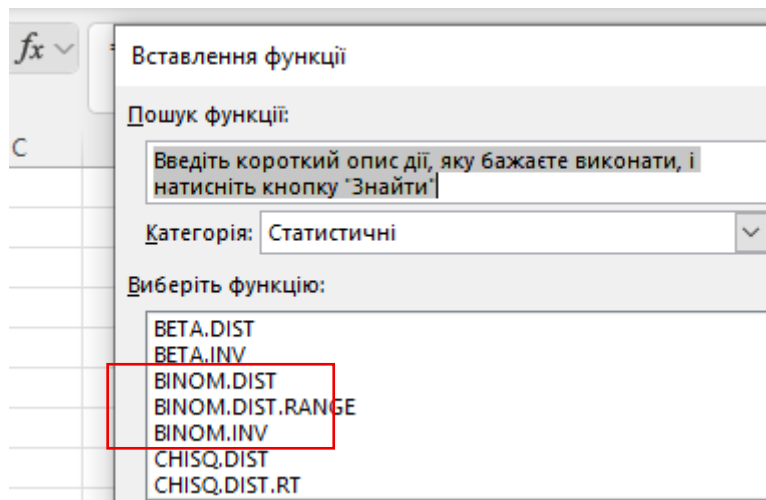


Рисунок 4.3

Для визначення параметрів біномінального розподілу у категорії Статистичні є три основні функції: **BINOM.DIST**, **BINOM.DIST.RANGE**, **BINOM.INV** (рис.4.3).

Функція **BINOM.DIST** має наступний формат (рис. 4.3):

- **кількість_успіхів** - це кількість випробувань зі сприятливим наслідком;
- **кількість_випробувань** - загальна кількість випробувань;
- **імовірність успіху** - ймовірність сприятливого наслідку в одному випробуванні;
- **функція** - логічна функція: **0(FALSE)** або **1(TRUE)**. Якщо **0**, функція повертає ймовірність того, що кількість сприятливих наслідків точно дорівнює значенню *кількість_успіхів*. Якщо **1**, функція повертає ймовірність того, що

кількість сприятливих наслідків буде не меншою ніж значення *кількість_успіхів* (рис.4.4).

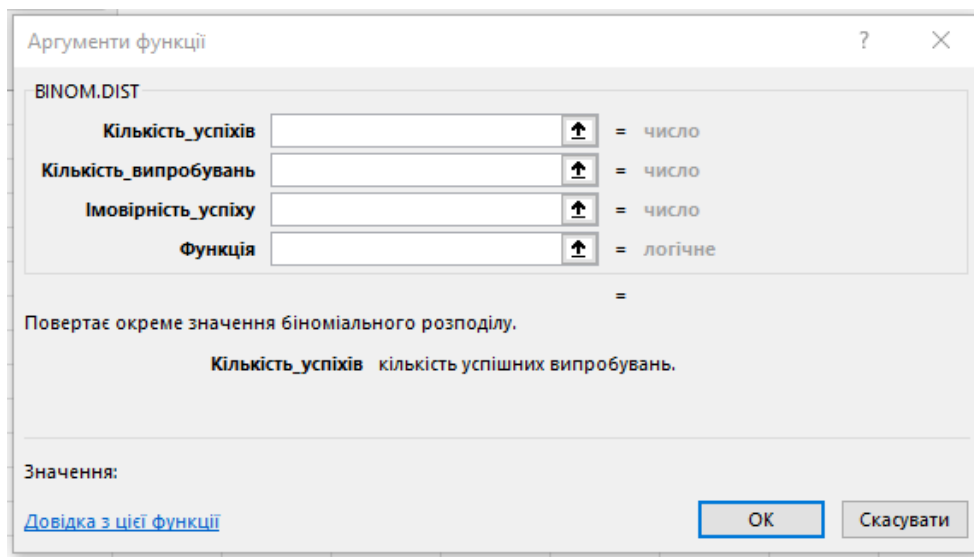


Рисунок 4.4

Так, щоб побудувати закон розподілу для прикладу з п.2.2, потрібно сформулювати таблицю, вказавши в першому рядку всі можливі значення випадкової величини. Потім вибрати функцію **BINOM.DIST** та ввести значення як показано на рис. 4.5.

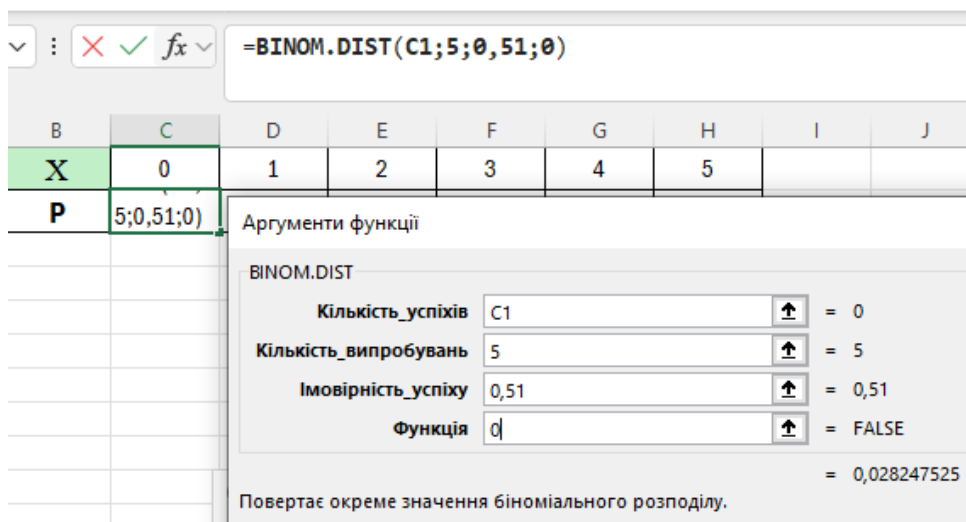


Рисунок 4.5

Після цього натиснути **ОК** і розтягнути формулу для обчислення ймовірностей інших значень випадкової величини.

Функція **BINOM.DIST.RANGE** використовується для обчислення ймовірності біномного розподілу в заданому діапазоні. Так, якщо нам необхідно знайти ймовірність, що з п'яти новонароджених народиться або три, або чотири, або п'ять хлопчиків, вводимо дані так, як показано на рис. 4.6.

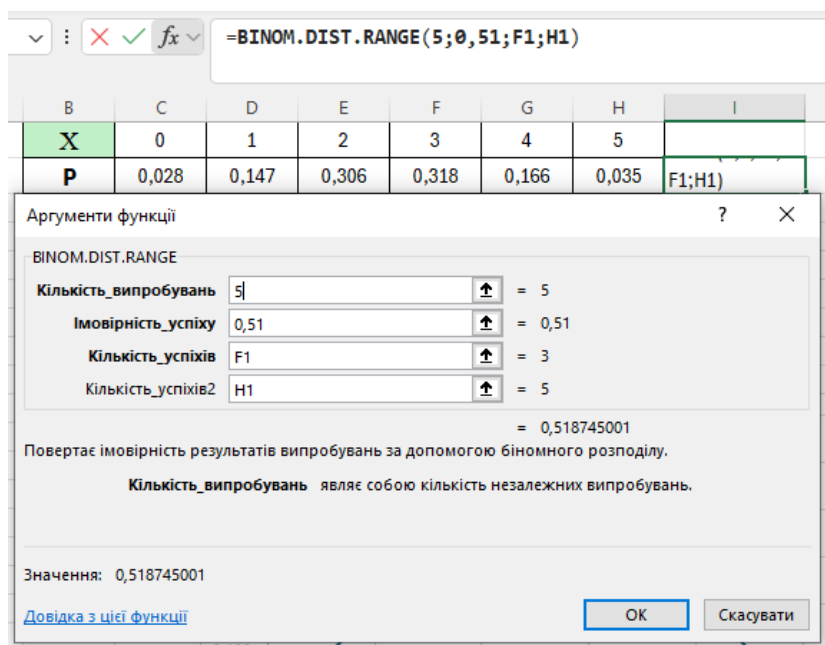


Рисунок 4.6

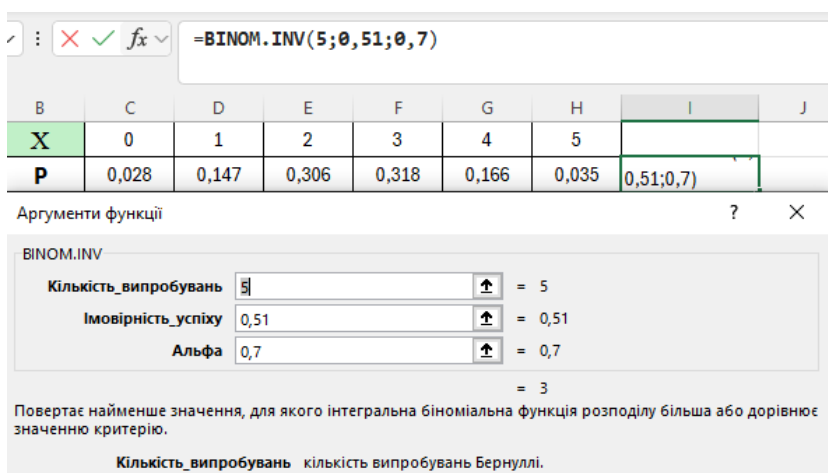


Рисунок 4.7

Функція **BINOM.INV** повертає найменше значення, при якому ймовірність інтегрального біномного розподілу дорівнює або перевищує задане значення. Так, якщо нам необхідно знайти мінімальну кількість новонароджених хлопчиків, при

якій ймовірність появи цієї кількості буде не менше **0,7**, вводимо дані, як показано на рис. 4.7. З результатів розрахунку бачимо, що така кількість дорівнює трьом.

3.2 Розподіл Пуассона

Для визначення параметрів біноміального розподілу у категорії **Статистичні** використовується функція **POISSON.DIST** (рис.4.8).

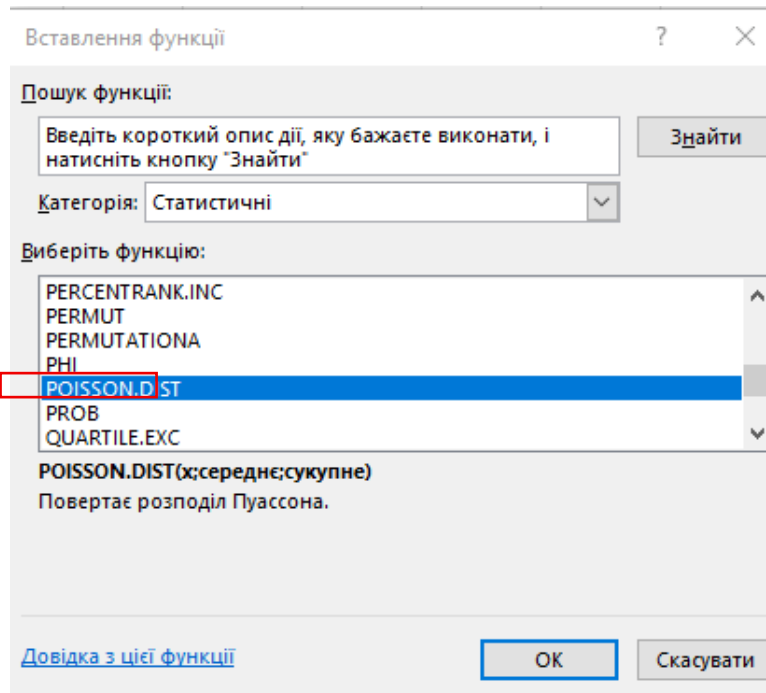


Рисунок 4.8

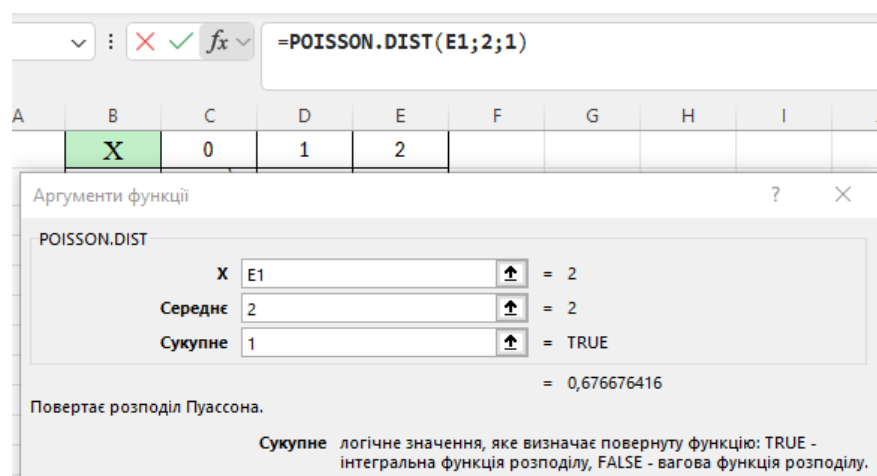


Рисунок 4.9

Для знаходження ймовірності з прикладу з п.2.3 слід внести дані так, як показано рис. 4.9. Для параметра **Сукупне** ми використовували логічне значення

1(TRUE), щоб знайти ймовірність того, що кількість подій буде від нуля до кількості, яка визначається параметром **X** (це клітинка **E2**, зі значенням **2**).

3.3 Гіпергеометричний розподіл

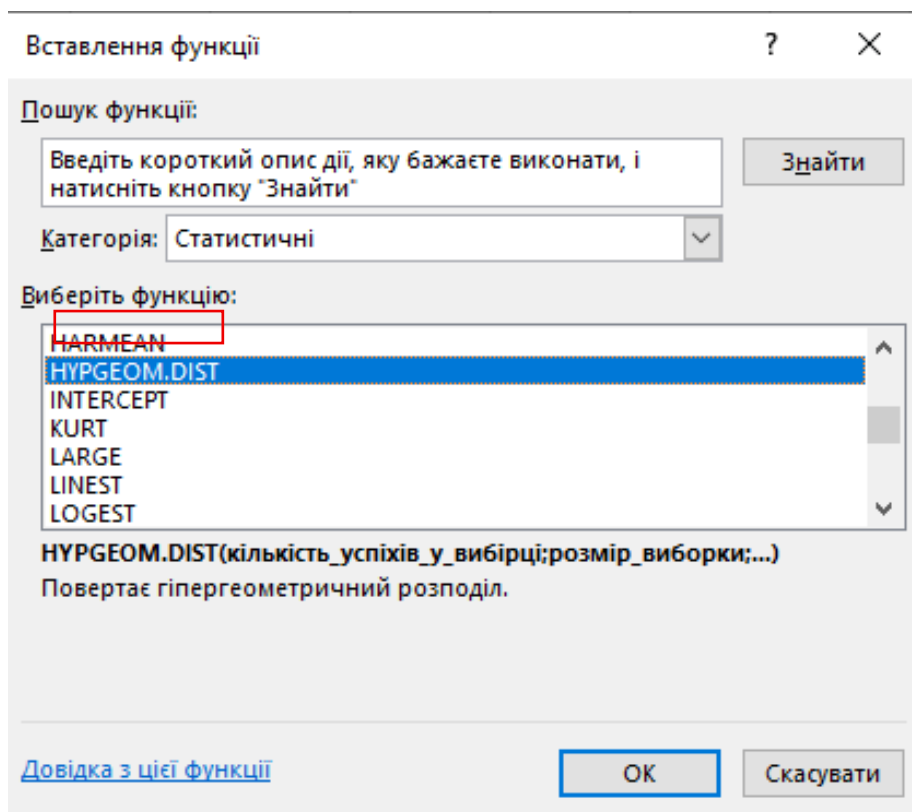


Рисунок 4.10

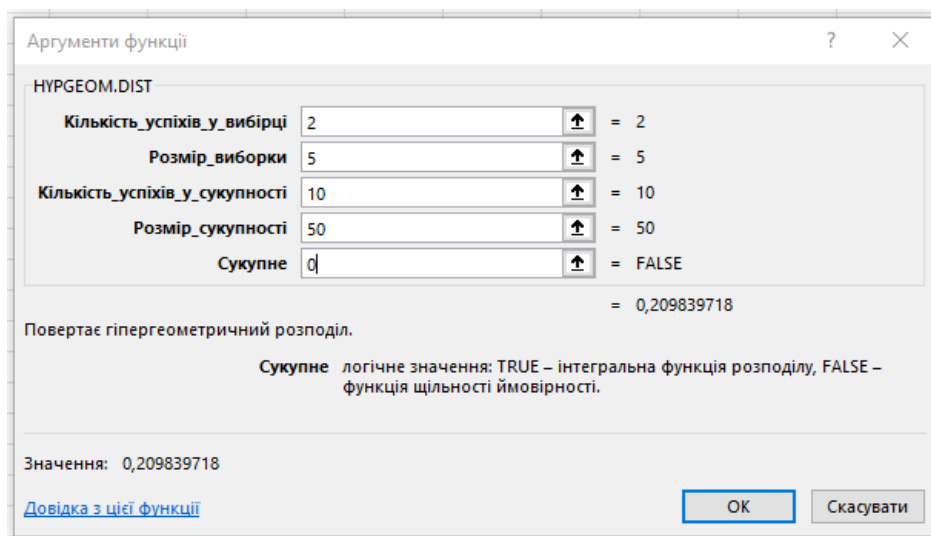


Рисунок 4.11

Для обчислення ймовірності випадкових величин при гіпергеометричному розподілі у категорії **Статистичні** використовується функція **HYPGEOM.DIST** (рис.4.10).

Для обчислення ймовірності з прикладу з п.2.5 слід внести дані так, як показано рис. 4.11.

Результат збігається з одержаним у п.2.5 і становить

$$P(X = 2) = \frac{C_{10}^2 C_{40}^3}{C_{50}^5} = \frac{45 \cdot 9880}{2118760} \approx 0,21.$$

Практичне завдання

Увага! Завдання виконуються у програмі Excel.

1. У медичній лабораторії відомо, що ймовірність правильного визначення позитивного результату експрес-тесту становить $p = 0,9$. У серії з $n = 20 + k$ (k – номер варіанту) незалежних тестів випадкова величина X — кількість правильних результатів. Потрібно побудувати закон розподілу, багатокутник розподілу, знайти найімовірніше число.

2. У лікарні середньодобово у відділення інтенсивної терапії надходить

$$\lambda = k + 1 \text{ (} k \text{ – номер варіанту)}$$

викликів «швидкої допомоги». Нехай випадкова величина X — це кількість викликів за одну добу. Потрібно побудувати закон розподілу кількості викликів від **0** до **10**, та багатокутник розподілу.

3. У цеху фармацевтичного підприємства перевіряють якість виробництва таблеток. В одній партії $N = 50 + k$ (k – номер варіанту) таблеток. Відомо, що $K = 20$ таблеток мають відхилення у вмісті діючої речовини (тобто «непридатні»). Випадково вибирають $n = 10$ таблеток для контролю **без повернення** (тобто вибірка без заміни). Нехай X - число непридатних таблеток у вибірці. Потрібно побудувати закон розподілу непридатних таблеток у вибірці від **0** до **10**, та багатокутник розподілу.

4. Файл Excel із виконаними завданнями прикріпіть до завдання.

ПРАКТИЧНА РОБОТА 5

Щільність розподілу ймовірностей випадкової величини.

Нормальний закон розподілу випадкової величини.

Питання, які розглядаються:

1. Щільність розподілу ймовірностей випадкової величини.
2. Нормальний закон розподілу випадкової величини.
3. Перевірка даних на відповідність нормальному розподілу.

За допомогою функції програми **Excel** проводиться перевірка даних на відповідність нормальному розподілу.

1. Щільність розподілу ймовірностей випадкової величини

Для безперервної випадкової величини вводиться поняття *щільності розподілу* ймовірностей випадкової величини. *Щільністю розподілу* безперервної випадкової величини називається функція, яка є похідною її функції розподілу, тобто

$$f(x) = F'(x).$$

Властивості щільності розподілу є наступними:

1. $f(x) \geq 0$, тому що $F(x)$ неспадюча функція;
2. $F(x) = \int_{-\infty}^x f(t)dt$;
3. $P(x_1 \leq X < x_2) = \int_{x_1}^{x_2} f(t)dt = F(x_2) - F(x_1)$, за формулою Ньютона-Лейбніца;
4. $\int_{-\infty}^{+\infty} f(t)dt = 1$.

2. Нормальний закон розподілу випадкової величини

Випадкова величина розподілена за *нормальним законом*, якщо її щільність розподілу має вигляд

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}},$$

де параметр a - математичне очікування (середнє значення), а також медіана та мода розподілу, параметр σ - середньоквадратичне відхилення.

Щільності розподілу відповідає функція розподілу. Вона має вигляд:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(t-a)^2}{2\sigma^2}} dt.$$

Графіки щільності розподілу та функції розподілу показані на рис. 5.1.

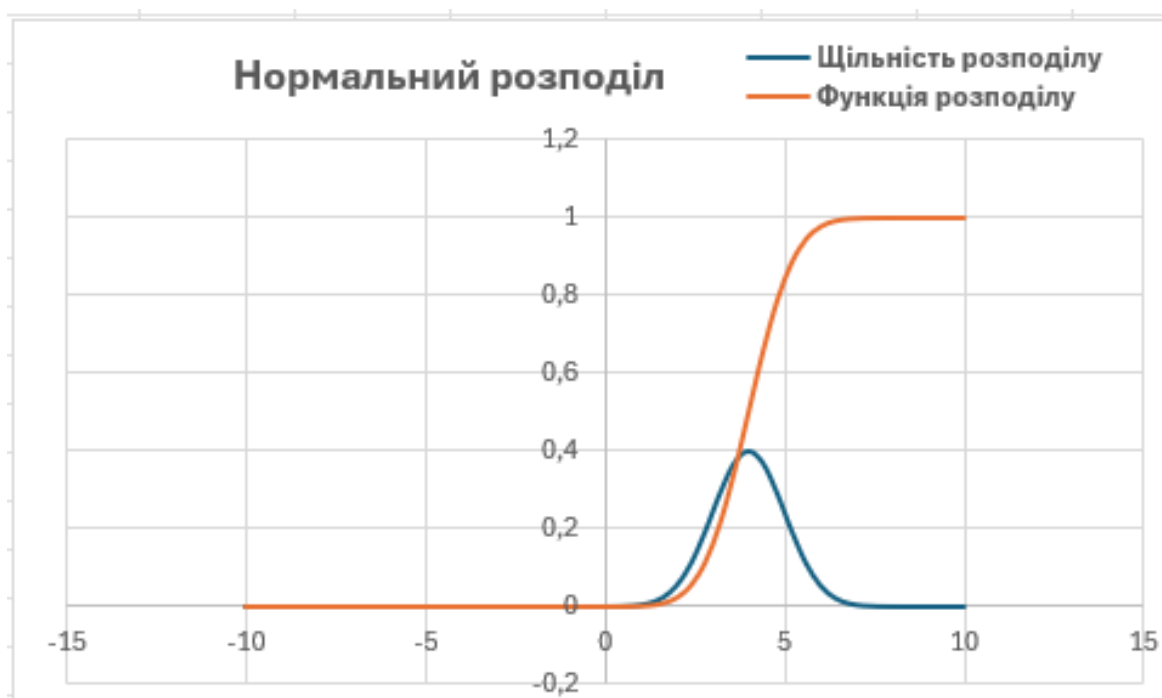


Рисунок 5.1

Нормальний розподіл за значеннями $a = 0$, $\sigma = 1$ називається **стандартним нормальним розподілом**. Щільність такого розподілу є **функцією Гауса**

$$\varphi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

а сама функція розподілу – *функцією Лапласа* (рис. 5.2)

$$\Phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt.$$

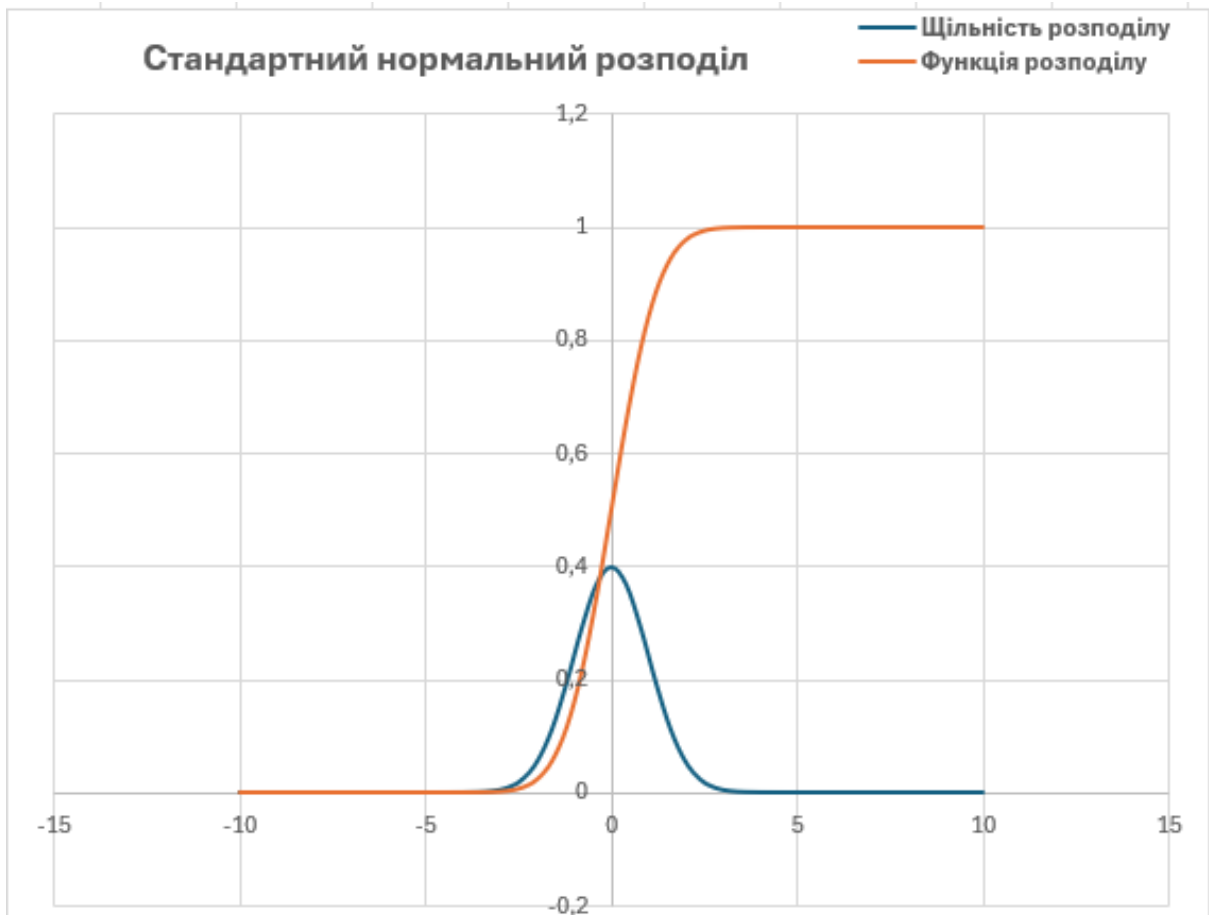


Рисунок 5.2

Імовірність знаходження випадкової величини, розподіленої за нормальним законом, в інтервалі (α, β) чисельно дорівнює площі криволінійної трапеції, обмеженої зверху графіком щільності розподілу, знизу - віссю абсцис, ліворуч і праворуч - вертикальними прямими $x = \alpha$, $x = \beta$ відповідно.

У зв'язку з цим така імовірність визначається формулою:

$$P(\alpha \leq X < \beta) = F(\beta) - F(\alpha),$$

або через функцію Лапласа, значення якої добре відомі

$$P(\alpha \leq X < \beta) = \Phi\left(\frac{\beta - a}{\sigma}\right) - \Phi\left(\frac{\alpha - a}{\sigma}\right).$$

Скористаємося цією формулою та обчислимо:

$$P(a - \sigma \leq X < a + \sigma) = \Phi(1) - \Phi(-1) = 2\Phi(1) = 0,683,$$

$$P(a - 2\sigma \leq X < a + 2\sigma) = \Phi(2) - \Phi(-2) = 2\Phi(2) = 0,954,$$

$$P(a - 3\sigma \leq X < a + 3\sigma) = \Phi(3) - \Phi(-3) = 2\Phi(3) = 0,997.$$

Отриманий результат дозволяє зробити висновок, що при нормальному законі розподілу в інтервалі $(a - \sigma, a + \sigma)$ міститься 68,3% значень випадкових величин, в інтервалі $(a - 2\sigma, a + 2\sigma)$ - 95,4%, в інтервалі $(a - 3\sigma, a + 3\sigma)$ - 99,7%. Цю властивість нормального розподілу називають **правилом трьох сигм**.

3. Перевірка даних на відповідність нормальному розподілу

При статистичному аналізі вибірки випадкових величин важливо знати, чи є їх розподіл нормальним. У разі нормального розподілу застосовують параметричні критерії аналізу даних. В Excel є п'ять функцій для розрахунків, пов'язаних із нормальним розподілом.

Функція **NORM.DIST(*x; середнє; стандартне відхилення; сукупне*)** повертає значення функції щільності розподілу (при *сукупне* = 0) або функції розподілу (при *сукупне* = 1) залежно від значення *x*.

Функція **NORM.INV(*імовірність; середнє; стандартне відхилення*)** - обернена функція, повертає обернене значення нормального інтегрального розподілу для вказаного середнього та стандартного відхилення. Функції **NORM.S.DIST(*z; сукупне*)** та **NORM.S.INV (*імовірність*)** - функції, аналогічні першим двом, для стандартного нормального розподілу при $a = 0$, $\sigma = 1$.

Функція **STANDARDIZE(*x; середнє; стандартне відхилення*)** повертає

нормалізоване значення за формулою:

$$\bar{x} = \frac{x - a}{\sigma}.$$

Розглянемо різноманітні способи перевірки розподілу на нормальність. Нехай задано 30 значень випадкової величини (табл. 5.1).

Таблиця 5.1

1,33	0,79	1,47	1,21	0,23	0,35	0,53	0,29	0,52	1,04	0,52	0,35	0,41	0,88	1,89
1,01	1,81	1,26	0,01	0,90	1,58	1,65	0,68	0,30	0,02	0,45	0,68	1,87	0,06	0,46

3.1 Перевірка виконання правила трьох сигм

За допомогою пакету **Аналіз даних** (див. практичну роботу 3, п.4.2) обчислюємо числові характеристики вибірки.

Column1	
Mean	-0,02
Standard Error	0,18
Median	0,14
Standard Deviation	1,01
Sample Variance	1,02
Kurtosis	-0,66
Skewness	-0,21
Range	3,75
Minimum	-1,89
Maximum	1,87
Sum	-0,60
Count	30

Рисунок 5.3

З рис. 5.3 визначаємо $a = \text{Mean} = -0.02$, $\sigma = \text{Standart Deviation} = 1,01$.

Тоді в **Excel** нескладно побудувати наступну таблицю 5.2. Зазначимо, що при знаходженні кількості значень у кожному інтервалі використовувалася функція **COUNTIFS**.

Таблиця 5.2

Інтервал	від	до	Кількість	%	
$(a - \sigma, a + \sigma)$	-1,03	0,99	19	63,3	>62,5%
$(a - 2\sigma, a + 2\sigma)$	-2,04	2,00	30	100	>68,3%
$(a - 3\sigma, a + 3\sigma)$	-3,05	3,01	30	100	>99,7%

3.2 Порівняння діаграми частот розподілу та теоретичної нормальної кривої

На аркуші **Excel** розміщуємо стовпець із даними. За допомогою пакету **Аналіз даних** визначаємо спочатку числові характеристики, потім будуємо гістограму (рис.5.4). На рис. 5.5 показано, які дані слід внести для побудови гістограми.

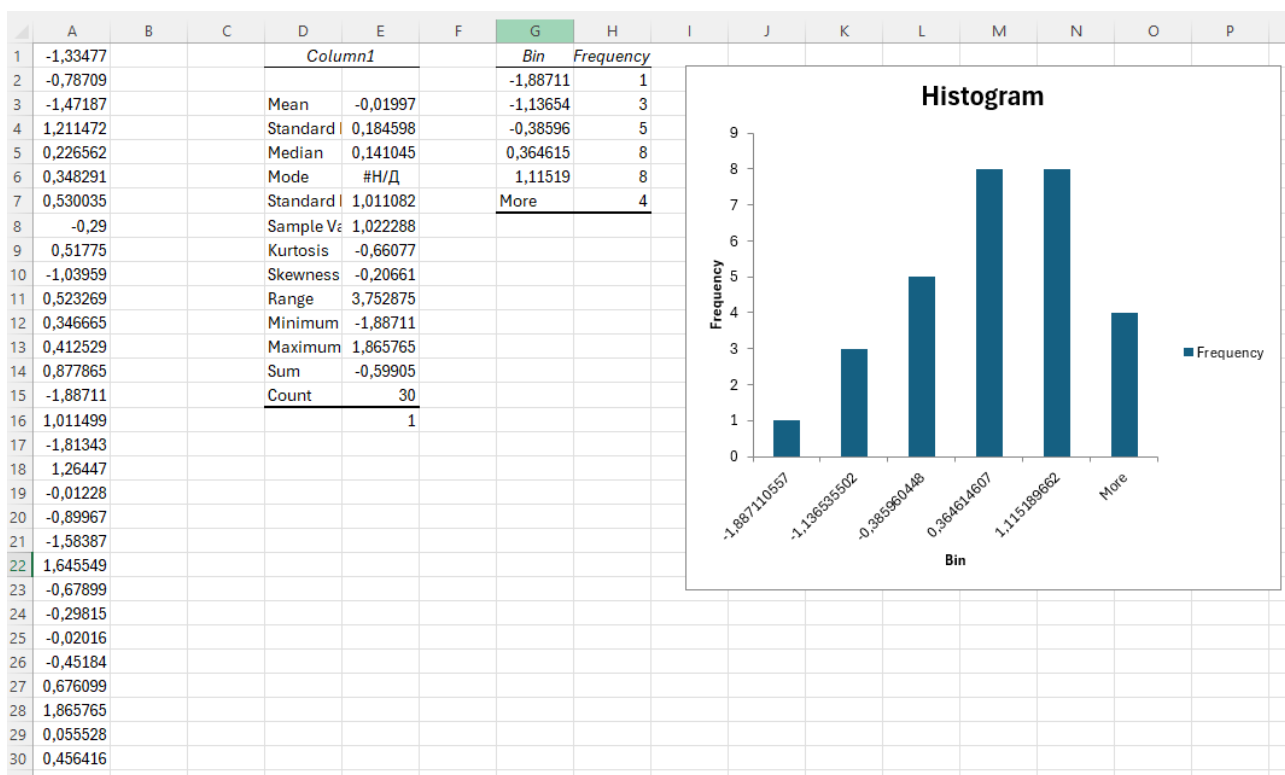


Рисунок 5.4

Наступний етап – підготовка даних для побудови теоретичної кривої нормального розподілу. У таблиці **Bin** (рис. 5.6) вказані значення кінців інтервалів, на яких розміщені основи стовпців гістограми. Використовуємо ці значення для визначення середин інтервалів.

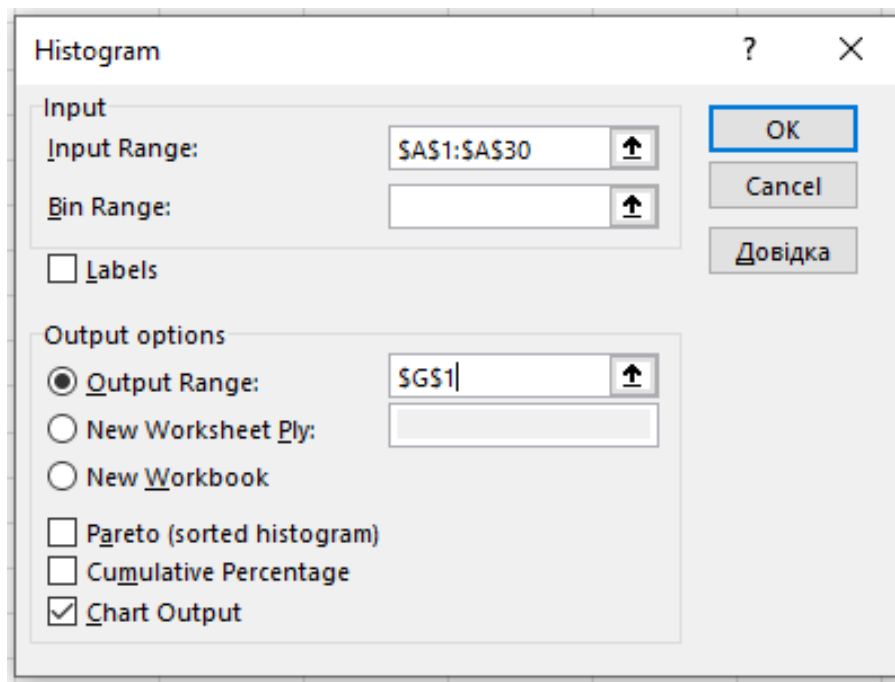


Рисунок 5.5

D	E	F	G	H
Column1			Bin	Frequency
			-1,88711	1
Mean	-0,036395273		-1,13654	4
Standard Error	0,190318551		-0,38596	5
Median	0,055528062		0,364615	8
Mode	#Н/Д		1,11519	7
Standard Deviation	1,024896763		More	4
Sample Variance	1,050413374			
Kurtosis	-0,7209017			
Skewness	-0,158162637			
Range	3,752875273			
Minimum	-1,887110557			
Maximum	1,865764716			
Sum	-1,055462917			
Count	29			
	1			
Середини інтервалів	Ймовірності	Частота		
=(G2+G3)/2				

Рисунок 5.6

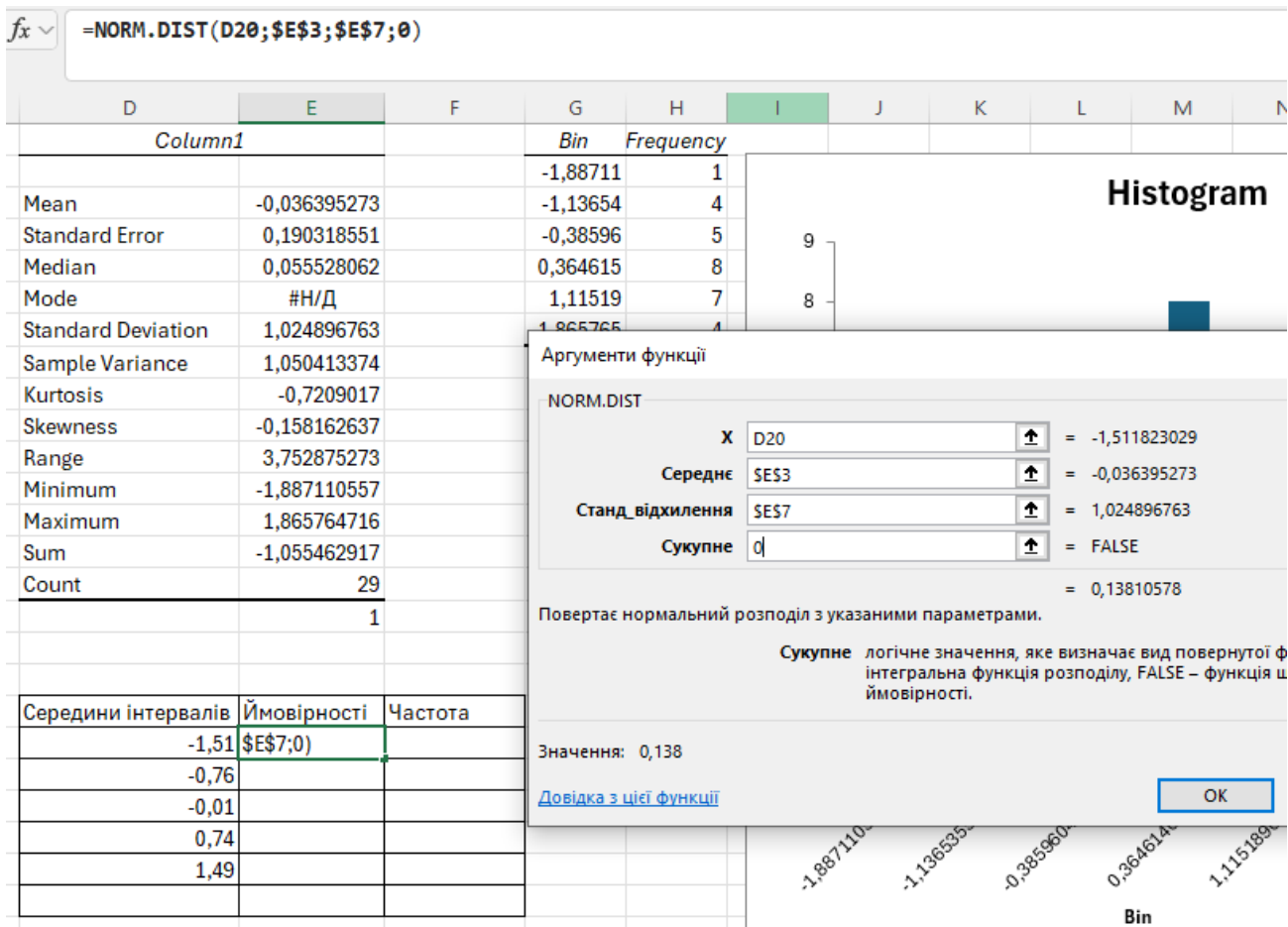


Рисунок 5.7

У стовпці **Ймовірності** за допомогою функції **NORM.DIST** розрахуємо значення щільності розподілу (рис.5.7).

Для розрахунку частот значення ймовірностей слід помножити на ширину інтервалу та на загальну кількість значень вибірки (рис.5.8).

Середини інтервалів	Ймовірності	Частота
-1,51	0,138	$=E20*SE$27*SE28
-0,76		
-0,01		
0,74		
1,49		
Ширина інтервалів	0,751	
Усього значень	30	

Рисунок 5.8

Значення частот можна округлити до цілого числа і записати в нижню комірку нуля.

Середини інтервалів	Ймовірності	Частота
-1,51	0,138	3
-0,76	0,303	7
-0,01	0,389	9
0,74	0,292	7
1,49	0,128	3
		0

Рисунок 5.9

Для додавання теоретичної кривої правою кнопкою миші натиснемо на стовпчик гістограми і потім на пункт **Виберіть дані** (рис.5.10).

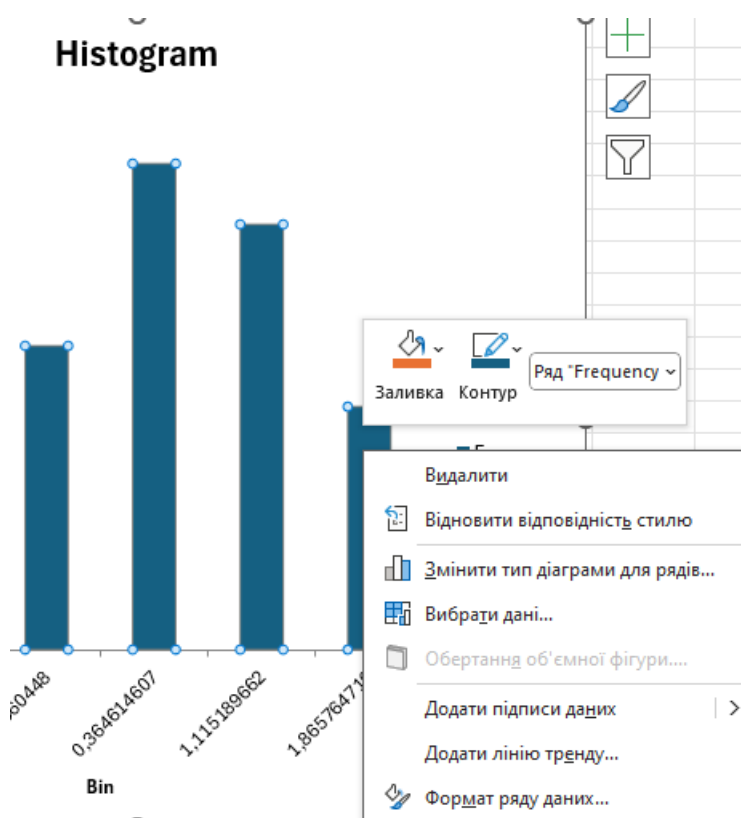


Рисунок 5.10

У вікні потрібно вибрати кнопку **Додати**, задати **Ім'я ряду** та **Значення ряду** (рис 5.11).

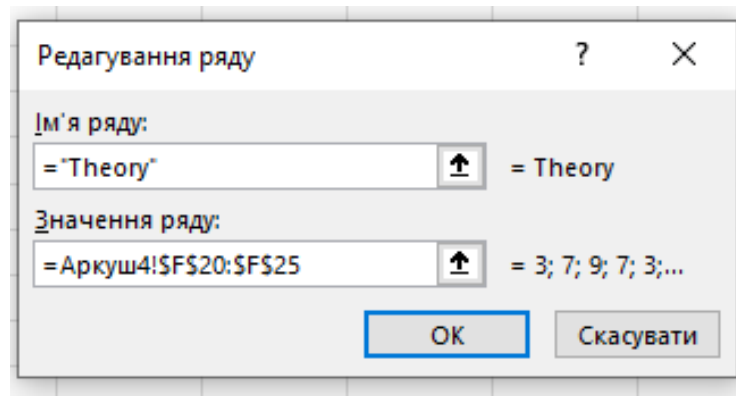


Рисунок 5.11

Потім правою кнопкою миші слід натиснути на новій діаграмі та вибрати пункт **Змінити тип діаграми для рядів** (рис.5.12).

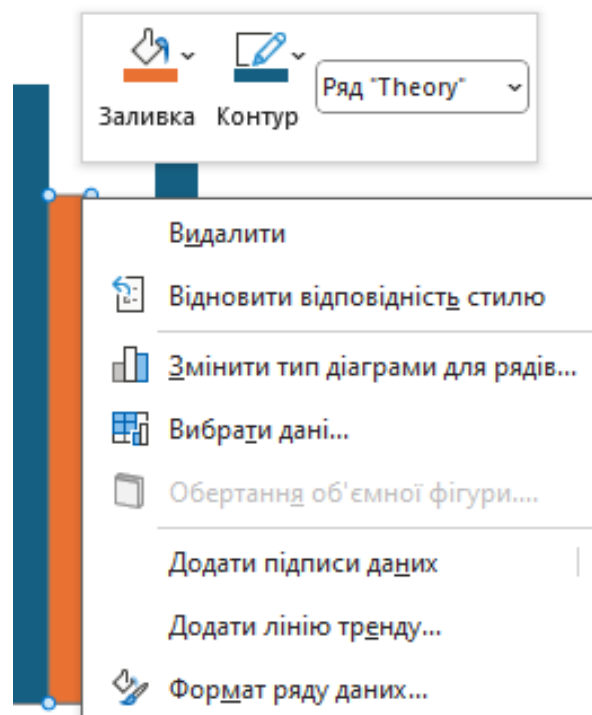


Рисунок 5.12

У випадяючому вікні для ряду вибираємо **Точкову діаграму з плавними лініями** (рис.5.13). Натискаємо ОК, результат показаний на рис.5.14.

Зауваження. Зазначимо, що перші два методи перевірки на нормальність розподілу є досить грубими. Нижче будуть розглянуті більш точні методи.



Рисунок 5.13

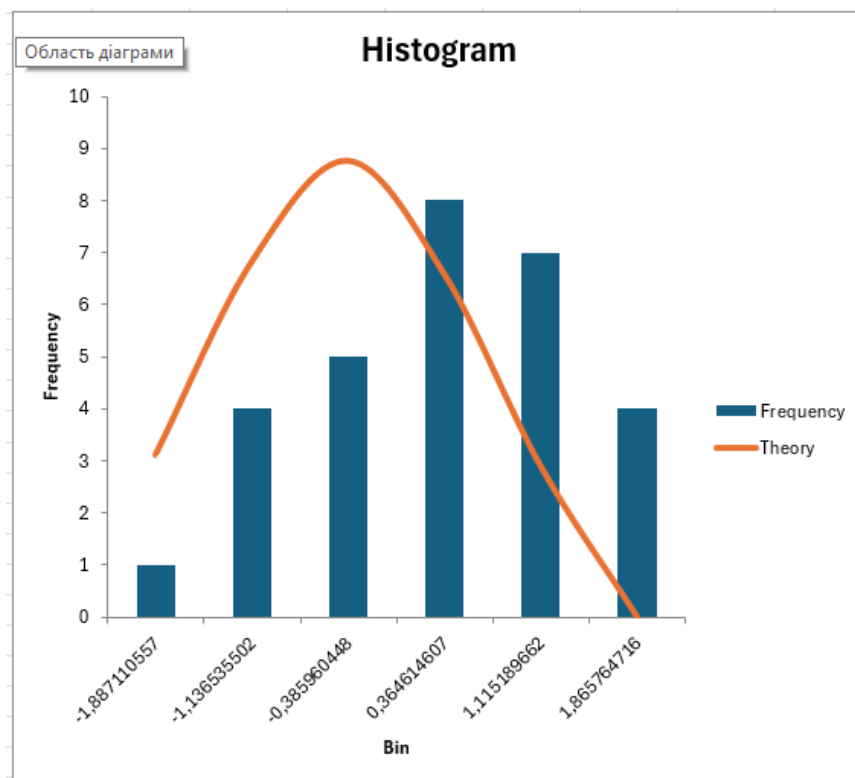


Рисунок 5.14

3.3 Метод квантиль-квантильних графіків (Q-Q plot)

У таблиці 5.1 задана вибірка, яку потрібно перевірити на нормальність розподілу. Якщо відсортуємо вибірку в порядку зростання, отримаємо набір значень x_i , які називають *емпіричними квантилями*. Для перевірки нормальності емпіричні квантилі порівнюються з *теоретичними квантилями* стандартного нормального розподілу q_i , які знаходять через обернену функцію розподілу:

$$q_i = F^{-1}(p_i).$$

На практиці при побудові QQ-plot, рівні p_i обирають наступним чином:

$$p_i = \frac{i - 0,5}{n}, \quad i = 1, 2, \dots, n.$$

В останній формулі ми віднімаємо від номера число 0,5, оскільки, якщо взяти i/n , то крайні значення (мінімум і максимум) надто чутливі до викидів, а зміщення на 0,5 дає більш «згладжену» оцінку. Також зазначимо, що ймовірності попадання випадкової величини між сусідніми теоретичними квантилями однакові, тобто $p_{i+1} - p_i = \text{const}$.

Далі будуємо на графіку пари точок (q_i, x_i) . Якщо розподіл вибірки збігається з теоретичним нормальним, то пари (q_i, x_i) повинні лежати на прямій

$$y = ax + b,$$

де b - середнє значення емпіричної вибірки, а a – її стандартне відхилення.

Пряма $y = ax + b$ - це лінія узгодження, яка показує, наскільки емпіричні квантилі збігаються з теоретичними.

Проведемо обчислення та побудови в програмі *Excel*. У діапазоні **A1:A30** розрахуємо значення p_i (рис.5.15). Потім в діапазоні **B1:B30** за допомогою функції **NORM.S.INV** знайдемо теоретичні квантилі (рис.5.16). В діапазоні **C1:C30** розмістимо елементи вибірки (табл. 5.1) в порядку зростання (рис.5.17).

Виділимо стовпці **B** і **C** та за допомогою пункту меню **Вставлення** побудуємо точкову діаграму з лінією тренду (рис.5.18).

	A	B	C
1	0,01666667	-2,12804523	-1,89
2	0,05	-1,64485363	-1,81
3	0,08333333	-1,38299413	-1,58
4	0,11666667	-1,19181617	-1,47
5	0,15	-1,03643339	-1,33
6	0,18333333	-0,90273479	-1,04
7	0,21666667	-0,78350038	-0,90
8	0,25	-0,67448975	-0,79
9	0,28333333	-0,57296755	-0,68
10	0,31666667	-0,47704043	-0,45
11	0,35	-0,38532047	-0,30
12	0,38333333	-0,29673784	-0,29
13	0,41666667	-0,21042839	-0,02
14	0,45	-0,12566135	-0,01
15	0,48333333	-0,0417893	0,06
16	0,51666667	0,041789298	0,23
17	0,55	0,125661347	0,35
18	0,58333333	0,210428394	0,35
19	0,61666667	0,296737838	0,41
20	0,65	0,385320466	0,46
21	0,68333333	0,477040428	0,52
22	0,71666667	0,572967548	0,52
23	0,75	0,67448975	0,53
24	0,78333333	0,783500375	0,68
25	0,81666667	0,902734792	0,88
26	0,85	1,036433389	1,01
27	0,88333333	1,191816172	1,21
28	0,91666667	1,382994127	1,26
29	0,95	1,644853627	1,65
30	0,98333333	2,128045234	1,87

Рисунок 5.17

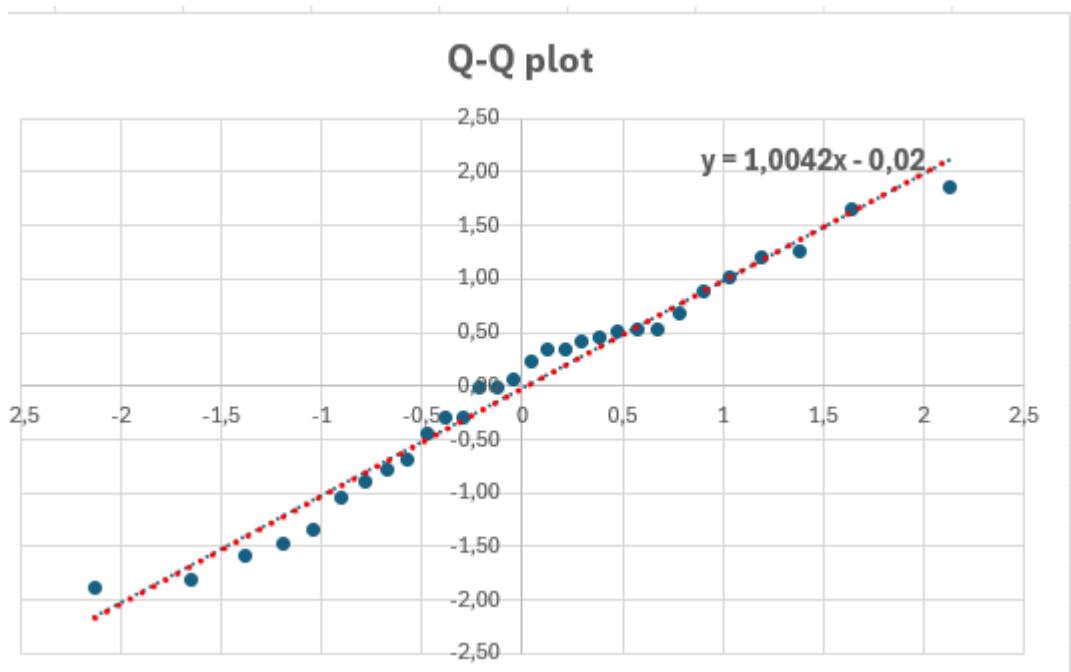


Рисунок 5.18

Як і передбачалося, параметри прямої - зсув та кутовий коефіцієнт - приблизно дорівнюють середньому та дисперсії нашої емпіричної вибірки. Точки

на графіку близькі до точок прямої, тому можна зробити висновок, що наш розподіл швидше за все є нормальним.

Практичне завдання

1. Перевірити вибірку на нормальність за допомогою:

- правила трьох сигм;
- методу квантиль - квантильних графіків (*Q-Q plot*);

Для виконання розрахунків використовувати данні з таблиці 5.3.

2. Файл **Word** зі звітом прикріпити до завдання.

Таблиця 5.3

n	Варіант №														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	- 0,05	0,52	0,4	- 0,45	0,36	- 0,47	1,08	0,03	- 1,54	- 0,57	0,64	- 0,28	- 1,55	- 0,31	1,61
2	0,37	- 0,55	- 0,51	- 0,15	0,56	- 0,04	- 1,04	1,52	- 0,09	0,82	1,36	- 0,44	0,31	0,83	-1,39
3	- 0,72	- 0,76	0,23	1,2	0,39	0,06	0,99	- 1,54	- 0,36	0,47	- 0,81	- 1,28	0,23	0,33	-0,65
4	- 0,33	- 0,47	1,22	- 0,45	- 0,12	- 0,87	0,16	0,33	- 0,26	0,07	- 1,96	- 0,28	1,42	0,35	1,92
5	- 0,05	- 0,2	0,32	- 0,21	0,77	- 0,28	0,98	- 0,42	0,84	- 0,81	0,06	- 0,06	- 1,06	- 0,49	-0,34
6	1,07	- 0,47	0,2	- 0,15	1,02	- 0,57	- 0,72	1,37	- 0,99	- 1,06	1	1	0,99	- 0,01	-0,03
7	0,7	0,66	- 0,93	0,46	- 0,41	0,79	0,13	- 1,57	0,1	- 0,66	- 0,66	- 0,19	- 0,08	- 0,77	0,25
8	-0,1	0	0,56	- 0,21	0,11	- 0,22	- 0,92	- 1,17	- 1,63	- 0,95	- -0,2	- 0,18	- 1,21	0,95	0,94
9	0,38	- 0,79	0,56	0,18	0,6	- 0,57	- 0,27	- 0,38	0,83	- 0,21	- 0,07	- 1,52	- 0,28	- 0,25	-0,18
10	0,1	0,19	- 0,39	-0,9	0,33	0,47	- 1,38	- 0,12	- 1,03	- 0,24	- 0,19	- 1,83	- 0,57	- 0,62	-0,14
11	- 0,52	- 0,79	- 0,58	0,06	- 0,73	- -0,5	- 0,07	- 0,35	- 0,79	- 0,15	- 0,51	- 0,01	- 1,76	0,19	0,43
12	- 0,41	0,84	- 0,26	0,76	0,78	- 1,34	- 0,69	- 0,51	- -0,8	- 0,16	- 0,1	- 0,39	- 1,22	- 0,99	1,52
13	- 0,53	- 0,66	- 0,75	0,37	1,07	- 0,33	- 0,56	- 0,33	- 1,27	- 0,07	- 1,36	- 0,11	- 1,83	1,35	-0,25

14	-	-		-		-		-	-			-		-	
	0,04	0,25	0,33	0,15	1,25	0,57	0,66	0,21	0,81	0	1,78	0,68	0,98	1,14	-1,99
15						-		-		-			-		
	0,08	0,11	0,63	0,56	0,24	0,96	-1,9	0,81	0,92	0,07	0,69	0,35	0,62	0,1	-0,55
16	-	-					-					-		-	
	0,89	0,43	0,07	0,55	1,51	0,15	0,76	0,16	0,33	0,23	0,08	1,06	0,14	0,09	0,95
17		-	-				-						-		
	0,53	0,54	1,09	0,26	0,46	0,09	0,75	0,4	0,7	-1,2	0,14	1,1	0,38	-0,2	0,19
18										-	-				
	0,02	0,14	0,96	0,45	0,26	0,1	0,3	0,96	0,04	0,85	0,97	1,18	0,11	1,73	-0,13
19	-	-								-	-	-	-		
	0,08	0,86	0,29	0,68	0,05	1,68	0,91	0,49	0,66	0,64	1,14	0,21	1,43	1,54	-0,06
20	-			-	-	-			-		-	-			
	0,16	0,57	0,14	0,47	0,49	0,39	-0,2	0,3	1,58	1,46	0,85	1,51	-0,4	0,83	1,2
21		-	-			-		-	-				-	-	
	0,79	0,71	0,97	-0,4	-0,2	1,83	-0,4	0,06	0,86	-0,7	0,75	-0,1	0,32	0,09	-1,03
22	-	-	-			-			-			-	-	-	
	0,09	0,69	0,07	0,42	0,06	0,48	0,9	1,16	0,83	0,07	1,75	1,63	1,15	1,92	0,69
23	-	-			-	-		-	-	-		-			
	0,09	0,08	0,03	1,11	0,26	0,43	0,34	1,11	0,58	0,44	0,94	0,01	0,05	0,8	-0,93
24		-		-			-	-	-	-	-			-	
	0,36	1,22	0,06	0,07	0,01	0,27	0,01	0,47	0,08	0,73	0,24	0,34	0,46	0,03	0,72
25		-		-	-			-						-	
	0	0,01	0,35	0,67	1,53	0,66	0,9	0,41	0,4	-1	0,66	0,08	0,3	0,87	-1,59
26			-					-					-		
	0,39	-0,4	0,51	0,8	0,71	0,11	0,08	0,69	0,55	0,44	1,42	1,95	0,27	0,27	0,45
27			-	-		-	-		-		-		-		
	0,04	0,18	0,29	0,21	-1,2	1,56	0,49	0,2	1,34	0,38	0,67	0,74	0,99	0,26	-1,63
28		-	-	-		-		-				-		-	
	0,03	0,06	0,46	0,26	0,14	0,34	1	0,23	0,11	0,08	1,16	0,44	0,43	0,19	-1,88
29			-	-		-			-			-	-		
	0,6	0,49	0,14	1,55	0,08	0,23	0,27	0,52	0,54	1,01	0,78	1,06	0,35	0,5	-0,57
30	-	-		-	-	-		-	-			-	-	-	
	0,03	0,49	-1,1	0,09	0,08	1,06	0,53	0,94	0,86	0,54	-0,8	0,59	1,51	0,38	-0,37

ПРАКТИЧНА РОБОТА 6

Перевірка статистичних гіпотез. Тест Шапіро–Уїлка.

Критерій Пірсона

Питання, які розглядаються:

1. Статистичні гіпотези.
2. Тест Шапіро–Уїлка
3. Критерій Пірсона.

Актуальною є задача визначення закону розподілу випадкової величини або параметрів вже відомого розподілу. Для її вирішення висувається статистична гіпотеза, яка після дослідження або приймається або відкидається. При цьому для перевірки відповідності даних нормальному закону розподілу часто застосовуються методи Шапіро- Уїлка та Пірсона χ^2 (ксі-квадрат).

1. Статистичні гіпотези

Статистичною називають гіпотезу про вид або окремі параметри розподілу випадкової величини.

Нульовою (основною) називають гіпотезу H_0 , що висувається.

Альтернативною (конкуруючою) називають гіпотезу H_1 , яка суперечить нульовій гіпотезі H_0 .

Перевірка гіпотези ґрунтується на обчисленні деякої випадкової величини, яку називають критерієм. **Критерієм** називають випадкову величину K , яка слугує для перевірки нульової гіпотези.

При формулюванні нульової гіпотези фіксують деяке граничне значення, яке називається **рівнем значущості** і позначається α .

Рівень значущості є ймовірністю відхилити нульову гіпотезу, за умови, що вона вірна, тобто зробити помилку першого роду. На практиці часто вважають $\alpha = 0,05$ или $\alpha = 0,01$. Можна сказати, що приймаючи $\alpha = 0,05$, мы заздалегідь

зафіксували ризок *помилки першого роду*. Приймаючи нульову гіпотезу треба розуміти, що можливість отримати такі дані знаходиться в інтервалі $(\alpha; 1)$.

Існує також *помилка другого роду*, яка полягає в тому, що приймається нульова гіпотеза, хоч вона невірна. Імовірність такої помилки позначається β . Знижуючи рівень значущості, ми зменшуємо ймовірність помилки першого роду, але при цьому зростає ймовірність помилки другого роду. У табл. 6.1 показані випадки виникнення помилок під час формулювання гіпотез про зв'язки явищ.

Таблиця 6.1

Статистика Реальність	Зв'язок є	Зв'язку немає
Зв'язок є H_1	узгоджується	помилка другого роду (β)
Зв'язку немає H_0	помилка першого роду (α)	узгоджується

Помилка першого роду (α) — ми відхиляємо нульову гіпотезу, хоча вона насправді вірна. Тобто помилку першого роду можна трактувати за допомогою терміну "переглянули".

Помилка другого роду (β) — ми не відхиляємо нульову гіпотезу, хоча вона хибна. Можна трактувати за допомогою терміну "недогляділи".

Перевірка гіпотез зазвичай проходить такі етапи.

- 1) Визначення статистичної моделі щодо закону розподілу випадкової величини та її параметрів;
- 2) формулювання H_0 и H_1 ;
- 3) вибирають критерій для статистичної моделі;
- 4) вибирають рівень значущості α ;
- 5) за вибраним критерієм знаходять критичне значення (заздалегідь відомо) і визначають критичні області (рис. 1.1);
- 6) розраховують значення обраного статистичного критерію за даними;

7) якщо значення критерію не лежить у критичній ділянці (значення критерію за модулем менше критичного) приймається нульова гіпотеза. В іншому випадку нульова гіпотеза відкидається.

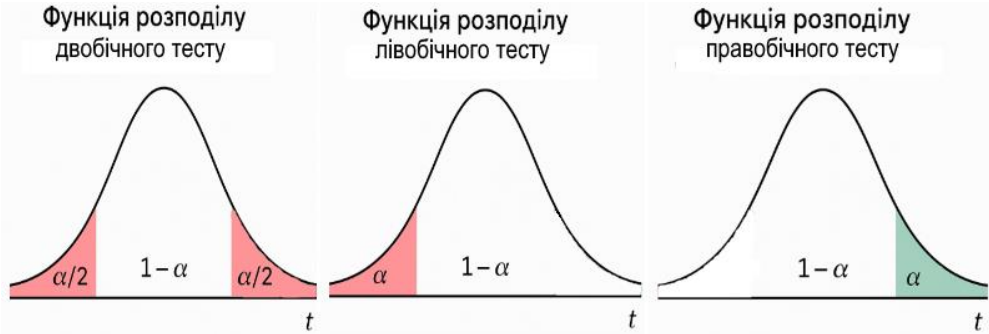


Рисунок 6.1

2. Тест Шапіро–Уїлка

Нехай задано 30 значень випадкової величини (табл. 6.2). Проведемо перевірку даного розподілу на нормальність за допомогою тесту *Шапіро–Уїлка* (*Shapiro–Wilk test*).

Таблиця 6.2

1,33	0,79	1,47	1,21	0,23	0,35	0,53	0,29	0,52	1,04	0,52	0,35	0,41	0,88	1,89
1,01	1,81	1,26	0,01	0,90	1,58	1,65	0,68	0,30	0,02	0,45	0,68	1,87	0,06	0,46

Спочатку формулюються гіпотези:

- H_0 : дані у вибірці розподілені нормально,
- H_1 : дані у вибірці не розподілені нормально.

Якщо розрахована статистика $W \geq W_{theory}$, приймається нульова гіпотеза H_0 . Якщо $W < W_{theory}$, приймається альтернативна гіпотеза H_1 .

Статистика *тесту Шапіро–Уїлка* розраховується за формулою:

$$W = \frac{(\sum_{i=1}^n a_i x_i)^2}{\sum_{i=1}^n (x_i - X_{cp.})^2},$$

де x_i – елемент вибірки, яка відсортована за зростанням;

$X_{\text{ср.}}$ – її вибіркове середнє;

n – кількість елементів вибірки;

α_i - коефіцієнти *Шаніро–Уїлка*.

Ці коефіцієнти табульовані, залежать лише від n , їх значення можна розрахувати самостійно або взяти в опублікованих таблицях. Для $n = 30$ використовується 15 значень, оскільки коефіцієнти симетричні за абсолютною величиною. Ці значення наведені в табл. 6.2.

Таблиця 6.2

k	n									
	21	22	23	24	25	26	27	28	29	30
1	0,4643	0,4590	0,4542	0,4493	0,4450	0,4407	0,4366	0,4328	0,4291	0,4254
2	0,3185	0,3156	0,3126	0,3098	0,3069	0,3043	0,3018	0,2992	0,2968	0,2944
3	0,2578	0,2571	0,2563	0,2554	0,2543	0,2533	0,2522	0,2510	0,2499	0,2487
4	0,2119	0,2131	0,2139	0,2145	0,2148	0,2151	0,2152	0,2151	0,2150	0,2148
5	0,1736	0,1764	0,1787	0,1807	0,1822	0,1836	0,1848	0,1857	0,1864	0,1870
6	0,1399	0,1443	0,1480	0,1512	0,1539	0,1563	0,1584	0,1601	0,1616	0,1630
7	0,1092	0,1150	0,1201	0,1245	0,1283	0,1316	0,1346	0,1372	0,1395	0,1415
8	0,0804	0,0878	0,0941	0,0997	0,1046	0,1089	0,1128	0,1162	0,1192	0,1219
9	0,0530	0,0618	0,0696	0,0764	0,0823	0,0876	0,0923	0,0965	0,1002	0,1036
10	0,0263	0,0368	0,0459	0,0539	0,0610	0,0672	0,0728	0,0778	0,0822	0,0862
11	–	0,0122	0,0228	0,0321	0,0403	0,0476	0,0540	0,0598	0,0598	0,0697
12	–	–	–	0,0107	0,0200	0,0284	0,0358	0,0424	0,0483	0,0537
13	–	–	–	–	–	0,0094	0,0178	0,0253	0,0320	0,0381
14	–	–	–	–	–	–	–	0,0084	0,0159	0,0227
15	–	–	–	–	–	–	–	–	–	0,0076

Проведемо необхідні розрахунки в програмі **Excel**.

а) у діапазоні **A2:A31** розмістимо значення нашої вибірки з табл. 6.1 у порядку зростання (рис.6.2).

б) у діапазон **B2:B16** запишемо в порядку спадання табульовані коефіцієнти з таблиці 6.2, а в діапазон **B17:B31** ті самі значення, тільки в порядку зростання і зі знаком мінус (рис.6.3).

в) розрахуємо за формулами значення в стовпцях **C** і **D**, і потім у комірках **C32** і **D32** за допомогою вбудованої функції **SUM** знайдемо суми значень із діапазонів **C2:C31**, **D2:D31** відповідно (рис.6.4).

	A	B	C	D
1	x_i	a_i	$x_i a_i$	$(x_i - X_{cp.})^2$
2	-1,89			
3	-1,81			
4	-1,58			
5	-1,47			
6	-1,33			
7	-1,04			
8	-0,90			
9	-0,79			
10	-0,68			
11	-0,45			
12	-0,30			
13	-0,29			
14	-0,02			
15	-0,01			
16	0,06			
17	0,23			
18	0,35			
19	0,35			
20	0,41			
21	0,46			
22	0,52			
23	0,52			
24	0,53			
25	0,68			
26	0,88			
27	1,01			
28	1,21			
29	1,26			
30	1,65			
31	1,87			

Рисунок 6.2

	A	B	C	D
1	x_i	a_i	$x_i a_i$	$(x_i - X_{cp.})^2$
2	-1,89	0,4254		
3	-1,81	0,2944		
4	-1,58	0,2487		
5	-1,47	0,2148		
6	-1,33	0,1870		
7	-1,04	0,1630		
8	-0,90	0,1415		
9	-0,79	0,1219		
10	-0,68	0,1036		
11	-0,45	0,0862		
12	-0,30	0,0697		
13	-0,29	0,0537		
14	-0,02	0,0381		
15	-0,01	0,0227		
16	0,06	0,0076		
17	0,23	-0,0076		
18	0,35	-0,0227		
19	0,35	-0,0381		
20	0,41	-0,0537		
21	0,46	-0,0697		
22	0,52	-0,0862		
23	0,52	-0,1036		
24	0,53	-0,1219		
25	0,68	-0,1415		
26	0,88	-0,1630		
27	1,01	-0,1870		
28	1,21	-0,2148		
29	1,26	-0,2487		
30	1,65	-0,2944		
31	1,87	-0,4254		

Рисунок 6.3

	A	B	C	D
1	x_i	a_i	$x_i a_i$	$(x_i - X_{cp.})^2$
2	-1,89	0,4254	-0,8028	3,49
3	-1,81	0,2944	-0,5339	3,22
4	-1,58	0,2487	-0,3939	2,45
5	-1,47	0,2148	-0,3162	2,11
6	-1,33	0,1870	-0,2496	1,73
7	-1,04	0,1630	-0,1695	1,04
8	-0,90	0,1415	-0,1273	0,77
9	-0,79	0,1219	-0,0959	0,59
10	-0,68	0,1036	-0,0703	0,43
11	-0,45	0,0862	-0,0389	0,19
12	-0,30	0,0697	-0,0208	0,08
13	-0,29	0,0537	-0,0156	0,07
14	-0,02	0,0381	-0,0008	0,00
15	-0,01	0,0227	-0,0003	0,00
16	0,06	0,0076	0,0004	0,01
17	0,23	-0,0076	-0,0017	0,06
18	0,35	-0,0227	-0,0079	0,13
19	0,35	-0,0381	-0,0133	0,14
20	0,41	-0,0537	-0,0222	0,19
21	0,46	-0,0697	-0,0318	0,23
22	0,52	-0,0862	-0,0446	0,29
23	0,52	-0,1036	-0,0542	0,30
24	0,53	-0,1219	-0,0646	0,30
25	0,68	-0,1415	-0,0957	0,48
26	0,88	-0,1630	-0,1431	0,81
27	1,01	-0,1870	-0,1892	1,06
28	1,21	-0,2148	-0,2602	1,52
29	1,26	-0,2487	-0,3145	1,65
30	1,65	-0,2944	-0,4844	2,77
31	1,87	-0,4254	-0,7937	3,56
32			-5,3563	29,6463

Рисунок 6.4

г) Остаточно обчислюємо статистику тесту за формулою (рис. 6.5):

$$W = \frac{(\sum_{i=1}^n a_i x_i)^2}{\sum_{i=1}^n (x_i - X_{\text{ср.}})^2} = 0,968.$$

	D	E	F	G
	$(x_i - X_{\text{ср.}})^2$			
128	3,49		W=	0,968

Рисунок 6.5

д) порівнюємо отримане значення з критичним $W_{\text{кр.}} = 0,927$ з таблиці 6.3.

Таблиця 6.3

<i>n</i>	<i>p</i>	
	0,01	0,05
26	0,891	0,920
27	0,894	0,923
28	0,896	0,924
29	0,898	0,926
30	0,900	0,927

е) Робимо остаточний висновок: оскільки $0,968 > 0,927$, тобто $W > W_{\text{кр.}}$, розподіл досліджуваних даних можна вважати нормальним.

3. Критерій Пірсона

Щільність розподілу χ^2 має вигляд

$$f(x) = \frac{1}{2^{\frac{n}{2}}} \frac{1}{\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{\left(-\frac{x}{2}\right)},$$

де $\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt, x > 0$ - гама-функція.

Математичне очікування розподілу χ^2 дорівнює n , а дисперсія – $2n$.

Щільності розподілу χ^2 відрізняються щодо різних значень ступеня свободи. На рис. 6.6 показаний графік щільності розподілу для ступеня свободи $df = 30$, а також критичні значення при рівні значущості $\alpha = 0,05$.

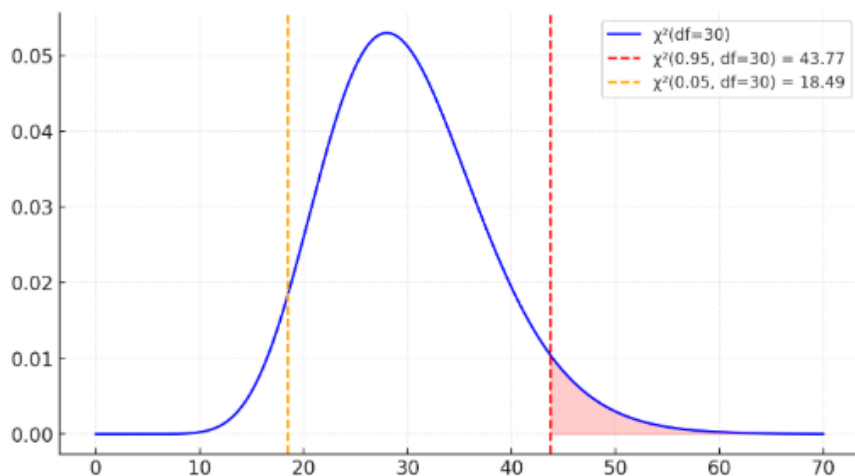


Рисунок 6.6

Для перевірки відповідності спостережуваного закону розподілу теоретичному (критерій згоди) застосовується формула

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - n_i^{\text{теор.}})^2}{n_i^{\text{теор.}}},$$

де n_i – частоти, що спостерігаються;

$n_i^{\text{теор.}}$ – теоретичні частоти, k - кількість частот.

Обчислене значення χ^2 порівнюється з теоретичним при заданих рівні значущості та ступені свободи. Якщо χ^2 менше, ніж теоретичне значення (критерій Пірсона правосторонній), приймаємо гіпотезу про відсутність різниць між спостережуваними та теоретичними даними.

При застосуванні **критерія Пірсона** для перевірки нормальності вибірки слід дотримуватися таких умов:

- дані мають бути згруповані в інтервали;

- частоти, що спостерігаються ≥ 5 ;
- обсяг вибірки має бути не меншим 30.

Розглянемо з прикладу конкретного набору даних алгоритм застосування **критерію Пірсона** для перевірки гіпотези про відповідність вибірки нормальному закону розподілу. Обчислення будемо проводити за допомогою засобів програми **Excel**.

Для прикладу виборки з табл.6.4 візьмемо 30 значень рівня систолічного артеріального тиску (мм рт. ст.).

Таблиця 6.4

118	122	115	128	110	134	126	120	124	118	130	117	133	111	127
116	119	125	129	121	123	112	132	128	114	136	122	119	131	124

	A	B	C	D	E
1	№	Тиск			
2	1	118		Column1	
3	2	122			
4	3	115		Mean	122,8
5	4	128		Standard Error	1,296
6	5	110		Median	122,5
7	6	134		Mode	118
8	7	126		Standard Deviation	7,097
9	8	120		Sample Variance	50,372
10	9	124		Kurtosis	-0,846
11	10	118		Skewness	0,009
12	11	130		Range	26
13	12	117		Minimum	110
14	13	133		Maximum	136
15	14	111		Sum	3684
16	15	127		Count	30
17	16	116			
18	17	119			
19	18	125			
20	19	129			
21	20	121			
22	21	123			
23	22	112			
24	23	132			
25	24	128			
26	25	114			
27	26	136			
28	27	122			
29	28	119			
30	29	131			
31	30	124			

Рисунок 6.7

1. Сформулюємо наступні гіпотези:

- H_0 : дані у вибірці розподілені нормально,
- H_1 : дані у вибірці не розподілені нормально.

Якщо розрахована статистика $\chi^2 \leq \chi^2_{theory}$ приймається нульова гіпотеза H_0 .

Якщо $\chi^2 > \chi^2_{theory}$ приймається альтернативна гіпотеза H_1 .

2. За допомогою пакету **Data Analysis** (див. практична робота 3, п.4.2, рис. 3.9) обчислюємо числові характеристики виборки (рис.6.7).

3. Обчислюємо кількість інтервалів (рис. 6.8) та їх ширину за формулами:

$$k = 1 + 1,443 \ln n \approx 6,$$

$$\Delta = \frac{x_{max} - x_{min}}{k} \approx 4,4.$$

4. Будуємо інтервальний ряд та за допомогою функції **FREQUENCY** визначаємо частоти в кожному інтервалі.

В першому полі вказаний діапазон вихідних даних вибірки, у другому - масив правих меж інтервалів; в результаті ми визначили частоти, що спостерігалися, в кожному з інтервалів (рис. 6.9). Легко перевірити, що сума частот дорівнює 30 (рис.6.10).

D	E	F	G	H
Column1			k=	6
			Δ =	4,4
Mean	122,8			
Standard Error	1,296			
Median	122,5			
Mode	118			
Standard Deviation	7,097			
Sample Variance	50,372			
Kurtosis	-0,846			
Skewness	0,009			
Range	26			
Minimum	110			
Maximum	136			
Sum	3684			
Count	30			

Рисунок 6.8

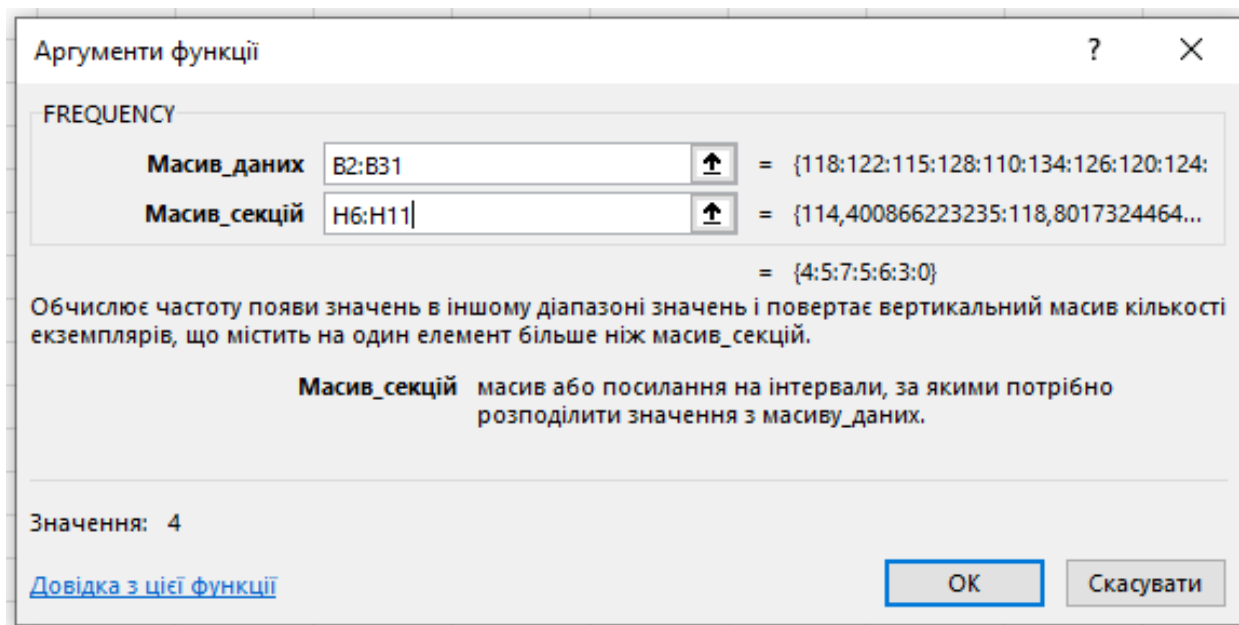


Рисунок 6.9

G	H	I
k=	6	
Δ =	4,4	
від	до	n_i
110	114,4	4
114,4	118,8	5
118,8	123,2	7
123,2	127,6	5
127,6	132,0	6
132,0	136,4	3

Рисунок 6.10

На цьому етапі можна побудувати діаграму частот, для того, щоб візуально порівняти її вигляд із діаграмою нормального розподілу (рис. 6.11).

5. Раніше було зазначено, що для коректності застосування **критерію Пірсона** необхідно, щоб кожен інтервал містив число частот не менше п'яти. Оскільки в

останньому шостому інтервалі число частот дорівнює трьом, об'єднаємо п'ятий і шостий інтервали (рис.6.12).

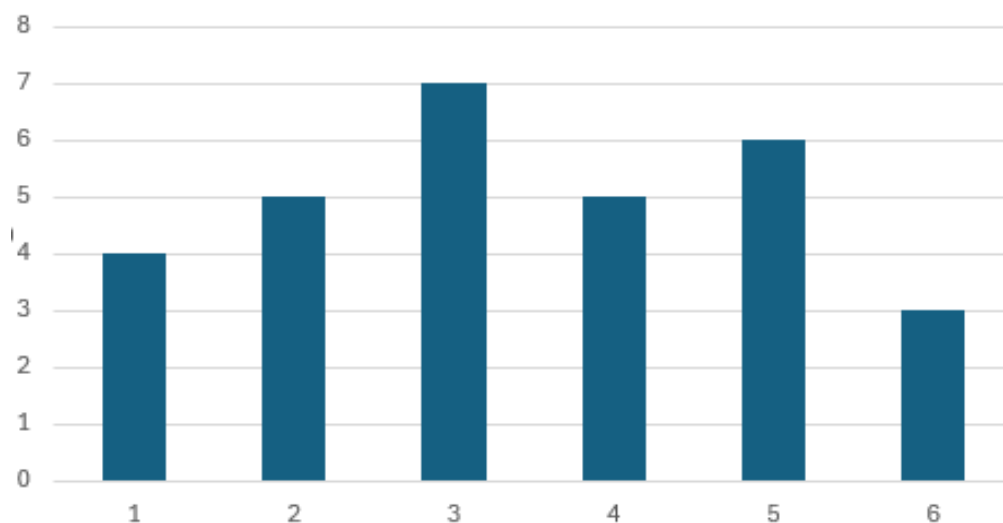


Рисунок 6.11

від	до	n_i
110	114,4	4
114,4	118,8	5
118,8	123,2	7
123,2	127,6	5
127,6	136,4	9

Рисунок 6.12

6. За допомогою функції **NORM.DIST** визначаємо теоретичні частоти. Для цього застосовуємо формулу:

$$=(\text{NORM.DIST}(\text{верхня_межа}; \text{середнє}; \text{стандартне_відхилення}; \text{TRUE}) - \text{NORM.DIST}(\text{нижня_межа}; \text{середнє}; \text{стандартне_відхилення}; \text{TRUE}) * 30).$$

Оскільки результатом функції **NORM.DIST** є значення ймовірності, щоб отримати значення частот, необхідно різницю помножити на обсяг вибірки, тобто на **30**.

Зауваження. Коли ми обчислюємо теоретичні частоти нормального розподілу через функцію **NORM.DIST**, потрібно враховувати особливості крайніх інтервалів.

Функція розподілу нормального закону при наближенні аргументу до мінус нескінченності наближається до нуля, а при наближенні до плюс нескінченності - до одиниці. Тому:

- для першого інтервалу використовуємо формулу

$$= \text{NORM.DIST}(\text{верхня_межа}; \text{середнє}; \text{стандартне_відхилення}; \text{TRUE}) * 30,$$
- для останнього інтервалу

$$= (1 - \text{NORM.DIST}(\text{нижня_межа}; \text{середнє}; \text{стандартне_відхилення}; \text{TRUE})) * 30.$$

Отриманий результат бачимо на рис.6.13.

7. Для того, щоб зробити остаточні висновки, розглянемо два можливих варіанти.

від	до	n_i	$n_i^{\text{теор.}}$
110	114,4	4	3,55
114,4	118,8	5	5,05
118,8	123,2	7	7,08
123,2	127,6	5	6,84
127,6	136,4	9	7,48

Рисунок 6.13

Перший варіант. Застосовуємо функцію **CHISQ.TEST** (рис. 6.14). В полі **Фактичний інтервал** вказуємо діапазон частот, що спостерігаються, а у полі **Очікуваний інтервал** – діапазон теоретичних частот. В результаті знаходимо ймовірність $p\text{-value} = 0,93$. Якщо рівень значущості $\alpha = 0,05$, отримуємо: $0,93 > 0,05$. Це означає, що гіпотеза про нормальність не відкидається — розподіл узгоджується з нормальним.

Другий варіант. Застосовуємо формулу

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - n_i^{\text{теор.}})^2}{n_i^{\text{теор.}}}$$

В Excel неважко зробити наступні розрахунки (рис.6.14).

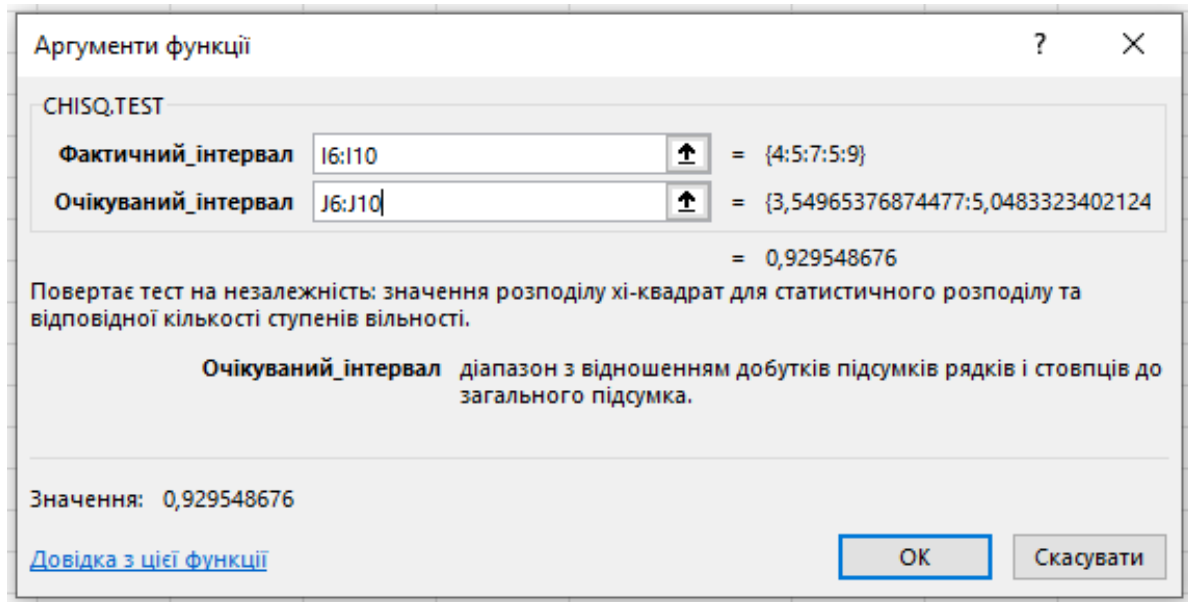


Рисунок 6.14

від	до	n_i	$n_i^{\text{теор.}}$	$\frac{(n_i - n_i^{\text{теор.}})^2}{n_i^{\text{теор.}}}$
110	114,4	4	3,55	0,057
114,4	118,8	5	5,05	0,000
118,8	123,2	7	7,08	0,001
123,2	127,6	5	6,84	0,497
127,6	136,4	9	7,48	0,310
			Сума:	0,865

Рисунок 6.15

Таким чином ми знайшли спостережуване значення $\chi^2 = 0,865$, яке потрібно порівняти з теоретичним (рис.6.15). Теоретичне значення можна знайти з таблиці або за допомогою функції **CHISQ.INV**. Для цього потрібно задати рівень значущості, наприклад. $\alpha = 0,05$ та обчислити число ступенів свободи за формулою:

$$df = k - p - 1,$$

де k — кількість інтервалів, p — кількість оцінених параметрів розподілу (для нормального розподілу $p = 2$: середнє та стандартне відхилення).

Використовуємо для розрахунків функцію **CHISQ.INV** (рис. 6.16).

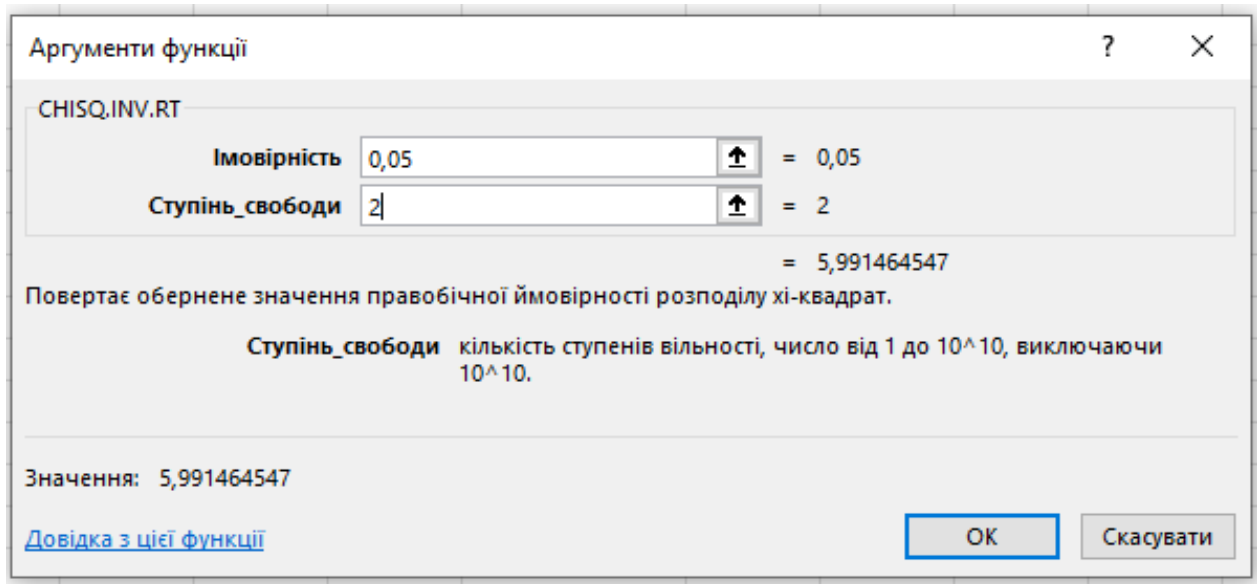


Рисунок 6.16

В полі **Імовірність** вводимо

$$1 - \alpha = 1 - 0,05 = 0,95,$$

в полі **Ступінь_свободи**

$$df = k - p - 1 = 5 - 2 - 1 = 2.$$

Отримуємо теоретичне значення критерію $\chi^2 = 5,991$.

Оскільки $0,865 < 5,991$, нормальність підтверджена.

Практичне завдання

Перевірити вибірку ($n = 30$) на нормальність за допомогою:

- 1) тесту Шапіро-Уїлка;
- 2) критерію Пірсона.

Під час виконання завдання використовуйте Додаток Б та Додаток В.

Варіант №1 Систолічний артеріальний тиск (мм рт. ст.)

220	174	211	261	184	210	154	228	228	180	171	187	162	216	232
204	146	248	214	207	152	197	201	203	217	175	185	177	193	145

Варіант №2 Діастолічний артеріальний тиск (мм рт. ст.)

73	70	78	72	82	80	76	74	78	82	83	71	89	74	71
71	87	85	84	74	84	77	70	74	70	83	85	72	77	80

Варіант №3 Частота серцевих скорочень (уд./хв)

81	62	58	64	68	63	84	72	58	75	58	87	60	68	74
64	62	75	56	82	59	59	71	50	72	65	75	71	81	63

Варіант №4 Рівень глюкози натще (мг/дл)

99	78	67	83	62	65	66	82	75	41	120	131	64	100	92
68	73	107	48	99	113	64	121	99	108	76	57	69	39	61

Варіант №5 Загальний холестерин (мг/дл)

220	174	211	261	184	210	154	228	228	180	171	187	162	216	232
204	146	248	214	207	152	197	201	203	217	175	185	177	193	145

Варіант №6 Індекс маси тіла (кг/м²)

29,8	27,9	29	25,4	27,1	24	32,3	21	20,5	25,5	26,7	35,4	30,2	28,2	22,5
22,1	22	29,1	29,7	29,5	22,2	20,3	26,3	27,9	26,8	29,1	29,3	31	29,1	19,5

Варіант №7 Температура тіла (°C)

36,5	36,7	37,5	36,5	36,7	36,7	37,1	36,7	36,9	36,2	36,8	37,3	37	36,7	37,1
37,1	37	37,1	37,2	36,5	36,6	37,1	37,5	36,8	36,4	37,3	36,7	36,6	35,8	36,7

Варіант №8 Частота дихання (вдихів/хв)

18	13	19	13	14	13	13	11	15	15	15	16	11	17	18	13
15	18	16	21	19	18	12	16	18	19	16	22	17	14	15	18

Варіант №9 Гемоглобін (г/дл)

14,6	15	14,7	13,8	15,3	12,2	13,3	14,2	15	14,6	12,7	15,3	15,8	16,1	14,4
16,6	14,8	14,4	14,1	13,2	13,7	14,1	13,6	14	11,4	15,2	11,9	14,2	14,1	13,9

Варіант №10 Кількість тромбоцитів (10³/мкл)

194	281	260	276	192	296	174	242	170	272	179	296	146	163	213
232	318	234	264	174	207	180	214	230	189	253	266	138	163	182

Варіант №11 Креатинін (мг/дл)

0,68	0,8	0,84	0,58	0,73	0,65	1,12	0,75	0,65	0,99	0,67	0,48	0,72	0,65	0,56
0,87	0,85	0,85	1,6	0,7	0,39	0,68	0,69	0,91	1,07	0,86	0,44	0,71	0,75	0,54

Варіант №12 С-реактивний білок (мг/л)

0,63	3,76	1,13	0,24	1,4	0,25	0,85	4,7	0,79	0,89	2,61	5,86	0,54	5,75	0,2
1,74	0,26	0,83	3,53	0,85	1,23	0,14	0,85	0,42	0,37	0,41	0,3	1,14	0,16	0,27

Варіант №13 Тривалість перебування в лікарні (днів)

2	5	3	5	10	8	11	6	2	6	7	5	5	3	5
5	3	3	1	5	5	3	2	6	8	4	7	6	5	7

Варіант №14 Оцінка болю (0–10)

2	10	8	6	9	1	8	0	6	8	0	6	5	1	0
0	6	6	8	8	4	9	2	9	5	6	2	3	4	7

Варіант №15 Насичення крові киснем (SpO₂, %)

98	93,1	99	97,9	95	98,7	95,6	99,5	96,9	96,5	96,4	93,9	98,2	96,9	98,9
96,9	95,2	94,4	97,3	100,8	95,5	98,5	98	98,6	97,3	95,2	98,3	97,2	99	94,9

ПРАКТИЧНА РОБОТА 7

Критерії Фішера і Стюдента

Питання, які розглядаються:

1. Критерій Фішера.
2. Критерії Стюдента.
3. Критерії Стюдента (Уелча)

В роботі розглядаються нормально розподілені вибірки. За допомогою параметричних методів перевіряються гіпотези про рівність дисперсій та середніх заданих вибірок.

1. Критерій Фішера

Критерій Фішера перевіряє гіпотезу про рівність дисперсій. Якщо гіпотеза правильна, тобто дисперсії рівні, то для перевірки рівності середніх застосовується **тест Стюдента**. Якщо дисперсії різняться – **тест Уелча**.

Припускаємо, що ці дві вибірки розподілені за нормальним законом. Нехай розміри вибірок n_1 і n_2 . Формулюємо гіпотези, нульову про рівність дисперсій та альтернативну.

$$H_0: S_1^2 = S_2^2,$$

$$H_1: S_1^2 \neq S_2^2.$$

Для перевірки гіпотез потрібно виконати такі дії. Знаходимо відношення більшої дисперсії до меншої

$$F_{\text{exp}} = \frac{S_1^2}{S_2^2} > 1.$$

Потім визначаємо табульоване значення, наприклад, з таблиці

$$F_{\text{theory}} = F\left(\frac{\alpha}{2}, k_1, k_2\right),$$

де α – рівень значущості,

$k_1 = n_1 - 1, k_2 = n_2 - 1$ – ступені свободи двох вибірок.

Якщо $|F_{\text{exp}}| < F_{\text{theory}}$, то гіпотеза H_0 приймається, робиться висновок, що дисперсії рівні. Якщо $|F_{\text{exp}}| \geq F_{\text{theory}}$, то гіпотеза H_0 відкидається, робиться висновок, що дисперсії відрізняються.

В Excel є два варіанти виконати *тест Фішера* для двох вибірок (табл. 7.1).

Таблиця 7.1

B1	5,0	6,1	5,5	4,8	5,3	6,2	5,6	4,9	5,8	5,1	6,0	5,4	5,7	5,2	6,3	4,7	5,9	5,0	6,4	5,3
B2	5,2	6,5	5,8	4,9	5,4	6,8	5,9	4,7	6,3	5,3	6,6	5,7	6,0	5,1	6,9	4,6	6,1	5,0	7,0	5,5

Розглянемо перший варіант виконання *тесту Фішера* за допомогою вбудованої функції F.TEST (рис.7.1).

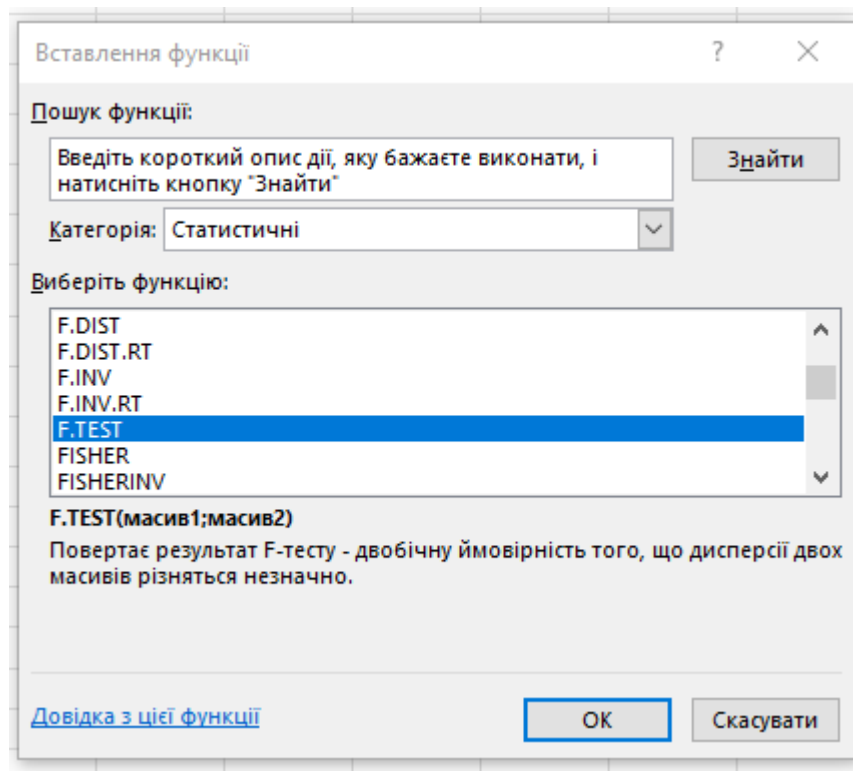


Рисунок 7.1

Після виклику функції відкривається вікно, в якому потрібно ввести діапазони вибірок **B1** і **B2** і натиснути **ОК** (рис.7.2).

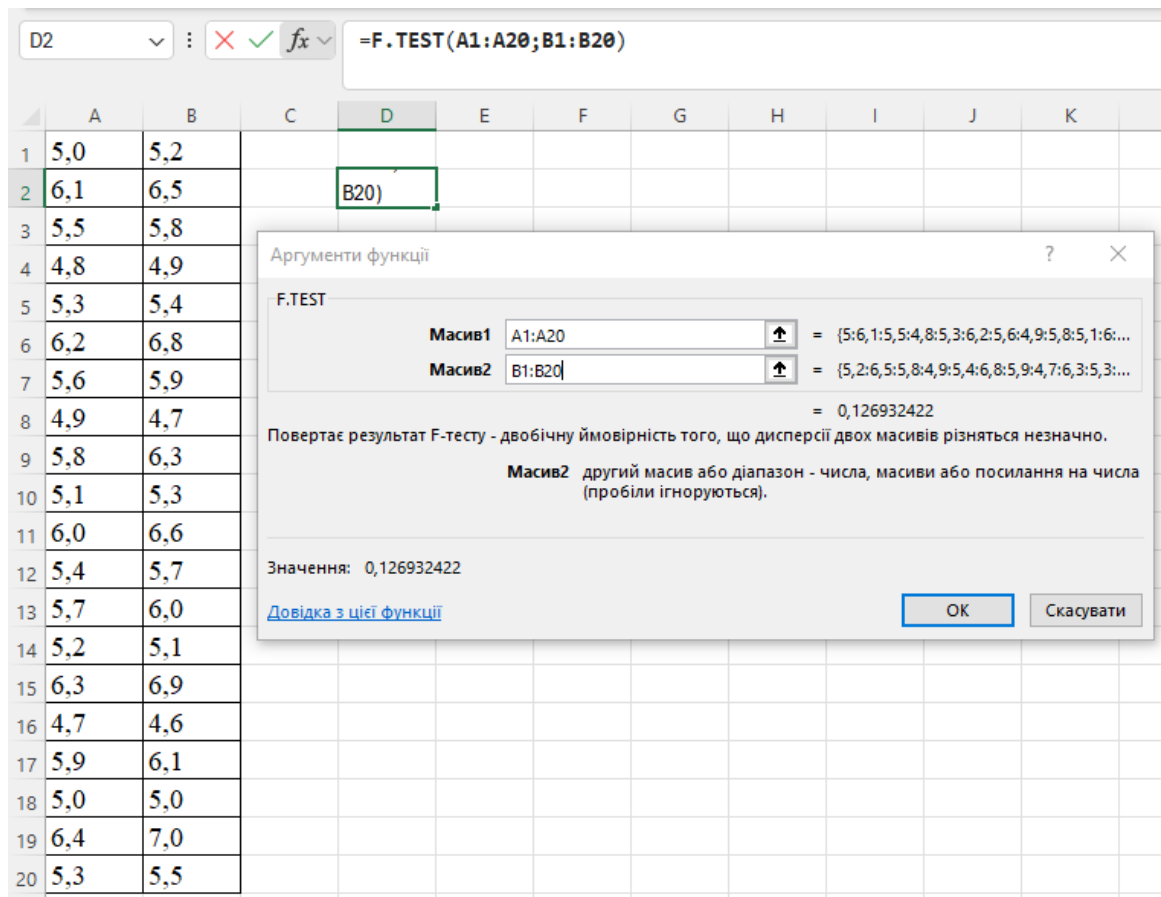


Рисунок 7.2

У результаті в осередку з'явиться розраховане значення ймовірності того, що дисперсії двох вибірок відрізняються статистично не значуще (рис.7.3).

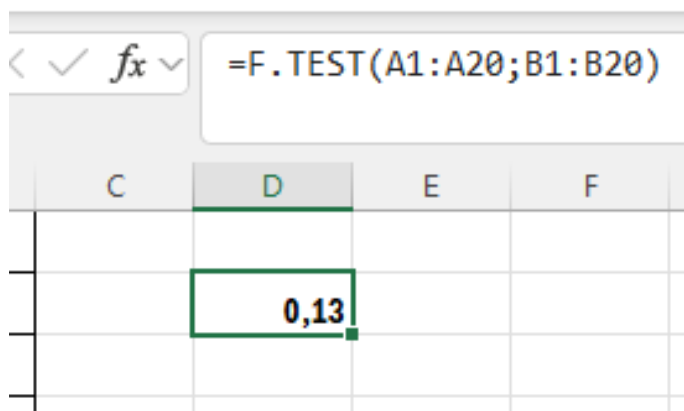


Рисунок 7.3

Оскільки значення **0,13** більше рівня значущості **0,05**, приймається нульова гіпотеза – дисперсії двох вибірок рівні.

Розглянемо другий варіант виконання *тесту Фішера*. Обираємо пункт меню **Дані**, потім **Data Analysis** та **F-Test Two-Sample for Variances** (двохвибірковий F-тест для дисперсії) (рис.7.4).

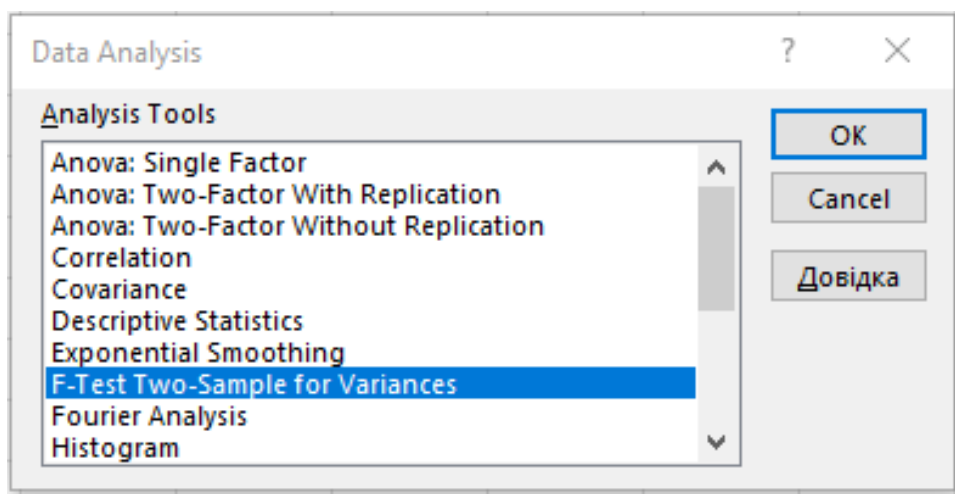


Рисунок 7.4

Відкриється вікно для введення даних. У першому полі потрібно вказати діапазон вибірки з більшою дисперсією, у другому полі - з меншою. Також вказуємо рівень значущості та адресу комірки, в якій будуть розміщуватися вихідні дані, та натискаємо **ОК** (рис.7.5).

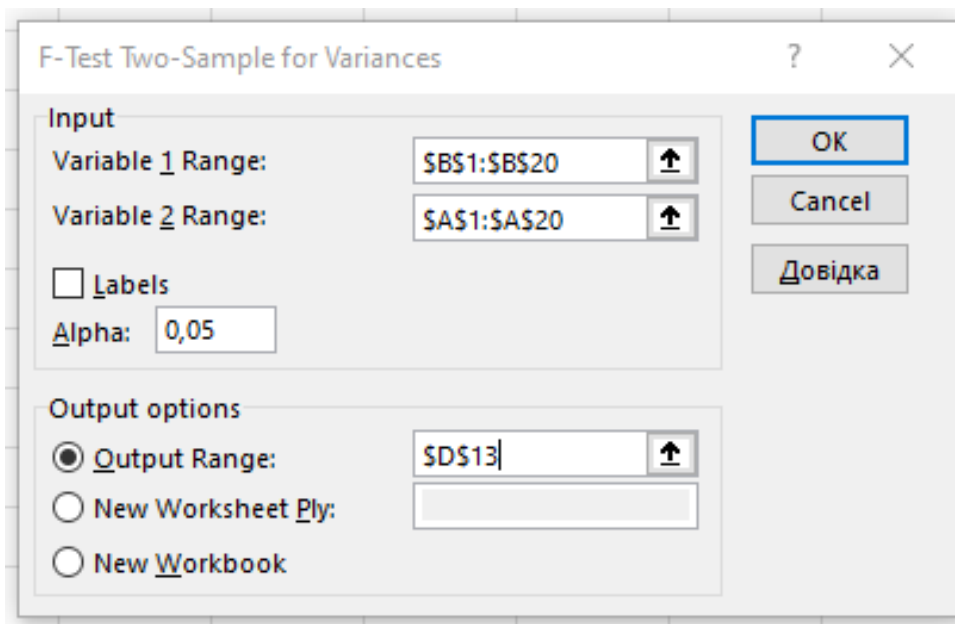


Рисунок 7.5

В результаті отримуємо вихідну таблицю (рис.7.6), яка містить усі результати розрахунку: середні вибірки, дисперсії, обсяги вибірок, ступінь свободи, розраховане значення критерію F_{exp} , половину значення ймовірності того, що дисперсії двох вибірок різняться, а також **критичне теоретичне значення критерію F_{theory}** .

F-Test Two-Sample for Variances		
	Variable 1	Variable 2
Mean	5,765	5,51
Variance	0,558184	0,272526
Observations	20	20
df	19	19
F	2,048185	
P(F<=f) one-tail	0,063466	
F Critical one-tail	2,168252	

Рисунок 7.6

Оскільки $2,048 < 2,168$, тобто $|F_{\text{exp}}| < F_{\text{theory}}$, приймаємо нульову гіпотезу.

2. Критерій Стьюдента

Критерій Стьюдента застосовується для перевірки рівності середніх значень двох вибірок. При цьому застосовуються такі обмеження: вибірки мають бути незалежними та нормально розподіленими. При перевірці гіпотези про рівність середніх фактично перевіряється чи розрізняються розподіли. Відмінність середніх означає, що розподіли зміщені щодо одне одного, тобто різні за центром.

Після проведення **тесту Фішера** вибирають один із двох типів тесту: **тест Стьюдента** для вибірок з рівними дисперсіями або **тест Сьюдента (Уелча)**, якщо дисперсії різняться. Для перевірки гіпотези за допомогою **тесту Стьюдента** для вибірок з рівними дисперсіями формулюються нульова та альтернативна гіпотези:

$$H_0: \overline{X}_1 = \overline{X}_2,$$

$$H_1: \overline{X}_1 \neq \overline{X}_2.$$

Потім розраховуємо статистику

$$t_{\text{exp}} = \frac{\overline{X}_1 - \overline{X}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

де n_1, n_2 – об’єми виборок,

$$S_p = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2} - \text{об'єднана оцінка дисперсії},$$

Потім визначається табульоване значення, наприклад, з таблиці

$$t_{\text{theory}} = t(\alpha, k), \text{ де } k = n_1 + n_2 - 2.$$

Якщо $|t_{\text{exp}}| < t_{\text{theory}}$, то гіпотеза H_0 приймається, робиться висновок, що середні рівні. Якщо $|t_{\text{exp}}| \geq t_{\text{theory}}$, то гіпотеза H_0 відкидається, робиться висновок, що середні відрізняються.

В програмі **Excel** є два варіанта виконати *тест Стьюдента* для двох вибірок з однаковими дисперсіями. Дані для тесту беремо із таблиці 7.1, як і раніше, за допомогою *тесту Фішера* було показано, що дисперсії цих вибірок не відрізняються.

Розглянемо перший варіант виконання тесту за допомогою вбудованої функції **T.TEST** (рис.7.7). Після виклику функції відкривається вікно, в якому потрібно ввести діапазони вибірок **B1** і **B2**, у полі **Боки** число **2** (тест двосторонній), у полі **Тип** також число **2** (тип тесту з рівними дисперсіями) і натиснути **ОК** (рис.7.8).

В результаті в клітинці з’явиться розраховане значення ймовірності того, що середні двох вибірок відрізняються статистично не значуще (рис.7.9). Оскільки значення **0,22** більше рівня значущості **0,05**, приймається нульова гіпотеза про рівність середніх двох виборок.

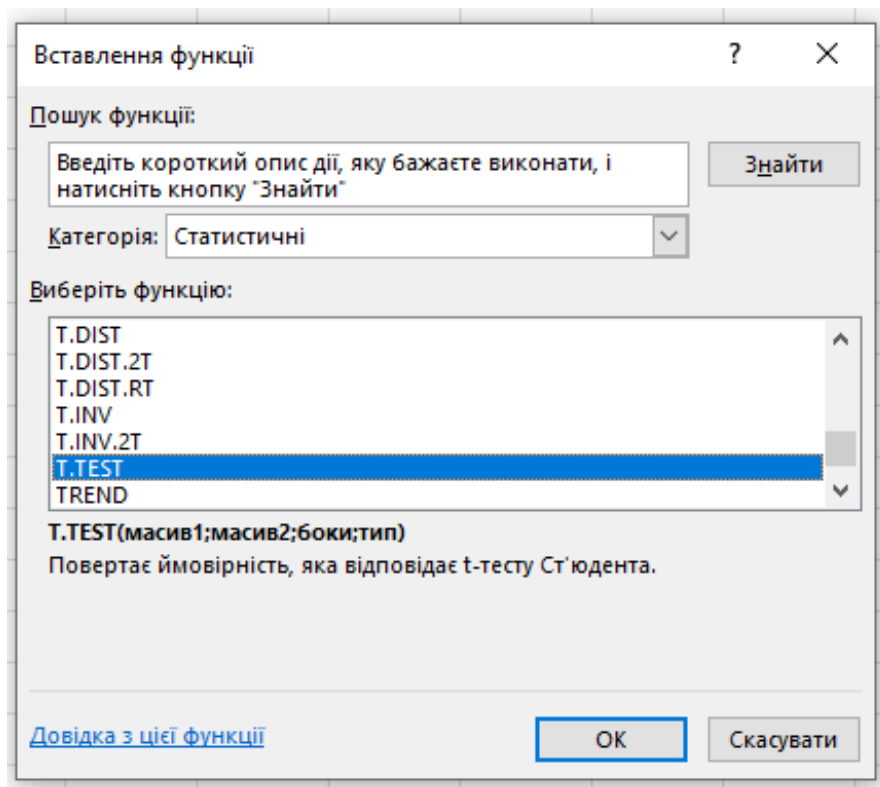


Рисунок 7.7

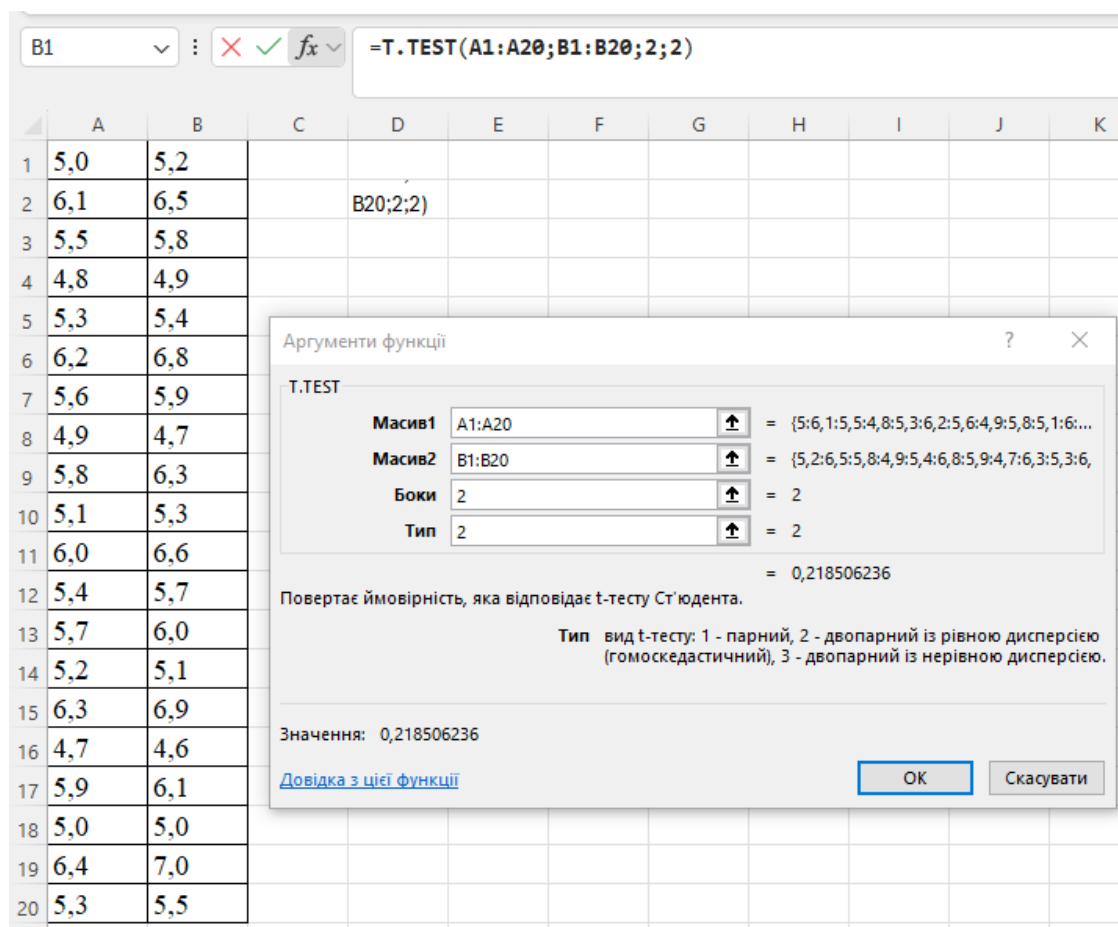


Рисунок 7.8

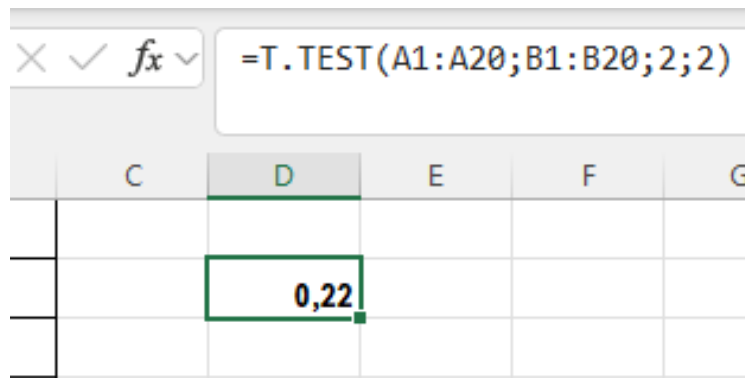


Рисунок 7.9

Розглянемо другий варіант виконання тесту (рис.7.10). Обираємо пункт меню **Дані**, потім **Data Analysis** та **T-Test: Two-Sample Assuming Equal Variances** (Двовибірковий t-тест із рівними дисперсіями). Відкривається вікно для вводу даних (рис.7.11). У першому та другому полях треба вказати діапазони комірок першої та другої вибірок відповідно. Також вказуємо рівень значущості і адресу комірки, в якій будуть розташовані вихідні дані, та натискаємо **ОК**.

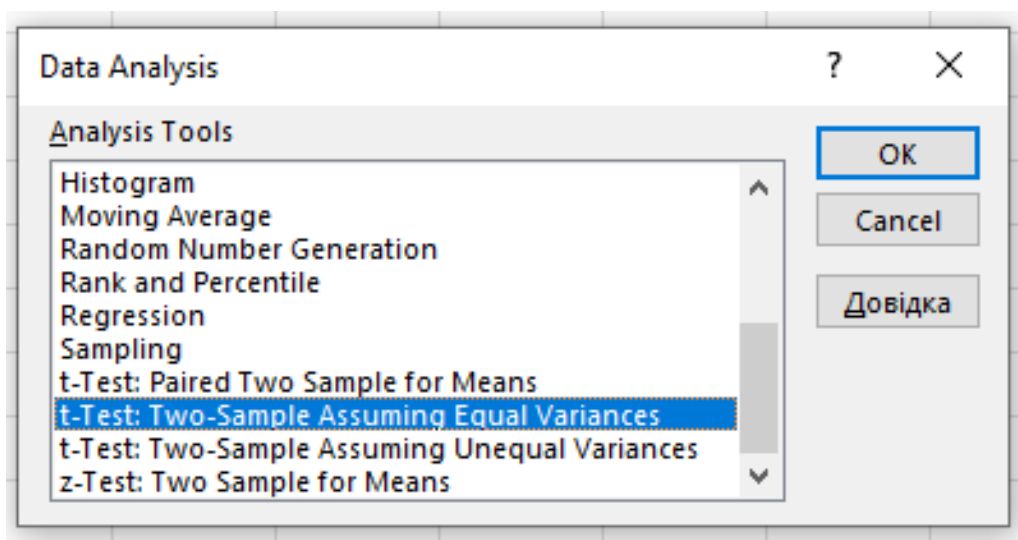


Рисунок 7.10

В результаті отримуємо вихідну таблицю, яка містить всі результати розрахунку (рис.7.12): середнє вибірок, дисперсії, об'єми вибірок, об'єднану оцінку дисперсій, гіпотетичну середню різницю, число ступенів свободи, **розраховане значення критерію (t_{Stat})**, половину значення ймовірності того, що середнє двох вибірок не розрізняються, критичне теоретичне значення критерію

для односторонньої області, значення ймовірності того, що середнє двох виборок не розрізняються а також критичне теоретичне значення критерію для двосторонньої області.

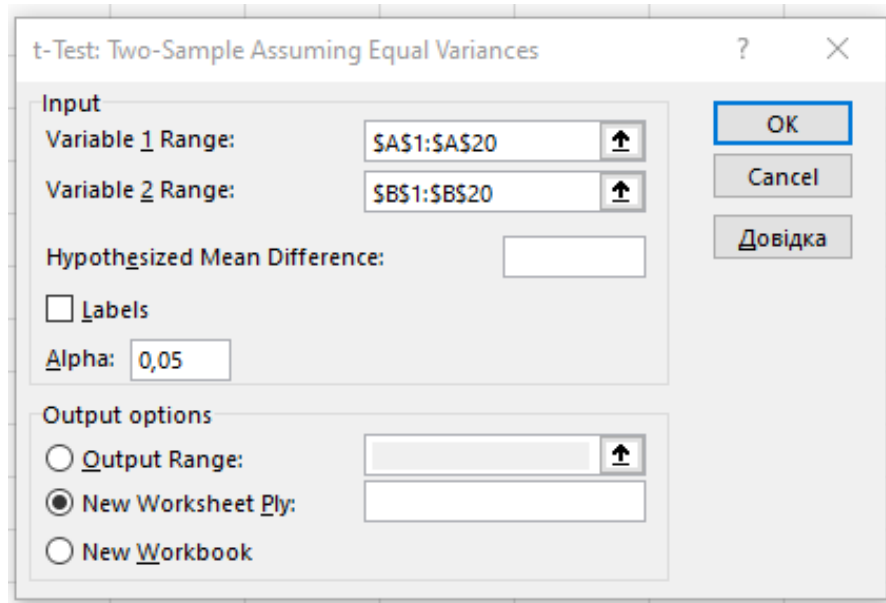


Рисунок 7.11

Оскільки $|-1,25| < 2,02$, тобто $|t_{\text{exp}}| < t_{\text{theory}}$, приймаємо нульову гіпотезу

t-Test: Two-Sample Assuming Equal Variances		
	Variable 1	Variable 2
Mean	5,51	5,765
Variance	0,272526	0,558184
Observations	20	20
Pooled Variance	0,415355	
Hypothesized Mean Difference	0	
df	38	
t Stat	-1,25121	
P(T<=t) one-tail	0,109253	
t Critical one-tail	1,685954	
P(T<=t) two-tail	0,218506	
t Critical two-tail	2,024394	

Рисунок 7.12

3. Критерій Стьюдента (Уелча)

Критерій Стьюдента (Уелча) призначений для перевірки гіпотези про рівність середніх двох незалежних, нормально розподілених вибірок у разі, коли встановлено (наприклад, за допомогою **тесту Фішера**), що дисперсії двох вибірок розрізняються.

Спочатку формулюються нульова та альтернативна гіпотези:

$$H_0: \overline{X}_1 = \overline{X}_2,$$

$$H_1: \overline{X}_1 \neq \overline{X}_2.$$

Потім розраховується статистика

$$t_{\text{exp}} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}},$$

де n_1, n_2 – об'єми вибірок,

Наступний крок - обчислення **числа ступенів свободи**. Це число визначається за формулою

$$k = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Потім визначається значення

$$t_{\text{theory}} = t(\alpha, k).$$

Оскільки число ступенів може бути дробовим числом, визначити значення t_{theory} з таблиць важко. Одним з варіантів знайти t_{theory} є використання вбудованої функції Excel для двостороннього тесту **T.INV.2T()**. Якщо $|t_{\text{exp}}| < t_{\text{theory}}$, то гіпотеза H_0 приймається, робиться висновок, що середні рівні. Якщо

$|t_{\text{exp}}| \geq t_{\text{theory}}$, то гіпотеза H_0 відкидається, робиться висновок, що середні відрізняються.

Далі будуть показані два варіанти за допомогою програми **Excel** виконати **тест Стьюдента (Уелча)** для двох вибірок з нерівними дисперсіями.

У таблиці 7.2 задані значення двох вибірок.

Таблиця 7.2

B1	5,8	6,3	5,9	6,0	5,7	6,1	6,2	5,6	6,4	5,5
B2	3,5	4,8	7,9	5,2	8,8	6,0	2,1	4,2	10,1	7,3
B1	6,0	5,9	6,1	5,8	6,2	5,7	6,0	6,1	6,3	5,8
B2	5,5	4,5	8,5	3,9	9,2	4,0	6,7	2,9	9,5	6,4

Насамперед за допомогою **тесту Фішера** (п.2) перевіряється гіпотеза про рівність дисперсій двох вибірок, наприклад, за допомогою пакета **Data Analysis**.

F-Test Two-Sample for Variances		
	Variable 1	Variable 2
Mean	6,05	5,97
Variance	5,61	0,061157895
Observations	20	20
df	19	19
F	91,7298	
P(F<=f) one-tail	8,9E-15	
F Critical one-tail	2,16825	

Рисунок 7.13

Оскільки $91,73 < 2,168$ (рис.7.13), тобто $|F_{\text{exp}}| > F_{\text{theory}}$, нульова гіпотеза відкидається, тобто дисперсії відрізняються. Це дає підставу застосувати **тест Стьюдента (Уелча)**.

Розглянемо перший варіант виконання тесту за допомогою вбудованої функції **T.TEST** (рис.7.14). Після виклику функції відкривається вікно, в якому потрібно ввести діапазони вибірок **B1** і **B2**, у полі **Боки** число **2** (тест двосторонній), у полі **Тип** число **3** (тип тесту з нерівними дисперсіями) і натиснути **ОК** (рис.7.15).

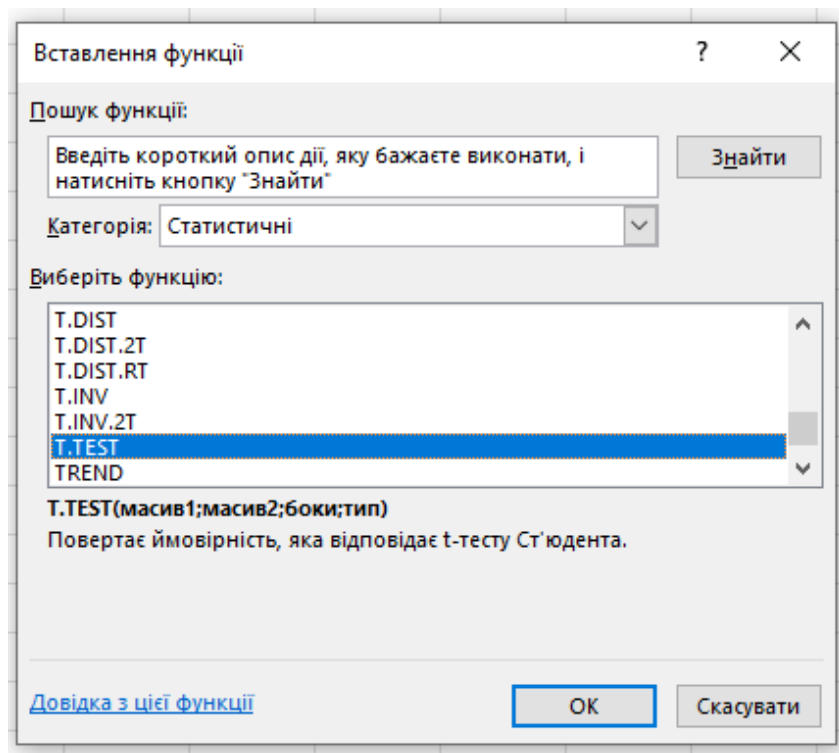


Рисунок 7.14

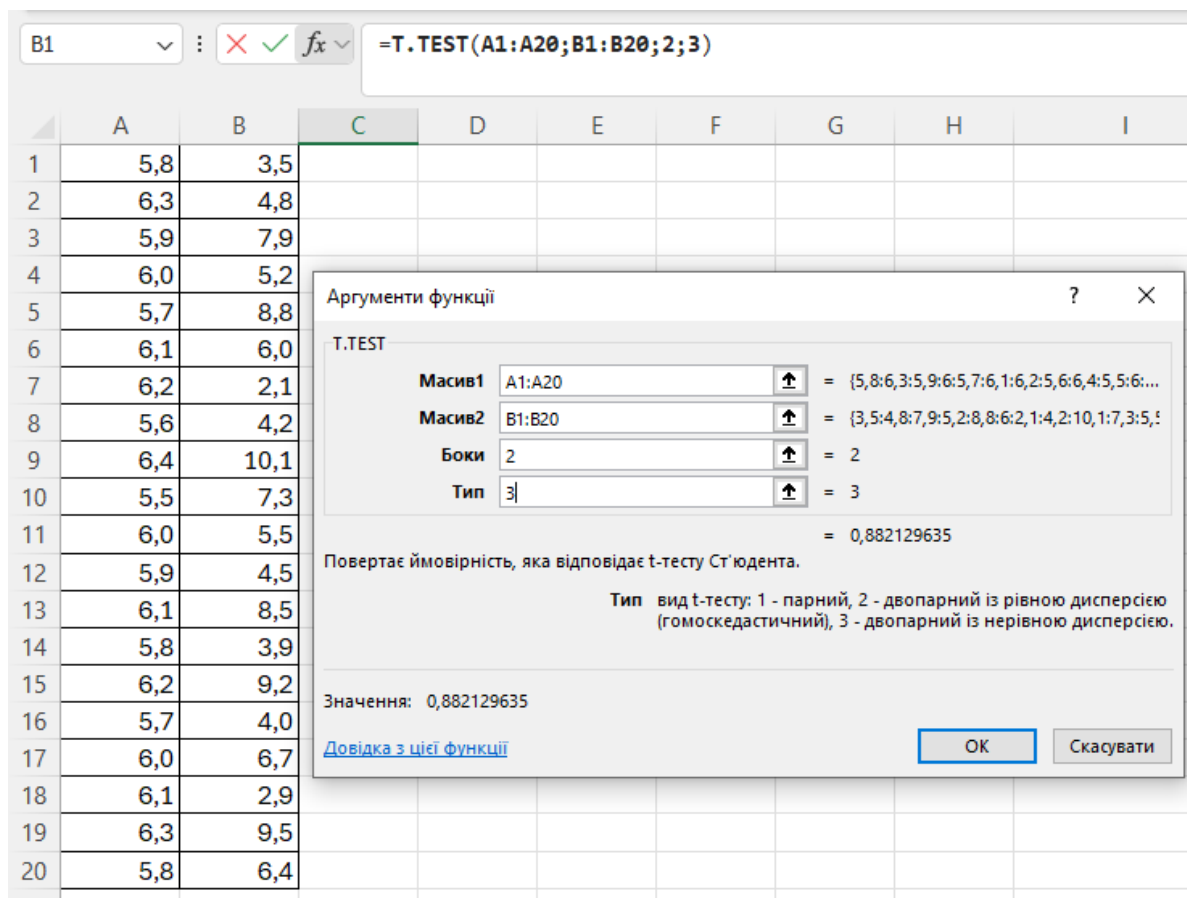


Рисунок 7.15

В результаті в осередку з'явиться розраховане значення ймовірності того, що середні двох вибірок відрізняються статистично не значуще (рис.7.16).

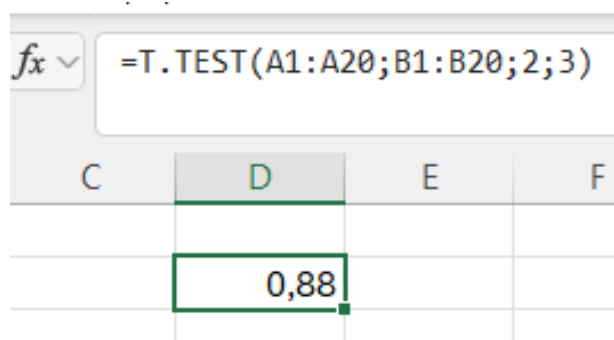


Рисунок 7.16

Оскільки значення **0,88** більше рівня значущості **0,05**, приймається нульова гіпотеза про рівність середніх двох вибірок.

Розглянемо другий варіант виконання тесту. Обираємо пункт меню **Дані**, потім **Data Analysis** та **T-Test: Two-Sample Assuming Unequal Variances** (Двовибірковий t-тест з нерівними дисперсіями) (рис.7.17).

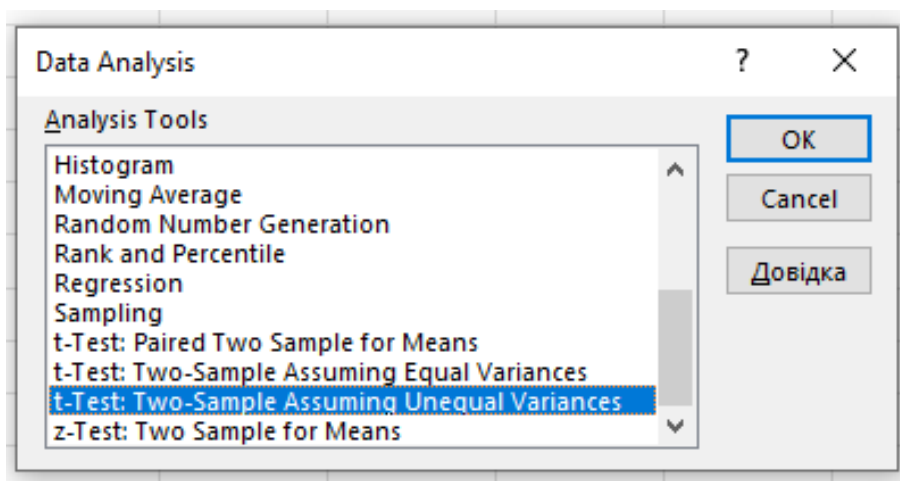


Рисунок 7.17

Відкриється вікно для введення даних (рис.7.18). У першому і другому полях потрібно вказати діапазони осередків першої та другої вибірок відповідно. Також вказуємо рівень значущості, та адресу осередку, в якій будуть розміщуватися вихідні дані та натискаємо **ОК**.

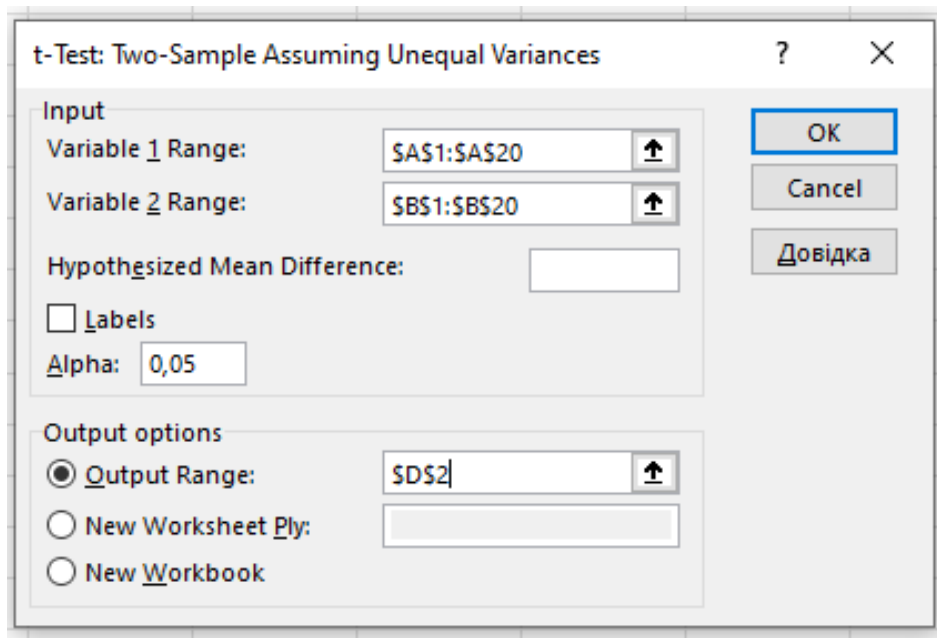


Рисунок 7.18

В результаті отримуємо вихідну таблицю, яка містить усі результати розрахунку (рис.7.19): середні вибірки, дисперсії, обсяги вибірок, гіпотетичну середню різницю, число ступенів свободи, **розраховане значення критерію (t_{Stat})**, половину значення ймовірності того, що середні двох вибірок не різняться, критичне теоретичне значення критерію для односторонньої області, значення ймовірності того, що середні двох вибірок не відрізняються, а також критичне теоретичне значення критерію для двосторонньої області.

t-Test: Two-Sample Assuming Unequal Variances		
	Variable 1	Variable 2
Mean	5,97	6,05
Variance	0,06116	5,61
Observations	20	20
Hypothesized Mean Difference	0	
df	19	
t Stat	-0,1502	
P(T<=t) one-tail	0,44108	
t Critical one-tail	1,72913	
P(T<=t) two-tail	0,88216	
t Critical two-tail	2,09302	

Рисунок 7.19

Оскільки $|-1,15| < 2,09$, тобто $|t_{\text{exp}}| < t_{\text{theory}}$, приймаємо нульову гіпотезу.

Практичне завдання

У завданнях 1 та 2 для кожного варіанту задані дві виборки. Потрібно:

- 1) перевірити гіпотезу про рівність дисперсій.
- 2) використовуючи результат, отриманий у п.1, перевірити гіпотезу про рівність середніх.

Завдання 1.

Варіант 1		Варіант 2		Варіант 3		Варіант 4		Варіант 5		Варіант 6		Варіант 7		Варіант 8	
-		27,4	33,6	12,9	12,7	8,3	1,0	33,2	31,8	-4,6	0,2	4,3	4,9	-6,3	-4,3
12,6	0,1														
-8,5	1,1	28,4	31,2	13,9	11,7	2,2	7,0	27,5	29,8	2,7	-2,2	1,5	4,9	-7,7	-5,9
-1,5	-0,9	28,9	34,5	14,3	11,4	7,1	7,7	25,5	29,0	-2,3	0,2	3,5	5,2	-7,1	-6,4
-9,3	-6,8	25,6	31,4	15,5	12,8	13,7	7,9	29,4	34,1	-2,5	-1,9	3,9	4,1	-9,0	-4,8
-2,5	0,6	30,5	30,8	13,1	13,5	3,4	9,8	30,0	31,0	-8,9	4,2	3,3	5,5	-8,2	-6,3
-4,6	0,4	26,5	31,7	14,6	11,5	6,9	4,1	24,2	34,5	-1,7	1,8	4,1	5,3	-8,8	-7,6
-5,4	2,7	25,6	32,4	10,1	10,7	-0,6	5,5	29,4	31,7	0,7	1,0	4,9	4,7	-8,1	-6,0
0,7	-0,6	30,0	31,7	14,3	11,0	9,4	4,4	24,5	32,7	-4,5	-6,0	4,1	5,1	-7,7	-6,4
-1,5	5,5	29,4	34,9	14,6	16,5	9,3	6,6	24,7	35,4	1,6	-1,6	4,9	5,0	-8,6	-7,0
-		29,0	33,0	18,1	9,8	3,0	0,2	30,2	32,7	-4,6	0,7	3,3	3,8	-7,3	-5,3
10,4	0,1														
-5,5	-0,1	26,4	30,8	15,0	10,5	1,7	11,3	28,1	26,3	-3,3	1,5	5,2	4,2	-7,1	-5,4
-2,6	3,7	29,2	33,6	15,7	9,7	3,8	9,4	26,9	33,1	-4,4	-1,7	3,8	4,3	-	-5,4
														10,1	
-1,6	-1,7	27,3	30,8	14,3	13,7	0,6	10,5	29,5	32,6	-4,3	4,9	3,1	5,1	-9,3	-3,0
-9,1	2,9	25,3	35,4	14,1	15,2	7,8	6,8	27,6	33,0	-6,2	0,0	2,6	5,4	-8,8	-6,1
-4,7	-2,7	26,5	31,1	15,6	12,0	9,8	8,6	31,3	33,8	-2,5	-3,5	3,1	3,7	-8,5	-8,3
-3,4	-1,1	25,5	32,3	12,3	16,2	6,1	5,5	29,1	33,3	-2,7	-3,6	3,3	4,0	-7,9	-6,2
-4,8	8,3	28,9	33,3	14,2	10,9	-1,6	13,8	23,4	29,3	-2,4	0,2	3,7	4,6	-7,7	-6,2
-4,2	1,4	29,4	31,8	13,6	15,4	12,0	2,5	30,1	35,4	-0,8	1,2	0,7	4,8	-9,1	-5,1
-9,2	-1,9	25,9	31,5	17,7	12,2	7,5	2,0	22,2	34,1	-6,8	0,4	2,3	4,2	-8,2	-4,5
-4,3	-0,8	27,3	33,4	14,5	10,1	6,5	6,9	30,6	33,9	0,1	-2,0	2,5	4,5	-8,7	-5,3

Варіант 9		Варіант 10		Варіант 11		Варіант 12		Варіант 13		Варіант 14		Варіант 15		Варіант 16	
3,6	5,4	19,3	23,0	6,7	0,3	-6,3	-3,2	14,5	16,8	- 4,8	1,4	9,9	16,9	18,6	20,7
7,2	5,4	17,5	25,6	-6,5	-5,6	-3,8	-5,4	16,2	19,9	0,9	-4,4	18,3	12,5	16,1	20,3
7,9	5,8	18,6	24,3	-3,8	0,3	-6,8	- 13,1	14,5	18,1	- 3,2	8,3	13,4	4,3	17,5	22,7
12,1	5,9	16,7	22,9	0,3	2,2	-6,8	- 11,3	15,1	21,2	1,4	1,2	15,0	14,8	16,3	19,1
8,7	4,4	19,0	22,4	-1,2	0,2	-6,4	-8,1	12,3	24,1	0,2	3,7	14,3	16,7	20,6	24,2
5,5	6,4	17,9	23,7	2,1	4,0	-5,5	-3,0	17,5	22,7	- 1,1	2,0	11,7	19,7	21,1	19,7
7,6	7,6	18,1	24,2	-0,8	3,9	-7,3	- 11,6	13,9	15,5	- 4,0	1,0	15,4	13,6	16,7	19,7
6,1	5,6	20,2	23,5	1,9	1,4	4,4	-6,6	10,7	19,0	- 4,9	0,2	8,7	14,9	18,5	21,9
7,8	11,8	15,8	23,1	0,5	5,3	-7,7	-7,5	15,4	18,3	- 3,5	4,9	14,0	17,9	19,6	17,8
10,6	5,3	17,8	23,9	0,6	5,9	-5,5	-7,1	19,2	20,1	- 4,5	4,7	14,2	4,5	16,5	22,2
12,7	6,4	17,8	26,2	-3,0	0,5	0,9	-3,3	15,8	15,2	- 0,7	3,7	15,9	5,8	20,6	20,8
6,1	2,0	19,6	24,4	3,4	5,8	-4,5	-2,8	14,4	20,0	- 2,1	6,8	18,1	15,4	16,8	22,8
5,6	3,0	20,6	24,6	4,9	1,6	-4,4	-5,9	17,1	21,0	- 0,9	5,3	11,5	16,4	16,1	22,0
9,3	10,9	18,7	23,7	2,0	2,4	- 10,8	- 11,9	13,9	18,7	- 1,9	2,0	17,5	8,8	17,1	24,0
11,7	9,0	18,4	24,4	9,6	-1,8	-2,3	-3,4	11,4	20,0	- 1,1	3,6	14,6	6,1	18,1	20,4
7,1	12,9	20,6	23,9	0,9	3,7	-3,2	-2,7	15,7	18,6	- 1,4	2,4	14,3	18,9	17,1	23,8
7,0	5,8	19,7	24,8	3,2	3,0	-4,7	-9,0	13,5	16,9	- 1,6	9,2	15,2	14,4	21,2	22,2
2,3	7,7	19,1	23,6	-2,7	-1,7	-7,8	-9,1	16,7	16,7	- 0,6	3,9	15,8	15,1	18,8	21,2
8,0	4,2	20,6	24,7	-5,6	2,9	-6,3	-1,8	16,9	20,8	- 5,3	-3,0	8,1	18,3	16,0	22,6
6,9	10,4	19,2	24,0	-5,1	-4,3	-4,9	-9,1	12,5	23,5	- 4,2	9,9	18,7	15,2	19,5	20,7

Завдання 2.

Варіант 1		Варіант 2		Варіант 3		Варіант 4		Варіант 5		Варіант 6		Варіант 7		Варіант 8	
-10,0	-1,2	10,8	0,5	-5,1	2,4	22,1	13,8	19,3	23,8	-0,3	4,3	16,2	7,8	1,8	-19,7
-7,1	4,9	3,1	4,6	-6,6	-1,3	17,6	6,5	20,0	21,3	-0,7	2,7	13,8	21,7	3,9	-4,1
-5,0	0,1	1,5	8,1	-4,4	-0,5	16,7	19,2	19,5	24,8	-0,7	6,1	15,1	9,3	4,4	-3,5
3,6	5,4	9,7	4,5	-5,3	4,3	18,2	17,0	20,6	24,7	-1,3	3,5	16,2	24,2	-0,3	4,6
0,7	0,7	8,7	19,9	-3,2	-5,8	15,9	18,3	20,7	23,3	0,0	1,8	14,8	3,3	2,7	9,1
-5,4	0,5	7,9	10,9	-7,0	3,4	21,0	14,1	21,5	24,2	-0,1	1,7	16,1	6,6	1,3	2,8
-1,7	8,2	3,0	0,2	-3,5	5,0	22,5	16,2	20,6	25,4	0,0	3,7	9,1	14,5	8,2	-16,6
-3,8	-0,3	8,3	13,4	-5,1	-2,1	19,8	12,6	16,0	19,7	0,6	4,2	19,2	17,0	3,0	-2,5
-8,5	-6,0	4,7	0,2	-6,2	5,9	14,3	22,0	17,8	20,1	-1,6	3,8	10,6	26,3	1,3	-1,3
-7,7	-8,4	0,9	22,3	-4,7	3,3	24,0	9,1	18,7	24,1	0,9	2,5	16,6	17,3	6,5	-10,8
-8,6	0,5	3,0	1,9	-6,7	4,4	20,8	8,6	22,6	27,5	1,0	4,6	16,4	20,8	1,6	7,6
-4,9	-6,5	1,3	7,6	-6,3	-0,1	20,1	14,2	17,8	21,7	0,2	0,4	15,7	8,6	-2,6	-2,6
-4,0	5,1	7,8	12,3	-6,2	-3,6	14,9	15,6	18,5	18,7	1,0	1,8	14,1	13,6	2,5	-7,0
-11,4	2,0	8,6	5,1	-4,7	-1,1	19,2	25,6	18,7	25,9	0,3	3,1	18,7	9,6	1,2	0,0
-0,5	-2,0	1,9	3,5	-5,3	-8,3	19,6	19,6	20,5	21,0	2,4	4,2	17,8	14,6	-0,7	-7,5
-4,4	-4,3	4,6	12,8	-9,5	-2,9	19,8	17,3	18,5	20,7	1,6	3,6	18,5	16,7	4,9	1,6
-5,2	-0,4	6,2	-0,8	-2,1	2,8	21,1	10,8	19,6	14,4	1,3	1,1	13,9	10,9	4,6	7,6
-5,8	4,1	1,6	18,6	-1,5	-2,9	17,1	10,4	16,5	21,6	-1,2	4,7	19,4	19,5	4,7	-0,9
1,5	5,9	7,9	1,2	-6,4	1,7	18,1	10,3	19,3	23,9	0,4	5,3	12,7	21,0	12,7	-8,3
-5,3	-8,0	2,0	1,5	-3,4	8,0	17,3	18,4	21,3	18,8	1,2	5,8	16,8	1,7	2,2	-10,6
Варіант 9		Варіант 10		Варіант 11		Варіант 12		Варіант 13		Варіант 14		Варіант 15		Варіант 16	
19,4	30,1	25,2	24,4	24,8	17,5	18,6	21,2	24,4	35,2	10,3	32,1	-7,0	-3,2	10,3	0,2
17,7	19,1	22,7	21,4	18,4	20,9	17,4	21,2	22,2	25,1	11,5	17,8	-8,5	-6,2	3,1	4,6
18,5	29,6	21,0	25,4	19,1	14,3	18,0	25,3	25,4	23,0	6,0	-1,0	-6,3	-6,7	1,5	8,1
17,7	25,3	25,6	13,1	18,5	18,9	18,5	21,2	22,9	32,1	7,3	33,9	-6,1	-5,9	9,2	4,5
18,3	29,6	23,3	18,7	20,8	24,0	19,9	24,6	22,5	31,5	8,1	4,9	-6,0	-1,6	8,7	19,7
20,9	20,8	20,9	32,4	18,5	15,0	19,0	23,4	22,5	32,8	1,5	33,1	-4,7	1,1	7,9	10,9
17,8	25,2	19,9	31,4	18,9	16,5	15,9	24,2	25,9	24,5	16,1	16,9	-8,8	-1,1	3,5	0,2
19,0	19,7	22,2	36,5	17,3	19,6	16,6	22,9	27,7	33,8	7,9	17,2	-10,9	-4,2	8,3	12,4
17,3	23,1	23,1	23,8	19,9	15,8	17,9	24,1	25,6	27,0	10,8	19,6	-7,1	-1,8	4,7	0,2
18,2	17,3	21,8	19,2	20,5	15,9	19,9	19,7	20,3	34,5	8,9	15,3	-5,4	0,2	0,9	22,8
18,4	24,2	21,1	26,1	22,0	16,3	16,5	21,1	23,3	32,6	7,8	21,7	-8,5	-5,4	3,0	1,9
18,4	20,9	22,6	17,8	20,7	17,4	18,5	21,2	29,7	30,4	6,8	13,9	-7,6	-0,5	1,3	7,6
18,8	21,6	26,7	27,1	20,2	15,3	18,1	21,6	25,0	25,8	12,2	5,0	-7,1	-1,5	7,8	12,3
17,2	27,9	23,4	14,8	19,3	29,8	18,3	22,6	23,3	24,4	12,0	19,8	-7,6	-5,8	8,6	5,1
18,6	14,6	23,8	25,5	22,5	14,7	19,8	23,6	24,5	26,5	10,9	18,9	-6,3	-2,0	1,9	3,5
21,6	20,6	22,3	10,8	16,5	17,5	20,0	23,9	21,4	25,0	14,4	16,9	-7,1	2,3	4,1	12,5
17,6	20,6	23,5	21,0	18,2	25,4	18,8	20,1	23,7	31,2	12,6	26,4	-9,8	-4,2	6,2	-0,8
20,4	26,0	22,5	31,6	18,3	18,7	16,4	23,0	22,5	28,8	8,9	20,4	-3,9	-0,8	1,6	19,6
22,8	29,2	24,2	9,8	17,6	18,9	19,8	21,7	23,4	30,9	10,8	21,0	-6,0	-4,3	7,4	1,2
18,7	23,4	22,1	23,2	19,9	10,9	20,1	21,1	25,4	29,3	9,4	18,1	-7,1	-1,1	2,0	1,5

ПРАКТИЧНА РОБОТА 8

Непараметричні методи перевірки гіпотез

Питання, які розглядаються:

1. Критерій Манна-Уїтні.
2. Застосування критерію Манна-Уїтні засобами програми Excel.
3. Критерій Уїлкоксона.
4. Застосування критерію Уїлкоксона засобами програми Excel.

В медичних дослідженнях дані часто не відповідають вимогам параметричних методів. До порушення таких вимог належать: малі обсяги вибірки, закон розподілу випадкової величини, відмінний від нормального, наявність викидів. У таких випадках використовуються непараметричні методи перевірки гіпотез.

1. Критерій Манна-Уїтні

Критерій Манна-Уїтні застосовується для оцінки різниці між двома незалежними вибірками. Цей метод визначає, наскільки сильно розподіли двох вибірок перекриваються. Чим меншою є область значень, що перехрещуються, тим більш ймовірно, що відмінності достовірні, тобто статистично значущі.

Нехай задані дві незалежні вибірки: x_i з об'ємом n_1 та y_i з об'ємом n_2 . Сформулюємо наступні гіпотези:

- H_0 : розподіли однакові,
- H_1 : один розподіл зрушено щодо іншого.

Для перевірки гіпотез за допомогою *критерію Манна-Уїтні* застосовується така схема:

1. Вибірки поєднуються.
2. Кожному значенню надається ранг. Якщо зустрічаються однакові значення, їм присвоюється середнє арифметичне їх рангів.
3. Підсумовуємо ранги для кожної вибірки:

$$R_1 = \sum_{i=1}^{n_1} R(x_i), \quad R_2 = \sum_{i=1}^{n_2} R(x_i).$$

4. Обчислюємо статистики:

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1,$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_2.$$

5. Обираємо найменше значення: $U = \min(U_1, U_2)$.

6. Для малих вибірок ($n_1, n_2 < 20$) із таблиці (Додаток Г) визначаємо теоретичне критичне значення U_{theory} .

Якщо $U > U_{theory}$, приймаємо нульову гіпотезу H_0 про те, що розподіли однакові. Якщо $U \leq U_{theory}$, приймаємо альтернативну гіпотезу H_1 про те, що один розподіл зрушено щодо іншого.

Зауваження. Якщо $n_1, n_2 \geq 20$, дії щодо перевірки гіпотез слід продовжити.

7. Використовуючи вже знайдене значення U обчислюємо

$$Z = \frac{U - \mu_U}{\sigma_U},$$

де

$$\mu_U = \frac{n_1 n_2}{2}, \quad \sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 1)}{12}}.$$

Якщо серед рангів є однакові значення (зв'язки), то для обчислення σ_U додається поправка:

$$\sigma_U = \sqrt{\frac{n_1 n_2}{12}} \left[(n_1 + n_2 - 1) - \frac{\sum_{i=1}^m (t^3 - t)}{(n_1 + n_2)(n_1 + n_2 - 1)} \right],$$

де m – кількість зв'язок,

t – кількість однакових значень у кожній групі зв'язок.

Використовуючи стандартне нормальне відхилення за заданим рівнем значущості α знаходимо Z_{theory} , яке можна порівняти із значенням Z , що спостерігається.

8. Перевіряємо наступні гіпотези:

- якщо для двосторонньої перевірки $|Z| < Z_{theory,\alpha/2}$, (для односторонньої $Z < Z_{theory,\alpha}$ або $Z > -Z_{theory,\alpha}$) приймається нульова гіпотеза H_0 про те, що розподіли однакові.

- якщо для двосторонньої перевірки $|Z| \geq Z_{theory,\alpha/2}$, (для односторонньої $Z \geq Z_{theory,\alpha}$ або $Z \leq -Z_{theory,\alpha}$) приймається альтернативна гіпотеза H_1 про те, що один один розподіл зрушено відносно другого.

2. Застосування критерію Манна-Уїтні засобами програми Excel

	A	B	C
1	№	A	B
2	1	5,39	4,89
3	2	5,12	4,86
4	3	5,32	4,68
5	4	5,76	4,66
6	5	4,93	4,87
7	6	5,39	5,40
8	7	5,13	4,37
9	8	5,77	5,18
10	9	5,04	5,54
11	10	5,04	5,11
12	11	5,78	4,96
13	12	4,76	4,64
14	13	4,76	4,69
15	14	4,76	4,98
16	15	5,14	4,82

Рисунок 8.1

Порівнюються два препарати (вибірка А та вибірка В) за впливом на рівень холестерину (ммоль/л). Кожна вибірка містить 15 значень. Через 30 днів лікування

отримано такі дані (рис.8.1): Помістимо вибірки в один стовпець: спочатку вибірка **A**, потім вибірка **B** (рис.8.2).

	A	B	C	D	E
1	№	A	B		
2	1	5,39	4,89		5,39
3	2	5,12	4,86		5,12
4	3	5,32	4,68		5,32
5	4	5,76	4,66		5,76
6	5	4,93	4,87		4,93
7	6	5,39	5,40		5,39
8	7	5,13	4,37		5,13
9	8	5,77	5,18		5,77
10	9	5,04	5,54		5,04
11	10	5,04	5,11		5,04
12	11	5,78	4,96		5,78
13	12	4,76	4,64		4,76
14	13	4,76	4,69		4,76
15	14	4,76	4,98		4,76
16	15	5,14	4,82		5,14
17					4,89
18					4,86
19					4,68
20					4,66
21					4,87
22					5,40
23					4,37
24					5,18
25					5,54
26					5,11
27					4,96
28					4,64
29					4,69
30					4,98
31					4,82

Рисунок 8.2

Для обчислення рангів використовуємо статистичну функцію **RANK.AVG** (рис.8.3). У полі **Число** ми вказуємо комірку зі значенням скомпонованої вибірки, у полі **Посилання** – весь діапазон значень двох вибірок, у полі **Порядок** записуємо число 1, щоб ранг відсортувався у порядку зростання та натискаємо ОК.

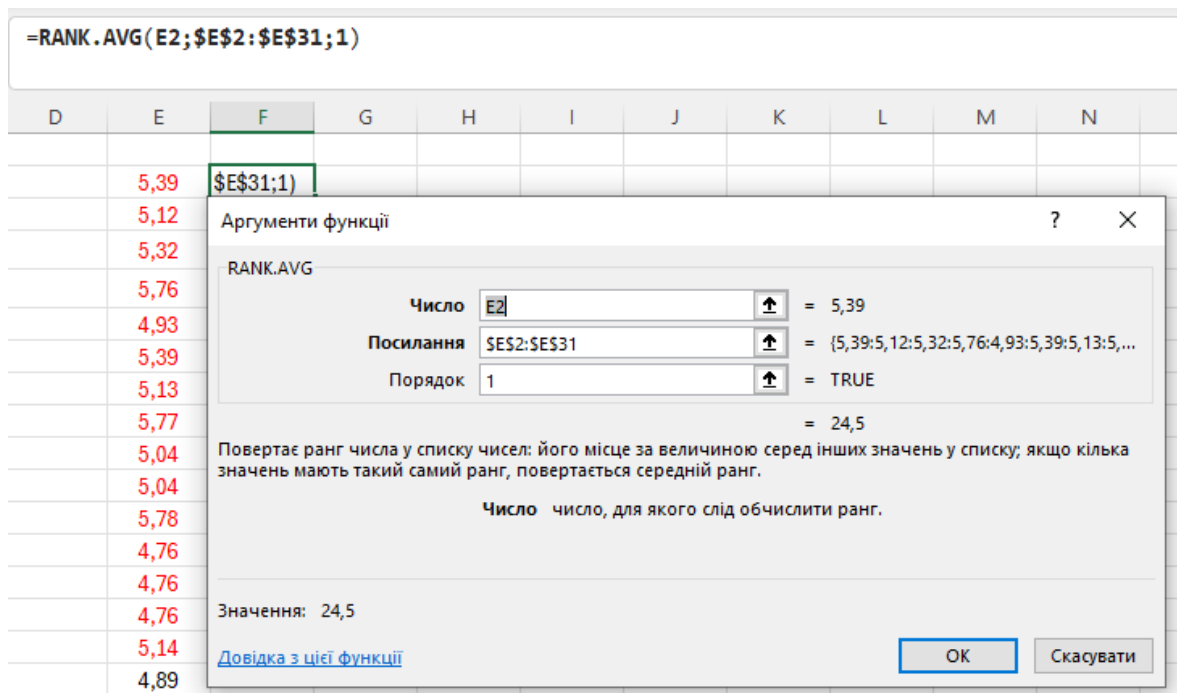


Рисунок 8.3

E	F	G	H
5,39	24,5		
5,12	19		
5,32	23		
5,76	28		
4,93	13		
5,39	24,5		
5,13	20		
5,77	29		
5,04	16,5		
5,04	16,5		
5,78	30		
4,76	7		
4,76	7		
4,76	7		
5,14	21	$S_1 = 286$	
4,89	12		
4,86	10		
4,68	4		
4,66	3		
4,87	11		
5,40	26		
4,37	1		
5,18	22		
5,54	27		
5,11	18		
4,96	14		
4,64	2		
4,69	5		
4,98	15		
4,82	9	$S_2 = 179$	

Рисунок 8.4

Щоб розрізняти ранги вибірок **A** та **B**, навпроти значень вибірки червоного кольору, таким самим кольором зафарбуємо значення рангів (рис.8.4). Потім за формулами п.1 можемо легко вирахувати суми рангів S_1 та S_2 (рис.8.4). І далі розраховуємо значення U_1 і U_2 (рис.8.5).

E	F	G	H	I	J
5,39	24,5				
5,12	19				
5,32	23				
5,76	28				
4,93	13				
5,39	24,5				
5,13	20				
5,77	29				
5,04	16,5				
5,04	16,5				
5,78	30				
4,76	7				
4,76	7				
4,76	7				
5,14	21	$S_1 = 286$		$U_1 = 59$	
4,89	12				
4,86	10				
4,68	4				
4,66	3				
4,87	11				
5,40	26				
4,37	1				
5,18	22				
5,54	27				
5,11	18				
4,96	14				
4,64	2				
4,69	5				
4,98	15				
4,82	9	$S_2 = 179$		$U_2 = 166$	

Рисунок 8.5

Таким чином маємо $U_1 = 59, U_2 = 166$. Тоді $U = \min(U_1, U_2) = 59$.

Якщо заздалегідь до аналізу даних є обґрунтоване припущення про напрямок ефекту, наприклад, перша група має більше значення, ніж друга, то застосовується односторонній тест. Оскільки ми не знаємо, в яку сторону може відрізнятись друга вибірка від першої – застосовуємо двосторонній тест. Для цього з відповідної таблиці (Додаток Г) для $\alpha = 0,05, n_1 = n_2 = 15$ визначаємо $U_{theory} = 64$.

Оскільки $59 < 64$, тобто $U < U_{theory}$, нульова гіпотеза H_0 відкидається.

Для візуального порівняння вибірок **A** і **B** у **Excel** можна побудувати діаграму **Розмах (boxplot)** (рис.8.6).

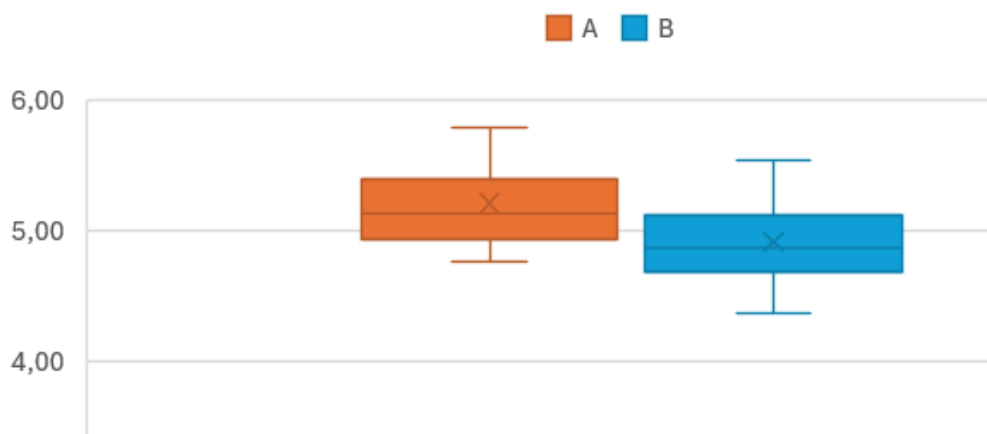


Рисунок 8.6

З рис.8.6 видно, що медіана у групи А вища, ніж у групи В, тобто препарат А в середньому призводить до вищих значень холестерину. Це візуально підтверджує результат *критерію Манна-Уїтні*, де відмінність виявилася статистично значущою.

3. Критерій Уїлкоксона

Критерій Уїлкоксона також відноситься до непараметричних критеріїв, але на відміну від *критерію Манна-Уїтні* застосовується до залежних (парних) даних (до і після лікування, права і ліва сторона тіла, дві методики на тих самих пацієнтах і т. д.). *Критерій Уїлкоксона* перевіряє, чи відрізняються медіанні значення до і після впливу, а також напрямок зміни показників (вгору або вниз).

Нехай задані дві пов'язані вибірки: x_i і y_i об'ємом n . Формулюємо наступні гіпотези:

- H_0 : відмінностей між парними вимірами немає,
- H_1 : відмінності є.

Для перевірки гіпотез за допомогою *критерію Уїлкоксона* застосовується така схема:

1. Обчислюються різниці $d_i = x_i - y_i$.
2. Виключаються пари, для яких $d_i = 0$. Кількість ненульових різниць позначаємо через k .
3. Знаходимо ранги R_i для модулів різниці $|d_i|$.
4. Відзначити, які ранги відповідають позитивним і які негативним різницям.
5. Обчислюємо суму рангів окремо для позитивних та негативних різниць

$$T_+ = \sum R_i, \text{ для } d_i > 0,$$

$$T_- = \sum R_i, \text{ для } d_i < 0.$$

6. Для визначення статистики беремо меншу із двох сум:

$$T = \min(T_+, T_-).$$

Для малих вибірок $k < 20$ із таблиці (Додаток Г) визначаємо теоретичне критичне значення T_{theory} . Якщо $T > T_{theory}$, приймається нульова гіпотеза H_0 про те, що відмінностей немає. Якщо $T \leq T_{theory}$, приймається альтернативна гіпотеза H_1 про те, що відмінності значущі.

Зауваження. Якщо $k \geq 20$, дії по перевірці гіпотез слід продовжити.

7. Використовуючи вже знайдене значення T обчислюємо

$$Z = \frac{T - \frac{k(k+1)}{4} \mu_U}{\sqrt{\frac{k(k+1)(2k+1)}{24}}}.$$

Використовуючи стандартне нормальне відхилення за заданим рівнем значущості α знаходимо Z_{theory} , яке можна порівняти із значенням Z , що спостерігається.

Перевіряємо гіпотези: якщо для двосторонньої перевірки $|Z| < Z_{theory, \alpha/2}$, приймається нульова гіпотеза H_0 про те, що відмінностей немає; якщо для двосторонньої перевірки $|Z| \geq Z_{theory, \alpha/2}$, приймається альтернативна гіпотеза H_1 про те, що відмінності значущі.

4. Застосування критерію Уїлкоксона засобами програми Excel

	A	B	C
1	№	A	B
2	1	112	104
3	2	139	136
4	3	141	141
5	4	108	108
6	5	149	151
7	6	126	126
8	7	165	161
9	8	104	111
10	9	128	128
11	10	136	134
12	11	137	137
13	12	107	109
14	13	164	164
15	14	116	116
16	15	170	179

Рисунок 8.7

	A	B	C	D
1	№	A	B	d_i
2	1	112	104	8
3	2	139	136	3
4	3	141	141	0
5	4	108	108	0
6	5	149	151	-2
7	6	126	126	0
8	7	165	161	4
9	8	104	111	-7
10	9	128	128	0
11	10	136	134	2
12	11	137	137	0
13	12	107	109	-2
14	13	164	164	0
15	14	116	116	0
16	15	170	179	-9

Рисунок 8.8

Нехай порівнюються дві залежні (парні) вибірки А та В. Наприклад, деякі медичні показники до та після лікування (рис.8.7). Перевірку гіпотези про відмінність вибірок відповідно до **критерію Уїлкоксона** проводимо в наступній

послідовності. Спочатку знаходимо різниці відповідних значень вибірки (рис.8.8). Різниці містять нульові, позитивні та негативні значення. Видаляємо нульові різниці, а у негативних значень за допомогою функції **ABS** видаляємо знак мінус (рис.8.9).

B	C	D	E
A	B	d_i	$ d_i $
112	104	8	8
139	136	3	3
149	151	-2	2
165	161	4	4
104	111	-7	7
136	134	2	2
107	109	-2	2
170	179	-9	9

Рисунок 8.9

Далі використовуємо функцію **RANK.AVG** для обчислення рангів модулів різниць d_i (рис.8.10).

fx				
=RANK.AVG(E2;\$E\$2:\$E\$9;1)				
B	C	D	E	F
A	B	d_i	$ d_i $	ранг $ d_i $
112	104	8	8	7,0
139	136	3	3	4,0
149	151	-2	2	2,0
165	161	4	4	5,0
104	111	-7	7	6,0
136	134	2	2	2,0
107	109	-2	2	2,0
170	179	-9	9	8,0

Рисунок 8.10

Потім за допомогою логічної функції **IF** визначаємо знак різниці d_i (рис.8.11). Далі, використовуючи функцію **SUMIF** знаходимо суму рангів окремо для позитивних та негативних різниць (рис.8.12).

Отримуємо $T_+ = 18$, $T_- = 18$, тобто, $\min(T_+, T_-) = 18$.

fx =IF(D2>0;"+";"-")

B	C	D	E	F
A	B	d_i	$ d_i $	знак d_i
112	104	8	8	+
139	136	3	3	+
149	151	-2	2	-
165	161	4	4	+
104	111	-7	7	-
136	134	2	2	+
107	109	-2	2	-
170	179	-9	9	-

Рисунок 8.11

fx =SUMIF(G2:G9;"+";F2:F9)

B	C	D	E	F	G	H	I	J
A	B	d_i	$ d_i $	ранг $ d_i $	знак d_i			
112	104	8	8	7,0	+			
139	136	3	3	4,0	+		$T_+ =$	18
149	151	-2	2	2,0	-		$T_- =$	18
165	161	4	4	5,0	+			
104	111	-7	7	6,0	-			
136	134	2	2	2,0	+			
107	109	-2	2	2,0	-			
170	179	-9	9	8,0	-			

Рисунок 8.12

Отже, маємо, що суми рангів позитивних і негативних різниць збігаються. Отже, різниці розподілені симетрично щодо нуля, і немає систематичного зсуву в один чи інший бік.

У нашому випадку загальна кількість пар – 15, але з них 7 пар із нульовими різницями. Вони не дають інформації про напрям змін, тобто не впливають на суму

позитивних чи негативних рангів. Тому для знаходження теоретичного значення за таблицею будемо використовувати

$$n - 15 - 7 = 8.$$

Із таблиці (додаток Д) для $\alpha = 0,05$, $n = 8$ визначаємо $T_{theory} = 5$.

Оскільки $18 > 5$, тобто $T > T_{theory}$, приймається нульова гіпотеза H_0 , тобто *тест Уїлкоксона* не виявляє статистично значущих відмінностей.

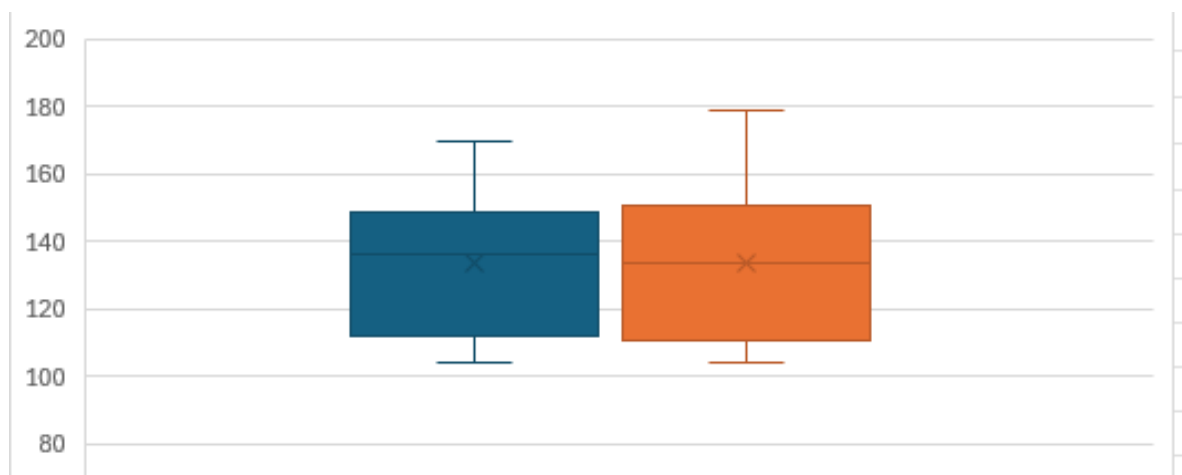


Рисунок 8.13

Для візуального порівняння вибірок **A і B** у **Excel** можна побудувати діаграму **Розмах (boxplot)**. На діаграмі видно, що медіани та розмахи ящиків майже збігаються, і що немає явного зсуву у бік підвищення чи зниження показників (рис.8.13).

Практичне завдання

У завданнях 1,2 за варіантом надані дві вибірки. Потрібно перевірити гіпотезу про відмінність вибірок:

- у завданні 1 за допомогою *критерію Манна-Уїтні*;
- у завданні 2 за допомогою *критерію Уїлкоксона*.

Завдання 1.

Варіант №1

A	5,37	5,62	4,97	5,2	5,37	6,04	5,82	5,34	5,63	5,46	5,09	5,15	5,08	5,37	5,45
B	4,36	5,22	4,91	4,85	4,80	5,38	4,84	4,85	5,12	4,90	5,13	4,93	4,92	5,26	4,88

Варіант №2

A	4,48	4,36	4,82	4,46	5,06	4,90	4,69	4,57	4,77	5,01	5,10	4,38	5,47	4,57	4,39
B	5,67	5,56	5,48	5,00	5,52	5,17	4,78	5,00	4,82	5,47	5,55	4,89	5,15	5,32	4,85

Варіант №3

A	5,23	4,75	4,65	4,79	4,91	4,79	5,30	5,00	4,64	5,09	4,64	5,38	4,70	4,89	5,05
B	4,79	4,73	5,04	4,59	5,24	4,66	4,67	4,94	4,43	4,98	4,80	5,06	4,95	5,20	4,76

Варіант №4

A	5,48	5,00	4,69	5,13	4,53	4,64	4,68	5,11	4,93	3,72	5,86	6,04	4,62	5,50	5,33
B	4,91	4,96	5,21	4,68	5,16	5,25	4,87	5,29	5,16	5,22	4,98	4,79	4,92	4,55	4,83

Варіант №5

A	5,13	5,07	5,13	5,07	5,13	5,07	5,13	5,07	5,13	5,07	5,13	5,07	5,13	5,07	5,13
B	4,83	3,91	4,83	3,91	4,83	3,91	4,83	3,91	4,83	3,91	4,83	3,91	4,83	3,91	4,83

Варіант №6

A	4,76	4,76	4,76	5,04	5,04	5,39	5,39	4,81	4,79	4,98	5,02	5,35	5,16	5,08	4,87
B	4,76	4,76	5,04	5,39	5,39	4,76	4,68	4,91	4,97	4,93	5,02	5,02	5,09	5,02	4,66

Варіант №7

A	6,50	5,21	6,50	5,21	6,50	5,21	6,50	5,21	6,50	5,21	6,50	5,21	6,50	5,21	6,50
B	4,94	5,10	4,94	5,10	4,94	5,10	4,94	5,10	4,94	5,10	4,94	5,10	4,94	5,10	4,94

Варіант №8

A	5,15	4,94	5,15	4,94	5,15	4,94	5,15	4,94	5,15	4,94	5,15	4,94	5,15	4,94	5,15
B	4,75	3,40	4,75	3,40	4,75	3,40	4,75	3,40	4,75	3,40	4,75	3,40	4,75	3,40	4,75

Варіант №9

A	5,73	5,34	5,73	5,34	5,73	5,34	5,73	5,34	5,73	5,34	5,73	5,34	5,73	5,34	5,73
B	5,80	4,96	5,80	4,96	5,80	4,96	5,80	4,96	5,80	4,96	5,80	4,96	5,80	4,96	5,80

Варіант №10

A	4,45	5,43	4,45	5,43	4,45	5,43	4,45	5,43	4,45	5,43	4,45	5,43	4,45	5,43	4,45
B	4,82	5,74	4,82	5,74	4,82	5,74	4,82	5,74	4,82	5,74	4,82	5,74	4,82	5,74	4,82

Варіант №11

A	4,84	5,05	4,84	5,05	4,84	5,05	4,84	5,05	4,84	5,05	4,84	5,05	4,84	5,05	4,84
B	4,99	5,03	4,99	5,03	4,99	5,03	4,99	5,03	4,99	5,03	4,99	5,03	4,99	5,03	4,99

Варіант №12

A	5,18	5,26	5,18	5,26	5,18	5,26	5,18	5,26	5,18	5,26	5,18	5,26	5,18	5,26	5,18
B	4,88	4,81	4,88	4,81	4,88	4,81	4,88	4,81	4,88	4,81	4,88	4,81	4,88	4,81	4,88

Варіант №13

A	5,00	5,00	5,00	5,00	5,00	5,00	5,00	5,00	5,00	5,00	5,00	5,00	5,00	5,00	5,00
B	5,00	5,00	5,00	5,00	5,00	5,00	5,00	5,00	5,00	5,00	5,00	5,00	5,00	5,00	5,00

Варіант №14

A	4,80	4,90	4,80	4,90	4,80	4,90	4,80	4,90	4,80	4,90	4,80	4,90	4,80	4,90	4,80
B	5,10	5,10	5,10	5,10	5,10	5,10	5,10	5,10	5,10	5,10	5,10	5,10	5,10	5,10	5,10

Варіант №15

A	5,07	4,71	5,07	4,71	5,07	4,71	5,07	4,71	5,07	4,71	5,07	4,71	5,07	4,71	5,07
B	6,30	4,94	6,30	4,94	6,30	4,94	6,30	4,94	6,30	4,94	6,30	4,94	6,30	4,94	6,30

Завдання 2

Варіант №1

A	151	114	171	160	120	174	174	123	102	121	152	101	129	137	101
B	151	106	179	169	120	174	168	127	102	121	153	101	125	145	101

Варіант №2

A	120	172	138	117	103	159	113	108	152	101	159	170	143	107	146
B	120	172	138	117	94	164	116	108	151	94	163	170	144	107	153

Варіант №3

A	171	177	161	139	179	152	123	125	159	140	128	114	144	164	170
B	171	175	161	135	175	150	124	125	159	137	128	114	134	164	161

Варіант №4

A	147	122	161	136	143	134	164	146	177	102	100	104	113	126	108
B	147	126	161	135	134	127	164	154	177	102	102	104	109	126	108

Варіант №5

A	144	161	156	105	127	127	143	129	161	174	161	100	126	161	176
B	135	162	156	114	127	127	143	129	161	180	170	98	132	157	176

Варіант №6

A	102	155	158	101	101	153	100	118	101	152	143	131	169	131	167
B	102	163	151	93	109	153	100	118	110	147	143	123	169	131	172

Варіант №7

A	119	170	151	132	139	138	100	110	156	149	122	130	141	106	115
B	119	170	143	132	144	130	98	119	156	149	122	130	146	114	113

Варіант №8

A	123	153	132	123	174	171	135	137	124	117	165	153	134	179	160
B	123	161	134	123	174	172	143	134	124	120	160	153	134	179	155

Варіант №9

A	126	126	120	129	127	163	168	160	147	118	103	134	163	148	116
B	125	135	120	119	127	163	175	152	147	110	93	134	163	145	116

Варіант №10

A	145	123	131	146	122	165	126	101	116	132	108	142	147	138	141
B	151	123	131	154	113	168	126	101	116	132	99	150	148	142	141

Варіант №11

A	151	141	169	114	153	159	107	152	159	104	167	105	146	154	139
B	151	141	169	114	151	160	98	157	153	104	171	105	146	144	129

Варіант №12

A	100	115	160	163	162	168	121	166	175	125	115	150	156	128	177
B	100	106	169	163	162	162	115	159	174	125	115	150	164	128	167

Варіант №13

A	103	115	123	179	101	131	123	111	149	134	132	132	160	150	142
B	103	115	123	176	101	130	121	111	145	142	123	132	153	150	149

Варіант №14

A	174	128	135	120	135	109	172	123	163	148	135	123	122	161	136
B	174	136	135	121	128	110	172	129	163	150	135	123	122	169	134

Варіант №15

A	108	170	127	177	151	115	168	111	124	151	152	122	115	156	138
B	106	169	130	178	151	116	168	117	119	151	148	122	115	156	138

ПРАКТИЧНА РОБОТА 9

Дисперсійний аналіз. Критерій Краскела–Уолліса

Питання, які розглядаються:

1. Дисперсійний аналіз параметричними методами.
2. Дисперсійний аналіз засобами програми Excel.
3. Непараметричні методи. Критерій Краскела-Уолліса.
4. Застосування критерію Краскела-Уолліса засобами програми Excel.

Дисперсійний аналіз ANOVA (Analysis of Variance) - це метод, який застосовується для порівняння кількох (більше двох) груп даних. Він перевіряє чи відрізняються середні значення (або медіани, якщо застосовується **критерій Краскела-Уолліса**). Цей метод часто використовується у дослідженнях для перевірки впливу різних умов, методів лікування чи інших факторів на певний результат.

1. Дисперсійний аналіз параметричними методами

Розрізняють **однофакторний ANOVA**, **двофакторний та двофакторний з повтореннями**. Прикладом **однофакторного ANOVA** може бути порівняння трьох методів навчання за балами, набраними у процесі тестування.

Якщо до фактора методу навчання додати фактор спеціальності (медицина, фармація), то застосовується **двофакторний ANOVA**. Якщо у кожній комбінації факторів (наприклад, Метод А та Фармація) є лише одне спостереження, то потрібно використовувати **двофакторний ANOVA** без повторень, а якщо кілька спостережень, то - **двофакторний ANOVA із повтореннями**.

Дисперсійний аналіз (ANOVA) передбачає, що:

- дані в кожній групі розподілені нормально (перевірка за допомогою, наприклад, **критерію Шапіро-Уїлка**);
- дисперсії вибірок рівні, тобто дані зосереджені навколо середнього значення (гомогеність дисперсій). Ця умова може бути перевірена за допомогою **критерію**

Фішера, а якщо вибірок більше трьох, за допомогою **критерію Левена або Брауна-Форсайта**;

- вибірки є незалежними.

Розглянемо алгоритм застосування **ANOVA**. Нехай задано k незалежних виборок $X_1, X_2, \dots, X_k$. Кількість значень в i -й виборці дорівнює n_i .

Формулювання гіпотез:

- H_0 : всі вибірки (групи) походять із генеральних сукупностей з однаковими середніми значеннями,
- H_1 : хоча б одне середнє відрізняється.

Далі застосовується наступна схема:

1. Визначається міжгрупова дисперсія $SS_{\text{між}}$ (відмінності між середніми груп)

$$SS_{\text{між}} = \sum_{i=1}^k n_i (\bar{X}_i - \bar{X}_{\text{заг}})^2.$$

де $\bar{X}_i$ - середнє в i -й групі, $\bar{X}_{\text{заг}}$ - загальне середнє всіх даних.

2. Визначається внутрішньогрупова дисперсія $SS_{\text{внут}}$ (розкид всередині кожної групи)

$$SS_{\text{внут}} = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2.$$

3. Обчислюємо середні дисперсії:

$$MS_{\text{між}} = \frac{SS_{\text{між}}}{k - 1}, \quad MS_{\text{внут}} = \frac{SS_{\text{внут}}}{N - k},$$

де N – загальна кількість спостережень.

4. Обчислюємо статистику :

$$F = \frac{MS_{\text{між}}}{MS_{\text{внут}}}.$$

Якщо $F < F_{theory}$, приймається нульова гіпотеза H_0 про рівність середніх. Якщо $F \geq F_{theory}$, приймається альтернативна гіпотеза H_1 про те, що хоча б одне середнє відрізняється.

Значення F_{theory} можна взяти із таблиці Фішера–Снедекора для рівня значущості $\alpha = 0,05$ (додаток Е). При цьому використовуємо ступені свободи $k_1 = k - 1, k_2 = N - k$.

2. Дисперсійний аналіз засобами програми Excel

Наведемо конкретний приклад, результати розрахунку якого потім порівняємо з результатом, отриманим засобами програми **Excel**. У трьох групах проведені заняття з різних методик: А, В та С. Після занять відбулося тестування, результати якого наведено у таблиці 9.1.

Таблиця 9.1

А	70	65	72	68	74
В	80	78	82	84	79
С	90	88	92	85	91

Потрібно порівняти середній бал у трьох груп, які навчаються за різними методиками.

1. Обчислюємо середні значення

$$\bar{X}_A = 69,8, \quad \bar{X}_B = 80,6, \quad \bar{X}_C = 89,2.$$

Тоді

$$\bar{X}_{\text{заг}} = \frac{(69,8 + 80,6 + 89,2)}{3} = 79,87.$$

2. Визначаємо міжгрупову суму дисперсій

$$SS_{\text{між}} = 5[(69,8 - 79,87)^2 + (80,6 - 79,87)^2 + (89,2 - 79,87)^2] = 944,9.$$

Оскільки міжгрупова ступінь свободи $k_1 = k - 1 = 3 - 1 = 2$, середня міжгрупова дисперсія дорівнює

$$MS_{\text{між}} = \frac{944,9}{2} = 472,47.$$

1. Визначаємо внутрішньогрупову суму дисперсій. Оскільки дисперсії груп

$$s_A^2 = 12,2, \quad s_B^2 = 5,8, \quad s_C^2 = 7,7,$$

то

$$SS_{\text{внут}} = 12,2(5 - 1) + 5,8(5 - 1) + 7,7(5 - 1) = 102,8.$$

Оскільки внутрішньогруповий ступінь свободи $k_2 = N - k = 15 - 3 = 12$, середня внутрішньогрупова дисперсія

$$MS_{\text{внут}} = \frac{102,8}{12} = 8,57.$$

Обчислюємо статистику:

$$F = \frac{MS_{\text{між}}}{MS_{\text{внут}}} = \frac{472,47}{8,56} = 55,15.$$

2. З таблиці Фішера–Снедекора для рівня значущості $\alpha = 0,05$, ступенях свободи $k_1 = 2, k_2 = 12$ знаходимо $F_{\text{theory}} = 3,89$ (додаток Е).

Оскільки $55,15 > 3,89$, тобто $F \geq F_{\text{theory}}$, відмінності статистично значимі, і нульова гіпотеза відкидається. Висновок: середні бали за трьома методиками різняться статистично значуще.

Порівняємо отримані результати з результатами застосування однофакторного *ANOVA* в програмі **Excel**. Перенесемо вихідну таблицю на аркуш програми **Excel**.

	A	B	C
1	A	B	C
2	70	80	90
3	65	78	88
4	72	82	92
5	68	84	85
6	74	79	91

Рисунок 9.1

Через пункт меню **Дані** запускаємо пакет **Data Analysis** (рис.9.2).

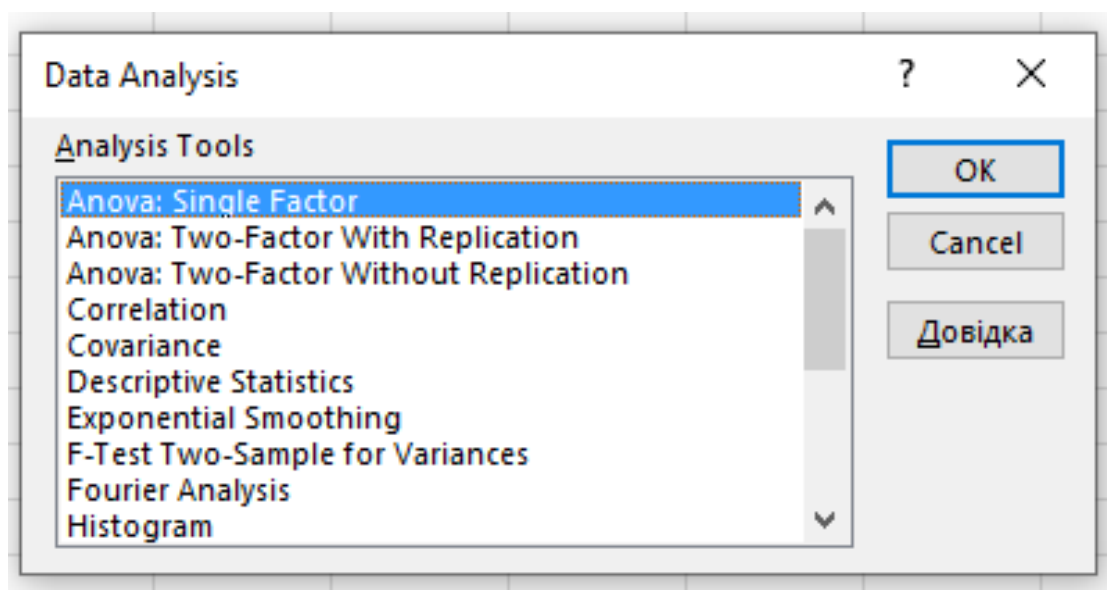


Рисунок 9.2

Перші три позиції призначені для проведення тестів *ANOVA*. Обираємо однофакторний **ANOVA (Anova: Single Factor)**. В полі **Input Range** вказуємо діапазон даних, в полі **Output Range** – комірку, в якій будуть розміщуватись результати і натискаємо **OK** (рис.9.3).

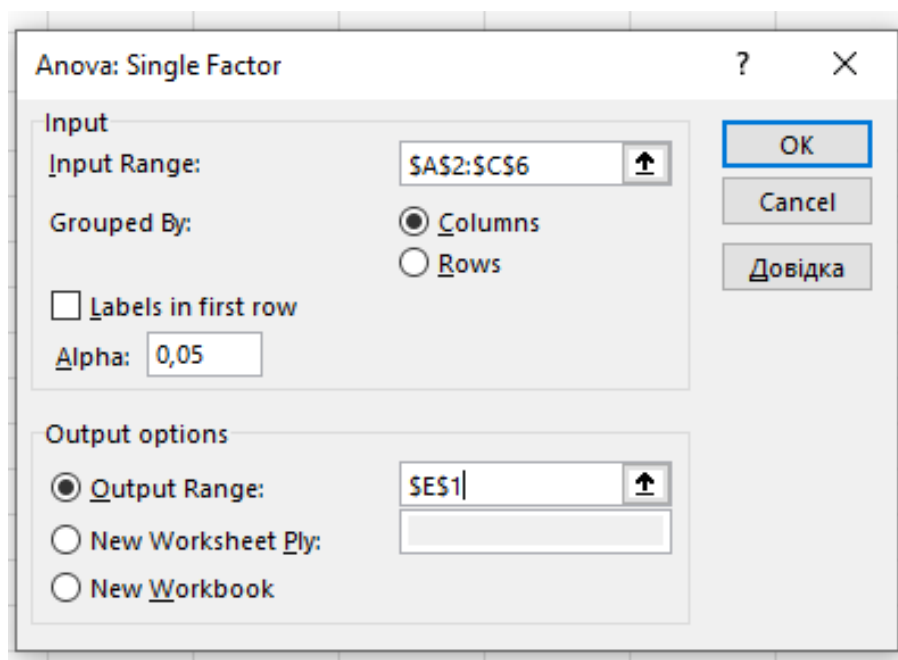


Рисунок 9.3

В результаті на аркуш виводяться всі проміжні та кінцеві результати, які співпали із обчисленими раніше (рис.9.4). Зазначимо, що додатково зроблено розрахунок $P - value = 8,92 \cdot 10^{-7}$. Так як він значно менший за рівень значимості, нульова гіпотеза відкидається, тобто також підтверджується раніше зроблений висновок.

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
Column 1	5	349	69,8	12,2		
Column 2	5	403	80,6	5,8		
Column 3	5	446	89,2	7,7		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	944,9333	2	472,4667	55,15175	8,92E-07	3,885294
Within Groups	102,8	12	8,566667			
Total	1047,733	14				

Рисунок 9.4

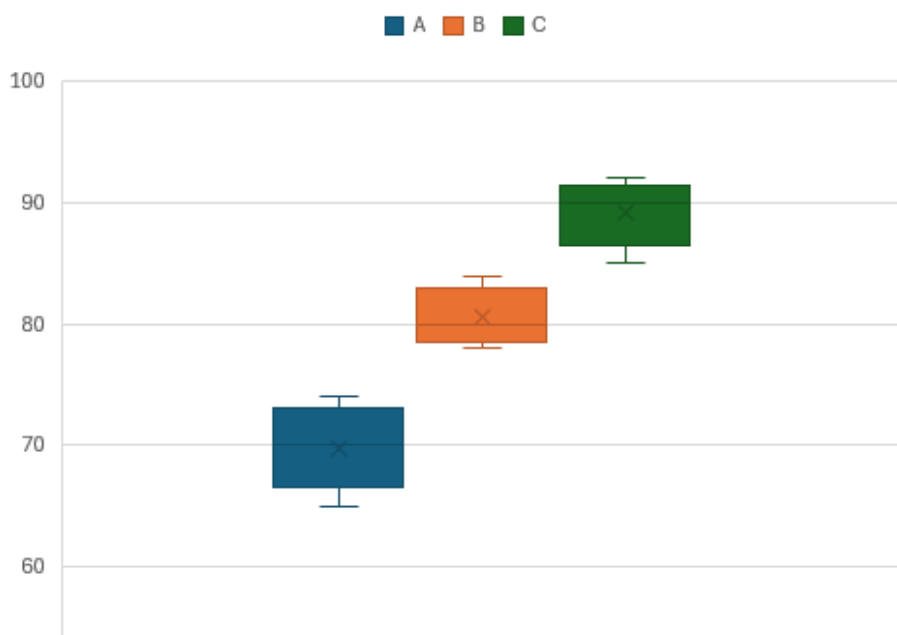


Рисунок 9.5

Для наочності за допомогою діаграм **Excel** побудуємо діаграму **Розмах (boxplot)** (рис.9.5). Діапазони значень майже не перекриваються - видно точний поділ між групами.

3. Непараметричні методи. Критерій Краскела–Уолліса

Критерій Краскела–Уолліса є непараметричним аналогом *однофакторного дисперсійного аналізу (ANOVA)*, який застосовується, коли дані не відповідають нормальному розподілу або мають різну дисперсію. Нехай задано k незалежних виборок $X_1, X_2, \dots, X_k$. Кожна виборка X_i має об'єм n_i , а сумарна кількість спостережень:

$$N = n_1 + n_2 + \dots + n_k.$$

Формулювання гіпотез:

- H_0 : всі виборки походять з однієї і тієї ж генеральної сукупності (медіани рівні),
- H_1 : хоча б одна вибірка має відмінну медіану.

Для перевірки гіпотез за допомогою **критерія Краскела–Уолліса** застосовується наступна схема:

1. Значення вибірок об'єднуються у загальний вибір, після чого кожному значенню присвоюється ранг.
2. Знаходимо суму рангів R_i окремо для кожної i -й групи.
3. Обчислюємо статистику критерія

$$H = \frac{12}{N(N+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(N+1).$$

4. Якщо у виборках є t груп однакових значень, і при цьому число збіглих значень дорівнює t_i , то вводиться поправка

$$C = 1 - \frac{\sum_{j=1}^m (t_j^3 - t_j)}{N^3 - N}.$$

Тоді уточнений критерій

$$H' = \frac{H}{C}.$$

5. Критичне значення **критерія Краскела–Уолліса** беруть з таблиці розподілу χ^2 (хи-квадрат) Пірсона. Для цього визначемо кількість груп k , обчислюємо ступінь свободи $df = k - 1$ та обраємо рівень значущості α (додаток Є).

Якщо $H' \leq \chi_{theory}^2$, приймається нульова гіпотеза H_0 про те, що відмінностей немає (медіани рівні). Якщо $H' > \chi_{theory}^2$, приймається альтернативна гіпотеза H_1 про те, що відмінності значущі (хоча б одна виборка має відмінну медіану).

4. Застосування критерія Краскела–Уолліса засобами програми Excel

Маємо показники рівня гемоглобіну (г/л) у трьох різних груп пацієнтів (рис.9.6).

	А	В	С
1	Група 1	Група 2	Група 3
2	104,5	90,6	110,7
3	117,8	93,2	105,9
4	93,3	109,9	128,9
5	83,1	103,3	103,2
6	106,6	98,1	100,9

Рисунок 9.6

Перевірку гіпотези про відмінність вибірок відповідно до **критерію Краскела-Уолліса** проводимо в наступній послідовності. Об'єднаємо значення трьох вибірок і розташуємо в одному стовпці (рис.9.7).

	A	B	C	D
1	Група 1	Група 2	Група 3	Загал.
2	104,5	90,6	110,7	104,5
3	117,8	93,2	105,9	117,8
4	93,3	109,9	128,9	93,3
5	83,1	103,3	103,2	83,1
6	106,6	98,1	100,9	106,6
7				90,6
8				93,2
9				109,9
10				103,3
11				98,1
12				110,7
13				105,9
14				128,9
15				103,2
16				100,9

Рисунок 9.7

За допомогою функції **RANK.AVG** обчислимо ранги (рис.9.8) .

	A	B	C	D	E
1	Група 1	Група 2	Група 3	Загал.	Ранг
2	104,5	90,6	110,7	104,5	9
3	117,8	93,2	105,9	117,8	14
4	93,3	109,9	128,9	93,3	4
5	83,1	103,3	103,2	83,1	1
6	106,6	98,1	100,9	106,6	11
7				90,6	2
8				93,2	3
9				109,9	12
10				103,3	8
11				98,1	5
12				110,7	13
13				105,9	10
14				128,9	15
15				103,2	7
16				100,9	6

Рисунок 9.8

Далі сумуємо ранги для кожної групи окремо (рис.9.9).

G16		: X ✓ f _x		=SUM(E12:E16)			
	A	B	C	D	E	F	G
1	Група 1	Група 2	Група 3	Загал.	Ранг		
2	104,5	90,6	110,7	104,5	9		
3	117,8	93,2	105,9	117,8	14		
4	93,3	109,9	128,9	93,3	4		
5	83,1	103,3	103,2	83,1	1		
6	106,6	98,1	100,9	106,6	11	R ₁ =	39
7				90,6	2		
8				93,2	3		
9				109,9	12		
10				103,3	8		
11				98,1	5	R ₂ =	30
12				110,7	13		
13				105,9	10		
14				128,9	15		
15				103,2	7		
16				100,9	6	R ₃ =	51

Рисунок 9.9

Значень, що повторюються, немає, тому використовуємо формулу розрахунку статистики без уточнень

$$H = \frac{12}{N(N+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(N+1).$$

В результаті отримуємо $H = 2,22$ (рис. 9.10).

У нашому випадку $k = 3$, $df = 3 - 1 = 2$, $\alpha = 0,05$. Тоді з таблиці (додаток Є) знаходимо $\chi_{theory}^2 = 5,99$.

Оскільки $2,22 \leq 5,99$, тобто $H \leq \chi_{theory}^2$, приймається нульова гіпотеза H_0 , про рівність медіан.

Для наочного порівняння вибірок трьох груп в **Excel** можна побудувати діаграму **Розмах (boxplot)**. З діаграми видно (рис.9.11), що медіани трьох груп розташовані близько одна до одної. Це наочно підтверджує результат тесту: значущих відмінностей між групами немає.

G2		: X ✓ f _x		=12/(15*(15+1))*(G6^2/5+G11^2/5+G16^2/5)-3*(15+1)			
	A	B	C	D	E	F	G
1	Група 1	Група 2	Група 3	Загал.	Ранг		
2	104,5	90,6	110,7	104,5	9	$H =$	2,22
3	117,8	93,2	105,9	117,8	14		
4	93,3	109,9	128,9	93,3	4		
5	83,1	103,3	103,2	83,1	1		
6	106,6	98,1	100,9	106,6	11	$R_1 =$	39
7				90,6	2		
8				93,2	3		
9				109,9	12		
10				103,3	8		
11				98,1	5	$R_2 =$	30
12				110,7	13		
13				105,9	10		
14				128,9	15		
15				103,2	7		
16				100,9	6	$R_3 =$	51

Рисунок 9.10

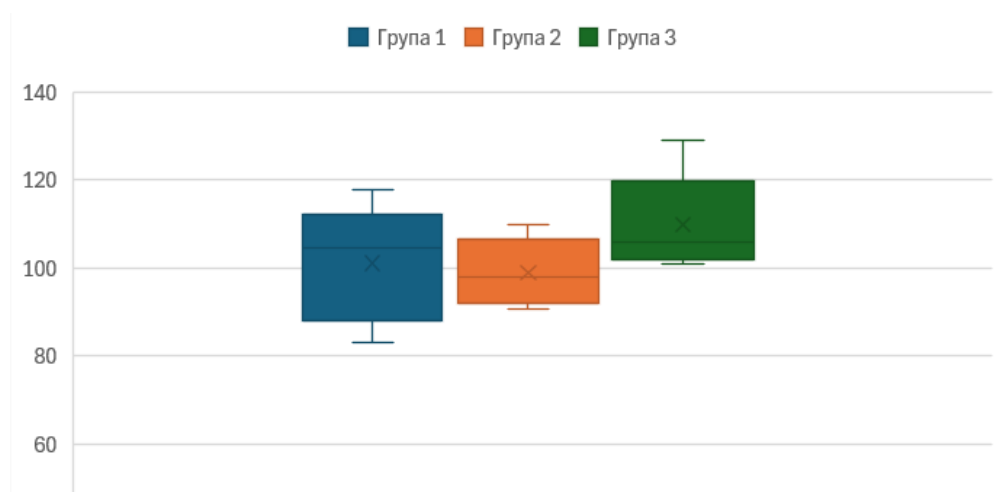


Рисунок 9.11

Практичне завдання

Завдання 1. Перевірити за допомогою *критерію Фішера* гіпотезу про рівність дисперсій (попарно). Застосувати однофакторний (*ANOVA*) для перевірки гіпотези про рівність середніх.

Варіант №1 (рівень тромбоцитів (10⁹/л))

Група 1	100,7	94	93,6	91,7	91
Група 2	127,9	109,9	112,4	115,8	126,3
Група 3	123,3	120,9	120,2	126,9	124,6

Варіант №2 (рівень холестерину (ммоль/л))

Група 1	90,5	92,3	102,9	90,7	93,8
Група 2	129,5	117,9	124,3	120,7	123,7
Група 3	104	109,1	114,5	122,6	117,1

Варіант №3 (рівень тромбоцитів (10⁹/л))

Група 1	95,9	86,4	98,8	97,2	95,9
Група 2	111	114,2	108,5	115,4	124,2
Група 3	102,7	110,5	105	105,9	112,4

Варіант №4 (частота серцевих скорочень (уд/хв))

Група 1	95,6	92,5	99,4	98,5	104,7
Група 2	104,5	111,6	114,4	109,4	110,9
Група 3	108,3	115,5	107,6	115	116

Варіант №5 (індекс ваги тіла (кг/м²))

Група 1	117,8	106,5	101,4	117,2	110
Група 2	104,4	100,8	97	111,3	101,6
Група 3	123,3	119	113,4	109,6	111,3

Варіант №6 (індекс ваги тіла (кг/м²))

Група 1	96,9	108,8	96,7	103,9	108,9
Група 2	96,3	103,2	92,6	91,2	99,9
Група 3	102,1	107,2	94,1	107,6	117

Варіант №7 (рівень глюкози у крові (ммоль/л))

Група 1	126,9	127,2	127,9	117,7	129,1
Група 2	84	89,9	98,2	88,3	94,5
Група 3	113,1	109,4	101,1	106,3	117,8

Варіант №8 (рівень холестерину (ммоль/л))

Група 1	83,9	99,8	78	86,3	91,4
Група 2	101	92,7	97,9	99,9	94,7
Група 3	92,1	86,2	99,3	91,5	102,3

Варіант №9 (артеріальний тиск (мм рт. ст.))

Група 1	122,6	128,6	125,4	119,3	123,9
Група 2	111,2	111,4	108,9	107,9	103,8
Група 3	115,5	119,2	117	123,3	105,6

Варіант №10 (рівень тромбоцитів (10⁹/л))

Група 1	88,5	116,9	98,1	96,6	102
Група 2	110,4	100,5	107,6	111,9	116
Група 3	108,1	117,8	105,8	110,6	106,8

Варіант №11 (частота серцевих скорочень (уд/хв))

Група 1	118,8	122,5	122,1	120	115
Група 2	94,1	91,8	90,5	95,2	97,7
Група 3	131,1	135,3	135,2	132,9	124,1

Варіант №12 (рівень холестерину (ммоль/л))

Група 1	128,4	129,6	137,8	120,6	133
Група 2	130,7	123,8	121,2	123,7	120,2
Група 3	88,7	89,6	91,1	97,6	103,8

Варіант №13 (частота серцевих скорочень (уд/хв))

Група 1	86,1	88,8	92,7	89,8	95,5
Група 2	100,1	100	106	111,7	100,7
Група 3	113,6	118,2	120,4	121,3	120,1

Варіант №14 (рівень тромбоцитів (10⁹/л))

Група 1	113,3	118,8	115,4	120,7	121,1
Група 2	93,2	91,9	90	101,3	99,8
Група 3	92	104,6	97,3	107,5	95,9

Варіант №15 (рівень холестерину (ммоль/л))

Група 1	103,2	97,7	94,6	102,2	101,3
Група 2	112,7	110,8	109,8	116	115,3
Група 3	117,2	120,6	118,5	124,8	116,9

Завдання 2. Застосувати критерій Краскела-Уолліса для перевірки гіпотези про рівність медіан.

Варіант №1 (рівень тромбоцитів ($10^9/\text{л}$))

Група 1	94,5	105,2	104,7	113,7	90,8
Група 2	103,8	84,9	100,1	108,9	95,7
Група 3	110,8	108,4	110,2	105,7	104,7

Варіант №2 (частота серцевих скорочень (уд/хв))

Група 1	98,8	93	78,6	93,7	106
Група 2	130,6	108,9	106,2	99,8	99
Група 3	119,5	112,9	103,6	99,8	108,4

Варіант №3 (рівень холестерину (ммоль/л))

Група 1	96,3	103,8	88,1	95,9	95,5
Група 2	120,2	108,2	91,1	83,2	94,6
Група 3	111,7	113,2	117,5	91,6	115,6

Варіант №4 (рівень холестерину (ммоль/л))

Група 1	100,3	85,2	92,8	95,4	110,6
Група 2	108,4	87,4	108,2	101,1	98,2
Група 3	116,1	120,3	119,3	101,6	106,9

Варіант №5 (рівень тромбоцитів ($10^9/\text{л}$))

Група 1	98,3	111,6	97,5	92,3	112,1
Група 2	117,3	89,3	111,4	84,6	115,7
Група 3	100,7	117,8	103,9	113,3	96,2

Варіант №6 (рівень лейкоцитів (10⁹/л))

Група 1	97,2	85,8	110,4	109	100,2
Група 2	99,7	90	97,1	112,4	102,9
Група 3	105,7	115	121,6	112,6	113,1

Варіант №7 (частота серцевих скорочень (уд/хв))

Група 1	109,6	97,9	92,3	96,4	107,2
Група 2	102,4	113,5	91,9	96,3	99,9
Група 3	96,9	139,4	99	119,1	103,3

Варіант №8 (рівень лейкоцитів (10⁹/л))

Група 1	94,9	118,9	101,7	102,6	99,3
Група 2	85,8	104,7	105,6	129,6	103,1
Група 3	113	109,7	98,3	121,4	117,5

Варіант №9 (рівень глюкози у крові (ммоль/л))

Група 1	107,3	93,8	103,6	99,8	112,2
Група 2	101,1	100,6	108,8	124,3	105,8
Група 3	105,8	118,6	117,2	100,5	104,6

Варіант №10 (рівень тромбоцитів (10⁹/л))

Група 1	111	105,4	90,9	106,2	98,4
Група 2	101,1	96,1	101,4	110,6	115,4
Група 3	115,3	123,6	135,4	106,8	107,9

Варіант №11 (частота серцевих скорочень (уд/хв))

Група 1	110	101,5	94	91,5	108,2
Група 2	109,7	102,1	115,7	113,5	122,9
Група 3	115,7	108,4	108,6	90,7	111,5

Варіант №12 (частота серцевих скорочень (уд/хв))

Група 1	99,8	90,3	107,9	111,6	91,8
Група 2	114,6	109,1	113,2	124	102,5
Група 3	102,5	101,1	101,8	109,2	113,4

Варіант №13 (індекс маси тіла (кг/м²))

Група 1	98,6	108,7	113	86,6	108,5
Група 2	108,9	105	89,3	119,8	109
Група 3	92,2	107,6	113,3	115,1	108,7

Варіант №14 (індекс маси тіла (кг/м²))

Група 1	103,4	95,5	107,8	110,5	96,6
Група 2	95,7	99,9	112,1	105,9	111,3
Група 3	127,6	112,3	101,9	120,6	110,5

Варіант №15 (індекс маси тіла (кг/м²))

Група 1	95,4	106,9	83,9	95,3	110,9
Група 2	105,6	94,2	97,8	111,8	97,7
Група 3	112,2	110,5	103,5	131,4	116,3

ПРАКТИЧНА РОБОТА 10

Довірчі інтервали

Питання, які розглядаються:

1. Основні поняття та формули.
2. Довірчі інтервали для нормально розподілених вибірок. приклади.
3. Довірчі інтервали для вибірок із невідомим законом розподілу. Метод багаторазової генерації (**bootstrap**).

Застосування статистичних критеріїв, засноване на порівнянні значення статистики, що спостерігається, з теоретичним, або *p-value* - з рівнем значущості. При цьому результатом тесту є відповідь із двома можливими варіантами: приймається нульова гіпотеза або відкидається. На відміну від статистичних критеріїв довірчі інтервали дозволяють оцінити діапазон істинних можливих значень параметра, що цікавить нас, а також оцінити і інтерпретувати точність результату.

1. Основні поняття і формули

Історично спочатку вчені-математики (зокрема Стюдент, Фішер, Пірсон та ін.) вивели розподіл статистики і знайшли її теоретичні квантильні значення (критичні точки). Потім на цій основі була розроблена та розвинена концепція оцінок та довірчих інтервалів. Розглянемо нормально розподілені вибірки.

1.1 Довірчий інтервал для середнього із генеральної сукупності

Формула довірчого інтервалу для середнього $\bar{X}$ - вибірки з n значень із середньоквадратичним відхиленням S має вигляд:

$$\bar{X} \pm t_{1-\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}}$$

де $t_{1-\frac{\alpha}{2}, n-1}$ - квантиль розподілу Стюдента при рівні значущості α .

1.2 Довірчий інтервал для медіани із генеральної сукупності

Для медіани m використовується формула:

$$m \pm z_{(1-\frac{\alpha}{2})} \cdot 1,253 \cdot \frac{S}{\sqrt{n}}.$$

де $z_{(1-\frac{\alpha}{2})}$ - квантиль стандартного нормального розподілу при рівні значущості α .

1.3 Довірчий інтервал для різниці середніх двох незалежних вибірок

Формула довірчого інтервалу для різниці середніх $\bar{X}_1$ і $\bar{X}_2$ двох незалежних вибірок розміром n_1 і n_2 при рівних дисперсіях має вигляд

$$(\bar{X}_1 - \bar{X}_2) \pm t_{1-\frac{\alpha}{2}, n_1+n_2-2} \cdot SE,$$

де стандартна помилка

$$SE = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

У випадку, коли дисперсії S_1^2 і S_2^2 різні для знаходження довірчого інтервалу використовується формула:

$$(\bar{X}_1 - \bar{X}_2) \pm t_{1-\frac{\alpha}{2}, k} \cdot SE,$$

де стандартна помилка

$$SE = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}},$$

А ступінь свободи k знаходимо із виразу

$$k = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}.$$

1.4 Довірчий інтервал щодо відносної частоти (частки) з генеральної сукупності

Для відносної частоти W використовується формула:

$$W \pm z_{(1-\frac{\alpha}{2})} \cdot \sqrt{\frac{W(1-W)}{n}}.$$

де $z_{(1-\frac{\alpha}{2})}$ - квантиль стандартного нормального розподілу при рівні значущості α .

1.5 Довірчий інтервал для відносних частот (часток) із двох незалежних вибірок

$$(W_1 \pm W_2) \pm z_{(1-\frac{\alpha}{2})} \cdot \sqrt{\frac{W_1(1-W_1)}{n_1} + \frac{W_2(1-W_2)}{n_2}},$$

де $z_{(1-\frac{\alpha}{2})}$ - квантиль стандартного нормального розподілу при рівні значущості α .

Зауваження. Якщо число 0 входить в довірчий інтервал різниці будь-яких параметрів, отже, нульова гіпотеза (відмінностей немає) сумісна з даними, що спостерігаються, відмінність статистично незначуща (**p – value** > α). Якщо довірчий інтервал для різниці часток не включає число 0, відмінність статистично значуща (**p – value** < α).

2. Довірчі інтервали для нормально розподілених вибірок. Приклади

2.1 Нехай із генеральної сукупності отримано дві нормально розподілені вибірки: **A** та **B** (табл. 10.1). Побудуємо довірчі інтервали середнього значення кожної вибірки окремо, а також довірчий інтервал різниці середніх значень вибірок **A** та **B**.

Таблиця №10.1

A	130,7	123,2	139,7	128,5	121,7	126,7	134,1	124,3	121,7	132,4
B	133,5	126,0	126,4	133,4	116,6	117,0	123,9	128,9	134,2	145,8

Перенесемо дані таблиці на аркуш програми **Excel** і з допомогою інструмента **Data Analysis** розрахуємо числові характеристики вибірок (рис.10.1).

	A	B	C	D	E	F	G	H
1	A	B		Виборка A			Виборка B	
2	130,7	133,5		Column1			Column1	
3	123,2	126						
4	139,7	126,4		Mean	128,30		Mean	128,57
5	128,5	133,4		Standard Error	1,88		Standard Error	2,76
6	121,7	116,6		Median	127,60		Median	127,65
7	126,7	117		Mode	121,70		Mode	#Н/Д
8	134,1	123,9		Standard Deviation	5,94		Standard Deviation	8,73
9	124,3	128,9		Sample Variance	35,32		Sample Variance	76,29
10	121,7	134,2		Kurtosis	-0,30		Kurtosis	0,51
11	132,4	145,8		Skewness	0,65		Skewness	0,44
12				Range	18		Range	29,2
13				Minimum	121,7		Minimum	116,6
14				Maximum	139,7		Maximum	145,8
15				Sum	1283		Sum	1285,7
16				Count	10		Count	10

Рисунок 10.1

Також для обчислення знадобиться квантиль *розподілу Стюдента* для рівня значущості $\alpha = 0,05$ та ступеня свободи $k = n - 1 = 10 - 1 = 9$. Для його знаходження використовуємо функцію **T.INV**. У полі **Імовірність** вводимо значення $1 - \frac{\alpha}{2} = 0,975$, а в поле **Ступінь_свободи** $k = 9$. Отримуємо $t_{0,975,9} = 2,26$ (рис. 10.2).

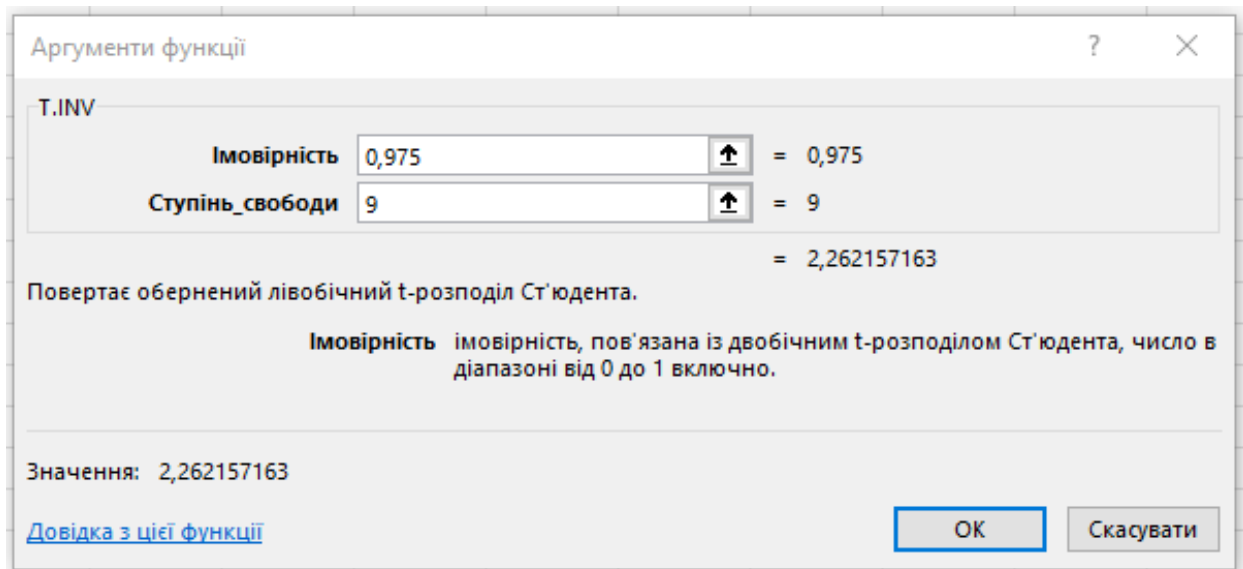


Рисунок 10.2

Для знаходження кордонів довірчого інтервалу використовуємо відповідну формулу:

$$\bar{X} \pm t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}}$$

Результат бачимо на рис.10.3, тобто для виборки **A** довірчий інтервал для середнього: **(124,05;132,55)**, для виборки **B** – **(122,32; 134,82)**. Оскільки довірчі інтервали перетинаються, можна зробити висновок, що відмінностей вибірок не виявлено і вони можуть належати одній чи близьким генеральним сукупностям.

Проведемо перевірку за довірчим інтервалом для різниці середніх. Для його розрахунку використовуємо формулу:

$$(\bar{X}_1 - \bar{X}_2) \pm t_{1-\frac{\alpha}{2}, k} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

D19

:

✕

✓

fx

=E\$4-\$B\$16*\$E\$8/SQRT(\$E\$16)

	A	B	C	D	E	F	G	H
1	A	B		Виборка А			Виборка В	
2	130,7	133,5		Column1			Column1	
3	123,2	126						
4	139,7	126,4		Mean	128,30		Mean	128,57
5	128,5	133,4		Standard Error	1,88		Standard Error	2,76
6	121,7	116,6		Median	127,60		Median	127,65
7	126,7	117		Mode	121,70		Mode	#Н/Д
8	134,1	123,9		Standard Deviation	5,94		Standard Deviation	8,73
9	124,3	128,9		Sample Variance	35,32		Sample Variance	76,29
10	121,7	134,2		Kurtosis	-0,30		Kurtosis	0,51
11	132,4	145,8		Skewness	0,65		Skewness	0,44
12				Range	18		Range	29,2
13				Minimum	121,7		Minimum	116,6
14				Maximum	139,7		Maximum	145,8
15				Sum	1283		Sum	1285,7
16	t=	2,26		Count	10		Count	10
17								
18				від	до		від	до
19				124,05	132,55		122,32	134,82

Рисунок 10.3

В результаті розрахунку в **Excel** знаходимо довірчий інтервал: **(-7,83;7,29)**. Довірчий інтервал для різниці включає **0**, тобто різниця між середніми статистично не значуща.

2.2 Проведено дослідження, в результаті якого виявлено, що у чоловіків **90** успішних результатів операцій із **100**, а у жінок - **80** із **100**. Побудуємо **95%**-й довірчий інтервал для різниці часток успішних операцій у чоловіків та жінок. Розрахунок проводимо за формулою

$$(W_1 \pm W_2) \pm z_{(1-\frac{\alpha}{2})} \cdot \sqrt{\frac{W_1(1-W_1)}{n_1} + \frac{W_2(1-W_2)}{n_2}},$$

де $z_{(1-\frac{\alpha}{2})}$ - квантиль стандартного нормального розподілу при рівні значущості α .

Квантиль знаходимо в **Excel** за допомогою функції **NORM.S.INV**. Для цього в полі **Імовірність** вводимо значення **0,975**.

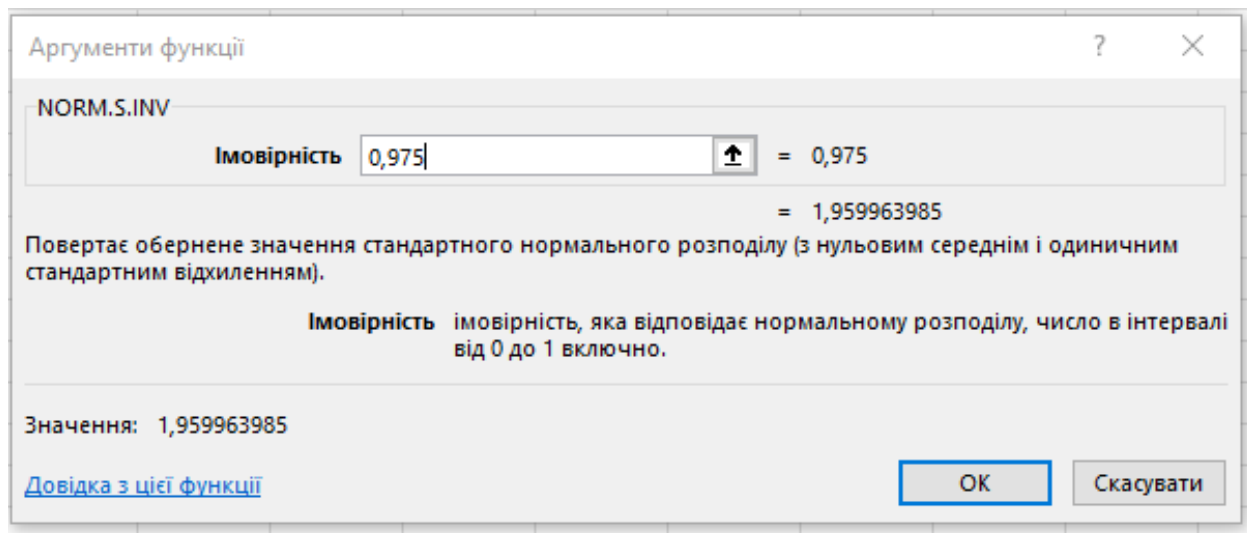


Рисунок 10.4

Знаходимо результат (рис.10.4): $z_{0,975} = 1,96$. Враховуючи наступні умови

$$W_1 = \frac{90}{100} = 0,9; \quad W_2 = \frac{80}{100} = 0,8; \quad n_1 = n_2 = 100,$$

за вищенаведеною формулою знаходимо довірчий інтервал для різниці долей. Він дорівнює: **(0,002;0,198)**. Тобто можемо стверджувати, що 95%-й довірчий інтервал для різниці долей успішних операцій між чоловіками та жінками складає від **0.2%** до **19.8%**. Оскільки інтервал не включає **0**, відмінність статистично значуща, і чоловіки мають більш високу частку успішних операцій із ймовірністю **95 %**.

3. Довірчі інтервали для вибірок із невідомим законом розподілу.

Метод багаторазової генерації (bootstrap)

Метод багаторазової генерації (бутстреп) застосовується у випадку, коли дані не відповідають нормальному розподілу або кількість значень вибірки мала. Нехай задана виборка из n значень

$$X = (x_1, x_2, \dots, x_n).$$

Ідея методу полягає у використанні вихідної вибірки, як заміника генеральної сукупності, для імітації багаторазового повторення експерименту.

Таким чином, формуються нові вибірки такого ж розміру n , причому кожне значення може потрапити у вибірку кілька разів, або жодного разу. Загальна кількість варіантів таких вибірок дорівнює n^n . На практиці використовується кількість вибірок від 1000 до 10000. Далі для кожної вибірки обчислюється числова характеристика, наприклад, середнє чи медіана. Потім одержаний розподіл статистики використовується для побудови довірчих інтервалів.

Розглянемо приклад побудови довірчого інтервалу медіани методом *бутстреп*, для вибірки рівня С-реактивного білка (мг/л) (табл. 10.2).

Таблиця №10.2

6,90	8,40	4,06	7,89	5,49	7,95	5,33	3,90	4,11	6,80
------	------	------	------	------	------	------	------	------	------

Розмістимо дані на аркуші програми **Excel**. У комірці **A2** розрахуємо нове випадкове значення, взяте із значень вибірки за допомогою функції **INDEX**. В якості аргументів цієї функції використовуємо: масив значень вибірки та випадковий номер значення у загальному списку (рис.10.5).

A2

✓

f_x

=INDEX(\$A\$1:\$J\$1;RANDBETWEEN(1;10))

	A	B	C	D	E	F	G	H	I	J
1	6,90	8,40	4,06	7,89	5,49	7,95	5,33	3,90	4,11	6,80
2	8,4									

Рисунок 10.5

A2

⌵

:

✖

✓

f_x

⌵

=INDEX(\$A\$1:\$J\$1;RANDBETWEEN(1;10))

	A	B	C	D	E	F	G	H	I	J
1	6,90	8,40	4,06	7,89	5,49	7,95	5,33	3,90	4,11	6,80
2	4,06	6,9	6,9	4,06	7,95	3,9	6,9	5,33	4,06	5,33

Рисунок 10.6

Продовжуємо формулу до стовпця **J** включно (рис.10.6). У другому ряду ми одержали новий варіант вибірки. Далі виділяємо діапазон другої вибірки та

розтягуємо формулу вниз, щоб згенерувати нові вибірки. Загальна кількість вибірок має бути не менше ніж 1000 (рис.10.7).

	A	B	C	D	E	F	G	H	I	J
1	6,90	8,40	4,06	7,89	5,49	7,95	5,33	3,90	4,11	6,80
2	5,49	6,9	6,8	3,9	4,11	4,11	5,49	8,4	8,4	4,06
3	7,95	4,11	8,4	4,11	7,95	5,33	5,49	5,33	6,8	7,95
4	5,33	4,06	7,89	7,89	7,89	3,9	8,4	6,9	3,9	4,06
5	6,9	6,8	3,9	7,95	5,49	7,95	4,06	7,89	7,95	5,33
6	8,4	5,49	3,9	7,95	3,9	6,9	6,9	4,11	4,11	6,8
7	7,89	5,33	7,89	5,49	4,11	6,8	7,95	7,89	5,49	4,06
996	5,33	4,06	3,9	7,89	5,33	5,49	6,9	7,89	7,95	3,9
997	6,8	5,49	5,49	4,11	6,9	5,33	5,49	7,89	5,49	7,89
998	5,49	5,33	6,8	6,8	5,33	4,11	7,95	6,8	4,11	5,33
999	8,4	7,89	7,95	8,4	7,95	7,89	5,33	6,8	4,06	5,49
1000	4,06	7,95	6,9	7,95	7,95	8,4	8,4	8,4	6,8	6,9

Рисунок 10.7

У стовпці **K** для кожної вибірки з допомогою функції **MEDIAN** розраховуємо значення медіани (рис.10.8).

K1

⌵

:

✕

✓

fx

⌵

=MEDIAN(A1:J1)

	A	B	C	D	E	F	G	H	I	J	K
1	6,90	8,40	4,06	7,89	5,49	7,95	5,33	3,90	4,11	6,80	6,145
2	5,33	3,9	5,33	5,33	8,4	5,33	8,4	3,9	7,95	7,89	5,330
3	5,33	4,11	5,49	4,11	7,89	6,8	4,06	6,8	7,89	8,4	6,145
4	7,89	5,49	6,9	4,11	7,95	6,8	4,11	7,95	8,4	5,49	6,850
5	4,06	3,9	6,9	5,49	7,89	7,95	5,49	8,4	6,8	6,9	6,850
6	8,4	4,06	5,33	4,11	5,33	6,9	4,06	6,9	4,11	5,49	5,330
7	6,9	7,95	4,11	4,06	5,33	3,9	4,11	7,95	8,4	5,49	5,410

Рисунок 10.8

Отже ми отримали розподіл *бутстреп*-медіан у кількості 1000 значень. Ліву та праву межі 95%-го довірчого інтервалу знаходимо за допомогою функції **PERCENTILE.INC**. Як аргументи задаємо: діапазон масиву медіан та значення ймовірності (для лівої межі – **0,025**, для правої – **0,975**).

=PERCENTILE.INC(\$K\$1:\$K\$1000;0,025)

D	E	F	G	H	I	J	K	L	M	N
7,89	5,49	7,95	5,33	3,90	4,11	6,80	6,145		від	до
4,06	4,06	7,95	5,49	5,49	7,89	7,89	5,490		4,085	7,890
7,95	5,33	4,11	3,9	7,95	4,06	5,49	5,330			

Рисунок 10.9

Таким чином, ми побудували довірчий інтервал: **(4,085;7,890)**, в якому з ймовірністю **0,95 (95%)** знаходиться медіана генеральної сукупності (рис.10.9).

Аналогічним чином можна збудувати довірчий інтервал для середнього. Для цього в стовпці **K** треба замість функції **MEDIAN** використовувати функцію **AVERAGE**.

Практичне завдання

Завдання 1

Побудуйте 95%-й довірчий інтервал для середнього значення у кожній групі та для різниці середніх між групами **A** і **B**. Дані представляють рівень гемоглобіну (г/л) у двох групах пацієнтів.

Варіант №1	A	138,0	146,7	129,1	129,1	147,3	139,2	126,8	136,9	126,9	126,8
	B	135,8	114,2	116,1	127,7	123,2	136,5	124,3	119,2	148	131,1
Варіант №2	A	124,8	131,3	118,7	134,0	124,2	127,3	124,2	148,7	130,1	119,6
	B	139,0	118,6	132,9	111,2	117,5	132,7	138,2	132,5	129,6	127,8
Варіант №3	A	121,0	136,1	129,0	107,9	128,8	121,7	118,8	131,7	135,9	134,9
	B	118,5	123,8	130,2	136,6	122,1	125,0	115,8	114,9	135,0	140,4
Варіант №4	A	133,4	123,3	133,4	145,2	129,4	145,4	103,6	138,0	130,7	126,8
	B	133,7	112,9	130,6	136,4	147,6	127,6	124,7	127,8	141,9	136,1
Варіант №5	A	129,4	138,1	121,4	125,1	124,5	113,8	131,4	131,0	128,5	126,1
	B	116,8	126,7	127,5	122,9	129,3	135,0	149,8	132,7	133,5	130,2
Варіант №6	A	124,8	148,9	122,3	127,3	123,9	112,6	135,7	131,8	132,2	115,1
	B	140,2	112,2	132,1	148,1	116,3	120,6	127,2	121,2	110,7	126,9
Варіант №7	A	117,6	142,3	119,0	123,6	134,9	114,5	129,1	139,9	110,7	128,7
	B	131,9	137,1	116,9	116,1	134,5	132,3	131,8	132,8	122,5	131,6

Варіант №8	A	149,5	135,6	119,0	137,4	121,1	138,8	142,5	122,7	140,5	135,0
	B	140,4	151,1	129,7	124,6	123,3	124,0	131,4	135,6	134,9	140,4
Варіант №9	A	127,4	157,2	136,3	121,5	119,3	134,9	127,8	137,2	134,8	129,3
	B	125,0	118,3	129,0	142,1	135,6	121,0	135,2	137,3	124,7	135,0
Варіант №10	A	133,8	135,8	141,0	140,7	116,4	120,8	135,3	135,3	135,3	168,7
	B	136,7	142,4	140,6	137,5	127,9	138,6	123,3	128,7	126,2	131,9
Варіант №11	A	143,8	120,8	132,2	147,8	137,6	126,2	129,8	143,7	129,6	139,1
	B	137,5	130,6	158,5	143,4	116,8	138,9	130,5	145,6	129,2	135,9
Варіант №12	A	119,5	128,2	126,8	125,0	149,2	135,6	118,9	140,7	152,7	141,8
	B	119,2	129,5	147,0	127,3	138,8	142,1	125,1	133,8	102,0	124,1
Варіант №13	A	145,6	114,9	124,8	130,5	143,7	114,9	140,9	129,3	119,4	133,9
	B	132,0	124,0	130,7	126,1	131,1	136,6	145,9	117,6	151,3	110,5
Варіант №14	A	132,4	123,3	127,5	124,6	123,7	138,0	133,1	122,6	138,5	132,6
	B	140,3	138,4	123,8	126,5	139,6	138,2	131,9	133,3	144,9	126,2
Варіант №15	A	129,5	142,6	139,9	139,8	144,7	131,9	138,5	128,5	134,9	130,3
	B	134,4	139,4	125,3	154,4	123,4	121,3	145,0	141,4	139,7	139,7

Завдання 2

У клінічному дослідженні оцінюється ефективність нової терапії. У групі 1 зі 100 чоловіків успішно проліковано 40+N пацієнтів (N – номер варіанту). У групі зі 100 жінок - 35 пацієнтів. Потрібно: розрахувати 95% довірчий інтервал різниці часток успіху між групами.

Завдання 3

Використовуючи метод багаторазової генерації бутстреп (*bootstrap*) (1000 псевдовибірок), побудуйте 95%-й довірчий інтервал для середнього та медіани рівня С-реактивного білка (мг/л).

Варіант №1	7,67	8,00	5,96	8,01	7,38	8,87	6,28	9,41	12,75	7,79
Варіант №2	9,46	6,82	8,23	7,72	7,92	6,10	7,75	8,93	11,89	10,26
Варіант №3	7,28	11,91	9,68	17,68	7,19	7,12	7,66	4,85	7,83	7,30
Варіант №4	8,31	13,17	9,98	6,31	5,73	8,69	5,05	12,99	10,69	6,52
Варіант №5	9,35	6,02	9,03	3,86	5,20	6,24	6,31	5,44	7,50	4,52
Варіант №6	7,55	8,50	9,02	5,20	4,60	10,69	8,05	5,82	11,60	7,54
Варіант №7	8,48	15,43	14,08	7,72	11,13	10,09	12,53	6,22	10,21	11,42

Варіант №8	4,35	3,36	5,72	7,69	9,73	6,34	10,10	4,10	3,72	6,09
Варіант №9	7,60	4,13	7,48	5,19	9,39	8,57	5,79	6,58	5,59	7,54
Варіант №10	6,05	9,46	6,93	6,41	7,87	5,96	6,89	5,68	14,66	8,22
Варіант №11	7,35	6,66	6,45	8,28	8,65	5,88	5,80	6,34	3,45	4,37
Варіант №12	13,88	7,86	10,07	9,30	21,34	11,85	8,15	6,36	5,23	9,00
Варіант №13	4,47	5,64	4,95	11,36	8,93	6,83	10,68	7,01	5,29	10,82
Варіант №14	5,71	7,37	6,00	5,15	10,30	13,83	5,13	9,23	6,42	6,74
Варіант №15	5,37	7,07	5,43	7,55	6,86	6,48	5,30	5,86	8,74	8,09

ПРАКТИЧНА РОБОТА 11

Тема №11. Регресійний и кореляційний аналіз

Питання, які розглядаються:

1. Регресійний аналіз. Приклад.
2. Кореляційний аналіз. Коефіцієнт Пірсона.
3. Кореляційний аналіз. Коефіцієнт Спірмена. Приклад.
4. Кореляційний аналіз. Коефіцієнт Кендалла. Приклад.

Регресійний та кореляційний аналізи є методами, що дозволяють виявити та оцінити кількісно зв'язок між показниками, які є значеннями вибірок. Це особливо актуально при медичних дослідженнях для встановлення наявності залежностей між фізіологічними, лабораторними та клінічними показниками, прогнозування результатів лікування, перебігу хвороби, ефективності терапії, а також для обґрунтування діагностичних критеріїв.

1. Регресійний аналіз

Нехай задані дві виборки $X = (x_1, x_2, \dots, x_n)$ и $Y = (y_1, y_2, \dots, y_n)$. Кожній парі чисел $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ в Декартовій системі координат на площині відповідає певна крапка (рис.11.1).

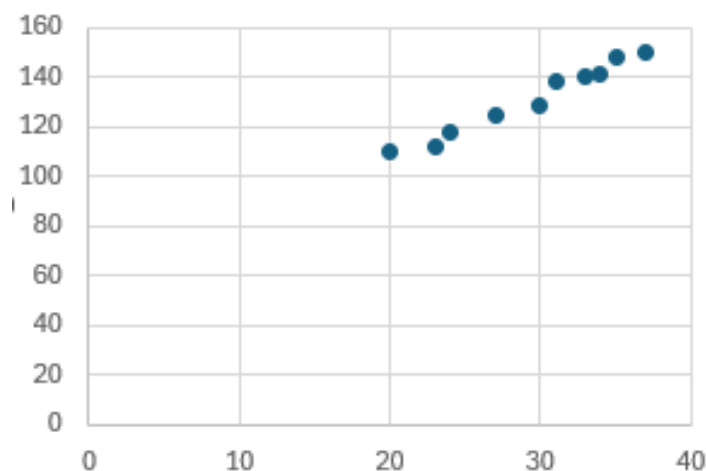


Рисунок 11.1

Потрібно побудувати пряму лінію таким чином, щоб мінімізувати вертикальні відстані від заданих точок до прямої. Оскільки пряму на поверхні описує лінійна функція $y = a + bx$, таке завдання відноситься до лінійної регресії та вирішується методом найменших квадратів.

Для реалізації методу з ординат точок вибірки віднімаються відповідні ординати точок прямої: $y_i - (a + bx_i)$. Щоб позитивні та негативні різниці при підсумовуванні не компенсували одна одну, різниці зводяться у квадрат і після цього підсумовуються. Отриманий вираз є деякою функцією, яка залежить від двох поки невідомих параметрів a і b .

$$F(a, b) = \sum_{i=1}^n (y_i - (a + bx_i))^2.$$

Щоб знайти значення a і b , при яких функція $F(a, b)$ приймає найменше значення, застосовується апарат математичного аналізу. Знаходяться частинні похідні і потім прирівнюються до нуля.

$$\begin{cases} \frac{\partial F}{\partial a} = 0, \\ \frac{\partial F}{\partial b} = 0 \end{cases} \Rightarrow \begin{cases} 2 \sum_{i=1}^n (ax_i + bx_i^2 - x_i y_i) = 0, \\ 2 \sum_{i=1}^n (a + bx_i - y_i) = 0. \end{cases}$$

Після перетворень отримуємо систему лінійних алгебраїчних рівнянь відносно параметрів a і b .

$$\begin{cases} a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i, \\ an + b \sum_{i=1}^n x_i = \sum_{i=1}^n y_i. \end{cases}$$

Визначаючи із системи значення a і b отримуємо рівняння прямої, яка є **рівнянням лінійної регресії**. Якщо побудувати графік цієї прямої, його називають **лінією тренда**. Коефіцієнт b рівняння показує, наскільки у середньому змінюються значення виборки Y при зміні значень виборки X на одиницю.

Після того як рівняння лінійної регресії $y = a + bx$ отримано, важливо з'ясувати, наскільки достовірно, що зв'язок між X і Y існує, а не вийшов випадково. І тому сформулюємо наступні гіпотези:

- $H_0: b = 0$, тобто між X і Y немає лінійного зв'язку;
- $H_1: b \neq 0$, лінійний зв'язок є – нахил статистично значущий.

За відомими значеннями кутового коефіцієнта b і стандартної похибки SE знаходимо значення статистики:

$$t = \frac{b - 0}{SE} = \frac{b}{SE},$$

де

$$SE = \frac{s}{\sqrt{\sum_{i=1}^n (x_i - \bar{X})^2}}, \quad s = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{Y})^2}{n - 2}}.$$

Значення t підкоряється t -розподілу із $n - 2$ ступенями свободи.
Зауваження. Треба відмітити, що ***t-тест*** є окремим випадком лінійної регресії. При порівнянні середніх застосовується ***t-тест*** для двох груп, наприклад, **Group 1** и **Group 2**, а при лінійній регресії ми будемо статистику, порівнюючи кутовий коефіцієнт b прямої з нулем.

Якщо розглянути окремий випадок і в якості значення вибірки **Group 1** взяти ординати точок регресії $(0, y_1), (0, y_2), \dots, (0, y_{n_1})$, а в якості значень виборки **Group 2** - $(1, y_1), (1, y_2), \dots, (1, y_{n_2})$, то можна помітити наступне (рис. 11.2).

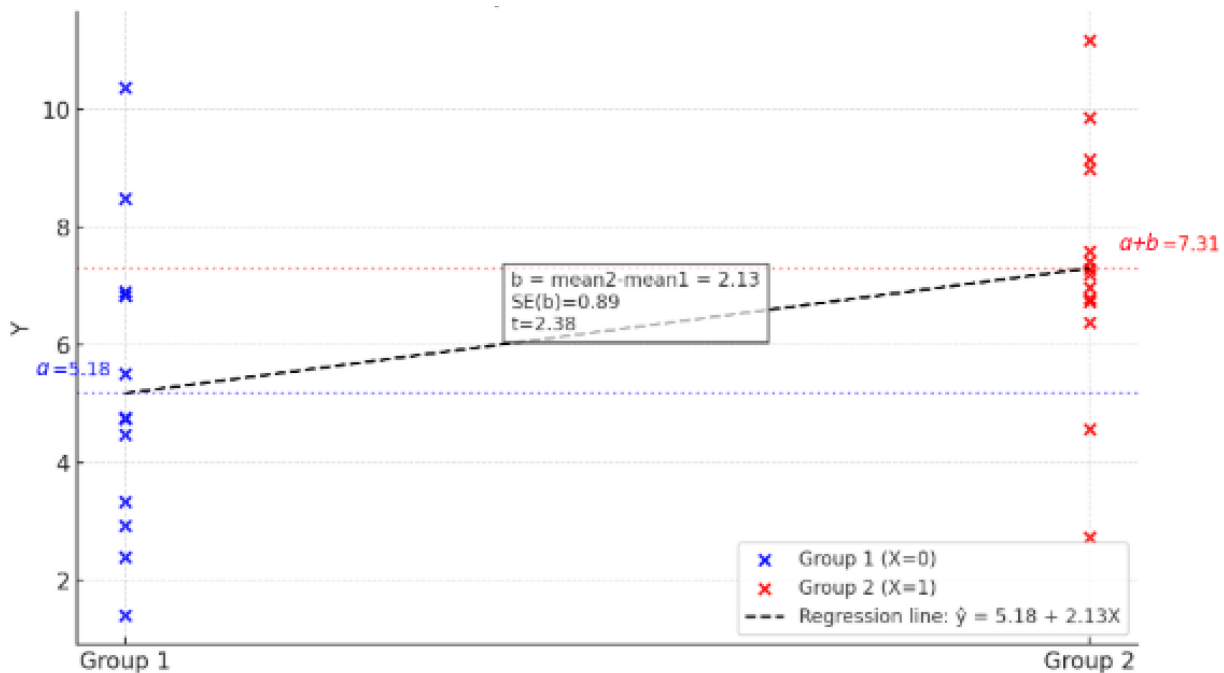


Рисунок 11.2

Нехай середня вибірка **Group 1** дорівнює a . Середня вибірка **Group 2** - $(a + b)$. Тоді очевидно, що пряма, яка найближче розташована до заданих точок (пряма регресії), проходить через точки з координатами $(0, a)$ і $(1, a + b)$. Кутовий коефіцієнт цієї прямої визначається через тангенс кута нахилу прямої. Він буде дорівнювати b . Але оскільки різниця середніх теж дорівнює

$$(a + b) - a = b,$$

можна дійти висновку про те, що **t-тест** - це окремий випадок лінійної регресії. Тобто **t-тест** та тест для лінійної регресії ідентичні, **t-тест** перевіряє чи різні середні вибірок, а тест для лінійної регресії - чи є лінійна залежність між значеннями вибірок.

Розглянемо приклад, на якому буде показано, як за допомогою засобів **Excel** знайти рівняння лінійної регресії та провести тест на достовірність зв'язку між вибірками.

Таблиця 11.1

Кількістю супутніх захворювань	4	4	5	4	2	3	1	2	4	5
Тривалість госпіталізації (дні)	8,4	8,8	8,2	7,8	1,1	5,2	3,2	3,5	9,8	7,6

Перенесемо дані таблиці 11.1 на аркуш програми **Excel** (рис.11.3).

	A	B
1	4	8,4
2	4	8,8
3	5	8,2
4	4	7,8
5	2	1,1
6	3	5,2
7	1	3,2
8	2	3,5
9	4	9,8
10	5	7,6

Рисунок 11.3

Для визначення параметрів прямої a і b лінійної регресії в **Excel** існує два варіанти.

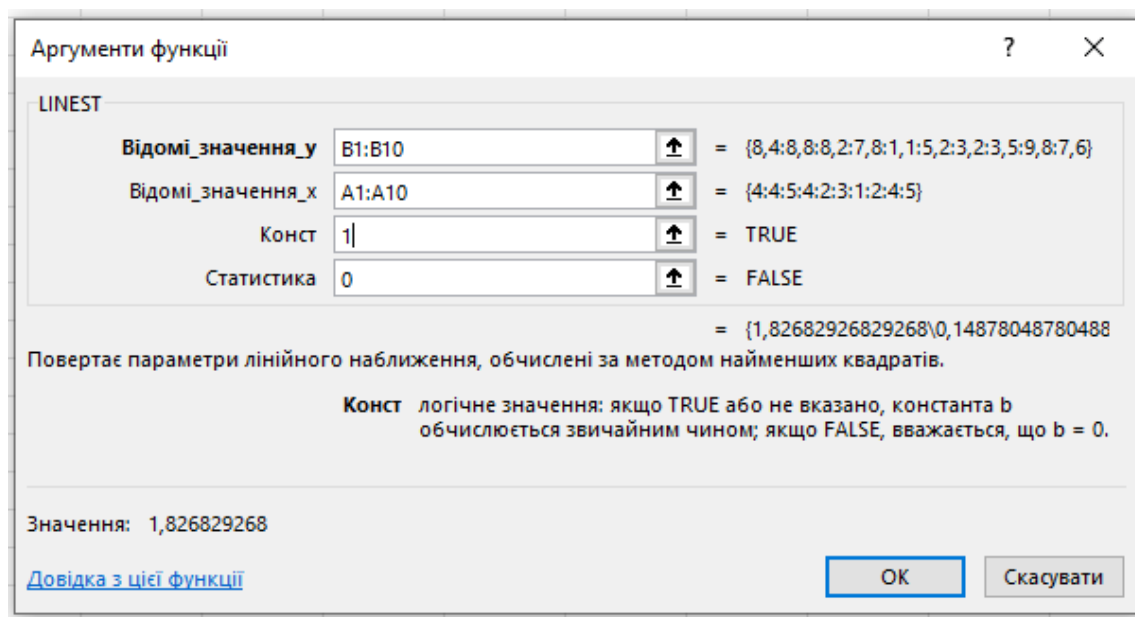


Рисунок 11.4

1 варіант. Використовуємо функцію **LINEST** (рис. 11.4). В першому полі розміщуємо діапазон значень виборки Y, в другому - діапазон значень виборки X. В третьому полі, щоб отримати обидва параметри прямої, вказуємо **1**. В четвертому

полі, якщо ввести **1**, разом з параметрами прямої a і b , будуть розраховані інші показники, включаючи статистику. Оскільки розширений варіант розрахунків доцільніше робити за допомогою пакету **Data Analysis**, в четвертому полі вводимо **0** і **ОК**. В результаті знаходимо параметри a і b (рис.11.5).

D	E
b	a
1,83	0,15

Рисунок 11.5

Тобто рівняння прямої лінійної регресії має вигляд:

$$y = 0,15 + 1,83x.$$

2 варіант. Через пункт меню **Дані** відкриваємо пакет **Data Analysis** і потім – **Regression** (рис.11.6).

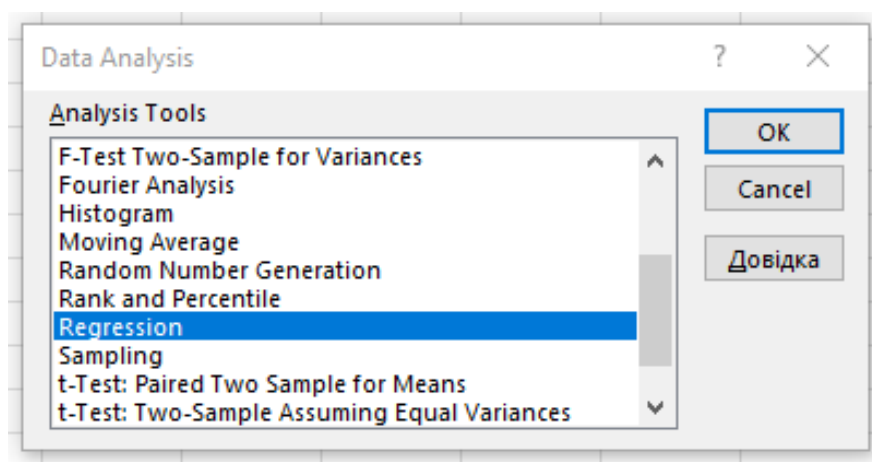


Рисунок 11.6

Відкриється вікно для введення даних (рис.11.7). У полі **Input Y Range** вказуємо діапазон значень вибірки Y , у полі **Input X Range** - діапазон значень вибірки X , відзначаємо пункт **Output Range**, у полі навпроти вказуємо адресу клітинки для виведення підсумкової таблиці та натискаємо **ОК**.

Рисунок 11.7

Підсумкова таблиця наведена на рис.11.8. Вона складається із трьох блоків (табл.11.2).

H	I	J	K	L	M	N	O	P
SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0,84792821							
R Square	0,71898225							
Adjusted R Square	0,68385503							
Standard Error	1,63524444							
Observations	10							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	1	54,73180488	54,7318	20,46795	0,001939259			
Residual	8	21,39219512	2,674024					
Total	9	76,124						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95,0%	Upper 95,0%
Intercept	0,14878049	1,467059485	0,101414	0,921717	-3,234264752	3,5318257	-3,2342648	3,53182573
X Variable 1	1,82682927	0,403794834	4,524152	0,001939	0,895676712	2,7579818	0,89567671	2,75798182

Рисунок 11.8

Таблиця 11.2

Блок 1 — Regression Statistics		
Показник	Розшифровка	Що означає
Multiple R	Коефіцієнт кореляції	Вказує на наявність зв'язку та його спрямованість.
R Square (R²)	Коефіцієнт детермінації	Частка варіації Y, що пояснюється X. Наприклад, R² = 0.80 означає, що 80% розкиду Y пояснюється моделлю.
Adjusted R Square	Скорельований R²	Те ж саме, але з поправкою на кількість змінних (важливо при множинній регресії). Для однієї змінної майже збігається з R².
Standard Error	Стандартна помилка регресії (s)	Середня помилка передбачення Y. Чим менше, тим краще припасування.
Observations	Кількість спостережень (n)	Скільки рядків даних брало участь.
Блок 2 — ANOVA (аналіз дисперсії)		
Стовпець	Що показує	
df	Ступені свободи: для Regression (1 при простій регресії), Residual (n-2), Total (n-1)	
SS (Sum of Squares)	Сума квадратів відхилень - пояснена (Regression SS), залишкова (Residual SS) і загальна (Total SS).	
MS (Mean Square)	Середні квадрати = SS/df.	
F	F-статистика (тест значущості всієї моделі).	
Significance F	p-value для F-тесту. Якщо <0.05 → модель загалом статистично	
Блок 3 — Coefficients (коефіцієнти моделі)		
Показник	Розшифровка	Інтерпретація
Intercept	Константа (a)	Значення Y при X = 0.
X Variable 1	Нахил (b)	Як змінюється Y зі збільшенням X на 1 одиницю.
Standard Error	SE(a), SE(b)	Стандартна помилка відповідного коефіцієнта.
t Stat	t = Коефіцієнт / SE	Показує, наскільки сильно коефіцієнт відрізняється від 0.
P-value	Імовірність випадково отримати таке t при H₀: коефіцієнт = 0.	Якщо < 0.05 → коефіцієнт значимий.
Lower 95% / Upper 95%	95% довірчий інтервал для коефіцієнта	Діапазон можливих істинних значень коефіцієнта з 95% впевненістю. Якщо інтервал не включає 0 → коефіцієнт значимий.

Для основних висновків достатньо знань параметрів прямої a і b , а також значення $p\text{-value}$. Як і в першому варіанті $b = 1,83$, $a = 0,15$, тобто рівняння

прямої регресії $y = 0,15 + 1,83x$. Оскільки кутовий коефіцієнт $b > 0$, при зростанні значень вибірки X , значення вибірки Y також зростають.

Оскільки $p - value = 0,0019 < 0,05$, нахил (кутовий коефіцієнт b) є статистично значущим. Це означає, що існує достовірна лінійна залежність між кількістю супутніх захворювань та тривалістю госпіталізації. Тобто чим більше кількість супутніх захворювань, тим довше пацієнт перебуває у стаціонарі.

Для візуалізації збудуємо графік розсіювання з лінією тренду (рис.11.9). Для цього виділяємо таблицю даних, у пункті **Вставлення** вибираємо точкову діаграму. За допомогою інструментів форматування діаграми до області побудови можна додати лінію тренда.

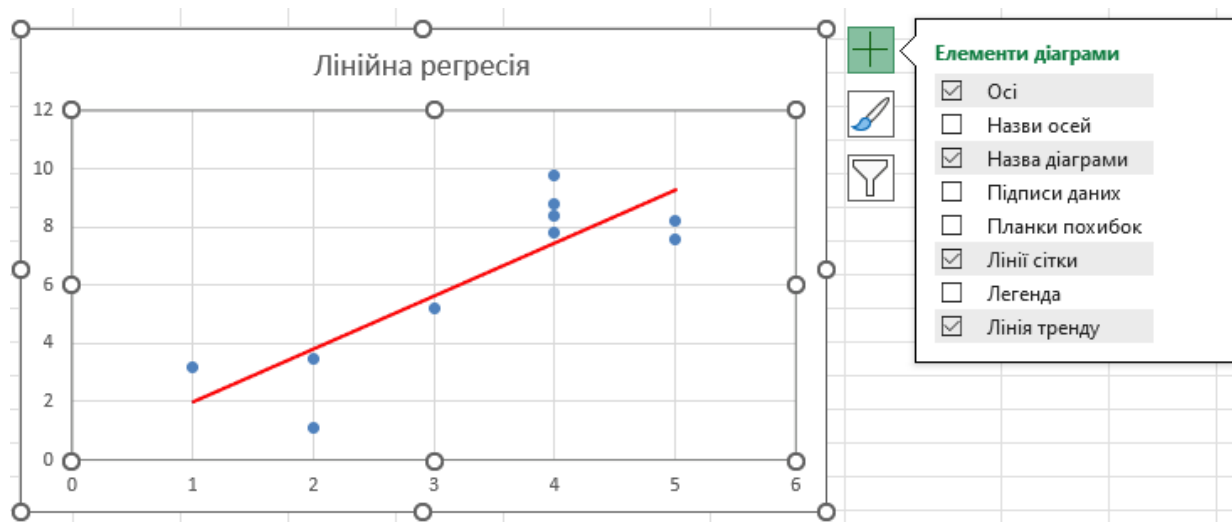


Рисунок 11.9

2. Кореляційний аналіз. Коефіцієнт Пірсона

Кореляція - це статистична залежність між значеннями вибірок, що показує ступінь спільної зміни показників. На відміну від регресії при кореляції вибірки рівноправні. Розглянемо історичну хронологію цього поняття.

На початку 19 століття вченими *Лежандром* та *Гаусом* було розроблено метод найменших квадратів. Як було розглянуто вище, кутовий коефіцієнт b показує наскільки змінюються значення вибірки Y при зміні значень вибірки X на одиницю. Наприклад, на скільки збільшиться тиск, якщо маса тіла зросте на 1 кг/м^2 .

Пізніше, наприкінці 19 століття математик **Пірсон**, щоб прибрати одиниці виміру та порівняти різні залежності, запропонував нормувати коефіцієнт регресії b на розкид значень вибірок. Таким чином, він отримав **коефіцієнт кореляції**

$$r = b \cdot \frac{S_x}{S_y},$$

де S_x і S_y стандартні квадратичні відхилення вибірок X та Y відповідно.

У розгорнутому вигляді вираз для коефіцієнта кореляції має вигляд:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{X})^2 (y_i - \bar{Y})^2}}.$$

Оскільки коефіцієнт кореляції показує нормовану залежність, його значення перебуває у проміжку $[-1; 1]$. Якщо $r > 0$, то говорять про прямий зв'язок між вибірками, тобто зі збільшенням значень однієї вибірки, також збільшуються значення іншої. При $r < 0$ – зворотний зв'язок, коли зі збільшенням значень однієї вибірки значення іншої вибірки зменшуються. Залежність значень коефіцієнта кореляції із спрямованістю та ступенем зв'язку представлена в табл. 11.3.

Таблиця 11.3

Значення коефіцієнта кореляції r	Направленість зв'язку	Сила зв'язку
$[-1; -0,76]$	зворотна	дуже сильна
$[-0,75; -0,51]$	зворотна	сильна
$[-0,5; -0,26]$	зворотна	помірна
$[-0,25; 0]$	зворотна	слабка чи зв'язку немає
$[0; 0,25]$	пряма	слабка чи зв'язку немає
$[0,26; 0,5]$	пряма	помірна
$[0,51; 0,75]$	пряма	сильна
$[0,75; 1]$	пряма	дуже сильна

В програмі **Excel** є два варіанти визначення коефіцієнта кореляції. Перший за допомогою функції **CORREL**, другий – за допомогою пакету **Data Analysis**. В

пакеті **Data Analysis** можна використовувати розділ **Correlation** (рис.11.10). Результат його застосування такий самий як при функції **CORREL**.

Доцільніше у пакеті **Data Analysis** використовувати розділ **Regression** (рис. 11.6). Вихідні дані при застосуванні цієї опції докладно описані в попередньому розділі (табл. 11.2). Зокрема, вихідна таблиця містить значення коефіцієнта кореляції r та $p - value$ для оцінки статистичної значущості. Так, наприклад з попереднього розділу $r = 0,85$, $p - value = 0,0019$ (рис.11.11).

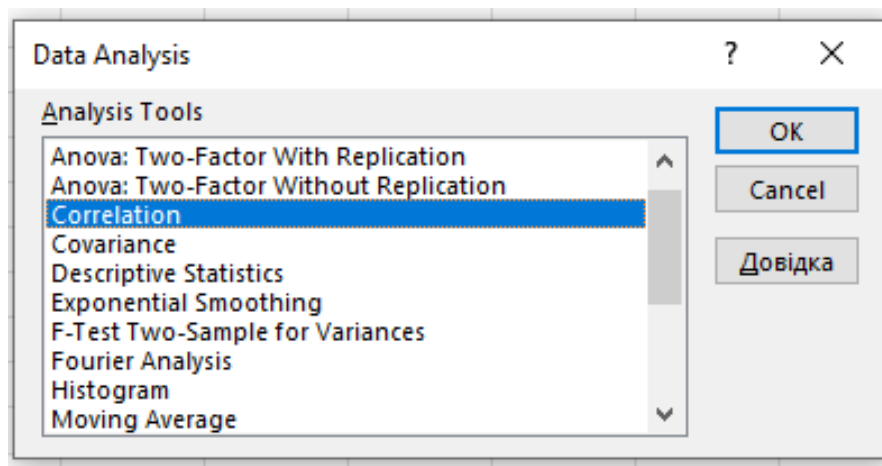


Рисунок 11.10

	H	I	J	K	L	M	N	O	P
SUMMARY OUTPUT									
<i>Regression Statistics</i>									
Multiple R		0,84792821							
R Square		0,71898225							
Adjusted R Square		0,68385503							
Standard Error		1,63524444							
Observations		10							
<i>ANOVA</i>									
		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression		1	54,73180488	54,7318	20,46795	0,001939259			
Residual		8	21,39219512	2,674024					
Total		9	76,124						
		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95,0%</i>	<i>Upper 95,0%</i>
Intercept		0,14878049	1,467059485	0,101414	0,921717	-3,234264752	3,5318257	-3,2342648	3,53182573
X Variable 1		1,82682927	0,403794834	4,524152	0,001939	0,895676712	2,7579818	0,89567671	2,75798182

Рисунок 11.11

Оскільки $r = 0,85 \in [0,76; 1]$ та $p - value = 0,0019 < 0,05$, слід, що між числом супутніх захворювань і тривалістю госпіталізації існує дуже сильний

прямий статистично значущий зв'язок: чим більша кількість супутніх захворювань, тим довше пацієнт перебуває у лікарні.

3. Кореляційний аналіз. Коефіцієнт Спірмена

Кореляція Спірмена - це непараметрична міра зв'язку між двома кількісними або порядковими змінними. Вона ґрунтується на рангах, а не на вихідних значеннях вибірки.

Коефіцієнт Спірмена розраховують у випадках, коли дані не розподілені нормально, змінні мають порядковий характер, зв'язок між значеннями вибірок приблизно монотонний, кількість значень вибірок менше **30**, є викиди.

Для визначення **коефіцієнта кореляції Спірмена ρ** слід виконати такі дії:

- 1) обчислити ранги вибірок R_X и R_Y .
- 2) застосувати формулу Пірсона для рангів

$$\rho = \frac{\sum_{i=1}^n (R_{Xi} - \overline{R_X}) (R_{Yi} - \overline{R_Y})}{\sqrt{\sum_{i=1}^n (R_{Xi} - \overline{R_X})^2 (R_{Yi} - \overline{R_Y})^2}}$$

- 3) якщо необхідно перевірити статистичну значущість результату, t-статистику, як і для кореляції Пірсона, знаходять за формулою

$$t = \rho \sqrt{\frac{n-2}{1-\rho^2}},$$

і потім порівнюють з теоретичним значенням із заданими рівнем значущості та ступенем свободи ($n-2$). Якщо обчислена статистика менша від теоретичної, приймають нульову гіпотезу.

Зауваження. Якщо ранги значень вибірок не повторюються, **коефіцієнт Спірмена** можна знайти за спрощеною формулою:

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2-1)},$$

де $d_i = R_{Xi} - R_{Yi}$.

Залежність значень *коефіцієнта кореляції Спірмена* із спрямованістю та ступенем зв'язку представлена в табл. 11.4.

Таблиця 11.4

Значення коефіцієнта кореляції ρ	Направленість зв'язку	Сила зв'язку
$[-1; -0,8]$	зворотна	дуже сильна
$[-0,79; -0,6]$	зворотна	сильна
$[-0,59; -0,4]$	зворотна	помірна
$[-0,39; 0,2]$	зворотна	слабка
$[-0,19; 0]$	зворотна	дуже слабка
$[0; 0,19]$	пряма	дуже слабка
$[0,2; 0,39]$	пряма	слабка
$[0,4; 0,59]$	пряма	помірна
$[0,6; 0,79]$	пряма	сильна
$[0,8; 1]$	пряма	дуже сильна

Розглянемо приклад, як за допомогою засобів **Excel** знайти *коефіцієнт кореляції Спірмена* та зробити висновок про зв'язок між вибірками. У таблиці 11.5 подано значення двох вибірок.

Таблиця 11.5

Частота нападів мігрені за місяць	2	5	4	7	9	8	3	6	4	7
Інтенсивність головного болю (0–10)	3	5	4	7	9	8	3	6	4	7

Перенесемо значення вибірок на аркуш програми **Excel** (рис.11.12).

	A	B
1	X	Y
2	2	3
3	5	5
4	4	4
5	7	7
6	9	9
7	8	8
8	3	3
9	6	6
10	4	4
11	7	7

Рисунок 11.12

За допомогою функції **RANK.AVG** розраховуємо ранги вибірок (рис.11.13).

C2		✕ ✓ <i>fx</i>		=RANK.AVG(A2;A\$2:A\$11;1)			
	A	B	C	D	E	F	
1	X	Y	R _x	R _y			
2	2	3	1	1,5			
3	5	5	5	5			
4	4	4	3,5	3,5			
5	7	7	7,5	7,5			
6	9	9	10	10			
7	8	8	9	9			
8	3	3	2	1,5			
9	6	6	6	6			
10	4	4	3,5	3,5			
11	7	7	7,5	7,5			

Рисунок 11.13

Потім відкриваємо пакет **Data Analysis**, вибираємо **Regression**, у перші два поля вводимо діапазони значень вибірок, відзначаємо пункт **Output Range**, в полі

навпаки вказуємо адресу комірки для виведення підсумкової таблиці та натискаємо **ОК** (рис.11.14).

Рисунок 11.14

F	G	H	I	J	K	L	M	N
SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0,9969278							
R Square	0,993865							
Adjusted R Square	0,9930982							
Standard Error	0,25							
Observations	10							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	1	81	81	1296	3,883E-10			
Residual	8	0,5	0,0625					
Total	9	81,5						
Coefficients								
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95,0%	Upper 95,0%
Intercept	0	0,172020491	0	1	-0,39668	0,39667996	-0,39667996	0,396679964
X Variable 1	1	0,027777778	36	3,88E-10	0,9359443	1,06405567	0,93594433	1,06405567

Рисунок 11.15

За результатами розрахунку визначено $\rho = 0,997$; $t = 36$; $p - value = 3,88 \times 10^{-10}$ (рис.11.15).

За отриманими результатами можна зробити наступні висновки:

- оскільки $\rho = 0,997 \in [0,8; 1]$ - між частотою мігрени та інтенсивністю головного болю існує дуже сильний прямий зв'язок;
- оскільки $p - value = 3,88 \times 10^{-10} < 0,05$ зв'язок статистично значущий.

4. Кореляційний аналіз. Коефіцієнт Кендала

Кореляція Кендала є непараметричною мірою зв'язку між двома змінними x і y . Запропонований метод ґрунтується на порівнянні рангів. Він показує, наскільки узгоджено змінюються дві змінні. **Кореляція Кендала** використовується, коли: дані представлені у вигляді рангів, розподіл не є нормальним, багато однакових значень (пов'язаних рангів), малі вибірки, досліджуються монотонні, але не обов'язково лінійні зв'язки.

Нехай у кожного з n об'єктів є по два спостереження: $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Тобто одне із спостережень задано вибіркою $X = (x_1, x_2, \dots, x_n)$, а друге - вибіркою $Y = (y_1, y_2, \dots, y_n)$. **Кореляція Кендала** встановлює, чи є зв'язок між вибірками X та Y .

Для цього складаються різні комбінації пар об'єктів (по два в кожній комбінації) і порівнюються їх спостереження. Загальна кількість таких пар дорівнює

$$N = \frac{n(n-1)}{2}.$$

Далі обчислюються:

- C - кількість узгоджених пар у випадку, коли обидва спостереження одного об'єкта більші за відповідні спостереження іншого об'єкта.
- D - кількість неузгоджених пар у випадку, коли одне із спостережень об'єкта більше відповідного спостереження іншого об'єкта, а друге спостереження менше.

- T_x - число пар, пов'язаних по X , тобто кількість об'єктів, у яких значення вибірки X збігаються.
- T_y - число пар, пов'язаних по Y , тобто кількість об'єктів, у яких значення вибірки Y збігаються.

Остаточно *коефіцієнт Кендала* обчислюється за формулою:

$$\tau = \frac{C - D}{\sqrt{(C + D + T_x)(C + D + T_y)}}$$

Залежність значень коефіцієнта кореляції Кендала з спрямованістю та ступенем зв'язку наведена в табл. 11.5.

Для того, щоб зробити остаточний висновок щодо залежності вибірок слід провести перевірку статистичної значущості. Нульова гіпотеза $H_0: \tau = 0$ – зв'язок відсутній. Альтернативна гіпотеза $H_1: \tau \neq 0$ – зв'язок між спостереженнями є значимим.

Таблиця 11.5

Значення коефіцієнта кореляції ρ	Направленість зв'язку	Сила зв'язку
$[-1; -0,8]$	зворотна	дуже сильна
$[-0,79; -0,6]$	зворотна	сильна
$[-0,59; -0,4]$	зворотна	помірна
$[-0,39; 0,2]$	зворотна	слабка
$[-0,19; 0]$	зворотна	дуже слабка
$[0; 0,19]$	пряма	дуже слабка
$[0,2; 0,39]$	пряма	слабка
$[0,4; 0,59]$	пряма	помірна
$[0,6; 0,79]$	пряма	сильна
$[0,8; 1]$	пряма	дуже сильна

Формула для Z – статистики має вигляд:

$$Z = \frac{\tau \sqrt{9n(n-1)}}{\sqrt{2(2n+5)}}.$$

Оскільки коефіцієнт τ може приймати як позитивні і негативні значення, розглядається двосторонній критерій. Критичне значення Z_{theory} визначається через стандартний нормальний розподіл. Після порівняння, якщо:

- $Z < Z_{theory}$, приймається нульова гіпотеза про те, що зв'язок відсутній;
- $Z \geq Z_{theory}$, приймається альтернативна гіпотеза - зв'язок є.

Розглянемо алгоритм обчислення **коефіцієнта Кендала** на конкретному прикладі. Для п'яти пацієнтів отримано дані рівня вітаміну D та щільності кісткової тканини (табл. 11.6).

Таблиця 11.6

Пацієнт	A	B	C	D	E
Рівень вітаміну D (нг/мл)	20	30	25	35	22
Щільність кісткової тканини	1,00	1,20	1,10	1,30	1,05

Потрібно встановити, чи існує зв'язок між цими даними. Складемо всілякі комбінації пар пацієнтів та порівняємо їх показники (табл. 11.7).

Оскільки $n = 5$, загальна кількість пар

$$N = \frac{n(n-1)}{2} = \frac{5 \cdot 4}{2} = 10.$$

Порівняння показало, що всі пари узгоджені, тобто $C = 10$, $D = 0$, $T_x = 0$, $T_y = 0$. Тоді **коефіцієнт Кендала**

$$\tau = \frac{C - D}{\sqrt{(C + D + T_x)(C + D + T_y)}} = \frac{10 - 0}{\sqrt{(10 + 0 + 0)(10 + 0 + 0)}} = 1.$$

Це максимальне значення позитивної кореляції у таблиці 11.7. Тобто чим вищий рівень вітаміну D, тим більша щільність кісткової тканини.

Для перевірки статистичної значущості результату сформулюємо гіпотезу.

$H_0: \tau = 0$ – зв'язок відсутній,

- $H_1: \tau \neq 0$ – зв'язок між спостереженнями є значущим.

Таблиця 11.7

N	Пари пацієнтів	Порівняння показників	Висновок
1	A-B	(20<30;1,00<1,20)	узгоджена пара
2	A-C	(20<25;1,00<1,10)	узгоджена пара
3	A-D	(20<35;1,00<1,30)	узгоджена пара
4	A-E	(20<22;1,00<1,05)	узгоджена пара
5	B-C	(30>25;1,20>1,10)	узгоджена пара
6	B-D	(30<35;1,20<1,30)	узгоджена пара
7	B-E	(30>22;1,20>1,05)	узгоджена пара
8	C-D	(25<35;1,10<1,30)	узгоджена пара
9	C-E	(25>22;1,10>1,05)	узгоджена пара
10	D-E	(35<22;1,30<1,05)	узгоджена пара

Обчислимо значення Z – статистики.

$$Z = \frac{\tau \sqrt{9n(n-1)}}{\sqrt{2(2n+5)}} = \frac{1 \cdot \sqrt{9 \cdot 5 \cdot 4}}{\sqrt{2 \cdot (2 \cdot 5 + 5)}} = \frac{\sqrt{180}}{\sqrt{30}} = \sqrt{6} \approx 2,45.$$

Критичне значення Z_{theory} для рівня значущості $\alpha = 0,05$ знаходимо, наприклад, з таблиці стандартного нормального розподілу: $Z_{theory} = 1,96$. Оскільки $|Z| = 2,45 > 1,96$, нульова гіпотеза відхиляється, тобто зв'язок між спостереженнями є значущим.

Остаточний висновок: кореляція між рівнем вітаміну D і щільністю кісткової тканини максимально висока, не випадкова і статистично значуща при рівні значущості $\alpha = 0,05$.

Розглянемо реалізацію обчислень за допомогою програми **Excel**. Перенесемо дані прикладу на аркуш програми **Excel** (рис.11.16).

	А	В
	Рівень вітаміну D (нг/мл)	Щільність кісткової тканини
1		
2	20	1
3	30	1,2
4	25	1,1
5	35	1,3
6	22	1,05
7		

Рисунок 11.16

Можливості версії **Excel 365** дозволяють обчислити *коефіцієнт Кендала* за допомогою однієї формули. Для цього використовується функція **LET**. Функція **LET** надає ім'я результату обчислення, що дозволяє використовувати це ім'я надалі.

Вигляд формули наведений у табл.11.8.

Таблиця 11.8

Формула	Розшифровка
=LET(	- функція
X; B2:B6;	- діапазон X
Y; C2:C6;	- діапазон Y
n; ROWS(X);	- кількість спостережень
C; SUM(--(X<TRANSPOSE(X)) * --(Y<TRANSPOSE(Y)));	- узгоджені пари
D; SUM(--(X<TRANSPOSE(X)) * --(Y>TRANSPOSE(Y)));	- неузгоджені пари
Tx; (SUM(--(X=TRANSPOSE(X))) - n)/2;	- число пар, пов'язаних по X
Ty; (SUM(--(Y=TRANSPOSE(Y))) - n)/2;	- число пар, пов'язаних по Y
(C - D) / SQRT((C + D + Tx)*(C + D + Ty)))	- коефіцієнт Кендала

Зауваження. У формулі «--» (два мінуси) використовуються для перекладу логічних значень (**TRUE**; **FALSE**) у числові (1;0).

Вставляємо вищенаведену формулу в комірку F1 і знаходимо **коефіцієнт Кендала** (рис.11.17).

F1		=LET(X; B2:B6;Y; C2:C6;n; ROWS(X);C; SUM((X<TRAN + Tx)*(C + D + Ty)))				
	A	B	C	D	E	F
1		Рівень вітаміну D (нг/мл)	Щільність кісткової тканини		$\tau =$	1
2		20	1			
3		30	1,2			
4		25	1,1			
5		35	1,3			
6		22	1,05			
7						

Рисунок 11.17

Застосовуючи формулу для Z-статистики, отримуємо її значення в клітинці F3.

F3

✕

✓

f_x

=F1*SQRT(9*ROWS(B2:B6)*(ROWS(B2:B6)-1))/SQRT(2*(2*ROWS(B2:B6)+5))

	A	B	C	D	E	F	G	H	I
1		Рівень вітаміну D (нг/мл)	Щільність кісткової тканини		$\tau =$	1			
2		20	1						
3		30	1,2		$Z =$	2,45			
4		25	1,1						
5		35	1,3						
6		22	1,05						

Рисунок 11.18

За допомогою функції **NORM.S.INV(0.975)**, яка є зворотною до стандартної нормальної функції розподілу, визначаємо критичне значення Z_{theory} для двостороннього тесту з рівнем значущості $\alpha = 0,05$. Таким чином, всі значення, знайдені за допомогою програми **Excel**, повністю збігаються з отриманими раніше.

F5		✕ ✓ f_x		=NORM.S.INV(0,975)		
	A	B	C	D	E	F
1		Рівень вітаміну D (нг/мл)	Щільність кісткової тканини		$\tau =$	1
2		20	1			
3		30	1,2		$Z =$	2,45
4		25	1,1			
5		35	1,3		$Z_{\text{theory}} =$	1,96
6		22	1,05			
7						

Рисунок 11.19

Практичне завдання

Завдання 1

- 1) визначити параметри a і b рівнення лінійної регресії $y = a + bx$.
- 2) знайти ***p-value***, оцінити статистичну значущість нахилу кутового коефіцієнта b) зробити висновок про клінічну (або практичну) значущість.
- 3.) обчислити ***коефіцієнт кореляції Пірсона*** і зробити висновок про спрямованість та характер зв'язку.
- 4) побудувати графік розсіювання з лінією тренду (регресії).

Варіант №1(X: Дозування (мг), Y: Зниження систоліч. артеріального тиску (мм рт. ст.))

X	58	71	48	34	62	27	80	40	58	77
Y	166,5	177	155,9	151,3	161,1	145,6	151,3	165,3	170,8	158,9

Варіант №2(X: Вік (роки), Y: Холестерин (мг/дл))

X	46	96	64	84	19	66	66	51	66	55
Y	14	26	16,5	25,7	1,8	19,5	19	12	19,5	15,3

Варіант №3 (X: ІМТ (кг/кв.м), Y: Глюкоза натщесерце (мг/дл))

X	21,8	26,6	22	34,6	27,4	22,6	30,5	18,7	36,5	27,9
---	------	------	----	------	------	------	------	------	------	------

Y	114,8	138,7	125,4	166,3	135	132	143,5	113,5	145,8	138,9
---	-------	-------	-------	-------	-----	-----	-------	-------	-------	-------

Варіант №4 (X: Тижні терапії, Y: Оцінка симптомів (0-100))

X	1	2	3	4	5	6	7	8	9	10
Y	52,9	48	37,9	38,1	43	38,1	34,5	28,3	27	21,3

Варіант №5 (X: Добова кількість кроків (тис.), Y: ЧСС спокою (уд/хв))

X	9	8	12	2	1	7	7	14	8	5
Y	61,4	65,8	59,1	68	72,2	68,6	63,9	60,3	54,9	68,8

Варіант №6 (X: Години сну, Y: Бал когнітивного тесту)

X	4,1	4,5	4,2	7,2	5,6	6,5	8,5	5,2	6,1	7,8
Y	53,9	56,9	64,6	69,5	59,8	68,6	71,7	67,7	60,2	67,2

Варіант №7 (X: Алкогольні одиниці/тиждень, Y: Фермент печінки АЛТ (од/л))

X	24	26	29	24	19	12	8	2	6	5
Y	38,1	40,2	45,6	39,1	36,5	21,8	22,9	27,4	24,9	17,5

Варіант №8 (X: Сигарети/день, Y: ОФВ1 (% від прогнозованого))

X	36	16	1	1	27	22	36	31	32	0
Y	47	74,5	98,3	92,5	56,5	68,3	43,6	57,3	55,9	101,6

Варіант №9 (X: Споживання кальцію (мг/день), Y: Щільність кісткової тканини (г/кв.см))

X	1406	1010	1251	643	1108	700	623	686	825	963
Y	1,401	1,269	1,345	1,115	1,284	1,173	1,068	1,11	1,208	1,209

Варіант №10 (X: Годин фізичних вправ/тиждень, Y: VO2 max (мл/кг/хв))

X	8,2	5,6	5,3	2,4	0,9	9	9	6,3	3,4	3,5
Y	59,48	47,74	44,31	32,83	26,07	61,22	66,38	51,21	34,57	38,73

Варіант №11 (X: Вік (роки), Y: Систолічний артеріальний тиск (мм рт. ст.))

X	52	23	52	33	40	67	39	75	27	26
Y	166,8	147,5	155,1	145,1	150,9	180,9	153,3	189,8	122,2	127,5

Варіант №12 (X: Час з моменту вакцинації (місяці), Y: Титр антитіл (AU))

X	0	14	0	4	15	18	3	2	16	16
Y	18,6	1,4	18,6	0	4,7	0,5	11,2	10,5	-1,8	-3

Варіант №13 (X: Цукор у їжі (г/день), Y: HbA1c (%))

X	86	147	37	44	242	73	77	86	123	193
Y	221,1	320,6	135,2	148,5	478,9	200,6	204,6	211,1	280,6	401,4

Варіант №14 (X: Рівень вітаміну D (нг/мл) Y: Оцінка імунної відповіді)

X	46,1	28	27,6	40,9	34,2	36,6	40,8	45,1	20,2	21,9
Y	81,5	57,14	50,38	77,2	65,24	70,61	70,83	78,82	58,18	53,07

Варіант №15 (X: Дотримання режиму прийому ліків (%), Y: Покращення симптомів (%))

X	93	93	54	88	53	55	94	81	79	96
Y	-289,6	-293	-	-	-	-145	-	-	-230	-
			142,2	260,3	125,2		294,5	240,1		291,4

Завдання 2

- 1) обчислити коефіцієнт кореляції Спірмена та зробити висновок про спрямованість та характер зв'язку.
- 2) знайти p-value та оцінити статистичну значимість залежності двох вибірок.

Варіант №1

Рівень С-реактивного білка (мг/л)	5	12	7	20	15	9	18	11	6	13
Шкала болю (0–10)	2	4	3	7	6	5	6	4	3	5

Варіант №2

Глюкоза крові натще (ммоль/л)	4,8	5,2	6,1	5,9	7,3	6,8	5,1	4,9	6,5	5,7
Індекс маси тіла	21	24	28	27	30	29	23	22	31	26

Варіант №3

Частота серцевих скорочень (уд/хв)	62	75	68	81	70	77	65	72	85	74
Рівень тривоги (шкала HADS)	6	10	8	12	9	11	7	8	13	10

Варіант №4

Ферритин (нг/мл)	15	25	18	40	33	29	22	37	20	27
Ступінь слабкості (0–10)	3	5	4	8	7	6	4	7	5	6

Варіант №5

Систолічний АТ (мм рт. ст.)	120	136	130	145	150	138	128	135	142	155
Рівень стресу (0–10)	3	6	5	7	8	6	4	6	7	9

Варіант №6

Холестерин загальний (ммоль/л)	4,3	5,1	6,2	5,8	4,9	5,5	6,8	4,7	5,9	6,1
Ступінь атеросклерот.бляшок (%)	12	18	32	28	15	21	35	14	27	30

Варіант №7

Температура тіла (°C)	36,7	37,5	38,2	37,9	38,5	37,2	36,9	38,1	37,8	38,4
Частота дихання (вдих/хв)	16	22	26	24	28	18	17	25	23	27

Варіант №8

Рівень вітаміну D (нг/мл)	12	18	24	20	15	27	30	16	22	25
Частота інфекцій за рік	6	5	4	5	6	3	2	6	4	3

Варіант №9

ALT (од/л)	22	35	28	40	45	38	30	42	33	47
Індекс стеатозу печінки	5	9	7	12	14	10	8	13	9	15

Варіант №10

GFR (мл/хв)	95	88	76	82	65	72	90	85	78	69
Стадія ниркової недостатності (1–5)	1	1	2	2	3	3	1	2	2	3

Варіант №11

Тригліцериди (ммоль/л)	0,9	1,2	1,8	1,5	2,1	2,4	1,7	1	2,2	2,3
Окружність талії (см)	72	80	94	88	102	105	90	75	99	103

Варіант №12

HbA1c (%)	5,4	6,1	7,2	6,8	8	7,5	6,4	5,7	7	8,3
Тривалість діабету (роки)	1	3	8	6	10	7	4	2	6	11

Варіант №13

Лейкоцити ($10^9/\text{л}$)	5,2	6,8	7,4	8,3	9,1	7,7	6,2	8	7,1	8,7
Температура тіла ($^{\circ}\text{C}$)	36,8	37,2	37,5	38,1	38,4	37,7	37,1	38	37,4	38,3

Варіант №14

Швидкість осідання еритроцитів (ШОЕ, мм/год)	4	10	15	22	18	25	20	12	16	28
Кількість симптомів (шт.)	1	3	4	7	5	8	6	3	4	9

Варіант №15

Рівень кортизолу (мкг/дл)	9	14	11	18	21	16	12	10	19	17
Рівень стресу (0–40)	8	15	10	22	28	18	14	9	25	20

Завдання 3

- 1) обчислити Коефіцієнт кореляції Кендалла та зробити висновок про спрямованість та характер зв'язку.
- 2) знайти Z-статистику та оцінити статистичну значущість залежності двох вибірок.

Варіант №1

Пацієнт	A	B	C	D	E
Рівень болю за шкалою 0–10	2	5	7	4	6
Обмеження рухливості, 0–3	1	2	3	2	3

Варіант №2

Пацієнт	A	B	C	D	E
---------	---	---	---	---	---

Артеріальний тиск (мм рт.ст.)	120	130	125	135	140
Рівень холестерину (ммоль/л)	4,5	5	4,8	5,2	5,5

Варіант №3

Пацієнт	A	B	C	D	E
Кількість днів з високою температурою	1	3	2	4	5
Інтенсивність головного болю (1–5)	2	4	3	5	5

Варіант №4

Пацієнт	A	B	C	D	E
Рівень цукру в крові (ммоль/л)	5,1	6,2	5,8	6,5	5,5
Індекс маси тіла (BMI)	22	25	24	26	23

Варіант №5

Пацієнт	A	B	C	D	E
Кількість годин сну на добу	6	7	5	8	7
Рівень втоми болю (1–10)	8	7	9	6	7

Варіант №6

Пацієнт	A	B	C	D	E
Артеріальний тиск	115	120	110	125	130
Частота серцевих скорочень	70	72	68	75	78

Варіант №7

Пацієнт	A	B	C	D	E
Рівень гемоглобіну (г/л)	130	125	135	128	140
Кількість лейкоцитів (10^9 /л)	6	5	7	6	8

Варіант №8

Пацієнт	A	B	C	D	E
Число кроків на день	4000	6000	5000	7000	4500
Рівень втоми	6	4	5	3	6

Варіант №9

Пацієнт	A	B	C	D	E
Вік пацієнта	25	40	35	50	30
Артеріальний тиск	120	130	125	135	122

Варіант №10

Пацієнт	A	B	C	D	E
Рівень стресу (1–10)	3	6	4	7	5
Тривалість сну	8	6	7	5	6

Варіант №11

Пацієнт	A	B	C	D	E
Кількість спожитих ліків на день	1	3	2	4	2
Рівень самопочуття (1–5)	5	3	4	2	4

Варіант №12

Пацієнт	A	B	C	D	E
Тиск систолічний	110	120	115	125	130
Рівень холестерину	4,5	5	4,8	5,2	5,5

Варіант №13

Пацієнт	A	B	C	D	E
Тривалість фізичних вправ (години на тиждень)	2	4	3	5	1
Індекс маси тіла (BMI)	25	23	24	22	26

Варіант №14

Пацієнт	A	B	C	D	E
Кількість сигарет на день	0	5	10	3	7
Рівень захворюваності дихальних шляхів (1–5)	1	2	4	2	3

Варіант №15

Пацієнт	A	B	C	D	E
Вік пацієнта	20	35	30	50	40
Кількість днів із застудою за рік	2	3	4	5	3

ПРАКТИЧНА РОБОТА 12

Дискримінантний аналіз

Питання, які розглядаються:

1. Алгоритм застосування.
2. Приклад розрахунку за допомогою програми Excel.

Дискримінантний аналіз є статистичним методом, що дозволяє класифікувати об'єкти (пацієнтів, біологічні зразки, клінічні випадки) за заздалегідь відомими групами на основі набору кількісних ознак. Внаслідок його застосування визначається лінійна комбінація ознак (дискримінантна функція), знання якої дозволяє віднести об'єкт до тієї чи іншої групи. Наприклад, можна визначити, до якої групи належить новий пацієнт, наприклад, здоровий або з патологією.

1. Алгоритм застосування

Нехай задані дві групи: A (n_1 – елементів), B (n_2 – елементів). І нехай у кожній групі її елементи мають по два показники: X и Y . Таким чином, задані чотири вибірки:

- X_A – вибірка значень показника X в групі A ;
- Y_A – вибірка значень показника Y в групі A ;
- X_B – вибірка значень показника X в групі B ;
- Y_B – вибірка значень показника Y в групі B .

Розглянемо алгоритм дій покроково.

КРОК 1. Обчислюємо середнє значення $\bar{X}_A, \bar{Y}_A, \bar{X}_B, \bar{Y}_B$ та будуємо вектор різниці

$$d = \begin{pmatrix} \bar{X}_A - \bar{X}_B \\ \bar{Y}_A - \bar{Y}_B \end{pmatrix}.$$

Початок цього вектора знаходиться в точці з координатами $(\bar{X}_B; \bar{Y}_B)$, а кінець в точці $(\bar{X}_A; \bar{Y}_A)$. Вектор d показує напрямок, уздовж якого групи різняться найсильніше (рис.12.1).

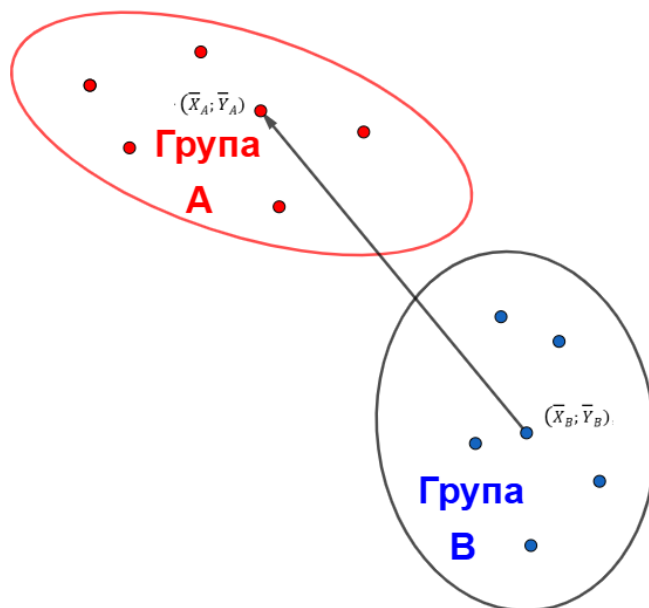


Рисунок 12.1

КРОК 2. Обчислюємо коваріаційні матриці для кожної групи окремо.

$$S_A = \begin{pmatrix} cov(X_A, X_A) & cov(X_A, Y_A) \\ cov(Y_A, X_A) & cov(Y_A, Y_A) \end{pmatrix}, \quad S_B = \begin{pmatrix} cov(X_B, X_B) & cov(X_B, Y_B) \\ cov(Y_B, X_B) & cov(Y_B, Y_B) \end{pmatrix}.$$

Для обчислення коваріації використовуємо формулу:

$$cov(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y}).$$

КРОК 3. Знаходимо об'єднану внутрішньогрупову коваріаційну матрицю. Вона розраховується як середньозважена коваріація за всіма групами:

$$S_{pooled} = \frac{(n_1 - 1)S_A + (n_2 - 1)S_B}{n_1 + n_2 - 2}.$$

КРОК 4. Знаходимо матрицю S_{pooled}^{-1} , яка є зворотною до матриці S_{pooled} .

КРОК 5. Коефіцієнти дискримінантної функції $D(x, y)$ знаходимо через добуток матриць S_{pooled}^{-1} и d :

$$b = S_{pooled}^{-1} \cdot d,$$

де $b = (b_x, b_y)$ і тоді дискримінантна функція

$$D(x, y) = b_x x + b_y y + a.$$

КРОК 6. Для знаходження постійної a використовуємо умову:

$$D(\bar{X}_A, \bar{Y}_A) = -D(\bar{X}_B, \bar{Y}_B).$$

Тоді

$$a = -\frac{1}{2} [b_x(\bar{X}_A + \bar{X}_B) + b_y(\bar{Y}_A + \bar{Y}_B)].$$

Таке значення вільного доданку **a** дозволяє визначити належність довільного елемента з показниками (x, y) до групи А або В за ознакою дискримінантної функції $D(x, y)$.

Отже дискримінантна функція має вигляд:

$$D(x, y) = b_x x + b_y y - \frac{1}{2} [b_x(\bar{X}_A + \bar{X}_B) + b_y(\bar{Y}_A + \bar{Y}_B)].$$

При $D(x, y) = 0$ дискримінантна функція геометрично визначає пряму лінію, яка на координатній площині поділяє групи А та В (рис.12.2).

Таким чином, якщо відомий розподіл елементів за групами, і за їх показниками побудовано дискримінантну функцію, то для класифікації довільного елемента достатньо підставити його показники в дискримінантну функцію та визначити її знак.

Наприклад, відомо, що для елементів групи A функція $D(x, y) > 0$, а для групи B функція $D(x, y) < 0$. Після підстановки показників нового елемента N визначили: $D(x_N, y_N) < 0$, отже елемент N належить до групи B .

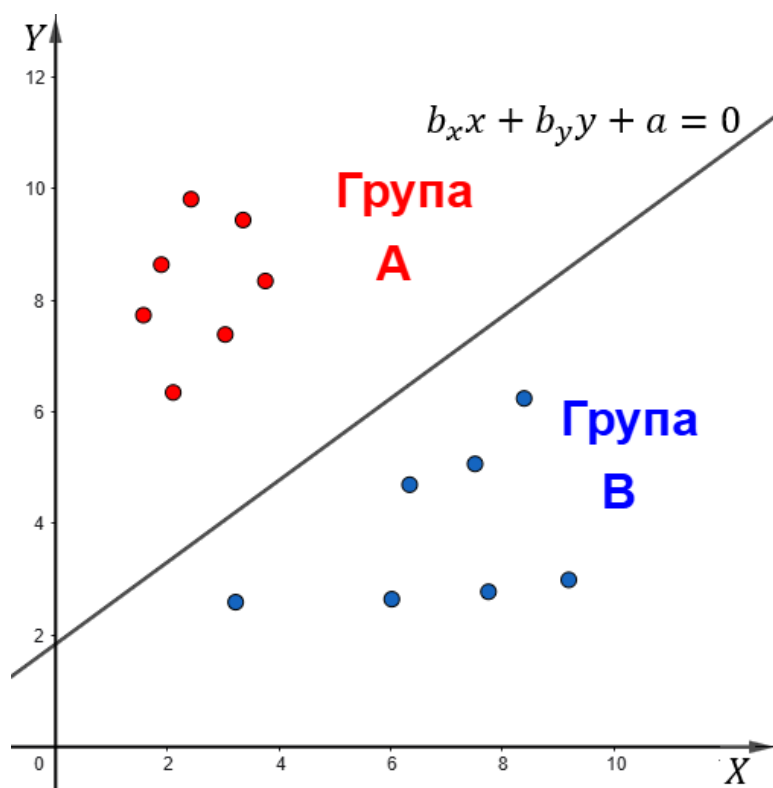


Рисунок 12.2

2. Приклад розрахунку за допомогою програми Excel

Розглянемо на конкретному прикладі, як за допомогою засобів програми *Excel* побудувати дискримінантну функцію. В групах A та B задані показники їх пацієнтів: альбумін та глобулін (табл. 12.1). Для цих двох груп потрібно побудувати дискримінантну функцію.

Таблиця 12.1

	Група А				Група В			
Альбумін (g/dL)	4,2	4,0	4,3	4,1	2,8	3,0	2,6	2,9
Глобулін (g/dL)	2,5	2,3	2,6	2,4	3,8	4,1	3,6	4,0

Розмістимо дані на аркуші програми *Excel* (рис.12.3).

	A	B	C
1			
2	Група	Альбумін	Глобулін
3	A	4,2	2,5
4		4	2,3
5		4,3	2,6
6		4,1	2,4
7			
8	B	2,8	3,8
9		3	4,1
10		2,6	3,6
11		2,9	4
12			

Рисунок 12.3

КРОК 1. В комірках **F6, F7, F11, F12** за допомогою функції **AVERAGE** обчислюємо середні значення $\bar{X}_A, \bar{Y}_A, \bar{X}_B, \bar{Y}_B$ (рис.12.4).

F	G
Середні	
Альбумін	Глобулін
4,15	2,45
2,825	3,875

Рисунок 12.4

Потім в комірках **I3, I4** знаходимо значення вектору різниці (рис.12.5) за формулою:

$$d = \begin{pmatrix} \bar{X}_A - \bar{X}_B \\ \bar{Y}_A - \bar{Y}_B \end{pmatrix}.$$

Н	І	Ј
	Різниця середніх	
	d	
	1,325	
	-1,425	

Рисунок 12.5

КРОК 2. Використовуючи функцію **COVARIANCE.S** обчислюємо коваріаційні матриці

$$S_A = \begin{pmatrix} cov(X_A, X_A) & cov(X_A, Y_A) \\ cov(Y_A, X_A) & cov(Y_A, Y_A) \end{pmatrix}, \quad S_B = \begin{pmatrix} cov(X_B, X_B) & cov(X_B, Y_B) \\ cov(Y_B, X_B) & cov(Y_B, Y_B) \end{pmatrix}.$$

Значення елементів матриці S_A розташовуємо в комірках **K3:L4**, а матриці S_B в комірках **K8:L9** (рис.12.6).

Ј	К	Л	М
	Матриці коваріації		
	S_A		
	0,017	0,017	
	0,017	0,017	
	S_B		
	0,029	0,038	
	0,038	0,049	

Рисунок 12.6

КРОК 3. В комірках **N3:O4** знаходимо значення елементів об'єднаної внутрішньогрупової коваріаційної матриці (рис.12.7) за формулою:

$$S_{pooled} = \frac{(n_1 - 1)S_A + (n_2 - 1)S_B}{n_1 + n_2 - 2}.$$

M	N	O	P
	Об'єднана матриця		
	S_{pooled}		
	0,023	0,027	
	0,027	0,033	

Рисунок 12.7

КРОК 4. Виділяємо комірку **Q3** та за допомогою функції **MINVERSE(N3:O4)** обчислюємо значення елементів матриці S_{pooled}^{-1} , яка є зворотною до матриці S_{pooled} (рис.12.8).

P	Q	R	S
	Зворотня матриця		
	S_{pooled}^{-1}		
	1580	-1300	
	-1300	1100	

Рисунок 12.8

КРОК 5. Коефіцієнти дискримінантної функції $D(x,y)$ є результатом добутку матриць S_{pooled}^{-1} та d . Для їх обчислення в комірці **T3** вводимо формулу: **=MMULT(Q3:R4;I3:I4)** (рис.12.9).

S	T	U
	Коефіцієнти b	
	3946	
	-3290	

Рисунок 12.9

КРОК 6. Щоб знайти остаточний вираз для дискримінантної функції $D(x, y)$ визначаємо значення постійної a

$$a = -\frac{1}{2} [b_x(\bar{X}_A + \bar{X}_B) + b_y(\bar{Y}_A + \bar{Y}_B)].$$

Для цього в комірку **V3** записуємо формулу (рис.12.10):

$$=-0,5*(T3*(F6+F11)+T4*(G6+G11)).$$

U	V	W
	a	
	-3357,05	

Рисунок 12.10

Отже, всі невідомі параметри знайдені, і дискримінантна функція має вигляд:

$$D(x, y) = b_x x + b_y y - a = 3946x - 3290y - 3357,05.$$

Підставимо значення показників пацієнтів груп **A** та **B** в формулу для дискримінантної функції. Отримуємо (рис.12.11):

- $D(x, y)$ в групі A – 4991,15; 4859,95; 5056,75; 4925,55 > 0;
- $D(x, y)$ в групі B - -4810,25; -5008,05; -4941,45; -5073,65 < 0.

Слід звернути увагу на те, що для пацієнтів групи **A** дискримінантна функція набуває позитивних значень, а для групи **B** – негативні.

Нехай група **A** – група пацієнтів, що страждають на ішемічну хворобу серця, а група **B** – група здорових пацієнтів. Нехай отримано аналізи нового пацієнта: альбумін – **4,4** и глобулін – **2,1**. Обчислимо значення дискримінантної функції:

$$D(x, y) = 3946 \cdot 4,4x - 3290 \cdot 2,1 - 3357,05 = 7096,35 > 0.$$

Послідовність застосування методу дискримінантного аналізу за допомогою програми **Excel** відображена у Додатку Ж.

	A	B	C	D
1				
2	Група	Альбумін	Глобулін	$D(x,y)$
3	A	4,2	2,5	4991,15
4		4	2,3	4859,95
5		4,3	2,6	5056,75
6		4,1	2,4	4925,55
7				
8	B	2,8	3,8	-4810,25
9		3	4,1	-5008,05
10		2,6	3,6	-4941,45
11		2,9	4	-5073,65
12				

Рисунок 12.11

Робимо висновок: оскільки $D(x, y) > 0$, нового пацієнта можна віднести до групи **A**, пацієнтів, що страждають на ішемічну хворобу серця (рис.12.12).

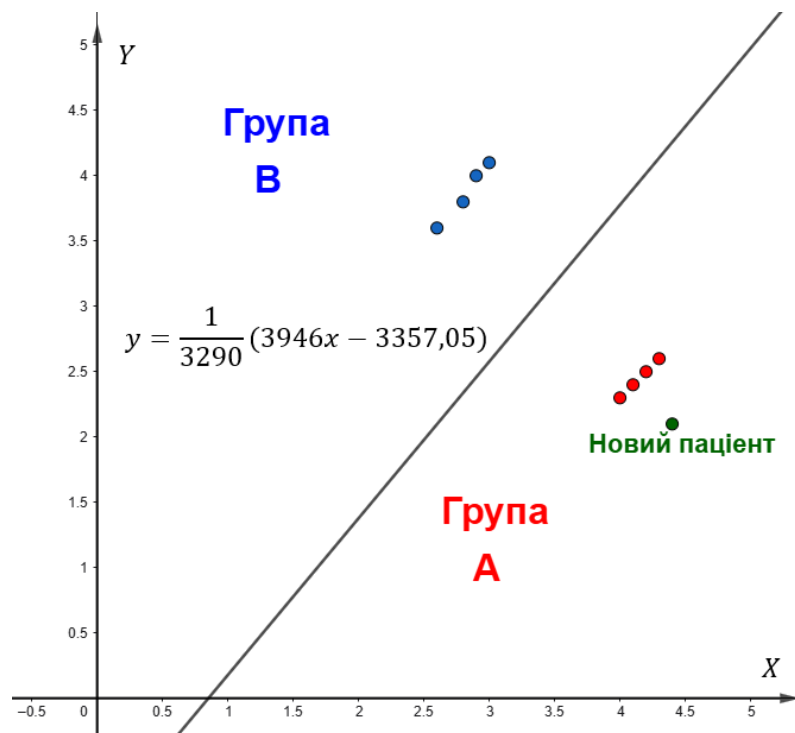


Рисунок 12.12

Зауваження: у випадку, коли груп більше, ніж дві, знаходять дискримінантну функцію окремо для кожної групи. Потім елемент відносять до тієї групи, у якої

дискримінантна функція під час встановлення показників елемента приймає максимальне значення.

Практичне завдання

- 1) за підсумками показників елементів двох груп побудувати дискримінантну функцію.
- 2) за показниками нового елемента С визначити, до якої групи він належить.

Варіант №1

	Група А				Група В				С
Систолічний тиск (мм рт.ст.)	118	122	115	124	142	150	136	148	130
Рівень холестерину (mg/dL)	180	192	175	187	230	245	218	237	200

Варіант №2

	Група А				Група В				С
Частота серцевих скорочень (уд/хв)	64	70	68	66	92	88	96	90	75
ІМТ	22,1	21,4	23	22,6	29,5	31	28,2	30,1	28

Варіант №3

	Група А				Група В				С
Глюкоза в крові натще (ммоль/л)	4,8	5,1	4,6	5	8,9	9,6	10,2	9,1	8
Інсулін (μIU/mL)	8	10	9	7	22	28	25	20	15

Варіант №4

	Група А				Група В				С
Гемоглобін (g/dL)	13,8	14,2	13,5	14	10,2	9,8	10,6	9,5	12
Кількість тромбоцитів ($\times 10^3/\mu\text{L}$)	250	270	240	260	110	95	125	105	180

Варіант №5

	Група А				Група В				С
Частота дихання (дих/хв)	14	16	15	13	22	20	24	21	19
Насичення киснем (SpO₂, %)	98	97	99	98	88	85	82	86	105

Варіант №6

	Група А				Група В				С
Кальцій (ммоль/л)	2,25	2,3	2,2	2,28	1,8	1,75	1,82	1,78	2,1
Магній (ммоль/л)	0,85	0,9	0,82	0,88	0,6	0,55	0,62	0,58	0,7

Варіант №7

	Група А				Група В				С
Натрій (ммоль/л)	139	137	141	138	125	128	123	127	132
Калій (ммоль/л)	4,1	3,9	4,3	4	5,6	5,2	5,8	5,4	6,2

Варіант №8

	Група А				Група В				С
Креатинін (мг/дл)	0,85	0,78	0,9	0,82	2,1	2,4	1,9	2,2	1,5
Сечовина (ммоль/л)	5,2	4,8	5,5	5	12	14,5	11,3	13,2	6,1

Варіант №9

	Група А				Група В				С
АЛТ (U/L)	22	18	25	20	135	160	120	148	90
АСТ (U/L)	19	21	17	20	110	95	130	125	80

Варіант №10

	Група А				Група В				С
ЛПНЩ (LDL, mg/dL)	95	88	102	90	160	175	148	168	140
ЛПВЩ (HDL, mg/dL)	58	62	55	60	38	34	40	36	70

Варіант №11

	Група А				Група В				С
Лейкоцити ($\times 10^3/\mu\text{L}$)	6,2	5,8	6,5	6	14,5	12,8	16,2	13,9	8,3
Еритроцити ($\times 10^6/\mu\text{L}$)	4,6	4,8	4,5	4,7	3,8	3,5	3,9	3,6	7,2

Варіант №12

	Група А				Група В				С
Вітамін D (ng/mL)	32	28	35	30	8	12	6	10	22
Кальцій (ммоль/л)	2,28	2,22	2,3	2,25	1,85	1,78	1,9	1,82	2,5

Варіант №13

	Група А				Група В				С
Тригліцериди (mg/dL)	110	95	125	105	320	285	350	300	246
Глюкоза (ммоль/л)	4,9	5,3	5	5,1	9,8	10,6	8,9	9,4	6,3

Варіант №14

	Група А				Група В				С
ТТГ (mIU/L)	1,8	2,1	1,6	2	8,5	7,2	9	8	5,7
Т3 (ng/dL)	1,1	1,3	1,2	1	0,6	0,5	0,7	0,55	0,9

Варіант №15

	Група А				Група В				С
С-реактивний білок (СРБ, mg/L)	2,1	1,5	2,8	1,9	45	38	52	41	28
ШОЕ (мм/год)	8	10	7	9	68	75	60	70	35

РЕКОМЕНДОВАНА ЛІТЕРАТУРА

Основна

1. Антомонов М.Ю. Математичне оброблення та аналіз медико-біологічних даних. 3-тє вид. – К.: 2026. – 450с.
2. Біостатистика : підруч. з навч. дисципліни "Соціальна медицина, громадське здоров'я", згідно з вимогами навч. плану підгот. фахівців другого (магістерського) рівня вищ. освіти кваліфікації освітньої "Магістр медицини" кваліфікації проф. "Лікар" галузі знань 22 "Охорона здоров'я" спец. 222 "Медицина" / Т.С. Грузєва [та ін.] ; за ред.: Т.С. Грузєвої ; рец.: А. Г. Шульгай, В.М. Михальчук, О.В. Любінець. - Вінниця : Нова книга, 2021. - 384 с.

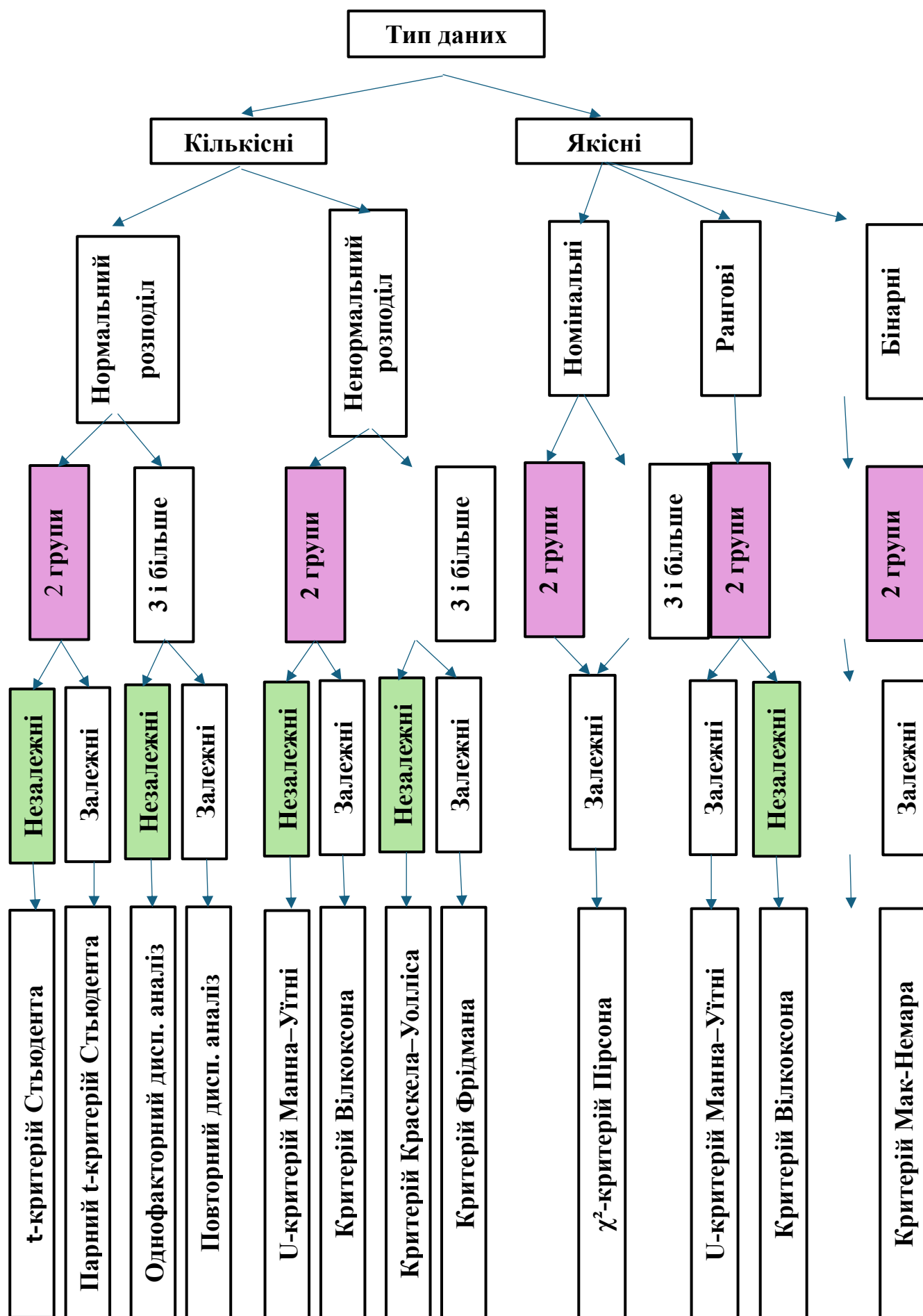
Додаткова

1. Соціальна медицина, громадське здоров'я. Модуль 1 "Біостатистика" : метод. рек. з проведення практичних занять для студентів III курсу підгот. фахівців другого (магістерського) рівня вищ. освіти кваліфікації освітньої "Магістр медицини" кваліфікації проф. "Лікар" галузі знань 22 "Охорона здоров'я", спец. 222 "Медицина" / Т. С. Грузєва [та ін.] ; рец.: В. А. Огнєв, В. М. Лехан ; МОЗ України, Нац. мед. ун-т ім. О. О. Богомольця, Кафедра соціальної медицини та громадського здоров'я. - Київ : Книга плюс, 2019. - 92 с.
2. Посібник з біостатистики. Аналіз результатів медичних досліджень у пакеті EZR (R–statistics) / В. Г. Гур'янов, Ю. Є. Лях, В. Д. Парій, О. В. Короткий, О. В. Чалий, К. О. Чалий, Я. В. Цехмістер : Навчальний посібник. – К. : Вістка, 2018. – 208 с.
3. Мармоза А. Т. Практикум з математичної статистики: підручник / А.Т.Мармоза – 2ге вид. перераб. Та доп. – К.: «Центр учбової літератури», 2013. – 592 с.
4. Володарський Є.Т., Кошева Л.О. Статистична обробка даних. Навч. посіб. – К.: НАУ.– Книжкове видавництво, 2008.– 308 с.
5. Мінцер О.П., Вороненко Ю.П., Власов В.В. Оброблення клінічних і експериментальних даних у медицині.– К.: Вища школа, 2003.– 350 с.

Інформаційні ресурси

1. Сайт дисципліни «Основи біостатистики та методи статистичного моделювання» <https://zsmu.sharepoint.com/sites/BioStat>
2. Національна бібліотека України ім. Вернадського.
<http://www.nbuv.gov.ua/>
3. Наукова бібліотека ЗДМФУ: <https://lib.mphu.edu.ua/>
4. Національна Медична Бібліотека США (National Library of Medicine) – вільний та безкоштовний доступ до MEDLINE <http://www.ncbi.nlm.nih.gov/PubMed>
5. Кафедра медичної інформатики та комп'ютерних технологій навчання Національного медичного університету імені О. О. Богомольця
<http://nmuofficial.com/zagalni-vidomosti/kafedri/departament-medical-biological-physics/>
6. Віртуальна хімічна лабораторія <http://chemcollective.org/vlab>

Схема вибору статистичного критерію



Коефіцієнти критерію Шапіро–Уїлка

<i>k</i>	<i>n</i>									
								8	9	10
1	—	—	—	—	—	—	—	0,6052	0,5888	0,5739
2	—	—	—	—	—	—	—	0,3164	0,3244	0,3291
3	—	—	—	—	—	—	—	0,1743	0,1976	0,2141
4	—	—	—	—	—	—	—	0,0561	0,0947	0,1224
5	—	—	—	—	—	—	—	—	—	0,0399
	11	12	13	14	15	16	17	18	19	20
1	0,5601	0,5475	0,5359	0,5251	0,5150	0,5056	0,4968	0,4886	0,4808	0,4734
2	0,3315	0,3325	0,3325	0,3318	0,3306	0,3290	0,3273	0,3253	0,3232	0,3211
3	0,2260	0,2347	0,2412	0,2460	0,2495	0,2521	0,2540	0,2553	0,2561	0,2565
4	0,1429	0,1586	0,1707	0,1802	0,1878	0,1939	0,1988	0,2027	0,2059	0,2085
5	0,0695	0,0922	0,1099	0,1240	0,1353	0,1447	0,1524	0,1587	0,1641	0,1686
6	—	0,0303	0,5359	0,0727	0,0980	0,1005	0,1009	0,1197	0,1271	0,1334
7	—	—	—	0,0240	0,0433	0,0593	0,0725	0,0737	0,0932	0,1013
8	—	—	—	—	—	0,0196	0,0359	0,0496	0,0612	0,0711
9	—	—	—	—	—	—	—	0,0163	0,0303	0,0422
10	—	—	—	—	—	—	—	—	—	0,0140
	21	22	23	24	25	26	27	28	29	30
1	0,4643	0,4590	0,4542	0,4493	0,4450	0,4407	0,4366	0,4328	0,4291	0,4254
2	0,3185	0,3156	0,3126	0,3098	0,3069	0,3043	0,3018	0,2992	0,2968	0,2944
3	0,2578	0,2571	0,2563	0,2554	0,2543	0,2533	0,2522	0,2510	0,2499	0,2487
4	0,2119	0,2131	0,2139	0,2145	0,2148	0,2151	0,2152	0,2151	0,2150	0,2148
5	0,1736	0,1764	0,1787	0,1807	0,1822	0,1836	0,1848	0,1857	0,1864	0,1870
6	0,1399	0,1443	0,1480	0,1512	0,1539	0,1563	0,1584	0,1601	0,1616	0,1630
7	0,1092	0,1150	0,1201	0,1245	0,1283	0,1316	0,1346	0,1372	0,1395	0,1415
8	0,0804	0,0878	0,0941	0,0997	0,1046	0,1089	0,1128	0,1162	0,1192	0,1219
9	0,0530	0,0618	0,0696	0,0764	0,0823	0,0876	0,0923	0,0965	0,1002	0,1036
10	0,0263	0,0368	0,0459	0,0539	0,0610	0,0672	0,0728	0,0778	0,0822	0,0862
11	—	0,0122	0,0228	0,0321	0,0403	0,0476	0,0540	0,0598	0,0598	0,0697
12	—	—	—	0,0107	0,0200	0,0284	0,0358	0,0424	0,0483	0,0537
13	—	—	—	—	—	0,0094	0,0178	0,0253	0,0320	0,0381
14	—	—	—	—	—	—	—	0,0084	0,0159	0,0227
15	—	—	—	—	—	—	—	—	—	0,0076
	31	32	33	34	35	36	37	38	39	40
1	0,4220	0,4188	0,4156	0,4127	0,4098	0,4068	0,4040	0,4015	0,3989	0,3964
2	0,2921	0,2898	0,2786	0,2854	0,2834	0,2813	0,2794	0,2774	0,2755	0,2737
3	0,2475	0,2463	0,2451	0,2439	0,2427	0,2415	0,2403	0,2391	0,2380	0,2368
4	0,2155	0,2141	0,2137	0,2132	0,2127	0,2121	0,2116	0,2110	0,2104	0,2098
5	0,1874	0,1878	0,1880	0,1882	0,1883	0,1883	0,1883	0,1881	0,1880	0,1878
6	0,1641	0,1651	0,1660	0,1667	0,1673	0,1678	0,1683	0,1686	0,1689	0,1691

7	0,1433	0,1449	0,1463	0,1475	0,1487	0,1496	0,1505	0,1513	0,1520	0,1526
8	0,1243	0,1265	0,1284	0,1301	0,1317	0,1331	0,1344	0,1356	0,1366	0,1376
9	0,1066	0,1009	0,1118	0,1140	0,1160	0,1179	0,1196	0,1211	0,1225	0,1237
10	0,0899	0,0931	0,0961	0,0988	0,1013	0,1036	0,1056	0,1075	0,1092	0,1108
11	0,0739	0,0739	0,0812	0,0844	0,0873	0,0900	0,0924	0,0947	0,0967	0,0986
12	0,0585	0,0585	0,0669	0,0706	0,0739	0,0770	0,0798	0,0824	0,0848	0,0870
13	0,0435	0,0435	0,0530	0,0572	0,0610	0,0645	0,0677	0,0706	0,0733	0,0759
14	0,0289	0,0289	0,0395	0,0441	0,0484	0,0523	0,0559	0,0592	0,0622	0,0651
15	0,0144	0,0144	0,0262	0,0314	0,0361	0,0404	0,0444	0,0481	0,0515	0,0546
16	—	—	0,0131	0,0187	0,0239	0,0287	0,0331	0,0372	0,0409	0,0444
17	—	—	—	0,0062	0,0119	0,0172	0,0220	0,0264	0,0305	0,0343
18	—	—	—	—	—	0,0057	0,0110	0,0158	0,0203	0,0244
19	—	—	—	—	—	—	—	0,0053	0,0101	0,0146
20	—	—	—	—	—	—	—	—	—	0,0049
	41	42	43	44	45	46	47	48	49	50
1	0,3940	0,3917	0,3894	0,3872	0,3850	0,3830	0,3808	0,3789	0,3770	0,3651
2	0,2719	0,2701	0,2684	0,2667	0,2651	0,2635	0,2620	0,2604	0,2589	0,2574
3	0,2357	0,2345	0,2334	0,2323	0,2313	0,2302	0,2291	0,2281	0,2271	0,2260
4	0,2091	0,2085	0,2078	0,2072	0,2065	0,2058	0,2052	0,2045	0,2038	0,2032
5	0,1876	0,1874	0,1871	0,1868	0,1868	0,1862	0,1869	0,1855	0,1851	0,1847
6	0,1693	0,1694	0,1695	0,1695	0,1695	0,1695	0,1695	0,1693	0,1692	0,1691
7	0,1531	0,1535	0,1539	0,1542	0,1545	0,1548	0,1550	0,1551	0,1553	0,1554
8	0,1384	0,1392	0,1398	0,1405	0,1410	0,1415	0,1420	0,1423	0,1427	0,1430
9	0,1249	0,1259	0,1269	0,1278	0,1286	0,1293	0,1300	0,1306	0,1312	0,1317
10	0,1123	0,1136	0,1149	0,1160	0,1170	0,1180	0,1189	0,1197	0,1205	0,1212
11	0,1004	0,1020	0,1035	0,1049	0,1062	0,1073	0,1085	0,1095	0,1105	0,1113
12	0,0891	0,0909	0,0927	0,0943	0,0959	0,0972	0,0986	0,0998	0,1010	0,1020
13	0,0782	0,0804	0,0824	0,0842	0,0860	0,0876	0,0892	0,0906	0,0919	0,0932
14	0,0677	0,0701	0,0724	0,0745	0,0765	0,0783	0,0801	0,0817	0,0832	0,0846
15	0,0575	0,0602	0,0628	0,0651	0,0673	0,0694	0,0713	0,0731	0,0748	0,0764
16	0,0476	0,0506	0,0534	0,0560	0,0584	0,0607	0,0628	0,0648	0,0667	0,0685
17	0,0379	0,0411	0,0442	0,0471	0,0497	0,0522	0,0546	0,0568	0,0588	0,0608
18	0,0283	0,0318	0,0352	0,0383	0,0412	0,0439	0,0465	0,0489	0,0511	0,0532
19	0,0188	0,0227	0,0263	0,0296	0,0328	0,0357	0,0385	0,0411	0,0436	0,0459
20	0,0094	0,0136	0,0175	0,0211	0,0245	0,0277	0,0307	0,0335	0,0361	0,0386
21	—	0,0045	0,0087	0,0126	0,0163	0,0197	0,0229	0,0229	0,0288	0,0314
22	—	—	—	0,0042	0,0081	0,0118	0,0153	0,0185	0,0215	0,0244
23	—	—	—	—	—	0,0039	0,0076	0,0111	0,0143	0,0174
24	—	—	—	—	—	—	—	0,0037	0,0071	0,0104
25	—	—	—	—	—	—	—	—	—	0,0035

Критичні значення критерію Шапіро–Уїлка

<i>n</i>	<i>p</i>		<i>n</i>	<i>p</i>	
	0,01	0,05		0,01	0,05
			26	0,891	0,920
			27	0,894	0,923
			28	0,896	0,924
			29	0,898	0,926
			30	0,900	0,927
			31	0,902	0,929
			32	0,904	0,930
8	0,749	0,818	33	0,906	0,931
9	0,764	0,829	34	0,908	0,933
10	0,781	0,842	35	0,910	0,934
			36	0,912	0,935
11	0,792	0,850	37	0,914	0,936
12	0,805	0,859	38	0,916	0,938
13	0,814	0,866	39	0,917	0,939
14	0,825	0,874	40	0,919	0,940
15	0,835	0,881			
			41	0,920	0,941
16	0,844	0,887	42	0,922	0,942
17	0,851	0,892	43	0,923	0,943
18	0,858	0,897	44	0,924	0,944
19	0,863	0,901	45	0,926	0,945
20	0,868	0,905			
			46	0,927	0,945
21	0,873	0,908	47	0,928	0,946
22	0,878	0,911	48	0,929	0,947
23	0,881	0,914	49	0,929	0,947
24	0,884	0,916	50	0,930	0,947
25	0,888	0,918			

Критичні значення Мана-Уїтні.

Критичні значення наводяться для двустороннього тесту при $\alpha = 0,05$

N_2	N_1													
	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4	3	4	4	5	6	7	8	9	10	11	11	12	13	13
5	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
9	12	15	17	20	23	26	28	31	34	37	39	42	45	48
10	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14	22	26	31	36	40	45	50	55	59	64	67	74	78	83
15	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17	28	34	39	45	51	57	63	67	75	81	87	93	99	105
18	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20	34	41	48	55	62	69	76	83	90	98	105	112	119	127

Критичні значення Уїлкоксона.

Критичні значення наводяться для тесту при $\alpha = 0,01$, $\alpha = 0,05$

<i>n</i>	α	
	0,05	0,01
5	0	-
6	2	-
7	3	0
8	5	1
9	8	3
10	10	5
11	13	7
12	17	9
13	21	12
14	25	15
15	30	19
16	35	23
17	41	27
17	47	32
19	53	37
20	60	43
21	67	49
22	75	55
23	83	62
24	92	69
25	100	76
26	110	84
27	119	92

Таблиця розподілу Фішера-Снедекора

(k_1 — число ступенів свободи більшої дисперсії, k_2 — число ступенів свободи меншої дисперсії при рівні значущості $\alpha = 0,05$)

k_2	k_1											
	1	2	3	4	5	6	7	8	9	10	11	12
1	161	200	216	225	230	234	237	239	241	242	243	244
2	18,51	19,00	19,16	19,25	19,30	19,33	19,36	19,37	19,38	19,39	19,40	19,41
3	10,13	9,55	9,28	9,12	9,01	8,94	8,88	8,84	8,81	8,78	8,76	8,74
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,93	5,91
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,78	4,74	4,70	4,68
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,63	3,60	3,57
8	5,32	4,45	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,34	3,31	3,28
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,13	3,10	3,07
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,97	2,94	2,91
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,86	2,82	2,79
12	4,75	3,88	3,49	3,26	3,11	3,00	2,92	2,85	2,80	2,76	2,72	2,69
13	4,67	3,80	3,41	3,18	3,02	2,92	2,84	2,77	2,72	2,67	2,63	2,60
14	4,60	3,74	3,34	3,11	2,96	2,85	2,77	2,70	2,65	2,60	2,56	2,53
15	4,54	3,68	3,29	3,06	2,90	2,79	2,70	2,64	2,59	2,55	2,51	2,48
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,45	2,42
17	4,45	3,59	3,20	2,96	2,81	2,70	2,62	2,55	2,50	2,45	2,41	3,38

Критичні значення χ^2 ($\alpha = 0,05$)

Кількість груп k	Ступені свободи $df = k - 1$	Критичне значення $\chi^2_{0,05}$
2	1	3,84
3	2	5,99
4	3	7,81
5	4	9,49
6	5	11,07
7	6	12,59
8	7	14,07
9	8	15,51
10	9	16,92

Приклад застосування дискримінантного аналізу в Excel

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1																						
2	Група	Альбумін	Глобулін			Середні			Різниця середніх													
3		4,2	2,5			Альбумін	Глобулін		1,325		0,017	0,017		0,023	0,027		1580	-1300		3946		
4		4	2,3						-1,425		0,017	0,017		0,027	0,033		-1300	1100		-3290		
5		4,3	2,6																			
6		4,1	2,4			4,15	2,45															
7																						
8		2,8	3,8																			
9		3	4,1																			
10		2,6	3,6			2,825	3,875															
11		2,9	4																			
12																						
13																						