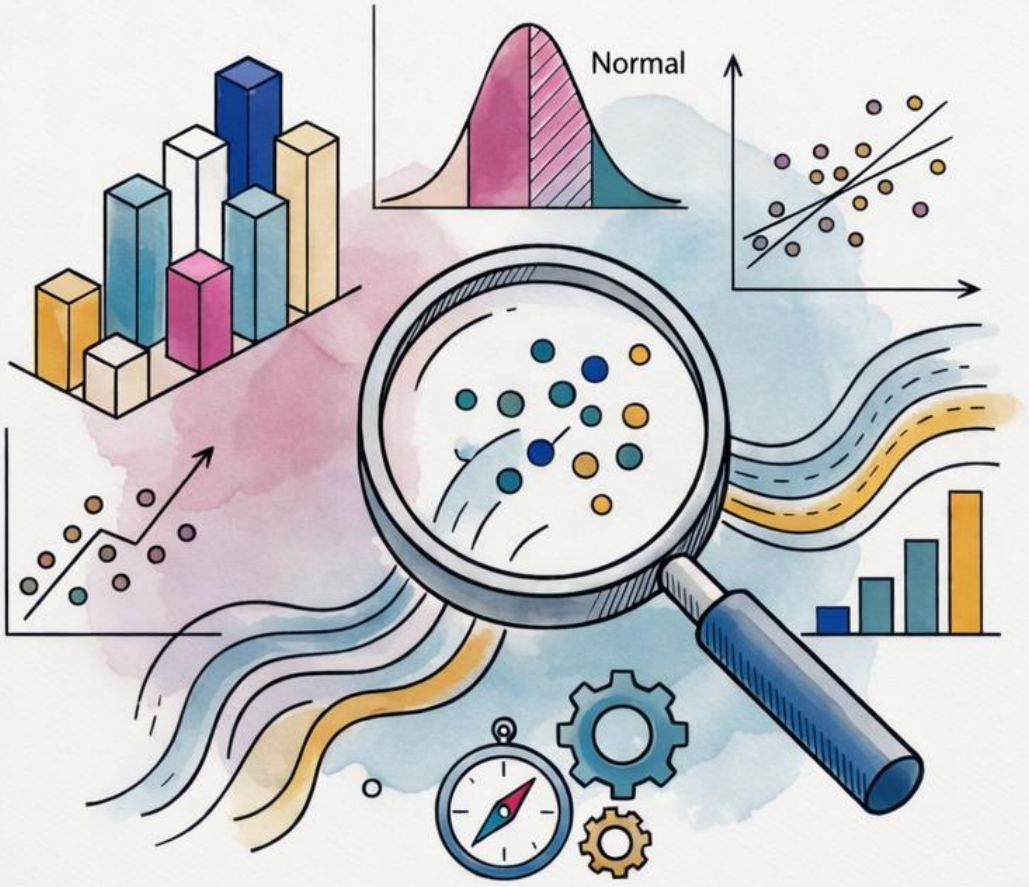
Є. С. СОВА

СТАТИСТИЧНИЙ АНАЛІЗ ДАНИХ

Від концепцій до практики



НАВЧАЛЬНИЙ ПОСІБНИК

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
«КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»

Є.С. СОВА

СТАТИСТИЧНИЙ АНАЛІЗ ДАНИХ

Від концепцій до практики

Навчальний посібник

Київ
2026

УДК 519.22:311:330.43(075.8)
С 56

Затверджено до друку рішенням Вченої ради факультету економічних наук Національного університету «Києво-Могилянська академія» (протокол № 7 від 11 червня 2026 року)

Рецензенти:

Дубина М. В., доктор економічних наук, професор,
завідувач кафедри фінансів, банківської справи та страхування
Національного університету «Чернігівська політехніка»

Макогон В. Д., доктор економічних наук, професор,
професор кафедри фінансів
Державного торговельно-економічного університету

Сова Є. С.

С 56 Статистичний аналіз даних: Від концепцій до практики : навч. посіб.
Київ : НаУКМА, 2026. 104 с.
ISBN 978-617-8640-14-9

Навчальний посібник присвячено фундаментальним концепціям статистики та практичним підходам до застосування інструментів кількісного аналізу для вирішення фінансово-економічних та інших дослідницьких завдань та формування обґрунтованих аналітичних висновків.

У посібнику розглянуто сутність і базові принципи статистичного аналізу; класифікацію статистичних показників і даних; методи графічного аналізу; міри центральної тенденції, варіації, форми та ступеня однорідності розподілу; характеристику нормального та інших поширених статистичних розподілів. Значну увагу приділено прикладам і практичним розрахункам, зокрема із використанням мови програмування Python.

Видання адресовано студентам закладів вищої освіти, аспірантам, викладачам, фахівцям-практикам та всім, хто цікавиться статистичними методами дослідження.

УДК 519.22:311:330.43(075.8)

ISBN 978-617-8640-14-9

ЗМІСТ

ПЕРЕДМОВА

У сучасному світі «великих даних» здатність їх розуміти і перетворювати на конкретну цінність стає стратегічною перевагою, а не лише допоміжною навичкою – як в академічній, так і у професійній сфері. Цей навчальний посібник покликаний відкрити світ статистики не лише для студентів, а й для всіх, хто прагне опанувати прикладні навички первинного аналізу та обробки даних.

Фундаментом більшості досліджень та аналітичних завдань є методи описової статистики. Саме вони дозволяють аналітикам та менеджерам у фінансово-економічній сфері дослідити великі масиви інформації: зрозуміти їхню структуру, виявити приховані тенденції та презентувати складні закономірності у зрозумілій формі. Опанування цих методів – це не просто вивчення формул, а формування базису для подальшого економіко-математичного моделювання та прийняття зважених управлінських рішень.

Особливості викладу матеріалу:

- Традиційні академічні концепції інтерпретовано через призму **авторського досвіду викладання** університетського курсу статистики та проілюстровано зрозумілими і практичними прикладами.
- Одна з ідей цього посібника – **полегшити сприйняття** абстрактних теорій та математичних методів за допомогою **ефективних візуалізацій**, що дозволяють навіть інтуїтивно зрозуміти сутність ключових концепцій.
- Посібник структуровано за логікою **поступового переходу** від фундаментальних статистичних понять до їхньої практичної реалізації. Особливу увагу приділено прикладам і розрахункам, у тому числі з використанням мови програмування **Python**, яка є особливо актуальною для аналізу великих масивів даних.

Опанування викладених у посібнику концепцій і методів покликане сформувати у читачів статистичний фундамент і сприяти розвитку аналітичного мислення, які є особливо необхідними для професійної діяльності фінансових аналітиків, а також усіх, хто цікавиться питаннями статистичного аналізу даних, в умовах цифровізації.

РОЗДІЛ 1. СУТНІСТЬ І МЕТОДИ СТАТИСТИЧНОГО АНАЛІЗУ

1.1. Сутність та цілі вивчення статистики. Базові принципи статистичного аналізу

У сучасному світі значення статистики важко переоцінити. З одного боку, статистика є фундаментальною дисципліною, яка лежить в основі різних фінансово-економічних та соціальних дисциплін, машинного навчання й науки про дані (Data Science. З іншого, статичний аналіз має безпосереднє практичне значення у професійному житті фінансистів, економістів, соціологів, маркетологів, журналістів, психологів, політологів, істориків та багатьох інших дослідників, аналітиків та керівників різних структурних підрозділів. Окрім цього, кожен з нас використовує статистичні методи під час прийняття рішень у повсякденному житті, інколи навіть не усвідомлюючи цього, наприклад:

- Який потрібно одержати середній бал за результатами національного тестування, щоб вступити до омріяного університету за рахунок державного бюджету? Наскільки імовірно вступити на обрану спеціальність у цьому році, якщо набрав N балів?
- Як розподілені доходи населення залежно від професії, країни, освіти тощо?
- Яка структура моїх щомісячних витрат, і як вона змінюється залежно від сезону?

- Як корелюють між собою кількість пройдених кроків за день та кількість спалених калорій?
- Як змінюються ціни на одяг залежно від дати виходу нової колекції, на смартфони – залежно від дати виходу нових моделей, абонементи у спортзал – залежно від періоду року?

Для відповіді на ці та багато інших запитань ми зазвичай збираємо історичні дані й аналізуємо їхній розподіл, досліджуємо взаємозв'язок між різними показниками, проводимо опитування, будуємо гіпотези й експериментально їх перевіряємо, моделюємо гіпотетичні ситуації та оцінюємо імовірність перебігу різних сценаріїв. Хоча багато людей робить це інтуїтивно або навіть несвідомо, усі ми так чи інакше використовуємо методи, що безпосередньо пов'язані зі статистичним аналізом даних. Навіть народна мудрість нерідко ґрунтується на багаторічних спостереженнях і узагальненнях, що дозволили виявити стійкі закономірності.

Отже, **предметом** статистики є розміри і кількісні співвідношення між масовими суспільними явищами, закономірності їх формування, розвитку та взаємозв'язку.

Метою цього посібника є допомогти студентам, науковцям, практикам та всім, хто цікавиться статистикою, оволодіти теоретико-методичними та практичними навичками *статистичного аналізу* соціально-економічних процесів для виявлення *об'єктивних закономірностей* і обґрунтування рішень у повсякденному й професійному житті та науковій діяльності.

Серед основних **практичних цілей** оволодіння навичками статистичного аналізу для студентів можна визначити такі:

- Навчитися застосовувати статистичні методи на практиці для прийняття зважених та обґрунтованих рішень, а також виявлення маніпуляцій і помилок у повсякденному житті.

- Опанувати навички ефективної обробки та статистичного аналізу даних, необхідні для написання курсових і кваліфікаційних робіт, а також при підготовці наукових публікацій.
- Сформувати базові навички з програмування й алгоритмічного мислення, які знадобляться при вивченні інших курсів, пов'язаних зі статистикою, фінансовим і економіко-математичним моделюванням.
- Опанувати навички самостійного та впевненого використання методів кількісного аналізу під час складання вступних іспитів на міжнародні освітні програми або проходження тестування при працевлаштуванні.
- Одержати конкурентні переваги при подальшому працевлаштуванні завдяки сформованим практичним навичкам з обробки, узагальнення, аналізу й візуалізації даних, формулюванню ключових висновків, а також презентації результатів дослідження, з використанням ефективних інструментів, зокрема MS Excel та Python.

Перед тим, як перейти до вивчення власне статистичних концепцій і формул, необхідно засвоїти базові визначення, які допоможуть краще зрозуміти сутність статистики.

Закономірність – це повторюваність, послідовність, або порядок у масових процесах. У разі *великої кількості подій (даних)* вплив випадкових причин взаємно врівноважується, завдяки чому певний статичний закон стає видимим і зрозумілим. Особливістю статистичного аналізу є те, що він опирається на кількісні дані й намагається їх *узагальнити*. Якість та надійність висновків прямо залежить від обсягу та якості даних, доступних для аналізу.

Статистичні дані – системні кількісні характеристики явищ чи процесів, які об'єднуються у сукупності, найчастіше у вигляді таблиць. Наприклад, це може бути таблиця, де зібрано дані щодо ВВП, рівня інфляції, темпів економічного зростання, обсягу золотовалютних резервів для країн ЄС.

Статистична сукупність – це певна множина елементів (статистичних спостережень), поєднаних за спільними характеристиками. Прикладом

статистичної сукупності є стовпчик у таблиці (див. рис.1.1), що може містити фінансові показники, які характеризують результати діяльності комерційних банків, компаній ІТ галузі, споживачів послуг певної компанії тощо.

Статистичне спостереження – це одиниця даних, що описує певний процес. Прикладом статичного спостереження є рядок у таблиці, який записаний відповідно до правил її заповнення (один із варіантів подано у таблиці 1.1).

Ознака – характеристика або властивості, притаманні одиницям статистичної сукупності, які фіксуються під час статистичного спостереження. Ознаки можуть бути кількісними (вік, довжина, обсяг доходу у грошових одиницях) та якісними (стать, країна, назва компанії). Ознаки з певними значеннями утворюють статистичну сукупність.

Варіація ознаки – коливання (зміна) значень обраної ознаки.

Варіанта – це окреме індивідуальне значення ознаки в статистичній сукупності. Тобто кожне конкретне значення ознаки x_j у сукупності є варіантою.

Таблиця 1.1. Приклад таблиці зі статистичними даними

1 № 3	Країна	ВВП на душу населення, дол. США	2 Темп зростання ВВП, %	Інфляція, індекс споживчих цін
1	Канада	56 730	1,1	1,258
2	Данія	72 383	2,4	1,171
3	Японія	46 811	1,3	1,140
4	Норвегія	90 318	0,2	1,121
5	Польща	45 633	3,3	1,445
6	Швейцарія	85 621	1,1	1,070
7	США	76 976	2,0	1,316

Примітки: 1 – статистичні дані; 2 – статистична сукупність; 3 – статистичне спостереження.

Статистичний ряд – це *впорядкований* набір елементів обраної сукупності за певною ознакою (характеристикою). У цьому навчальному посібнику найчастіше розглядаються ряди розподілу (структури) та динаміки.

Ряд розподілу – статистичний ряд, що відображає частотність появи різних елементів сукупності, згрупованих за якісними (групи/класи) або кількісними (інтервали значень) характеристиками. Ряд розподілу дозволяє описати структуру сукупності, яка досліджується, і складається з двох основних елементів – групувальної ознаки (групи або інтервали значень у розподілі) та частоти (кількість елементів у кожній групі). На рис.1.1 наведено один з можливих прикладів такого ряду.

Рис. 1.1. Приклад ряду розподілу спортсменів за зростом

Ряд динаміки – статистичний ряд, що відображає зміни сукупності за певною ознакою у часі. Приклад такого ряду наведено на рис.1.2.

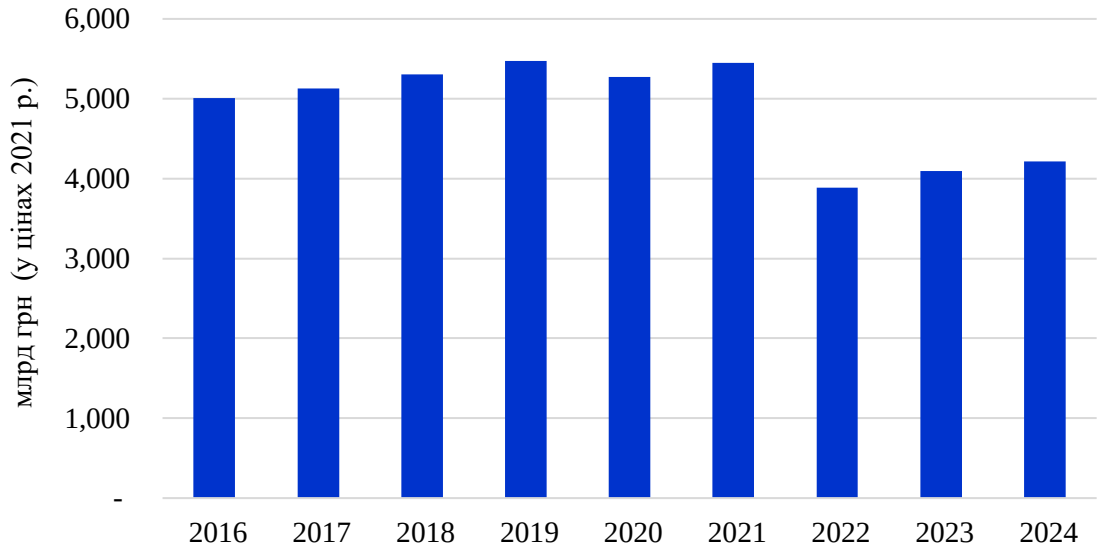


Рис. 1.2. Приклад ряду динаміки – ВВП країни у 2016-2024 рр.

Хоча поділ статистики на розділи є досить умовним через їхній тісний взаємозв'язок, за функціональними завданнями можна виділити кілька ключових напрямів статистичного аналізу, зокрема:

- *Описовий* (descriptive): описати та узагальнити усі наявні дані.
- *Індуктивний, або інференційний* (inferential): зробити висновки щодо усієї сукупності даних на основі дослідження певної вибірки даних.
- *Діагностичний* (diagnostic): дослідити взаємозв'язок (кореляцію) між різними сукупностями даних та зробити припущення щодо причин і наслідків такого зв'язку.
- *Прогностичний* (predictive): проєкція/прогнозування майбутніх значень показників на основі їхніх історичних значень.
- *Прескриптивний* (prescriptive): визначення найкращих стратегій дій на основі сценарного аналізу, оптимізаційних алгоритмів та інших аналітичних інструментів.

У цьому посібнику основну увагу приділено першому напрямку (див. табл. 1.2), тоді як інші висвітлюються лише оглядово та становлять предмет поглибленого вивчення в межах інших тем та економіко-математичних дисциплін.

Таблиця 1.2. Обрані напрями статистичного аналізу

	Описова статистика	**Індуктивна (інференційна) статистика**	**Діагностична статистика**
Основне завдання	▪ Узагальнення та візуалізація характеристик наявних даних (статистичної сукупності)	▪ Формування висновків щодо усієї (генеральної) сукупності на основі вибірки даних	▪ Виявлення взаємозв'язку (кореляції) між показниками; формування припущень щодо причинно-наслідкового зв'язку
Методи	▪ Розрахунок показників, які характеризують середнє значення, відхилення, симетрію тощо ▪ Візуалізація даних: таблиці та графіки	▪ Перевірка статистичних гіпотез; розрахунок довірчих інтервалів; визначення похибки вибірки	▪ Розрахунок коефіцієнтів кореляції, побудова регресійної моделі, виявлення аномальних значень
Приклади	▪ Розподіл учнів за академічною успішністю; динаміка середніх температур та їхнього коливання протягом останніх 30 років.	▪ Опитування 1 000 респондентів для прогнозу результатів виборів; перевірка гіпотези, що використання нового добрива суттєво збільшує середню урожайність пшениці порівняно зі стандартним добривом.	▪ Аналіз залежності середнього балу студентів від кількості інвестованих годин на підготовку; дослідження, як зміна ціни на продукт впливає на обсяг його продажів

На практиці ці напрями статистики інтегруються у взаємопов'язані етапи єдиного аналітичного процесу, що лежить в основі вирішення дослідницьких проблем або слугує підґрунтям для прийняття управлінських рішень і який схематично відображено на рис.1.3.

Рис. 1.3. Приклад етапів процесу статистичного аналізу

Нижче наведемо приклади, як статистичний аналіз дозволяє об'єктивно перевірити певні твердження й факти, а також виявити потенційно неправдиву інформацію. Навіть у засобах масової інформації та аналітичних звітах кількісні дані або текст можуть містити неточності через недостатню статистичну підготовку авторів, або навмисне маніпулювання фактами.

Приклад 1 «Ефект Моцарта»

У 1993 році було опубліковано дослідження про так званий «ефект Моцарта». Автори стверджували, що 36 студентів-респондентів продемонстрували покращення просторового мислення протягом наступних 15 хвилин після того, як прослухали обрані сонати Моцарта. Засоби масової інформації тиражували результати дослідження під заголовками у стилі «Моцарт робить вас розумнішими», поширивши це твердження також і на немовлят. Це призвело до зростання ажіотажу серед батьків, які прагнули придбати компакт-диски з відповідною музикою для своїх дітей. Губернатор Джорджії виділив окремий бюджет для фінансування закупівель компакт-дисків із класичною музикою для кожного немовляти. Уряд Флориди прийняв закон щодо обов'язкового прослуховування класичної музики протягом щонайменше однієї години на день у державних дитячих садках. *Подальші ґрунтовні дослідження, здійснені іншими науковцями, продемонстрували відсутність доказів статистично значущого покращення просторового мислення після прослуховування класичної музики.*

Потенційні причини появу цього міфу в результаті первинного дослідження:

- В оригінальному дослідженні брало участь лише 36 студентів, що робить цю вибірку нерепрезентативною через вкрай обмежену кількість спостережень і некоректну екстраполяцію результатів, отриманих від дорослих, на когнітивні навички дітей.
- При тестуванні когнітивних навичок використовувалися тести на просторове мислення, що не вимірюють повноцінно загальний рівень інтелекту (IQ).
- Міф посилила активність засобів масової інформації та політичні рішення керівництва деяких штатів, які могли переслідувати власні цілі.
- *Які інші причини ви помітили?*

Приклад 2 «Помилка середнього – середня тривалість життя у Середньовіччі»

У побуті часто можна почути, що середня тривалість життя в Середньовіччі становила лише 30–40 років. Проте цей показник не відображає реального віку дорослої людини того періоду. Наприклад, якщо усереднити життя немовляти (0 років) та літньої людини (70 років), отримаємо «середнє» у 35 років, але це не означає, що люди масово доживали лише до такого віку. Насправді найпоширеніший вік дожиття (статистична мода) у Середньовіччі, за деякими даними, сягав 60–70 років. Низький середній показник був зумовлений насамперед надзвичайно високою дитячою смертністю через інфекції та відсутність належної гігієни.

Узагальнимо типові причини появи подібних міфів або непідтверджених концепцій:

- *Невелика кількість спостережень.* Зазвичай для формування надійних висновків необхідно зібрати щонайменше 100 спостережень.
- *Нерепрезентативна вибірка*, яка сформована не випадковим чином, а містить представників лише деяких груп, що спотворює результати опитування.
- *Упередженість*, яка проявляється у схильності висвітлювати лише ті факти, які підтверджують власні переконання, водночас ігноруючи або недооцінюючи інформацію, яка суперечить поглядам дослідника.
- *Неточність*. Використання непідтверджених особистих історій або розповідей інших осіб як доказу певних концепцій чи міфів, попри відсутність достатньої кількості спостережень; при цьому окремі деталі можуть бути спотворені, пропущені або передані з помилками.

Питання для самоперевірки

1. *Що є предметом вивчення статистики як науки?*
2. *Надайте визначення поняттю «статистична сукупність» та наведіть приклад.*
3. *Чим відрізняється кількісна ознака від якісної? Наведіть приклади обох типів.*
4. *Що таке статистична закономірність і чому для її виявлення потрібна велика кількість даних?*
5. *Поясніть різницю між рядом розподілу та рядом динаміки.*
6. *Які існують основні напрями статистичного аналізу (від описового до прескриптивного)?*

1.2. Види статистичних показників. Класифікація даних за типами й методи їх групування

Статистичні дані – це фундамент аналітики в будь-якій сфері діяльності. У певному сенсі вони є «паливом», що дає змогу виявляти приховані закономірності, відстежувати тренди та формулювати обґрунтовані висновки (*англ.* “insights”). Якими б просунутими та складними не були економіко-математичні моделі чи окремі формули («інструменти» статистичного аналізу), вони не матимуть практичної цінності, якщо дані будуть неналежної якості або в недостатній кількості

У цій темі розглянемо базові поняття та поширені класифікації даних за типом і видами структур, а також наведемо відповідні приклади.

Базовими одиницями статистичних даних є змінна (*англ. “variable“)* та константа (*англ. “constant”*).

Змінна – це певний статистичний показник, що може набувати ***різних*** значень залежно від об'єкта та часового періоду. Наприклад, дохід компанії, зріст людини, темп інфляції.

Константа – показник, який набуває лише ***одного*** значення. Наприклад, у деяких контекстах константою може бути інфляційна ціль НБУ, комісія банку за транзакцію тощо. Відомими математичними константами є число Пі ($\pi \approx 3{,}14159$), число Ейлера ($e \approx 2{,}71828$).

Поняття даних є досить широким, адже вони можуть набувати різних форм, як показано у табл.1.3: числові, текстові, графічні, звукові, зображення, відео тощо. Сучасні технології дають змогу перетворювати всі ці види даних у числовий формат, що відкриває можливість їх обробки та аналізу за допомогою поширених статистичних інструментів.

У межах курсу «Статистика» ми зосередимося на методах аналізу даних, представлених у табличному вигляді — у формі числових або текстових змінних.

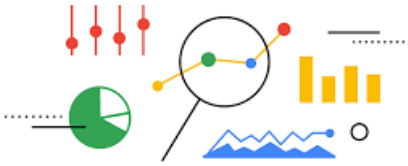
Для практичного засвоєння матеріалу посібника радимо використовувати MS Excel (галузевий стандарт у сфері фінансів протягом багатьох років) та мову програмування Python. Теоретичні концепції буде проілюстровано конкретними прикладами в зазначених середовищах.

Вивчення основ мови програмування Python має на меті посилити аналітичні й практичні навички читача під час опанування методів обробки й статистичного аналізу даних. Опанування базових навичок роботи з Python надає читачу низку переваг:

- підготовка до вивчення інших курсів, пов'язаних з економіко-математичним моделюванням;
- розширення можливостей аналізу даних для досліджень і написання курсових і кваліфікаційних робіт;
- розширення кругозору, розвиток навичок аналітичного й алгоритмічного мислення, що може стати перевагою при працевлаштуванні на посаду фінансового аналітику, або споріднені ролі;

- Python є доступною (безкоштовною) та відносно нескладною мовою п рограмування, для вивчення якої існує безліч ресурсів, онлайн-курсів і технічної документації;
- Python дозволяє вирішувати широкий спектр задач, значно спрощує й автоматизує технічну обробку даних та моделювання.

Таблиця 1.3. Форми представлення даних

Форма даних	Ілюстрація
Числові	A B C D E F G 99 30 44 22 17 19 40 42 51 27 16 78 25 95 21 16 14 11 35 45 96 86 8 11 19 64 51 36 100 77 70 77 42 98 80 50 97 13 96 70 56 84 12 44 34 20 54 65 1
Текстові	" 2 B R b r Φ (8 H X h x Σ # 3 C S c s Γ) 9 I Y i y Θ $ 4 D T d t Λ * : J Z j z Ξ % 5 E U e u Ω + ; K [k { & 6 F V f v Π , < L \ l | ' 7 G W g w Ψ - = M] m }
Графічні	
Звукові	
Зображення	
Відео	

За однією з класифікацій розрізняють такі типи даних:

1. ***Числові дані***

1.1. *Дискретні* – мають лише окремі цілочислові значення (наприклад, кількість укладених на біржі угод, кількість дітей у сім'ї тощо).

1.2. *Неперервні* - мають незліченну кількість значень в межах певного інтервалу (наприклад, вік людини в межах від 0 до 100, або температура повітря від – 15 до + 25 градусів за Цельсієм).

2. Категорійні (якісні) дані

2.1. Номінальні – набір категорій, що не потребують логічного впорядкування (наприклад, код виду економічної діяльності (КВЕД), текстові коди країн, коди валют).

2.2. Порядкові – категорії, що підлягають впорядкуванню (наприклад, індекси кредитних агенцій S&P, Fitch: AAA, AA+, AA, AA-, …, BBB+, BBB і т.д.).

Класифікацію даних за видами в узагальненому вигляді наведено на рис.1.4.

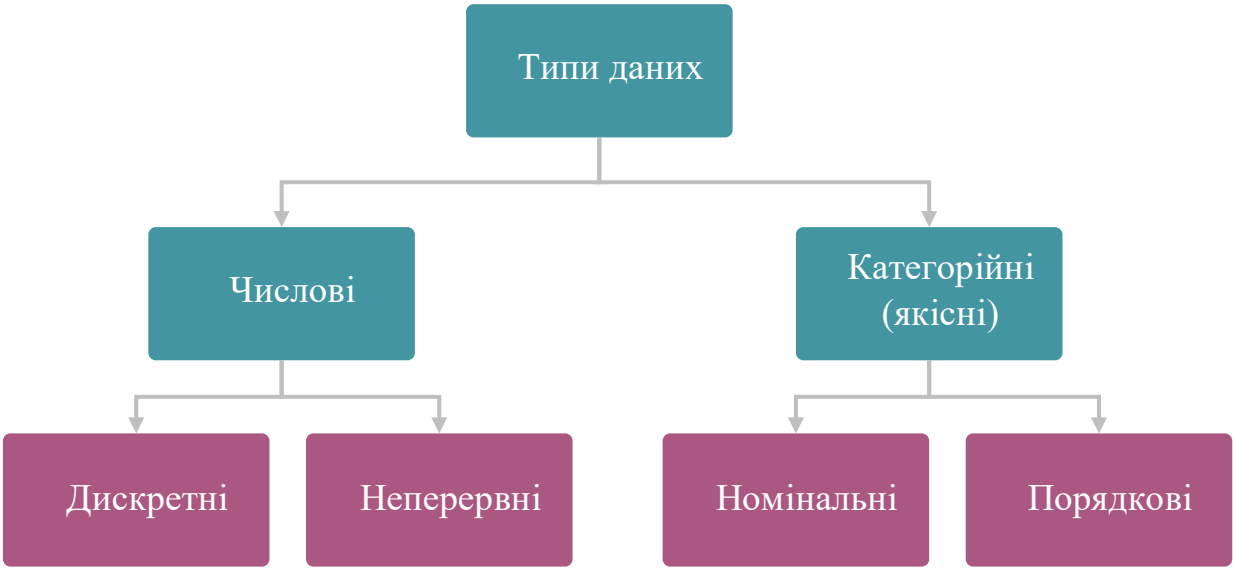


Рис. 1.4. Класифікація даних за видами

Серед числових даних також розрізняють абсолютні й відносні величини.

Абсолютні величини – це кількісні показники, які відображають свої безпосередні одиниці вимірювання: штуки, тонни, кіловати, гривні тощо.

Відносні величини є результатом порівняння, який характеризує міру кількісного співвідношення споріднених або різних показників (зазвичай вимірюються у відсотках, проміле, коефіцієнтах або індексах), наприклад:

- Показники *динаміки*: на скільки відсотків у цьому році виросли доходи, порівняно з попереднім?
- Показники *структури*: яку частку ринку займає обрана фірма?
- Показники виконання *нормативу*: який відсоток бюджету фактично виконано, порівняно із прогнозним значенням?

Деякі приклади абсолютних і відносних величин наведено у табл. 1.4.

Таблиця 1.4. Приклади абсолютних і відносних величин

	1	2	1	1	3	4
Вид продукції	**План на місяць, тис. грн**	**Планова структура, %**	**Міс. 1, тис. грн**	**Міс. 2, тис. грн**	**Темп приросту, %**	**Виконання плану у Міс. 2, %**
Матча (чай)	55 000	*59,1*	51 000	61 000	*19,6*	*110,9*
Енергетичні батончики	23 000	*24,7*	21 000	20 000	*(4,8)*	*87,0*
Протеїновий коктейль	15 000	*16,1*	18 000	14 500	*(19,4)*	*96,7*
Всього	**93 000**	***100,0***	**90 000**	**95 500**	***6,1***	***102,7***

Примітки: 1 – абсолютні змінні; 2 – відносний показник структури *(наприклад, для Енергетичних батончиків: 24,7% = 23 000 / 93 000 * 100)*; 3 – відносний показник динаміки *(наприклад, для Матча: 19,6% = (61 000 / 51 000 – 1)* 100*; 4 – відносний показник виконання нормативу/плану *(наприклад, для Протеїнового коктейлю: 96,7% = 14 500 / 15 000 * 100)*.

Залежно від *ступеня обробки* й *формату представлення* розрізняють:

- **Структуровані дані** – інформація, яку можна легко представити у вигляді таблиць, готових до подальшого традиційного аналізу. Приклади: дані щодо цін акцій на фондовому ринку, фінансова звітність, аналітичні таблиці з прогнозом грошових потоків.
- **Неструктуровані дані** - дані, що зберігаються у нестандартному, часто необробленому вигляді. Для їхнього аналізу використовують специфічні статистичні й математичні методи, зокрема алгоритми на основі машинного навчання та штучного інтелекту Приклади: фінансові новини у вигляді тексту, відгуки споживачів у соціальних мережах, супутникові знімки сільськогосподарських угідь.

У процесі обробки та трансформації статистичні сукупності групуються в набори даних, що формують такі базові структури (рис. 1.5):

- **Крос-секційні дані** – набір спостережень певної змінної для *множини об'єктів* (компаній, ринків, індивідів, регіонів) *у деякий момент часу* (наприклад, за останній рік). Наприклад, значення інфляції за січень для країн ЄС.
- **Часові ряди –** це набір спостережень деякої змінної для *певного об'єкта* для *однакових часових інтервалів* або моментів часу (щоденно, щомісячно, щорічно тощо). Наприклад, щоденні ціни на акції компанії ABC за січень.
- **Панельні дані –** комбінація *часових рядів* і *крос-секційних* даних. Дані представлено *у трьох вимірах*: змінні, об'єкти та часові інтервали/моменти.

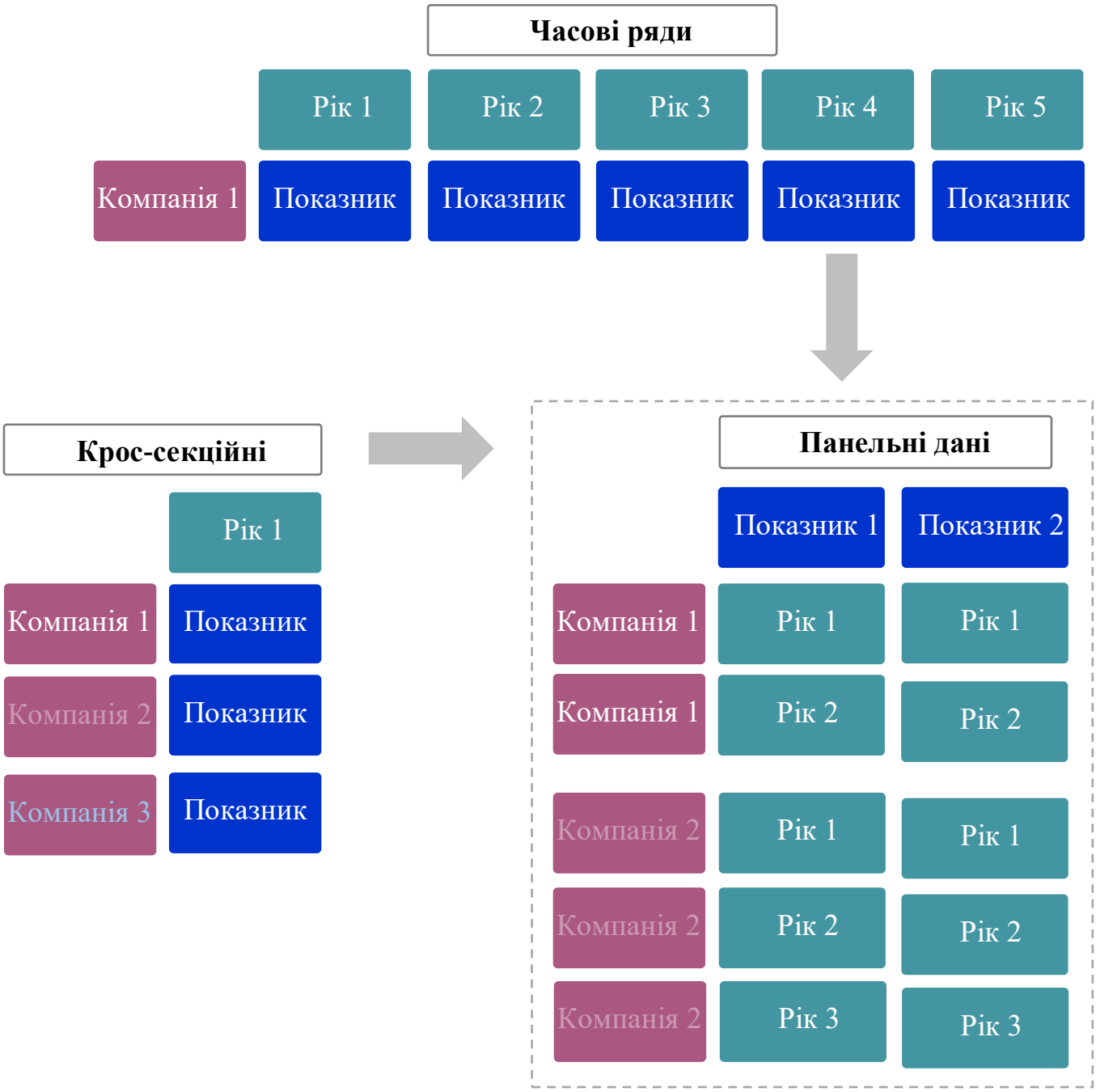


Рис. 1.5. Види структур даних

На практиці набори даних обробляють, трансформують та аналізують за допомогою спеціалізованих статистичних і фінансових інструментів. Найбільш поширеними в корпоративному середовищі є програма MS Excel (зокрема її потужні аналітичні надбудови, такі як Power Query та Power Pivot) та мова програмування Python.

Важливо розуміти, що процес збору, очищення (обробки) та трансформації даних зазвичай займає левову частку часу – близько 80% від загального обсягу

проєкту. На безпосередній аналіз і презентацію результатів залишається лише 20% ресурсів. Саме тому оволодіння сучасними методиками автоматизації та використання спеціалізованого програмного забезпечення є критично важливим: це дозволяє скоротити рутинні витрати часу на підготовку даних і зосередитися на формуванні якісних висновків.

Для первинної обробки статистичної інформації дані необхідно трансформувати до *незгрупованого (дезагрегованого)* вигляду. Саме такий формат, іноді званий *«пласкими даними» (flat data)*, дозволяє найбільш ефективно й автоматизовано очищувати інформацію, готуючи її до подальшого моделювання.

Основні правила трансформації даних до *незгрупованого* вигляду (див. табл.1.5):

- Кожна *змінна* представлена в окремому стовпці.
- Кожне *спостереження* представлено в окремому рядку.
- Таблиця представляє собою набір векторів, кожен з яких містить однакову кількість елементів (спостережень).

Таблиця 1.5. Приклад вибірки даних у *незгрупованому* вигляді

ID транзакції	Час	1 Сума, грн	Тип операції
TR-001	10:05	450	Поповнення карти
TR-002	10:12	12 400	Переказ
TR-003	10:15	800	Оплата послуг
TR-004	10:22	2 100	Поповнення карти
2 TR-005	10:30	15 000	Переказ
TR-006	10:35	350	Оплата послуг
TR-007	10:42	9 800	Кредитний платіж
TR-008	10:50	1 200	Поповнення карти
TR-009	10:55	25 000	Депозит
…	…	…	…

Примітки: 1 – змінна; 2 – спостереження

Для цілей подальшого аналізу до набору даних застосовують зворотну операцію – *групування* й представлення у *зведеному* вигляді (див. приклад, наведений у табл. 1.6). Саме цей вигляд зручний для побудови гістограм та ухвалення управлінських рішень. Принципи статистичного зведення:

- *Систематизація:* елементи сукупності (спостереження) за певними ознаками об'єднують у групи, класи, типи, а інформацію про них агрегують як у межах груп, так і в цілому по сукупності.
- *Типізація:* основне завдання зведення – виявити типові риси та закономірності масових явищ чи процесів.
- *Аналітика:* на основі зведених даних обчислюють узагальнювальні показники, здійснюють порівняльний аналіз, а також аналіз причин групових відмінностей та вивчають взаємозв'язки між ознаками.

Таблиця 1.6. Приклад вибірки даних у *зведеному* вигляді

Діапазон суми, грн	**Кількість транзакцій (абсолютна частота)**	**Питома вага (відносна частота)**	**Питома вага (кумулятивна* відносна частота)**
0 – 1 000	3	33.3%	33.3%
1 001 – 5 000	2	22.2%	55.56%
5 001 – 15 000	3	33.3%	88.89%
понад 15 000	1	11.1%	100.00%
Всього	**9**	**100.0%**	**-**

Примітки: * накопичувальним підсумком

Питання для самоперевірки

1. *Поясніть різницю між дискретними та неперервними числовими даними.*
2. *Наведіть приклади абсолютних та відносних статистичних величин.*
3. *Охарактеризуйте три базові структури наборів даних: крос-секційні, часові ряди та панельні дані.*
4. *Які основні правила трансформації даних до незгрупованого (дезагрегованого) вигляду?*

1.3. Ефективна візуалізація (графічне представлення) даних

Окрім розрахунків та побудови економіко-математичних моделей, важливим елементом статистичного аналізу є графічне представлення даних (візуалізація). Це є особливо актуальним при роботі з великими обсягами інформації, оскільки дозволяє уникнути когнітивного перевантаження при спробі осягнути необроблені масиви чисел і тексту.

Як окремий аналітичний інструмент, візуалізація перетворює громіздкі масиви інформації, оптимізовані для опрацювання комп'ютером, у зрозумілі й стислі структури, які дозволяють виявити ключові закономірності та полегшують комунікацію між дослідником і читачем (споживачем, керівником, аудиторією тощо).

Отже, **візуалізація** – метод графічного представлення даних у вигляді різноманітних діаграм, мап і графіків, що забезпечує її зрозумілу й об'єктивну інтерпретацію завдяки таким чинникам:

- дозволяє виявити взаємозв'язки й закономірності, що приховані за великим масивом даних;
- структурує складну для сприйняття інформацію для полегшення розуміння зовнішніми користувачами (читачами);
- має на меті представити дані у візуально привабливому вигляді, але без втрати об'єктивності, зберігаючи фактичні пропорції між показниками, щоб уникнути маніпуляцій.

Ефективність графічного аналізу залежить великою мірою від правильного вибору формату візуалізації, що визначається цілями дослідження (пошук взаємозв'язків, виявлення трендів, вивчення структури), архітектурою даних (обсяг і рівень деталізації, кількість змінних і спостережень) і природою показників (якісні або кількісні дані, неперервні або дискретні величини).

Нижче розглянемо найбільш поширені види графіків й діаграм, їх інтерпертацію та сферу застосування:

I. Візуалізації для аналізу розподілу даних

a. Гістограма (рис. 1.6) – зображує *розподіл даних* за частотністю (*f*) кожної категорії або інтервалу (*x*) у вигляді стовпчикової діаграми. На осі X представлено категорії (для дискретних даних) або інтервали (для неперервних даних). На осі Y відображено абсолютну або відносну (у %) частотність для кожної категорії або інтервалу.

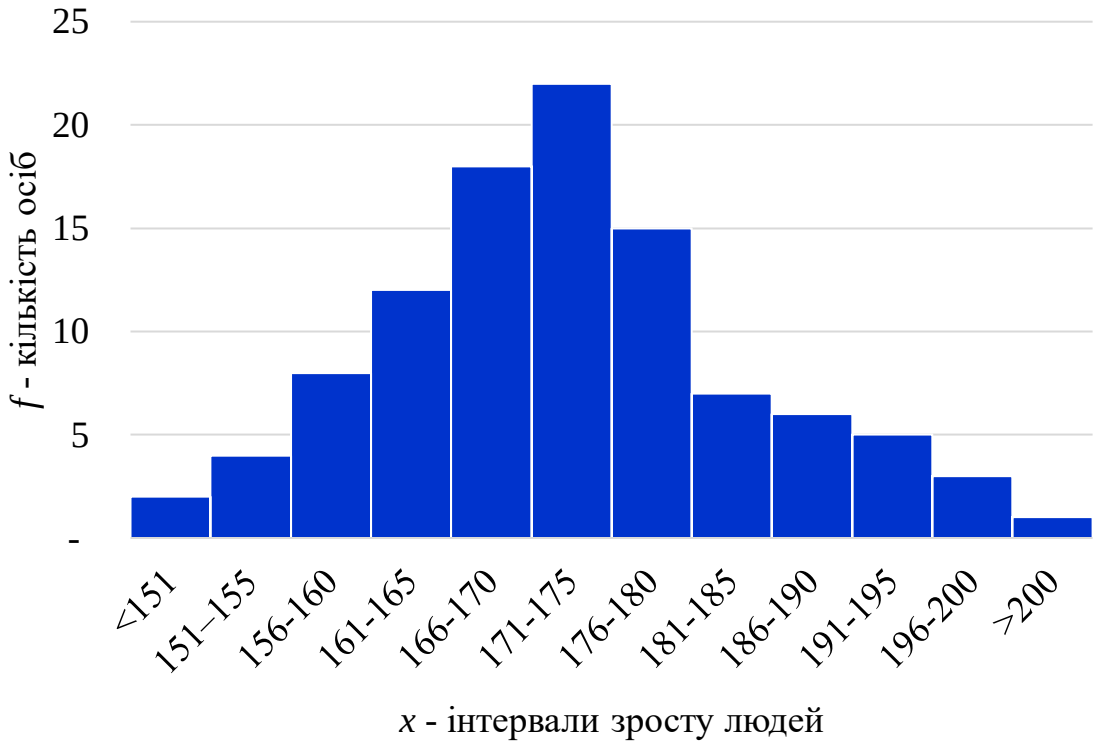


Рис. 1.6. Приклад гістограми: розподіл людей за зростом

b. Полігон частот (рис.1.7) є лінійною формою представлення розподілу даних. Цей графік є дуже близьким до гістограми, однак замість стовпців розподіл зображується за допомогою сполучених відрізків, акцентуючи увагу на зміні частот між інтервалами.

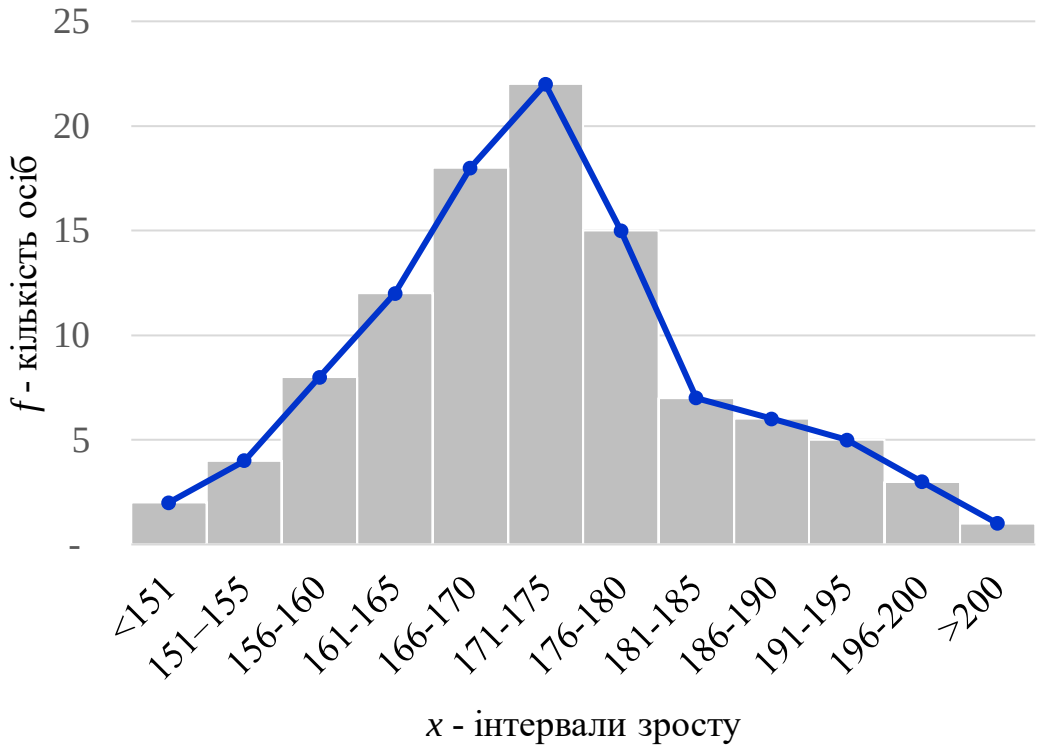


Рис. 1.7. Приклад полігону частот: розподіл людей за зростом

Розподіли на двох попередніх рисунках показують, що для даної групи людей найбільш поширеним був зріст у діапазоні 170-175 сантиметрів, а понад 70% людей з групи мають зріст в діапазоні від 161 до 185 сантиметрів.

c. **Діаграма «хмара слів», англ. “word cloud” (рис.1.8)** відображає розподіл ключових **словосполучень** за частотою їх згадування: що більше згадувань має словосполучення, то більшим є розмір його шрифту.

Рис. 1.8. Приклад «хмари слів»: результати опитування щодо обраної професії або сфери діяльності

d. **Картограма (рис.1.9)** – це карта, на якій певні географічні зони позначені кольорами або індикаторами різної інтенсивності відповідно до значення показника, який досліджується. На рисунку (1.9) видно, що в Ірландії та Люксембурзі рівень ВВП в розрахунку на душу населення був найвищим.

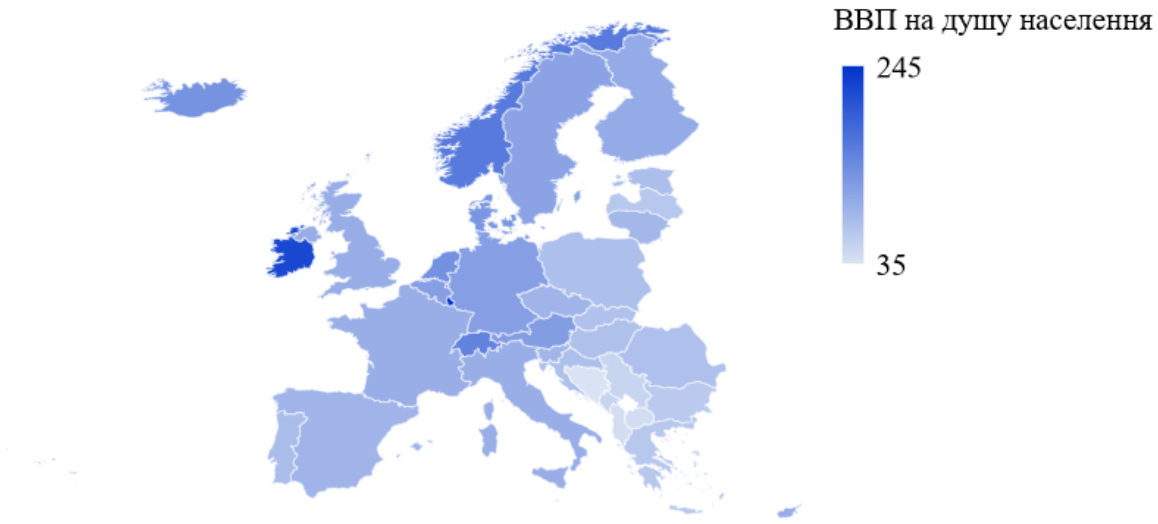


Рис. 1.9. Приклад картограми: розподіл деяких країн Європи за ВВП на душу населення у 2024 р. за паритетом купівельної спроможності

II. Візуалізації для аналізу структури даних і співвідношення груп

a. **Кругова діаграма, англ. "pie chart" (рис. 1.10)** дозволяє зобразити структуру сукупності, відображаючи співвідношення між частинами цілого, де розмір кожного сектору пропорційний до питомої ваги відповідної категорії станом на конкретний момент часу або за обраний період (сума всіх часток дорівнює 100%). Рисунок 1.10 наочно показує, що витрати на житло (30%) й одяг (20%) формували половину сімейного бюджету, а питома вага кожної з інших семи статей витрат коливалася у межах від 5% до 11%.

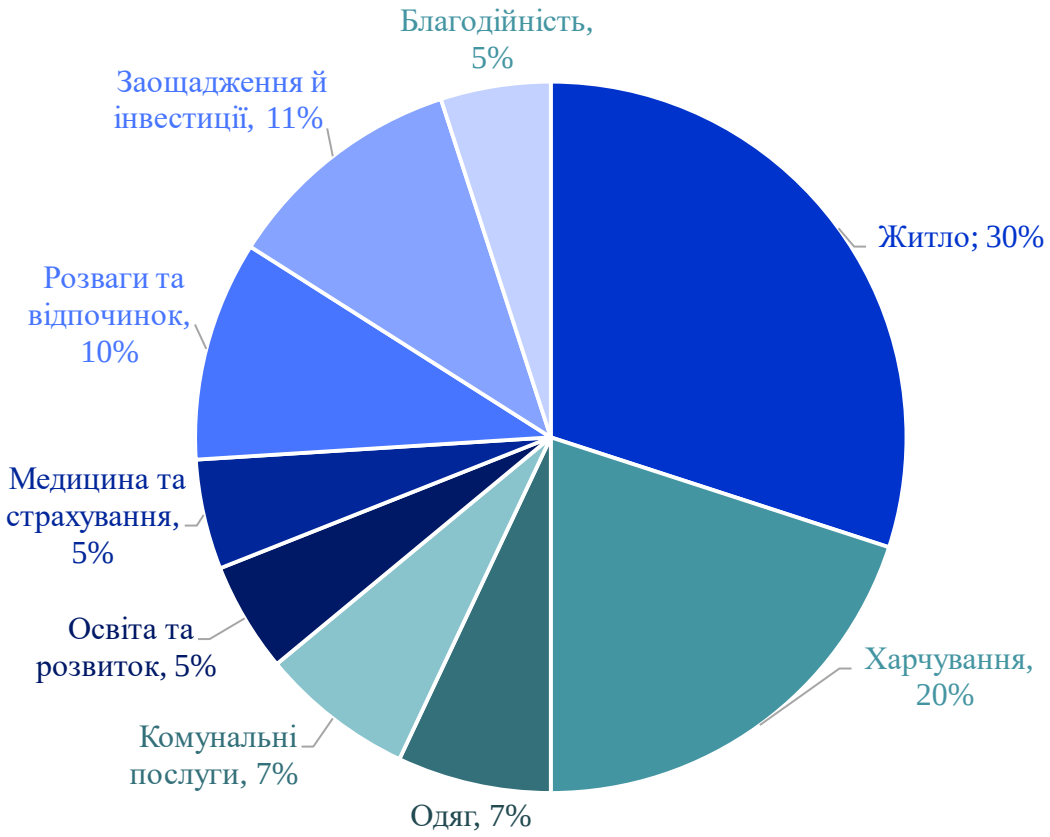


Рис. 1.10. Приклад кругової діаграми: структура місячного сімейного бюджету

b. Діаграма «деревовидна карта», англ. “treemap” (рис. 1.11) подібно до кругової діаграми показує структуру сукупності за категоріями. Основна перевага цієї візуалізації полягає в тому, що вона дає змогу легше порівнювати пропорції між категоріями, зокрема й на різних ієрархічних рівнях. Наприклад, на рисунку 1.11 кольором показано різні сектори економіки, в межах кожного виділено різні галузі пропорційно до їхньої валової доданої вартості.

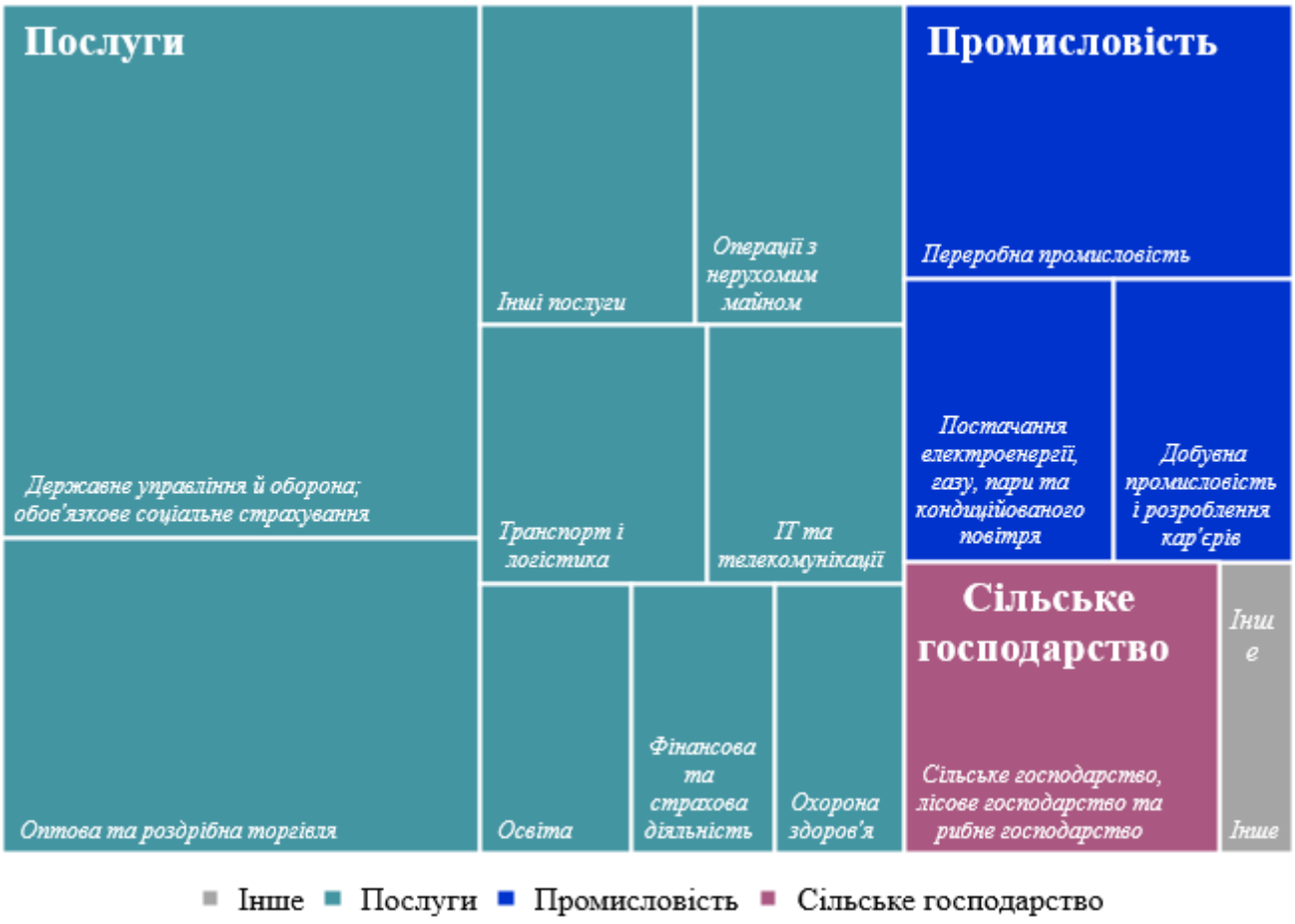


Рис. 1.11. Приклад деревовидної карти: структура валової доданої вартості за секторами економіки й галузями

c. **Діаграма «сонячні промені», англ. “sunburst” (рис. 1.12)** є розширеною версією кругової діаграми для відображення структури даних. На відміну від звичайної кругової діаграми, вона показує структуру відповідного сегменту одразу на кількох ієрархічних рівнях – від загального до деталізованого.

d. **Стовпчикова діаграма, англ. “bar chart” (рис. 1.13)** – це діаграма, призначена для порівняння показників за різними категоріями або часовими періодами (рис. 1.13 (а)). Стовпчикова діаграма з накопиченням (рис. 1.13 (б)) відображає зміну структури певного показника у динаміці, зокрема зміну доходів у розрізі продуктів протягом трьох років. Як ілюструє відповідний рис.1.13, найвищі доходи у розрізі всієї продуктової лінійки зафіксовано у 20X5 році. Водночас структура портфеля продуктів була диверсифікованою і динамічною: частка окремих сегментів суттєво

варіювалася протягом досліджуваного періоду, що виключало домінування будь-якого окремого продукту в загальному обсязі доходів.

Рис. 1.12. Приклад діаграми «сонячні промені»: індикативна структура генерації електроенергії в ЄС за типом палива у 2024 р.

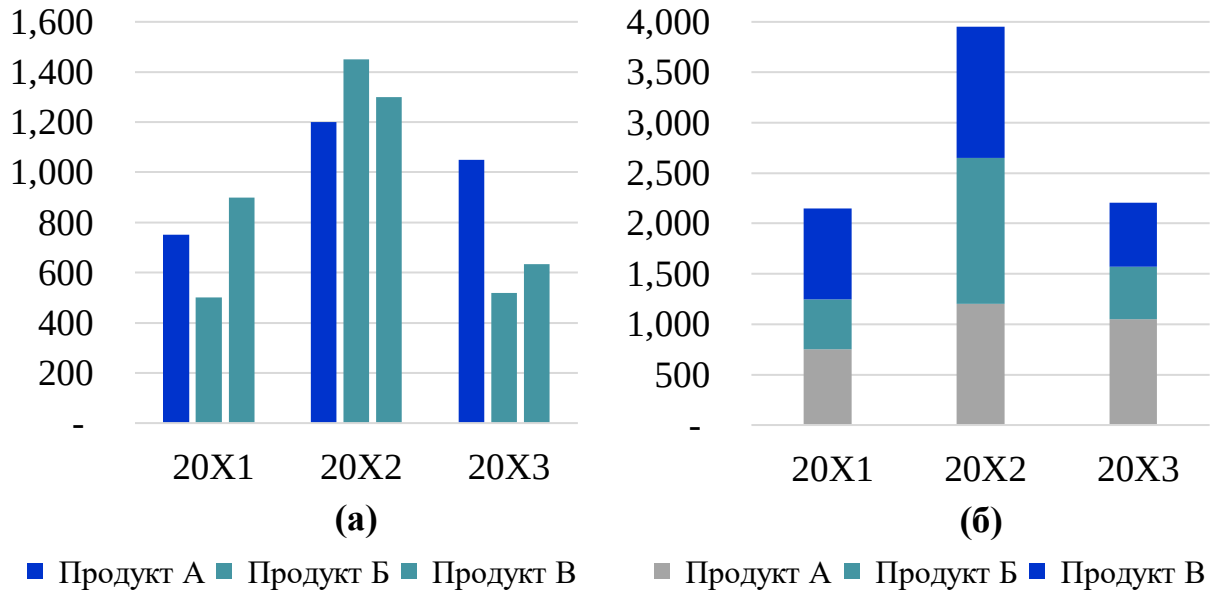


Рис. 1.13. Приклад стовпчикової діаграми: розподіл доходу компанії за роками й видами продукції

III. Візуалізації для аналізу динаміки показників

а. Лінійний графік (рис. 1.14) є основним видом візуалізації часових рядів. Він дозволяє зобразити та порівняти динаміку кількох показників протягом періоду аналізу для виявлення трендів, сезонності й екстремальних значень (піків і падінь). На рис. 1.14 показано, що дохід від продажу морозива має виразну сезонність: основний період високих продажів припадає на квітень–вересень, тоді як у решту місяців він залишається мінімальним. Водночас дохід від реалізації класичної кави є відносно стабільним протягом року. Також спостерігається тренд поступового зростання виручки від продажу круасанів, що, ймовірно, зумовлено розширенням асортименту цієї продукції.

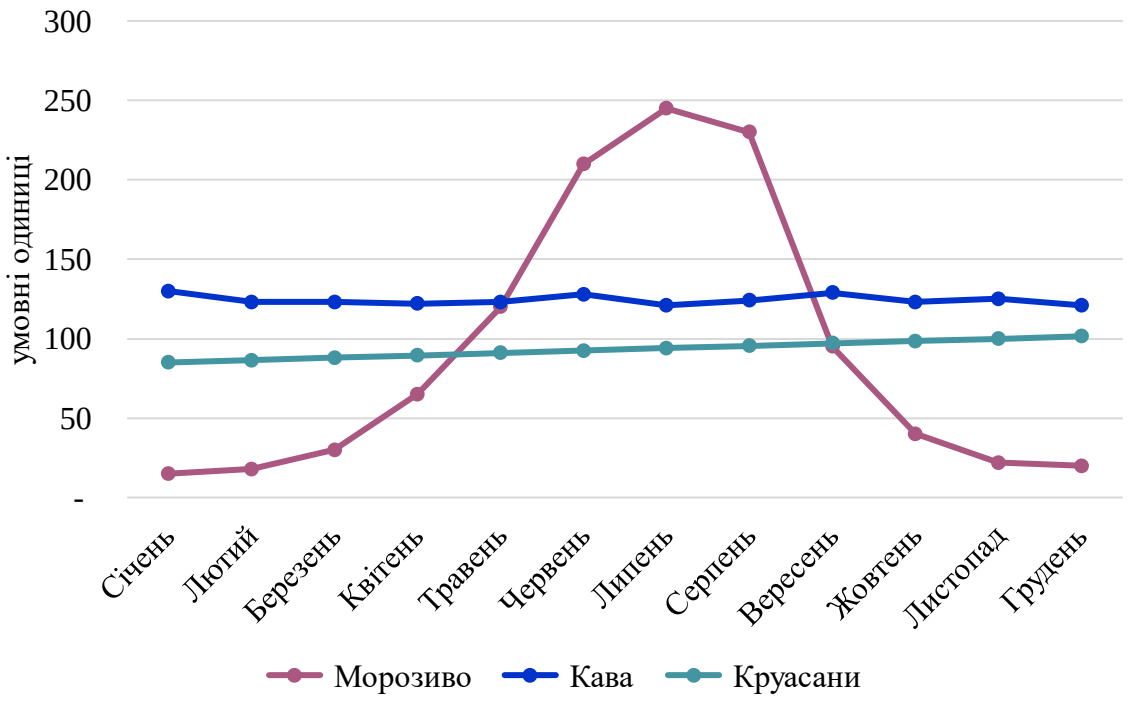


Рис. 1.14. Приклад лінійної діаграми: щомісячні доходів кондитерської у розрізі категорій продукції

IV. *Візуалізації для аналізу взаємозв'язків*

a. Діаграма розсіювання, англ. scatterplot (рис. 1.15) візуалізує щільність та напрям взаємозв'язку між двома показниками, тобто відображає їхню кореляцію. Кожна пара спостережень зображена на графіку у вигляді окремої точки. На рисунку 1.15 простежується пряма лінійна залежність між витратами на маркетинг та обсягом нових продажів: зростання маркетингових інвестицій супроводжується пропорційним збільшенням доходу (лінія, проведена між точками, має висхідний тренд). Значна щільність розташування точок відносно лінії тренду свідчить про високий ступінь кореляції, тобто стійкий статистичний зв'язок між показниками.

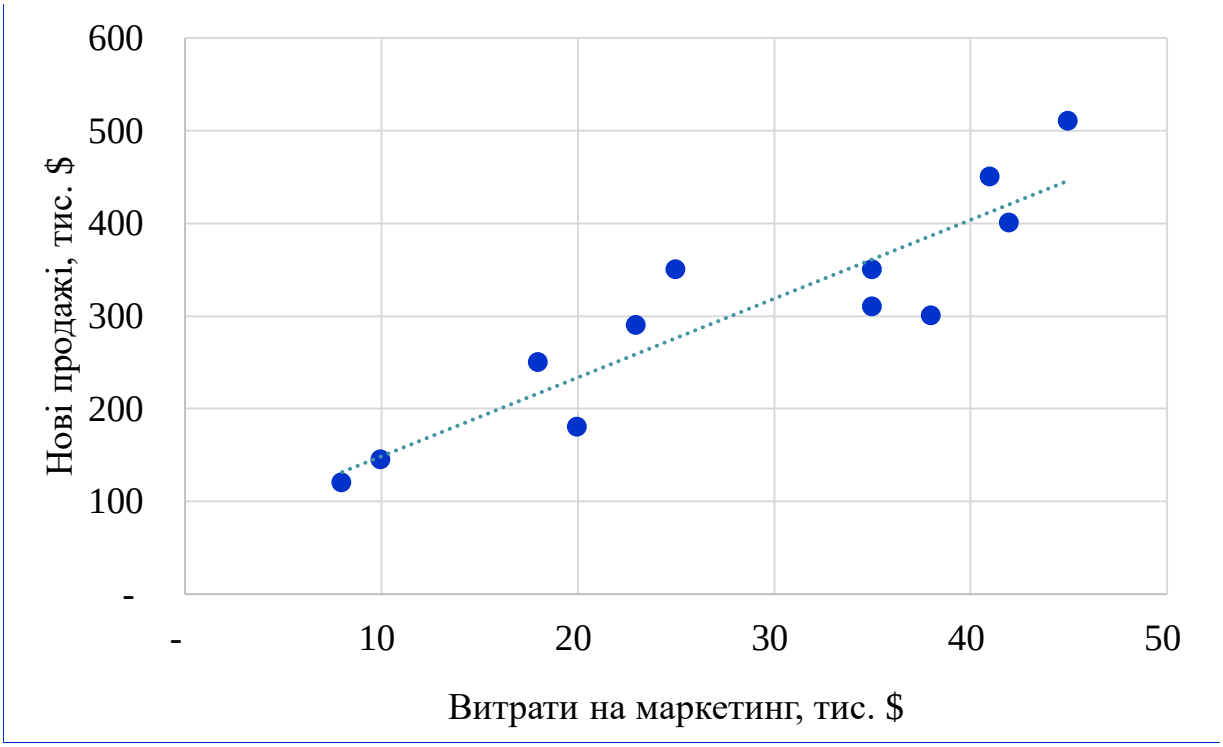


Рис. 1.15. Приклад діаграми розсіювання: витрати на маркетинг і нові продажі

b. **Теплова карта, англ. heatmap (рис. 1.16)** за допомогою кольорового кодування відображає числові значення у таблиці, що дає змогу аналітику швидко й зручно виявляти закономірності у великих масивах даних: зазвичай насичені або теплі кольори позначають вищу інтенсивність, а холодні чи менш насичені відтінки – нижчу.

75	49	65	39	32	36	58	28	48	39
29	75	60	43	23	31	38	75	60	54
85	47	75	74	26	79	54	31	65	31
41	40	29	34	86	65	75	41	58	26
28	61	71	49	77	67	61	84	70	63
40	67	37	71	70	25	63	82	72	27
73	59	65	82	57	50	44	80	34	75
59	39	75	36	73	66	68	64	68	36
67	62	54	26	33	85	41	78	72	40
62	33	83	37	29	38	37	29	59	76

Рис. 1.16. Приклад теплової карти

V. ***Візуалізація для факторного аналізу зміни показників***

a. **Каскадна діаграма, англ. waterfall, (рис. 1.17)** відображає основні чинники зміни значень показника за певний період. Вона демонструє «шлях» показника від базового значення до кінцевого результату через послідовні позитивні та негативні ефекти. Наприклад, на рисунку 1.17 показано ключові фактори зміни доходів протягом року. Загальне зростання доходу на 300 умовних одиниць зумовлене насамперед збільшенням продажів продукту А (+250 умовних одиниць), тоді як продукт Б мав відносно невеликий вплив (+50 умовних одиниць). Вирішальним чинником стало нарощування *кількості* проданих одиниць продукту А, а також підвищення *цін* на продукт Б. Водночас *ціновий* фактор продукту А та *кількісний* фактор продукту Б

спричинили негативний ефект на доходи. При подальшому аналізі варто окремо дослідити, чи зміна обсягів проданих товарів відбувалася внаслідок підвищення/зниження цін.

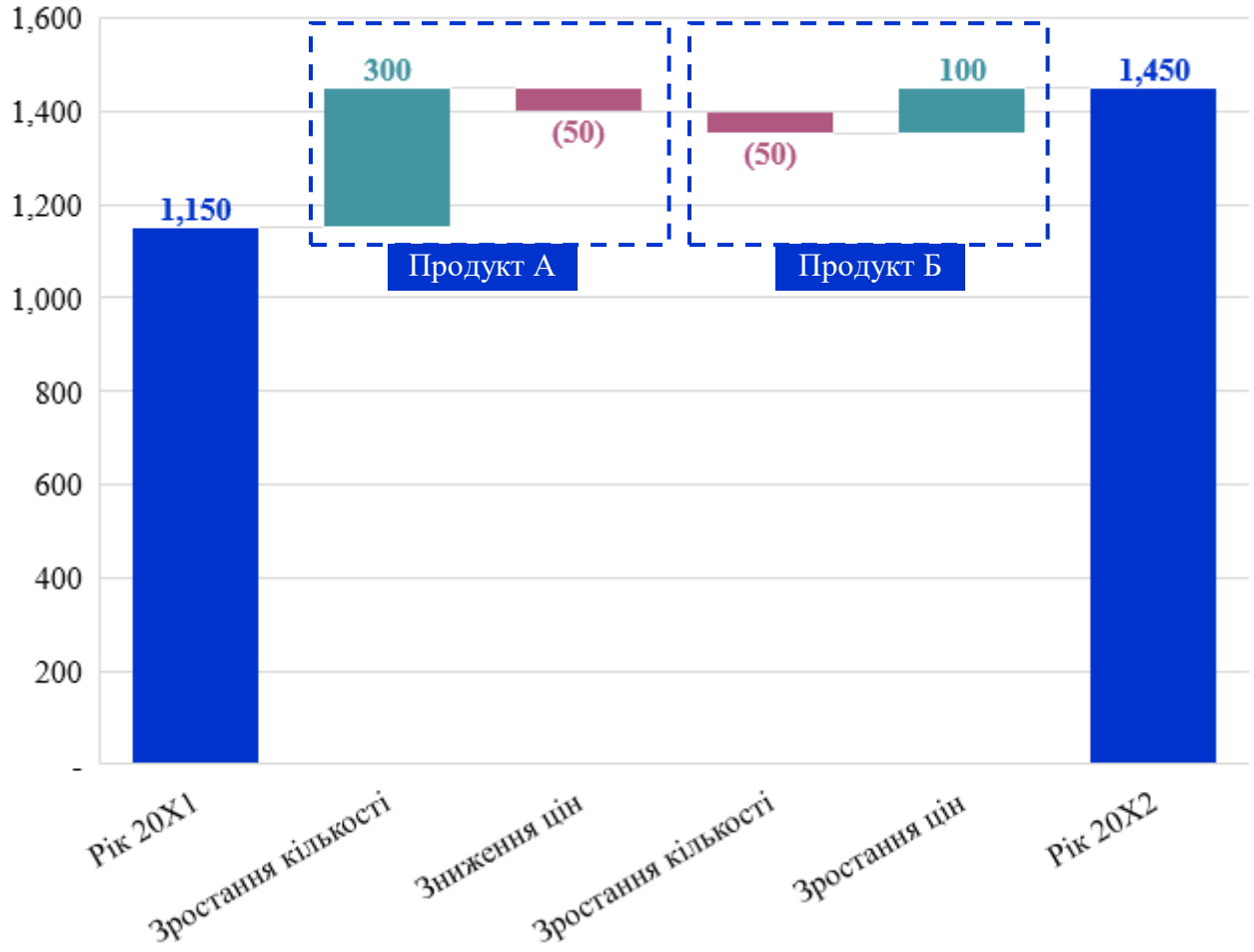


Рис. 1.17. Приклад каскадної діаграми: фактори зміни доходу за рік

Важливо зазначити, що під час вибору графіка варто насамперед орієнтуватися на цілі статистичного аналізу або презентації даних, а не лише на візуальну привабливість тієї чи іншої діаграми. На рисунку 1.18 наведено алгоритм, що допомагає обрати відповідну форму візуалізації залежно від аналітичних завдань.

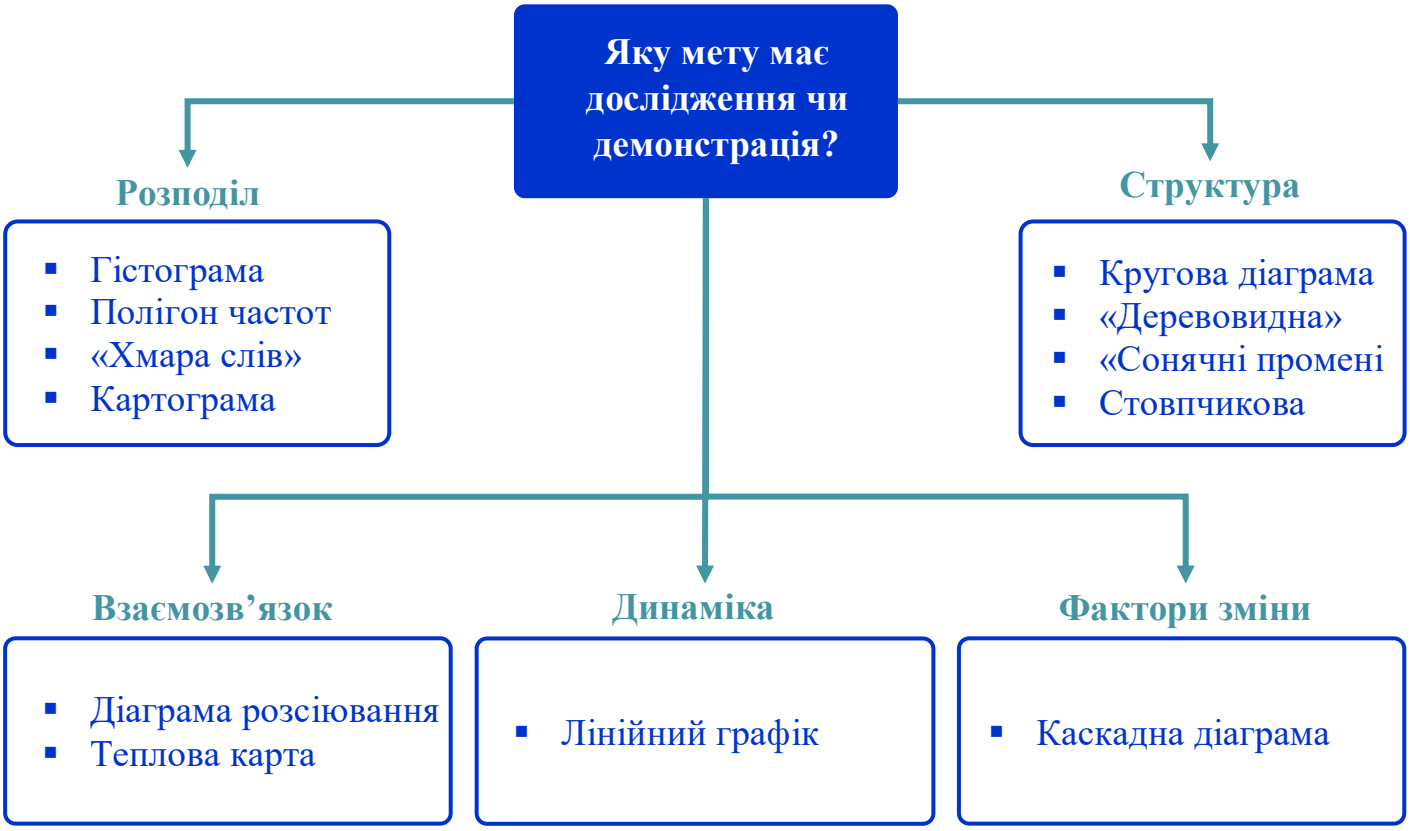
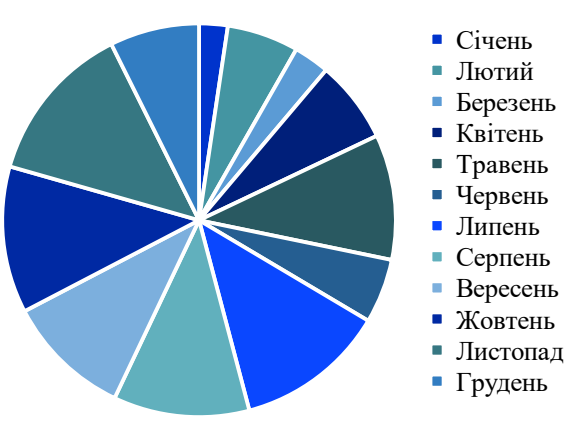


Рис. 1.18. Блок-схема для вибору відповідної візуалізації залежно від цілей аналізу

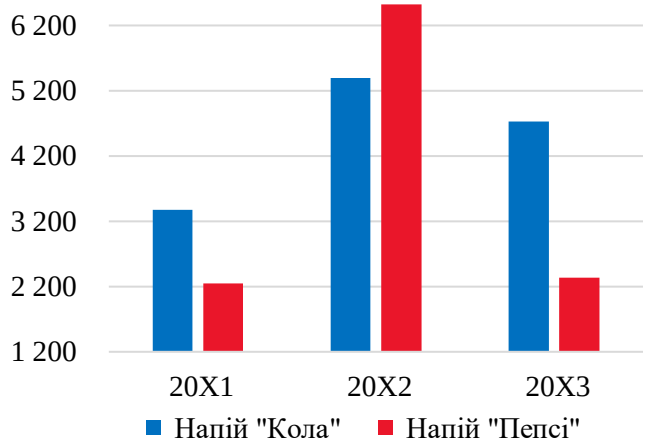
Графічний аналіз є невід'ємною складовою будь-якого статистичного дослідження, як на етапі виявлення прихованих тенденцій і взаємозв'язків, так і як засіб комунікації результатів цільовій аудиторії. Візуалізація часто дозволяє значно ефективніше презентувати отримані інсайти (від англ. «insights»), ніж текстові або числові дані в таблицях. Важливо пам'ятати, що адекватність висновків на основі графічного аналізу залежить передусім від таких чинників:

- відповідності візуалізації структурі даних і цілям дослідження;
- об'єктивності відображення даних, а також відсутності свідомих маніпуляцій чи випадкових помилок у презентації результатів;
- інформативності та зрозумілості подання матеріалу для цільової аудиторії.

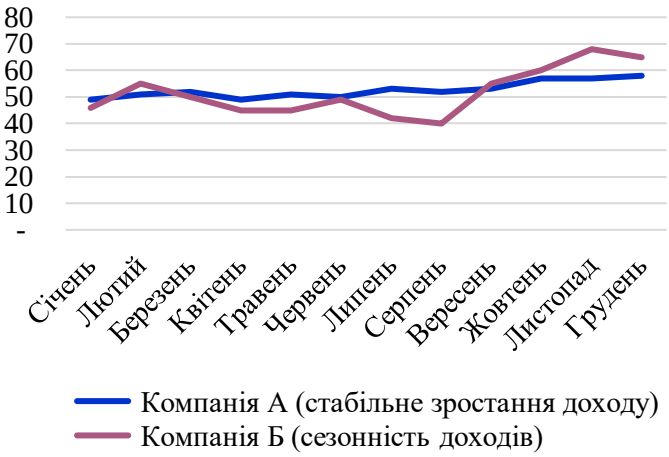
На рисунку 1.19 наведено приклади типових помилок при візуалізації даних.



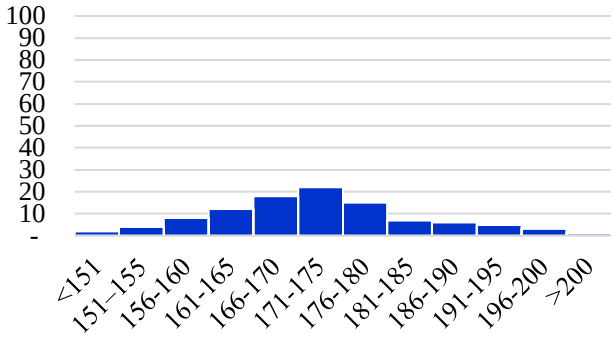
(а) Бюджет на маркетинг



(б) Динаміка продажів



(в) Динаміка доходів



(г) Розподіл групи за зростом

Рис. 1.19. Приклади типових помилок при візуалізації даних: діаграми (а) – (г)

Для ілюстрації поширених недоліків при візуалізації даних розглянемо кілька типових прикладів на рис. 1.19:

- *приклад (а):* для візуалізації часового ряду (наприклад, маркетингового бюджету з січня по грудень) доцільніше було б використати лінійний графік або стовпчикову діаграму для відображення динаміки витрат. Натомість кругова діаграма призначена для демонстрації структури, а не відображення змін у часі.
- *приклад (б):* початок відліку вертикальної осі (наприклад, для обсягу продажів напоїв) із позначки 1 200 замість нуля спотворює сприйняття даних. Таке масштабування штучно перебільшує різницю між показниками:

періоди із помірним зниженням продажів візуально виглядають як дуже суттєве падіння порівняно з піковими роками. Окрім цього, більшість людей асоціює умовний напій «Кола» із червоним кольором, а напій «Пепсі» - із синім. Подібні асоціації існують і для футбольних клубів, політичних партій чи державної символіки, де закріплені певні кольорові гами. Використання невідповідних кольорів, зокрема для візуалізації напоїв «Кола» і «Пепсі», може спотворити або принаймні ускладнити сприйняття та аналіз даних.

- *приклад (в):* На лінійному графіку представлено динаміку доходів Компанії А (стабільне зростання) та Компанії Б (виражена сезонність). На графіку спостерігається надмірне горизонтальне розтягування. Це призвело до згладжування природних коливань, через що реальна різниця в динаміці доходів компаній та сезонні відмінності виглядають менш суттєвими, ніж вони є насправді.
- *приклад (г):* інтервал позначок на вертикальній осі (від 0 до 100) не відповідає реальному діапазону значень у представленому розподілі (максимальне значення по вертикальній осі – 22). Унаслідок цього графік виглядає «сплющеним» відносно горизонтальної осі, що зменшує видиму варіативність між різними стовпцями та ускладнює інтерпретацію даних.

Питання для самоперевірки

1. *Для яких цілей найкраще використовувати гістограму, а для яких – лінійний графік?*
2. *Який вид візуалізації використовують для аналізу кореляції (взаємозв'язку) між двома показниками?*
3. *Що відображає каскадна діаграма (waterfall)?*
4. *Які основні причини появи статистичних міфів?*
5. *Назвіть типові помилки при візуалізації даних, які можуть спотворити сприйняття результатів.*

РОЗДІЛ 2. ОПИСОВА СТАТИСТИКА ЯК ПЕРВИННИЙ ЕТАП ДОСЛІДЖЕННЯ

2.1. Міри центральної тенденції розподілу та їх інтерпретація

Майже кожне кількісне дослідження розпочинається з етапу ознайомлення з певною сукупністю даних. Враховуючи, що кількісні дані в первинному необробленому табличному вигляді можуть бути значними за обсягом і складними для сприйняття, виникає потреба у певних універсальних інструментах, які дозволили б системно й структуровано описати й візуалізувати наявні кількісні дані. Адже складно перевіряти статистичні гіпотези або будувати економіко-математичну модель, не маючи глибокого розуміння джерел інформації та їхніх статистичних характеристик.

Розділ описової статистики пропонує вирішення цієї проблеми на первинних етапах дослідження, використовуючи універсальний аналітичний інструментарій, що має на меті узагальнити усю сукупність даних за певними статистичними характеристиками, серед яких: середні значення; показники варіації, однорідності й форми розподілу; представлення даних у вигляді графіку щільності розподілу тощо.

Перша група статистичних метрик, яку ми розглянемо – це міри центральної тенденції. Середнє значення є базовим і найпоширенішим показником для характеристики сукупності значень. Він широко використовується у

повсякденному побутовому й професійному житті, тому є інтуїтивно зрозумілим для багатьох людей.

Середнє арифметичне просте – це вид середнього значення, який розраховується для *незгрупованих даних* як сума індивідуальних значень певної ознаки (x_j), поділена на загальну кількість значень у сукупності (n). Приклади: середній бал за навчальний предмет у школі; ВВП на душу населення у певній країні; кількість спожитих чашечок кафи за місяць в розрахунку на одного працівника в офісі. Для розрахунку середнього арифметичного простого використовують таку формулу:

$$\bar{x} = \frac{\sum_{j=1}^{n} x_j}{n} \qquad (2.1)$$

де $\bar{x}$ — *просте середнє арифметичне;*

x_j — *значення ознаки у j-го елемента сукупності;*

j — *порядковий номер елемента сукупності;*

n — *кількість елементів у сукупності;*

$\sum_{j=1}^{n}(\dots)$ – *оператор послідовного сумування елементів сукупності.*

На рисунку 2.1 наведено приклад розрахунку середнього арифметичного простого шляхом написання коду у Python, використовуючи готові функції з бібліотек «numpy» та «statistics». Вхідні дані містяться у векторі *x* (список або «list» у Python): [-100, 10, 20, 30, 40, 50, 75, 100, 125, 150]. Щоб порахувати середнє значення вручну за формулою (2.1) для заданих десяти чисел, виконаємо такі арифметичні дії: *(-100 + 10 + 20 + 30 + 40 + 50 + 75 + 100 + +125 + 150) / 10 =* ***50.***

Код у Python

```python
import numpy as np
import statistics

x = [-100, 10, 20, 30, 40, 50, 75, 100, 125, 150]

# Спосіб 1
mean1 = np.mean(x)
print("Середнє арифметичне просте: ", mean1)

# Спосіб 2
mean2 = statistics.mean(x)
print("Середнє арифметичне просте: ", mean2)

### Результат ###
# Середнє арифметичне просте:  50
```

Рис. 2.1. Код у Python: розрахунок середнього арифметичного простого

Середнє арифметичне зважене – це вид середнього значення, який розраховується для *згрупованих даних* як сума індивідуальних значень певної ознаки (x_j), зважених на відносні частоти ($f_j \ / \ \sum f_j$) індивідуальних ознак у сукупності. Цей підхід до розрахунку середнього арифметичного використовується у випадках, коли окремі елементи сукупності мають різну вагу, значущість, імовірність, або частоту. Приклади: середній бал за навчальний предмет в університеті, де у кожного предмета є своя «вага» (на основі кредитів ECTS) у загальному рейтингу залежно від відповідного навчального навантаження; розрахунок середньої дохідності інвестиційного портфеля, що складається з різних за розміром інвестицій у набір активів; показник інфляції, або індекс споживчих цін, що розраховується на основі кошику окремих товарів з різною вартістю (або часткою у витратах домогосподарства). Для розрахунку середнього арифметичного зваженого використовують таку формулу:

$$\bar{x} = \frac{\sum_{j=1}^{n} x_j f_j}{\sum_{j=1}^{n} f_j} = \sum_{j=1}^{n} x_j w_j \qquad (2.2)$$

де $\bar{x}$ — *зважене середнє арифметичне;*

x_j — *значення ознаки у j-го елемента сукупності;*

f_j — *частота, або кількість елементів j-го елемента сукупності;*

w_j — *відносна вага j-го елемента сукупності, де* $w_j = f_j / \sum f_j$

$\sum f_j = n$ — *кількість елементів у всій сукупності.*

На рис. 2.2 наведено приклад розрахунку середнього арифметичного зваженого у Python, використовуючи готову функції з бібліотеки «numpy». Скористаємося формулою (2.2), щоб порахувати середнє зважене вручну: *(10*2 + 20*3 + 30*5) / 10 =* ***23.***

Код у Python

```
import numpy as np

x = [10, 20, 30] # значення ознаки у групі j
f = [2, 3, 5] # частота (кількість елементів) у групі j

mean_weighted = np.average(x, weights = f)

print("Середнє арифметичне зважене: ", mean_weighted)

### Результат ###
# Середнє арифметичне зважене:  23.0
```

Рис. 2.2. Код у Python: розрахунок середнього арифметичного зваженого

Математичний зв'язок між простою та зваженою формами середнього арифметичного можна простежити, якщо уявити, що *зважене* середнє агрегує однорідні значення ознаки (x_j) в окремі групи. При цьому «вага» кожної групи (f_j) визначається кількістю зібраних у ній однакових елементів. Інтуїтивно цей принцип

можна представити у вигляді розрахунку вартості набору монет різного номіналу, як зображено на рис. 2.3. У верхньому рядку представлено розрахунок за формулою *простого* середнього, а у нижньому – *зваженого*, об'єднавши ознаки (монети за номіналом) у групи.

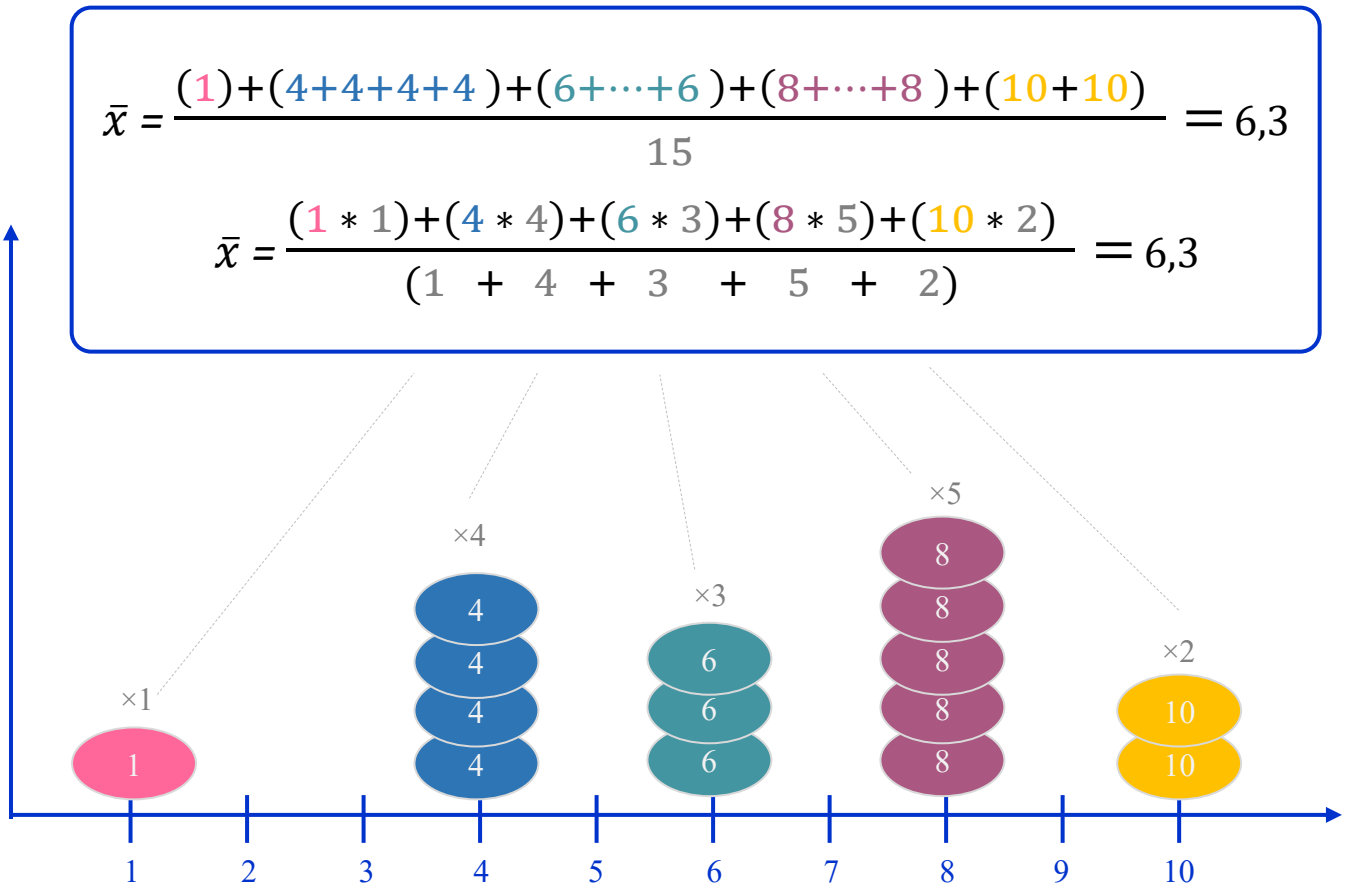


Рис. 2.3. Код у Python: зв'язок між середнім арифметичним простим і зваженим

Хоча середнє арифметичне широко використовується на практиці для первинних аналітичних розрахунків, воно має певні обмеження. Основний недолік цього методу усереднення – це висока чутливість до екстремальних значень (від англ. «outliers» – викиди). У побуті це явище іронічно називають «середньою температурою по палаті». Наприклад, показник середньої заробітної плати у компанії на рівні 50 тис. грошових одиниць є мало інформативним, якщо десять співробітників отримують по 10 тис. грошових одиниць, а керівник – 450 тис. У такому разі середнє значення не відображає адекватно реальний рівень винагороди жодного з працівників. Звичайно, якщо сукупність є відносно однорідною, без

елементів з екстремальними значеннями, середнє арифметичне залишається надійним інструментом усереднення.

Коли аномальні значення все ж присутні у вибірці, то можна розрахувати **середнє «відсічене» (від англ. «trimmed mean»)** – показник, який розраховується як середнє арифметичне *після видалення* певної частки найбільших та найменших значень у сукупності, наприклад, 5% або 10% найбільших і найменших значень. Обсяг відсікання (k) визначають за формулою:

$$k = \text{n} * \text{p}\% \qquad (2.3)$$

де k – *обсяг відсікання;*

n – *обсяг сукупності;*

$p\%$ – *частка сукупності, що підлягає відсіканню;*

Наприклад, розрахуємо цей показник (з відсіченням 10%) для вектору чисел [-100, 10, 20, 30, 40, 50, 75, 100, 125, 150]. За умовою виходить, що p = 10%, n = 10. Відповідно, $k = 10 * 10\% = 1$. Це означає, що при розрахунку відсіченого середнього необхідно прибрати з впорядкованої сукупності чисел одне найбільше та одне найменше значення, тобто -100 і 150 у наведеному прикладі. Після цього розраховуємо середнє арифметичне за формулою (2.1): (10 + 20 + 30 + 40 + 50 + 75 + 100 + 125) / 8 = **56,25**. Отже, цей метод дозволяє виключити з розрахунку екстремальні значення, які викривлюють розрахунок середнього арифметичного. Завдяки симетричному відсіканню «хвостів» ми одержуємо більш об’єктивну оцінку центру розподілу. Автоматизований розрахунок відсіченого середнього у Python наведено на рисунку 2.4.

Код у Python

```
from scipy import stats

x = [-100, 10, 20, 30, 40, 50, 75, 100, 125, 150]

mean_trimmed = stats.trim_mean(x, proportiontocut = 0.1)
print("Середнє арифметичне відсічене: ", mean_trimmed)

### Результат ###
# Середнє арифметична відсічена:  56.25
```

Рис. 2.4. Код у Python: розрахунок середнього арифметичного відсіченого

У формулах середнього арифметичного ми розглядали усереднення абсолютних показників. До них належать, зокрема, кількість балів, обсяг ВВП, кількість вироблених одиниць продукції, обсяг витрат у грошових одиницях тощо.

Існують задачі, коли необхідно усереднити не абсолютні величини, а відносні показники інтенсивності, серед яких кількість вироблених деталей за одиницю часу, затрати часу на збирання зернових культур з 1 га поля, швидкість руху (км/год), ціни на товари (грн/одиницю), ефективність маркетингових витрат на залучення клієнтів (грн/клік, грн/клієнта). У таких випадках використання простого середнього арифметичного є некоректним, адже воно не враховує різний масштаб спостережень (різна вартість, площа, тривалість).

Середнє гармонійне – це статистичний показник, який використовується для усереднення показників інтенсивності, зокрема тоді, коли частоти (f_j) невідомі у явній формі, а в умові задачі дано лише значення ознаки (x_j) та добутку $x_j f_j$ (масштаб залежно від контексту: загальна відстань, вартість продукції, обсяг витрат, фонд інвестицій тощо). Для кращого розуміння сутності цього показника наведемо приклад. Формули для розрахунку середнього гармонійного мають вигляд, обернений до простого середнього:

- Середнє гармонійне *просте* – при *однакових* S_j для кожного j:

$$\bar{x} = \frac{n}{\sum_{j=1}^{n} \frac{1}{x_j}} \qquad (2.4)$$

- Середнє гармонійне *зважене* – якщо S_j *відрізняються* для різних елементів:

$$\bar{x} = \frac{\sum_{j=1}^{n} S_j}{\sum_{j=1}^{n} \frac{1}{x_j} S_j} \qquad (2.5)$$

де $\bar{x}$ — *середнє гармонійне;*

x_j — *значення ознаки у j-го елемента сукупності;*

S_j — *загальний обсяг/вартість для j-го елемента, де* $S_j = x_j * f_j$

$n = \sum f_j$ — *кількість елементів у всій сукупності.*

Формула (2.4) є частковим випадком формули (2.5): при $S_1 = S_2 = S_j$ суму всіх вартостей $\sum S_j$ можна записати у вигляді nS_j. Тоді загальний обсяг S_j скоротиться у чисельнику й знаменнику формули (2.5).

Приклад 1 «Закупівля яблук на однакову суму в кожного продавця»

Уявімо, що всього ми витратили 2 000 гр.од. на закупівлю яблук: в одного продавця ми придбали яблука в рамках бюджету 1 000 гр.од. за ціною 50 гр.од. / кг (грошових одиниць за кілограм), а іншому заплатили стільки ж, але за ціною 100 гр.од. / кг. Виникає питання: якою була середня ціна одного кілограма яблук за результатами здійснених закупівель? Для наочності представимо умови задачі у таблиці 2.1.

Таблиця 2.1. Задача: визначення середньої закупівельної ціни яблук (при *однакових обсягах* закупівель в кожного продавця)

Контрагент	**Ціна 1 кг яблук, гр.од.**	**Вартість закупівель, гр.од.**	***Розрахунок:* обсяг закупівель, кг**
Продавець 1	50	1 000	20 = 1 000 / 50
Продавець 2	100	1 000	10 = 1 000 / 100
Всього	**X**	**2 000**	**30**

У декого може виникнути інтуїтивне припущення, що середня ціна склала 75 гр.од. / кг (подумки використавши формулу простого арифметичного: (50 + 100) / 2). Однак це некоректна відповідь, адже вона не враховує, що в першого продавця ми придбали 20 кг яблук (1 000 / 50), а в другого – 10 кг яблук (1 000 / 100). Відповідно, при усередненні потрібно враховувати різні обсяги закупівель, адже більшість яблук у нашому кошику мають ціну 50 грн, тому середня ціна має бути ближчою до 50 гр.од./кг, ніж до 100 гр.од./кг. Застосуємо формулу гармонійного середнього простого, одержавши: 2 / (1/50 + 1/100) = **66,7 гр.од./кг**, що суттєво відрізняється від 75 гр.од./кг при використанні простого середнього арифметичного. Приклад розрахунку в Python наведено на рис. 2.5.

Код у Python

```
import statistics

x = [50, 100] # Ціни на яблука в різних продавців

mean_harmonic = statistics.harmonic_mean(x)
print("Середнє гармонійне: ", mean_harmonic)

### Результат ###
# Середнє гармонійне:  75
```

Рис. 2.5. Код у Python: розрахунок середнього гармонійного простого

Приклад 2 «Закупівля яблук на різні суми у двох продавців»

У цій задачі припустимо, що на закупівлі в першого продавця ми витратили 1 000 гр.од., придбавши яблука за ціною 50 гр.од./кг, а в другого – 3 000 гр. од. за ціною 100 гр.од./кг. Аналогічно до попереднього прикладу, представимо умови задачі у таблиці 2.2 для наочності.

Таблиця 2.2. Задача: визначення середньої закупівельної ціни яблук (при *різних обсягах* закупівель в кожного продавця)

Контрагент	**Ціна 1 кг яблук, гр.од.**	**Вартість закупівель, гр.од.**	***Розрахунок:* обсяг закупівель, кг**
Продавець 1	50	1 000	20 = 1 000 / 50
Продавець 2	100	3 000	30 = 3 000 / 100
Всього	**X**	**4 000**	**50**

Відповідно, в першого продавця ми придбали 20 кг яблук (1 000 / 50), а в другого – 30 кг яблук (3 000 / 100). Застосуємо формулу гармонійного середнього зваженого, одержавши: 4 000 / (1/50 * 1 000 + 1/100 * 3 000) = **80,0 гр.од./кг**, що відрізняється від 75 гр.од./кг при використанні простого середнього арифметичного. Приклад розрахунку в Python наведено на рис. 2.6.

Код у Python

```
import statistics

x = [50, 100] # Ціни на яблука в різних продавців
S = [1000, 3000] # Вартість закупівель в різних продавців

mean_harmonic = statistics.harmonic_mean(x, weights = S)
print("Середнє гармонійне: ", mean_harmonic)

### Результат ###
# Середнє гармонійне:  80
```

Рис. 2.6. Код у Python: розрахунок середнього гармонійного зваженого

Якщо нам необхідно усереднити показники, що змінюються пропорційно один до одного, зокрема *темпи зростання* якогось економічного показника у %, то варто застосовувати *середнє геометричне*.

Середнє геометричне – це статистичний показник, який використовується для усереднення *відносних показників динаміки*, зокрема дохідності, темпів зміни ВВП, витрат, інвестицій тощо. Темпи приросту можуть вимірюватися у відсотках (наприклад, на +5% або -5%), коефіцієнтах (у 1,05 разів або 0,95 разів) або індексах (105 пунктів або 95 пунктів). Базова формула середнього геометричного має такий вигляд:

$$\bar{x} = \sqrt[n]{\prod_1^n x_j} = \sqrt[n]{x_1 * x_2 * \dots * x_n} \qquad (2.6)$$

де $\bar{x}$ – *середнє геометричне;*

x_j – *значення ознаки у j-го елемента сукупності (невід'ємні значення);*

$\prod_{j=1}^{n}(\dots)$ – *оператор послідовного множення елементів сукупності.*

При застосуванні формули середнього геометричного важливо зважати на формат представлення значень ознаки, які використовуються як вхідні дані у формулі (x_j). Формула (2.6) припускає, що темпи приросту записані у вигляді коефіцієнтів (тобто 1,1 замість +10%, або 0,9 замість -10%). Якщо вхідні дані записані у форматі відсотків, то формулу необхідно скоригувати, записавши у вигляді:

$$1+\bar{x} = \sqrt[n]{\prod_1^n (1+x_j)} = \sqrt[n]{(1+x_1) * (1+x_2) * \dots * (1+x_{n)}}$$

$$\bar{x} = \sqrt[n]{\prod_1^n (1+x_j)} - 1 = \sqrt[n]{(1+x_1) * (1+x_2) * \dots * (1+x_{n)}} - 1 \qquad (2.7)$$

де $\bar{x}$ – *середнє геометричне;*

x_j – *значення ознаки у j-го елемента сукупності (невід'ємні значення);*

$\prod_{j=1}^{n}(\dots)$ – *оператор послідовного множення елементів сукупності.*

Така трансформація у вигляді (1 + Δ%) пов'язана з деякими математичними та практичними причинами:

- Добуток під коренем парного степеню має бути *невід'ємною величиною* за визначенням. Додавання одиниці перетворює відсоток на коефіцієнт: наприклад, -5% можна записати у вигляді коефіцієнта як 0,95 = 1 + (-0,05).
- Формула середнього геометричного припускає, що ми множимо не просто відсотки зростання, а розраховуємо, у скільки разів змінюється початкова сума в кожному періоді, враховуючи *накопичений ефект* капіталізації (або «наслоювання»). Наприклад, відсоток зміни ціни акції кожного періоду нараховується на вже змінену базу, яка враховує прирости попередніх років (наприклад, 1,01 * 1,05 * 0,95). Перемножування власне темпів приросту не має практичного змісту (наприклад, +1% * +5% * -5%).

Код у Python

```
import statistics

x_percentage = [-35, 15, 20] # Річні темпи приросту у % (-35%, +15%, +20%)
x = []

# Трансформація вхідних даних у формат коефіцієнтів
for i in x_percentage:
    coef_format = ( i / 100 ) + 1
    x.append(coef_format)

print(x) # Річні темпи приросту у форматі коефіцієнтів (на 0,65; 1,15; 1,20)

mean_geom = (statistics.geometric_mean(x) - 1) * 100
print("Середнє геометричне: ", mean_geom)

### Результат ###
# Середнє геометричне:  1,036 (або +3,6%)
```

Рис. 2.7. Код у Python: розрахунок середнього геометричного

На прикладі наступного розрахунку можна переконатися, що застосування простого середнього арифметичного для усереднення темпів приросту є некоректним, адже не враховує мультиплікативний ефект складних відсотків.

Приклад розрахунку середнього темпу приросту

Розглянемо ілюстративний приклад, наведений у таблиці 2.3. У *рядку 1* наведено ціну ціни акції однієї з компаній, що зростала або знижувалася у кожному періоді під впливом ринкових факторів. У *рядку 2* розраховано темп такої зміни у кожному році, порівняно з попереднім періодом (пригадаємо підхід до розрахунку відносних показників динаміки). Темп зміни ціни акції у першому році склав: (65 / 100 – 1) * 100 = -35%. Використовуючи аналогічну формулу, розрахуємо, що у наступному році ціна зросла на 15%, а в крайньому періоді – на 20%.

Трансформувавши темп приросту у відсотках до коефіцієнтного вигляду у форматі (1 + %), одержимо у *рядку 3*: (1 + -0,35) = **0,65**; (1 + 0,15) = **1,15**; (1 + 0,2) = **1,2**.

Формат запису темпів приросту у *рядку 2* використаємо для розрахунку середнього арифметичного, а у *рядку 3* – середнього геометричного:

- Середньорічний приріст ціни акції на основі *середнього арифметичного* складає: (-35% + 15% + 20%) / 3 = **0%**. З розрахунку випливає, що в середньому ціна акції не змінювалася протягом трьох років. Однак, як видно з таблиці, це не відповідає дійсності: ціна акції знизилася зі 100 ум.од. у базовому періоді до 90 ум.од. у третьому році, тобто приблизно на 10,3%. Окрім цього, можна помітити, що ціна акції суттєво коливалася протягом досліджуваного періоду: -35%↓, +15%↑, +20%↑.
- Середньорічний приріст ціни акції на основі *середнього геометричного* складає: $(0{,}65 * 1{,}15 * 1{,}20)^{1/3} - 1 = - 0{,}036$, або **-3,6%**. Результати розрахунку демонструють, що в середньому ціна акції за рік знижувалася на 3,6%. Враховуючи загальне падіння ціни на 10,3%, результат у -3,6% за в середньому за рік виглядає співставним.

Отже, саме середнє геометричне дозволяє коректніше відобразити мультиплікативний ефект темпу зміни статистичної величини, на відміну від

середнього арифметичного, яке може давати суттєву системну похибку, особливо за наявності різноспрямованих коливань (темпів приросту з різними знаками).

Таблиця 2.3. Задача: розрахунок середнього темпу приросту ціни акції

№	Показник	Рік 0 (базовий період)	Рік +1	Рік +2	Рік +3
1	Ціна акції компанії, *ум.од.*	100	65	75	90
2	Темпи зміни ціни, *%*	X	-35,0%	15,0%	20,0%
3	Темпи зміни ціни, *разів (коефіцієнт 1 + %)*	X	0,65	1,15	1,20

Медіана є ще одним важливим характеристичним показником міри центральної тенденції статистичного розподілу. Однак, на відміну від розглянутих раніше середніх значень, методика її розрахунку базується на іншому підході.

Медіана – це середина *впорядкованого* ряду даних, яка ділить його на дві рівні за обсягом частини. Тобто для розрахунку медіани необхідно спочатку відсортувати значення від найбільшого до найменшого (або навпаки) і порахувати їхню загальну кількість. У разі непарної кількості елементів сукупності, медіаною є число у середині розподілу. Якщо обсяг сукупності є парним, то медіаною є середньоарифметичне двох сусідніх чисел у центрі розподілу.

Для наочної ілюстрації методу розрахунку розглянемо приклад, що представлений на рисунку 2.8

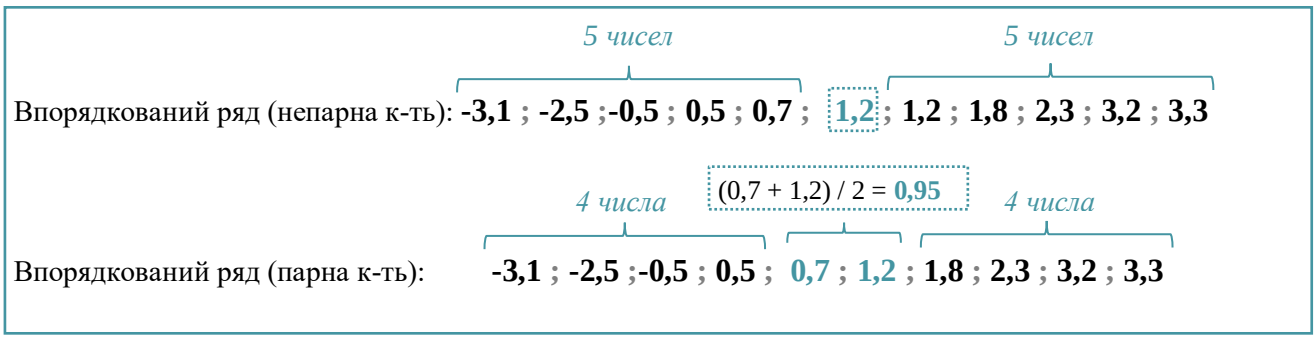


Рис. 2.8. Приклад визначення медіани

Впорядкувавши ряд чисел (див. рис. 2.8), визначимо медіану як **1,2** (для рядку з парною кількістю елементів), адже по обидва боки від медіани знаходиться однакова кількість чисел. Для рядку з непарною кількістю елементів визначаємо медіану як середнє арифметичне для двох чисел посередині ряду: (-0,7 + 1,2) = **0,95**.

Цікаво, що при використанні Python наявні функції автоматично впорядковують ряд чисел при розрахунку медіани, як представлено на рисунку 2.9.

Код у Python

```
import statistics

x_парний= [1.2,  3.3, -3.1, 0.7, -2.5, 0.5, 3.2, 0.7, 1.2, -0.5, 1.8, 2.3, 3.2]
x_непарний= [1.2,  3.3, -3.1, 0.7, -2.5, 0.5, 3.2, 0.7,-0.5, 1.8, 2.3, 3.2]

median_парн = statistics.median(x_парний)
print("Медіана (парний ряд): ", median_парн)

median_непарн = statistics.median(x_непарний)
print("Медіана (непарний ряд): ", median_непарн)

### Результат ###
# Медіана для парного ряду:  1,2
# Медіана для непарного ряду:  0,95
```

Рис. 2.9. Код у Python: розрахунок медіани для дискретного ряду даних

Звернімо увагу: оскільки на медіану впливають лише ті значення, що знаходяться у центрі розподілу, цей показник є стійким до впливу «хвостів розподілу», зокрема екстремальних значень (від англ. «outliers»). Ця властивість суттєво відрізняє медіану від середнього арифметичного. Тому за наявності аномальних значень у розподілі медіана, як і відсічене середнє, дозволяє отримати більш об'єктивну оцінку центру розподілу (пригадаємо типовий недолік середнього арифметичного, що ілюструється іронічною приказкою «середнє по палаті»).

Розглянувши принцип визначення медіани для дискретного ряду, перейдемо до випадку з неперервними даними, що згруповані у числові інтервали. У такому разі необхідно спочатку визначити *медіанний інтервал* – тобто той проміжок у розподілі (стовпець), що містить спостереження, чий порядковий номер розділяє

сукупність на дві рівні за кількістю частини, як представлено у таблиці 2.4. У крайньому правому стовпці таблиці розраховано кумулятивну кількість спостережень (накопичувальним підсумком) у поточному інтервалі та всіх попередніх. Наприклад, для рядку з номером $j = 3$ кумулятивну кількість спостережень розраховуємо як **30 = 5** (*к-ть спостережень в інтервалі 1*) + **10** (*к-ть спостережень в інтервалі 2*) + **15** (*к-ть спостережень в інтервалі 3*).

Таблиця 2.4. Задача: визначення медіани та моди для неперервних даних

Номер стовпця, *j*	**Інтервал**	**К-ть спостережень в інтервалі, f_j**	**Кумулятивна к-ть спостережень, S_{fj}**
1	[2; 7)	5	5
2	[7; 12)	10	15
3	[12; 17)	15	30
4	[17; 22)	12	42
5	[22; 27)	8	50
Всього	**X**	**50**	**X**

Якщо у вибірці всього 50 спостережень, то серединні значення мають порядкові номери 25 та 26, що відповідає інтервалу [12; 17), який ми визначаємо як *медіанний*. Далі ми аналітично розраховуємо точку в межах медіанного інтервалу, яку і вважатимемо медіаною. Для цього використаємо формулу (2.8):

$$Me = x_0 + h\frac{0{,}5\sum_{j=1}^{m} f_j - cumf_{Me-1}}{f_{Me}} \qquad (2.8)$$

де Me — *медіана;*

j – *порядковий номер інтервалу;*

x_0 — *нижня межа медіанного інтервалу;*

h — *ширина медіанного інтервалу;*

f_{Me} — *частота, або кількість спостережень в медіанному інтервалі;*

f_j — *частота, або кількість спостережень в інтервалі j;*

$cumf_{Me-1}$ — *сума частот (кумулятивна) усіх інтервалів, що передують медіанному;*

$\sum_{j=1}^{m} f_j$ — *загальна сума частот в усіх інтервалах з 1 до j.*

На рисунку 2.10 схематично зображено розподіл та ключові елементи формули (2.8) для більшої наочності. Інтуїтивно цей розрахунок має такий алгоритм: ми беремо нижню межу інтервалу x_0 і додаємо до неї певну частку (пропорцію) його ширини h. Ця частка визначається тим, скільки ще спостережень нам бракує до середини ряду, зважаючи на частоту самого медіанного інтервалу. Розрахуємо медіану для прикладу у табл. 2.4: *Me* = 12 + 5 * (0,5 * 50 – 15) / 15 = **15,3**.

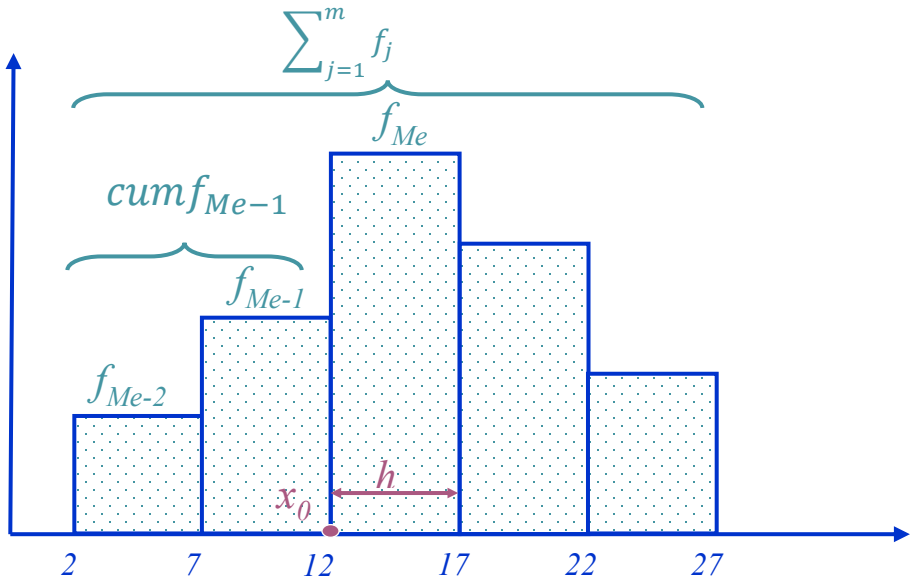


Рис. 2.10. Схематичне зображення підходу до визначення медіани для неперервних даних

Визначення моди для розподілу з неперервними даними розпочинається з пошуку *модального інтервалу* – того, що має найбільшу частоту. На гістограмі розподілу йому відповідає найвищий стовпець. Подальший розрахунок моди здійснюється за формулою (2.9) за принципом, подібним до алгоритму визначення медіани.

$$Mo = x_0 + h\frac{f_{Mo} - f_{Mo-1}}{(f_{Mo} - f_{Mo-1}) + (f_{Mo} - f_{Mo+1})} \qquad (2.9)$$

де Mo — *мода;*

x_0 — *нижня межа модального інтервалу;*

h — *ширина модального інтервалу;*

f_{Mo} — *частота, або кількість спостережень в модальному інтервалі;*

f_{Mo-1} — *частота інтервалу, що передує модальному;*

f_{Mo+1} — *частота інтервалу, наступного за модальним.*

Для кращого розуміння формули (2.9) на гістограмі нижче (див. рис. 2.11) графічно представлено основні елементи, які використовуються в розрахунку. Подібно до медіани, підхід до визначення моди для неперервних даних базується на такій логіці: ми беремо нижню межу модального інтервалу (x_0) і додаємо до неї певну частку його ширини h. Ця частка залежить від співвідношення частот інтервалів, які межують з модальним: чим більше спостережень міститься у попередньому інтервалі (f_{Mo-1}), порівняно з наступним (f_{Mo+1}), тим ближчим є точкове значення моди до нижньої межі x_0, і навпаки. Розрахуємо моду для прикладу у табл. 2.4: *Mo* = 12 + 5 * (15-10) / ((15-10) + (15-12)) = **15,125**.

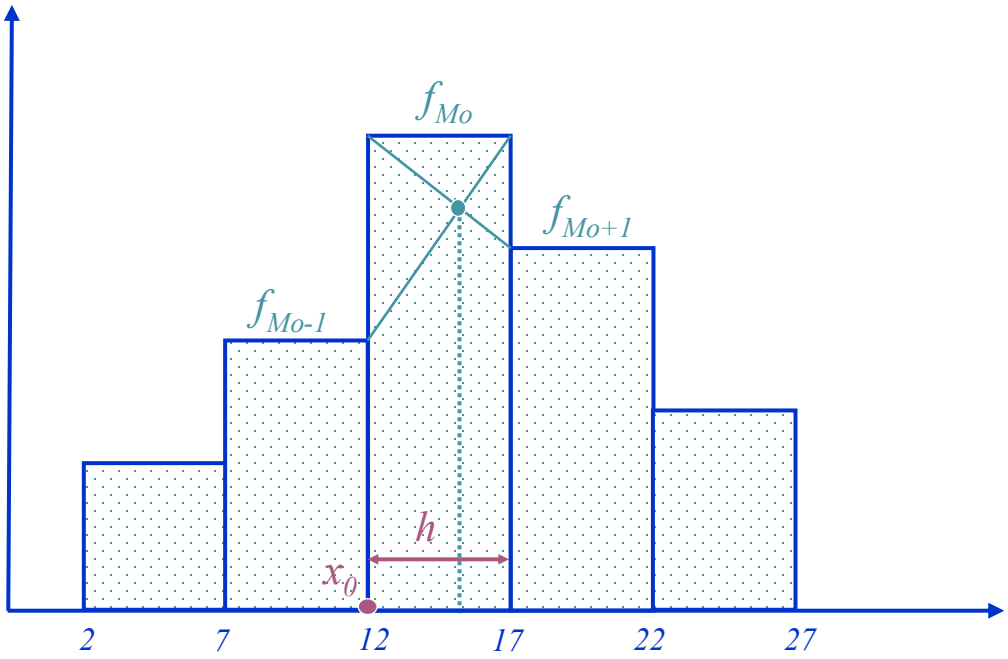


Рис. 2.11. Схематичне зображення визначення моди для неперервних даних

У таблиці 2.5 наведено порівняльну характеристику різних видів статистичних показників, що описують міру центральної тенденції, їхню чутливість до екстремальних значень у вибірці, а також сферу застосування.

Таблиця 2.5. Порівняльна характеристика показників, що описують міру центральної тенденції у статистичному розподілі

Показник	**Зміст**	**Сфера застосування**
Середнє арифметичне	Сума всіх значень, поділена на їхню кількість	▪ Дані з відносно симетричним розподілом відносно центру. ▪ ***Висока чутливість*** до аномальних значень.
Середнє геометричне	Корінь n-го ступеня з добутку значень у форматі (1+%)	▪ Відображення мультиплікативних процесів. ▪ Розрахунок темпів приросту, складних відсотків, дохідності фінансових активів.
Середнє гармонійне	Обернене до середнього арифметичного обернених величин	▪ Аналіз для відносних показників інтенсивності, що виражаються у вигляді «за одиницю»: ціна за одиницю продукції, витрат на 1 квадратний метр, швидкість, продуктивність тощо.
Медіана	Значення, що ділить впорядкований ряд навпіл за кількістю елементів	▪ Найбільш ***стійкий*** показник до екстремальних значень. Підходить для аналізу асиметричних розподілів, де присутні аномальні значення: доходи фірм, ціни на нерухомість, доходи домогосподарств тощо.
Мода	Значення, що зустрічається найчастіше у вибірці даних	▪ Аналіз найпоширеніших значень у сукупності: маркетинговий аналіз попиту, оцінка типового значення розподілу тощо.

Для кращого засвоєння принципових відмінностей між середнім арифметичним та медіаною корисно розглянути їхній геометричний зміст за допомогою візуалізацій, представлених на рис. 2.12:

- *Медіана* – це «серединне» або типове спостереження, яке ділить упорядкований набір елементів навпіл, зважаючи на їхні порядкові номери, а не числові значення. На ілюстрації нижче (рис. 2.12) – це той елемент сукупності (виділений кубик), справа і зліва від якого знаходиться по 24 елементи.
- *Середнє арифметичне* - це точка рівноваги розподілу, або його «центр маси». Якщо уявити елементи розподілу у вигляді кубиків з різною вагою на важелі, то середнім арифметичним значенням буде та точка опори на горизонтальній осі, яка врівноважує кубики, що знаходяться ліворуч і праворуч від неї.

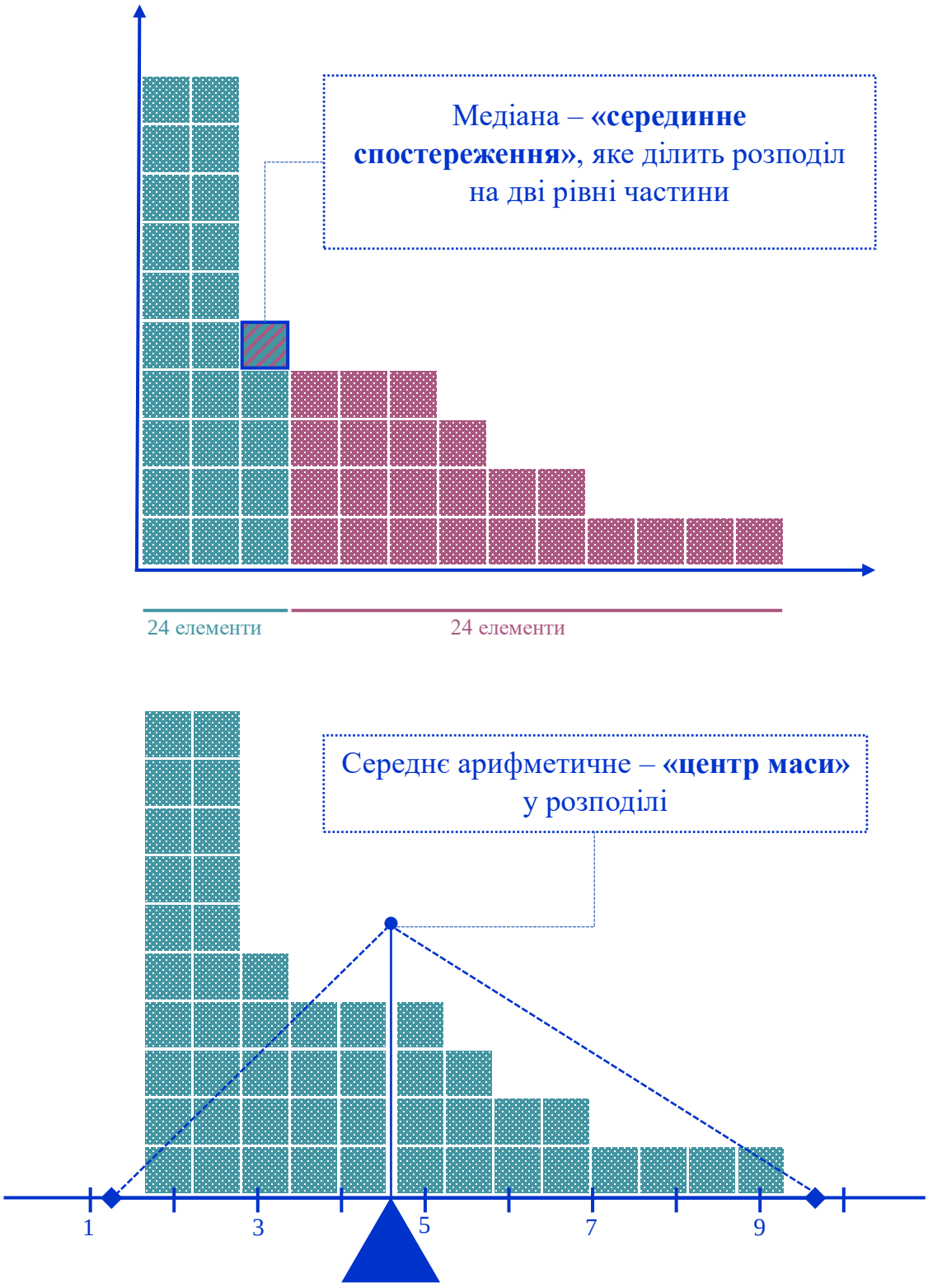


Рис. 2.12. Порівняння геометричного змісту медіани й середнього арифметичного

Квантилі

Медіану можна розглядати не лише як один з показників центральної міри розподілу, а й як значення (елемент сукупності x_j), яке дозволяє розділити сукупність на дві різні частини. Подібні показники називають квантилями. **Квантилі** – це варіанти (значення ознаки x_j), які розподіляють *упорядкований* ряд даних на визначену кількість частин:

- **квартилі** – варіанти, які поділяють обсяг сукупності на *чотири* рівні частини. Наприклад, формула для визначення третього квартиля має такий вигляд:

$$Q_3 = x_0 + h\frac{0{,}75\sum_{j=1}^{m} f_j - cumf_{Q_3-1}}{f_{Q_3}} \qquad \textbf{(2.10)}$$

де Q_3 – *третій квартиль;*

j – *порядковий номер інтервалу;*

x_0 – *нижня межа квартильного інтервалу;*

h – *ширина квартильного інтервалу;*

f_{Q_3} – *частота квартильного інтервалу;*

f_j – *частота, або кількість спостережень в інтервалі j;*

$cumf_{Q_3-1}$ – *сума частот (кумулятивна) усіх інтервалів, що передують квартильному;*

$\sum_{j=1}^{m} f_j$ – *загальна сума частот в усіх інтервалах з 1 до j.*

- **децилі** поділяють обсяг сукупності на *десять* рівних частин;
- **перцентилі** поділяють обсяг сукупності на *сто* рівних частин.

Децилі та перцентилі визначаються за формулою, подібною до інтервальної формули медіани або квартиля, з урахуванням порядкового номера відповідного дециля або перцентиля. Зокрема, у формулі для третього квартиля загальний обсяг сукупності множиться на 0,75; для першого дециля – на 0,1; для 95 перцентиля – на 0,95. Тобто множник відповідає частці сукупності, яку охоплює відповідний квантиль.

Питання для самоперевірки

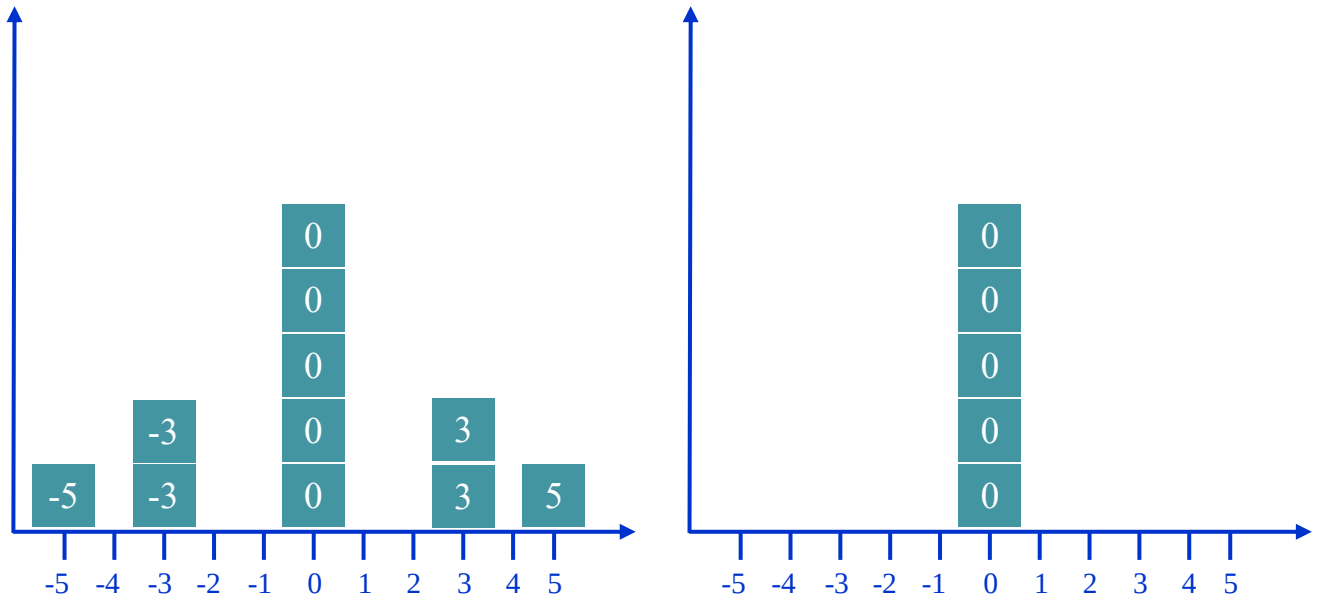
1. *У чому полягає різниця між простим та зваженим середнім арифметичним?*
2. *Чому середнє арифметичне вважається чутливим до екстремальних значень (викидів)?*
3. *У яких випадках доцільно використовувати «відсічене середнє»?*
4. *Чому для аналізу темпів приросту або дохідності фінансових активів використовують середнє геометричне, а не арифметичне?*
5. *Дайте визначення медіані та поясніть, як вона розраховується для парної та непарної кількості елементів.*
6. *Чому медіана вважається більш стійким показником до аномальних значень, ніж середнє арифметичне?*
7. *Поясніть геометричний зміст медіани як «серединного спостереження» та середнього арифметичного як «центру маси».*

2.2. Показники варіації розподілу

Ознайомившись із сутністю міри центральної тенденції як характеристики розподілу, підходами до її визначення за допомогою різних показників, а також ключовими відмінностями між середнім арифметичним, геометричним, гармонійним, медіаною та модою, може виникнути закономірне запитання: чи достатньо цих показників для повного опису певного ряду даних або розподілу?

Для того, щоб це дізнатися, дослідимо два розподіли, що представлені на рис.2.13. Розрахуємо середнє значення для першого графіку, скориставшися формулою середнього арифметичного зваженого: (-5*1) + (-3*2) + (0*5) + (3*2) + (5*1) = **0**. Аналогічно для розподілу на другому графіку маємо: 0 * 5 = **0.**

Попри однакове середнє значення обох розподілів, вони суттєво відрізняються. Постає питання: яка додаткова характеристика дає змогу розрізнити розподіли, окрім середнього значення?



(а) **(б)**

Рис. 2.13. Приклад розподілів з однаковим середнім, але різною варіацією

Для відповіді на це питання розглянемо сутність поняття *«варіації»*, тобто ступеня розсіювання індивідуальних значень ознаки x_j (одиниць сукупності) відносно центру розподілу.

В одних сукупностях індивідуальні значення ознаки щільно групуються навколо середнього значення, в інших – суттєво відхиляються, як продемонстровано на попередньому графіку (див. рис.2.13). Чим менше розсіювання, тим однорідніше є досліджувана сукупність і тим надійніші характеристики центру розподілу. Наприклад, якщо бали у групі студентів, заробітна плата співробітників у компанії, ВВП на душу населення з-поміж групи країн суттєво відрізняються, коливаючись в рамках широкого діапазону, то складно визначити «типовий» рівень успішності або рівня доходу внаслідок суттєвої варіації.

Отже, як і міра центральної тенденції, ступінь варіації є однією із ключових характеристик розподілу.

Для кількісного визначення ступеню розсіювання даних у розподілі розрізняють такі основні статистичні показники:

Абсолютні:

- Варіаційний розмах, R
- Середнє лінійне відхилення, $\bar{l}$
- Середнє квадратичне відхилення, σ
- Дисперсія, σ^2

Відносні:

- Лінійний коефіцієнт варіації, $V_{\bar{l}}$
- Квадратичий коефіцієнт варіації, $V_{\bar{l}}$
- Коефіцієнт осциляції, V_R
- Квартильний коефіцієнт варіації V_Q

Варіаційний розмах – це найпростіший показник варіації, що визначається як різниця між максимальним і мінімальним значеннями ознаки, використовуючи формулу:

$$R = x_{max} - x_{min} \qquad (2.11)$$

де R – *розмах;*

x_{max} – *максимальне значення ознаки в розподілі;*

x_{min} – *мінімальне значення ознаки в розподілі.*

Інакше кажучи, розмах відображає ширину усього діапазону, в межах якого знаходиться значення досліджуваної вибірки. Розмах для лівого графіку на рис. 2.13 становить: 5 - (-5) = **10**. Для правого графіка розмах, відповідно, дорівнює нулю.

Окрім загального варіаційного розмаху для усієї сукупності також виділяють його підвиди – квартильний і децильний розмахи згідно з такими формулами:

$$R_Q = Q_3 - Q_1 \qquad (2.12)$$

де R_Q – *квартильний розмах;*

Q_1 – *перший квартиль;*

Q_3 – *третій квартиль.*

$$R_D = D_8 - D_2 \qquad (2.13)$$

де R_D– *децильний розмах;*

D_2 – *другий дециль;*

D_8 – *восьмий дециль.*

Попри просту формулу та інтуїтивну зрозумілість, розмах має суттєвий недолік: він враховує лише два значення, зокрема найбільше й найменше, ігноруючи решту елементів сукупності. Наприклад, на рис. 2.14 зображено розподіли з однаковим розмахом, хоча візуально помітно, що цей показник не

описує варіацію вичерпно. Тому виникає потреба в додаткових показниках, які враховували б не лише крайні значення розподілу, а й усі інші.

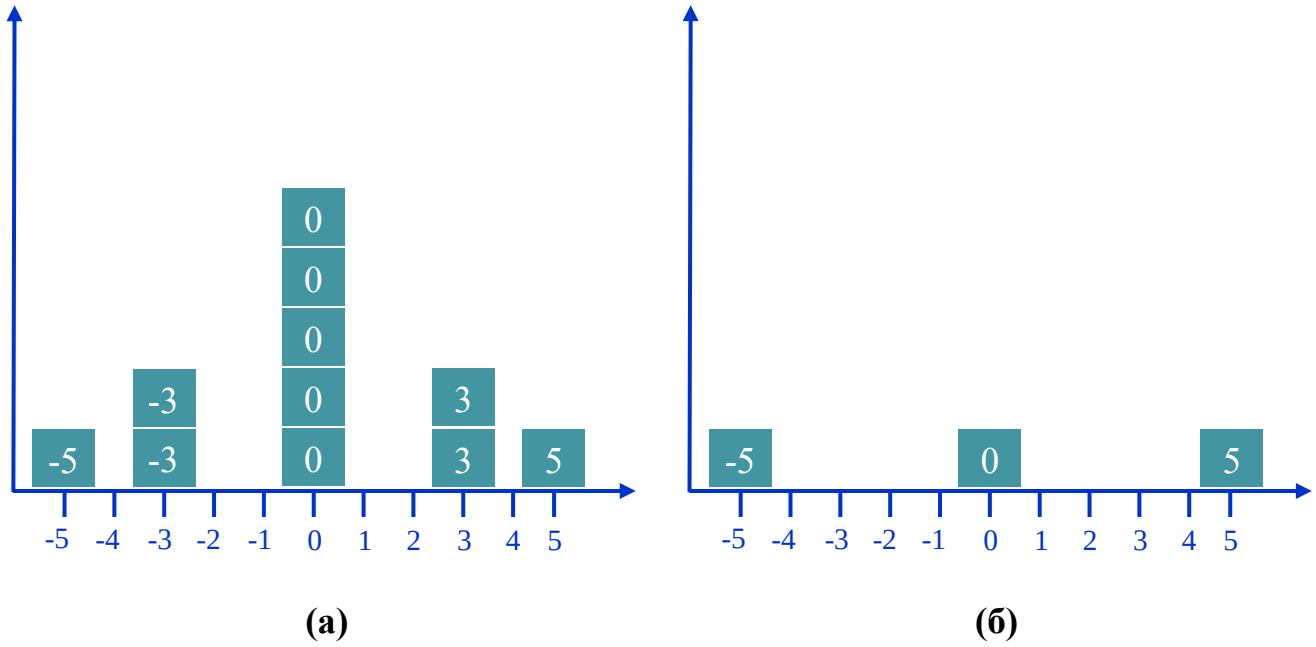


Рис. 2.14. Приклад різних розподілів з однаковим розмахом даних

Наступний показник варіації – це середнє лінійне відхилення, яке, на відміну від розмаху, враховує кожне значення сукупності при визначенні ступеня розсіювання. За геометричним змістом (див. рис. 2.15) **середнє лінійне відхилення** показує, наскільки далеко в середньому індивідуальні значення досліджуваної ознаки (x_j) знаходяться від центру розподілу ($\bar{x}$). Згідно з формулою, відстань кожного елемента сукупності від центру береться за модулем і усереднюється на основі простого (формула (2.14)) або зваженого (формула (2.15)) середнього. Середнє лінійне відхилення вимірюється у тих самих одиницях, що і значення ознаки x_j.

$$\bar{l} = \frac{\sum_{j=1}^{n} |x_j - \bar{x}|}{n} \qquad (2.14)$$

$$\bar{l} = \frac{\sum_{j=1}^{n} |x_j - \bar{x}| f_j}{\sum_{j=1}^{n} f_j} \qquad (2.15)$$

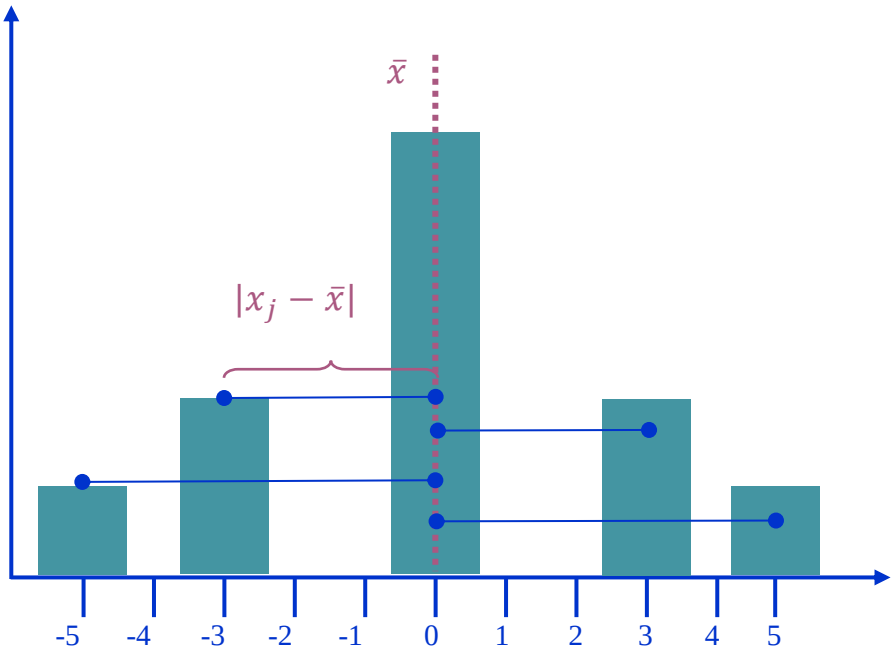


Рис. 2.15. Ілюстрація геометричного змісту середнього лінійного відхилення

Середнє квадратичне (стандартне) відхилення за принципом розрахунку та геометричним змістом є подібним до середнього лінійного відхилення, однак вимірює ступінь розсіювання (відстань) за іншою методикою: замість модуля відхилення зводяться до квадрата, а наприкінці з усього виразу добувається квадратний корінь. Як і середнє лінійне, середнє квадратичне відхилення вимірюється у тих самих одиницях, що і значення ознаки x_j. Завдяки вищій чутливості до розсіювання середнє квадратичне відхилення за математичними властивостями завжди перевищує середнє лінійне для того самого набору даних: $\sigma > \bar{l}$.

Нижче наведено формулу (2.16) для розрахунку цього коефіцієнту на основі середнього зваженого.

$$\sigma = \sqrt{\frac{\sum_{j=1}^{n} (x_j - \bar{x})^2 f_j}{\sum_{j=1}^{n} f_j}} \qquad (2.16)$$

Дисперсія – це показник варіації, що визначається як середнє значення квадратів відхилень індивідуальних значень ознаки від центру розподілу. Фактично дисперсія є середнім квадратичним (стандартним) відхиленням у квадраті, тому вимірюється у квадратних одиницях ознаки.

$$\sigma^2 = \frac{\sum_{j=1}^{n} (x_j - \bar{x})^2 f_j}{\sum_{j=1}^{n} f_j} \qquad (2.17)$$

За геометричним змістом дисперсію можна уявити як середній квадрат відстані від індивідуальних значень ознаки до центру розподілу ($\bar{x}$), зважений на частоту (вагу) відповідної ознаки (див. рисунок 2.16). Тобто на першому етапі визначаються квадрати відстаней, далі їх множать на висоту відповідного стовпця гістограми та знаходять середнє значення отриманих добутків.

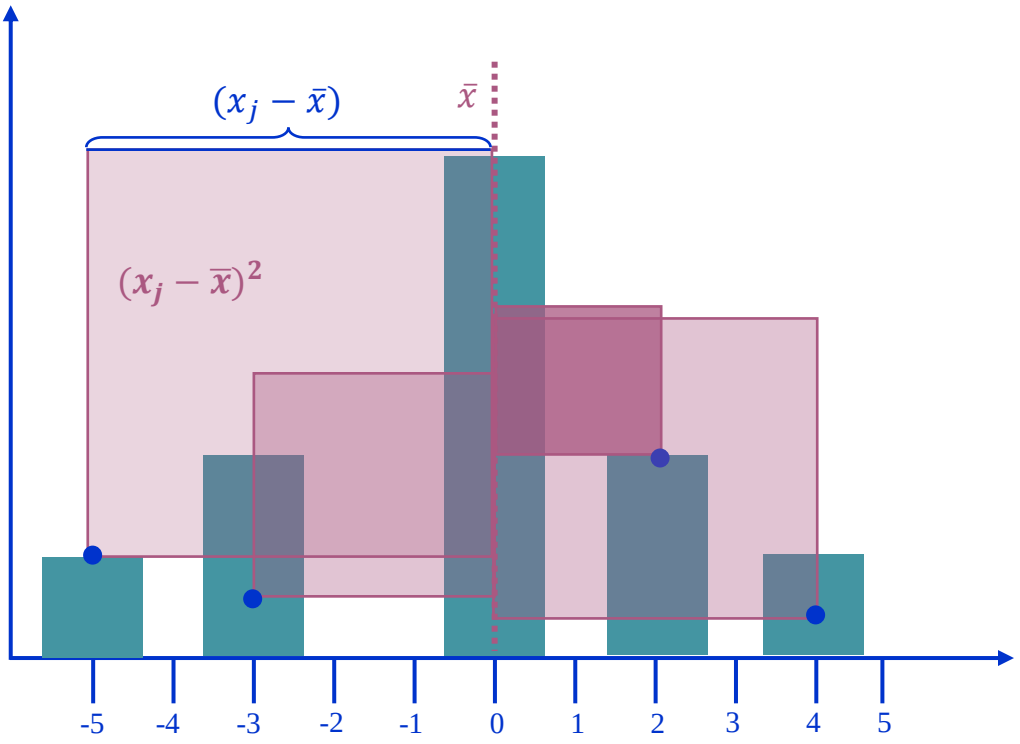


Рис. 2.16. Ілюстрація геометричного змісту дисперсії

Дисперсія і середнє квадратичне відхилення є одними з найпоширеніших показників варіації у практиці аналізу даних та моделювання. Розглянемо деякі математичні властивості дисперсії:

- Якщо всі значення варіанти x_j зменшити на сталу величину A, тобто перенести графік розподілу паралельно вліво на A одиниць, то дисперсія не зміниться: $\sigma^2_{x-A} = \sigma^2_x$
- Якщо всі значення варіант x_j збільшити або зменшити в A разів, тобто розтягнути графік розподілу відносно його центру, то дисперсія зміниться в A^2 разів: $\sigma^2_{x*A} = \sigma^2_x * A^2$
- Якщо частоти f_j замінити на ваги (відносні частоти) w_j, де $w_j = f_j / \sum f_j$, то дисперсія не зміниться

Правило декомпозиції (розкладання) дисперсії

Перш ніж перейти до декомпозиції дисперсії та її видів, проілюструємо навчальну задачу на рис.2.17, для кращого розуміння її сутності, а також визначимо ключові позначення:

- на рисунку зображено сукупність елементів (кружечки зелено-синього кольору), які утворюють три кластери (виділені синім пунктиром);
- кожен елемент сукупності позначено як $x_j^{[i]}$, де [i] – номер кластеру (групи), j – порядковий номер елемента сукупності; $n^{[i]}$ – кількість елементів у групі [i]; наприклад, $\bar{x}^{[1]}$ – це середнє значення елементів першої групи, а $\bar{x}$ – це загальне середнє серед усіх елементів трьох груп; n – загальна кількість елементів у сукупності. Відповідні середні значення позначено на графіку рожево-фіолетовим кольором;
- прикладом подібної задачі на практиці можуть слугувати різні відділи в компанії або студентські групи, а ознакою – рівень заробітної плати або навчальний бал відповідно.

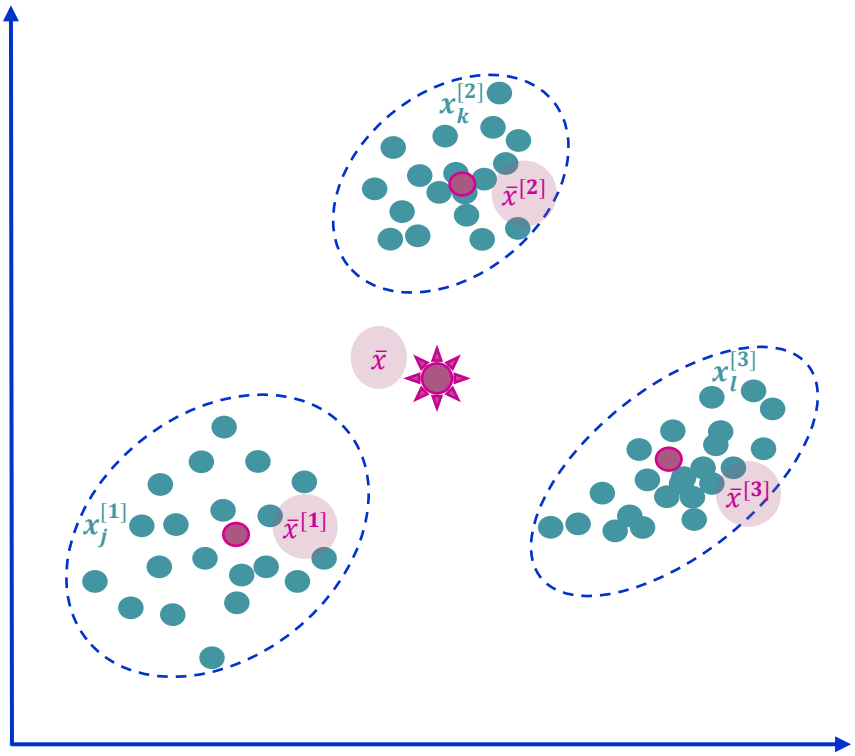


Рис. 2.17. Ілюстрація основних умов задачі на визначення дисперсії

Зауважимо, що ступінь розсіювання можна оцінити як для кожної групи окремо, так і для всієї сукупності загалом, об'єднавши всі елементи в єдину групу. Залежно від цього можемо виокремити такі види дисперсії, як схематично зображено на рис.2.18 (зверніть увагу на кольорові позначення для наочності):

- **загальна дисперсія** $\sigma^{2^{[*]}}$ – показник, що характеризує варіацію ознаки x_j навколо загального середнього значення $\bar{x}$:

$$\sigma^2 = \frac{\sum_{j=1}^{n} (x_j - \bar{x})^2}{n} = \overline{x^2} - \bar{x}^2 \qquad (2.18)$$

- **внутрішньогрупова дисперсія** $\sigma^{2^{[i]}}$ – показник, що характеризує варіацію ознаки x_j навколо групового середнього значення $\bar{x}^{[i]}$ для групи [i]. Оскільки в групи об'єднуються схожі елементи сукупності, то варіація всередині груп, як правило, менша, ніж загалом по сукупності:

$$\sigma^{2[i]} = \frac{\sum_{j=1}^{n} \left(x_j^{[i]} - \bar{x}^{[i]}\right)^2}{n^{[i]}} \qquad (2.19)$$

- **середня з внутрішньогрупових дисперсій** – узагальнювальна міра внутрішньогрупової варіації серед m груп, що визначається як зважене середнє арифметичне внутрішньогрупових дисперсій:

$$\overline{\sigma^2} = \frac{\sum_{j=1}^{n} \sigma^{2[i]} * n^{[i]}}{\sum_{i=1}^{m} n^{[i]}} \qquad (2.20)$$

- **міжгрупова дисперсія** δ^2 – показник варіації групових середніх $\bar{x}^{[i]}$ навколо загального середнього $\bar{x}$. Квадрати відхилень групових середніх від загального середнього зважуються на кількість елементів у кожній групі $n^{[i]}$, як представлено у формулі:

$$\delta^2 = \frac{\sum_{j=1}^{n} \left(\bar{x}^{[i]} - \bar{x}\right)^2 * n^{[i]}}{\sum_{i=1}^{m} n^{[i]}} \qquad (2.21)$$

Правило декомпозиції дисперсії – загальну дисперсію можна розкласти як суму середньої з внутрішньогрупових дисперсій та міжгрупової дисперсії за формулою.

$$\sigma^{2[*]} = \overline{\sigma^2} + \delta^2 \qquad (2.22)$$

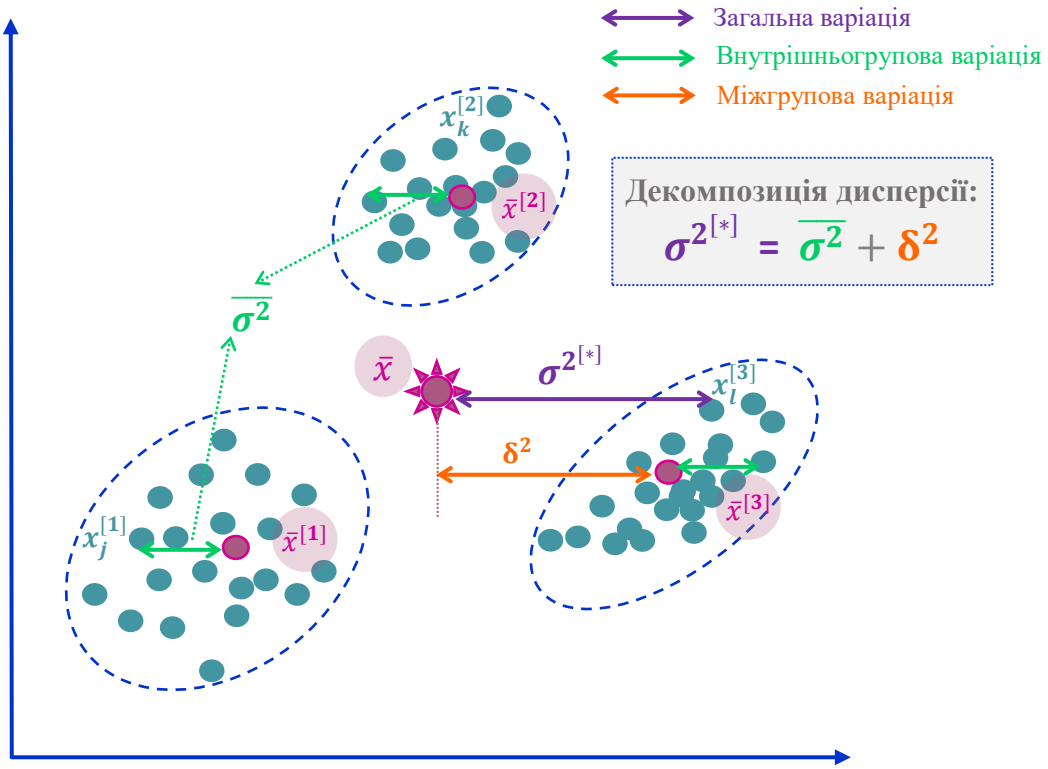


Рис. 2.18. Ілюстрація правила декомпозиція дисперсії

Для оцінювання варіації та порівняння розподілів за ступенем розсіювання також застосовують графічний аналіз на основі коробкового графіку (від англ. «boxplot»), зображеного на рисунку 2.19, де 50% елементів сукупності знаходяться в межах «коробки», а більшість елементів сукупності, окрім аномальних, – у межах «хвостів/вусів». Коробковий графік також дає змогу виявити **аномальні (або екстремальні)** значення у розподілі (від англ. «*outliers*») – тобто ті значення ознаки x_j, які перевищують третій квартиль більш ніж на півтора квартильного розмаху ($1{,}5*R_Q$), або є меншими за перший квартиль більш ніж на півтора квартильного розмаху. На практиці такі елементи сукупності додатково вивчають, перевіряються і нерідко виключаються з вибірки під час подальшого дослідження й моделювання.

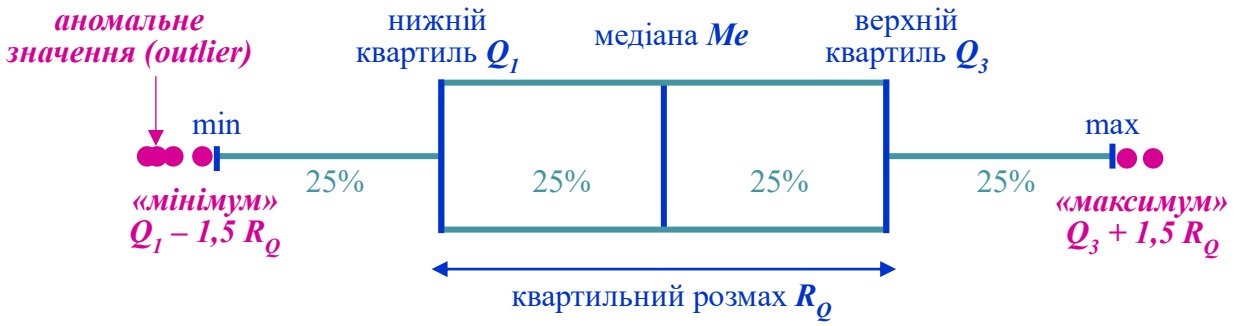


Рис. 2.19. Схема коробкової діаграми («boxplot») для аналізу розподілу даних

Окрім абсолютних показників варіації, зокрема варіаційного розмаху, середнього лінійного та квадратичного відхилення, дисперсії, часто застосовують відносні показники варіації. Це особливо доречно при порівнянні ступеня варіації у двох або більше розподілах, де значення ознаки суттєво відрізняються внаслідок різної розмірності даних.

Відносні показники варіації вимірюють ступінь розсіювання не в одиницях досліджуваної ознаки (наприклад, гривнях, кілометрах, кілограмах, кіловатах тощо), а у відсотках, що дозволяє порівнювати різні ряди даних. Типовими відносними показниками варіації є лінійний коефіцієнт варіації ($V_{\bar{l}}$); квадратичний коефіцієнт варіації (V_σ); коефіцієнт осциляції (V_R) та квартильний коефіцієнт варіації (V_Q), згідно з формулами (2.23):

$$V_{\bar{l}} = \frac{\bar{l}}{\bar{x}} * 100\,; \quad V_\sigma = \frac{\sigma}{\bar{x}} * 100\,; \quad V_R = \frac{R}{\bar{x}} * 100\,; \quad V_Q = \frac{Q_3 - Q_1}{2\,Me} * 100; \qquad \textbf{(2.23)}$$

Як правило, відносні показники варіації розраховуються за спільним принципом: у знаменнику береться середнє значення, а в чисельнику – відповідний абсолютний показник варіації. Таким чином, коефіцієнт показує, яку частку (у %) від середнього становить відповідний показник варіації.

Іноді абсолютні показники варіації (наприклад, стандартне відхилення) не дозволяють адекватно порівняти два ряди даних. Розглянемо приклад визначення абсолютних та відносних показників варіації (дані наведено у табл. 2.6).

Припустимо, що було зібрано дані річних доходів для двох вибірок – локальних магазинів і великих супермаркетів, для кожної яких розраховано середні доходи та показники варіації. З'ясуємо, у якій групі торгових точок середньорічні доходи коливаються сильніше, тобто який тип бізнесу є більш ризиковий з точки зору стабільності доходів:

- якщо орієнтуватися на *середнє квадратичне відхилення*, може здатися, що у вибірці супермаркетів доходи були більш волатильними. Однак ця відповідь ігнорує факт, що у таких великих точок як супермаркети доходи є природньо вищими, ніж у маленьких локальних магазинів. Це закономірно впливає і на абсолютний показник варіації;
- якщо розрахувати *квадратичний коефіцієнт варіації*, то зможемо коректніше порівняти різні за розмірністю вибірки і відповісти на дослідницьке питання. Для локальних магазинів коефіцієнт становитиме 750 / 6 000 * 100 = **12,5%** . Для супермаркетів: 1 000 / 60 000 = **1,7%**. Отже, ми визначали, що в досліджуваних вибірках доходи локальних магазинів мали відносно більшу варіацію, ніж доходи супермаркетів.

Таблиця 2.6. Задача: визначення абсолютних і відносних показників варіації

Показник	**Позначення**	**Локальний магазин**	**Великий супермаркет**
Середній дохід, тис. грн за рік	$\bar{x}$	6 000	60 000
Середнє квадратичне відхилення, тис. грн	σ	750	1 000
Квадратичний коефіцієнт варіації	V_σ	750 / 6 000 = 12,5%	1 000 / 6 000 = 1,7%

Для практичного засвоєння матеріалу на рис.2.20 наведено код Python для розрахунку основних показників варіації – абсолютних і відносних.

Код у Python

```
import numpy as np

# Вибірка даних
data = np.array([20.43, 5.42, 17.67, 19.98, 14.97, 9.5, 8.96, 20.21,
                 8.38, 6.77, 18.71, 24.07, 5.08, 15.24, 21.25, 17.25,
                 19.44, 10.84, 23.36, 19.29])


# Абсолютні показники варіації
R = np.max(data) - np.min(data) # варіаційний розмах
l = np.mean( np.abs(data - np.mean(data)) ) # лінійний коефіцієнт варіації
sigma = np.std(data) # стандартне відхилення
variance = np.var(data) # дисперсія

# Відносні показники варіації
V_r = ( R / np.mean(data) ) * 100 # коефіцієнт осциляції
V_l  = ( l / np.mean(data) ) * 100 # лінійний коефіцієнт варіації
V_sigma = (sigma / np.mean(data) ) * 100 # квадратичний коефіцієнт варіації

print("Абсолютні: ", "\n", R, "\n", l, "\n", sigma, "\n", variance)
print("Відносні: ","\n", V_r, "\n", V_l, "\n", V_sigma)

### Результати:
# Абсолютні показники варіації: 18.996, 5.2909, 6.001, 36.0
# Відносні показники варіації: 123.786, 34.489, 39.117
```

Рис. 2.20. Код у Python: розрахунок показників варіації

Розглянуті показники, зокрема міра центральної тенденції і ступінь варіації, є ключовими параметри для характеристики розподілу або будь-якого набору даних і широко використовуються у фінансовому й економічному аналізі. Разом з тим лише кількісних статистичних показників може бути недостатньо для формування надійних висновків: покладання лише на середнє значення та дисперсію може привести до хибних результатів. Щоб уникнути таких ситуацій, кількісний аналіз доцільно доповнювати графічним.

Важливість графічного аналізу переконливо продемонстрував британський статистик Френсіс Анскомб, опублікувавши у 1973 році чотири набори даних, відомі як «Квартет Анскомба» (англ. Anscombe's Quartet), з майже ідентичними описовими статистиками: середнім значенням, дисперсією та кореляцією. Попри схожість кількісних характеристик, ці набори кардинально відрізняються при

графічному відображенні на діаграмі розсіювання, зокрема внаслідок впливу аномальних значень. Графічний аналіз дає змогу дослідникові не лише краще зрозуміти структуру даних і взаємозв’язки між ними, а й сформулювати початкові гіпотези та наочно представити результати аналізу своїй цільовій аудиторії.

Розрахувавши ключові показники розподілу для кожної з чотирьох груп (див. табл.2.7), можна помітити, що вони мають однакові середні значення, однакову дисперсію та дуже близькі значення кореляції. На перший погляд, без графічного аналізу може здатися, що всі чотири групи є статистично подібними, як відображено в таблиці 2.7.

Однак, як видно з рис. 2.21, кожен набір даних «Квартету Анскомба» має відмінну структуру: перший демонструє помірний лінійний зв'язок між *x* та *y*; другий — нелінійний; третій — щільний лінійний; четвертий — майже повну відсутність залежності *x* від *y*.

Таблиця 2.7. Описова статистика для наборів даних «Квартету Анскомба»

Набір даних	**Середнє арифметичне для *x***	**Дисперсія для *x***	**Кореляція між *x* та *y***
I	9,0	11,0	0,8164
II	9,0	11,0	0,8162
III	9,0	11,0	0,8163
IV	9,0	11,0	0,8165

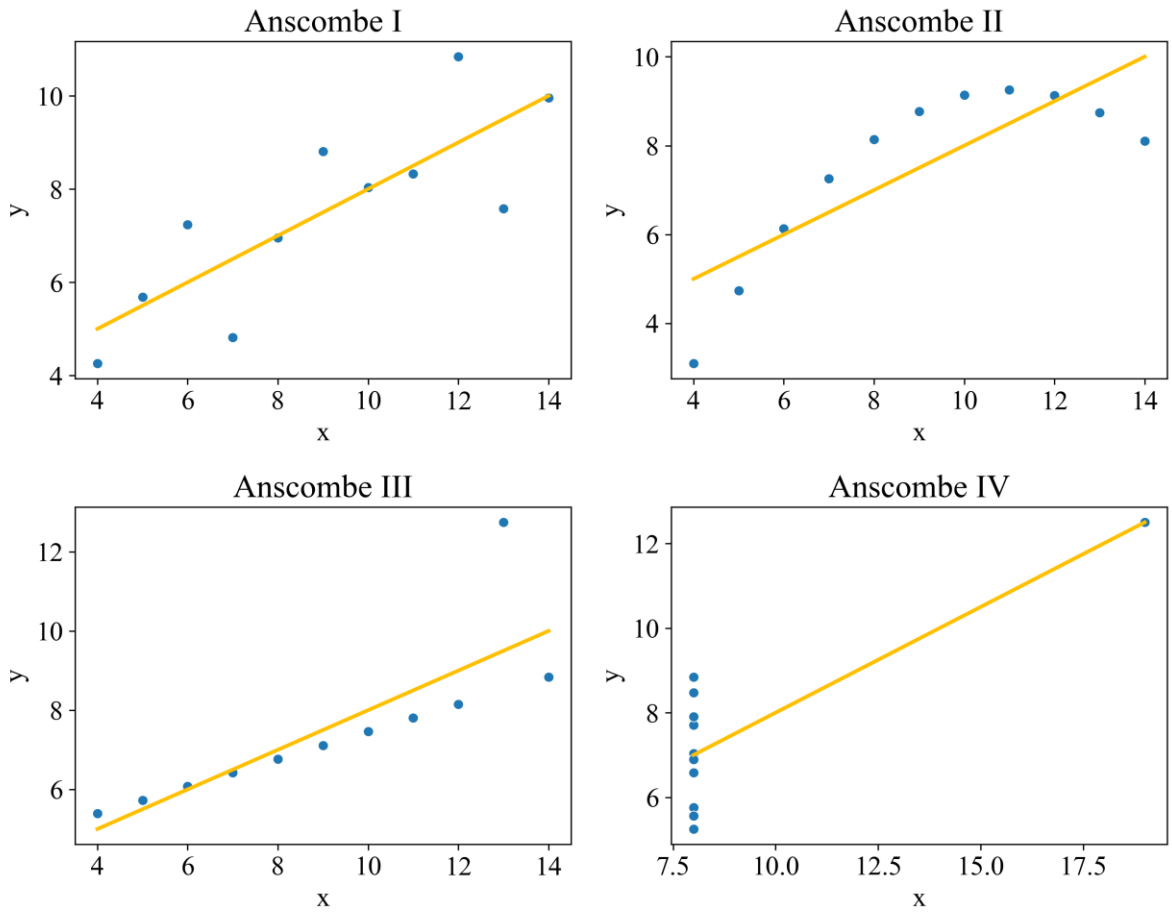


Рис. 2.21. Діаграми розсіювання для наборів даних з «Квартету Анскомба»

Для кращого опрацювання й глибшого розуміння описаного вище прикладу читач має змогу відтворити наведені розрахунки та побудувати відповідні графіки самостійно за допомогою коду Python, поданого на рисунках 2.22 - 2.23.

Код у Python

```
# Завантаження бібліотек
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np

# Завантаження даних
df = sns.load_dataset("anscombe")
df.head()

# Розрахунок показників для кожної групи
df.groupby(by='dataset').mean() # середнє арифметичне
df.groupby(by='dataset').std() # стандартне відхилення
df.groupby(by='dataset').var() # дисперсія
df.groupby(by='dataset').corr() # кореляція
```

Рис. 2.22. Код у Python: описова статистика для Квартету Анскомба

Код у Python

```
## Побудова графіків
fig, axes = plt.subplots(2, 2, figsize=(10, 8))
axes = axes.flatten()

for i, group in enumerate(df['dataset'].unique()):
    subset = df[df['dataset'] == group]
    sns.scatterplot(x="x", y="y", data=subset, ax=axes[i])
    sns.regplot(x="x", y="y", data=subset, ax=axes[i], scatter=False,
                color="#FFC000", ci = False)
    axes[i].set_title(f"Anscombe {group}")

plt.tight_layout()
plt.show()
```

Рис. 2.23. Код у Python: графіки розсіювання для Квартету Анскомба

Питання для самоперевірки

1. *Що таке квартилі, децилі та перцентилі, і на скільки частин вони ділять сукупність?*
2. *Чому середнього значення недостатньо для повного опису розподілу (на прикладі сукупностей з однаковим середнім, але різною варіацією)?*
3. *Назвіть основні абсолютні показники варіації та поясніть недолік варіаційного розмаху.*
4. *У чому полягає різниця між середнім квадратичним відхиленням та дисперсією (зокрема в одиницях вимірювання)?*
5. *Сформулюйте правило декомпозиції дисперсії: з яких компонентів складається загальна дисперсія?*
6. *Яку інформацію про розподіл надає коробковий графік (boxplot) і як за його допомогою виявити аномальні значення?*
7. *Для чого використовують відносні показники варіації (наприклад, коефіцієнт варіації Vσ)?*

2.3. Оцінювання форми та ступеня однорідності розподілу

У попередній темі ми з'ясували, що базові описові статистики є необхідними, але недостатніми інструментами аналізу даних і мають бути доповнені графічним аналізом. Поглибити аналіз дають змогу також інші характеристики розподілу – показники однорідності, асиметрії, ексцесу та рівномірності, які розглянуто в цій темі.

Критерій однорідності

В **однорідних** сукупностях елементи мають спільні властивості і належать до одного класу. Такі сукупності характеризуються одновершинними розподілами – одномодальні розподіли, які мають одну моду. Приклад однорідного розподілу представлено на рисунку 2.24.

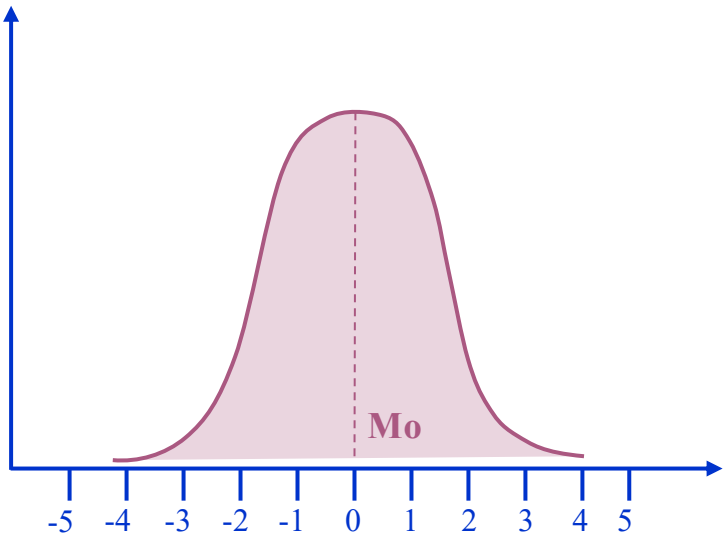


Рис. 2.24. Приклад однорідного розподілу з однією вершиною (модою)

На відміну від однорідних, **неоднорідні сукупності** мають кілька вершин, тому їх ще називають бімодальними (дві вершини), тримодальними (три вершини) тощо розподілами, як показано на рисунку 2.25. Перш ніж розпочинати аналіз, доцільно перегрупувати дані, розділивши неоднорідну багатомодальну сукупність на декілька груп з подібними характеристиками.

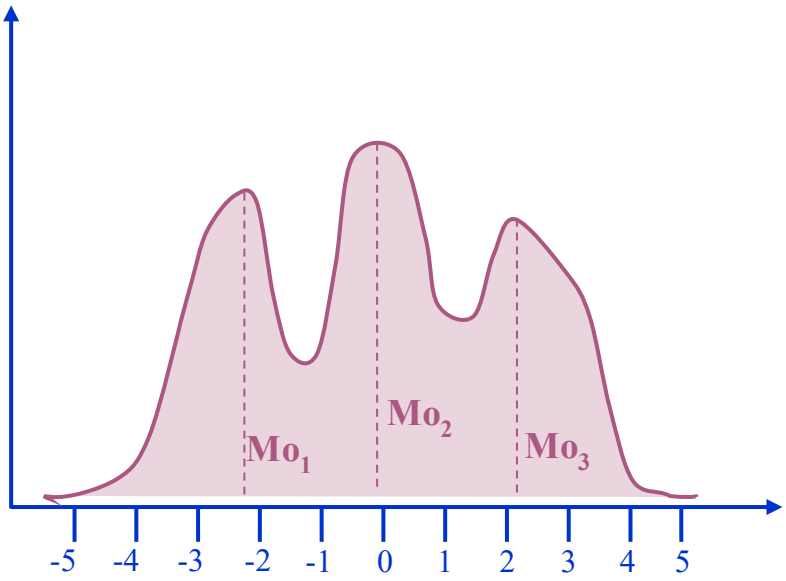


Рис. 2.25. Приклад неоднорідного розподілу з трьома вершинами (модами)

Однак навіть одномодальний розподіл може бути неоднорідним унаслідок значного ступеня розсіювання. Окрім власне графічного аналізу, ступінь однорідності можна виміряти кількісно, використовуючи квадратичний коефіцієнт варіації V_σ. Загальноприйнятим орієнтиром однорідності на практиці є порогове значення 33%. Якщо квадратичний коефіцієнт не перевищує приблизно 33%, то сукупність вважається однорідною, а середнє значення надійно відображає типовий рівень ознаки.

Ступінь симетричності (англ. «skewness»)

Одновершинні розподіли можна також охарактеризувати за ступенем симетричності. У симетричному розподілі середнє значення, мода та медіана збігаються ($\bar{x} = Mo = Me$), як показано на рисунку 2.26.

Чим вищий ступінь асиметрії, тим більшим є відхилення між модою – найвищою точкою розподілу – і середнім значенням: $\bar{x} - Mo$.Тому найпростішою мірою асиметрії може слугувати такий показник:

$$\frac{(\bar{x} - Mo)}{\sigma} \qquad (2.24)$$

За геометричним змістом цей індикатор відображає відстань між модою і середнім значенням, виражену в одиницях середнього квадратичного (стандартного) відхилення. Така «стандартизація» (ділення відстані від центру на стандартне відхилення) дає змогу порівнювати асиметрію різних розподілів незалежно від одиниць вимірювання ознаки.

Хоча найпростіша міра асиметрії дозволяє інтуїтивно зрозуміти сутність цього показника, для кількісного оцінювання ступеня **асиметрії (*As*)** на практиці використовують таку формулу:

$$As = \frac{1}{n} \frac{\sum_{j=1}^{n}(x_j - \bar{x})^3}{\sigma^3} \qquad (2.25)$$

У формулі показника асиметрії подібним чином вимірюється відстань між кожним значенням ознаки та середнім. Далі ця відстань усереднюється ($1/n$) і стандартизується ($1/\sigma^3$). Також цю формулу можна записати для випадку, коли за основу береться метод середнього *зваженого*:

$$As = \frac{\sum_{j=1}^{n}(x_j - \bar{x})^3 * f_j}{\sigma^3 \sum_{i=1}^{n} f_j} \qquad (2.26)$$

Схематичне зображення прикладів симетричного і асиметричних розподілів наведено на рис.2.26.

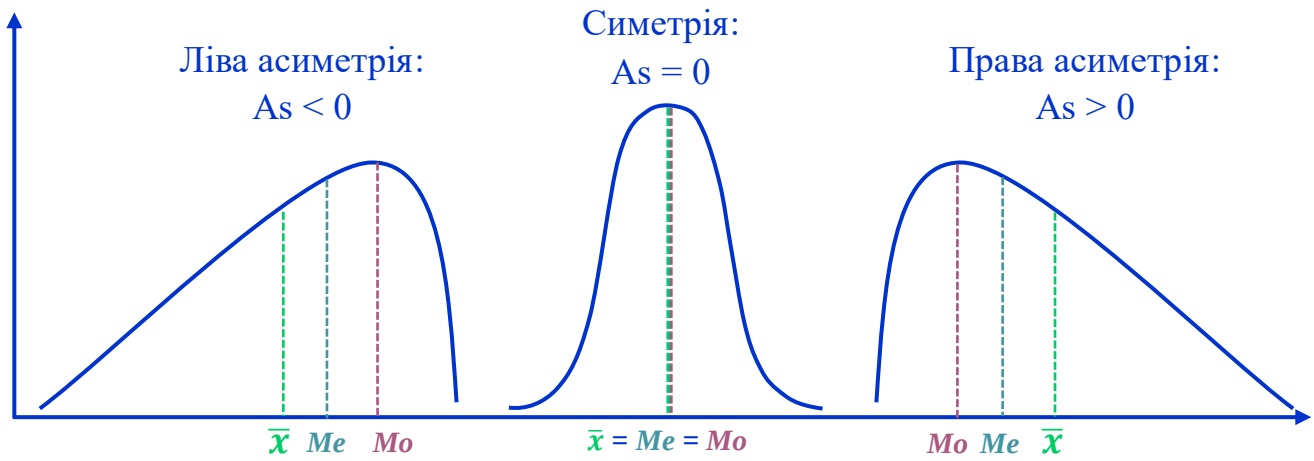


Рис. 2.26. Схематичне зображення симетричного і асиметричного розподілів

Відповідно до формального визначення, **симетричним** є той розподіл, в якому ознаки (x_j), що рівновіддалені від його центру (вершини), мають однакові частоти (f_j).

Відповідно, асиметричним є той розподіл, коли ця умова не виконується, і вершина є зміщеною. Серед асиметричних розподілів розрізняють:

- ***Лівосторонню*** асиметрію (при коефіцієнті асиметрії As < 0): вершина зміщена праворуч, довгий ***«хвіст» зліва***.
- ***Правосторонню*** асиметрію (при коефіцієнті асиметрії As > 0): вершина зміщена ліворуч, довгий ***«хвіст» справа***.

Ступінь ексцесу (англ. «kurtosis»)

Ексцес (*Es*) – ступінь сконцентрованості елементів сукупності навколо центру розподілу. На практиці ступінь ексцесу заданого розподілу порівнюють з нормальним, у якого цей показник становить 3. З урахуванням цього **надмірний ексцес (*Es**)** визначається за формулою (2.27).

$$Es^* = \frac{\sum_{j=1}^{n}(x_j - \bar{x})^4 * f_j}{\sigma^4 \sum_{j=1}^{n} f_j} - 3 \qquad (2.27)$$

Зверніть увагу, що у формулі *ексцесу*, на відміну від *надмірного ексцесу*, не потрібно віднімати 3.

На рис. 2.27 наочно показано розподіли з різним ступенем надмірного ексцесу:

- Якщо надмірний ексцес є ***додатнім*** (тобто показник ексцесу перевищує 3: Es > 3), то розподіл є порівняно ***гостровершинним,*** а хвости є вужчими, ніж у нормального.
- Якщо надмірний ексцес є ***від'ємним,*** то розподіл є ***плосковершинним*** з ширшими хвостами, порівняно з нормальним.

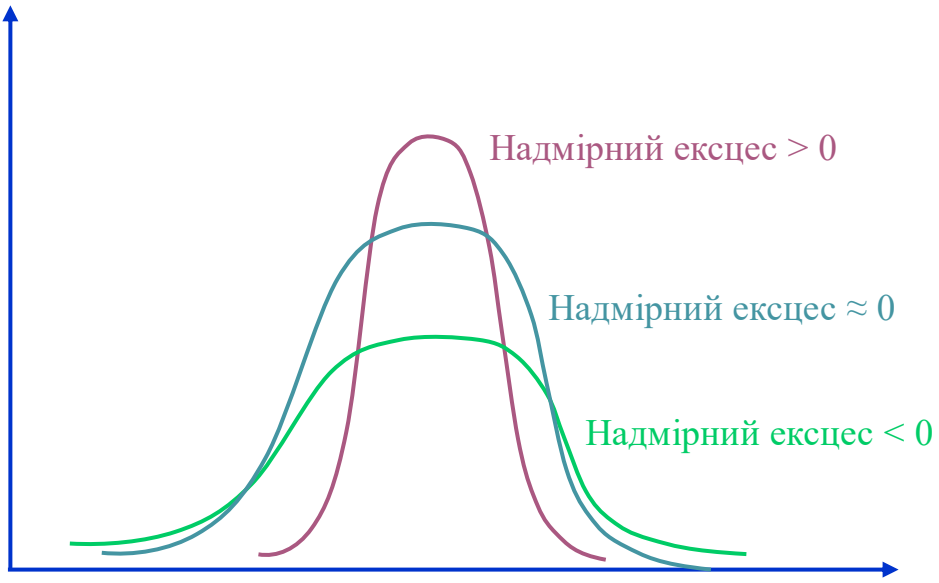


Рис. 2.27. Схематичне зображення розподілів з різним ексцесом

Зауважимо, що формула ексцесу є технічно подібною до формули асиметрії: визначається відхилення від центру розподілу у певному степені, яке потім стандартизується та усереднюється. Основна відмінність полягає у степені, в який підноситься різниця в чисельнику та середнє квадратичне відхилення у знаменнику. Крім того, простежується подібність із формулою дисперсії. Таким чином, основні характеристики розподілу (дисперсія, асиметрія та ексцес) є середніми арифметичними k-го порядку відхилень індивідуальних значень ознаки від середнього арифметичного і називаються **центральними моментами розподілу**. Загальна формула має такий вигляд:

$$\mu_k \approx \frac{\sum_{j=1}^{n}(x_j - \bar{x})^k * f_j}{\sum_{j=1}^{n} f_j} \qquad (2.28)$$

Відповідно, формули для показників асиметрії (центральний момент 3-го порядку) та ексцесу (центральний момент 4-го порядку) можна записати у такому вигляді:

$$As = \frac{\mu_3}{\sigma^3}\ ; Es^* = = \frac{\mu_4}{\sigma^4} \qquad (2.29)$$

Оцінювання рівномірності розподілу

Щоб оцінити, наскільки рівномірно розподілені дані для однієї або кількох груп показників, розрахуємо декілька коефіцієнтів на основі типових прикладів:

- Коефіцієнт ***концентрації K*** відображає, наскільки заданий розподіл відхиляється від рівномірного. Розрахований показник коливається у межах від 0 до 1, де у абсолютно рівномірного розподілу $K = 0$, а при зростанні ступеня нерівномірності K наближається до 1. Коефіцієнт концентрації розраховується за формулою:

$$K = \frac{1}{2}\sum_{j=1}^{n}\frac{1}{100}\left|d_j - D_j\right|, 0 \le K \le 1 \qquad \textbf{(2.30)}$$

де d_j *— частка об'єкту* j *у розподілі (у %) розрахована за кількістю елементів;* D_j *— частка об'єкту* j *у розподілі (у %) розрахована за обсягом значень ознаки.*

Використовуючи формулу (2.30), визначимо коефіцієнт концентрації на основі даних з таблиці 2.8, де компанії згруповано у шість категорій за кількістю штатних працівників. Для кожної групи визначено частку $\boldsymbol{d_j}$ за кількістю компаній і частку $\boldsymbol{D_j}$ за обсягом виробленої продукції. У крайньому стовпці справа для кожної групи (у рядку) розрахуємо модуль відхилення часток (із діленням на 100 для переходу від відсотків до коефіцієнтів). Знайшовши суму відхилень (0,52, тобто 52%, як розраховано у таблиці), підставимо її у формулу і в результаті одержимо: 1/2 * 0,52 = **0,26**, що відображає відносно незначний рівень концентрації.

Таблиця 2.8. Задача: розрахунок коефіцієнту концентрації

Групи компаній, к-ть працівників	**К-ть компаній у групі (частка групи d_j), %**	**Обсяг виробленої продукції (частка групи D_j), %**	***Модуль відхилення часток (1/100 * $\|d_j - D_j\|$)***
До 50	10	8	*0,02*
51-100	18	10	*0,08*
101-200	32	16	*0,16*
201-500	25	28	*0,03*
501-1000	10	20	*0,10*
Понад 1 000	5	18	*0,13*
Всього	**100**	**100**	***0,52***

- Якщо коефіцієнт концентрації вимірює ступінь концентрації загалом, то коефіцієнт ***локалізації*** L_j розраховується окремо для кожного елемента сукупності і дозволяє визначити, навколо якого суб'єкта або об'єкта локалізована концентрація. Якщо розподіл є рівномірним, то коефіцієнти L_j для кожного елемента є близькими до 1, а у разі локалізації в *j*-й складовій L_j перевищує 1. Коефіцієнт локалізації розраховується за такою формулою:

$$L_j = \frac{D_j}{d_j} \qquad (2.31)$$

Як видно з таблиці 2.9, найвищу концентрацію виробництва демонструють групи компаній із понад 500 та понад 1 000 працівників: на кожен відсоток компаній у цих групах припадає 2,0% та 3,6% виробленої продукції відповідно (з коефіцієнтами концентрації **2,0** та **3,6**).

Таблиця 2.9. Задача: розрахунок коефіцієнтів локалізації

Групи компаній, к-ть працівників	**К-ть компаній у групі (частка групи d_j), %**	**Обсяг виробленої продукції (частка групи D_j), %**	***Коефіцієнти локалізації,* $L_j = D_j / d_j$**
До 50	10	8	*0,80*
51-100	18	10	*0,56*
101-200	32	16	*0,50*
201-500	25	28	*1,12*
501-1 000	10	20	*2,00*
Понад 1 000	5	18	*3,60*
Всього	**100**	**100**	***X***

- Коефіцієнт ***подібності структур P*** дозволяє визначити, наскільки подібними є певні два об'єкти j та r (наприклад, регіони, компанії, галузі економіки тощо) за певними показниками, наприклад, доходами, рівнем ВВП, попитом тощо. Якщо структури об'єктів однакові, то коефіцієнт подібності P = 1, чим більше відрізняються – тим значення коефіцієнту ближче до 0. Коефіцієнт подібності розраховується за формулою:

$$P = 1 - \frac{1}{2} \sum_{j=1}^{n} \frac{1}{100} |d_j - d_r| \qquad (2.32)$$

де d_j — *частка об'єкту j у розподілі (у %);*

d_r — *частка об'єкту r у розподілі (у %).*

Розрахувавши суму модулів відхилень між частками кожної галузі у структурі економіки двох регіонів, отримаємо 0,74 (розраховано у таблиці 2.10.). Підставивши це значення у формулу (2.32), одержимо: 1 – 1/2 * 0,74 = 0,63, або **63%**, що відображає помірну подібність структури економіки досліджуваних регіонів. Результати розрахунків наведено в таблиці 2.10.

Таблиця 2.10. Задача: розрахунок коефіцієнту подібності структур

	Структура ВВП			
Регіон	**Транспорт та інфраструктура, %**	**Сільське господарство, %**	**Інформаційні технології, %**	**Всього, %**
Північний, d_j	12	32	56	**100**
Південний, d_r	36	45	19	**100**
Модуль відхилення, $1/100 * \lvert d_j - d_r \rvert$	0,24	0,13	0,37	**0,74**

- Коефіцієнти ***структурних зрушень***, лінійний $\bar{l}_d$ та квадратичний σ_d, показують, чи суттєво, в середньому, змінилася частка елементів у сукупності протягом обраного періоду (наприклад, у разі зміни структури основних клієнтів або постачальників). Для розрахунку структурних зрушень можна використати одну з таких формул:

$$\bar{l}_d = \frac{\sum_{j=1}^{n} \frac{1}{100} |d_{j1} - d_{jo}|}{m} \qquad (2.33)$$

$$\sigma_d = \sqrt{\frac{\sum_{j=1}^{n} \frac{1}{100} (d_{j1} - d_{jo})^2}{m}} \qquad (2.34)$$

де $\bar{l}_d$ – *лінійний коефіцієнт структурних зрушень;*

σ_d – *квадратичний коефіцієнт структурних зрушень;*

d_{j1} – *частка об'єкту* j *у поточному періоді (у %);*

d_{j0} – *частка об'єкту* j *у базисному періоді (у %);*

m – *кількість структурних елементів у сукупності.*

Розрахуємо зміну у структурі інвестиційного портфеля за один рік на основі даних таблиці 2.11 (суму модулів і квадратів відхилень часток уже розраховано в таблиці), використовуючи два показники. Лінійний коефіцієнт: 0,06 / 5 = 0,012 або **1,2%.** Квадратичний коефіцієнт: $\sqrt{0,08}$ / 5 = 0,057, або **5,7%.** Зауважимо, що квадратичний коефіцієнт структурних зрушень є чутливішим до змін, особливо значних. Відповідні дані та отримані результати розрахунку коефіцієнтів структурних зрушень наведено в табл. 2.11.

Таблиця 2.11. Задача: розрахунок коефіцієнтів структурних зрушень

Інвестиційний актив	**20X0, dj0 (%)**	**20X1, dj1 (%)**	**Модуль відхилення часток (1/100 * |dj1 − dj0|**	**Квадрат відхилення часток (1/100 * (dj2 − dj1)^2)**
Золото	22	23	0,01	0,01
Державні облігації	18	19	0,01	0,01
Акції компаній	20	21	0,01	0,01
Валютні депозити	25	23	0,02	0,04
Нерухомість	15	14	0,01	0,01
Усього	**100**	**100**	**0,06**	**0,08**

Питання для самоперевірки

1. *Яка сукупність вважається однорідною згідно з критерієм квадратичного коефіцієнта варіації?*
2. *Як співвідносяться між собою середнє, мода і медіана у суто симетричному розподілі?*
3. *Охарактеризуйте лівосторонню та правосторонню асиметрію (As).*
4. *Які властивості розподілу характеризує ступінь ексцесу (Es)?*
5. *Поясніть призначення коефіцієнта концентрації (K) та коефіцієнта локалізації (Lj).*
6. *Для чого розраховують коефіцієнти структурних зрушень?*

2.4. Нормальний та інші статистичні розподіли

Ознайомившись із основними описовими статистиками розподілу, зокрема мірами центральної тенденції, варіацією, однорідністю, асиметрією та ексцесом, у цій темі розглянемо найпоширеніші види теоретичних розподілів та їхні ключові характеристики.

Очевидно, що більшість розподілів, з якими стикаються на практиці, є унікальними, але не «ідеальними» з точки зору відповідності теоретичним розподілам – таким як нормальний, біноміальний або розподіл Пуассона.

У статистичному аналізі та моделюванні часто намагаються визначити, до якого теоретичного розподілу є близьким досліджуваний емпіричний, щоб використати властивості відомого емпіричного розподілу для перевірки припущень і статистичних гіпотез, моделювання та формування висновків, зокрема:

- Якщо відомо, якому розподілу підпорядковується вибірка, можна сформулювати обґрунтовані припущення щодо властивостей усієї генеральної сукупності.
- Використання властивостей теоретичних розподілів дає змогу відповідати на практичні запитання: яка частка студентів має середній бал понад 95; до якого перцентиля найуспішніших компаній належить фірма з доходом, що перевищує Х одиниць тощо.
- В основі більшості економіко-математичних моделей лежать припущення щодо розподілу даних. Розуміння характеристик відповідних розподілів дозволяє впевнено застосовувати такі статистичні методи як, перевірка гіпотез, довірчі інтервали, регресійний аналіз взаємозв'язків, прогнозування на основі економетричних моделей тощо.

Далі розглянемо графіки функції щільності та ключові характеристики декількох поширених теоретичних розподілів, а також сферу їхнього застосування.

Нормальний розподіл

Основну увагу у посібнику зосереджено на нормальному розподілі, який є еталонним і одним з найпоширеніших не лише у статистичному й економетричному аналізі, а й у повсякденному житті – для опису «випадкових» процесів (наприклад, розподіл людей за зростом, вагою, розміром взуття та одягу; розбризкування води з-під коліс автомобіля; зношеність парти, дошки або барної стійки).

Нормальний розподіл є симетричним (коефіцієнт асиметрії *As* дорівнює нулю) та має показник ексцесу *Es* = 3 (тобто надмірний ексцес *Es** дорівнює нулю), як зображено на рис. 2.28. З цього випливає, що, як і в будь-якому іншому симетричному розподілі, мода, медіана та середнє значення нормального розподілу збігаються: $Mo = Me = \bar{x}$. Ще однією властивістю є те, що лінійна комбінація випадкових величин є нормально розподіленою. Графік функції щільності нормального розподілу наведено на рис. 2.28.

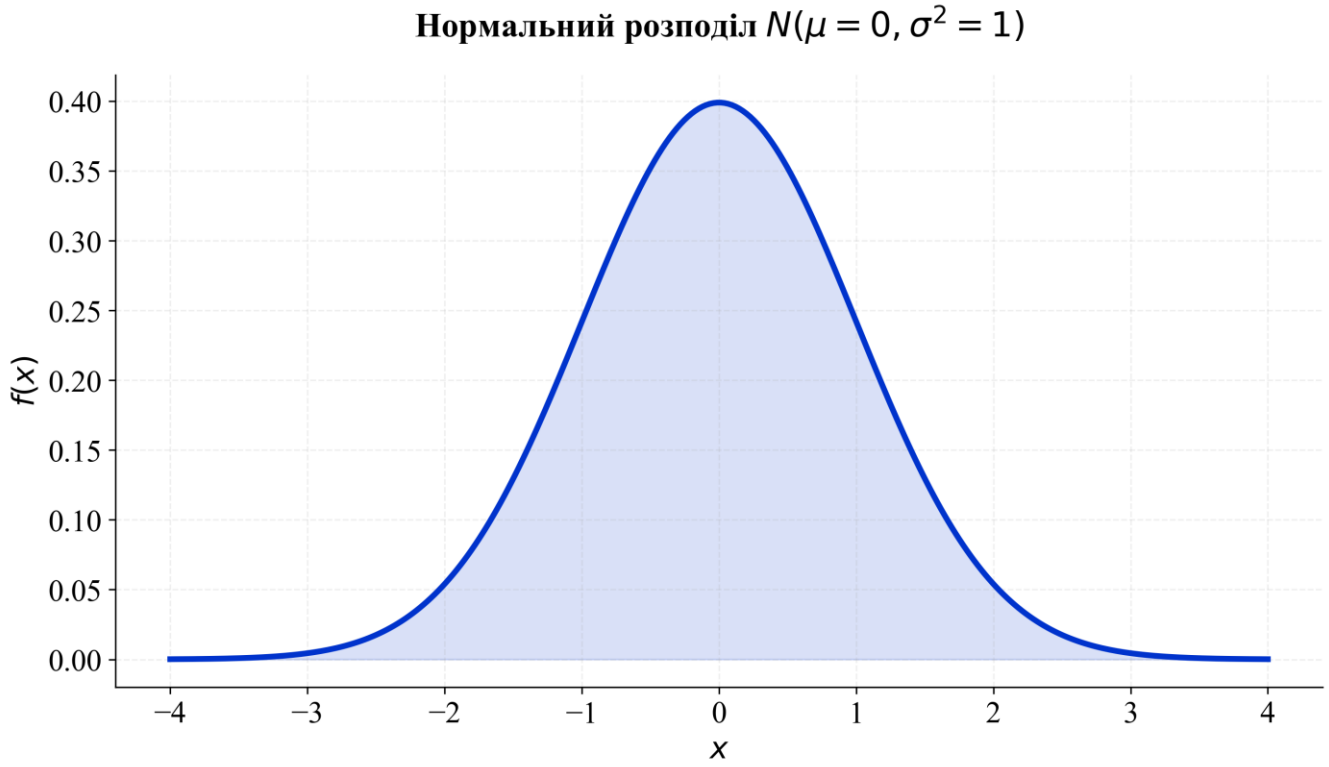


Рис. 2.28. Графік функції щільності нормального розподілу

Нормальний розподіл можна описати за допомогою двох параметрів – середнього значення μ та дисперсії σ^2. Зауважимо, що для конкретної вибірки

середнє значення позначають $\bar{x}$ (вибіркове середнє), тоді як μ використовують для позначення середнього значення у теоретичному розподілі генеральної сукупності. Нижче наведено формулу (2.35), що описує графік щільності нормального розподілу.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)} \text{ для } -\infty < x < +\infty \qquad \textbf{(2.35)}$$

де $f(x)$ – *функція щільності нормального розподілу.*

На практиці для зручного використання властивостей нормального розподілу, зокрема для переходу між імовірностями та квантилями, кожне значення ознаки нормального розподілу x_j трансформують у z_j (z-значення або z-score) шляхом стандартизації за наведеною нижче формулою (2.36). У результаті одержують **стандартний нормальний розподіл** із середнім значенням $\mu = 0$ та дисперсією $\sigma^2 = 1$.

$$z_j = \frac{x_j - \mu}{\sigma} \qquad \textbf{(2.36)}$$

Результати такої трансформації схематично відображено на графіку 2.29. По суті, за цією формулою ми знаходимо відстань від кожної ознаки x_j до середнього значення, виражену у стандартних відхиленнях σ. Розподіл, який отримуємо внаслідок такої трансформації, називається *стандартним* нормальним і має такі корисні властивості:

- У межах $\bar{x} \pm 3\sigma$ знаходиться ≈99,7% значень розподілу («**правило трьох сигм**»). Інакше кажучи, ≈99,7% елементів сукупності мають значення ознаки в цих межах; або імовірність, що довільно обране значення із сукупності потрапить у ці межі, становить ≈99,7%.
- У межах $\bar{x} \pm 2\sigma$ знаходиться ≈95,4% значень розподілу.

- У межах $\bar{x} \pm \sigma$ знаходиться ≈68,3% значень розподілу.

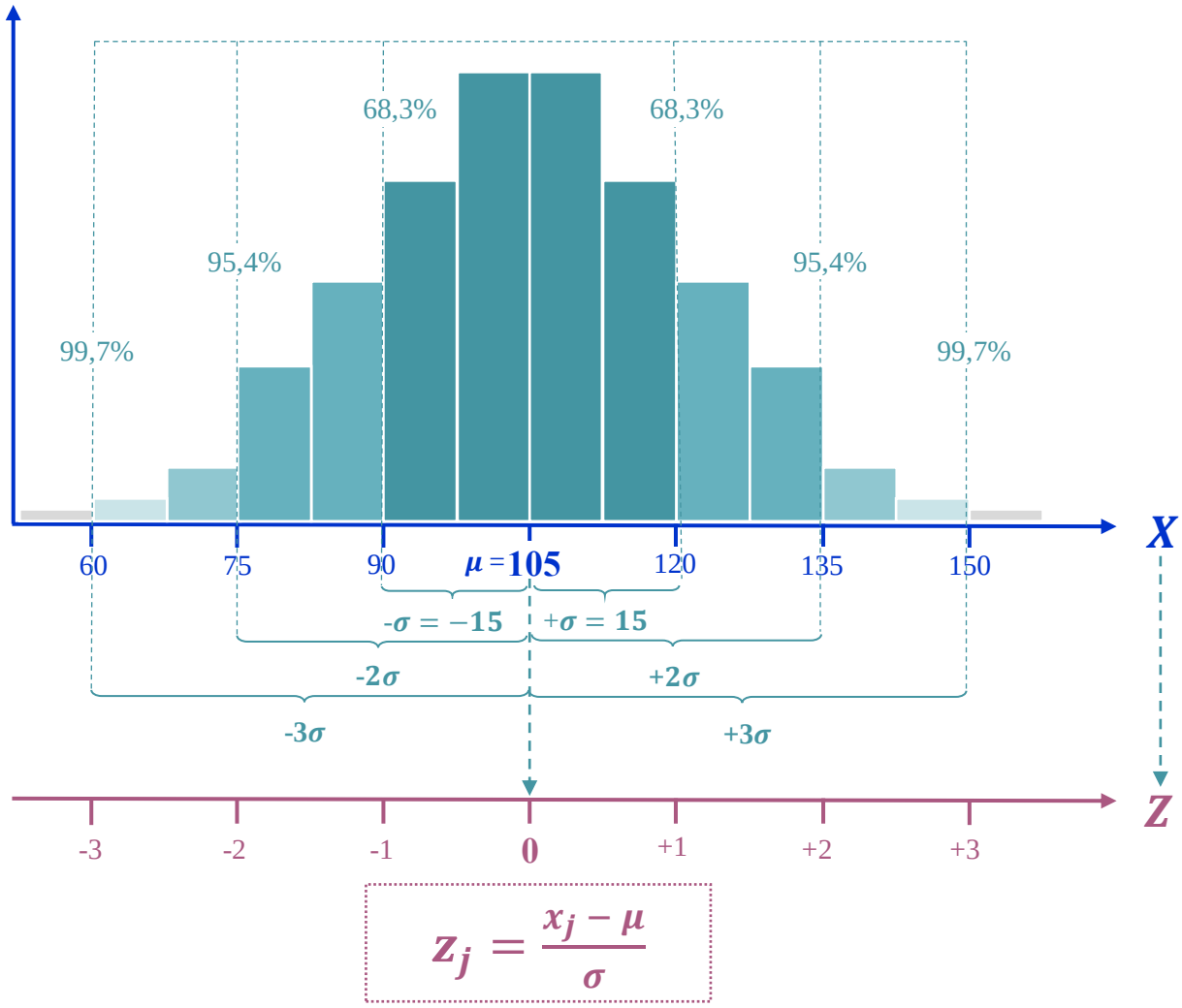


Рис. 2.29. Схематичне зображення переходу до *стандартного* нормального розподілу та його властивостей

Кумулятивна функція стандартного нормального розподілу $\Phi(z)$ розраховується за формулою (2.27):

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{\frac{-x^2}{2}} dx \qquad (2.37)$$

де $\Phi(z)$ – *кумулятивна функція стандартного нормального розподілу;*
z – *стандартизовані значення ознаки (див. формулу (2.36)).*

На практиці відповідні значення функції $\Phi(z)$ для кожного z можна знайти у статистичних таблицях, які також вбудовані в Microsoft Excel, пакети Python, інші

статистичні застосунки. У Microsoft Excel ця функція викликається командою *NORM.S.DIST([значення], TRUE)*, а у Python – методом *norm.cdf([значення])*, який можна знайти у бібліотеці scipy.stats (імпортувати її можна командою: *from scipy.stats import norm*).

Використовуючи властивості кумулятивної функції стандартного нормального розподілу, можна визначити імовірність потрапляння значення ознаки в інтервал $(-\infty,\ a)$, тобто «менше, ніж a». На рисунку 2.30 ця імовірність зображена заштрихованою областю ліворуч від a.

$$P(X < a) = \Phi(z) = \Phi\left(\frac{a-\mu}{\sigma}\right) \qquad \textbf{(2.38)}$$

де $\Phi(z)$ – *кумулятивна функція стандартного нормального розподілу*

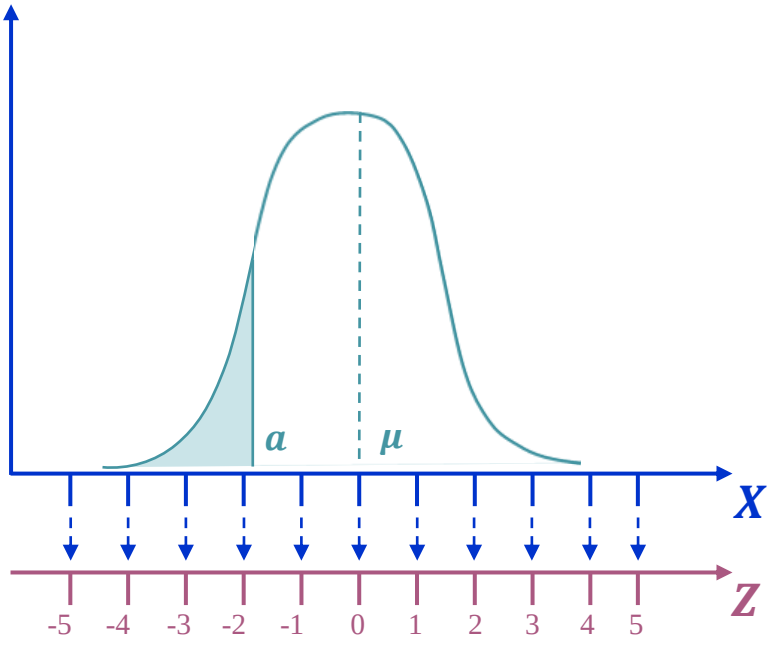


Рис. 2.30. Площа під кривою розподілу на інтервалі $(-\infty,\ a)$

Зауважимо, що одиниця тут позначає імовірність потрапляння в інтервал $(-\infty, +\infty)$, або, за геометричним змістом, площу під усією кривою функції розподілу, яка завжди дорівнює 1. Оскільки стандартний нормальний розподіл є симетричним, імовірність потрапляння в інтервал $(a, +\infty)$, тобто «більше, ніж a»,

можна визначити як різницю між повною площею під кривою та площею лівої частини (інтервалу $(-\infty,\ a)$):

$$P(X > a) = 1 - P(X < a) = 1 - \Phi\left(\frac{a-\mu}{\sigma}\right) \qquad (2.39)$$

На рисунку 2.31 ця імовірність зображена заштрихованою областю праворуч від a.

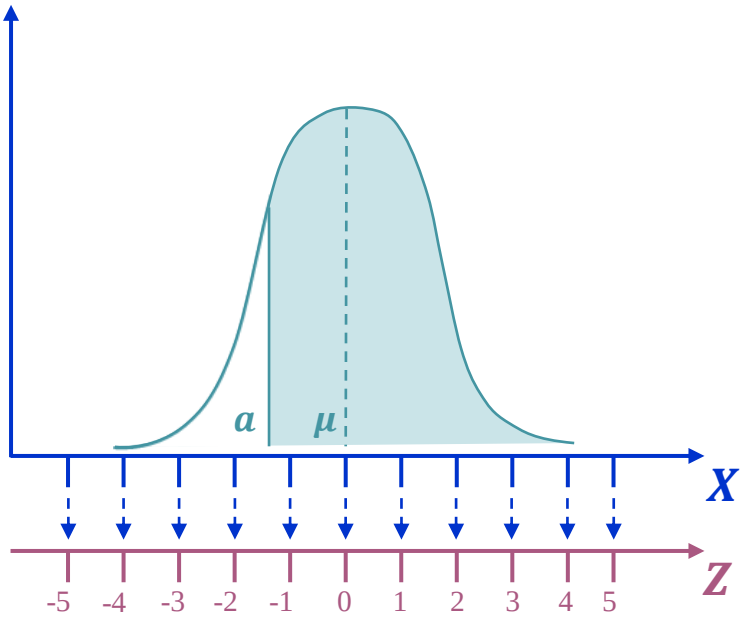


Рис. 2.31. Площа під кривою розподілу на інтервалі (a, +∞)

Використовуючи наведені формули, можна також визначити імовірність потрапляння значення ознаки в довільний інтервал (a , b) як різницю двох значень кумулятивної функції стандартного нормального розподілу (див. формулу (2.40)):

$$P(a < X < b) = P(X < b) - P(X < a) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \qquad (2.40)$$

Схематичне зображення знаходження площі в рамках цього інтервалу наведено на рис. 2.32.

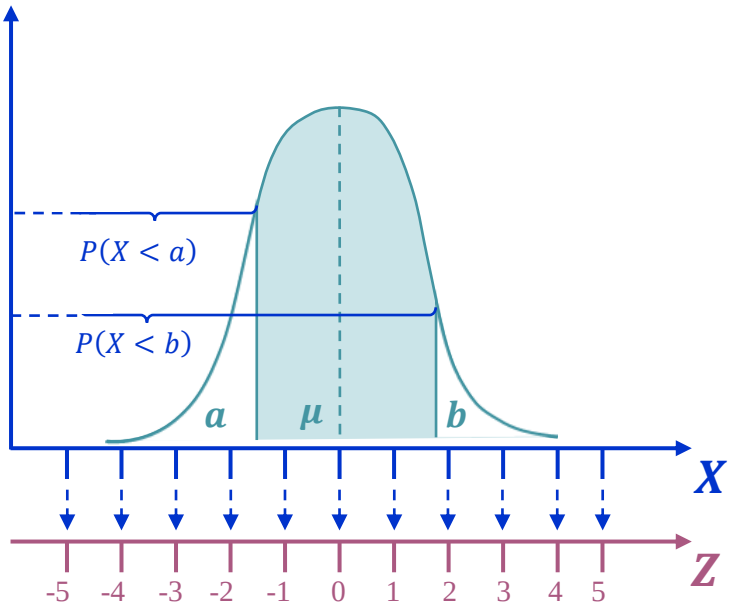


Рис. 2.32. Площа під кривою розподілу на інтервалі $(-\infty, +\infty)$

Розглянуті формули дають змогу визначити імовірність потрапляння значення ознаки в інтервал із заданими межами. Це стосується задач на визначення імовірності або частки сукупності з певними характеристиками. Наприклад: яку частку від усієї групи становлять студенти із загальним рейтингом від 71 до 95 балів.

Розглянуті формули дають змогу визначити імовірність за відомими межами інтервалу. Однак на практиці часто виникає обернена задача: знаючи імовірність, знайти відповідне значення ознаки z або квантиль розподілу. Наприклад, який мінімальний бал треба обрати, щоб у результаті бути серед 5% найкращих конкурсантів. Для цього використовують обернену кумулятивну функцію стандартного нормального розподілу $\Phi^{-1}(p)$. У Microsoft Excel ця функція називається *NORM.S.INV([значення імовірності])*, а у Python – метод *norm.ppf([значення імовірності])*, який можна викликати з бібліотеки *scipy.stats* (імпортувати бібліотеку можна такою командою: *from scipy.stats import norm*).

Задача: знаходження імовірності у заданому інтервалі

Припустимо, що розподіл результатів тесту є близьким до нормального із середнім значенням μ=700 та середнім квадратичним відхиленням σ=150, тобто $X \sim N(\mu=700, \sigma=150)$. Необхідно визначити частку результатів, що потрапляють в інтервал [510, 650].

Розв'язок:

- Стандартизуємо задані межі інтервалу, трансформувавши значення з нормального розподілу X у Z-значення стандартного нормального розподілу. Для x = 510: z = (510−700) / 150 ≈ −1,27. Для x = 650: z = (650−700) / 150 ≈ −0,33.
- За допомогою кумулятивної функції стандартного нормального розподілу $\Phi(z)$ знайдемо імовірність того, що результат не перевищує кожне зі стандартизованих значень, використовуючи таблиці для стандартного нормального розподілу, або функції в Microsoft Excel чи Python: $\Phi(-1{,}27) = P(Z \leq -1{,}27) \approx 0{,}102$; $\Phi(-0{,}33) = P(Z \leq -0{,}33) \approx 0{,}370$.
- Отже, визначимо імовірність потрапляння результату за тест в інтервал: $P(510 \leq X \leq 650) = P(Z \leq -0.33) - P(Z \leq -1.27) = 0{,}370 - 0{,}102 = 0{,}268$, або 26,8%.

Задача: знаходження порогового балу для топ-10% результатів

Використовуючи умови попередньої задачі, необхідно знайти мінімальну кількість балів, яку потрібно набрати, щоб увійти до 10% кращих результатів:

- Топ-10% результатів відповідають як 90%-перцентилю розподілу (квантиль $z_{0,9}$), праворуч від якого розташовано 10% найвищих балів. Для розв'язку цієї задачі нам необхідно скористатися оберненою кумулятивною функцією стандартного нормального розподілу $\Phi^{-1}(p)$.
- Використовуючи табличні дані, функції в Microsoft Excel або Python знайдемо $z_{0,9} = \Phi^{-1}(0{,}9) \approx 1{,}282$.

- Трансформуємо Z-значення стандартного нормального розподілу у вихідні величини нормального розподілу X (бали за тест), виразивши x із формули стандартизації: $x = \mu + z * \sigma$ = 700 + 1,282*150 = 892,3.
- Отже, для потрапляння до 10% найкращих результатів необхідно набрати щонайменше 893 бали.

Інші статистичні розподіли

Розглянемо оглядово інші поширені розподіли випадкових величин, зосередившись на графіках функцій щільності та сферах їх застосування:

- **t-розподіл** (розподіл Ст'юдента) є подібним до нормального, однак має ширші «хвости», тобто надає більшу імовірність крайнім значенням. Цей вид розподілу використовується, коли дисперсія генеральної сукупності σ^2 невідома і оцінюється за даними вибірки, що є типовим на практиці при роботі з невеликими вибірками (менше 30 спостережень). При збільшенні кількості спостережень t-розподіл наближається до стандартного нормального. Графік щільності t-розподілу Ст'юдента у порівнянні з функцією щільності нормального закону розподілу наведено на рис. 2.33.

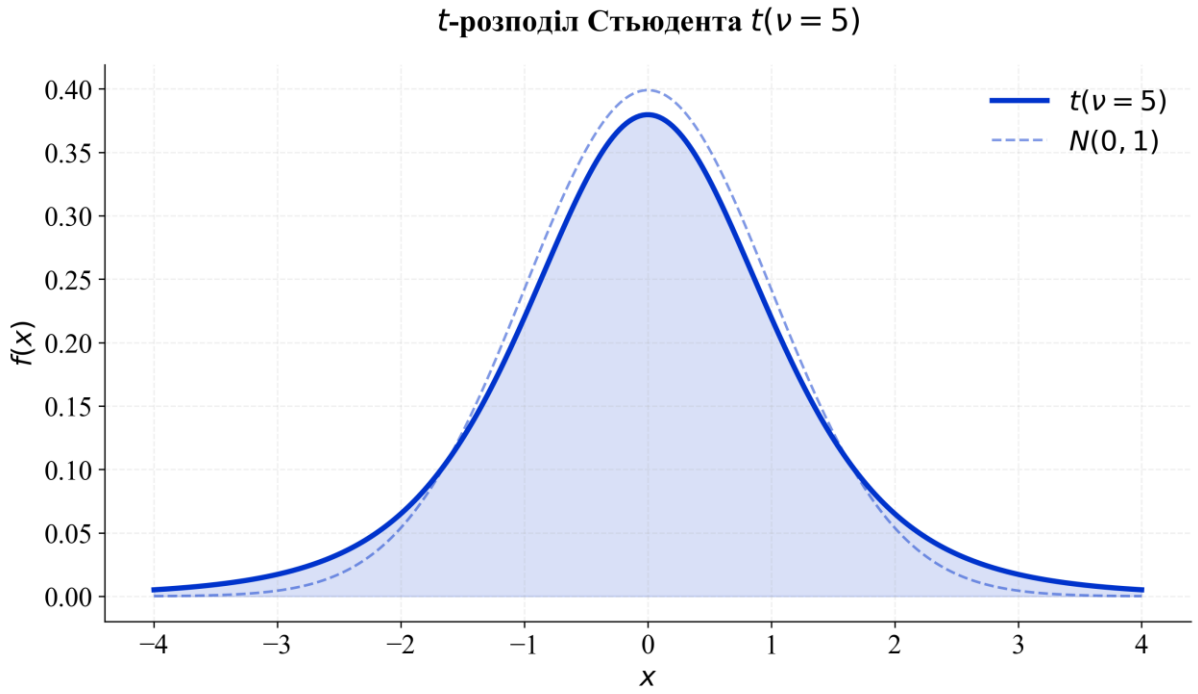


Рис. 2.33. Графік функції щільності **t-розподілу**, у порівнянні з нормальним

- **Рівномірний розподіл** застосовується для моделювання процесів, у яких кожне значення в межах діапазону є однаково ймовірним. Графік функції щільності рівномірного розподілу наведено на рис. 2.34. Цей розподіл доцільно використовувати на практиці, коли немає підстав надавати перевагу жодному конкретному значенню: генерування випадкових чисел для симуляцій, зокрема методом Монте-Карло; моделювання часу очікування у черзі, коли розклад невідомий.

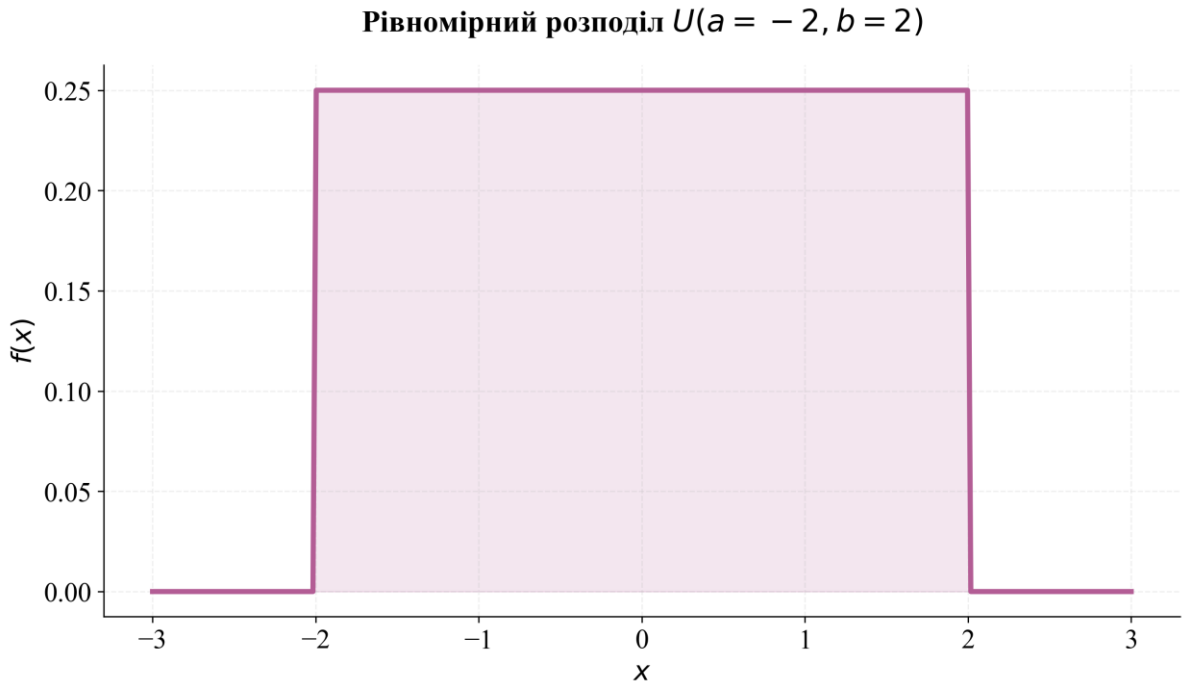


Рис. 2.34. Графік функції щільності **рівномірного розподілу**

- **Експоненційний розподіл** використовується для моделювання часу між послідовними незалежними подіями, що відбуваються з однаковою інтенсивністю. Графік функції щільності даного розподілу наведено на рис. 2.35. Приклади застосування: тривалість дзвінка у кол-центрі; час роботи пристрою до несправності; час для розвантаження товарів на складі; оцінка часу до дефолту.

- **Розподіл хі-квадрат** – це розподіл суми квадратів певної кількості незалежних стандартних нормальних випадкових величин. Цей розподіл є одним із ключових у статистиці, адже використовується для широкого

спектру задач: перевірки гіпотез щодо відповідності емпіричного розподілу теоретичному (наприклад, чи є розподіл балів нормальним); побудови F-розподілу; оцінювання дисперсії генеральної сукупності та побудови довірчих інтервалів для дисперсії; перевірки значущості зв'язку між категорійними змінними. Графік функції щільності даного розподілу наведено на рис. 2.36.

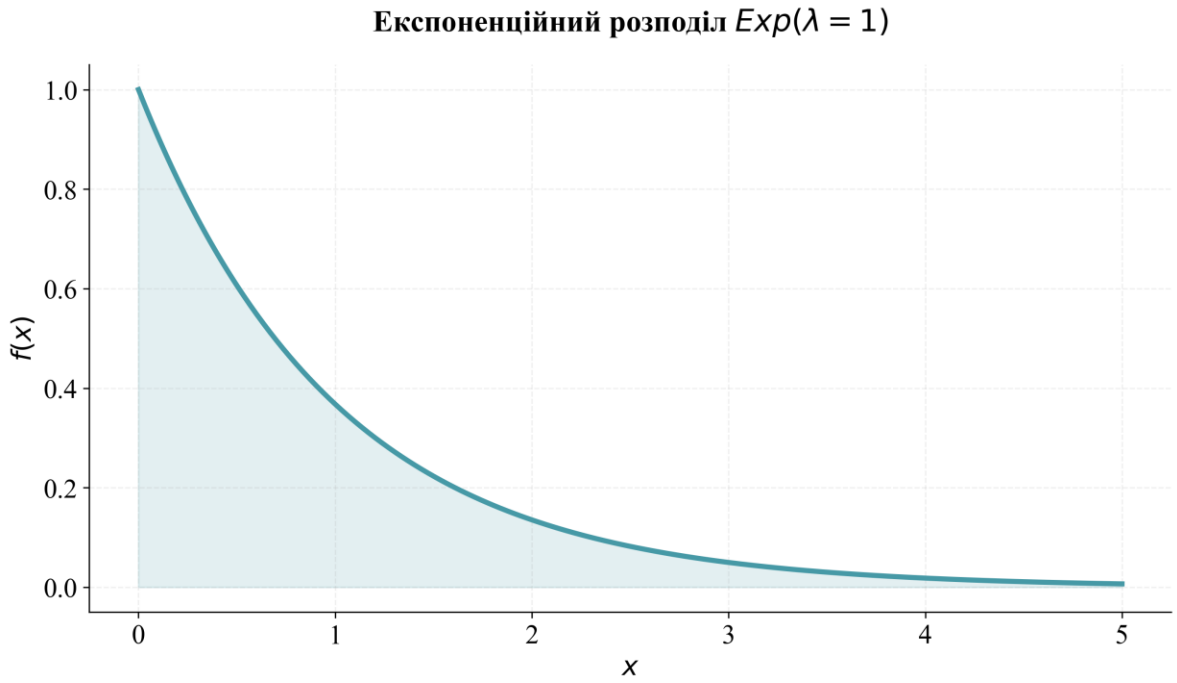


Рис. 2.35. Графік функції щільності **експоненційного розподілу**

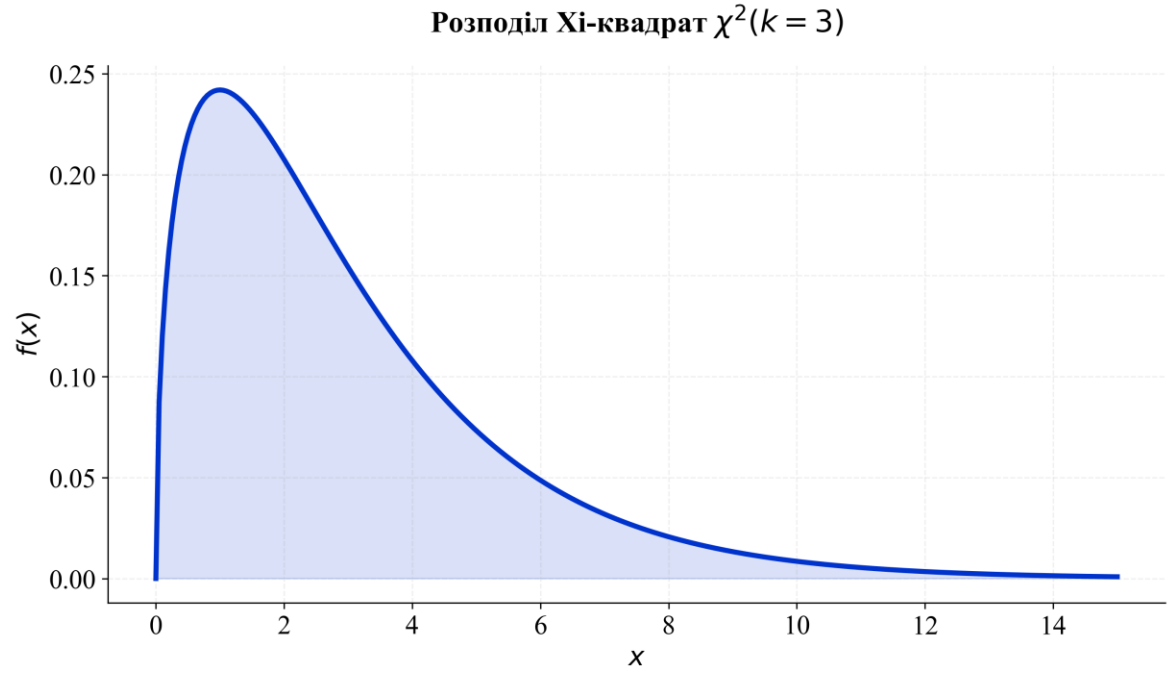


Рис. 2.36. Графік функції щільності розподілу **хі-квадрат**

Питання для самоперевірки

1. *Які два ключові параметри описують нормальний розподіл?*
2. *Сформулюйте «правило трьох сигм» для стандартного нормального розподілу.*
3. *Що таке стандартизація даних (z-score) і для чого вона проводиться?*
4. *У яких ситуаціях замість нормального розподілу використовують t-розподіл (Стьюдента)?*
5. *Для моделювання яких процесів використовують експоненційний розподіл та розподіл хі-квадрат?*

СПИСОК ДЖЕРЕЛ

1. Державна служба статистики України. Квартальні національні рахунки [Електронний ресурс]. – Режим доступу: https://stat.gov.ua/uk/explorer?urn=SSSU%3ADF_QUARTERLY_NATIONAL_ACCOUNTS%28~%29&filter=GROSS_DOM_PRODUCT.UA00000000000000000.IDX_VOLUME_CONST_2021._T.Q
2. Державна служба статистики України. Річні національні рахунки [Електронний ресурс]. – Режим доступу: https://stat.gov.ua/uk/explorer
3. Підручник / С. С. Герасименко, А. В. Головач, А. М. Єріна та ін. ; за наук. ред. д-ра екон. наук С. С. Герасименка. – 2-ге вид., перероб. і доп. – Київ : КНЕУ, 2000. – 467 с.
4. Статистика : опорний конспект лекцій / В. Б. Захожай, А. М. Єріна, І. А. Гончар та ін. – Київ : МАУП, 2006. – 160 с. – Бібліогр. : с. 159.
5. Теорія статистики : практикум / А. М. Єріна, З. О. Пальян. – 5-е вид. – Київ : Товариство «Знання», 2006. – 255 с.
6. Downey A. B. Think Stats : Exploratory Data Analysis in Python. – 2-е вид. – New York : O’Reilly Media, 2023. – 300 с.
7. Ember. European Electricity Review 2025 [Електронний ресурс]. – Режим доступу: https://ember-energy.org/app/uploads/2025/01/EER_2025_22012025.pdf
8. Eurostat. GDP per capita in PPS [Електронний ресурс]. – Режим доступу: https://ec.europa.eu/eurostat/databrowser/view/TEC00114/default/table?lang=en
9. Lancaster H. O. Expectations of Life : A Study in the Demography, Statistics and History of World Mortality. – New York : Springer, 1990.
10. Navarro D., Weed E. Learning Statistics with Python [Електронне видання]. – 2021. – Режим доступу: https://ethanweed.github.io/pythonbook/

11. OECD. Economic outlook [Електронний ресурс]. – Режим доступу: https://data-explorer.oecd.org/vis?fs[0]=Topic%2C1%7CEconomy%23ECO%23%7CEconomic%20outlook%23ECO_OUT%23&pg=0&bp=true&snb=6&vw=tb&df[ds]=dsDisseminateFinalDMZ&df[id]=DSD_EO%40DF_EO&df[ag]=OECD.ECO.MAD&df[vs]=1.4&dq=.GDPVD_CAP%2BCPI%2BYDRH%2BGDPV_ANNPCT.A&pd=2025%2C2025&to[TIME_PERIOD]=false&ly[cl]=MEASURE&ly[rw]=REF_AREA
12. University of Vienna. Mozart's music does not make you smarter, study finds. ScienceDaily, 2010, May 10 [Електронний ресурс]. – Режим доступу: www.sciencedaily.com/releases/2010/05/100510075415.htm
13. Wasserman L. All of Statistics. New York, NY : Springer New York, 2004. URL: https://doi.org/10.1007/978-0-387-21736-9
14. Witte R. S., Witte J. S. Statistics. – 11-е вид. – Hoboken, NJ : John Wiley & Sons, 2017. – 496 с.

STATYSTICAL ANALYSIS OF DATA:

FROM CONCEPTS TO PRACTICE

SUMMARY

Statistical data analysis offers essential tools for obtaining actionable insights from large volumes of information to support data-driven decisions in finance, economics, and related fields.

The text book “Statistical Analysis of Data: From Concepts to Practice” provides a structured course in foundational statistics, combining core theoretical concepts with applied methods and instruments for data transformation, grouping, and visualization, summary measurements and interpretation of findings. It also explains the theoretical logic and practical implications of statistical distributions and time series, developing the main toolkit for descriptive analysis – measures of central tendency, dispersion, shape (e.g. skewness and kurtosis), key characteristics of common distributions and areas of their applications, with emphasis on the standard normal distribution.

The book places strong focus on real-world application: each concept is illustrated with hands-on examples and exercises, as well as code snippets in Python so that students learn both the formulas and how to implement them on exemplary datasets using real-world analytical software. It also outlines the crucial stages of a data analysis workflow: data collection, cleaning and preparation, exploratory description and visualization, hypothesis testing, evaluation of relationships, and presentation of conclusions.

The textbook is designed for undergraduate and graduate students, instructors, and practitioners who need a well-structured and practice-oriented introduction to statistics and data analytics.

Навчальне видання

СОВА
Євгеній Станіславович

СТАТИСТИЧНИЙ АНАЛІЗ ДАНИХ:

Від концепцій до практики

Навчальний посібник

В авторській редакції