ОСНОВИ СИСТЕМНОГО АНАЛІЗУ ТА МОДЕЛЮВАННЯ СИСТЕМ У СФЕРІ КІБЕРБЕЗПЕКИ

Частина 1. Первинний аналіз даних: описова статистика, аналіз гіпотез, виявлення залежностей

Навчальний посібник

Д.М. ШАРАДКІН | А.В. МИКИТЮК | С.Г. ПЕТРОВ | Д.А. МІНОЧКІН

iscip.kpi.ua

Міністерство освіти і науки України
Національний технічний університет України
“Київський політехнічний інститут імені Ігоря Сікорського”
Інститут спеціального зв’язку та захисту інформації

ОСНОВИ СИСТЕМНОГО АНАЛІЗУ ТА МОДЕЛЮВАННЯ СИСТЕМ У СФЕРІ КІБЕРБЕЗПЕКИ

Частина 1. Первинний аналіз даних: описова статистика, аналіз гіпотез, виявлення залежностей

Навчальний посібник

Рекомендовано Методичною радою КПІ ім. Ігоря Сікорського
як навчальний посібник для здобувачів ступеня бакалавра
за освітньою програмою “Комп’ютерні системи і технології спеціального зв’язку”
спеціальності F3 Комп’ютерні науки

Укладачі: Д. М. Шарадкін, А. В. Микитюк, С. Г. Петров, Д. А. Міночкін

Електронне мережеве навчальне видання

Київ
ІСЗЗІ КПІ ім. Ігоря Сікорського
2026

УДК 004.942 (075.8)
Ш25

Укладачі: *Шарадкін Дмитро Михайлович* канд. техн. наук, доц.
Микитюк Артем В'ячеславович, PhD
Петров Станіслав Геннадійович, д-р юрид. наук
Міночкін Дмитро Анатолійович, канд. техн. наук, ст. наук. співроб.

Рецензенти *Горнійчук І. В.,* PhD, ІСЗЗІ КПІ ім. Ігоря Сікорського
Шевчук О. С., PhD, ІСЗЗІ КПІ ім. Ігоря Сікорського

Гриф надано Методичною радою КПІ ім. Ігоря Сікорського
(протокол № 9 від 26.06.2026 р.)
за поданням Вченої ради ІСЗЗІ КПІ ім. Ігоря Сікорського
(протокол № 13 від 22.06.2026 р.)

Д. М. Шарадкін, А. В. Микитюк, С. Г. Петров, Д. А. Міночкін

Ш25 **Основи системного аналізу та моделювання систем у сфері кібербезпеки. Частина 1. Первинний аналіз даних: описова статистика, аналіз гіпотез, виявлення залежностей.** [Електронний ресурс] : навч. посіб. для здобувачів ступеня бакалавра за освіт. програмою "Комп'ютерні системи і технології спеціального зв'язку" спец. F3 Комп'ютерні науки / ІСЗЗІ КПІ ім. Ігоря Сікорського. Київ : КПІ ім. Ігоря Сікорського, 2026. – 262 с.

Навчальний посібник "Основи системного аналізу та моделювання систем у сфері кібербезпеки" присвячений базовим методам обробки та сучасного аналізу даних, що є необхідними для побудови систем моніторингу та поведінкового аналізу систем реального світу, зокрема – систем забезпечення кібербезпеки. Частина 1 навчального посібника охоплює задачі класичного системного аналізу та його застосуванню до задач забезпечення кібербезпеки; охоплює задачі первинного аналізу даних: описову статистику, перевірку статистичних гіпотез, виявлення залежностей між змінними, а також застосування методів вирішення вказаних задач для аналізу різноманітних технічних, економічних, медичних, фінансових систем, систем кіберзахисту. Значна увага приділяється інтерпретації результатів та практичному використанню статистичних методів у задачах виявлення аномалій, оцінювання ризиків і підтримки прийняття рішень.
Навчальний посібник призначений для здобувачів вищої освіти першого (бакалаврського) рівня вищої освіти, які навчаються за спеціальністю F3 Комп'ютерні науки і вивчають освітній компонент "Основи системного аналізу та моделювання систем у сфері кібербезпеки", а також рекомендовано для використання при виконанні кваліфікаційних робіт за відповідною спеціальністю.

УДК 004.942(075.8)

Реєстр. № НП 25/26-498. Обсяг 9,8 авт. арк.
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
проспект Берестейський, 37, м. Київ, 03056
https://kpi.ua
Свідоцтво про внесення до Державного реєстру видавців, виготовлювачів
і розповсюджувачів видавничої продукції ДК № 5354 від 25.05.2017 р.

ЗМІСТ

ПЕРЕДМОВА

Сучасний розвиток інформаційних технологій, систем моніторингу та засобів забезпечення кібербезпеки супроводжується стрімким зростанням обсягів даних, що генеруються технічними системами, мережевими сервісами та користувачами. Проте самі по собі ці дані не мають цінності, якщо немає вміння їх використовувати. Ефективне використання цих даних неможливе без застосування методів системного аналізу, статистики, математичного моделювання та машинного навчання. Саме тому підготовка майбутніх фахівців у галузі комп'ютерних наук і систем кіберзахисту потребує не лише ознайомлення з окремими алгоритмами, технічними засобами та програмними інструментами, а й формування цілісного аналітичного мислення, здатності правильно інтерпретувати результати досліджень, оцінювати достовірність висновків і розуміти межі застосування математичних методів.

Навчальний посібник *"Основи системного аналізу та моделювання систем у сфері кібербезпеки. Частина 1. Первинний аналіз даних: описова статистика, аналіз гіпотез, виявлення залежностей"* створено для здобувачів ступеня бакалавра за освітньою програмою "Комп'ютерні системи і технології спеціального зв'язку" спеціальності F3 Комп'ютерні науки. Посібник присвячений базовим методам аналізу емпіричних даних, які лежать в основі сучасних систем виявлення аномалій, поведінкової аналітики, моніторингу подій та підтримки прийняття рішень у сфері кібербезпеки.

Важливе місце у посібнику відведено базовим питанням класичного системного аналізу як методологічній основі дослідження складних об'єктів і процесів. Розглядаються поняття системи, структури, елементів, зв'язків, цілей функціонування та критеріїв ефективності, а також підходи до декомпозиції, формалізації та моделювання складних систем. Особлива увага приділяється аналізу взаємозв'язків між компонентами систем, виявленню факторів, що впливають на їхню поведінку, та основам побудови математичних моделей, придатних для подальшого статистичного дослідження й програмної реалізації. Такий підхід сприяє формуванню у здобувачів цілісного уявлення про системне мислення як універсального інструменту дослідження технічних, інформаційних, економічних, соціальних систем та процесів, і необхідного для розуміння сучасних задач кібербезпеки, аналізу даних та розробки інтелектуальних інформаційних систем.

Статистика та її методи розглядаються в межах курсу не лише як інструмент для вирішення вузькоспеціалізованих задач кіберзахисту, а як універсальний

математичний апарат аналізу даних. Саме тому коло прикладів і задач у посібнику навмисно виходить за межі суто кібербезпекової тематики. Поряд із прикладами аналізу мережевого трафіку, журналів подій та поведінки користувачів розглядаються задачі з інженерних систем, економіки, біоінформатики, медицини, експериментальних досліджень і соціальних процесів. Такий міждисциплінарний підхід дозволяє краще зрозуміти загальні принципи та взаємозв'язок системного та статистичного мислення, навчитися відокремлювати математичний метод від конкретного прикладного контексту й усвідомити універсальність статистичних закономірностей.

Особлива увага приділяється не лише формальному застосуванню статистичних процедур, а й розумінню припущень, що лежать в їхній основі, обмежень та умов коректного використання. Вміння працювати з вибірками, оцінювати якість даних, перевіряти статистичні гіпотези та інтерпретувати результати аналізу є невід'ємною складовою професійної компетентності сучасного фахівця з інформаційних технологій незалежно від спеціалізації. Ці навички однаково важливі як для розробки програмних систем і алгоритмів, так і для аналізу результатів експериментів, оцінювання продуктивності систем, побудови моделей та дослідження поведінки складних технічних об'єктів.

При складанні навчального посібника автори насамперед орієнтувалися на практичне застосування викладеного матеріалу у майбутній професійній діяльності читача. Основною метою було подати матеріал у доступній та зрозумілій формі, спираючись на вже набуті знання з математики, теорії ймовірностей, математичної статистики, програмування та інших суміжних дисциплін. Водночас автори свідомо уникали надмірної математизації та ускладнення викладу. Такий підхід зумовлений переконанням, що для майбутнього інженера значно важливішим є глибоке розуміння ідей, принципів та логіки методів, ніж надмірна формальна суворість математичних доведень. Саме тому у посібнику основну увагу приділено осмисленню підходів, їх практичній інтерпретації та можливостям застосування у реальних задачах.

Оскільки навіть найсучасніший метод чи алгоритм практично втрачає цінність, якщо він не доведений до реалізації у вигляді відповідного програмного застосунку усі теоретичні положення в посібнику розглядаються через призму їхньої програмної реалізації, Основним інструментом обрано мову програмування *Python* та її екосистему бібліотек для статистичного аналізу, обробки даних і машинного

навчання. Хоча з об'єктивних причин посібник не може містити лістинги всіх програмних реалізацій, автори прагнули надати достатньо матеріалу для того, щоб здобувачі могли самостійно створювати, адаптувати та вдосконалювати відповідні програмні рішення.

Важливою складовою підготовки сучасного фахівця з комп'ютерних наук є володіння англійською мовою. Значна частина наукової, технічної та професійної інформації публікується саме англійською, тому знання відповідної термінології є необхідною умовою ефективної роботи з фаховими джерелами. З цією метою більшість професійних термінів у посібнику подається разом із їхніми англомовними відповідниками, що має сприяти розвитку навичок самостійної роботи з англомовною літературою та допоможе впевнено орієнтуватися в сучасному міжнародному професійному середовищі.

Професійна діяльність сучасного спеціаліста у сфері інформаційних технологій вимагає постійного саморозвитку та вміння працювати з науковою й технічною літературою. Тому у списку рекомендованих джерел поряд із базовими підручниками наведено сучасні наукові статті, інтернет-ресурси та інші матеріали, що відображають актуальний стан розвитку методів аналізу даних, машинного навчання та кібербезпеки.

Автори сподіваються, що цей навчальний посібник стане для читача не лише джерелом базових знань із системного аналізу, статистики та методів аналізу даних, а й основою для подальшого самостійного професійного розвитку у сфері інформаційних технологій, моделювання систем, машинного навчання, сучасних технологій штучного інтелекту, побудови інтелектуальних систем у сфері кібербезпеки й інших прикладних галузях. *Матеріал посібника покликаний сформувати у здобувачів цілісне бачення процесу дослідження даних – від постановки задачі та побудови моделі до програмної реалізації методів аналізу й інтерпретації отриманих результатів.* Подальше становлення майбутніх фахівців має ґрунтуватися на поглибленні теоретичної підготовки, розвитку алгоритмічного та системного мислення, удосконаленні навичок програмування, а також постійному освоєнні нових інструментальних засобів аналізу даних, моделювання та машинного навчання. Лише поєднання фундаментальних знань, практичних навичок і готовності до безперервного самонавчання дозволить ефективно працювати в умовах стрімкого розвитку сучасних інформаційних технологій та зростання складності задач забезпечення кібербезпеки.

1. СИСТЕМНИЙ АНАЛІЗ І МОДЕЛЮВАННЯ. МЕТОДОЛОГІЧНІ ОСНОВИ

1.1. Основні положення сучасного системного аналізу

Сучасний світ характеризується зростаючою складністю явищ і процесів, з яким людство стикається в природі, суспільстві і техніці. Економічні системи, інформаційні мережі, соціальні інституції, технологічні комплекси - всі ці об'єкти є надзвичайно складними конструкціями, їх поведінку неможливо вивчити, розглядаючи лише окремі частини. Саме тут на перший план виходить системний аналіз - міждисциплінарна наукова дисципліна, яка надає методологічний інструментарій для вивчення складних об’єктів у їх цілому та взаємозв’язку.

Системний аналіз як наукова дисципліна [1], [2], [3], [4] сформувався на стику декількох напрямів науки: загальної теорії систем, кібернетики, математичного програмування, теорії прийняття рішень. Його розвиток був мотивований практичними необхідностями: керування великими військовими та космічними проектами, оптимізацією виробничих процесів, плануванням складних соціально-економічних систем тощо. З часом системний аналіз став самостійною дисципліною зі своїм категоріально-понятійним апаратом, методами, сферою впровадження. Актуальність системного аналізу в наш час не те що не слабшає, а навпаки - постійно зростає. Цифрова трансформація економіки, прогрес штучного інтелекту, глобальні кліматичні виклики, ускладнення соціальних структур - усе це ставить перед дослідниками і практиками завдання, для вирішення яких потрібно системне мислення та відповідні аналітичні інструменти.

1.1.1. Визначення поняття «система»

Ключовим поняттям системного аналізу є поняття «система». Незважаючи на широке вживання цього терміну у повсякденному житті та різних наукових дисциплінах, його точне визначення є предметом наукових дискусій. У загальному сенсі ***система*** - *це сукупність взаємопов'язаних і взаємодіючих елементів, що утворюють цілісне утворення з певними властивостями, які не зводяться до властивостей окремих елементів.*

Будь-яка система характеризується певними ознаками:

– *цілісністю* - система має властивості, яких немає у жодного з її елементів окремо (емерджентність);

- *структурованістю* - елементи системи пов'язані між собою певними відношеннями та утворюють структуру;
- *ієрархічністю* - кожна система є підсистемою системи більш високого рівня і сама у своєму складі містить підсистеми;
- *взаємодією з зовнішнім середовищем* - система відокремлена від навколишнього середовища певною межею та обмінюється з ним речовиною, енергією або інформацією.

1.1.2. Класифікація систем

Системи класифікують за різними ознаками. *За природою елементів* розрізняють фізичні, біологічні, соціальні, технічні та змішані (соціотехнічні) системи. *За характером взаємодії з навколишнім середовищем* виділяють відкриті системи, що обмінюються речовиною, енергією та інформацією із зовнішнім середовищем, і закриті системи, що є ізольованими від нього. *За ступенем організованості* розрізняють добре організовані системи (піддаються точному математичному опису), погано організовані або дифузні системи (важко описати повністю), і системи, що самоорганізуються.

За характером поведінки виділяють *детерміновані системи*, в яких наступний стан однозначно визначається попереднім, і *стохастичні (імовірнісні) системи*, поведінка яких описується статистичними закономірностями. Системи також поділяють за розміром і складністю. Окрему категорію утворюють цілеспрямовані системи, здатні до вироблення та реалізації цілей - передусім соціальні та соціотехнічні системи.

1.1.3. Поняття мети та критерію в системному аналізі

Центральним поняттям системного аналізу є поняття цілі. ***Мета*** *- це бажаний стан системи або бажаний результат її функціонування, якого прагне досягти суб'єкт управління.* У системному аналізі принципово важливим є чітке та операціональне визначення цілей, без якого неможливо оцінити альтернативні варіанти рішень.

У системному аналізі ***критерій*** - це формалізоване правило, показник або міра, за допомогою яких оцінюють стан, ефективність, якість функціонування чи ступінь досягнення мети системи. Критерій використовується для порівняння альтернатив, прийняття рішень та вибору найкращого варіанта функціонування

або розвитку системи. Іншими словами, критерій відповідає на питання: «За якою ознакою ми будемо оцінювати систему або результати її роботи?»

Вибір критеріїв є одним із найвідповідальніших кроків у системному аналізі, оскільки від нього безпосередньо залежать результати порівняння альтернатив. Складність полягає в тому, що реальні системи, як правило, є *багатоцільовими*, тобто мають кілька цілей, які подекуди суперечать одна одній. У таких випадках застосовують методи багатокритеріальної оптимізації, лексикографічного упорядкування цілей, методи компромісу тощо.

Дерево цілей - важливий інструмент структурування цілей складної системи. Воно *являє собою ієрархічну структуру, в якій генеральна мета розкладається на підцілі першого рівня, ті, у свою чергу, - на підцілі другого рівня і так далі до рівня конкретних заходів і дій.* Побудова дерева цілей дозволяє виявити всі необхідні умови досягнення генеральної мети та розподілити відповідальність між різними підсистемами.

1.1.4. Системний підхід як методологічна основа дослідження об’єктів реального світу

Системний підхід *- це методологічний напрям, що передбачає розгляд будь-якого об'єкта дослідження як системи, тобто як цілісної множини взаємопов'язаних елементів*. На відміну від редукціонізму, що намагається пояснити ціле через властивості його частин, системний підхід підкреслює, що ціле є більшим від суми своїх частин і має власні, емерджентні властивості. *Якщо системний аналіз є загальною методологічною позицією, то системний підхід це сукупність методів, процедур і технік практичного вирішення проблем на основі системного підходу.*

Предметом системного аналізу є складні системи та проблемні ситуації, що виникають у процесі їх функціонування та розвитку. Системний аналіз спрямований на вирішення так званих «погано структурованих» або «слабко структурованих» проблем - задач, у яких цілі не повністю визначені, альтернативні варіанти рішень не є очевидними, а критерії оцінки неоднозначними та часто суперечливими.

Основні завдання системного аналізу включають:

- виявлення і чітке формулювання задачі аналізу, який планується провести;
- визначення системи та її меж;

- структурування системи (декомпозицію на підсистеми і елементи);
- виявлення зв'язків між елементами та підсистемами;
- побудову моделей системи;
- аналіз альтернативних варіантів рішень;
- оцінку варіантів за обраними критеріями;
- вироблення рекомендацій щодо вирішення проблеми.

Коли говорять про перший з перелічених етапів, а саме формулювання задачі, як правило мається на увазі виявлення проблеми, заради якої проводиться аналіз, її «симптомів» та прояв, її актуальність (наприклад, для розуміння системи вищого рівня ієрархії). Тільки зрозумівши (і описавши) це можливий подальший перехід до безпосереднього аналізу, формулювання та опису цілей системи, виявлення її структури та функцій, аналіз та виявлення можливих дефектів у структурі чи функціях, виявлення потреб у ресурсах, що необхідні для існування та/або функціонування системи.

Системний аналіз ґрунтується на сукупності принципів, що визначають загальну логіку дослідження та вимоги до його проведення. Ці принципи виступають нормативними правилами системно-аналітичної діяльності:

- принцип кінцевої мети стверджує, що абсолютний пріоритет належить глобальній меті функціонування системи. Будь-які локальні рішення мають оцінюватись з точки зору їх внеску у досягнення глобальної мети. Підвищення ефективності підсистеми неприпустиме, якщо воно веде до погіршення функціонування системи в цілому.
- принцип вимірювання вимагає кількісного або якісного оцінювання характеристик системи, її стану та ступеня досягнення цілей. Без вимірювання неможливо порівняти альтернативні рішення та обрати кращий варіант. При цьому система вимірювань має бути узгоджена з цілями дослідження.
- принцип еквіфінальності означає, що система може досягати одного й того самого кінцевого стану з різних початкових умов і різними шляхами. Це принципово відрізняє відкриті системи від закритих і означає, що пошук єдиного «правильного» шляху до мети є хибним підходом - слід розглядати множину альтернативних траєкторій.

- принцип єдності передбачає спільний розгляд системи як цілого і як сукупності частин (елементів). Ні суто «цілісний», ні суто «елементний» підхід самостійно не є достатнім - необхідне їх поєднання.
- принцип зв'язності вимагає розгляду будь-якого елемента системи разом із його зв'язками з іншими елементами та зовнішнім середовищем. Відокремлення елемента від системи і дослідження його у ізоляції дає спотворені результати.
- принцип модульної побудови стверджує, що при аналізі та синтезі систем доцільно виявляти відносно незалежні підсистеми (модулі) та досліджувати їх окремо. Це спрощує аналіз складних систем і дозволяє використовувати спеціалізовані методи для кожної підсистеми.
- принцип ітеративності відображає циклічний характер системного аналізу. Дослідження рідко виконується за один прохід від постановки проблеми до кінцевого рішення - як правило, необхідні повернення до попередніх етапів, уточнення формулювань, перегляд критеріїв і альтернатив. Кожна ітерація поглиблює розуміння системи та проблеми.
- принцип декомпозиції та синтезу передбачає аналітичне розкладання системи на частини (декомпозицію) з наступним відтворенням цілісної картини (синтезом). Декомпозиція дозволяє подолати складність системи, синтез - не втратити цілісність.

1.1.5. Методи та інструменти системного аналізу

Методи системного аналізу утворюють розгалужену систему, що охоплює як якісні (евристичні), так і кількісні (формалізовані) підходи. У загальному вигляді їх можна поділити на три великі групи: методи структуризації проблеми та цілей; методи отримання та обробки експертних оцінок; формалізовані методи (математичні, оптимізаційні, імітаційні).

До методів структуризації цілей можна віднести:

- *метод дерева цілей*, являє собою граф, що відображає ієрархію цілей системи. Корінь дерева - генеральна мета; вузли - підцілі різних рівнів; листя - конкретні завдання та заходи. Дерево проблем будується аналогічно і відображає причинно-наслідкові зв'язки між проблемами системи.

– *морфологічний аналіз* за допомогою матриці взаємовідповідності полягає у систематичному переборі всіх можливих комбінацій значень ключових параметрів системи. Морфологічна матриця дозволяє виявити всі теоретично можливі рішення задачі та відібрати найбільш перспективні.
– *SWOT-аналіз* - метод структурованої оцінки системи через виявлення її сильних сторін (англ. *Strengths*), слабких сторін (англ. *Weaknesses*), зовнішніх можливостей (англ. *Opportunities*) і загроз (англ. *Threats*). Незважаючи на простоту, SWOT-аналіз є ефективним інструментом первинного структурування проблемної ситуації та вироблення стратегій.
– *діаграми причинно-наслідкових зв'язків* призначені для систематичного виявлення всіх можливих причин проблеми та їх класифікації. Метод широко застосовується в управлінні якістю та аналізі відмов.

Серед методів роботи з експертами варто зазначити:

– *метод Дельфі* - ітераційна процедура узгодження думок групи експертів. Кожен експерт анонімно відповідає на запитання анкети; результати узагальнюються та надсилаються всім учасникам для перегляду своїх оцінок з урахуванням думок колег. Процедура повторюється декілька разів (зазвичай 3–5 раундів) до досягнення прийнятного рівня узгодженості. Анонімність запобігає домінуванню авторитетів, а ітераційність дозволяє уточнювати судження.
– *метод мозкового штурму* (англ. *Brainstorming*). Суть методу - в групі відбуваються дві фази: спочатку вільне висування ідей без будь-якої критики, потім - критичний аналіз та відбір найбільш продуктивних ідей. Метод ефективний для генерування нестандартних рішень і виходу за межі стереотипного мислення.
– *метод сценаріїв* - якісний метод прогнозування, що полягає у побудові кількох альтернативних описів можливого майбутнього стану системи (сценаріїв). Кожен сценарій є внутрішньо узгодженою картиною можливого розвитку подій при певному поєднанні ключових факторів. Аналіз сценаріїв дозволяє розробити стратегії, адекватні широкому спектру можливих майбутніх умов.

– *метод аналізу ієрархій* призначений для вирішення задач багатокритеріального вибору. Метод передбачає побудову ієрархічної структури задачі (мета - критерії - альтернативи) і попарне порівняння елементів кожного рівня за відповідними критеріями. За допомогою матриць попарних порівнянь і власних векторів обчислюються пріоритети альтернатив.

1.1.6. Формалізовані методи та їх місце у системному аналізі

Математичне програмування, включаючи лінійне, нелінійне, динамічне, дозволяє знаходити оптимальні рішення задач з чітко сформульованою цільовою функцією та системою обмежень. В системному аналізі вони застосовуються для оптимізації розподілу ресурсів, планування виробництва, маршрутизації тощо.

Теорія графів надає потужний формальний апарат для аналізу структури мереж, потоків, зв'язків між елементами системи. Мережеве планування (метод критичного шляху - CPM, метод PERT) засноване на теорії графів і дозволяє оптимізувати управління складними проектами.

Теорія черг (масового обслуговування) вивчає системи, в яких вимоги (заявки) надходять на обслуговування з певною інтенсивністю, а обслуговуючі пристрої мають обмежену продуктивність. Теорія черг дозволяє розраховувати характеристики таких систем (довжину черги, час очікування, завантаженість обслуговуючих пристроїв) і оптимізувати їх параметри.

Теорія ігор вивчає ситуації прийняття рішень в умовах конфлікту, тобто коли результат дій кожного учасника залежить від дій інших. Концепція рівноваги Неша, кооперативні та некооперативні ігри, ігри з нульовою та ненульовою сумою - ці поняття активно використовуються при аналізі конкурентних і кооперативних взаємодій у соціально-економічних системах.

Як бачимо, системний аналіз спирається в методології своїх досліджень на різноманітні та абсолютно різнопланові наукові дисципліни, підходи, методи і алгоритми. В свою чергу, використання вказаних інструментів не розрізнено, а в рамках загального системного підходу дозволяє в повній мірі використати їх потужність, в кінцевому разі вирішити ті задачі і проблеми, які неможливо було виконати, використовуючи ці інструменти ізольовано.

1.1.7. Моделювання в системному аналізі

Моделювання займає центральне місце в системному аналізі. ***Модель*** - *це спрощене відображення реального об'єкта або явища, що зберігає суттєві риси оригіналу і дозволяє досліджувати його властивості та поведінку*. Оскільки безпосередній «натурний» експеримент зі складними реальними системами часто є неможливим, надто дорогим або небезпечним, моделювання стає незамінним інструментом пізнання.

Будь-яка модель є компромісом між точністю відображення дійсності та простотою, зрозумілістю і керованістю. Надмірно детальна модель може бути так само важко досліджуваною, як і сам оригінал. Модель завжди відображає лише ті аспекти реальної системи, що є суттєвими з точки зору цілей дослідження.

Різноманіття моделей, які використовуються в сучасних дослідженнях, надзвичайно широке. Варто коротко згадати такі моделі, як

- *концептуальні моделі* - описові моделі у вигляді словесних описів, схем, діаграм, що відображають загальну структуру системи та взаємозв'язки між її елементами без точного кількісного опису. Прикладами є організаційні схеми, діаграми потоків даних, концептуальні схеми баз даних.
- *математичні моделі* - формалізовані описи системи мовою математики (рівняннями, нерівностями, функціями). За характером математичного апарату розрізняють аналітичні моделі (система рівнянь, розв'язок яких описує поведінку системи) та чисельні моделі (наближені методи розв'язання рівнянь). За врахуванням часу розрізняють статичні (описують систему в певний момент) і динамічні (описують зміни у часі) моделі.
- *імітаційні моделі* відтворюють процеси, що відбуваються у системі, шляхом їх «програвання» у часі. Імітаційне моделювання особливо корисне для складних систем, що важко піддаються аналітичному опису. Методи Монте-Карло, дискретно-подієве моделювання, агентне моделювання - основні різновиди імітаційного моделювання.

Кожен з згаданих класів моделей розпадається на незліченну кількість окремих підкласів, навіть перелічити які тут нема можливості. Це дає змогу досліднику спробувати різностороннє підійти до процесу моделювання

враховуючи наявний арсенал засобів, кожен з яких має свої сильні сторони, але й водночас обмеження.

Обов'язковими етапами будь-якого дослідження з аналізу систем та їх моделювання є етапи верифікації та валідації побудованих моделей систем.

Верифікація моделі - перевірка коректності її формальної реалізації (наприклад, програмного коду). Це процес перевірки того, чи правильно побудована модель, чи відповідає вона початковим вимогам, специфікаціям та чи коректно працюють її алгоритми.

Валідація моделі - перевірка адекватності моделі реальній системі, тобто того, чи дійсно модель відображає ті аспекти системи, для дослідження яких вона призначена. Іншими словами, якщо валідація відповідає на питання «Чи правильну модель побудовано?» то верифікація відповідає на питання «Чи правильно побудовано обрану модель?»

Методи валідації включають порівняння результатів моделювання з реальними даними (якщо такі є), перевірку поведінки моделі в екстремальних умовах, оцінку чутливості результатів до змін параметрів (аналіз чутливості). Модель вважається валідною, якщо вона достатньо точно відтворює поведінку реальної системи у тих умовах і для тих цілей, для яких вона призначена.

1.1.8. Застосування системного аналізу в прикладних галузях

Треба завжди пам'ятати, що ані системний аналіз взагалі, ані моделювання систем зокрема не робляться заради самого аналізу як самоцілі. Вони підпорядковуються загальній, глобальній меті - використати отримані результати для вирішення прикладних задач, які і поставлені перед дослідником. Саме заради цього вирішуються всі описані задачі та виконуються всі проміжні етапи, включаючи дослідження (виявлення) законів, за якими функціонує система, що досліджується, опис системи, побудова моделі, що описують поведінку системи, перевірку отриманих описів щодо відповідності та адекватності (включаючи верифікацію та валідацію). Без заключного кроку – прийняття рішень на основі побудованої моделі системи саме дослідження втрачає будь який сенс.

Системний аналіз знайшов широке застосування у найрізноманітніших прикладних галузях, при рішенні нескінченно широкого кола задач. Так у державному управлінні аналіз публічної політики спирається на методологію системного аналізу для оцінки ефективності державних програм та

альтернативних варіантів регуляторних рішень. Методи аналізу витрат та вигод дозволяють порівнювати суспільні переваги та витрати різних варіантів державних рішень. Планування розвитку регіонів і міст також широко використовує інструменти системного аналізу. Комплексні плани просторового розвитку, транспортні схеми, плани розвитку комунальної інфраструктури розробляються на основі системного аналізу потреб населення, ресурсних обмежень і довгострокових тенденцій розвитку. У сфері стратегічного менеджменту системний аналіз є методологічною основою для розробки корпоративних стратегій. Стратегічне планування, конкурентний аналіз, управління портфелем проектів, реінжиніринг бізнес-процесів - всі ці практики спираються на системне мислення та відповідні аналітичні інструменти. В управлінні ланцюгами постачань системний аналіз застосовується для оптимізації логістичних мереж, управління запасами, мінімізації операційних витрат з урахуванням вимог щодо рівня обслуговування клієнтів. Системний підхід дозволяє враховувати складні взаємозалежності між постачальниками, виробниками, дистриб'юторами та кінцевими споживачами.

Екологічні системи є класичним прикладом складних систем із численними зворотними зв'язками, нелінійними ефектами та невизначеністю. Системний аналіз надає інструменти для дослідження взаємодій у екосистемах, оцінки впливу антропогенної діяльності на довкілля та розробки стратегій сталого управління природними ресурсами. А також - вирішувати задачі з прогнозування погоди, що безпосередньо впливає, серед іншого, на об’єм та якість врожаїв сільськогосподарського виробництва.

Системи охорони здоров'я є складними соціотехнічними системами, що об'єднують медичний персонал, установи, технології, пацієнтів та механізми фінансування. Системний аналіз застосовується для оптимізації організації медичної допомоги, розподілу ресурсів між рівнями медичної допомоги, оцінки ефективності медичних технологій та моделювання поширення захворювань.

Епідеміологічне моделювання - окремий важливий напрям, що набув особливої актуальності під час пандемії COVID-19. Математичні моделі поширення інфекцій дозволяють прогнозувати динаміку епідемій та оцінювати ефективність різних заходів стримування.

Системна інженерія (англ. *Systems Engineering*) - прикладна дисципліна, що застосовує системний підхід до проектування, розробки, управління та утилізації

складних технічних систем. Вимоги до великих інженерних проектів (космічних, авіаційних, оборонних, телекомунікаційних) формулюються, аналізуються і узгоджуються з використанням формальних методів системної інженерії.

Управління ризиками в інженерних проектах спирається на системний аналіз для виявлення потенційних відмов, оцінки їх ймовірності та наслідків, розробки заходів з підвищення надійності. Аналіз видів і наслідків відмов, дерева відмов, аналіз дерев подій- класичні інструменти системного аналізу надійності.

1.1.9. Висновки

Системний аналіз - це потужна міждисциплінарна методологія, що пройшла шлях від прикладного інструменту до універсального фреймворку для дослідження і управління складними об'єктами найрізноманітнішої природи. Його теоретичні засади, що сягають загальної теорії систем, кібернетики і дослідження операцій, утворюють стрункy і внутрішньо узгоджену систему понять, принципів і методів.

Методологічна цінність системного аналізу полягає передусім у тому, що він надає мову і засоби для опису та дослідження складних цілісних утворень - систем - у їх взаємодії з навколишнім середовищем. Принципи системного аналізу - цілісності, ієрархічності, еквіфінальності, зв'язності - формують загальну логіку системно-орієнтованого мислення, що є надзвичайно цінним у сучасному складному світі.

Багатство методів системного аналізу - від якісних (дерево цілей, морфологічний аналіз, метод Дельфі, сценарний аналіз) до кількісних (математичне програмування, теорія ігор, імітаційне моделювання, системна динаміка) - забезпечує гнучкість і адаптивність цієї методології до задач різної природи та ступеня формалізованості.

Опанування системного аналізу є необхідною умовою для ефективної діяльності сучасного фахівця у будь-якій галузі - від державного управління до інженерії, від бізнесу до охорони здоров'я. Системне мислення дозволяє бачити ціле за окремими деталями, розуміти зворотні зв'язки та нелінійні ефекти, виявляти приховані можливості та ризики, приймати обґрунтовані рішення в умовах складності та невизначеності.

1.2. Системний аналіз у контексті забезпечення кібербезпеки

1.2.1. Система забезпечення кібербезпеки як об’єкт системного аналізу

Сучасна кібербезпека – це не сукупність окремих програмних, апаратних, організаційних засобів, не набір антивірусів чи фаєрволів, а складна динамічна система, де ефективність цілого залежить від функціонування та взаємодіє кожного окремого елемента. Водночас - це складна соціотехнічна система, в якій задіяні люди, технології, процеси, ресурси та зовнішнє середовище [1], [2], [5], [6].

Системний підхід до кібербезпеки ґрунтується на усвідомленні її емерджентної природи: рівень захищеності формується лише за умови скоординованої роботи компонентів мережевої інфраструктури, програмного забезпечення та користувачів. У цьому контексті об’єктом аналізу мають бути не тільки і навіть не стільки окремі елементи системи, але взаємозв’язки між ними [3]. Вразливість у системі - це часто не «дірка» в коді, а непередбачена взаємодія між двома цілком безпечними на перший погляд функціями.

Системи кіберзахисту є адаптивними. Загрози постійно еволюціонують, що змушує систему захисту перебувати у стані безперервного самовдосконалення. За таких умов ефективнішими є не статичні захисні бар’єри, а архітектури, здатні виявляти аномалії, реагувати на інциденти та відновлювати роботу після збоїв.

Забезпечення кібербезпеки можна розглядати як задачу управління складністю. Вона потребує переходу від фрагментарного мислення до цілісного бачення цифрового середовища та його взаємозв’язків.

Важливою, фундаментальною властивістю систем кіберзахисту є неможливість досягнення *абсолютного* захисту інформаційних систем. Це зумовлено складністю сучасних кіберсередовищ, динамічністю загроз, наявністю людського фактору та принциповою неповнотою інформації про майбутні атаки. Будь-яка система містить вразливості, частина з яких є невідомими або виявляються в процесі експлуатації, що робить повне усунення ризиків технічно та економічно недосяжним. За цих умов забезпечення кібербезпеки переходить від ідеї абсолютного захисту до парадигми керування ризиками [7], [8], [9]. *Метою стає не усунення всіх загроз, а зниження ймовірності їх реалізації та мінімізація можливих наслідків.* Керування ризиками передбачає систематичну ідентифікацію активів, загроз і вразливостей, кількісну або якісну оцінку ризиків, визначення прийнятного рівня ризику та вибір адекватних контрзаходів з

урахуванням ресурсних обмежень. Тобто - *глибокого, аналітичного, комплексного аналізу даних*, які різнобічно характеризують стан системи захисту та її розвиток у часі.

Традиційна модель безпеки роками була реактивною: стався інцидент - шукаємо рішення. Іншими словами, реактивний підхід базується на відповіді на вже існуючі загрози. Його основні інструменти - системи захисту (наприклад, на основі сигнатур) та запровадження захисних дій після того, як вразливість була використана зловмисником. Однак при цьому виникає проблема часового розриву між атакою та реакцією. Зловмисник завжди на крок попереду, а захист працює в режимі постійного стресу та дефіциту ресурсів.

Системний аналіз, перетворюючи забезпечення безпеки на керований, прогнозований процес, дозволяє докорінно змінити підхід. *Проактивність* означає усунення причин потенційних проблем ще до того, як вони будуть використані, зменшення ймовірності проблем ще до їх реалізації. Це досягається завдяки таким механізмам:

- Моделювання загроз: замість очікування атаки, аналітик будує *модель загроз*. Це дає змогу виявляти логічні слабкі місця системи на етапі проектування.
- Виявлення прихованих залежностей: системний аналіз допомагає зрозуміти, як зміна в одному модулі (наприклад, оновлення API) *впливає* на безпеку всієї мережі.
- Аналіз аномалій замість сигнатур: проактивний захист використовує системний *моніторинг для визначення «нормального стану» системи*. Будь-яке відхилення від цієї норми розцінюється як потенційна загроза, що дозволяє зупиняти атаку ще на етапі розвідки.
- Керування ризиками: системний підхід дозволяє *кількісно оцінити ризики*. Це допомагає інвестувати ресурси в захист критичних вузлів, а не намагатися «захистити все і нічого одночасно».

Перехід до проактивності - це перехід від стратегії «ізольованих захисних бар'єрів» до стратегії «стійких систем». Системний аналіз дає можливість не просто відбивати удари, а створювати таке середовище, де сама спроба атаки стає занадто дорогою або технічно неможливою для зловмисника.

Конкретизуючи основні поняття системного аналізу у контексті систем кіберзахисту маємо ще раз підкреслити, що в даному разі система лише зрідка

обмежується виключно технічними компонентами. До них належать апаратні засоби, програмне забезпечення, мережі як такі та мережеві пристрої зокрема, програмні сервіси або навіть окремі процеси, користувачі, адміністратори, політики безпеки та зовнішнє середовище. Важливим етапом аналізу є визначення меж системи [5], [10]. Наприклад, при аналізі безпеки корпоративної мережі межі можуть включати внутрішню інфраструктуру, віддалених користувачів, хмарні сервіси та партнерські системи. Неправильне визначення меж призводить до ігнорування важливих векторів атак, таких як компрометація облікових записів сторонніх підрядників або зловживання легітимним доступом. В загальному випадку межі такої системи є аналітичним вибором, що залежить від цілей дослідження, а не фіксованою технічною характеристикою.

Ключовим для системного аналізу є не тільки сам перелік елементів, що входять до складу системи, а *характер зв'язків* між ними. Саме взаємодія елементів формує властивості системи, зокрема її вразливості. Атаки часто є наслідком не відмови окремого елемента, а некоректної або непередбаченої взаємодії між елементами системи. Наприклад, окремо взятий сервер може бути налаштованим коректно, але його взаємодія з іншими компонентами через мережу або через спільні облікові дані створює нові ризики.

Представлення системи у *вигляді «вхід – стан – вихід»* при дослідженні систем кібербезпеки призводить до наступної деталізації:

- *Входи системи* включають мережевий трафік, дії користувачів, оновлення програмного забезпечення, адміністративні команди.
- *Стан системи* визначається конфігурацією, активними процесами, рівнем завантаження ресурсів, правами доступу та поточними з'єднаннями.
- *Виходи системи* можуть проявлятися у вигляді сервісів, відповідей на запити, журналів подій або побічних ефектів.

Важливо також взяти до уваги, що небезпечний вихід в системі кіберзахисту не завжди є прямим наслідком одного входу. Часто він виникає внаслідок накопичення змін стану системи, що ускладнює виявлення причин інциденту.

Інформаційні системи, що підлягають захисту, практично завжди є відкритими системами, тобто такими, що взаємодіють із зовнішнім середовищем. Це середовище включає інтернет, користувачів, сторонні сервіси та потенційних

зловмисників. Відкритість системи означає неможливість повного контролю над усіма впливами, що є принциповим обмеженням кібербезпеки. Системний аналіз дозволяє не усувати цю відкритість, а враховувати її, моделюючи можливі сценарії взаємодії із середовищем та оцінюючи їх ризики.

Одним із *ключових інструментів системного аналізу є декомпозиція*, тобто поділ складної системи на підсистеми. У кібербезпеці це може бути поділ на мережевий рівень, рівень хостів, рівень додатків і рівень користувачів [3], [10]. Ієрархічна структура дозволяє аналізувати систему на різних рівнях абстракції. Наприклад, на стратегічному рівні розглядається архітектура безпеки організації, тоді як на операційному - конкретні події, що відображаються в журналах (лог-файлах). Важливою особливістю є те, що проблеми на нижчих рівнях можуть проявлятися у вигляді ризиків на вищих рівнях.

Зворотні зв'язки є фундаментальним поняттям системного аналізу. ***Зворотний зв'язок*** - це механізм впливу *результатів функціонування системи на її подальшу поведінку або стан*. Інакше кажучи, система отримує інформацію про результати власної роботи та використовує її для коригування свого функціонування.

У системах кібербезпеки зворотні зв'язки проявляються, наприклад, у вигляді автоматичних реакцій на інциденти, оновлення правил виявлення або навчання моделей машинного навчання на нових даних. Негативні зворотні зв'язки спрямовані на стабілізацію системи (наприклад, блокування підозрілої активності), тоді як позитивні зворотні зв'язки можуть призводити до ескалації проблем (наприклад, надмірні хибні спрацювання, що перевантажують аналітиків, відволікаючи їх від боротьби з реальними загрозами).

Одним з проявів згаданої вище *емерджентності*, характерним для систем кібербезпеки, є складна багатокрокова атака, що виникає внаслідок поєднання легітимних дій, кожна з яких окремо не є шкідливою. Це ще раз підкреслює обмеженість суто локального аналізу та необхідність системного підходу.

Таким чином, основні поняття системного аналізу - система, елементи, зв'язки, межі, стани, ієрархія та зворотні зв'язки - у задачах кібербезпеки набувають конкретного практичного змісту. Вони дозволяють перейти від аналізу окремих технічних засобів до розуміння систем забезпечення кібербезпеки як складної адаптивної системи, у якій ефективний захист можливий лише за умови цілісного, багаторівневого підходу, орієнтованого на дані.

1.2.2. Моделювання в системному аналізі кібербезпеки

Моделювання є одним з інструментів системного аналізу в задачах забезпечення кібербезпеки, оскільки дозволяє формалізувати уявлення про складні кіберсистеми, досліджувати їхню поведінку та аналізувати можливі сценарії розвитку подій без втручання в реальну інфраструктуру. Моделі використовуються для опису структури системи, потоків інформації, взаємодії компонентів і дій потенційного порушника [5].

У системному аналізі кібербезпеки застосовуються різні типи моделей. *Структурні моделі* описують компоненти системи та зв'язки між ними, наприклад у вигляді графів мережевої топології або моделей довіри між вузлами. Такі моделі дозволяють виявляти критичні елементи, точки концентрації ризику та можливі шляхи поширення атак. *Функціональні моделі* зосереджуються на процесах і механізмах обробки інформації, відображаючи логіку роботи сервісів і захисних засобів. Широкого застосування набули *ймовірнісні та статистичні моделі*, які використовуються для оцінки ризиків, імовірності реалізації атак і очікуваних наслідків інцидентів. Такі моделі дозволяють працювати з невизначеністю, що є характерною рисою кібербезпеки, та підтримують прийняття рішень у межах керування ризиками. Вони також лежать в основі методів виявлення аномалій і прогнозування.

Окрему групу становлять *імітаційні моделі*, які відтворюють динаміку системи в часі та дозволяють аналізувати складні сценарії атак і захисту. Імітаційне моделювання дає змогу досліджувати наслідки різних стратегій захисту, оцінювати ефективність контрзаходів і виявляти емерджентні ефекти, що виникають унаслідок взаємодії компонентів системи.

Важливою особливістю моделювання в кібербезпеці є його тісний зв'язок з емпіричними даними. Моделі потребують постійної валідації та уточнення на основі даних моніторингу, журналів подій і результатів інцидентів. Без зворотного зв'язку моделювання втрачає практичну цінність і перетворюється на абстрактний опис.

Таким чином, моделювання в системному аналізі кібербезпеки слугує засобом інтеграції знань про структуру системи, її поведінку та загрози. Воно забезпечує основу для *прогнозування, оцінки ризиків і обґрунтованого вибору захисних рішень*, за умови усвідомлення обмежень моделей і необхідності їх постійного оновлення.

1.2.3. Роль аналітики та машинного навчання в задачах забезпечення кібербезпеки

Аналітика даних і методи ***машинного навчання*** (МН) відіграють ключову роль у сучасних системах кібербезпеки, оскільки дозволяють працювати з великими обсягами різнорідних даних, виявляти приховані закономірності та підтримувати прийняття рішень в умовах невизначеності. У сучасному, системному підході до кібербезпеки аналітика та МН розглядаються не як ізольовані інструменти, а як функціональні компоненти, інтегровані в загальну архітектуру [5], [11], [12].

Аналітика забезпечує *спостережуваність* кіберсистеми, *тобто можливість робити висновки про її внутрішній стан на основі доступних даних.* У практичних задачах кібербезпеки це означає аналіз мережевого трафіку, журналів подій, телеметрії кінцевих точок, поведінкових характеристик користувачів, тощо. Без аналітичної обробки ці дані залишаються фрагментованими та не дозволяють сформувати цілісне уявлення про стан системи. Аналітика виконує функцію перетворення *«сирих» спостережень* (англ. - *Row Data*), зібраних безпосередньо з джерела без будь-якої трансформації, очищення чи агрегації, на інформацію, придатну для подальшого використання, моделювання та управління. Наприклад, агрегація NetFlow-даних дозволяє перейти від даних, що описують окремі з'єднання до моделей мережевої поведінки, а від аналізу журналів автентифікації - до профілів користувацької активності.

Машинне навчання розширює можливості аналітики, дозволяючи *автоматично виявляти складні закономірності* у даних, які важко або неможливо формалізувати вручну. Важливо розглядати МН не як автономний механізм прийняття рішень, а як елемент більшої системи. У кібербезпеці МН широко застосовується для виявлення аномалій, класифікації подій, прогнозування ризиків і підтримки автоматизованого реагування. Одним із найпоширеніших застосувань машинного навчання в кібербезпеці є ***виявлення аномалій***. МН-моделі фактично формують модель нормального функціонування системи на основі історичних даних. Однак, оскільки кіберсистеми є динамічними, ця «норма» постійно змінюється. Системний аналіз дозволяє враховувати цю динаміку, поєднуючи результати МН із експертними знаннями та правилами бізнес-логіки. Треба окремо зазначити, що з системної точки зору

аномалія не є синонімом атаки, а лише сигналом про відхилення від очікуваного стану системи. Інтерпретація таких сигналів потребує урахування контексту, зворотних зв'язків і можливих помилок вимірювання.

1.2.4. Дані у задачах забезпечення кібербезпеки

Дані виступають фундаментальним базисом, джерелом для побудови моделей у межах системного аналізу. Саме дані *відображають* реальний стан системи, її динаміку та реакцію на зовнішні впливи. Без опори на емпіричні дані системний аналіз залишається теоретичним і не дозволяє перевіряти гіпотези щодо поведінки кіберсистеми [13].

Характерною рисою систем кібербезпеки є багатоджерельність і різнорідність даних. Інформація надходить з різних складових системи, відображаючи різні аспекти її функціонування. Мережеві дані, журнали подій та дані кінцевих точок не лише відрізняються за форматом і масштабом, але й описують систему з різних точок зору. Системний аналіз полягає не просто в обробці кожного типу даних окремо, а *в їх інтеграції в єдину аналітичну картину*.

У практичних задачах кібербезпеки дані надходять із різних рівнів системи.

Мережеві дані (NetFlow, sFlow, IPFIX) описують потоки трафіку між вузлами та дозволяють виявляти аномальну комунікаційну поведінку, наприклад сканування портів, DDoS-атаки або нетипові зовнішні з'єднання. Ці дані є агрегованими, але добре масштабуються і відображають глобальну структуру взаємодій у мережі.

Дані систем централізованого журналювання та кореляції подій (SIEM) об'єднують логи з серверів, мережевих пристроїв, додатків і засобів захисту. У системному аналізі SIEM-дані відіграють роль «хроніки станів системи», дозволяючи відстежувати причинно-наслідкові зв'язки між подіями, виявляти складні багатокрокові атаки та оцінювати ефективність політик безпеки.

Дані кінцевих точок (англ. *Endpoint Detection and Response, EDR*) фіксують поведінку процесів, користувачів і файлів на робочих станціях та серверах. Вони дозволяють аналізувати внутрішню динаміку системи, виявляти шкідливу активність, яка не проявляється на мережевому рівні, та будувати детальні поведінкові моделі атак.

Системний аналіз дозволяє інтегрувати ці різнорідні типи даних у єдину модель, де мережеві події, журнали та поведінка кінцевих вузлів розглядаються

як взаємопов’язані елементи. Саме на цьому рівні стає можливим виявлення емерджентних загроз, які не помітні при аналізі кожного джерела окремо.

Важливою умовою аналізу даних у кібербезпеці є робота в стані *невизначеності та наявності шуму*, що суттєво ускладнює як *інтерпретацію спостережень*, так і *прийняття обґрунтованих рішень*. Неповнота даних виникає внаслідок обмежень сенсорів, втрат журналів подій, використання шифрування або навмисного маскування активності зловмисником. Шум у даних може бути проявом як певних технічних причин, так і природної варіативності поведінки користувачів і систем. Легітимні дії можуть виглядати підозрілими, тоді як шкідлива активність може бути схожою на нормальну. Додатковим ускладненням є запізнення даних, коли інформація про події надходить із часовим лагом або асинхронно з різних джерел. Це порушує сприйняття причинно-наслідкових зв’язків та ускладнює реконструкцію сценаріїв атак.

У таких умовах аналітик або автоматизована система ніколи не має повної картини стану кіберсистеми, а працює лише з її частковими та непрямими відображеннями. Це означає, що аналіз даних неможливий – в тому числі - без використання статистичних і ймовірнісних підходів, які дозволяють формалізувати невизначеність. *Оцінювання параметрів, довірчі інтервали, перевірка статистичних гіпотез і байєсівські моделі* дають змогу кількісно описувати ступінь упевненості у отриманих висновках. Наприклад, замість бінарних рішень типу «атака / не атака» формувати ймовірнісну оцінку ризику, що є більш адекватною для складних кіберсистем.

1.2.5. Роль статистичних методів в аналітиці кібербезпеки

Розглянуті вище підрозділи підручника закладають концептуальну основу для розуміння кібербезпеки як складної динамічної системи, що функціонує в умовах невизначеності, відкритості та постійної зміни загроз. Однак системний аналіз, моделювання та машинне навчання не можуть бути реалізовані без використання *формального апарату роботи з емпіричними даними*. Саме методи *математичної статистики* забезпечують цей апарат і виступають фундаментальним засобом аналізу даних у системах кібербезпеки.

У системному аналізі методи математичної статистики використовується для формалізації спостережень за станами системи та переходу від окремих подій до узагальнених характеристик поведінки [14], [15]. Мережеві потоки, журнали подій і телеметрія кінцевих точок утворюють великі вибірки даних, які

потребують опису за допомогою *розподілів, оцінок параметрів і статистичних показників*. Без цього неможливо побудувати коректні моделі нормальної поведінки системи та визначити критерії відхилень.

Моделювання в кібербезпеці, зокрема ймовірнісне та імітаційне, безпосередньо спирається на статистичні припущення щодо розподілу випадкових величин, частоти подій і кореляцій між ними. Оцінювання ризиків, аналіз сценаріїв атак і прогнозування наслідків інцидентів вимагають використання статистичних методів оцінювання, перевірки гіпотез і аналізу невизначеності. Таким чином, *статистика слугує мостом між концептуальними моделями та кількісними висновками.*

Аналітика даних і машинне навчання, згадані в попередніх розділах, також ґрунтуються на статистичних принципах. Побудова ознак, оцінювання параметрів моделей, вибір порогів спрацювання та аналіз рівня помилок неможливі без розуміння статистичних властивостей даних. Навіть алгоритми машинного навчання, включаючи сучасні нейромережеві моделі та т.з. *Великі Мовні Моделі* (англ. *Large Language Models, LLM*) фактично реалізують статистичні підходи до узагальнення інформації на основі вибірок [15]. Особливого значення математична статистика набуває в умовах невизначеності та неповноти даних. І це є характерною рисою не лише в галузі задач забезпечення кібербезпеки, а й багатьох інших прикладних галузях. Вибір адекватного закону розподілу, коректна інтерпретація статистичних показників, розуміння ролі вибірки, похибок і ступенів свободи безпосередньо впливають на якість аналітичних висновків. У задачах кіберзахисту помилки на цьому етапі можуть призводити до хибних спрацювань систем виявлення, недооцінки ризиків або пропуску реальних атак. Аналогічні статистичні помилки мають критичні наслідки і в інших сферах - від аналізу експериментальних даних у науці до оцінювання ефективності алгоритмів, бізнес-процесів чи інженерних систем.

З огляду на це*, статистика в межах даного курсу розглядається не лише як інструмент для вирішення вузькоспеціалізованих задач кібербезпеки, а як універсальний математичний апарат аналізу емпіричних даних*. Хоча навчальний посібник орієнтований насамперед на майбутніх фахівців у галузі комп’ютерних наук та кіберзахисту, коло прикладів і задач навмисно розширюється. Поряд із прикладами з мережевої безпеки, аналізу журналів подій та поведінкової аналітики будуть використовуватися приклади з інших галузей:

експериментальних досліджень, інженерних систем, економіки, біоінформатики та аналізу соціальних процесів. Така міждисциплінарність дозволяє краще усвідомити загальні принципи системного та статистичного мислення та відокремити математичний метод від конкретного прикладного контексту. Наголос буде робитися не лише на формальному застосуванні статистичних методів, а й на розумінні їхніх припущень, обмежень і області коректного використання. *Розуміння статистичних закономірностей, вміння коректно працювати з вибірками та інтерпретувати результати аналізу є ключовими компетенціями сучасного фахівця з комп'ютерних наук незалежно від спеціалізації.* Ці навички є однаково важливими для розробки алгоритмів, аналізу експериментів, дослідження продуктивності систем або побудови моделей у суміжних галузях. Саме такий підхід дозволяє сформувати універсальне аналітичне мислення, необхідне як для ефективного забезпечення кібербезпеки, так і для ширшого професійного застосування статистичних і математичних методів у сучасних інформаційних технологіях.

1.2.6. Висновки

Забезпечення кібербезпеки в сучасних інформаційних середовищах носить складний, динамічний характер. Сама система повинна розглядатися як така, у якій технічні компоненти, програмне забезпечення, дані, користувачі та організаційні процеси перебувають у постійній взаємодії. Зміна будь-якого елемента або умов функціонування такої системи може призводити до появи нових ризиків і загроз, що унеможливлює застосування ізольованих або статичних підходів до захисту.

Застосування системного аналізу дозволяє перейти від фрагментарного розгляду окремих інцидентів або засобів захисту до цілісного розуміння функціонування кіберсистеми. По-перше, системний підхід дає змогу виявляти взаємозв'язки між елементами системи, подіями та процесами, що є критично важливим для аналізу складних багатокрокових атак і емерджентних загроз. По-друге, він забезпечує роботу з *невизначеністю, яка є невід'ємною характеристикою кібербезпеки, і робить це шляхом використання статистичних, імовірнісних і аналітичних методів.* По-третє, системний аналіз створює основу для прийняття обґрунтованих управлінських рішень, зокрема в межах керування ризиками, пріоритезації захисних заходів та оптимального використання ресурсів.

Таким чином, ефективне забезпечення кібербезпеки неможливе без системного мислення, яке дозволяє поєднати технічні, аналітичні та організаційні аспекти захисту в єдину адаптивну систему. Саме системний підхід формує методологічну основу для подальшого використання статистики, моделювання та машинного навчання як інструментів аналізу даних і підтримки рішень у галузі кіберзахисту.

Контрольні запитання

1. Що таке система в системному аналізі? Назвіть її основні властивості та наведіть приклад системи з галузі кібербезпеки?
2. Поясніть різницю між відкритими та закритими системами. Як ця відмінність впливає на моделювання та аналіз комп’ютерних систем?
3. Опишіть етапи системного аналізу при розробці інформаційної системи. Які задачі вирішуються на кожному етапі?
4. Чому забезпечення кібербезпеки доцільно розглядати як динамічну складну систему, а не як набір окремих технічних засобів?
5. У чому полягає значення роботи з невизначеністю та шумом у даних для задач кібербезпеки?
6. Яку роль відіграють зворотні зв’язки в системах захисту?
7. Яким чином статистичні та ймовірнісні методи підтримують прийняття обґрунтованих рішень у системах кіберзахисту?
8. Чому машинне навчання не може розглядатися як автономний засіб забезпечення кібербезпеки без системного підходу?
9. Як системне мислення впливає на ефективність керування ризиками в кібербезпеці?

2. ПЕРВИННИЙ АНАЛІЗ ДАНИХ

2.1. Дані, їх формати, шкали вимірювання

2.1.1. Формати представлення даних. Їх класифікація

У сучасному світі дані стали однією з найцінніших складових у будь-якій сфері діяльності людини - від науки й бізнесу до повсякденного життя. Саме через дані характеристик чи ознак, що отримуються від об’єктів аналізу та спостереження, інформація фіксується, системи описуються, аналізуються. На їх основі в разі потреби виконується передбачення явищ та події.

Почнемо від початку та здамо собі питання - а що таке «дані»? ***Дані - це сукупність результатів спостережень*** за об’єктами, явищами або процесами реального світу. Ці результати можуть мати безліч варіацій, від якісних до кількісних, від чітких та однозначних до розпливчастих, суб’єктивних і багатозначних, що залежать від контексту, інтерпретації та умов їх отримання. Дослідники збирають дані, отримані в ході експериментів, підприємці збирають дані своїх клієнтів, а ігрові компанії збирають дані щодо поведінки гравців. Якщо когось спитати, як можна класифікувати дані, які йому доводилося зустрічати, то можна почути перелік, на кшталт такого: «числові (вік, ціна), категоріальні (колір, країна), упорядковані категоріальні, булеві (так/ні), часові ряди (температура по днях), текстові (опис товару), аудіосигнали, зображення тощо». Для повсякденного життя такого поділу може і вистачить. Але при науковому дослідженні, при технічному аналізі об’єктів нам потрібно чітко знати *не тільки зовнішню форму представлення даних, але й особливості їх внутрішньої (семантичної, прикладної) сутності.* Без розуміння цих особливостей або при невірному їх трактуванні навіть формально правильний аналіз може дати некоректні результати. Спробуємо розібратися в тому, які-ж дані дійсно бувають, та як не помилитися, обираючи алгоритми та інструменти їх обробки.

У програмуванні ми призвичаєні до того, що дані мають чітко структуровані форми. Їх представляють у вигляді наперед заданих типів, структур і об'єктів, які зручно обробляти в алгоритмах і системах збереження. Наприклад, у прикладних застосунках, що створені з використанням мови програмування *Python*, можливо оперувати такими змінними, як «*age = 25*» (ціле число), «*temperature = 36.6*» (дійсне число), «*name = "Victory"*» (строка або

рядок), «*is_active = True*» (булевий тип) тощо. Але звернемо увагу, що тип змінної в мові програмування ще не розкриває, як саме ці дані можна аналізувати, які операції можна виконувати, або як саме можна порівнювати дані між собою. Наприклад, якщо змінна X має значення «*23*», а змінна *Y* – значення «*-4*», то чи можна записати *Z=X/Y*? Формально – так, ніщо нас не стримує. А якщо ці змінні зберігають значення температури? То про що буде говорити отримане значення змінної *Z*?

Існують й інші способи класифікації даних. Один з них – поділ даних на якісні та кількісні. *Якісні дані* є описовими/категоріальними. *Кількісні дані* мають певне числове значення. Відразу випливає перша суттєва відмінність якісних та кількісних даних: для кількісних даних такі статистичні показники, як середнє значення та стандартне відхилення (див. Розділ 2.3) можуть мати сенс, в той час як для якісних даних – ні. Цей тип класифікації може бути важливим для вибору правильного (тобто такого, застосування якого є семантично коректним) алгоритму подальшого статистичного аналізу. Наприклад, вибір між регресією (кількісні значення змінної *X*) та дисперсійним аналізом (*ANOVA*) (якісні значення змінної *X*) ґрунтується на розумінні такої класифікації змінної(-их) *X*.

Кількісні дані можна додатково класифікувати на *дискретні* та *безперервні*. В свою чергу, *дискретні дані* можуть набувати або скінченну кількості значень, або нескінченної але зліченної кількості значень. Кількість користувачів корпоративної інформаційної системи, у яких збільшився час отримання відповіді на запит, є прикладом дискретної випадкової величини, яка може набувати скінченної кількості значень. Кількість збійних пакетів в мережі є прикладом дискретної випадкової величини, яка може набувати зліченної нескінченної кількості значень (фіксованої верхньої межі цієї кількості немає). *Безперервні дані* можуть набувати нескінченної кількості значень, таких як кров'яний тиск, рівень води в річці або температура тіла. Навіть якщо фактичні вимірювання можна округлити з певною точністю, теоретично існує деяка «точна» температура тіла, що має багато знаків після коми. Саме це робить такі змінні, як кров'яний тиск і температура тіла, безперервними.

Від того, чи мається дискретна, чи неперервна змінна, залежить вибір статистичного розподілу, що буде використовуватися для моделювання даних. Так біноміальний та пуассонівський розподіли (див. Розділ 2.4) є найбільш

вживаними при роботі з дискретними даними, тоді як гаусовський або логарифмічний – при роботі з безперервними даних.

Таким чином, для ***ефективного аналізу потрібно розуміти не лише тип, формат даних, а й їхню семантичну природу.*** Наприклад, що означає число "*25*" ? Кількість, ранг чи просто ярлик (позначку)? Щоразу, коли постає задача обробки інформації, нам потрібно зрозуміти не тільки як її формально представити, але і як її правильно вимірювати та аналізувати. Якщо ми не можемо цього зробити, то ми і не зможемо побудувати жодної моделі та прийняти правильні рішення, підкріплені доказами. Наприклад, якщо системний адміністратор підозрює, що відбувається кібератака на корпоративну мережу, але не має даних про параметри мережевого трафіку, кількість спроб входу або про активність користувачів, він не може зробити обґрунтовані висновки, а отже і вжити відповідних заходів. Якщо ж почати збирати ці метрики - наприклад, виявити, що за останню годину була зафіксована аномальна велика кількість невдалих спроб входу з однієї IP-адреси, або нетиповий обсяг вихідного трафіку з певного вузла - можна припустити, що система під загрозою, і негайно активувати протоколи реагування. Отже, *стан системи* визначається сукупністю значень параметрів системи в певний момент часу, а *поведінка системи* – як зміна станів системи в часі. Тобто ***значення параметрів системи - це в першу чергу не набір чи послідовність деяких чисел, а прояв деяких внутрішніх процесів, що відбуваються в системі.*** (Рис 2.1)

Саме на цьому рівні системного аналізу з’являється поняття *шкал вимірювання*. В повсякденному житті, коли хтось використовує термін «вимірювання», як правило мають на увазі вимірювання значень певного параметру об’єкту (наприклад – довжини, об’єму, ваги)за допомогою відповідного інструменту вимірювання або підрахунок кількості чогось (наприклад, кількість студентів у групі, планет сонячної системи, грошей в гаманці тощо). Це побутове, досить обмежене використання терміну «вимірювання». У техніці, економетриці, наукових дослідженнях, в науці про дані, термін «вимірювання» використовується значно ширше.

Рис.2.1 Стан, поведінка, параметри системи. Їх аналіз та взаємозв’язок.

Отже, одна з перших задач, що постає при вивчення будь якої системи - навчитися вимірювати значення параметрів цієї системи, звертаючи увагу на семантичні, прикладні особливості цих даних, та на математичні операції (перетворення), що має сенс застосовувати до цих даних, а вже далі - на їх формальному представленні та обробці.

2.1.2. Вимірювання як аналітична операція. Шкали вимірювання

В основі спостереження та аналізу параметрів системи лежать ***виміри, які представляють собою алгоритмічні операції****: даному стану об’єкта, за яким ведеться спостереження, ставиться у відповідність (відображається) деяке визначене позначення, буть то число, номер, ранг, мітка, символ тощо.* При цьому мають бути дотримані певні умови, а саме - різним характеристикам стану об'єкта спостереження мають відповідати різні позначення, «не різним» (тобто таким, які неможливо відрізнити один від одного) - однакові. Множина позначень - знаків або чисел, що використовуються для реєстрації станів об'єкта, що спостерігається, а також ті операції які семантично обґрунтовані для опису стану об’єкту, називається *вимірювальною шкалою* або *шкалою вимірювання*.

Шкали вимірювання використовуються для представлення та роботи як з якісними, так і з кількісними даними (що дуже незвично в побутовому розумінні). Використовуються *чотири основні шкали вимірів*, дві з яких - номінальна та порядкова - використовуються для вимірювання якісних даних, а дві - інтервальна шкала та шкала відносин - для вимірювання кількісних.

Рис.2.2 Виміри та шкали вимірювання

Дослідник, людина, що проводить аналіз, як правило розуміє природу даних. Наприклад те, що температуру треба вимірювати і градусах, відстань - в метрах, час – в секундах, оцінку студента – в набраних рейтингових балах тощо. За потреби аналого-цифрове перетворення ставить кожному відліку вхідного сигналу у відповідність номер інтервалу на шкалі амплітуд, в який цей відлік потрапляє - і це є вимір сигналу у обраній вимірювальній шкалі (Рис 2.2). Так, колір зазвичай вимірюють в таких значеннях, як «червоний», «жовтий», «блакитний», «білий», «чорний»… тощо, темперамент людини - як «сангвінік», «холерик», «флегматик», «меланхолік», а техніка, на якій може працювати програмний застосунок, поділяється на «сервери», «деск-топ комп'ютери», «ноутбуки», «планшети», «мобільні телефони». Всі ці приклади демонструють, як надане вище визначення поняття «вимір» втілюється в результати вимірів ознак (або параметрів) об'єктів реального світу.

Вже тут ми стикаємося з різноманітністю отриманих результатів саме на понятійному, семантичному рівні. Так, температуру повітря виміряють у градусах, які можуть змінюватися в діапазоні від плюс *60°C* – (всесвітній рекорд був зафіксований в 1913 році в Долині Смерті, Каліфорнія, і становив *+56.7°C*), до приблизно мінус *90°C* (рекорд *-89.2°C* був зафіксований в Антарктиді, в 1982 році). Але якщо з таким же питанням звернутися до жителя США, то він відповість, що йому потрібен термометр зі шкалою від *+100* градусів до *-130*, бо він ту саму величину (параметр) того самого об'єкта спостереження (повітря)

звик вимірювати в *шкалі* Фаренгейта, де ті самі рекорди мають значення *+134.06°F* та *-128.6°F* градусів відповідно.

Якщо ми порівнюємо довжину двох об’єктів, то один буде довше другого, наприклад, в три рази в незалежності від того, в сантиметрах чи дюймах довжина буде виміряна. Тобто, чисельне значення параметру «довжина» в сантиметрах відрізняється від чисельного значення параметру «довжина» в дюймах, а незмінним лишається лише результат порівняння (відношення) довжин двох об'єктів. А от спроба зробити аналогічний висновок про температуру двох різних об’єктів (наприклад - в двох різних містах на Землі) буде позбавлена сенсу. Якщо досягнення учнів в школі вимірювати за 12-бальною шкалою, при цьому один з учнів отримав *12 балів*, а другий *8* - то хіба можна сказати, що перший вчиться в *1.5* рази краще за іншого? Якщо ми хочемо визначити вік людини, або кількість грошей на банківському рахунку, то ми маємо використовувати шкали, які не мають негативних значень. І можемо, вимірявши значення за цими шкалами для двох різних людей, зробити висновок, що один з них старший за іншого ***на*** *20* років, або багатший ***в*** *3* рази.

Ще одне питання, на яке треба звернути увагу при аналізі та виборі шкали вимірювання – чи існує для даних, які аналізуються, справжній, «природний» нуль, тобто *значення, яке характеризує відсутність вимірюваної ознаки*. Наприклад, *0* градусів за шкалою Кельвіна є таким абсолютним нулем, а *0* градусів за шкалою Цельсія ні, оскільки за нього прийнято одне із довільно взятих фізичних явищ - температура плавлення льоду.

Наведені приклади показують, що в різних випадках при вимірюванні параметрів об’єкта необхідно точно уявляти яка (або які) вимірювальна шкала може (має) бути використана при вимірі, та вміти з цією шкалою працювати. При цьому самі *шкали різняться за*:

- множиною можливих значень, які використовуються для позначення виміряних даних;
- відношеннями, які між цими значеннями об’єктивно існують;
- операціями (синонім - перетвореннями), які можливо виконувати при виконанні аналізу цих даних, тобто які мають сенс з прикладної точки зору.

Хоча шкала вимірювання не визначає єдиний, найкращий метод чи алгоритм, придатний для аналізу даних, вона визначає, які статистичні методи *не*

можуть використовуватися для аналізу. Іншими словами шкала вимірювання задає ***допустимі перетворення*** – тобто такі перетворення*, які не змінюють співвідношень між об'єктами вимірювання.*

2.1.3. Аналіз шкал вимірювання

Як вже було сказано, існують чотири основні *типи* (ще вживається термін – *рівні*) *шкал вимірювання* - це *номінальна, порядкова, інтервальна та відносна.* І кожна з них обмежує допустимий набір операцій перетворення даних. З іншого боку, вибір тієї чи іншої шкали визначається саме семантичними особливостями даних, якими оперують і вимагає обов'язкового осмислення на початкових етапах дослідження. *Використання неадекватної до природи об'єкту шкали та/або використання недопустимих операцій в більшості випадків призводить до отримання помилкових чи навіть безглуздих результатів.*

2.1.3.1. Якісні шкали вимірювання

Історично найперші результати вимірів, з яким познайомилося людство, відображали або кількість деяких об'єктів, або деяку прийняту числову величину результату вимірів. Сьогодні природа величин, що спостерігаються в завданнях, з яким стикаються дослідники, і які треба навчитися коректно обробляти, може бути самою різноманітною. Значення ознак можуть бути кількісними, а можуть бути вказівкою на категорію, до якої можна віднести об'єкт (колір, захворювання, професія тощо). У другому випадку говорять про якісну (нечислову) ознаку.

Почнемо з більш простого, якісного вимірювання.

Шкала найменувань (інша назва цієї шкали – *номінальна*) використовується у *найпростішому типі вимірювання – класифікації тотожних об'єктів та їх розрізнення* за значенням певної ознаки. В результаті вимірювання просто визначається, чи має об'єкт певну ознаку, чи ні. Відповідно, *всі об'єкти, що мають одне і те саме значення за цією шкалою, не розрізняються між собою.* А між самим об'єктами за такою ознакою не існують інших відношень, окрім відношень «подібності» та «відмінності» (або в математичних позначеннях, відношень «=» та «≠»). Як наслідок – для даних, що виміряні в шкалі найменувань, чинними операціями є еквівалентність, входження до множини, та інші операції, пов'язані з множинами.

У *шкалі найменувань допустимими є всі взаємно-однозначні перетворення*. Тобто, якщо два об'єкти мали значення «*червоний*», то допустимо перейменувати

усі об’єкти цього класу на «*red*» (якщо така мітка раніше не використовувалася). Зрозуміло, що єдиною осмисленою характеристикою (*статистикою* – див. Розділ 2.3) для набору таких даних може бути *мода* - значення, яке зустрічається в наданому наборі даних найчастіше. Ані середнє значення, ані медіану для таких даних визначити неможливо.

Раса, національність, колір очей, волосся, місто проживання, професія, сімейний стан, знак зодіаку, колір чи марка автомобіля, група крові людини, залік складений чи ні – все це номінальні ознаки, параметри відповідних об’єктів (Рис.2.3).

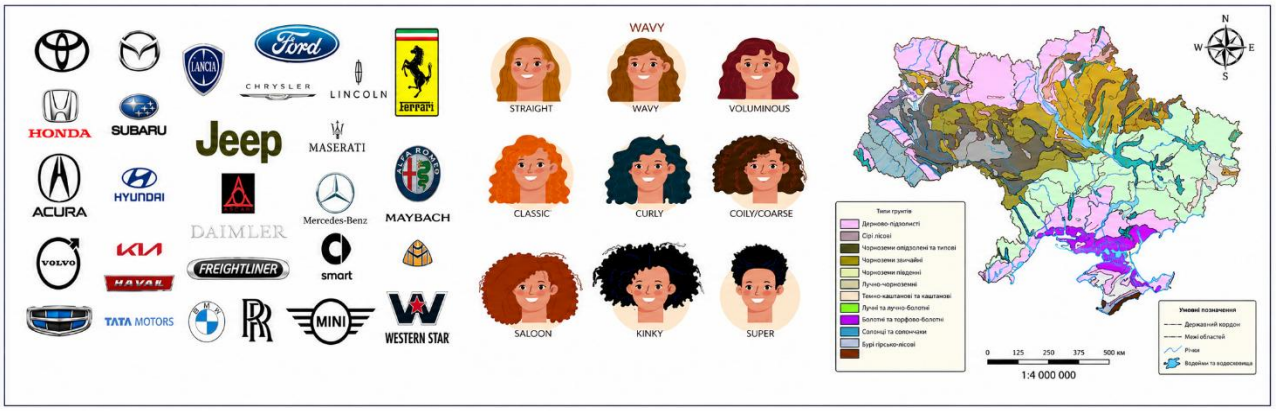


Рис.2.3 Приклади даних, що виміряні в номінальних шкалах.

Розглядаючи пакети мережевого трафіку можна визначити, який саме протокол використовується в кожному з пакетів, що пересилається мережею. Позначення протоколу в заголовках пакетів які передаються мережею інтернет - також номінальна ознака. За цією ознакою ми можемо сепарувати потік пакетів для подальшого аналізу, що є одним з найпростіших засобів системного адміністрування. Аналізуючи інформацію про протоколи можливо розглядати такі операції, як еквівалентність (чи однакові протоколи використовуються в двох пакетах), входження до множини (чи використовується протокол, що входить в деякий перелік - наприклад, допустимих - протоколів), підрахунок (скільки пакетів з заданим протоколом виявилося в наданому зразку мережевого трафіка), або знаходження моди (який з протоколів використовується найчастіше). Проте зрозуміло, що ми не можемо змістовно обчислити «суму» протоколів, або запитати, чи є протокол TCP «менший» або «більший» за протокол DNS. Аналогічно, для іншого прикладу, ми не в змозі змістовно

запитати, що в множині прізвищ певної групи людей є «середнім прізвищем» (середнє значення), або «прізвищем, найближчим до середини» (медіаною).

У шкалі найменувань виміряють, зокрема, стать людей. А результат такого вимірювання набуває лише двох значень – «чоловіча» або «жіноча». Різновид шкали найменувань, в якій використовуються тільки два можливих класи, іноді називають *дихотомічною*, або *бінарною шкалою*.

Навіть якщо класи об'єктів нумеруються - числа в даному разі виступають виключно як мітки, своєрідні ярлики для об'єктів цього класу. З таким самим успіхом для цієї мети можуть застосовуватися слова або літери. Номери телефонів, автомашин, паспортів, студентських квитків, номери шкіл у місті, номер гравця у команді, номери мережевих портів у комп'ютерах - хоча і представлені у вигляді чисел, нікому в здоровому глузді не спаде на думку складати або множити їх значення. Цікаве спостереження: шафки в роздягальнях в дитячих садках використовують малюнки, по суті - мітки, за якими діти знаходять свою шафку. Шафки для дорослих, наприклад, в басейнах, розрізняють за номерами, тобто за числами. А шафки в роздягальнях на виробництві підписують прізвищем власника шафки. Суть допустимих операцій – розрізнити шафки між собою - від того аж ніяк не змінюється. І це є єдина дія, яку можна застосувати до даних, що виміряні у такій шкалі. Штрих-коди товарів в магазині, які одночасно є і «картинкою» і числом – ще один приклад, що вказує на однакову природу таких позначок.

Яким би тривіальним не здавався вимір у номінальній шкалі, у науці ряд завдань починається саме з виділення класів об'єктів (Рис 2.4), знаходження способів ототожнення та розрізнення об'єктів дослідження, тобто з формування поняття певної *емпіричної множини*.

Маючи результати дослідження об'єктів у вигляді значень їх параметру, що виміряний у номінальній шкалі, можливо побудувати таблицю (або графік) розподілу значень, цього параметру (всім відомі стовпчикова діаграма, та кругова діаграма). В статистиці такі розподіли називають *категоріальним розподілом (англ. - Categorical Distribution,* інша назва - *«узагальнений розподіл Бернуллі»)*. В кібербезпеці до поділу на класи зводиться задача розрізнення нормального стану роботи комп'ютерної мережі, та стану компрометації системи захисту, а вивчення розподілу мережевих пакетів за протоколами - є досить дієвим методом поточного моніторингу трафіку. Задачі пов'язані з розпізнаванням об'єктів по

зображенню, диктора за голосом, автора за стилем викладання матеріалу тощо - також приклади застосування саме номінальних шкал вимірювання.

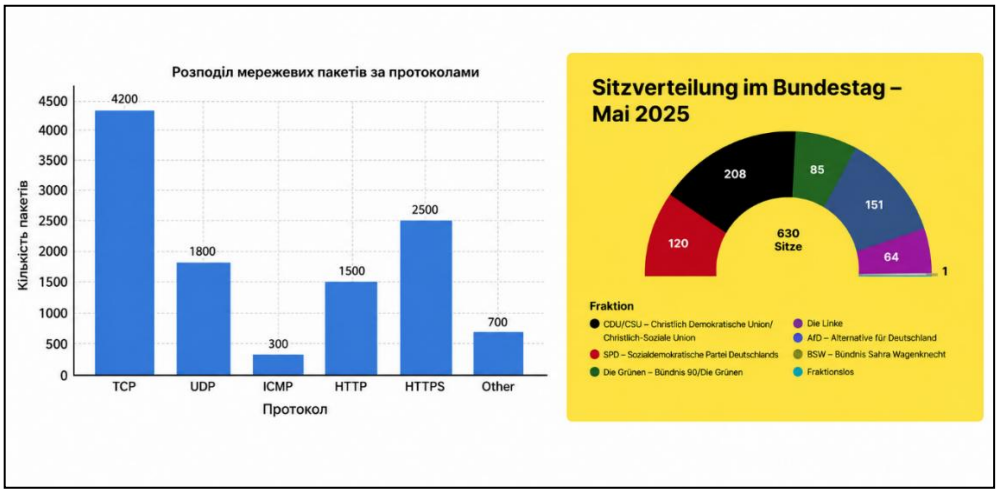


Рис.2.4 Приклади відображення даних, що виміряні в номінальній шкалі

Шкала порядку (інша назва - *рангова шкала*) застосовується для *впорядкування* об'єктів за однією чи декількома ознаками. У цій шкалі значення вимірів розбиваються на класи, *а самі класи упорядковуються між собою.* Кожному класу відповідає власна мітка, а порядок міток відповідає порядку класу за правилом «більше, ніж», «менше, ніж», «краще, ніж», «сильніше» тощо.

Прикладом вимірювання ознаки у цій шкалі є система військових звань. Оцінки на екзаменах «відмінно», «дуже добре», «добре», «задовільно», «достатньо», «незадовільно». Рівень володіння іноземною мовою. Шкала терористичної загрози – теж порядкова шкала, в якій ступінь загрози відображається кольором, але не довільними, а впорядкованим за послідовністю кольорів у веселці. Можливо «вимірювати» рівні освіти людей як «середня школа», «ступінь бакалавра», «ступінь магістра» та «доктор філософії». В усіх наведених прикладах категорії мають чіткий порядок, завжди зрозуміло, яка з них представляє вищий рівень а яка – нижчий. Але інтервали між ними не означають однакового зростання – наприклад - знань чи навичок. За ранговою ознакою часто вимірюють відгук кореспондентів у різного роду опитуваннях (Рис.2.5).

Досить часто при роботі з ранговою шкалою використовують числові мітки, як от для означення місця, що зайняла команда в чемпіонаті, важкість захворювання, сила землетрусу, сила вітру тощо. Всі зустрічалися з проханнями

дати відгук на питання «як ви оцінюєте якість обслуговування в нашій установі» з можливими оцінками від «1 - дуже погано» до «5 -дуже добре». Всі розуміють, що таке «краще обслуговування», але на скільки краще, або тим більше «у скільки разів краще» для даних, що виміряні за цією шкалою, не несуть смислового навантаження, хоча в даному разі виражені в деяких числах.

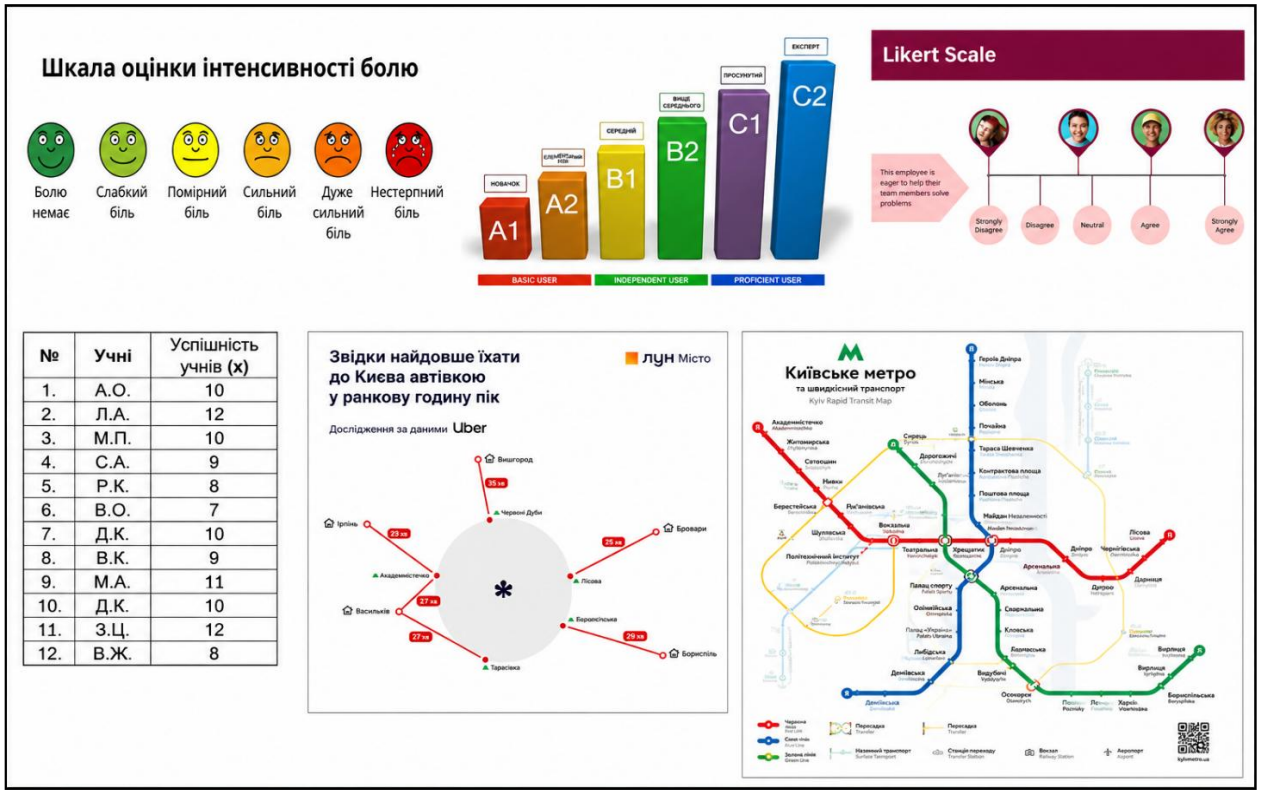
Отже, шкала порядку використовується в випадку, коли дані потрібно впорядкувати – в часі, просторі, або в будь якому іншому вимірі, але НЕ потрібно визначити точне значення. Номери будинків вздовж вулиці виміряні у порядковій шкалі – вони показують, у якому порядку стоять будинки, але ніяк не пояснюють, як далеко один дім знаходиться від іншого. Номери томів у зібранні творів письменника чи номери справ у архіві підприємства зазвичай пов'язані з хронологічним порядком їх (творів та справ) створення.

Як і номінальні дані, *величини, що виміряні в ранговій шкалі, навіть якщо вони зображені цифрами, не можна розглядати як числа.* Наприклад, не можна обчислювати вибіркове середнє, інші статистичні характеристики множини таких даних. Знаючи, що один студент отримав 5 рейтингових балів, а інший 3, використовувати різницю в балах (5-3=2 бали) для оцінки різниці в знаннях студентів сенсу не має. Говорячи строго математично, некоректним виявляється обчислення місця команди на Олімпіаді за сумою набраних місць чи медалей, або середньої оцінки учня по ряду дисциплін. Проте такі статистичні показники, як *мода і медіана* вже виявляється цілком коректним з семантичної точки зору. А також, оскільки у цих шкалах допустимі співвідношення (на додаток до тих, які справедливі і для номінальних шкал):

якщо A > B, то B < A;

якщо A > B та B > C, то A >C;

то це означає, що такі *дані можливо сортувати*. Якщо для об’єктів визначаються значення двох параметрів, причому обидва параметри вимірюються в порядковій шкалі (наприклад – оцінки з двох різних предметів, або важливість та складність політики безпеки системи захисту) за допомогою коефіцієнту рангової кореляції Спірмана (див. Розділ 4.5) можливо виявити, чи існує статистичний лінійний зв’язок між цими параметрами.



№	Учні	Успішність учнів (x)
1.	А.О.	10
2.	Л.А.	12
3.	М.П.	10
4.	С.А.	9
5.	Р.К.	8
6.	В.О.	7
7.	Д.К.	10
8.	В.К.	9
9.	М.А.	11
10.	Д.К.	10
11.	З.Ц.	12
12.	В.Ж.	8

Рис.2.5 Приклади відображення даних, виміряних у рангових шкалах

Іноді *додатково* вимагається, щоб кожен клас складався з лише одного елемента (тобто забороняється рівність між двома об’єктами за ознакою) . В такому разі говорять про *строге впорядкування елементів*, виміряних у порядковій шкалі. Позатим слід ще раз підкреслити, що для даних, які виміряні у шкалах порядку, неможливо визначити міру домінування, тобто виміряти, наскільки один об'єкт кращий (або гірший), важливіший, сильніший за інший.

Порядкові шкали не мають природньої абсолютної початкової точки або точки відліку, в якій вимірювана характеристика не існує. Ця відсутність «справжнього» нуля обмежує математичні операції, які можна виконувати з порядковими даними.

Використання порядкових шкал дозволяє застосовувати всі операції, які допустимі для номінальної шкали, а також перетворення отриманих значень за допомогою монотонно зростаючих відображень, тобто таких перетворень, в *результаті застосування яких зберігається обраний порядок в множині об’єктів.* Так, якщо значення представлені числами, наприклад - позитивними рейтинговими оцінками - значення можуть бути замінені їх квадратами,

логарифмами, або будь-який іншою функцією, що монотонно зростає. Значення самої ознаки буде змінено, але їх взаємопорядок при цьому лишиться без змін.

Порядкова шкала та шкала найменувань – основні шкали ***якісних ознак***. Тому в багатьох галузях науки та практичних завданнях результати якісного аналізу можна розглядати як вимірювання за цими шкалами. Шкали, що розглядаються нижче, відносяться до шкал ***кількісних ознак***.

2.1.3.2. Кількісні шкали вимірювання

Інтервальна шкала - це шкала, яка поєднує властивості номінальної та порядкової шкал, але йде далі, дозволяє *кількісно визначити інтервали (різницю) між значеннями ознак.* В цій шкалі також відсутня справжня, природня нульова точка початку вимірювання. Тобто нульове значення, навіть якщо і використовується на цій шкалі, не означає відсутність вимірюваної ознаки. Так, температура, виміряна в градусах Цельсія або Фаренгейта (Рис 2.6), є класичним прикладом даних, виміряних за інтервальними шкалами. Інтервали між градусами є однаковими, в незалежності від самих значень (наприклад, різниця між 10°C та 30°C така ж, як між 40°C та 60°C або між -10°C та +10°C), але температура 0°C не означає відсутність температури, це лише точка умовного початку відліку.

Хоча значення, що виміряні у інтервальних шкалах можна додавати або віднімати одне від іншого, їх неможна множити чи ділити. Наприклад, +40°C – це не в два рази більше, ніж +20°C. Використовуючи вимірювання показників в інтервальній шкалі можна точно сказати, наскільки одне значення більше за інше. Це дозволяє дослідникам систематизувати свої вимірювання і говорити про те, наскільки великі або малі відмінності. Наприклад - приріст продукції підприємств (в абсолютних одиницях) у поточному році порівняно з минулим, збільшення чисельності установ, кількість придбаної техніки протягом року тощо.

Саме *відсутність природнього нуля* призводить до існування багатьох різноманітних шкал вимірювання однієї і тієї величини. Прикладом може слугувати вимір висоти місцевості. Її можна виміряти відносно рівня моря, відносно найнижчої точки поверхні на Землі, відносно центру Землі тощо. Іншим прикладом вимірювання в інтервальній шкалі може бути ознака «дата здійснення події», оскільки для вимірювання часу в конкретній шкалі необхідно фіксувати масштаб *і початок відліку*. Григоріанський, ісламський, іудейський, китайський,

буддистський календарі є прикладами конкретизації шкал інтервалів за рахунок різних прийнятих точок початку відліку та різних одиниць вимірювання -тижнів, місяців, років (а був ще французький революційний календар, юліанський календар тощо).

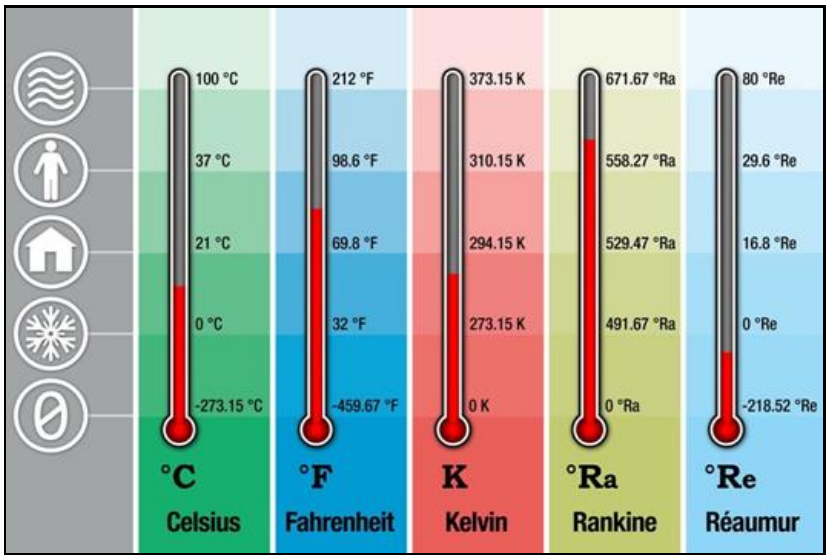


Рис 2.6 Температура, виміряна у різних інтервальних шкалах

Інтервальні дані можуть бути дискретними з цілими числами, (10 градусів, 20 років, 2 місяці тощо), або безперервними з дробовими числами, такими як 12.2 градуси, 3.5 тижнів або 4.2 сантиметри. Допустимими перетвореннями у шкалі інтервалів є лінійні зростаючі перетворення, тобто лінійні функції $y = ax + b$, $a > 0$, $-\infty < b < +\infty$. Основною властивістю цих шкал є збереження незмінними відносин інтервалів в еквівалентних шкалах:

$$\frac{X1 - X2}{X3 - X4} = \frac{\Phi(X1) - \Phi(X2)}{\Phi(X3) - \Phi(X4)}$$

Тобто шкали зберігають різницю, впорядкованість об’єктів, що досліджуються, а також відношення відстаней між парами об’єктів. Проте не зберігають початку відліку (параметр b) і масштабу вимірювань (параметр a).

Досить специфічним, але цікавим окремим випадком інтервальної шкали ***є циклічна шкала.*** Прикладом вимірювання в такій шкалі є вимір часу по

годиннику, або вимір кута за компасом або транспортиром. В цій шкалі значення не змінюється при довільній кількості зсувів, тобто при перетворенні:

$$y = x + nb,$$
$$n = 0, 1, 2,\ldots$$

В такій шкалі, наприклад, різниця (інтервал) між 20:00 та 23:00 годинами та між 23:00 та 02:00 годинами буде однаковим. Однак угода про хоч і довільний, але єдиний для нас початок відліку шкали дозволяє використовувати показання в цій шкалі як числа, застосовувати до них арифметичні дії (до тих пір, поки хтось не забуде про умовність нуля, наприклад при переході на літній час або назад).

Типова помилка при роботі з шкалою інтервалів - властивості, що вимірюються в такій шкалі, приймаються як показники для інших властивостей, монотонно пов'язаних з першими. Наприклад випробування розумових здібностей, у якому вимірюється час, що був витрачений на вирішення деякого завдання. Хоча фізичний час вимірюється в шкалі інтервалів, час, що використовується як міра розумових здібностей, належить шкалі порядку. Ігнорування цього факту часто призводить до неправильних результатів.

Інтервальні шкали зручні та звичні для нас. Їх використання для вимірювання дає можливість статистичного аналізу таких даних. Наприклад, центральну тенденцію (див. Розділ 2.3.4.1) для множини таких даних можна виміряти за допомогою моди, медіани або середнього значення; також можливо розрахувати відповідні дисперсію та стандартне відхилення.

Ключовою навичкою для спеціалістів з обробки даних є розуміння відмінностей між даними, що виміряні в інтервальній шкалі та даним, що виміряними в ***шкалі відношення***. Обидві ці шкали використовуються, якщо дані представлені в кількісній формі. Однак між цими шкалами існують суттєві відмінності. Як інтервальні шкали, так і шкали відношення показують нам порядок і точне, кількісне значення різниці між одиницями. Але, на відміну від інтервальних шкал*, шкали відношення мають абсолютне, природнє нульове значення*, тобто повну відсутність вимірюваної характеристики об’єкту дослідження, що дозволяє проводити з такими даними повний спектр операцій статистичного аналізу.

Зріст, вага, температура за шкалою Кельвіна, кількість дітей в родині, мешканців в місті, кількість голосів, отриманих на виборах, частота пульсу людини,

доза ліків, тривалість життя, інтервал між подіями, відстань між двома точками - все це є прикладами даних, що виміряні в шкалі відношень. Наприклад, ми можемо виявити, що один пакет трафіку в три рази довший за інший. Або сесія між двома об’єктами мережі тривала в чотири рази довше, аніж попередня. Допустиме перетворення шкал відношення - $y = ax$, $a > 0$, інакше кажучи, лінійні зростаючі перетворення без вільного члена.

Ще деякі приклади. Вимірюючи вагу об’єкту кілограмах, отримуємо одне чисельне значення, при вимірі ваги у фунтах - інше. Можна помітити, що в будь-якій системі одиниць *відношення* мас будь-яких об'єктів однаково і *при переході від однієї числової системи до іншої, еквівалентної, не змінюється*. Цю ж властивість мають і вимір відстаней, і вимір довжин предметів. Якщо у вас в кармані (або на банківському рахунку) певна сума грошей, а в іншої людини – в два рази більше, то таке співвідношення збережеться в незалежності від того, будете ви вимірювати своє багатство у гривнях, євро, тугриках або зімбабвійських доларах.

Дані, що виміряні в шкалі відношення є більш гнучкими, ніж усі попередні. *Такі дані можливо (семантично змістовно) додавати, віднімати, множити та ділити.* Як говорить сама їх назва, можливо визначити відношення між двома значеннями.

Окремим випадком шкали відношення є ***абсолютна шкала вимірювання***. Загалом кажучи, це шкала відношення *за додаткової умови наявності природної одиниці вимірювання*. В якості значень на цій шкалі використовуються як натуральні числа, коли об'єкти представлені цілими одиницями, так і дійсні числа, якщо крім цілих одиниць присутні і частини об'єктів. У цій шкалі приймається нульова точка відліку ($b = 0$) та одиничний масштаб ($a = 1$). Над даними, що виміряні в такій шкалі допускаються будь-які дії, а от *будь які перетворення - заборонені*. Тобто – це *безрозмірна шкала*. Це означає, що існує лише одне відображення об'єктів у числову шкалу. Наприклад, в абсолютній шкалі визначають *кількість* людей, об'єктів, подій тощо, яке можна виміряти єдиним способом за допомогою натуральних чисел. Інтервал дійсних чисел від 0 до 1, який застосовується для оцінки ймовірностей випадкових подій - це теж приклад абсолютної шкали. Ця шкала часто використовується як допоміжна шкала при вимірах та інших шкалах. Числа, одержані за такою шкалою, можна складати, віднімати, ділити, множити – всі ці дії будуть осмисленими.

2.1.4. Чи має значення шкала вимірювання для аналізу даних?

Правильний вибір шкали вимірювання має велике значення і залежить від наявності необхідної інформації та мети, яка переслідується під час проведення оцінювання. Так, використання кількісних шкал потребує повнішої інформації порівняно з номінальними або порядковими шкалами, а отримання цієї інформації пов'язане з додатковими витратами ресурсів і часу. Тому при виборі типу шкали завжди необхідно враховувати особливість задачі: як ця інформація буде використана надалі? Аналіз шкал, проведений вище, дає нам підстави розглядати взаємовідносини між шкалами. Чим вищій рівень шкали, в якій були виміряні та представлені дані, тим більше інформації несуть в собі дані, тим складніші математичні перетворення (операції), які допустимі при їх обробці (Рис 2.7). Чітке розуміння рівня шкали вимірювання в якій представлені дані може допомогти запобігти семантичним помилкам, таким як усереднення групи поштових індексів або співвідношення двох екзаменаційних оцінок. Оскільки тип шкали визначає множину допустимих математичних операцій над результатами виміру, остільки вплив початкового системного представлення об'єкта позначається на можливості і адекватності застосування тих чи інших математичних методів, що може визначити успіх чи неуспіх роботи. Крім того, знання шкали вимірювання для змінних насправді допомагає як планувати аналіз систем, так і інтерпретувати результати, отримані в результаті дослідження.

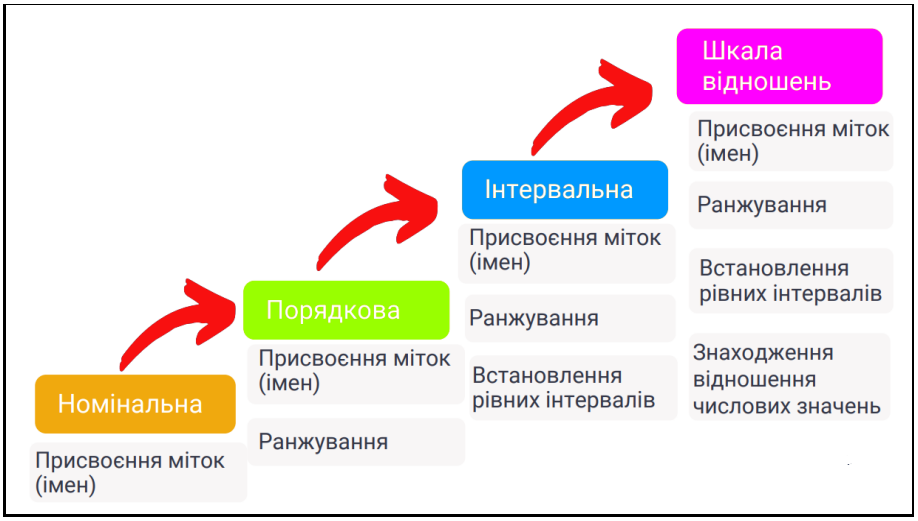


Рис 2.7 Рівні шкал вимірювання та семантика операцій, які мають сенс при роботі в цих шкалах

В наведеній нижче таблиці 2.1 представлені основні статистичні показники, які допустимо використовувати при роботі з даними, що виміряні в відповідних шкалах.

Підсумовуючи, номінальні шкали використовуються для позначення або опису значень без відносно до інших можливих значень. Порядкові шкали використовуються для надання інформації про конкретний порядок значень даних. Інтервальна шкала використовується для розуміння як факту наявності порядку між значеннями даних так та величини відмінностей між ними. Шкала відношення надає інформацію про ідентичність, порядок та відмінності, а також співвідношення різних значень даних.

Зауважте, що іноді шкала вимірювання для ознаки об'єкту не є чітко визначеною. В якій шкалі представлений колір? У психологічному дослідженні на особливості сприйняття зовнішніх подразників він може вважатися номінальною ознакою. При постановці медичного діагнозу кольори вважатимуться порядковими (ступень почервоніння обличчя). У фізичному дослідженні колір кількісно визначається довжиною хвилі, тому колір вважатиметься змінною шкали відношення. А вік людини - як він змінюється? Дискретно (хоча-б з точністю до дня) чи безперервно?

Таблиця 2.1

Допустимі математичні перетворення

Статистика та операції	Типи шкал			
	Номінальна	Рангова	Інтервальна	Відношень
Мода	Так	Так	Так	Так
Розподіл ймовірностей	Так	Так	Так	Так
Медіана	Ні	Так	Так	Так
Середнє арифметичне	Ні	Ні	Так	Так
Ранг	Ні	Ні	Так	Так
Дисперсія	Ні	Ні	Так	Так
Додавання/віднімання	Ні	Ні	Так	Так
Множення/ділення	Ні	Ні	Ні	Так

У шкал є одна зручна властивість: будь-які дані, виміряні в шкалі вищого рівня, можна легко перетворити на дані, виміряні в шкалі нижчого рівня.

Загальна порада полягає в тому, щоб завжди прагнути використовувати найбільш сильну шкалу, яку дозволяє джерело даних.

Значення аналізу та обґрунтування вибору відповідної шкали вимірювання полягає також в тому, що саме на їх основі багато в чому ґрунтується «місток» між семантично-прикладним формулюванням завдання дослідження та адекватним математичним апаратом його виконання і в решті-решт - його коректної програмної реалізації.

Бажаючим поглибити свої знання в питаннях, пов'язаних з теорією вимірювання даних та застосуванням шкал, можна порекомендувати звернутися до джерел [17], [18], [19], [20], [21].

Контрольні запитання

1. В якій шкалі вимірювання доцільно представлені наступні данні?
   - Результати шоу талантів (наприклад, перше місце, друге місце, 10-те місце...).
   - Зріст у см.
   - Вага у кілограмах, грамах.
   - Кольори за назвою.
   - Температура в градусах Цельсія.
   - Рейтингова оцінка студента по дисципліні за семестр.
2. В чому полягає значення ідентифікації типу шкали виміру для подальшого аналізу даних та моделювання системи?

2.2. Невизначеність в даних. Випадкові величини

У сучасному світі, який характеризується складністю, динамічністю та високим рівнем невизначеності, виникає об'єктивна потреба у формалізованому підході до прийняття рішень в умовах нестачі або варіативності інформації. У контексті сучасного аналізу складних систем - технічних, економічних, соціальних - ми маємо справу з великою кількістю даних, часто *неповних, зашумлених або невизначених*.

У повсякденному житті, професійній діяльності, бізнесі та наукових дослідженнях ми постійно стикаємося з ситуаціями, результат яких наперед не визначений. Торговець не знає, скільки відвідувачів прийде до магазину завтра; службовець - скільки часу триватиме дорога на роботу; підприємець - яким буде

курс валюти через місяць; банкір - чи поверне клієнт позику вчасно; страховик - коли і за яких обставин доведеться здійснювати страхову виплату. Ми не можемо точно сказати, скільки годин працюватиме лампа, як довго прослужить електронний прилад, скільки користувачів відвідає сайт в наступному місяці чи який буде обсяг переданих даних у мережі через п’ять хвилин. Ми не можемо точно передбачити, як довго працюватиме комп’ютерна система без перезавантаження, скільки часу триватиме DDoS-атака, скільки зловмисних пакетів пройде через фаєрвол або який рівень помилок виникне під час передачі даних по нестабільному каналу, чи проявиться дефект програмного забезпечення у критичний момент. Така ***невизначеність*** властива як природним, так і технічним, соціальним, економічним системам.(рис 2.8)

Незважаючи на це, ми змушені постійно приймати рішення. У простих ситуаціях ми часто керуємося інтуїцією, досвідом або здоровим глуздом. Ми часто намагаємося зробити якийсь «запас міцності про всяк випадок»: скажімо, «вийдемо з дому на десять хвилин раніше, щоб *майже напевне* не спізнитися на роботу». Проте в умовах сучасного бізнесу, конкуренції, у критично важливих сферах - кібербезпека, телекомунікації, енергетика, фінансові технології, медицина, військова справа - такий підхід є недостатнім. Інженер чи менеджер, лікар або агроном, проектувальник атомної станції чи системи захисту від крилатих ракет не може зупинитися на «може бути». Той, хто приймає рішення хоче знати, наскільки часто це «може бути»? Яка частка випадковості? Чи можна її виміряти? Як її враховувати? З яких причин вона змінюється? Чи можна нею керувати? Для отримання відповідей на ці питання необхідно спиратися на науково обґрунтовані методи аналізу, які дозволяють працювати з невизначеністю системно, і працювати об’єктивно та відтворювано. Рішення мають базуватись на *статистичних оцінках*, *аналітичних моделях* і, за потреби, *методах машинного навчання*.

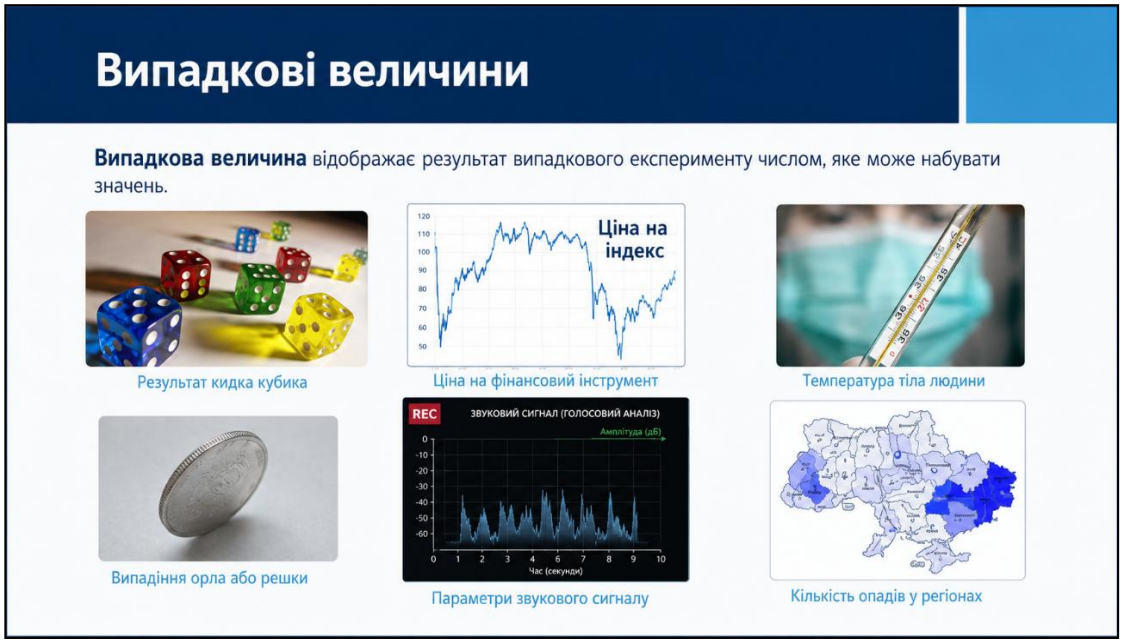


Рис. 2.8 Випадкові величини в навколишньому світі

Для того щоб адекватно описати поведінку систем, спрогнозувати їх розвиток або прийняти обґрунтоване рішення, ми повинні навчитися працювати з математичними абстракціями, які відображають цю невизначеність. Саме тут «з'являються» в дослідженнях такі поняття, як випадкові величини, їх параметри, вибірки, тобто методи та засоби, які дозволяють формалізувати імовірнісні характеристики явищ і процесів. Вони формують підґрунтя *статистичного моделювання, яке, своєю чергою, є необхідним інструментом системного аналізу та ключовим компонентом машинного навчання.*

Видатний британський математик Карл Пірсон (Karl (Carl) Pearson) сказав «*статистика – це граматика науки*». Сьогодні ми можемо до цього додати, що це особливо стосується комп'ютерних та інформаційних наук, а також фізичних, економічних, біологічних наук тощо. І безсумнівно - *статистика це основне підґрунтя науки про аналіз даних, машинного навчання, інтелектуальної обробки інформації, систем штучного інтелекту*. Без її розуміння розібратися в тому, як працюють більшість алгоритмів машинного навчання, а тим більше удосконалювати їх та розвивати, практично неможливо. Опанування цього матеріалу є необхідним для подальшого ефективного використання методів системного аналізу, оптимізації, прогнозування та машинного навчання в умовах реального світу, де повна визначеність - радше виняток, ніж правило.

Для подальшого більш глибокого занурення в основи статистики радимо звернутися до джерел: [22], [23], [24], [47]. Тим, хто цікавиться темою «Статистичні основи Data Science» буде цікаво ознайомитися з [25].

2.2.1. Випадкові величини. Визначення та класифікація

Явища, ситуації, результати яких повністю визначаються факторами, що на нього впливають, називаються *детермінованими*, а ті, в яких це не виконується - *недетермінованими* або *стохастичними*.

Випадковою називається величина, яка в ході її вимірювання (див. Розділ 2.1) може приймати деяке невідоме наперед значення.

Найбільш відомий, хоч і тривіальний, приклад вимірювання випадкової величини – результати підкидання монеті або кидання кубику. Більш складні випадки, з якими ми зустрічаємося чи не кожного дня: температура повітря в заданій точці в заданий момент часу; кількість відмов обладнання протягом місяця; кількість автомобілів, що проходять через перехрестя за хвилину; ймовірність активації аварійного протоколу; кількість сонячних днів в році або кількість спалахів на сонці за рік; час очікування наступного автобуса на зупинці; результат (наприклад час) який спортсмен покаже в наступних змаганнях тощо.

Зазвичай результати вимірів статистичної величини (а ми вже знаємо, що вимірювання в аналізі даних - це поняття більш широке, аніж щось поміряти лінійко, амперметрам чи секундоміром) зумовлені впливом фізичних факторів, які неможливо точно виміряти або проконтролювати. Наприклад, у випадку підкидання звичайної монети кінцевий результат - випадіння аверсу чи реверсу - визначається комплексом нестабільних фізичних параметрів, таких як початкова швидкість, кут обертання, сила тертя тощо. Оскільки ці параметри мають випадковий характер або змінюються в неконтрольований спосіб, результат експерименту набуває ознак невизначеності. Таким чином результат, який ми отримуємо, є величиною випадковою. *Областю визначення випадкової величини* в даному прикладі є множина всіх можливих елементарних результатів. Для ідеалізованої моделі монети ця множина обмежується двома елементами - "аверс" і "реверс" (ознака вимірюється в дихотомічній шкалі).

Якщо зрозуміло, хоча-б теоретично, від яких параметрів залежить результат нашого виміру, але невідомо як саме ця залежність описується, говорять, що *випадкова величина залежна* від вказаних параметрів. Якщо ж

залежність результату від будь яких інших параметрів невідома (або не береться до уваги) - говорять, що *випадкова величина є незалежною*. Зверніть увагу на те, що незалежність величин між собою говорить про те, що знання значення однієї з них *не дає жодної інформації про можливі значення іншої*. Або, що події, які описуються такими величинами, *не впливають одна на одну*.

Якщо-ж величини є залежними, це означає, що *існує статистичний або функціональний зв'язок між величинам, що розглядаються*. І *значення однієї з них впливає на можливі значення іншої*. Такі величини часто мають спільне походження або описують взаємозалежні елементи системи.

Простим прикладом залежних випадкових величин може бути температура процесора і швидкість обертання кулера в комп'ютерній системі. Інший приклад – аналіз росту людей з деякої групи. Зазвичай ми розуміємо, що зріст однієї особи ніяк не залежить від зросту інших осіб і є величиною незалежною. Але чи завжди це так? Не завжди, бо зріст двох людей все ж може перебувати в залежності між собою. Наприклад - якщо ці люди знаходяться в близькому родинному зв'язку, «батько-дитина», або «брат-брат». Отже ***визначення наявності/відсутності залежності між величинами, які досліджуються - важлива задача системного аналізу***, від результатів вирішення якої можуть залежати і перебіг, і всі подальші результати дослідження (див. Розділ 4).

Нагадаємо, що випадкові величини можуть бути або *дискретними*, тобто такими, які можуть приймати деякі конкретно задані значення з деякої обмеженої, визначеної наперед множини можливих (наприклад - кількість дітей; кількість запитів до веб-сервера за хвилину), зокрема ранговими, номінальними, дихотомічними, або можуть бути *безперервними*, тобто такими, можливі зазначення яких заповнюють деякий інтервал (наприклад - зріст, вага, час до відмови обладнання; швидкість завантаження файлу тощо).

Слід мати на увазі, що безперервні випадкові величини можуть з'явитися як в результаті безпосереднього вимірювання параметрів об'єкту спостереження, так і в результаті певних перетворень первинних (сирих значень. Наприклад, при вимірюванні параметрів мережевого трафіку ми майже завжди маємо справу з дискретними величинами: кількість байт в пакеті, кількість переданих пакетів, кількість з'єднань, встановлених за певний період, кількість помилок під час передачі, кількість пакетів, заблокованих фаєрволом тощо. Безперервні дані з'являються в похідних параметрах трафіку, наприклад - якщо вимірюється

середня кількість байт в пакеті, середня кількість пакетів на секунду, середній час між надходженням пакетів, ймовірність втрати пакету, частка заблокованих запитів тощо. Отже, як правило (хоча і не завжди) дискретні дані - це сировина (англ. *Raw Data*), з якої через математичну обробку (нормування, усереднення, диференціювання тощо) отримують неперервні величини, придатні для моделювання, прогнозування, тренд-аналізу та машинного навчання.

2.2.2. Розподіл випадкових даних. Функції розподілу

У системному аналізі та моделюванні складних об'єктів із використанням методів машинного навчання важливе значення має вивчення поведінки випадкових величин, зокрема їхніх розподілів. *Розподіл імовірностей - це основний математичний інструмент*, який дозволяє кількісно описати стан об’єкту дослідження. Формалізація розподілу є основою для прогнозування, оптимізації та прийняття рішень у умовах невизначеності.

Перш ніж перейти до розгляду розподілів, нагадаємо, що *ймовірність події* відображає частоту її виникнення серед великої (теоретично – нескінченої) кількості однотипних вимірів (експериментів, спостережень). У загальному випадку, якщо досліджується сукупність численних реалізацій випадкового процесу або явища, ми можемо побудувати *ряд розподілу ймовірностей* - тобто множину впорядкованих пар вигляду:

$$\left\{x_i, P\left(X = x_i\right)\right\}$$

де x_i - можливий результат виміру значення випадкової величини X, а $P\left(X = x_i\right)$ - ймовірність його появи. Така відповідність між значеннями величини та їх імовірностями формалізується у вигляді **закону розподілу**. Іншими словами, ***розподіл випадкової величини*** *- це математична модель, що описує, як значення випадкової величини пов’язані з імовірністю їх появи.* Це фундаментальна концепція, що визначає, з якою імовірністю той чи інший стан системи виникає в процесі її (системи) функціонування.

У випадку ***дискретної випадкової величини*** (наприклад, кількість запитів до сервера за хвилину, кількість збоїв у мережі за добу) розподіл задається

таблицею ймовірностей, де кожному можливому значенню параметра що вимірюються, відповідає певна ймовірність. В подальшому, при дослідження системи можна припустити, що розподіл цієї величини підпорядковується деякому відомому (теоретичному) закону розподілу. Наприклад - кількість підключень до мережі в окремий момент часу може бути - можливо з якоюсь похибкою - описаний розподілом Пуассона. Так це чи ні дослідник має впевнитися в результаті подальшого дослідження відповідності даних з таблиці ймовірностей та теоретичного визначення обраного закону розподілу.

Для ***безперервних випадкових величин*** (наприклад, час затримки пакету в мережі або температура процесора, рівень шуму в сигналі тощо) значення не можна перелічити, тому розподіл описується за допомогою *функції щільності* (англ. *Probability Density Function, скорочено* - PDF). Математично ймовірність потрапляння виміряного значення у деякий інтервал значень визначається площею під графіком цієї функції над відповідним інтервалом. Типовим прикладом є нормальний (гаусовський) розподіл, який часто моделює шум у даних або результат агрегації великої кількості малих випадкових впливів.

У контексті машинного навчання та моделювання систем, знання про форму (функцію щільності) розподілу є принципово важливим. Наприклад, алгоритми класифікації, регресії або оцінки ймовірностей можуть вимагати попередньої нормалізації даних або перевірки статистичних припущень щодо розподілу. Ігнорування таких вимог призводить до втрати точності або хибних висновків при прийнятті на основі цих висновків подальших рішень.

Нагадаємо, з точки зору теорії вимірювання «подія» це не обов'язково дія. Це - в загальному випадку - може бути будь яким значенням, отриманим в результати виконання процедури виміру, і в залежності від обраних типів шкал вимірювання кінцевий розподіл може проявлятися по-різному. Можна, наприклад, визначати розподіл чоловіків і жінок в деяких колективах (біноміальна шкала). Для номінальної ознаки можна побудувати таблицю розподілу кількості брюнетів, шатен і блондинів в деякій групі людей. Може бути побудований розподіл кількості курсантів, що прибули на навчання з кожної з областей країни (Рис.2.9). В останньому прикладі «подія» є факт того, що конкретна людина проживає в конкретній області. Сам по собі ряд або функція розподілу вже може представляти вельми інформативну основу для аналізу. Зрозуміло, що якщо номінальна ознака що досліджується за своєю природою

унікальна (наприклад - номер паспорта), то будувати таку таблицю сенсу мало. Але, наприклад, побудувати таблицю кількості дзвінків, що здійснюються з кожного з номерів телефонів в деякій організації - вельми цікава інформація, що дозволяє виявити "базік".

Рис. 2.9 Приклади графічного відображення розподілу даних

2.2.3. Гістограми. Побудова та використання

Після формального визначення закону розподілу випадкової величини як відповідності між емпіричними значеннями цієї величини та ймовірностями їх появи, постає практичне завдання - візуалізувати цей розподіл. Одним із найпоширеніших способів такої візуалізації є гістограма. Хоча найчастіше гістограми постають у вигляді деякого графіка, для практичного аналізу (наприклад, при програмній реалізації) нас будуть цікавити самі значення, що на цій гістограмі з'являються, по суті - сама *таблиця ймовірностей.*

Залежно від типу даних, побудова гістограми потребує різних підходів:

1. *Гістограма для номінальних (категоріальних) даних.*

У разі, якщо значення випадкової величини є номінальними (тобто приймають певні фіксовані, якісні значення - як-от тип протоколу, статус транзакції, клас об'єкта, діагноз захворювання, тощо), побудова гістограми зводиться до підрахунку частоти появи кожного унікального значення (Рис. 2.10).

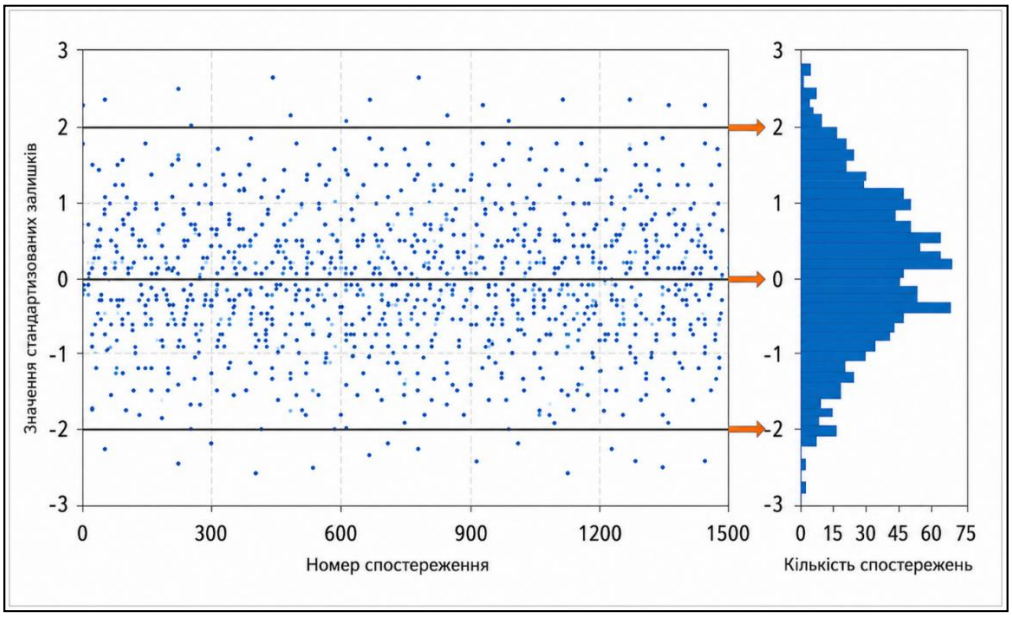


Рис. 2.10 Процедура побудови гістограми

При використанні мови програмування Python це можна реалізувати у кілька простих кроків:

– Визначення унікальних значень у наборі даних, наприклад, за допомогою функції *set()*.
– Підрахунок частоти кожного значення - через метод *.count()* або за допомогою словника частот.
– Візуалізація результатів, тобто побудова стовпчикової діаграми за допомогою функції *plt.bar()* із бібліотеки *matplotlib*.

Такий підхід дозволяє, зокрема, побачити, які категорії є найпоширенішими, що важливо, наприклад, у класифікаційних задачах машинного навчання, в задачах діагностики – як технічної, так і медичної.

2. *Гістограма для числових (безперервних) даних*

У випадку числових безперервних даних, змінна може набувати нескінченну кількість значень. Тому попередньо необхідно виконати *дискретизацію* наданих для аналізу первинних даних - тобто розбиття області значень на діапазони (інтервали, або «біни» - від англ. *Bean*).

Подальші кроки процедури аналізу включають:

– Визначення кількості бінів (кількість інтервалів, на які розбивається числова шкала). Ця задача – за деякими обмеженнями, серед яких вимога однаковості розмірів всіх інтервалів - рівнозначна задачі визначення розміру кожного з інтервалів.
– Підрахунок кількості спостережень кількість, що потрапляють у відповідний інтервал.
– Побудова графічної (візуальної) гістограми, яка відображає кількість (або частоту) даних у кожному діапазоні(інтервалі).

Для полегшення реалізації цієї процедури в Python та його бібліотеках існують відповідні функції:

– *np.histogram()* (з пакету *NumPy*) - виконує попередній розрахунок частот по інтервалах.
– *plt.hist()* (з пакету *Matplotlib*) - автоматизує всі кроки побудови гістограми, від розбиття на інтервали до виводу графіка.

Отриману гістограму зазвичай візуалізують в більш зручній, ніж показано на Рис 2.10 формі, а саме – у горизонтальному представленні, в якому діапазон можливих значень відображають по горизонтальній осі, а кількість значень, що потрапили в кожний з інтервалів – по вертикальній (Рис. 2.11).

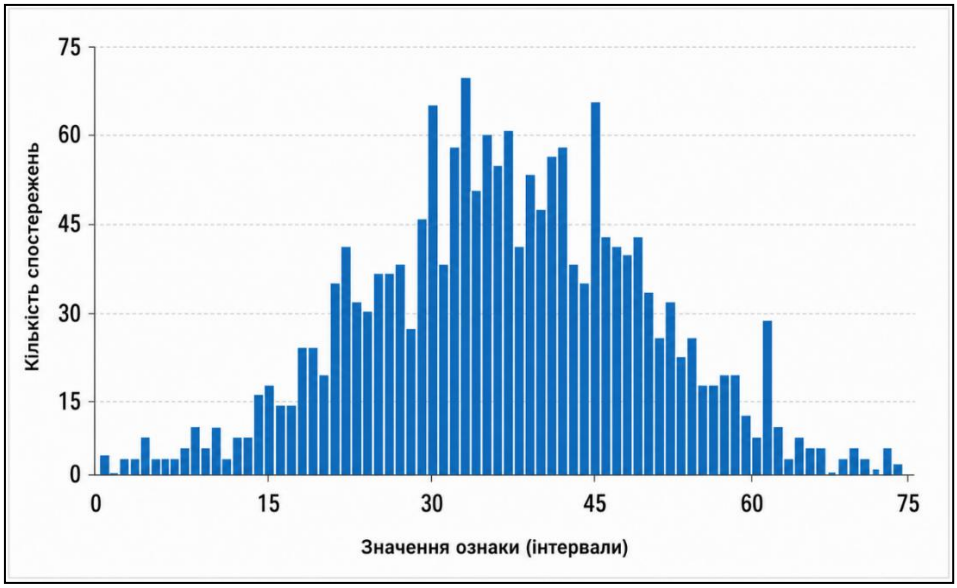


Рис. 2.11 Графічне представлення гістограми

Складність виникає, наприклад, тоді, коли нам треба порівняти дві гістограм, за умови що сумарна кількість експериментів в першому і другому випадку різні. Зрозуміло, що в такій ситуації висота стовпчиків на діаграмах буде принципово різна. Що б уникнути цього дані нормують, тобто виконують перехід від відображення абсолютної кількості повторень кожної з подій k_i до відображення його частки серед всіх K подій експерименту:

$$k_i^* = \frac{k_i}{\sum_i^K k_i}.$$

Зрозуміло, що $0 \le k_i^* \le 1, \sum_i^K k_i^* = 1.$ При такому перетворенні і при незмінній кількості інтервалів форма гістограми не зміниться, а от масштаб по вертикалі стане стандартизованим (тобто буде коливатися в діапазоні від 0 до 1) і не буде залежати від кількості вхідних даних (Рис.2.12). При певних умовах це дає можливість інтерпретувати отриманий опис набору даних як ймовірність того, що випадково взятий з набору наданих даних елемент буде мати значення, що лежить в тому чи іншому інтервалі.

Але формально - це і є визначення (емпіричної) ймовірності події, з якого ми почали!

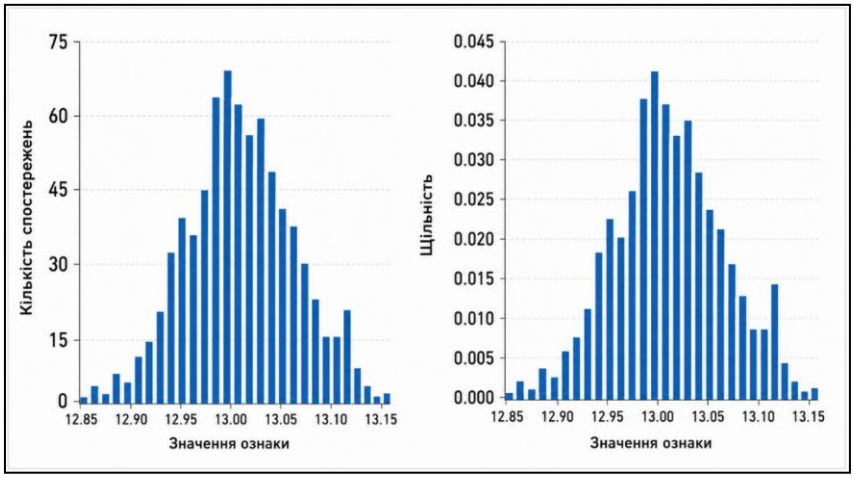


Рис. 2.12 Представлення даних на гістограмі в абсолютних значеннях та в нормованому вигляді

Детальніше про побудову гістограми за допомогою засобів бібліотеки *Matplotlib* можна ознайомитися в [26].

Головна *відмінність між гістограмою та PDF* полягає в тому, що замість частотних стовпчиків при візуалізації гістограми при відображенні PDF використовують неперервну криву, під якою площа над інтервалом відповідає ймовірності попадання випадкової величини в цей інтервал (Рис.2.13). Це узагальнення добре проявляється тоді, коли *ми уявимо*, як висоти стовпчиків гістограми стають все нижчими, кількість інтервалів зростає, а їх ширина стає все вужчою - у граничному випадку ми отримаємо неперервну функцію.

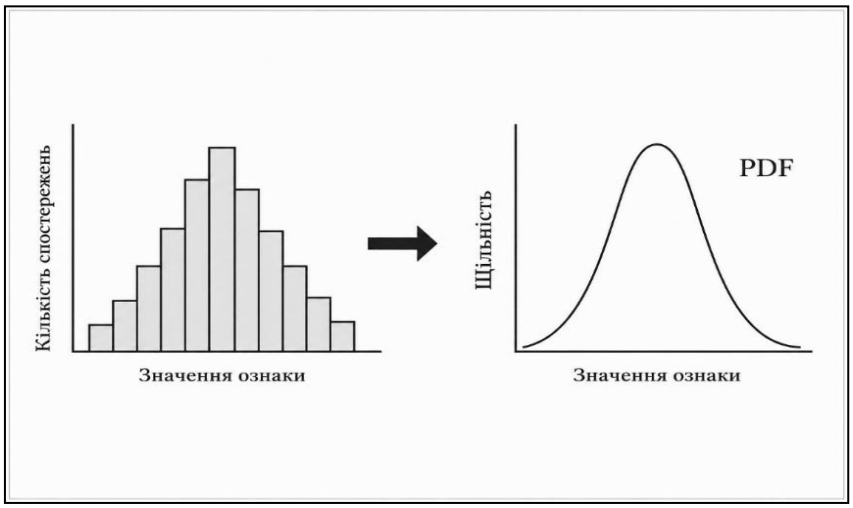


Рис. 2.13 Гістограма та функція щільності розподілу

2.2.4. Кумулятивна функція розподілу

Наступним кроком у математичному описі розподілу є побудова ***кумулятивної функції розподілу*** (англ. *Cumulative Distribution Function*), скорочено - CDF). Вона дає відповідь на запитання: *яка ймовірність того, що випадкова величина прийме значення, менше або рівне обраного значення*. Прикладом практичної задачі, яка по суті не вимагає нічного більшого, ніж знання кумулятивної функції розподілу, може слугувати запит «На основі попереднього досвіду та накоплених даних трафіку визначити, яка ймовірність того, що трафік мережі за секунду не перевищить 500 КБ».

Формально, для безперервної випадкової величини X, функція CDF визначається як:

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)dt$$

де *f(t)* - це функція щільності ймовірності (PDF).

Кумулятивна функція має низку корисних властивостей:

– вона є неспадаючою функцією, тобто, $x_i < x_j \Rightarrow F\left(x_i\right) \le F\left(x_j\right)$;

– вона змінюється в діапазоні від 0 до 1;

$F\left(-\infty\right) = 0; F\left(+\infty\right) = 1;$

Графік CDF має вигляд плавної кривої, що монотонно зростає від 0 до 1. Для прикладу, для нормального закону розподілу, графіки функцій PDF та CDF так, як показано на Рис. 2.14.

Імовірність того, що безперервна випадкова величина прийме значення між a і b дорівнює приросту CDF на цьому інтервалі:

$$P\left(a < x < b\right) = F\left(a\right) - F\left(b\right)$$

Стає зрозумілим, що імовірність того, що безперервна функція набуде одного, заздалегідь визначеного, значення дорівнює 0. Це випливає з того, що

$$F(a) - F(a) = 0$$

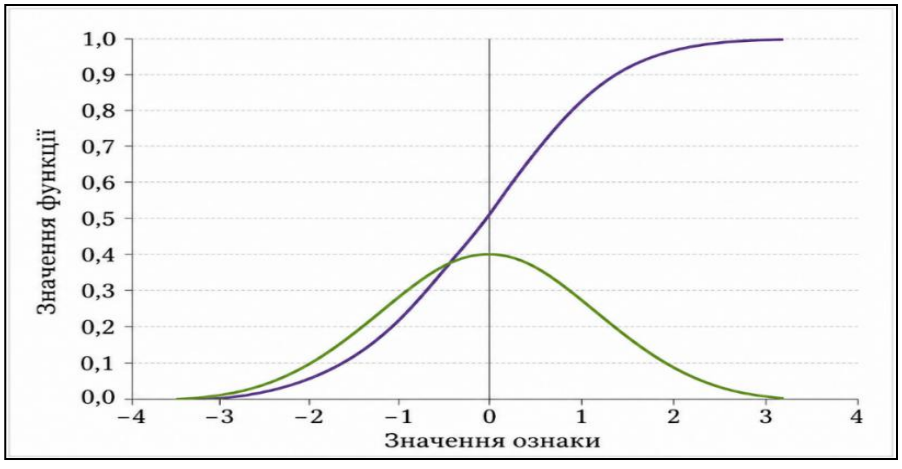


Рис.2.14 Кумулятивна функція розподілу та функція щільності розподілу

Важливе зауваження. *Формально, з точки зору теорії ймовірності, значення функції щільності в певній точці не є ймовірністю, а лише індикатором*

того, наскільки ймовірним є наближення значення випадкової величини до цієї точки. [22],[23]

У випадку дискретної випадкової величини, також можливо побудувати кумулятивну функцію розподілу, яка в даному разі набуває вигляду сходинкової функції. Вона визначається як сума ймовірностей усіх можливих значень, менших або рівних заданому:

$$F(x) = P(X \le x) = \sum_{x_i \le x} P(x_i)$$

де x_i - значення, які може набувати дискретна випадкова величина.

Графічно (див. Рис.2.15) така функція виглядає як послідовність горизонтальних відрізків (сходинок), що «підскакують» на розмір відповідної ймовірності при кожному новому значенні випадкової величини. На Рис. 3.8 також можна ознайомитися з виглядом, якого набуває графік кумулятивної функції розподілу в реальних задачах.

Треба розуміти і завжди пам’ятати, що вхідний набір вимірюваних значень (т.з. *«сирі данні»*) випадкової величини - це *найбільш повний (з точки зору наявної інформації) її опис.* Однак з ним працювати не завжди зручно, аналізувати – досить складно. Алгоритми обробки можуть вимагати надзвичайно багато пам’яті та/або за обчислювальною складністю можуть бути відчутно неефективними. Більшість методів, які будуть розглядатися далі, в той чи інший спосіб побудовані на скороченні інформації, яка обробляється. Гістограми та функції щільності розподілу – один з засобів, які зменшуючи об’єми даних видаляють з них найменшу кількість інформації, а отже залишають досліднику надію отримати точні та адекватні результати. У ситуації, коли розподіл значень дійсно досить точно описуються деякою функцією, працювати з функцією як з об'єктом при рішення задач виявляється і простіше, і майже так само ефективно, як з сирими даними. Особливо це проявляється у випадках, коли дані підпорядковуються одному з відомих *теоретичних законів* розподілу (див. Розділ 2.4). В даному випадку «підпорядковується» - це значить, що обрана функція розподілу досить точно описує значення, які ми спостерігаємо в експерименті (*емпіричні значення*). Зрозуміло, що при цьому відразу виникає питання - а як перевірити, чи *насправді* така відповідність існує? І таке

припущення потрібно попередньо довести – або спростувати - в ході аналізу, зокрема методами перевірки гіпотез(див. Розділ 3). Фіксуємо цю задачу для подальшого вирішення.

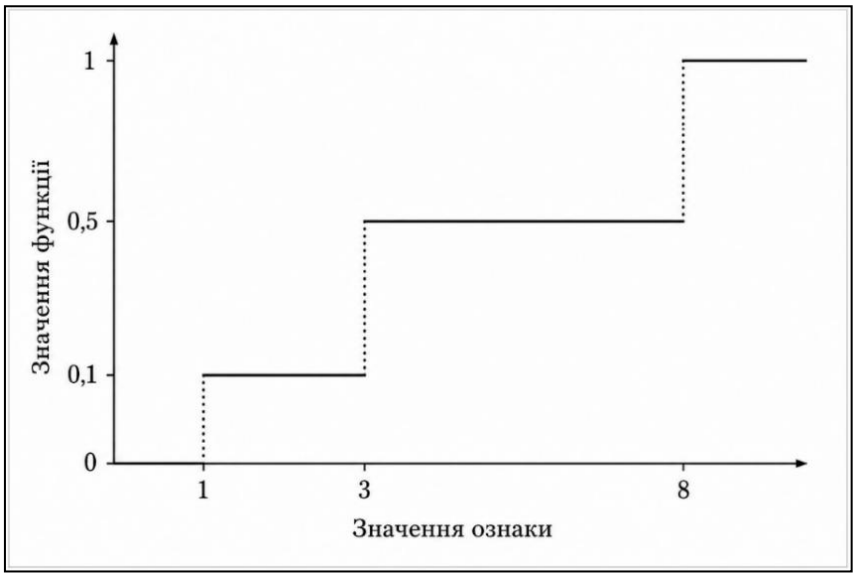


Рис. 2.15 Приклад кумулятивної функції розподілу для дискретної випадкової величини

Незважаючи на свою простоту, використання гістограм дозволяє вирішити досить серйозні задачі реального світу. Наприклад, навіть така проста і зрозуміла задача, як визначення фальшивості (тобто несиметричності) монети або кубика може бути вирішена виключно спираючись на аналіз гістограми отриманих результатів.

На Рис. 2.16 наведений більш складний приклад. За трафіком в мережі ведеться спостереження, в результаті якого отримано гістограму, що показує, яка кількість яких за довжиною пакетів курсують в мережі. При фіксації зміни в гістограмі логічно припустити, що ця зміна відображає певні зміни в поведінці трафіку, які в свою чергу виникли через певну подію, про яку ми можемо навіть не знати. В даному разі – хтось з користувачів почав користуватися torrent-ом. Протокол BitTorrent передбачає в процесі пошуку peer'а, який містить черговий шматок потрібного файлу, відсилку великої кількості коротких повідомлень-запитів, та частого отримання у відповідь ще більш коротких сигналів "Ні". Такі короткі пакети будуть відображені на гістограмі, що може трактуватися системним адміністратором як сигнал про подію і шукати хто саме змінив

гістограму, коли це відбулося, тощо. Іншими словами, *відхилення гістограми від очікуваного вигляду - це ознака проблеми.* Оскільки ми розуміємо, що наші дані – це певна скінчена вибірка, і ідеального співпадіння ані з очікуваною гістограмою, ані з «попередньою, нормальною» ніколи не буде, то як саме визначити межу, після якої можна вважати, що спостерігається певна аномалію? Поки що цю проблему ми також для себе фіксуємо, але пошук відповіді на неї поки відкладемо.

2.2.5. Вибір кількості інтервалів при побудові гістограм

Насамкінець треба обговорити питання, яке виникає при застосуванні гістограм при дослідження даних. Справа в тім, що вигляд гістограми істотно залежить від обраної кількості інтервалів. (Рис. 2.17). Якщо інтервалів занадто мало - розподіл виглядає надто грубо, втрачаються важливі деталі. Якщо ж інтервалів занадто багато, кількість значень, що потрапляють в кожен з інтервалів стає занадто малою, в деяких випадках дорівнює 0, і гістограма теж втрачає інформативність. Тому правильний вибір кількості інтервалів - ключ до коректної інтерпретації отриманих результатів.

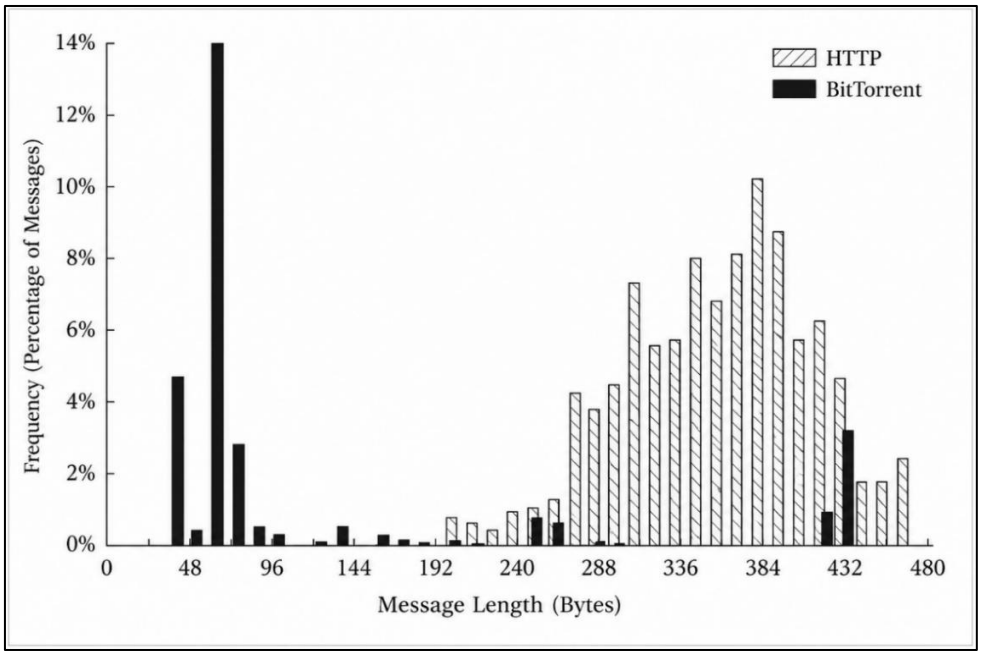


Рис. 2.16 Приклад використання гістограм при аналізі мережевого трафіку

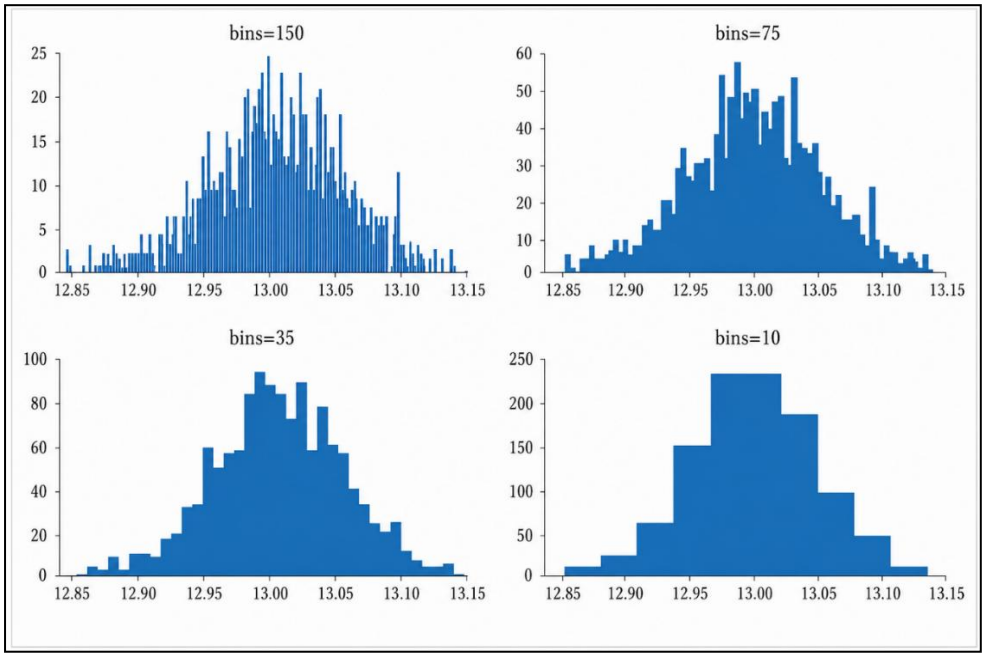


Рис. 2.17 Приклад гістограми при різній кількості інтервалів розбиття

Нажаль, єдиного формально обґрунтованого методу, що дозволяв би обрати оптимальну кількість інтервалів для довільної гістограми на сьогодні не існує. Випробуваний і застосовуваний більшістю дослідників метод є метод спроб та помилок, в якому дослідник намагається знайти деяку розумну - з певної точки зору - «золоту середину» між ситуаціями втрати інформативності гістограм, що наведені вище. Існує декілька найбільш поширених емпіричних підходів для визначення оптимальної кількості інтервалів k для вибірки розміром n:

– *Правило Стерджеса (*англ. *Sturges' rule)*:

$$k = \log_2 n + 1$$

Це одне з найпростіших і найпоширеніших правил. Воно добре працює для даних, близьких до нормального розподілу, однак для великих наборів даних цей підхід схильний до примусового зменшення кількості інтервалів.

– *Правило Райса (*англ. *Rice's rule)* пропонує простішу формулу, яка часто використовується для великих вибірок:

$$k = 2\sqrt[3]{n}$$

– *Метод Скотта (*англ. *Scott's rule)*:

$$h = \sigma * \sqrt[3]{\frac{24 * \sqrt{\pi}}{n}} \approx \frac{3.5\sigma}{\sqrt[3]{n}}$$

де σ - середньоквадратичне відхилення значень даних, h – ширина інтервалів. Тоді

$$k = \frac{\max(X) - \min(X)}{h}$$

Це правило також добре працює для даних, близьких до нормального розподілу.

– *Правило Фрідмана–Діаконіса (*англ. *Freedman–Diaconis rule)*:

$$h = \frac{2\left(Q_{75} - Q_{25}\right)}{\sqrt{N}}$$

де Q_{25} та Q_{75} – відповідно 25% та 75% процентілі множини значень (див.Розділ 2.3.2). Цей підхід краще підходить для розподілів, що суттєво відрізняються від нормального або мають значну кількість викидів.

Для деяких задач статистичного аналізу даних, де побудова гістограм (причому часто навіть без візуалізації) є одною з складових процесу дослідження (наприклад в задачах перевірки гіпотез однорідності/узгодженості законів розподілів, в задачах виявлення точок змін в часових рядах, тощо) використовуються і інші підходи до вибору кількості або розміру інтервалів [26], [27].

Отже, коротко, рекомендації по вибору кількості інтервалів гістограми можна сформулювати так:

- Для невеликих вибірок (n<200) зручно використовувати правило Стерджеса.
- Для великих обсягів даних - правило Райса або Фрідмана–Діаконіса.
- Для кращого сприймання людиною кількість інтервалів зазвичай вибирається в межах від 5 до 20.
- У всіх випадках інтервали бажано робити однакової ширини (крім спеціальних випадків).

– Інтервали повинні включати всі дані, навіть викиди.
– Завжди перевіряйте, чи не спотворює вибране розбиття загальний вигляд розподілу.

Контрольні запитання

1. Скільки подій ми можемо зафіксувати в експерименті, що складається з викидання двох гральних кубиків?
2. Які параметри впливають на вигляд гістограми?
3. Чим відрізняється закон розподілу дискретної випадкової величини від функції щільності ймовірності неперервної випадкової величини?
4. Що показує функція щільності ймовірності і як інтерпретується площа під графіком PDF на заданому інтервалі?
5. Яке значення має кумулятивна функція розподілу в точці x_i і як вона пов'язана з функцією щільності ймовірності?

2.3. Кількісні характеристики наборів даних

Одна з перших речей, яку має зробити дослідник, який вивчає стан або поведінку деякої системи - проаналізувати множину параметрів, які в сукупності дають уявлення про систему чи процес, а отже і про особливості функціонування та поведінки об'єкту, за яким ведеться спостереження. Вище був представлений один з найпоширеніших інструментів дослідження та аналізу даних – а саме їх гістограми та (для випадку неперервних змінних) функції розподілу ймовірностей. Нагадаємо: гістограма представляє або частоти різних значень, що зустрічаються в множині наданих даних, або частоти значень, що потрапляють в (здебільшого рівномірні) інтервали можливих значень. Гістограма може бути легко і зрозуміло візуалізована у вигляді відповідного графіку. Цей інструмент дозволяє побачити загальну форму розподілу, оцінити його симетрію, ширину, наявність "хвостів" чи піків. Але для аналітичних цілей цього недостатньо: модель або алгоритм, в подальшому реалізовані в вигляді програмного застосунку, потребують числового представлення значень параметрів, а не просто графіків.

Тому наступним логічним кроком дослідження є узагальнення розподілу через кількісні характеристики, які стисло передають його основні властивості - *положення, розсіювання, форму* тощо. Ці (і інші, які будуть розглянуті далі) характеристики даних, що описують їх «в цілому», в вигляді певного значення,

ще називають «статистиками». Формально, ***статистика – це довільна функція над множиною даних, областю значень якої є множина дійсних чисел і яка не залежить від параметрів розподілу даних.*** Тобто, статистика може бути однаково розрахована для будь якої наданої сукупності даних. А от ***параметри законів (функцій) розподілу*** визначаються на основі статистик, але вже по різному для різних законів розподілу. Це означає, що дослідник може за наявними у його розпорядженні даними, та незалежно від їх (даних) властивостей завжди знайти значення статистик, а отже - ґрунтуючись на цих значеннях зробити об'єктивні та статистично обґрунтовані оцінки та висновки.

2.3.1. Мода

Почнемо з найбільш універсальної статистики множини даних. І це буде не звичне нам з побутових прикладів «середнє», а така характеристика, як **мода** (англ. *Mode*. Термін був вперше введений Карлом Пірсоном ще в 1894 році. Для простоти сприймання трактуйте слово «мода» як синонім слова «типовість»). ***Мода - це значення (одне чи декілька), яке (які) найчастіше трапляється у множині даних.*** Тобто мода визначає *найбільш типове* значення випадкової величини.

Універсальність моди полягає в тому, що вона, *на відміну від усіх інших статистик положення* (вони розглядаються далі), може бути обчислена та матиме смислове навантаження незалежно від природи даних - кількісні вони, категоріальні чи номінальні. А відтак – за допомогою моди можна знаходити відповіді на досить серйозні та важливі прикладні задачі. Наприклад, маючи логи інцидентів кібербезпеки за місяць та визначивши моду, можна зрозуміти куди є сенс першочергово спрямувати зусилля для підвищення рівня захищеності системи. Проаналізувавши розміри пакетів, що передаються, можна визначити найтиповіший розмір пакета в мережевому трафіку і відповідним чином налаштувати фаєрволи. З експертної оцінки товарів, за допомогою моди визначають найпопулярніші типи продукту (наприклад - при прогнозі продажів чи плануванні їх виробництва) тощо.

Мода, як статистика множини даних, частіше вживається тоді, коли дані мають нечислову природу. В такому разі мода досить просто визначається навіть візуально, за відповідною гістограмою. Вона відповідає стовпчику, який є найвищим на гістограм. Однак у випадку, коли гістограма будувалася за визначеними інтервалами (тобто – дані виміряні у кількісній шкалі), точне значення моди описаним вище чином визначити вже не вдасться. В такому

випадку замість моди говорять про *«модальний інтервал»,* маючи на увазі інтервал, в якій потрапляє найбільша кількість даних. У випадку безперервних значень даних, що досліджуються, та за умовою, що відповідна функція щільності ймовірності (PDF) відома, ***модою вважається таке значення x, при якому значення PDF досягає максимуму.***

Мода є показником (статистикою), що стійкий до викидів, тобто не змінюється при додаванні кількох екстремальних значень. Залежно від кількості мод множини даних, що досліджується, можуть визначатися як «унімодальна», «бімодальна», та «мультимодальна», тобто такі, що мають одну, дві, або більше мод відповідно. Мультимодальність даних для дослідника часто є причиною розбиратися з джерелом її виникнення. В наведеному раніше прикладі гістограми розмірів пакетів в трафіку мережі (Рис 2.16) за результатами появи мультимодальності можна було рекомендувати почати пошук причини того, що змінилося в поведінці системи, моменту часу коли ця зміна відбулася тощо.

В певних випадках мультимодальність (або точніше - локальна модальність) є природною форму розподілу *і є індикатором того, що вибірка не є однорідною*, а самі спостереження (надані дані), можливо, породжені двома або більше «накладеними» розподілами. Тобто в наданих даних присутні декілька різних за своєю природою типів об'єктів. Так може виглядати вибірка даних про зріст людей в групі, в якій присутні і чоловіки і жінки. У соціологічних опитуваннях, в разі якщо надані данні відображають перевагу або ставлення до чогось, локальна модальність може означати, що існують кілька виразно різних думок.

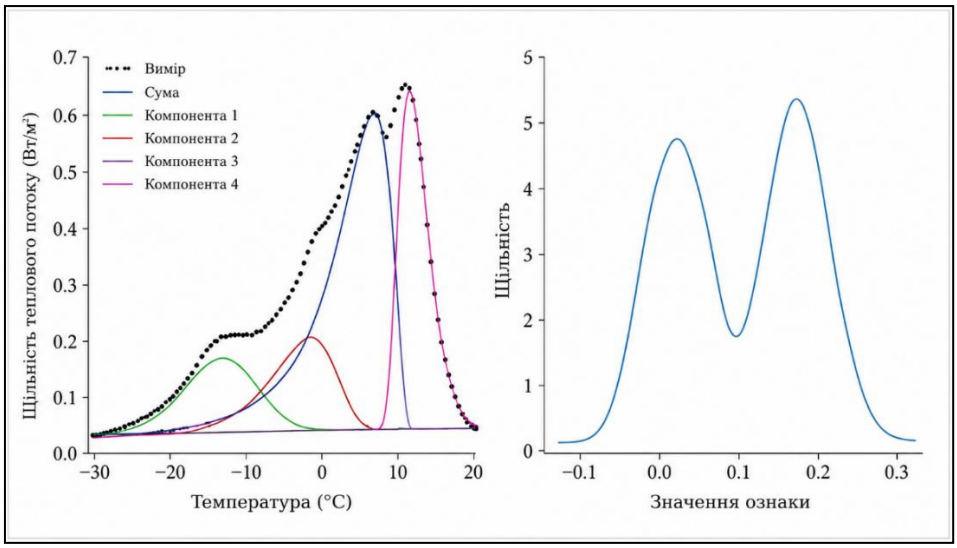


Рис. 2.18 Приклади мультимодальних розподілів даних

Мода дозволяє визначити домінуючі тенденції та найбільш типові значення набору даних, але в реальних аналітичних завданнях часто важливо розуміти не лише «найпоширеніший випадок», а й структуру розподілу значень у цілому. Наприклад, при аналізі затримок у мережевому трафіку чи часу відповіді web-сервера може виявитися, що більшість запитів обробляється швидко (мода), але є довгі «хвости» в розподілі затримок, саме які і впливають критично на якість сервісу або безпеку. Для більш глибокого аналізу структури розподілу даних, особливо коли вони не є симетричними або є мультимодальними, потрібні більш потужні статистичні інструменти.

2.3.2. Квантилі та статистики розкиду

Таку, більш повну характеристику розподілу, надають ***квантилі***. Якщо мода фокусується на одному значенні, то квантилі дозволяють поділити множину даних на частини, надаючи інформацію про те, які значення мають, наприклад, 25%, 50% або 75% найменших (найбільших) значень серед усіх наявних даних. Тобто, *мода відповідає на запитання: «Яке значення зустрічається найчастіше?», тоді як то квантилі відповідають на запитання: «Які значення відокремлюють певну частку даних від решти?».* Це надає певне розуміння *розкиду значень* наданої множини даних, допомагає в виявленні викидів, асиметрії розподілу та більш повного опису набору даних.

Формально, ***квантиль*** - це рішення відносно X рівняння:

$$F(X) = p$$

де p - задана ймовірність. Тобто, p-квантиль - це значення x_p, для якого виконується умова $P(X \leq x_p) = p$, де $p \in (0;1)$. Наприклад, *медіана* (див. далі) є 50%-квантилем. Квантилі 25%, 50% та 75%, іноді називають ***квартилями***, та позначають, відповідно, як перший квартиль ($Q1$), другий квартиль ($Q2$), третій квартиль ($Q3$)). Статистики $Q1, Q2$ та $Q3$ ділять дані з наданої множини на чотири рівновеликі (за кількістю елементів в них) частини.

(Для інформації. Іноді можна зустріти і такі статистики розподілу, як децилі. Децилями називають такі значення x, які ділять набір даних на групи по 10% від загального обсягу кожна*.)*

Похідною від названих статистик є статистика, яка називається «***міжквартильний розмах***» (англ. *Inter-Quartile Range, IQR*), і яка визначається за формулою:

$$IQR = Q3 - Q1$$

Міжквартильний розмах є мірою розсіювання даних і використовується, зокрема, для виявлення викидів. Відображаючи розкид середніх 50% спостережень, він є стійкішим до *викидів (аномально великих або аномально малих значень)* порівняно з *розмахом (різницею між максимумом і мінімумом)*, дозволяючи аналізувати дані без врахування впливу таких значень.

В найпростішому - хоча і дещо спрощеному - випадку, аномальними вважають дані, які лежать поза інтервалом:

$$\left[Q1 - 1.5 * IQR;\ Q3 + 1.5 * IQR\right]$$

тобто «далеко» від квартилю *Q2* вибірки, в так званих «*хвостах функції розподілу ймовірності*» (До речі, «хвіст розподілу» - не жаргон, а досить стійке, прийняте в статистиці позначення вказаної зони).

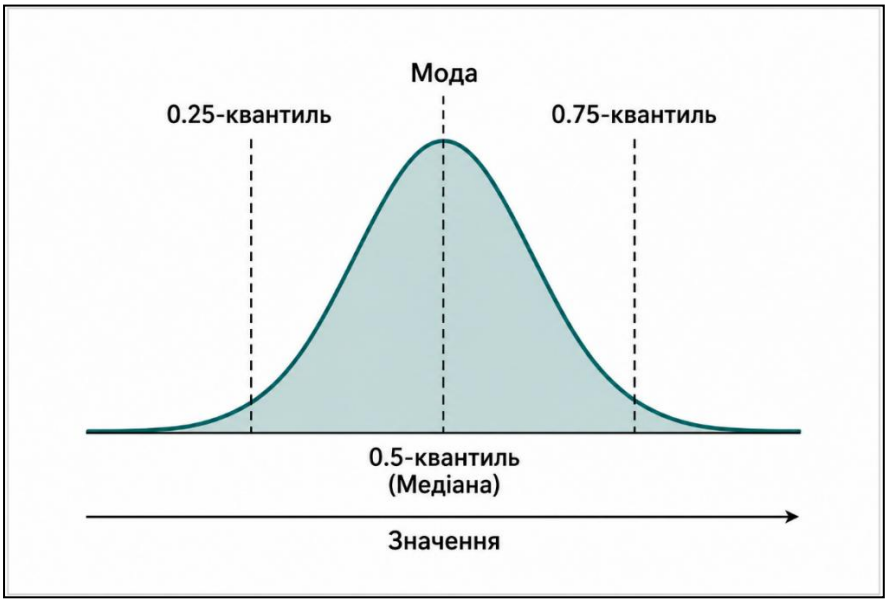


Рис. 2.19 Мода та квартилі на графіку функції щільності розподілу (Приклад)

Іноді вказані статистики відображають в вигляді графіку, знаному як «ящику з вусами» - візуальної діаграми (англ. *Box and Whisker Plot*, або скорочено «*boxplot*»), яка природним та зрозумілим чином пов’язана з гістограмою.

Окрім визначення розкиду даних, квантилі та *IQR* можна використати для *попереднього аналізу асиметрії та скосу* у розподілі даних. Так, асиметрія проявляється через нерівномірність відстаней між медіаною та крайніми квартилями:

– Правосторонній скос: медіана ближче до *Q3*, ніж до *Q1*:

–

$$(Q2 - Q1) > (Q3 - Q2);$$

– Лівосторонній скос: медіана ближче до *Q1*, ніж до *Q3*:

–

$$(Q3 - Q2) > (Q2 - Q1);$$

– Симетричний розподіл:

відстані від *Q2* до *Q1* та *Q3* приблизно рівні.

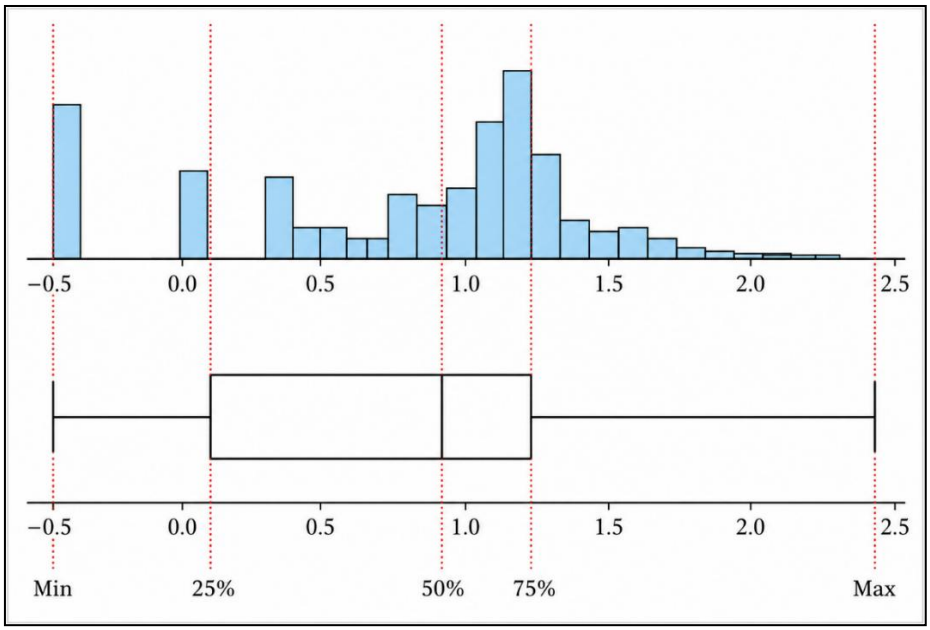


Рис.2.20 Діаграма «ящик з вусами» та її зв’язок з гістограмою.

В дійсності, асиметрія зустрічається на практиці дуже часто. Загальнозрозумілим прикладом асиметрично розподілених даних є, наприклад, розподіл доходів громадян у багатьох країнах, який часто має негативну асиметрію (лівосторонній скос): невелика кількість людей має дуже високі доходи, тоді як більшість має нижчі або помірні доходи. Хотілося б мати певний кількісний показник (статистику), на базі якого виконувати подальший аналіз даних. Такий показник носить назву ***коефіцієнт асиметрії.*** Він задає стандартизовану міру асиметрії (або відсутності симетрії) в розподілі, що дозволяє порівнювати різні набори даних або їх розподіли, показуючи, чи дані зміщені вліво, вправо чи є симетричними. Таких коефіцієнтів було запропоновано декілька.

Коефіцієнт асиметрії Боулі (англ. *Bowley's Coefficient of Skewness*)– це статистичний показник, який, на відміну від інших показників асиметрії, які базуються на середньому значенні та стандартному відхиленні, використовує виключно квартилі. Це надає йому переваги в питання стійкості до викидів у даних. Він визначається за формулою:

$$BC = (Q3 - Q2) - (Q2 - Q1)/(Q3 - Q1) = (Q3 + Q1 - 2Q2) / IQR$$

Коефіцієнт може приймати значення від мінус нескінченності до плюс нескінченності. Нульове значення вказує на ідеально симетричний розподіл, тоді як додатні значення вказують на позитивну асиметрію (зміщення праворуч), а від'ємні значення вказують на негативну асиметрію (зміщення ліворуч).

(Далі будуть розглянуті і інші коефіцієнтами асиметрії. Бажаючи можуть також ознайомитися з оглядом та порівнянням таких статистик, наприклад, за посиланням [28]).

2.3.3. Вибірка та генеральна сукупність

Перед тим, як продовжити розбиратися з статистиками наборів даних, необхідно ввести деякі поняття, які відіграють надзвичайно важливу роль у аналізі даних. Це такі поняття як вибірка та генеральна сукупність.

Дотепер ми висловлювалися дещо абстрактно – «набір» (або «множина») значень, які були надані (отримані) для проведення подальшого аналізу. Далі нам доведеться розібратися з тим, що в дійсності існують *два різних типи таких множин даних* і в залежності від завдань які визначені для дослідження,

доведеться користуватися одним чи іншим. І навпаки - різні типи множин будуть зумовлювати від дослідника вирішення різних завдань та застосування різних методів та алгоритмів.

Спочатку надаймо деякі визначення. ***Генеральна сукупність*** (англ. *Statistical Population*) - це *вся множина* об'єктів, явищ або подій, яка може бути розглянута (часто - тільки теоретично) у межах дослідження. (Іноді замість терміну «генеральна сукупність» вживають термін «*популяція*»). ***Вибірка*** (англ. *Sampling*) - це *частина генеральної сукупності*, що безпосередньо надана (отримана, доступна) для виконання дослідження. (Рис 2.21)

Прикладом спорідненості між вказаними різновидами наборів даних, але водночас і проявом їх відмінностей, є дані, що надаються при описі деяких «параметрів» громадян країни (вік, стать, освіта, місце проживання, відношення до певних подій тощо). В такому разі генеральна сукупність представляєтеся в вигляді результатів суцільного перепису населення, а вибірки – у вигляді результатів вибіркового опитування. У сфері кібербезпеки прикладом генеральної сукупності може бути весь мережевий трафік підприємства за рік, усі спроби входу в систему за місяць або всі пакети даних у конкретному сегменті мережі. А для аналізу трафіку за рік можна взяти дані за одну годину кожного дня або випадкові 10% пакетів, що проходять через маршрутизатор - і це буде вибірка (одна з можливих) з генеральної сукупності. Ще один приклад. Якщо перед дослідником поставлена задача визначити – наприклад - фізичні можливості курсантів інституту, можливо всіх курсантів примусити здати відповідні тести і отримати результати. В даному випадку це буде генеральна сукупність. Але можна обмежитися лише деякими випадково обраними курсантами і визначити їх результати - і це буде вибірка з генеральної сукупності. Однак, якщо поставлена задача визначити ті самі можливості усіх студентів країни, то отримати генеральну сукупність таких даних буде практично неможливо. І тоді інформація про результати курсантів інституту може розглядатися як одна з можливих вибірок і використовуватися в такому вигляді для пошуку відповіді на поставлене запитання.

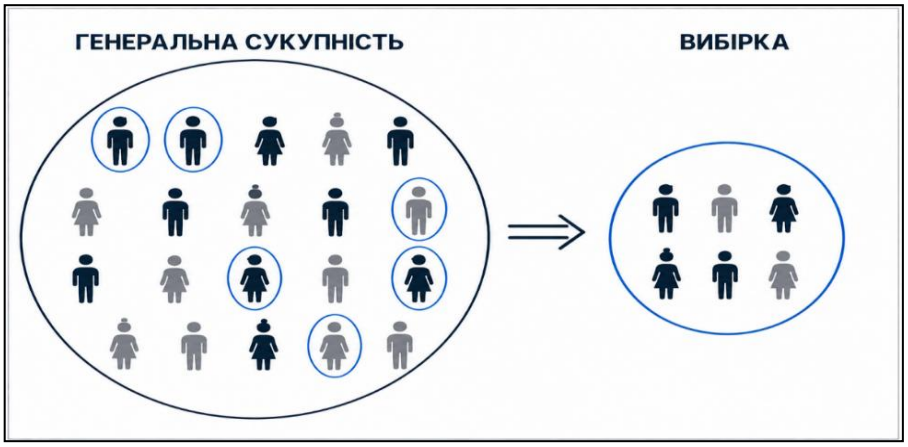


Рис 2.21 Генеральна сукупність та вибірка

Тобто, *ключова особливість генеральної сукупності полягає в тому, що вона і тільки вона описує повну картину процесу.* Проте через великий обсяг або складність збору даних, через їх нескінченність за своєю природою, через неможливість отримати безпосередній доступ до всіх даних тощо, а іноді – взагалі через те, що генеральна сукупність є деякою абстракцією, така інформація може виявитися практично недосяжною. В такому разі використовують деяку її доступну підмножину – вибірку. Звідси постає *перша і найважливіша задача аналізу* - ***як на основі обмежених даних вибірки зробити висновки про всю генеральну сукупність даних?***

Вибіркове дослідження - це метод статистичного дослідження, при якому узагальнюючі показники сукупності встановлюються тільки по окремо взятій її частині. Основна ідея вибіркового методу дослідження полягає в постулаті: «*те, що вірно для вибірки, вірно і для генеральної сукупності, з якої вибірка отримана».* Якщо це так, то це дає можливість поширити результати, зроблені для вибіркових даних на всю генеральну сукупність. Наприклад, було з'ясовано на певній вибірці пацієнтів, що якісь ліки дають позитивний ефект лікування. Можна припустити, що цей ефект буде спостерігатися у всіх людей на планеті. І відразу постає наступне питання: якщо ми спостерігаємо вибірку а результат дослідження хочемо розповсюдити на всю генеральну сукупність, то ***наскільки ми можемо бути впевненими в правдивості такого розповсюдження?*** Наприклад – чи допустимо розповсюдження результатів аудиту безпеки на всю інфраструктуру? Якщо - припустимо - з 1000 серверів випадково перевірити 100 і виявити при цьому 5 критичних вразливостей, чи можна з певною ймовірністю оцінити загальну кількість вразливостей у всій інфраструктурі підприємства?

Проблема полягає в тому, що зробити дійсно репрезентативну вибірку з генеральної сукупності також часто виявляється практично неможливо. У разі хворих нам довелося б, наприклад, задіяти людей, які проживають на різних континентах, належать до різних рас, ведуть різний спосіб життя тощо. Побудова репрезентативної вибірки - це окреме і дуже складне завдання, однак *якщо вибірка буде нерепрезентативною, поширення отриманих за допомогою вибірки результатів на всю генеральну сукупність можуть виявитися некоректним.* Ось майже анекдотичний приклад такої нерепрезентативної вибірки і відповідно - некоректного висновку: «Яку статистику не візьми, ймовірність смерті на лікарняному ліжку втричі перевершує ймовірність летального результату на дому. Висновок - ні в якому разі не лягати в лікарню». Тут підступно змішуються дві групи населення – здорові та завідомо хворі, і результат отримуємо абсолютно неадекватний реальності.

Ще один приклад, який показує як важливість так і складність створення репрезентативних вибірок. Журнал «Literary Digest» проводив опитування громадської думки в США перед виборами президента у 1920, 1924, 1928 та 1932 роках. Тоді вважалося, що чим більше людей буде опитано (чим ближче вибірка буде до генеральної сукупності), тим точніше буде прогноз. І це здавалося справджувалося: щоразу прогноз, складений з урахуванням опитування, виявлявся вірним. Але в 1936 році ситуація кардинально змінилася. Журнал передбачив, що кандидат від республіканської партії переможе в більшості штатів і стане президентом. Але насправді по результатам виборів він переміг тільки в 2 з 46 штатів, що тоді складали країну. Повний провал статистичного дослідження! Результати «аналізу помилок» показали, що основна помилка полягала в тому, що в гонитві за збільшення кількості опитуваних журнал перейшов на опитування з використанням телефонної книги. Тобто, з таких книг обирали випадковим чином номери та адреси власників, за якими і відправляли опитувальні листи. Було відправлено більше 10 мільйонів таких листів. З них 2,5 млн відповіли, що є неймовірно великою кількістю респондентів для будь-якого опитування, і здавалося б мало гарантувати точність результатів. Але наявність домашнього телефону на той час було найбільш вірогідною для людей з доходами, вищими за середні. Крім того, надіслали відповіді не всі, а люди не тільки досить впевнені в своїй думці, а ще й такі, що звиклі відповідати на листи. Тобто в значній частині - представники ділового світу, які здебільшого

підтримували республіканців. У результаті вибірка респондентів за своєю соціальною структурою помітно відрізнялася від генеральної сукупності виборців, що і вплинуло на результати. Репутація журналу була практично знищена, а сам він дуже швидко перестав виходити.

Явище, подібне до описаних вище, коли вибірка представляє не всю генеральну сукупність, а лише якийсь її шар, якусь її частину, називається *зміщенням (або зсувом) вибірки* (англ. *Sampling Bias*). Зсув - один з основних джерел помилок при використанні вибіркового методу. Іншою причиною зміщення може стати малий обсяг вибірки. Наприклад, поставлена задача визначити, чи дійсно хлопчики та дівчата народжуються з ймовірністю 50:50? Для проведення дослідження випадковим чином обрали деякий пологовий будинок. У ньому «сьогодні» народилося 7 хлопчиків і 3 дівчинки. Чи можна вважати, що сьогодні в усій Україні (або й у усьому світі) ймовірність народження хлопчиків дорівнює 0.7? В даному разі нерепрезантивність вибірки є наслідком малої кількості наявних даних. (Тут ми маємо занотувати ще одне питання, до якого в подальшому треба буде повернутися - *який саме об'єм вибірки має бути, щоб бути з певною ймовірністю впевненим в правдивості результатів дослідження*? В наведеному прикладі - скільки саме, 10, 100, 1000 чи мільйон немовлят треба дослідити, щоб отримати результати, яким можна було-б заданою мірою довіряти?)

Отже, якісна вибірка - запорука подальшого точного аналізу. Важливо, щоб вибірка репрезентувала генеральну сукупність, інакше висновки будуть спотвореними. Тому застосовують різні методи відбору даних вибірки:

- *Випадкова вибірка* (англ. *Random Sampling*) - кожен елемент генеральної сукупності має однакову ймовірність потрапити у вибірку.
- *Систематична вибірка (*англ. *Systematic Sampling)* - елементи відбираються за певним кроком. Цей метод часто використовується в техніці, де зразок вибирається для тестування з виробничої лінії (скажімо, кожні п'ятнадцять хвилин, або кожний двадцятий вироб), щоб переконатися, що машини та обладнання працюють відповідно до специфікацій. Аналізуючи кожен 100-й інформаційний пакет в комп'ютерній мережі можна скласти деяку характеристику трафіку, який по цій мережі передається.

– *Стратифікована вибірка* (англ. *Stratified Sampling*) - сукупність попередньо ділять на групи (страти) за певними ознаками, і вибірка береться з кожної групи пропорційно її розміру. У соціології це можуть бути різні професії, у медицині – вікова група пацієнтів, у фізіології – гендер особи тощо. У кібербезпеці це може бути, наприклад, трафік з різних сегментів мережі. Цей підхід дозволяє уникнути ситуації, коли в результаті випадкового відбору деяка група елементів зовсім не потрапляє в результуючу вибірку.

Перелічені методи мають сенс в разі, якщо дослідник має змогу вплинути на процес створення вибірок. Як вже було сказано, таке має місце далеко не завжди. Я такому випадку задача дослідника впевнитися в репрезентативності наданих даних, для чого часто доводиться заглиблюватися в вивчення прикладних аспектів поставлених завдань.

2.3.4. Вибіркові статистики та параметри генеральної сукупності

Як було зазначено вище, при аналізі даних, замість оперування всіма без виключення наданими значеннями набору даних, зручно оперувати деяким скороченим описом - можливо кількома числами. Вище ми застосовували для таких описів термін «статистика». Кожна статистика представляє тільки один з можливих аспектів набору даних. Отже, з одного боку статистики зручні в обробці, корисні в аналізі, а з іншого боку - не можна забувати, що вони є результатом «стискання» інформації, яка надана для дослідження. А отже при їх використанні деяка (іноді доволі велика) кількість інформації та подробиць опису даних стають недоступними для подальшого аналізу.

2.3.4.1. Показники центральної тенденції

При дослідженні даних часто цікавить питання: де “зосереджені” наші дані? Іншими словами*, яке значення можна вважати найбільш типовим або репрезентативним для всього набору спостережень?* Для цього використовують т.з. характеристики *центральної тенденції* - числові показники (статистики), які описують «центр», або «середній рівень» (як побачимо далі – в одному з багатьох можливих розумінь цих термінів) розподілу значень. Найвідоміші серед показників центральної тенденції - *середнє арифметичне, медіана та мода.* Кожен із цих показників має свої переваги та обмеження, і вибір між ними

залежить *від шкали виміру даних, від поставленої мети подальшого аналізу та методів його проведення*.

Про медіану та моду ми говорили в одному з попередніх розділів. Нагадаємо, що *мода* - це значення, яке найчастіше зустрічається в розподілі. Одна з переваг моди над медіаною та середнім значенням полягає в тому, що моду можна знайти (вирахувати) як для числових, так і для категоріальних (нечислових) даних. Існують деякі обмеження щодо використання моди як описової статистики набору даних. У деяких розподілах мода може бути досить віддалена від значень, про яку людина має уявлення про як про центр (хай і гіпотетичний) розподілу. Наприклад, якщо відобразити на графіку дані, що задаються наступною відсортованою множиною – кількістю рейтингових балів, отриманих курсантами групи:

65,65,65,65,70,71,72,72,73,75,80, 82,84,85,85,90,95,95,98

(див Рис.2.22) та спробувати попросити вказати їх «центральне» значення, то здається, що найлогічнішою відповіддю виглядало б значення «82», хоча мода при цьому дорівнює «65». Також можливо, що для одного й того ж розподілу даних існує більше однієї моди (бімодальний або мультимодальний розподіли), що обмежує здатність моди описувати центр або типове значення розподілу в завданнях, де потрібно визначити єдиний показник для опису центру. У таких випадках логічнішим виглядає використання медіани або середнього значення.

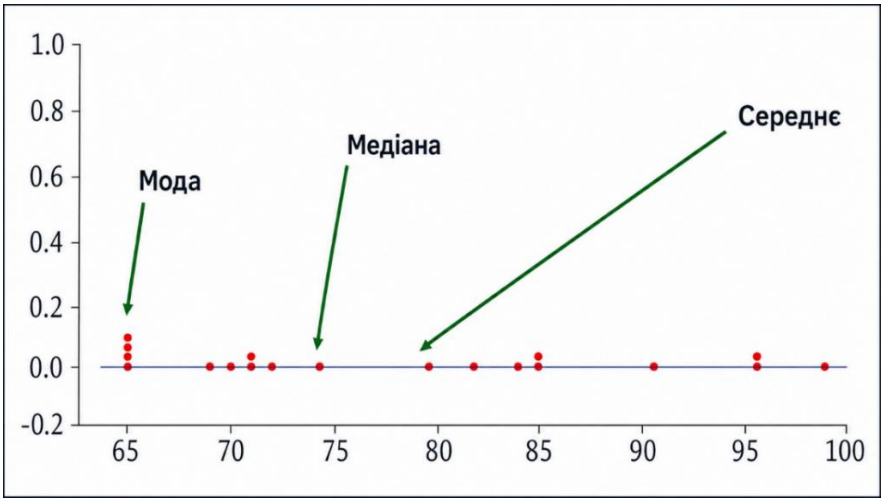


Рис 2.22 Значення набору даних та їх статистики центральної тенденції

Медіана (англ. *Median*) *– це серединне значення в послідовності значень які розташовані у порядку зростання або спадання.* Медіана ділить розподіл навпіл (по 50% спостережень знаходяться по обидва боки від медіанного значення). В наведеному вище прикладі з оцінками значенням медіани буде «75». (В наведеному вище наборі даних 9 значень менші за 75, а 9 - більші). У розподілі з непарною кількістю спостережень медіанне значення дорівнює середньому арифметичному двох серединних чисел.

Медіана менше зазнає впливу викидів та асиметрії даних, ніж розглянуте нижче *середнє значення.* Зазвичай медіана є кращим показником центральної тенденції у випадках, коли розподіл не є симетричним. Зрозуміло, що медіану неможливо визначити для номінальних даних, оскільки в такому разі значення неможливо логічним чином впорядкувати.

Мода і медіана визначаються однаково, що для генеральної сукупності, що для вибірок з неї. Зауважте, *мова йде про однакові алгоритми визначення цих показників, а не про однакові значення, яку будуть отримані в результаті роботи цих алгоритмів.* Наступний показник - середнє значення - має певну відмінність при роботі з генеральною сукупністю та вибірками.

2.3.4.2. Середнє арифметичне та математичне очікування.

Середнє арифметичне *(*часто - просто *«середнє»;* англ. *Mean; термін «Average» також вживається, але в даному контексті значно рідше)* є найпопулярнішим і найвідомішим показником центральної тенденції. Його можна використовувати як з дискретними, так і з безперервними даними, хоча найчастіше він використовується саме з безперервними. *Середнє арифметичне дорівнює сумі всіх значень у наборі даних, поділеній на кількість значень у наборі даних.* Воно позначається як $\bar{x}$ і визначається за формулою:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

де x_i-елементи вибірки, n - кількість елементів в вибірці.

В неведеному вище прикладі з оцінками курсантів значенням середнього є число «78.26315789473684».

При спробі таким чином описати генеральну сукупність, а саме застосувати визначення:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

де x_i-елементи генеральної сукупності, N - кількість елементів, відразу виникає питання. А як ми можемо визначити як значення N, так і всі можливі (!) значення елементів сукупності, коли генеральна сукупність визначається як сутність, всі значення якої ми отримати не можемо? Тобто, *значення статистики* μ *найчастіше на практиці визначити неможливо*. Воно може бути відомим з деяких теоретичних, часто позастатистичних, міркувань. Такі статистики називають «*параметрами*», маючи на увазі, що вони є параметрами функцій - а отже, математичних абстракцій - які описують теоретичний розподіл даних, що вивчаються. В даному раз йдеться про параметр, що зветься ***«математичне очікування»*** (англ. - *Mathematical Expectation; Expected Value*).

Отже, *математичне очікування - це параметр* генеральної сукупності, фіксована (але часто невідома) величина. *Середнє вибірки - це статистика*, яка обчислюється з конкретних, експериментальних, безпосередньо виміряних даних і може змінюватися від вибірки до вибірки навіть якщо дані вибираються з однієї і тієї самої генеральної сукупності. ***Важливо запам'ятати – середнє арифметичне вибірки,*** *як і будь яка інша статистика, що відносяться до вибірок* ***- це завжди випадкова величина.*** Так, в прикладі, що наводився раніше про визначення ваги курсантів інституту, в разі зміні значень в генеральній сукупності (наприклад, хтось із курсантів погладшав) математичне очікування теж об'єктивно змінилося. А якщо змінилося якесь значення у вибірці - просто було взято значення, пов'язане з іншим курсантом - то середнє вибіркове теж змінилося, а от математичне очікування – лишилося без змін.

Інший приклад. Уявімо, що ми вимірюємо середній розмір пакета в мережевому трафіку за рік. В цьому випадку математичне очікування - це значення, яке описує середній розмір усіх пакетів, що були передані мережею за рік. Середнє вибірки - це результат, який ми отримали, наприклад, з аналізу

трафіку, що був переданий через мережу за одну годину. Це наближення, і воно буде тим точнішим, чим більше даних ми врахуємо.

І відразу-ж постає чергове питання - а ***наскільки точно отримане в результаті аналізу наданих даних середнє значення відповідає математичному очікуванню генеральної сукупності, з якої ці значення були отримані*?** На це питання ми будемо пробувати знайти відповідь в розділі, що присвячений перевірці статистичних гіпотез (Розділ 3).

Однак і середнє значення не завжди точно відповідає уявленню про місцезнаходження центру даних. Зверніть увагу на гістограму на Рис.2.23, де відображене середнє значення в розподілі заробітної плати.

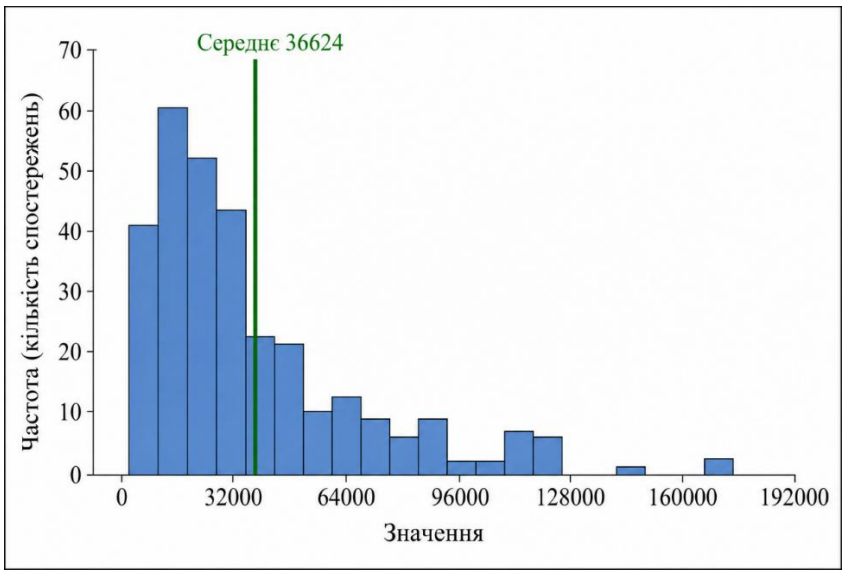


Рис. 2.23 Гістограма розподілу значень доходів представників деякої професійної групи

При наданні в якості опису центральної тенденції середнього значення вибірки у споживача інформації створюється хибне уявлення, що більшість заробляє набагато більше, ніж це є насправді.

Приклад: Нехай ми маємо виконати аналіз рівня продуктивності інтернет-серверу за часом завантаження сторінок. Якщо у більшості користувачів час завантаження становить 1–2 секунди, але у кількох з них через технічні збої сягає 30–40 секунд, середнє значення може становити приблизно 5 секунд, і це веде до висновку, що сайт «повільний». При цьому значення медіани може бути 1.5 секунди - досить непогане значення, що значно реальніше відображає досвід

більшості. Ще один життєвий приклад, з яким ми стикаємося майже щодня, але не замислюємося. У супермаркетах на касах клієнту пропонують кульок перед тим, як почати його обслуговування. Справа в тім, що в випадку, якщо клієнт раптом захоче купити кульок після того, як його основний чек буде закритий, з'явиться новий чек, в якому буде одна позиція – кульок, і цей чек буде мати дуже мале значення суми покупки. А одним з основних показників, за які особовий склад магазину отримує преміальні, є середній розмір чека. Наслідки використання такого показника – зрозумілі.

Наведені приклади показують, що рідкісні значення даних, але такі, що суттєво відрізняються від більшості значень (т.з. **«*викиди*»**, англ. - *Outliers*». Інший термін що застосовується – **«*аномалії*»,** англ. *Anomalies*), тобто такі значення, що формують хвости вибірки, можуть суттєво впливати на значення середнього. На відміну від значень медіани чи моди.

2.3.4.3. Модифікації середнього арифметичного

Хоча виявлення викидів є одним з частих завдань, що вирішуються у різноманітних застосуваннях статистики, машинного навчання тощо, в деяких випадках навпаки, їх наявність для подальшого аналізу неприпустима. Щоб усунути вплив викидів на статистику середнього, використовують певні модифікації наведеного вище визначення.

Так, в статистиці *усіченого середнього* (англ. *Trimmed Mean*) після того, як вхідні дані вибірки впорядковуються, усічене середнє визначається як середнє, після відкидання p найбільших та p найменших значень:

$$\overline{x}_T = \frac{\sum_{i=p+1}^{n-p} x_{(i)}}{n-2p}$$

Наприклад, в деяких видах спорту під час змагань, отримуючи оцінки арбітрів, приймають $p=1$ відкидаючи таким чином найвищу та найнижчу оцінки, і враховують думку тільки тих судів, що залишилися. Це заважає одному

недоброчесному судді маніпулювати результатами, можливо, на користь учасника своєї країни.

Уявимо завдання, в якій необхідно виконати тест на швидкість обробки задачі у кластері з 200 вузлів. Декілька вузлів тимчасово перевантажені або вийшли з ладу. Середнє значення враховує ці «проблемні» вузли, створюючи враження, що кластер працює повільніше, ніж є насправді. Усічене середнє дозволяє оцінити «нормальну» швидкість, виключивши проблемні вузли з розрахунку.

Інший модифікацією середнього є *зважене середнє* (англ. *Weighted Mean*) значення, розраховане шляхом множення кожного значення даних на деяку вагу розподілу і ділення їх суми на суму ваг. Формула для зваженого середнього:

$$\bar{x}_W = \frac{\sum_{i=1}^{n} w_i * x_i}{\sum_{i=1}^{n} w_i}$$

Основні причини використання зваженого середнього полягають в нерівнозначності даних, якими оперують. Наприклад, якщо приймають значення спостережень від декількох датчиків, причому один або декілька датчиків менш точні, можливо знизити вплив («внесок») даних від цього датчика. При розрахунку важливого показника економічного стану в державі - т.з. *Індексу споживчих цін*, враховують факт, що продукти мають не просто різну ціну, але й споживаються в різній кількості і мають різну необхідність для нормального існування людини. При оцінці ризику кіберінцидентів враховуються їх ймовірність та потенційний збиток, причому збиток має суттєво більший вплив на прийняття рішень, ніж сама ймовірність. Зважене середнє дозволяє збалансувати обидва фактори з урахуванням їх відносної важливості.

Ще одною проблемою при роботі з середнім значення вибірки є інтерпретованість результатів - тобто можливість їх пояснити. Щоб краще зрозуміти цю проблему – просто спробуйте самостійно пояснити, а що таке «середня оцінка» в групі або «середня зарплата» на підприємстві? Про анекдотичні приклади «середньої температури по палаті», чи «середньої кількості коліс в транспорті» годі і говорити.

Стає зрозумілим, для більш глибокого аналізу центральної тенденції варто розглядати не тільки середнє значення, але також медіану та моду (моди). *Вибір між середнім, медіаною та модою багато в чому залежить від форми розподілу даних.* Для симетричних розподілів даних, в яких ці три показника збігаються або дуже близькі між собою, можна користуватися середнім. При скошених або мультимодальних розподілах медіана та мода часто дають більш зрозумілу і стабільну картину.

2.3.5. Показники розкиду даних

При виконанні аналізу даних дослідника цікавить не лише те, *де* знаходиться «центр» розподілу (середнє, медіана, мода), але й *наскільки дані розкидані* відносно цього центру. Два набори даних можуть мати однакове середнє значення, але виглядати кардинально по-різному: один - компактно зосереджений навколо середнього, інший - сильно розкиданий (Рис.2.24).

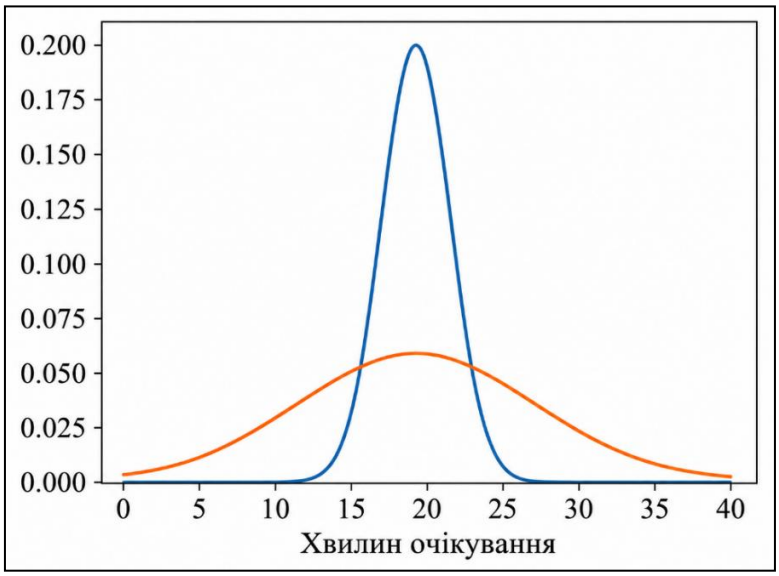


Рис.2.24 Функції щільності розподілу для двох випадкових величин
(на прикладі часу очікування доставки піци)

Розкид даних характеризує ***варіативність***, тобто ступінь відмінності між окремими спостереженнями. Розуміння варіативності даних вкрай важливе для розуміння природи даних та системи, що аналізується. Якщо розподіл значень має низьку варіативність, результати, як правило, є більш передбачуваними та

узгодженими. Навпаки, висока мінливість означає, що дані більш розкидані, а ймовірність появи екстремальних (нестандартних) значень зростає. Розуміння цього дозволяє точніше оцінити ризики, передбачити нестандартні події та вчасно реагувати на відхилення від норми.

Мінливість оточує нас скрізь: Час поїздки на роботу змінюється від дня до дня. Одне й те саме блюдо в ресторані може мати трохи інший смак або вигляд залежно від зміни кухаря. Деталі з конвеєра можуть здаватися ідентичними, але мати незначні відмінності у розмірах. Побутовий приклад: уявімо дві піцерії, які обидві рекламують середній час доставки 20 хвилин. На перший погляд, вони однаково привабливі. Але якщо в однієї піцерії доставка завжди займає від 15 до 25 хвилини, а в іншої - від 1 до 40 хвилин, то в ситуації, коли ви дуже голодні, різниця в мінливості (варіативності часу доставки) стає вирішальною (Рис.2.24). Приклад з ІТ та кібербезпеки: уявіть систему виявлення атак, яка аналізує час середній час реагування на інцидент, тобто проміжок часу від моменту проникнення зловмисника до його блокування. Дві команди кіберзахисту можуть мати однакове середнє значення часу реагування, наприклад, 15 хвилин. Але якщо в однієї команди цей час стабільно коливається від 14 до 16 хвилин, а в іншої - від 2 до 45 хвилин, то перша команда набагато надійніша. У випадку з кіберзагрозами навіть кілька хвилин затримки можуть стати критичними. Отже, *аналіз мінливості даних - ключовий інструмент не лише для статистики, а й для прийняття рішень у критичних сферах, де ціна помилки або затримки є високою.*

У математичній статистиці для аналізу варіативності використовують т.з. показники розсіювання (англ. *Measures of Dispersion*»). До основних показників розсіювання належать:

- Розмах
- Міжквартильний розмах
- Середня різниця Джині
- Дисперсія та стандартне відхилення

Розмах - найпростіший показник, що відображає широта розподілу, тобто різницю між найбільшим і найменшим значенням:

$$R = x_{\max} - x_{\min} = Q_{100\%} - Q_{0\%}$$

Варіаційний розмах є величиною нестійкою, при роботі з вибірками - надзвичайно залежною від випадкових обставин. До того-ж він ніяк не враховує, як саме розподілені дані між мінімальним і максимальним елементами. Тому цей показник, як правило, застосовується тільки для приблизної, початкової оцінки варіативності даних.

Міжквартильний розмах (*IQR*) - більш стійкий до викидів показник, що базується на квартилях, був розглянутий в розділі 2.3.2. *IQR* використовується, наприклад, при початковому аналізі даних на наявність викидів. В контексті задач кібербезпеки - як приклад - при виявленні аномальних піків трафіку, що сигналізують про DDoS-атаку, або при визначенні стабільності затримки пакетів у мережі.

Середня різниця Джині (англ. *Gini Mean Difference*, *GMD*) – міра розкиду даних, заснована на аналізі попарних різницях між усіма спостереженнями в вибірці. Оцінює варіативність без припущень про закон розподілу (на відміну від дисперсії, яка є чутливою до викидів). *GMD* визначається як середнє значення абсолютних різниць між усіма можливими парами спостережень у наборі даних:

$$GMD = \frac{1}{n(n-1)} \sum_{\substack{i,j \\ i \neq j}}^{n} \left| x_i - x_j \right|$$

Інтерпретація значення *GMD* не викликає суттєвих труднощів. Якщо всі значення однакові (тобто варіативність даних відсутня), то *GMD = 0*. Якщо всі дані набору помножити на константу, *GMD* масштабується так само. Чим більше різняться значення в наборі (чим більший розкид значень даних), тим більшим буде *GMD*. Так, при оцінюванні швидкості обробки запитів на сервері, *GMD* покаже, наскільки рівномірно система обробляє трафік. Якщо сервер іноді відповідає за 0.1 с, а іноді - за 2 секунди, *GMD* буде високим, сигналізуючи про проблеми з продуктивністю. В порівнянні з дисперсією (див. далі), цей показник менш чутливий до екстремальних значень (але все ж їх враховує), тому його іноді обирають для аналізу фінансових показників, часу обробки в системах, кіберінцидентів тощо. Використання *GMD* не потребує припущення про нормальність розподілу даних. Також *GMD* використовується тоді, коли важлива інтерпретація «середньої різниці» в натуральних одиницях, наприклад, у соціально-економічних дослідженнях для оцінки нерівності доходів. Тут *GMD*

легко пояснити: це середня різниця доходів між будь-якими двома людьми в досліджуваній групі. В машинному навчанні середня різниця Джині використовуються, зокрема, для оцінки розкиду категоріальних чи нечислових даних (де дисперсія виявляється непридатною- див. Розділ 2.1.4), для обчислення «відстаней» між розподілами, у метриках схожості, тощо. Однак через високу обчислювальну складність ($O(n^2)$ для прямого розрахунку), *GMD* частіше застосовують на невеликих вибірках або у випадках, коли потрібна саме така інтерпретація.

Ще одна причина, що обмежує поширення *GMD* в практиці наукових досліджень полягає в складності аналітичного дослідження цієї функції, а відтак – суттєве ускладнення для обґрунтування результатів, що отримані за його допомогою. Зацікавлені можуть ознайомитися з деякими останніми роботами в цьому напрямку за посиланням [29], [30].

2.3.6. Статистики розкиду, що аналізують відхилення від центральної тенденції.

Якщо піти шляхом вимірювання відхилень значень набору не один від одного, а від середнього значення (математичного очікування) та враховувати не кожне таке відхилення окремо, а усереднити їх, то можна ввести наступні оцінки: для генеральної сукупності т.з. «***середнє абсолютне відхилення***» (англ. *Average Absolute Deviation* або *Mean Absolute Aeviation* (*MAD*)):

$$\delta = \frac{\sum_{i-1}^{N} |X - \mu|}{N}$$

а також її вибірковий аналог (оцінку):

$$MAD = \frac{\sum_{i-1}^{n} |x_i - \bar{x}|}{n}$$

Ці статистики характеризують *розсіяння значень навколо центру розподілу або середнього значення вибірки*: більше значення показників відповідає більшому відхиленню значень даних відносно центру. Використання під сумою модулів(абсолютних значень) відхилень зумовлено тим, що при усередненні лінійних відхилень від середнього арифметичного чи математичного очікування

отримана сума буде дорівнювати «0» (за визначенням самого середнього арифметичного). Перехід від лінійних відхилень до їх абсолютних значень дозволяє уникнути такого «обнуління» міри.

В якості центру розподілу даних замість вибіркового середнього можуть бути взяті інші статистики вибірки - медіана чи мода. Слід зауважити, що при усередненні лінійних відхилень від медіани або моди сума відхилень (з урахуванням знаків) може й не дорівнювати нулю. Наприклад, в прикладі набору оцінок курсантів, що розглядався в розділі 2.3.4.1, середнє відхилення відносно середнього значення вибірки становить 9.4349, відносно медіани - 9.2632, відносно моди - 13.2631. В загальному випадку середнє абсолютне відхилення від медіани менше або дорівнює середньому абсолютному відхиленню від середнього арифметичного та менше або дорівнює середньому абсолютному відхиленню від будь-якого іншого фіксованого значення.

Для зацікавлених. За посиланням [31] можна знайти дослідження використання статистики ***медіанного абсолютного відхилення*** (англ. *Median Absolute Deviation, (MdAD)*):

$$MdAD = median\left(x_i - median\left(x\right)\right)$$

В роботі [32] надані приклади більш глибокого аналізу та порівняння методів визначення розкиду даних відносно різних параметрів центральної тенденції вибірки.

Середнє абсолютне відхилення як міра розсіяння в порівнянні зі стандартним відхиленням та дисперсією (див. нижче) стійкіша до викидів, простіша для інтуїтивного розуміння. По своїй суті воно дає інформацію, наскільки далеко від центру розподілу в середньому знаходяться значення випадкової величини. В порівнянні з розглянутими вище показниками розмаху - має перевагу в тому, що розраховується за всіма значеннями, тому репрезентує більше інформації про характер розподілу значень. Позатим, такі показники варіативності використовуються досить рідко. Це пояснюється тим, що хоча середнє абсолютне відхилення має природне інтуїтивне визначення як «середнє відхилення від середнього значення», та вимагає менших затрат обчислювальних ресурсів в порівнянні з дисперсією або стандартним відхиленням, його використання робить *аналітичні дослідження* з цією статистикою набагато

складнішими. Разом з тим, середнє абсолютне відхилення не дуже зручне і для математичних перетворень в комп'ютерних реалізаціях алгоритмів. Зокрема, з модулями досить складно працювати в операціях чисельного інтегрування та диференціювання.

Крім середнього абсолютного відхилення як міра розсіяння використовується також ***відносне середнє абсолютне відхилення*** (англ. *Relative Mean Absolute Deviation,* (*RMAD*)) - це показник варіативності, який дозволяє оцінити середнє відхилення значень від їхнього середнього *у відсотковому вираженні*. На відміну від *MAD*, що показує середню величину відхилення в тих самих одиницях, що й самі дані, *RMAD нормалізує цю величину* шляхом ділення на середнє значення вибірки:

$$RMAD = \frac{1}{n}\sum_{i-1}^{n}\frac{\left|x_i - \bar{x}\right|}{\bar{x}} = \frac{MAD}{\bar{x}} * 100\%$$

Цей показник зручний для порівняння варіативності наборів даних, що мають різні одиниці вимірювання або сильно відрізняються за масштабом. Вибіркове відносне *середнє абсолютне відхилення*, як і будь-яка інша статистична оцінка, *є величиною випадковою*.

2.3.7. Дисперсія та середнє квадратичне відхилення

Найбільш широковживаними показниками при роботі з кількісними даними є дисперсія та стандартне відхилення.

Дисперсія (позначається σ^2) - *це середній квадрат відхилень усіх значень генеральної сукупності від її математичного очікування.* Дисперсія є показником, що відображає наскільки «розтягнутими» або «сконцентрованими» є дані:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \mu\right)^2$$

В реальних умовах рідко мається доступ до всіх можливих значення сукупності даних. Так само, як і у випадку математичного очікування, значення дисперсії генеральної сукупності найчастіше нам невідоме, або відоме лише з теоретичних міркувань - тобто *дисперсія є параметром функції розподілу*

випадкової величини. Оцінкою дисперсії є вибіркова статистика, що називається ***«вибірковою дисперсією»***.

Однак при роботі з вказаними характеристиками виникає проблема: якщо просто підставити середнє вибірки та поділити на розмір вибірки, ми отримаємо систематично занижену оцінку дисперсії. Строго математичне пояснення цьому явищу надавалося в курсі теорії ймовірностей [22], де доводилося, що для того, щоб значення вибіркової дисперсії наближалося до значення дисперсії генеральної сукупності при $n \to \infty$, тобто для того, щоб виправити згадане зміщення, у формулі для вибіркової дисперсії використовується т.з «поправка Бесселя», яка полягає у використанні у визначенні дільника *n-1,* а не *n:*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$$

Особливо важливо це мати на увазі при роботі з вибірками малого об'єму. Описане зміщення зменшується із збільшенням об'єму вибірки, Причому зменшується у порядку, пропорційному до 1/n. Таким чином зміщення є найбільш відчутним для малих і середніх вибірок; для n<75 зміщення стає досить відчутним, перевищуючи 1 %. Чи можна ним знехтувати – залежить від необхідної точності результатів дослідження та визначається дослідником. Детальніше про вказану поправку можна дізнатися, відвідавши веб-сторінки [33], [34].

Принагідно зауважимо, що позбавлення від негативних значень - вагома причина для використання квадратичної функції, але не єдина. Зверніть увагу на те, що в процесі сумування *квадратів* різниць елементи, що лежать далі від середнього значення, набувають більшої ваги, ніж елементи, що лежать ближче до середнього. Говорять, що показник дисперсії квадратична збільшує «вагу» відхилення конкретного значення, з його (відхилення) зростанням.

Дисперсія - дуже важливий показник варіативності даних. Вона показує, наскільки значення розкидані відносно середнього. Але є ще одна проблема, пов'язана з використанням цієї характеристики: значення дисперсії *має розмірність квадрату одиниць вимірювання значень* даних. Якщо вимірюється час у тижнях, то дисперсія буде виміряна у «тижнях в квадраті». Якщо аналізується розмір файлів у мегабайтах, то дисперсія вимірюється у «мегабайтах

у квадраті», в інших задачах можуть виникати зовсім химерні одиниці - квадратний долар, квадратний відсоток тощо. Такі результати може бути м'яко кажучи досить складно інтерпретувати. Саме тому вводиться статистика, що називається ***стандартне відхилення*** (англ. *Standard Deviation, (STD)*, інша назва «***середнє квадратичне відхилення***») - додатній квадратний корінь із дисперсії:

$$\sigma = +\sqrt{\sigma^2};\ \ s = +\sqrt{s^2}.$$

Ці показники (відповідно - для генеральної сукупності та вибірки) *вимірюються у тих самих одиницях, що й самі дані*. Це робить інтерпретацію отриманих результатів значно зрозумілішою. Уявімо, що вимірюється час відгуку сервера (у мілісекундах) під час атаки "відмова в обслуговуванні" (DDoS). При цьому виявилося, що середній час відповіді сервера дорівнює 200 мс, а дисперсія - 2500мс^2. Такий показник важко інтерпретувати і суперечить людській інтуїції. А от стандартне відхилення при цьому набуває значення 50 мс., що означає, що «*більшість* значень часу відповіді сервера відхиляються від середнього на ±50 мс.»

Цікавий факт. Стандартне відхилення існує не для будь-якої випадкової величини. Як видно, умовою існування стандартного відхилення випадкової величини є існування її математичного очікування. Але існують такі розподіли наприклад т.з. розподіл Коші, для яких не існує математичне очікування [22], а отже не існує і стандартне відхилення. Це важко собі уявити, особливо враховуючи те, що візуально графік функції щільності для розподілу Коші, дуже сильно схожий на звичний графік функції нормального розподілу. Позатим це математично доведений факт.

Як зазначалося вище, останні роки спостерігається суттєве зростання досліджень, в яких для аналізу розкиду даних використовується *MAD* та *MdAD,* замість *STD.* Це пояснюється багатьма факторами, зокрема тим, що *MAD* та *MdAD* менш чутливі до викидів і краще відображають «типову» варіацію. Сучасна статистика та машинне навчання тяжіють до використання т.з. «*робастних*» методів аналізу, тобто таких, що не «ламаються» на викидах. Це особливо важливо у випадках із «шумними» даними (наприклад, при аналізі мережевого трафіку, часу відгуку серверів тощо), аналізі часових рядів – тобто

даних, що явним чином залежать від часу їх отримання (фінанси, кібербезпека, IoT).

Деякі статистичні явища і процеси характеризуються «нескінченною дисперсією», хоча мають кінцеве середнє абсолютне відхилення. І навпаки, всіх випадках, де існує *STD*, існує і *MAD*. Крім того, *MAD* часто простіший для обчислення та пояснення. Уявіть, що потрібно оцінити «середнє відхилення за день» для температури у місті (біржової вартості компанії, артеріального тиску людини тощо) протягом п'яти останніх днів. Дані *відхилення* в ці п'ять днів такі: (-23, 7, -3, 20, -1). Як оцінити середнє відхилення? Пропонуємо перед тим, як читати далі - самим спробувати дати відповідь на таке питання.

STD дасть оцінку, що дорівнює 15,7, *MAD* -10.8. Насправді друге значення більш пасує до інтуїтивного розуміння поняття «середнє відхилення» звичайною людиною без математичної чи навіть технічної освіти. Подекуди трапляється, що замовнику дослідження повідомляють середнє квадратичне відхилення, а він потім діє (приймає рішення) так, ніби прийняв цю величину за середнє відхилення. *Системний аналітик має мати на увазі таку особливість мислення замовника* та приймати заходи для уникнення вказаних непорозумінь.

З іншого боку, розглянемо Рис 2.25.

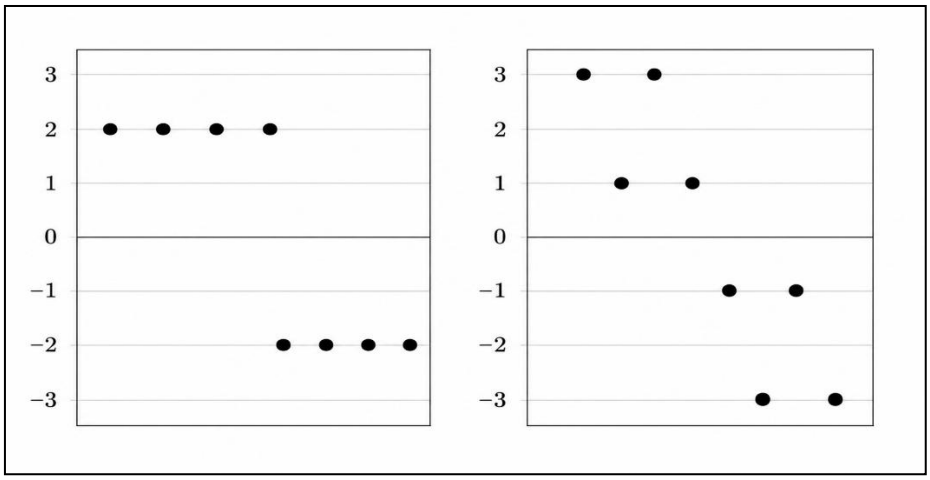


Рис 2.25 Приклад вибірок з однаковим значенням *MAD* та різними значеннями *STD*.

Очевидно, що у другому наборі даних більше варіативності, ніж у першому, але *MAD* цієї різниці не відобразить. Воно дорівнює 2 для обох вибірок. Однак *STD* для другої змінної буде вищою - 2,24, що краще відповідає нашим

інтуїтивним уявленням. Отже, нема і бути не може бути відповіді на питання– який з показників кращий (точніший). Відповідь – як досить часто буває при застосуванні статистичного аналізу та методів машинного навчання - залежить від більш глибокого, прикладного аналізу системи, що досліджується.

Для тих, хто хоче більш глибоко розібратися з сучасними методами аналізу розкидів та застосування цих методів в машинному навчанні, рекомендую ознайомитися з наступним джерелами: [35], [36] (До речі, автор останнього джерела відомий ще й як дослідник, що вперше запропонував концепцію «чорного лебедя» при вивченні та аналізі об'єктів, систем, процесів та явищ реального світу).

Нарешті, варто зупинитися на ще одному випадку, який досить часто зустрічається при аналізі даних. Уявимо, що дані надані вже в дещо обробленому вигляді, а саме у вигляді *інтервального ряду* (певною мірою - гістограми). Так може статися, наприклад, в результаті неможливості провести вимір даних з деякою фіксованою точністю; через попереднє автоматичне групування даних; якщо дані отримані в результаті опитування тощо. Отже - на вході існує таблиця кількості потраплянь даних в кожен з виділених діапазонів або частотний розподіл даних по інтервалах, але точних значень кожного спостереження не існує. У такому випадку для обчислення характеристик варіації можливо користуватися середніми значеннями інтервалів та відповідними частотами. Дисперсія інтервального ряду визначається наступним чином:

$$D = \frac{\sum_{j=1}^{K} f_j \left(\overline{x}_j - \overline{x} \right)^2}{\sum_{j=1}^{K} f_j}$$

де K - кількість інтервалів;

f_j -частота (кількість спостережень) в j-ому інтервалі;

$\overline{x}_j$ - середнє значення в j-ого інтервалу;

$\overline{x}$ - середнє значення інтервального ряду, що в свою чергу визначається за формулою:

$$\overline{x} = \frac{\sum_{j=1}^{K} f_j \overline{x}_j}{\sum_{j=1}^{K} f_j}$$

Відповідно, *середнє квадратичне відхилення для інтервального ряду* визначається як:

$$d = +\sqrt{D}$$

Дисперсія і середнє квадратичне відхилення – найбільш широко вживані показники варіації. Пояснюється це тим, що вони входять в більшість теорем теорії ймовірності, які служить фундаментом математичної статистики. Крім того, дисперсія може бути розкладена на складові елементи, що дозволяють оцінити вплив різних чинників, які обумовлюють варіацію ознаки. Ці показники лежать в основі багатьох методів аналізу даних, моделювання та машинного навчання. Вони допомагають зрозуміти стабільність процесів, виявляти відхилення й приймати рішення на основі реальних даних.

2.3.8. Показники форми розподілу

Для здобуття приблизного уявлення про форму розподілу будують графіки розподілу , наприклад - у вигляді гістограми. Або їх апроксимації. На практиці дослідникові доводиться зустрічатися з дуже різноманітними розподілами. Вище зазначалося, що однорідні сукупності характеризуються, як правило, одновершинними розподілами. Мультимодальність свідчить про неоднорідність сукупності, що вивчається.

З'ясування загального характеру розподілу передбачає оцінку міри його варіативності, які розглядалися вище, а також визначення показників асиметрії і ексцесу. Ці показники дають уявлення про *симетрію* та *гостровершинність/пологість* кривої функції щільності розподілу. По суті *вони дозволяють кількісно оцінити форму розподілу даних в порівнянні з нормальним розподілом*.

Асиметрією називають характеристику розподілу даних, яка описує відхилення функції розподілу від симетричності. *У симетричному розподілі середнє значення та квантиль $Q_{50\%}$ (медіана) дорівнюють одне одному*. Якщо

розподіл одночасно і симетричний і унімодальний (наприклад, нормальний), то середнє значення дорівнює медіані та дорівнює моді. *Це може слугувати найпростішим показником симетрії функції розподілу, навіть не малюючи відповідний графік.* Але в житті такого ідеального збігу, звичайно, не буває (навіть тіло людини трохи асиметричне). Тому на практиці у «майже симетричних» розподілах вказані показники мають бути дуже близькими один до одного. Якщо ж дані «зсунуті» вліво чи вправо, розподіл набуває асиметричності. Чим помітніше асиметрія, тим більше відхилення між вказаними характеристиками центру розподілу.

Розрізняють позитивну та негативну асиметрії розподілу. Позитивна асиметрія (правостороння, англ. *Positive Skew*; для усвідомлення - «довгий хвіст праворуч») виникає в разі, якщо більшість значень сконцентровані в області менших (нижчих) значень, але в наборі існують поодинокі великі спостереження. Негативна асиметрія (лівостороння, англ. *Negative Skew*; «довгий хвіст ліворуч») - більшість значень великі, але трапляються поодинокі малі значення. На практиці, при *попередньому ознайомленні* з даними, знак асиметрії визначають за положенням кривої відносно медіани: якщо «довша» частина кривої знаходиться правіше медіани, то асиметрія негативна, якщо лівіше - позитивна (Рис 2.26).

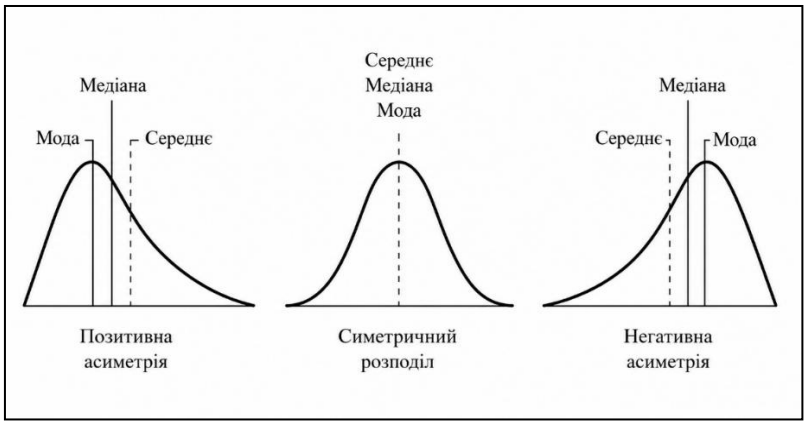


Рис.2.26 Різновиди асиметрії

На рис 2.27 наведено відображення гістограми даних про кількість дорослих людей, що проживають в одному помешканні в США.

Для практичного застосування потрібна характеристика, яка не тільки визначить напрям зсуву, але і дасть його відповідну кількісну характеристику. Таких характеристик запропоновано декілька.

Коефіцієнтом асиметрії Фішера-Пірсона теоретичного розподілу ймовірностей випадкової величини називають відношення центрального моменту третього порядку до кубу стандартного відхилення:

$$\beta = \frac{\mu_3}{\sigma^3}$$

де μ_3 - т.з. третій центральний момент розподілу (див. теорію ймовірностей[22], [23]). При роботі з вибіркою це визначення трансформується в наступну статистику:

$$B = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right]^{\frac{3}{2}}}$$

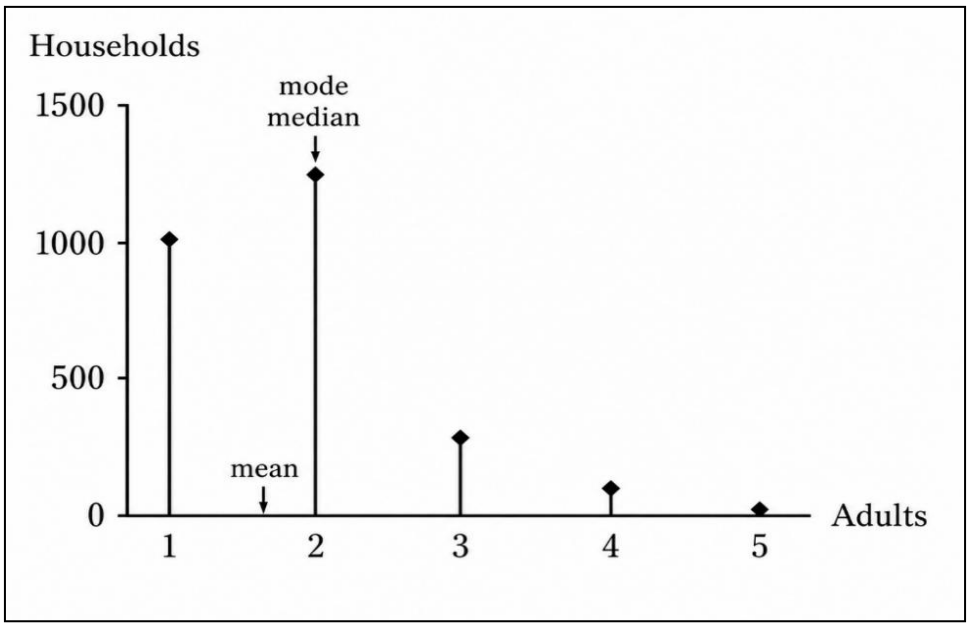


Рис. 2.27 Кількість дорослих, що проживають в одному помешканні (США)

Якщо значення коефіцієнту $B > 0$ то розподіл скошений вправо, якщо $B < 0$ – то вліво. Коефіцієнт B для нормального розподілу має значення 0. На практиці

чим ближче значення B наближається до 0 тим більше розподіл вибірки наближається до нормального закону розподілу (див. Розділ 2.4.2.1) з параметрами $\mu = \bar{x}$ та $\sigma = s$. Водночас, для експоненціального розподілу (див. Розділ 2.4.3.3) значення коефіцієнта B дорівнює 2, а логнормальний розподіл (див. Розділ 2.4.3.1) може мати асиметрію будь-якого додатного значення, залежно від його параметрів.

Потрібно мати на увазі, що в деяких програмах статистичного аналізу, зокрема в MS Excel, Minitab, SAS, SPSS тощо, використовується дещо інший варіант визначення коефіцієнту асиметрії, а саме:

$$G = \frac{n^2}{(n-1)(n-2)} B$$

На використанні саме цього визначення асиметрії заснований тест д'Агустініо (D'Agostino) – один з найбільш потужних тестів на визначення того, чи підпорядковується надана вибірка даних нормальному закону розподілу.

Ще один показник, що характеризує форму функції розподілу - ***коефіцієнт ексцесу*** (англ. *Kurtosis*). Він вимірює «гостровершинність» або «пласкість» розподілу, «крутість», тобто стрімкість підвищення кривої розподілу, та важкість його хвостів *у порівнянні з кривою нормального закону розподілу*. На рис.2.28 представлені розподіли, які мають однакове математичне очікування ($\mu = 0$), однакову дисперсію ($\sigma = 1$), симетричні, але явно відрізняються за вказаними ознаками**.**

Коефіцієнт ексцеса розподілу розраховується за формулою:

$$\kappa = \frac{\mu_4}{\sigma^4} - 3$$

де μ_4 – центральний момент четвертого порядку.

При аналізі експериментальної вибірки використовується визначення:

$$K = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^4}{\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right]^2} - 3$$

Ексцес може бути позитивним і негативним. У нормальному розподілі відношення $\frac{\mu_4}{\sigma^4} = 0$ (власне, саме через це в вищенаведену формулу і було введено поправку -3), тому для такого розподілу $K=0$. Якщо ексцес деякого розподілу відмінний від нуля, то крива щільності цього розподілу відрізняється від кривої щільності нормального розподілу: якщо ексцес додатній, то крива розподілу, що досліджується, має вищу та «гострішу» вершину ніж крива нормального; якщо ексцес від'ємний, то крива розподілу має нижчу та «плоскішу» вершину ніж крива нормального.

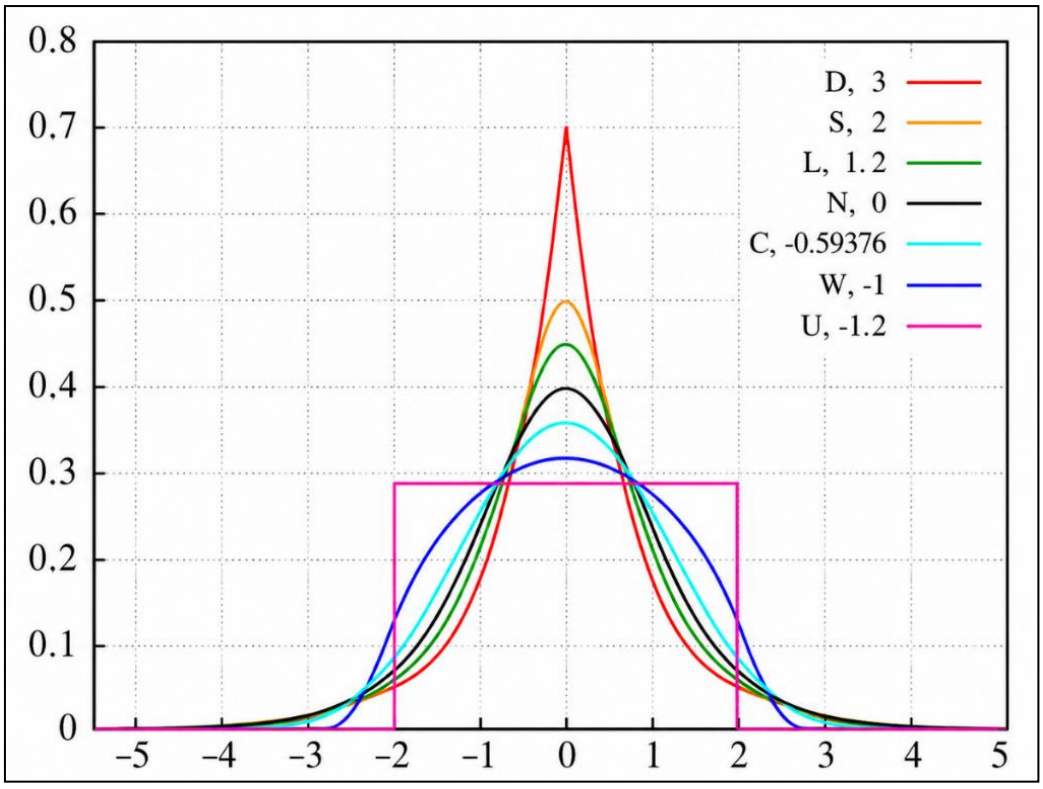
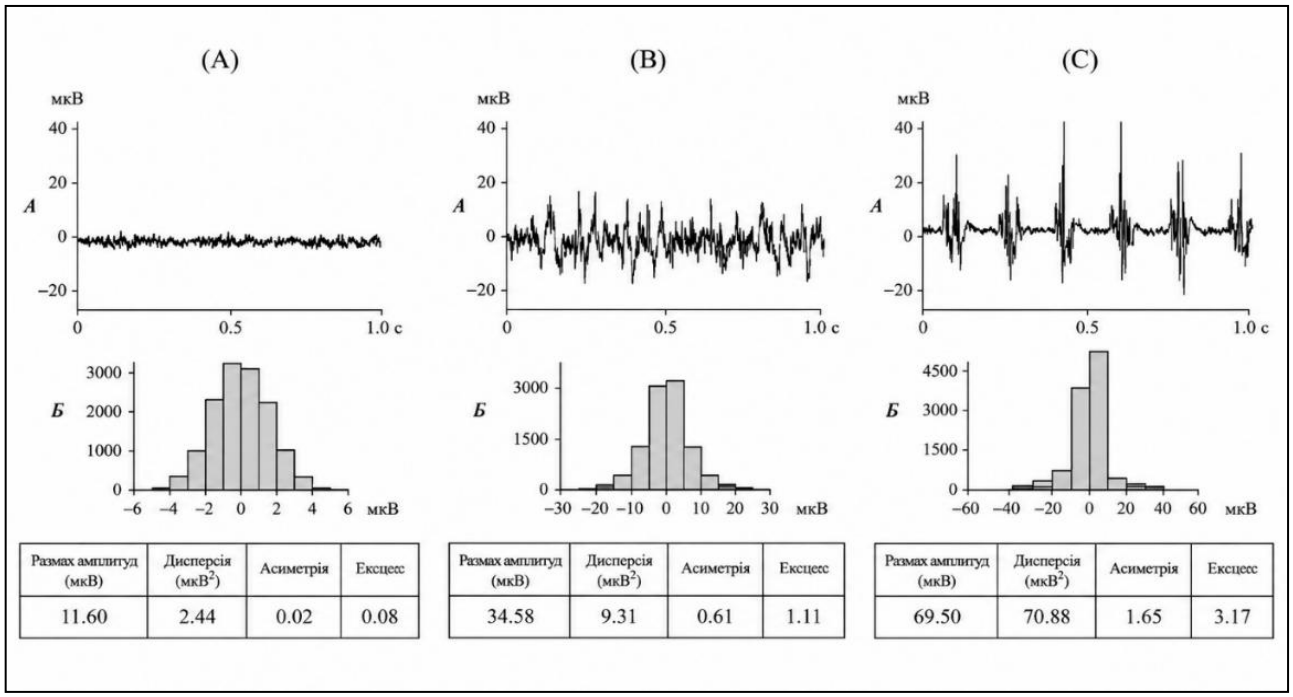


Рис.2.28 Приклади розподілів з однаковим математичним очікуванням та різними значеннями коефіцієнту ексцесу

Маючі в розпорядженні лише розглінуті в цьому і попередніх розділах статистики параметрів об'єктів досліждення, фахівці мають змогу робити достатньо серйозні висновки, діагностувати та класифікувати їх. Важливо ще раз наголосити, що ***різноманіття статистик дає можливість обрати саме ті з них, які відповідають природі даних, що досліджуються, в тому числі врахувати шакли вимірів, яки при цьому використовуються, та особливості задач аналізу, які постають перед дослідником***. Наприклад, при аналізу мережевого трафіку медіана, мода та середє дозволяють проаналізувати типовий час реакції систем, дисперсія та асиметрія часто вказують на наявність аномалій (зокрема – на затримки в трафіку), ексцес підказує, що більшість значень стабільні, але є ризик рідкісних великих збоїв.

Приклад наведений на рис.2.29 демонструє можливість використання розглянутих описових статистик даних для класифікації об'єктів (в даному випадку - пацієнтів), та вирішити відповідну діагностичну задачу практичної психіатрії.



Размах амплитуд (мкВ)	Дисперсія (мкВ^2)	Асиметрія	Ексцесс
11.60	2.44	0.02	0.08

Размах амплитуд (мкВ)	Дисперсія (мкВ^2)	Асиметрія	Ексцесс
34.58	9.31	0.61	1.11

Размах амплитуд (мкВ)	Дисперсія (мкВ^2)	Асиметрія	Ексцесс
69.50	70.88	1.65	3.17

Рис.2.29 Приклад застосування описової статистики вибірок до вирішення задач практичної психіатрії. Джерело (http://www.mif-ua.com/archive/issue-22763/article-22791/)

Контрольні запитання

1. У чому різниця між середньою та показниками варіації?
2. У чому полягає значення показників варіації?
3. Які показники (заходи) варіації застосовуються найчастіше?
4. Назвіть основні категорії статистики?
5. Дайте визначення поняттям «розмах варіації», «середнє лінійне відхилення», «середнє квадратичне відхилення» « коефіцієнт варіації», «дисперсія».
6. Для чого розраховується коефіцієнт асиметрії ?
7. Для чого розраховується ексцес розподілу?
8. Як впливає розмір вибірки на значення середнього квадратичного відхилення?

2.4 Короткий вступ до законів розподілу ймовірностей

При аналізу даних постає питання: як можна описати - а тим більше спробувати передбачити - поведінку систем, якщо дані мають випадкову, ймовірнісну природу? Тому виконуючи практично будь який дослідницький аналіз даних, перше, що потрібно зрозуміти - як саме *розподілені значення в наборі даних*, що наданий для дослідження, *яким закономірностям ці дані підпорядковуються*. Такі закономірності називають ***законами розподілу ймовірностей***.

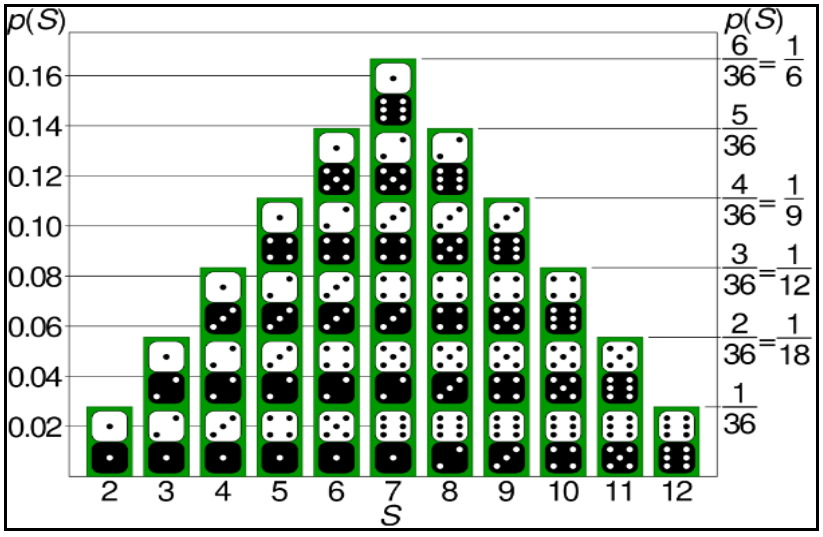
Треба усвідомлювати, що будь які дані – це «об'єктивна реальність», а закон розподілу– це завжди деяка *математична модель*, а отже - завжди деяка абстракція. Розглядаючи гістограму набору даних (див. розділ 2.2.3) можливо приблизно охарактеризувати форму функції (чи радше - графіку) розподілу його значень. В часи, коли не було комп'ютерів, це був абсолютно натуральний і по суті єдиний можливий шлях наукового дослідження - спробувати знайти, підібрати математичну функцію, що найкраще пасує до наданих даних, а потім вивчати властивості саме цієї функції, сподіваючись, що дослідивши їх ми можемо зробити висновки і про самі наші дані.

Коли постає питання аналізу закону розподілу ймовірнісних величин, необхідно перш за все визначити, до якого типу - дискретних чи неперервних випадкових величин - належать дані, що надані для дослідження. У випадку *дискретної випадкової величини* для опису достатньо задати функцію маси

ймовірності (англ. *Probability Mass Function, (PMF)*), яка ставить у відповідність *кожному можливому значенню цієї величини ймовірність його появи у вибірці*. Класичним прикладом є підкидання двох правильних шестигранних гральних кубиків, в яких кожне з можливих значень від 1 до 6 має однакову ймовірність появи. Це дає наступну картину (див. Рис. 2.30):

На малюнку по горизонтальній осі відображені всі можливі значення (суми граней кубика, що випали), які можуть бути отримані при киданні двох кубиків, по вертикалі – ймовірність того, що саме така сума буде отримана в результаті чергового кидання кубиків. Всього може випасти 36 різних комбінацій.

Зазначимо, що часто замість терміну «функція маси ймовірності» в літературі вживають більш звичний термін «функція щільності розподілу ймовірності». Однак слід зауважити, що більш коректно в контексті аналізу даних з дискретною кількістю можливих значень використовувати PMF, а в випадку неперервних даних - PDF (див. розділ 2.2.2). Обидві функції описують розподіл ймовірностей випадкової величини. В обох випадках «повна ймовірність» має дорівнювати 1 (через суму для PMF або інтеграл для PDF). Значення обох функцій не можуть бути меншими за нуль. Але при моделюванні систем ми використовуємо ці функції для різних задач. Наприклад, PMF допомагає моделювати кількісні показники: скільки вузлів мережі буде уражено, скільки пакетів даних втрачено. PDF критично важлива, зокрема, при роботі з часовими характеристиками: як швидко спрацює система виявлення вторгнень або який середній час відновлення системи після атаки тощо.



2.30 Приклад візуалізації функції маси ймовірності

2.4.1. Розподіли дискретних ймовірнісних величин. Біноміальний розподіл та розподіл Пуассона

Найпоширеніші приклади дискретних розподілів - біноміальний розподіл (кількість успіхів у серії експериментів "так/ні") та розподіл Пуассона (кількість подій за певний проміжок часу або в певній області).

Біноміальний розподіл є дискретним ймовірнісним розподілом, що характеризує кількість успіхів в послідовності експериментів, значення яких змінюється за принципом «так/ні», кожен з яких набуває успіху з ймовірністю p. У біноміальному розподілі випадкова величина визначається як кількість успіхів у n незалежно повторених спробах. Нехай ймовірність успіху дорівнює p, тоді формула для біноміального розподілу ймовірностей має вигляд:

$$f\left(k,n,p\right)=P(x=k)=C_n^k p^k \left(1-p\right)^{n-k}$$

$$C_n^k=\frac{n!}{\left(n-k\right)!k!}$$

Деякі приклади цієї функції при різних значеннях параметрів показані на Рис.2.31.

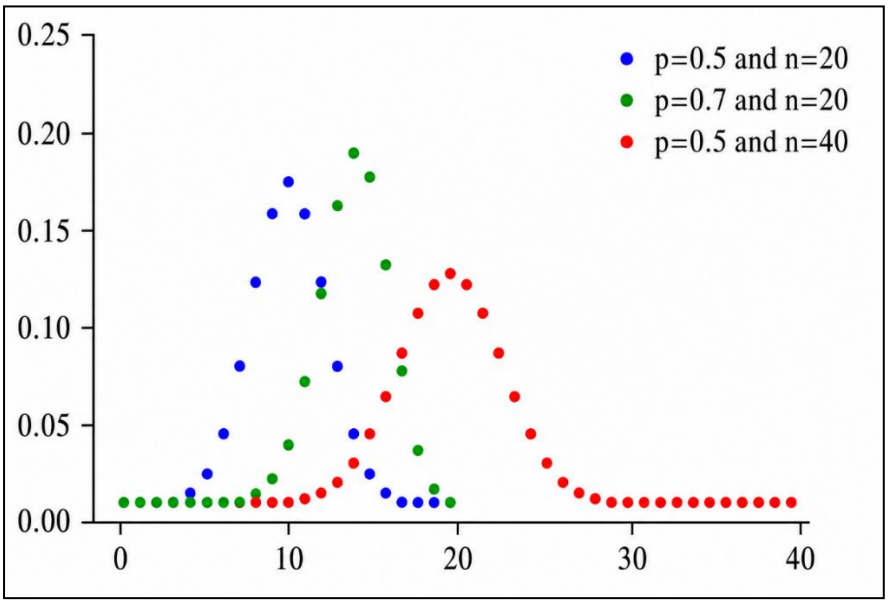


Рис.2.31 Приклади функцій маси ймовірності розподілу за біноміальним законом в залежності від значень параметрів

Відповідно, кумулятивна функція розподілу задається таким чином:

$$F(k,n,p) = P\left(x \le k\right) = \sum_{i=0}^{k} C_n^i p^i \left(1-p\right)^{n-i}$$

Якщо дані розподілені за біноміальним законом розподілу, то можна показати, що математичне очікування буде визначатися як $\mu = np$, дисперсія - як $\sigma^2 = np\left(1-p\right)$. Відповідно, отримавши оцінки для цих показників і знаючи значення n можна визначити значення параметру розподілу p.

Розподіл Пуассона часто називають «розподілом рідкісних подій». Нехай є подія, яка відбувається з фіксованою середньою повторюваністю (інтенсивністю) у часі μ. Наприклад, в середньому 5 людей входять в деяке приміщення щохвилини; або щоденно до відділу технічної підтримки надходить в середньому дві заявки; або на станції швидкої допомоги щогодини фіксується 7 викликів тощо. Ймовірність спостереження k подій за одиницю часу можна розрахувати за допомогою розподілу Пуассона за формулою:

$$f\left(k,\mu\right) = P(x=k) = \frac{\mu^k e^{-\mu}}{k!}$$

Цікавою особливістю даних, розподілених за законом Пуассона є рівність значень математичного очікування та дисперсії. Приклади функції розподілу цього закону наведені на Рис. 2.32.

Закон розподілу Пуассона - це розподіл ймовірностей масових подій. Його використовують у задачах статистичного контролю якості, теорії надійності, теорії масового обслуговування. Багато реальних явищ, таких як автомобільні аварії, дорожній рух, генетичні мутації та кількість помилок друку на сторінці, відповідають розподілу Пуассона. Наприклад, його використовують при аналізі необхідної кількості машин швидкої допомоги у місті, кількості обслуговуючого персоналу при роботі з потоком клієнтів. Деякі параметри мережевого трафіку, такі як кількість пакетів, що надходять до мережевого інтерфейсу за одиницю часу, кількість нових з'єднань/користувацьких підключень у фіксованому часовому інтервалі, кількість помилкових пакетів або перерваних передач у

певному інтервалі часу тощо також можна розглядати як випадкові величини, що розподілені за законом Пуассона. Головне обмеження при застосування даного формалізму - всі події виникають незалежно одна від одної і інтенсивність (тобто – *середня* кількість подій за фіксований проміжок часу) лишається незмінною. Для подальшого ознайомлення з цим розподілом та його властивостями раджу звернутися за посиланням [37].

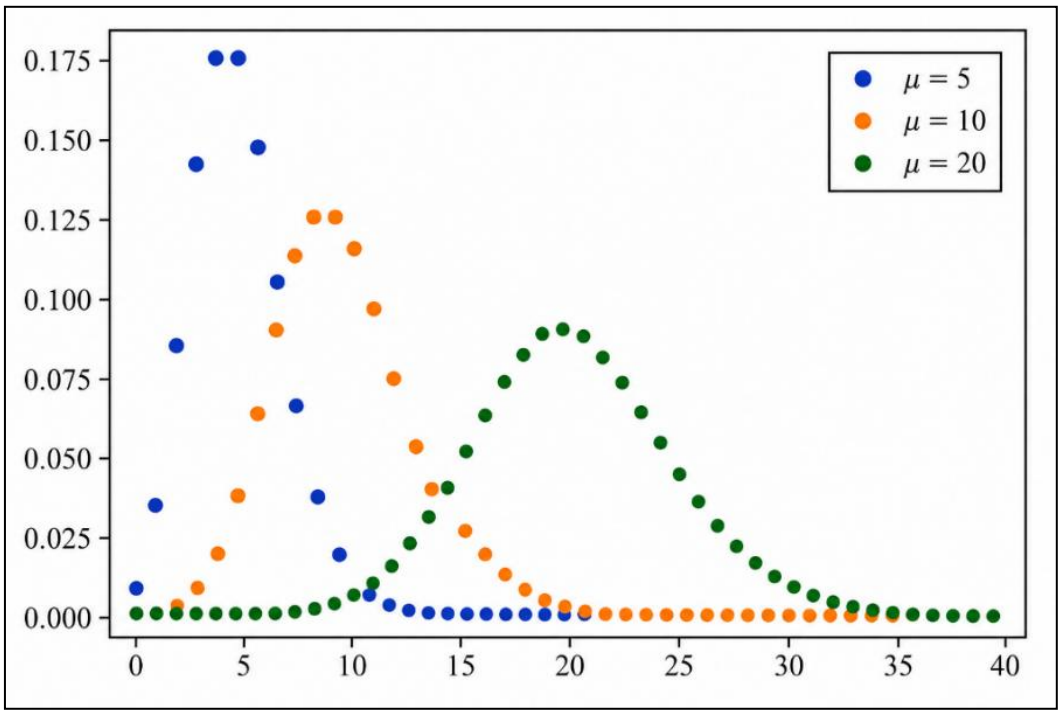


Рис.2.32 Приклади функцій маси ймовірності розподілу за законом Пуассона в залежності від значень параметрів

Окрім вказаних, при аналізі даних з дискретними значеннями використовуються і інші закони розподілу.

2.4.2. Розподіли безперервних ймовірнісних величин

При роботі з даними, які можуть приймати нескінчену кількість допустимих значень (тобто - безперервних випадкових величин), безсумнівно найвідомішим, найбільш часто (або коректніше сказати - традиційно) застосованим розподілом є так званий нормальний розподіл.

2.4.2.1. Нормальний розподіл імовірностей випадкової величини

Нормальний розподіл (або ***розподіл Гауса***, по імені видатного математика початку XIX сторіччя Карла Фрідріха Гауса). Висока вживаність в дослідженнях саме цього закону зумовлена тим, що нормальний розподіл (більш-менш) адекватно відображає розподіл багатьох безперервних змінних в реальних об'єктах та системах - від параметрів виробничого процесу до результатів перевірки розумових здібностей людей, від ваги людини при народженні до розподілу акцій на біржі.

Інший привід широкого використання нормального розподілу полягає в тому, що за певних умов (що задаються відомою *Центральною граничною теоремою*) можна вважати, що розподіл вибіркових статистик, серед яких і вибіркове середнє арифметичне, буде нормальним, *навіть якщо самі вибірки вибираються з генеральної сукупності, для якої нормальний розподіл не властивий.* Отже, працюючи з вказаними статистиками вибірок доводиться спиратися на властивості саме нормального закону розподілу.

Нагадаємо. В математиці серед інших елементарних функцій добре відома і часто зустрічається т.з. *функція Гауса*, яка має опис:

$$Y = e^{-x^2}$$

і вигляд, що показаний на рис. 2.33. Ця функція, дзвоноподібна (англ. *bell-shaped*), симетрична відносно нульового значення x, завжди позитивна, приймає своє максимальне значення, рівне 1 в тому випадку, коли x = 0.

За своїм виглядом ця функція досить сильно подібна до графіків щільності ймовірності багатьох даних реального світу, що і дало можливість Гаусу запропонувати її модифікацію як функцію, яка апроксимує нормальний закон розподілу ймовірностей, і тепер носить його ім'я:

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

де μ - математичне очікування, σ^2 - дисперсія випадкової величини.

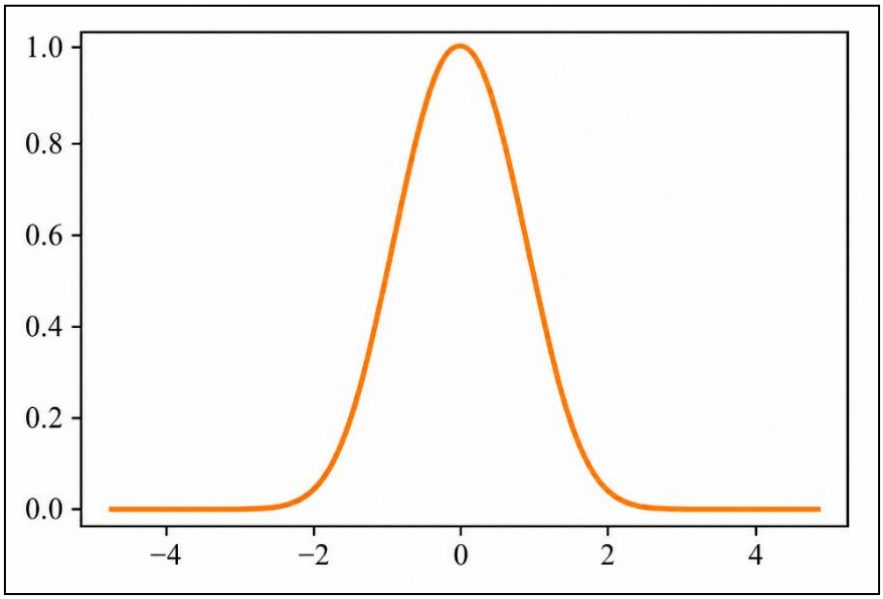


Рис.2.33 Функція Гауса

Часто той факт, що набір даних X підпорядковується нормальному закону розподілу з відповідними значеннями параметрів, позначають як $X \sim N(\mu, \sigma^2)$.

Графік щільності розподілу такої функції називають *нормальною кривою* або *кривою Гауса*. На рис.2.34 представлені графіки цієї функції при різних значеннях параметрів.

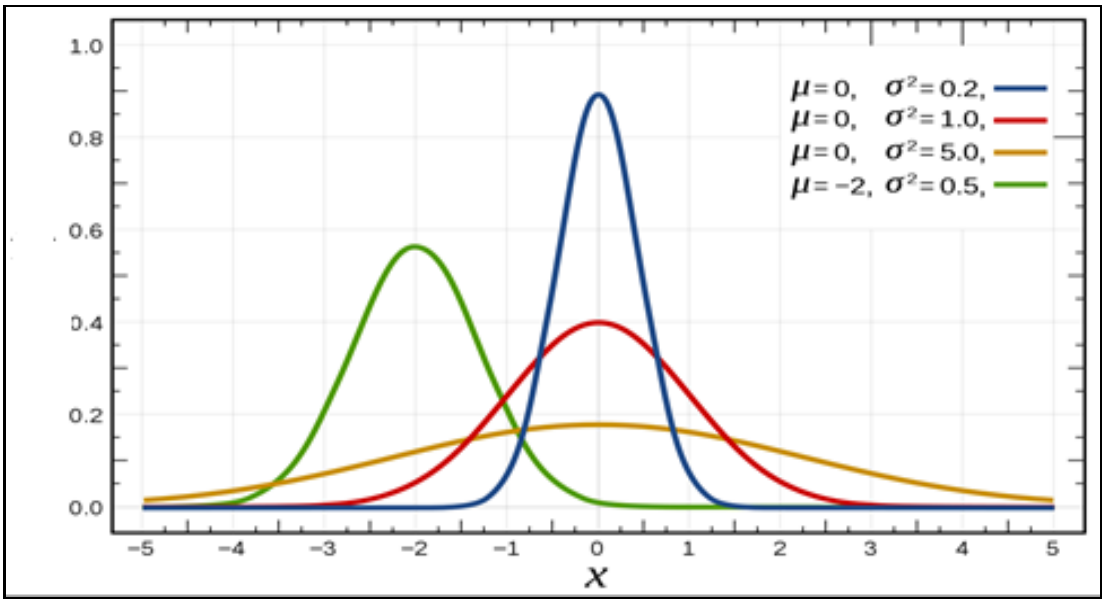


Рис 2.34 Крива Гауса при різних значеннях параметрів

Збільшення математичного очікування μ призводить до зсуву кривої на графіку вправо вздовж осі X, а її зменшення – до зміщення вліво. Із зростанням дисперсії максимальна ордината нормальної кривої убуває, а сама крива стає більш пологою. При зменшенні σ ордината точки максимуму необмежено зростає, при цьому крива пропорційно сплющується вздовж осі абсцис, але так, що площа під її графіком залишається рівною одиниці.

Розподіл, в якому значення параметрів $\mu = 0$ та $\sigma^2 = 1$ називають ***стандартним нормальним розподілом***. (Тобто, розподілом $N(0,1)$). Така функція симетрична відносно значення $x = 0$ при якому вона набуває максимального значення $\frac{1}{\sqrt{2\pi}}$.

Функції ймовірності - щоб запобігти певним непорозумінням їх ще часто називають *«кумулятивними функціями розподілу ймовірності»* (англ. *Cumulative Distribution Function* (CDF), див. розділ 2.3.4) - таких розподілів виглядають так, як це відображено на рис.2.35.

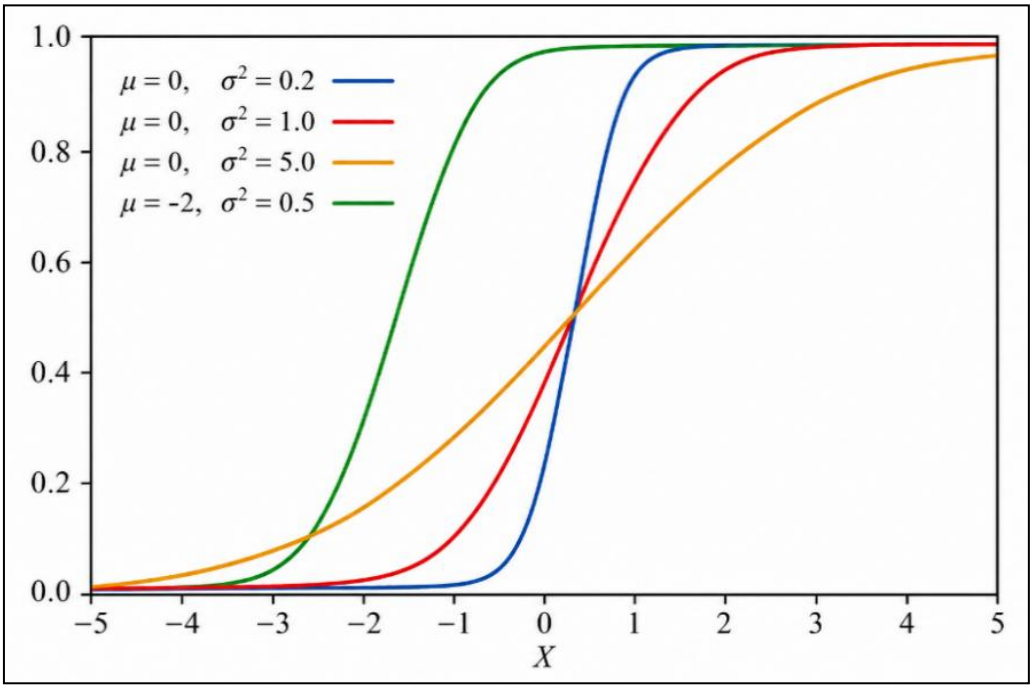


Рис.2.35 Кумулятивна функція розподілу ймовірностей при різних значеннях параметрів

Оскільки на значення параметрів жодних обмежень не накладається, теоретично існує безліч нормальних розподілів, які в цілому мають подібну «дзвоноподібну» форму функції щільності, але різняться через різні значення математичного очікування μ і стандартного відхилення σ. Для всіх нормальних розподілів незалежно від значень їх параметрів характерні деякі загальні властивості. До них відносяться:

- симетричність;
- унімодальність (єдине найбільш часте значення);
- безперервність значень в діапазоні від мінус нескінченності до плюс нескінченності;
- загальна площа під кривою дорівнює одиниці;
- рівність середнього, медіани і моди.

2.4.2.2. z-перетворення та викиди

Стандартний нормальний розподіл відіграє дуже важливу роль в аналізі даних. З ним тісно пов'язано поняття z***-перетворення*** (або *z-нормалізація*, англ. *z-score Transformation*) - базовий інструмент у статистиці та аналізі даних.

Розглянемо визначення:

$$z_i = \frac{x_i - \mu}{\sigma},\ x_i \in X$$

Або, дещо скорочено (у векторній формі):

$$Z = \frac{X - \mu}{\sigma}$$

Тобто, маючи на вході значення набору даних $X = \{x_i\}$ та застосувавши наведені перетворення, отримуємо новий набір даних $Z = \{z_i\}$, з тією-ж кількістю елементів, що і набір X. У формулі μ - математичне очікування генеральної сукупності (або середнє значення вибірки у разі, коли аналогічне перетворення застосовується до вибірки) X, σ - її стандартне відхилення (або його вибіркова оцінка). По суті, z-оцінка показує, наскільки далеко і в якому напрямку (вліво чи вправо) від середнього знаходиться конкретне значення, причому *ця відстань виражена в одиницях стандартного відхилення*. Таким чином, якими б не були

значення наших даних, за умови, що вони розподілені за нормальним законом, в результаті z-перетворення вони стають «*стандартизованими*», тобто такими, середнє значення яких дорівнює 0, а стандартне відхилення дорівнює 1. Форма функції щільності розподілу не змінить своєї дзвоноподібної форми, зміняться лише значення μ та σ. Відтак говорять, що z-перетворення *перетворює дані різних масштабів у єдину шкалу*. Див. рис.2.36. Треба розуміти, що z-перетворення НЕ зводить дані до нормального вигляду, якщо вони з самого початку не відповідали нормальному закону розподілу. z-перетворення *лише масштабує такі дані*.

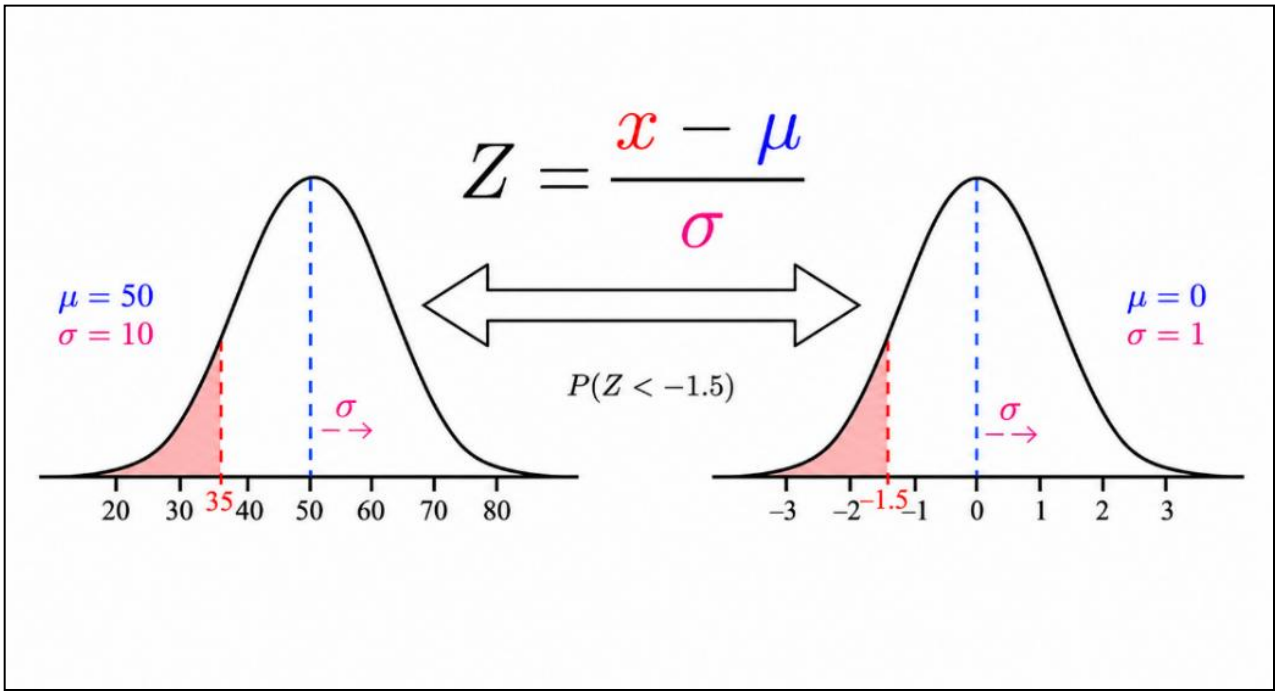


Рис.2.36 Застосування z-перетворення

Попереднє використання z-перетворення суттєво спрощує подальший аналіз даних обчислення ймовірностей, проведення різноманітних статистичних тестів, використання більшості алгоритмів машинного навчання. Зокрема, вони дають можливість порівняння значень генеральних сукупностей з різними середніми арифметичними і стандартними відхиленнями. Наприклад, розглядаючи дві генеральні сукупності: $x \sim N(100,25)$ та $y \sim N(50,100)$, важко зрозуміти, зустрічається число 95 в першій генеральній сукупності рідше або частіше числа 35 у другій генеральної сукупності. Однак таке порівняння можна

легко провести за допомогою Z-значень. В першому випадку отримуємо $Z = \frac{95-100}{5} = -1.0$, у другому випадку $-Z = \frac{35-50}{10} = -1.5$. Отже можемо побачити, що хоча обидва значення нижче середнього в відповідних генеральних сукупностях, друге значення виділяється від математичного очікування сильніше, оскільки -1.5 віддалено від 0 (маточікування стандартного нормального розподілу) далі, аніж 1.0.

Практичний приклад. Уявімо, що проводиться технічна діагностика промислового електродвигуна. Моніторингу піддаються два параметри:

- Вібрація корпусу: 4.2 мм/с;
- Температура обмотки: 105 °C.

Ці параметри важливі для виявлення потенційних збоїв, але вони мають різні одиниці вимірювання й діапазони допустимих значень. Пряме порівняння значень неможливе.

За історичними даними від виробника:

- Для вібрації: середнє μ=2.5 мм/с, стандартне відхилення σ=1.0.
- Для температури: середнє μ=90 °C, стандартне відхилення σ=10.

Обчислюється z-оцінки:

- Вібрація: (4.2−2.5)/1.0=+1.7
- Температура: (105−90)/10=+1.5

Інтерпретація результатів: В обох випадках параметри вище норми. Але у стандартизованій шкалі видно, що відхилення вібрації більш суттєве (+1.7) і, вочевидь, є для обладнання більш критичним, ніж відхилення температури (+1.5). Отже, слід в першу чергу звернути увагу на усунення причини виникнення більш суттєвого відхилення від нормального стану (т.з «пріоритизація обслуговування»).

Раніше (див. Розділ. 2.3.2) розглядався параметр «квантиль» ймовірнісного розподілу. Говорилося, що данні, які рідко зустрічаються в вибірці – це т.з. «викиди», тобто дані, що можуть генеруватися об’єктом в якихось надзвичайних ситуаціях. Наприклад – в випадках помилок, збоїв, аномалій, хвороб тощо. Виникає питання: а який квантиль розподілу обрати для того, щоб відділити такі аномальні значення від нормальних - тобто такі, які є найбільш вірогідними при

нормальному стані об'єкту, який досліджується. Об'єктивного обґрунтування вибору у нас поки що не було. Використання нормального закону розподілу (*в разі, якщо дані добре йому відповідають (!)*) дає можливість пов'язати відповідь на це питання з параметрами розподілу.

На рис.2.37 зроблена «прив'язка» графіку функції розподілу до значень параметрів μ та σ. Зверніть увагу, що на графіку значення даних замінені на їх же значення, але в одиницях, що дорівнюють значенню σ. Графік демонструє, як змінюється ймовірність тих чи інших значень в залежності від відношення цих даних до параметрів μ та σ. (Наприклад, якщо $x = 40, \mu = 20, \sigma = 10$ то на наведеному графіку вказаному значенню x буде відповідати точка, з значенням коефіцієнта при σ що дорівнюю 2. Якщо аналогічний графік побудувати для стандартизованого розподілу - значення μ та σ будуть дорівнювати 0 та 2 відповідно. А от форма графіку, а відтак і значення квантилів, тобто відповідних ймовірностей (площин під кривою значень функції) – не зміняться. (Нагадаємо, що *площа фігури, обмежена* графіком функції щільності розподілу, віссю абсцис і відрізками двох вертикальних прямих, $x = b$, $x = c$, є *ймовірність попадання випадкової величини в інтервал (b, c)*). Отже можливо працювати або в одній шкалі (наданих значень) або в іншій (z-значень). Цей зв'язок треба постійно мати на увазі. Можливо переходити до одної чи іншої форми представлення в залежності від того, де простіше отримати вирішення завдань, які перед будуть поставлені.

Аналіз показує, що випадкова величина, розподілена за стандартним нормальним законом *N(0,1)*, з ймовірністю, що дорівнює приблизно 0.68 потрапляє в інтервал (-1,1); з ймовірністю 0.94 потрапляє в інтервал (-2, 2); з ймовірністю, приблизно рівній 0.9973 - в інтервал (-3, 3). Підкреслимо, тут інтервали задані в значеннях . Звідси для довільної нормально розподіленої випадкової величини можна сформулювати правило, відоме в літературі як «правило трьох сигм». А саме, нормальна випадкова величина *N(a,σ)* з ймовірністю 0.9973 потрапляє в інтервал (a - 3σ, a + 3σ). Вихід за цей інтервал має ймовірність 0.0027. З іншого боку, значення, яке, наприклад, перевищує 3σ, може бути рідкісним винятком, але й може бути і помилкою. Отже правило трьох (або будь-якого іншого значення) сигм можна застосовувати для визначення кордонів, які відділяють нормальні значення від аномальних. Але підходити до

цього визначення формально - ризиковано. Його завжди треба супроводжувати відповідним семантичним аналізом.

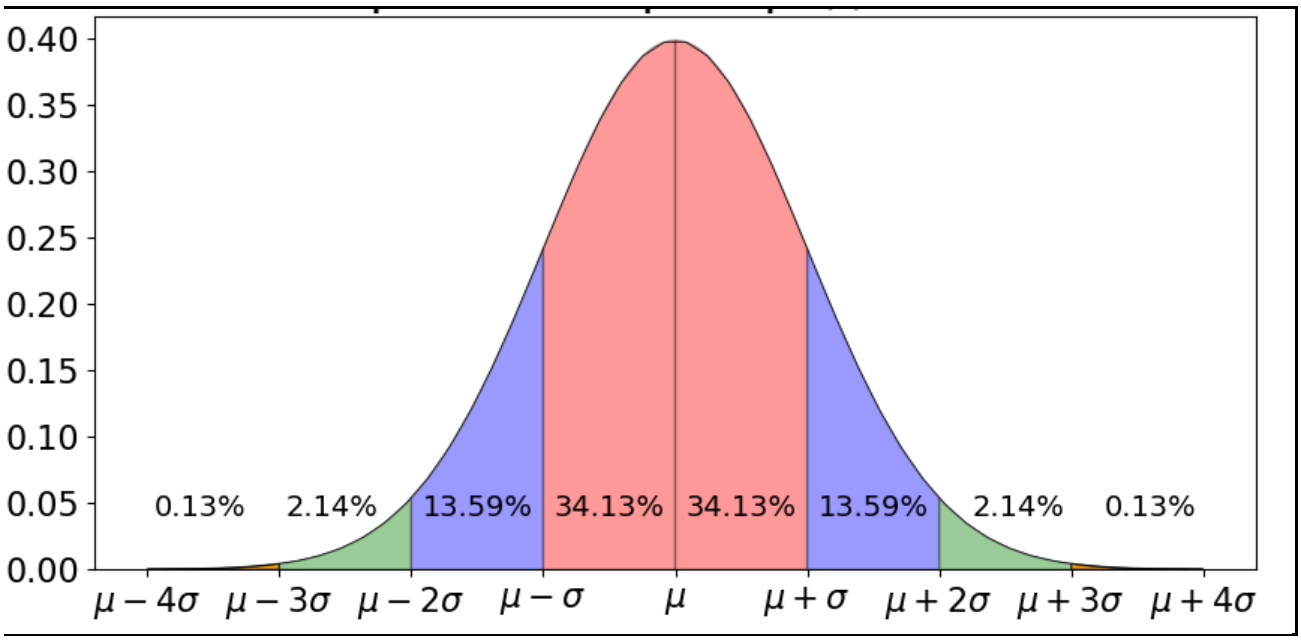


Рис 2.37 Представлення функції щільності розподілу з прив'язкою до параметрів μ та σ

(Повертаючись до останнього прикладу, можна відразу сказати, що значення CDF у точці $x = 40$ буде дорівнювати 0.13%+2.14%+13.59%+34.13%+34.13%+13.59%=100%-0.13%-2.14%=97.3% або 0.97. Спробуйте також самостійно розібратися в усіх діях, що були виконані).

Просто цікава інформація: значення розподілу при 5σ, дорівнює приблизно 10^{-6}; для 10σ приблизно 10^{-22} , а при $15\sigma \approx 10^{-50}$. Для порівняння вік всесвіту в даний час оцінюється приблизно в 15 мільярдів років, що становить близько $4 \cdot 10^{17}$ секунд. Таким чином, навіть якщо б і була можливість робити по тисячі випробувань в кожну секунду, що минула від початку часів, ми все одно не мали б шансів натрапити на значення, більше ніж значення при 10σ. Звичайно, за умови, що дані розподілені за нормальним законом.

З іншого боку ми знаємо, що для деяких систем великі викиди-нормальна практика, тобто викиди в реальності спостерігаються набагато частіше. Це означає, що для таких систем гаусова модель і засновані на ній теорії

неспроможні, а *невиправдане їх використання в таких випадках призводить до некоректних або взагалі помилкових результатів*.

Багато алгоритми і методів аналізу даних вельми чутливі до викидів в даних. Зокрема, результати їх застосування можуть виявитися некоректними (помилковими) у випадку присутності викидів у наданих вибірках. *Тому етап попереднього аналізу даних на наявність в них викидів рекомендується виконувати на самому початку будь-якого дослідження систем.* Причому при виконання такого аналізу можуть застосовуватися і формальні техніки (наприклад - метод трьох сигм), але нікуди не діється і від неформальних технік. (Наприклад, значення віку людини у «300 років» явно, за своєю семантикою, вказує на викид, помилку в даних).

Вище (див. розділ 2.3.4.3) говорилося про т.з. «усічені середні» - своєрідне використання відкидання потенціальних викидів - декількох крайніх за величинами значень з множини даних, що досліджується. Це досить примітивний підхід. Якщо продовжити ідею попереднього очищення даних, то в випадку застосування вона буде полягати в тому, що-б знайти і відкинути дані, значення яких більше (або менше) обраних *«порогових значень»,* (англ. *Threshold*). Як завжди, тут виникає нова проблема. Ми можемо так «розкидатися» даними, якщо їх у нас дійсно багато. А якщо даних відносно мало, такий підхід - з відкиданням наданих для дослідження даних - явно не годиться. Для таких випадків застосовують спеціальні алгоритми, які менш чутливі до викидів (т.з. *«робастні алгоритми»).*

На рис.2.38 показано, як викиди можуть проявлятися на графіку трафіку в мережі (червоні позначки). Але це найпростіший випадок. Бувають дані, в яких аномалії (або помилки) в даних можуть бути виявлені тільки аналізуючи сукупність значень параметрів для об'єкта (див. Рис.2.39). В даному разі по жодній з координат дані не виходять за межі «трьох сігм», однак червону точку можливо ідентифікувати як явний як викид. Добре, якщо у нас дві координати, а якщо їх більше, і, відповідно, нічого намалювати ми не можемо?

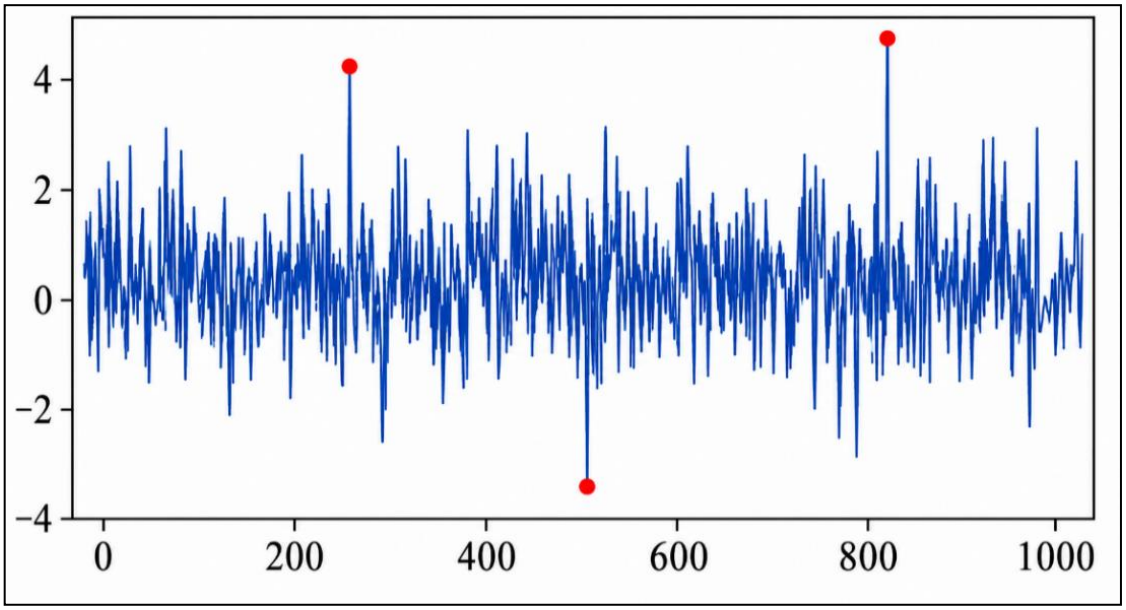


Рис 2.38 Приклад аномалій (викидів) в часовому ряду

Згадаємо також приклад з пологовим будинком, що наводився раніше. Зрозуміло, що кількість народжень в конкретному будинку в конкретний день - це певна оцінка генеральної сукупності всіх народжених. Якщо в конкретному будинку в певний день зафіксовано народження 7 хлопчиків і 3 дівчаток – наскільки можливий (ймовірний) такий випадок? Чи повинні такі дані викликати у нас сумніви? А якщо ми отримуємо дані, що у всіх пологових будинках країни в певний день народилося 700 хлопчиків і 300 дівчаток? Цей приклад показує, наскільки наша оцінка відрізняється в залежності від об'єму вибірки, навіть в побутових випадках. І цю залежність ми маємо собі занотувати, до її аналізу ми повернемося трохи пізніше (див. Розділ 3.2.3)

Окрім стандартизації через використання z-перетворення, застосовують і інші методи приведення значень параметрів до узгодженого вигляду. Такі методи найбільш широко використовуються в моделях машинного навчання, що базуються на використанні матриці відстаней, наприклад методах K-найближчих сусідів, SVM та в нейронних мережах. Зацікавлені читачі, а також тим, хто така нормалізація знадобиться в ході подальшого навчання (при виконанні бакалаврської та магістерської роботи) чи в подальшій професійній діяльності, можуть почати знайомитися з цими питаннями, наприклад, за посиланням [38].

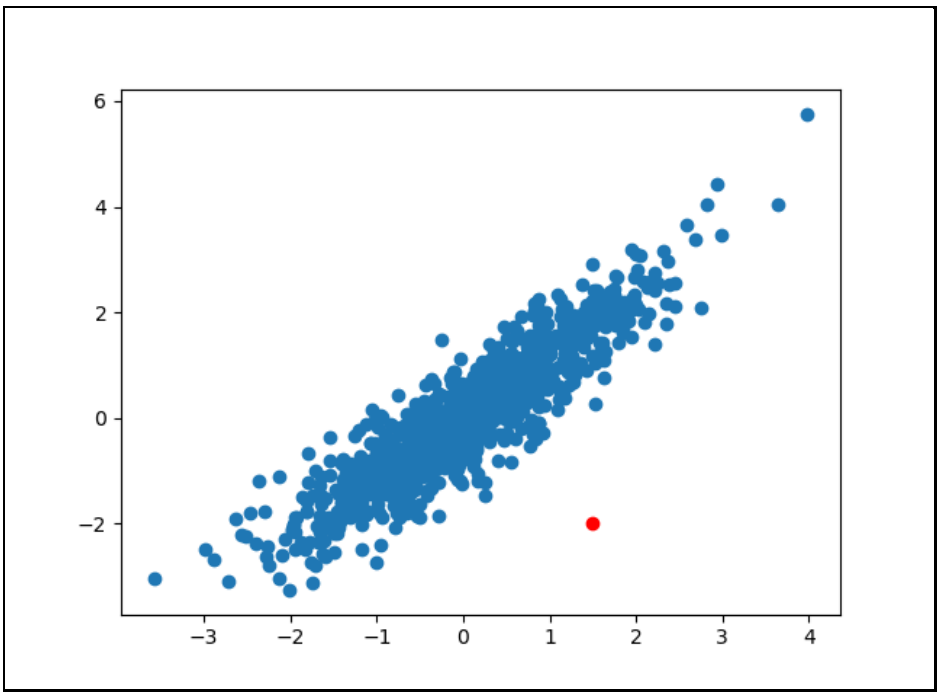


Рис.2.39 Приклад аномалій в двовимірній вибірці даних

2.4.3. Дані, що не підпорядковуються нормальному закону розподілу

В практичних завданнях значення експериментальних даних досить часто відповідають іншім, відмінним від нормального, законам розподілу. Наприклад відомо, що в задачах біостатистики лише 20-30% даних відповідають нормальному розподілу. Багато реальних даних мають явно виражену скошеність в розподілах. Скошеність має місце, наприклад, тоді, коли дані не можуть мати негативних значень, що зустрічається в надзвичайно великій кількості випадків. Багато фінансових експертів вважають, що зміна ціни акцій на біржі або вартості криптовалюти краще описуються розподілами з важкими хвостами, ніж нормальним розподілом. Час між підозрілими підключеннями (*brute-force атака*) часто має експоненційний розподіл: багато дуже коротких інтервалів (зловмисник намагається якнайшвидше перебирати паролі), а довгі інтервали трапляються рідко. Розмір файлів у мережевому трафіку (HTTP/FTP) має стиснутий степеневий розподіл: дуже багато маленьких файлів (іконки, запити сторінок), і дуже мало - надвеликих (відео, архіви). Довжини шкідливих пакетів у DDoS-атаках нерідко мають бімодальний розподіл (дві «піки»): один пік на мінімальних розмірах пакетів (шум), інший - на максимальних (UDP flood). Час до відмови обладнання зазвичай описується розподілом Вейбулла: на початку

ймовірність відмов дуже мала, з віком обладнання вона зростає. Розподіл вібраційних амплітуд у двигуні може мати логнормальний розподіл, бо вібрації часто зростають мультиплікативно (накладаються дрібні фактори). Час виживання пацієнтів після операції або початку терапії підпорядковується розподілу Вейбулла або експоненційному розподілу. (Причина: більшість пацієнтів одужують чи помирають у перші місяці - високий ризик на початку лікування, далі ризик зменшується і «хвіст» розподілу розтягується). Доходи населення - більшість людей мають низький або середній дохід, тоді як дуже заможні формують «важкий хвіст» розподілу. Приклади можна продовжувати дуже довго.

І з такими даними треба працювати.

2.4.3.1. Логнормальний розподіл

Коли дослідник бачить графік функції щільності розподілу приблизно такого вигляду, як на рис. 2.40, перше що приходить на думку – вияснити, чи це т.з. логнормальний розподіл.

Логнормальний розподіл - це двопараметричний, неперервний, унімодальний, лівоскошений розподіл ймовірностей даних, які мають виключно позитивні значення. (*Не плутати з логарифмічним розподілом !!*)

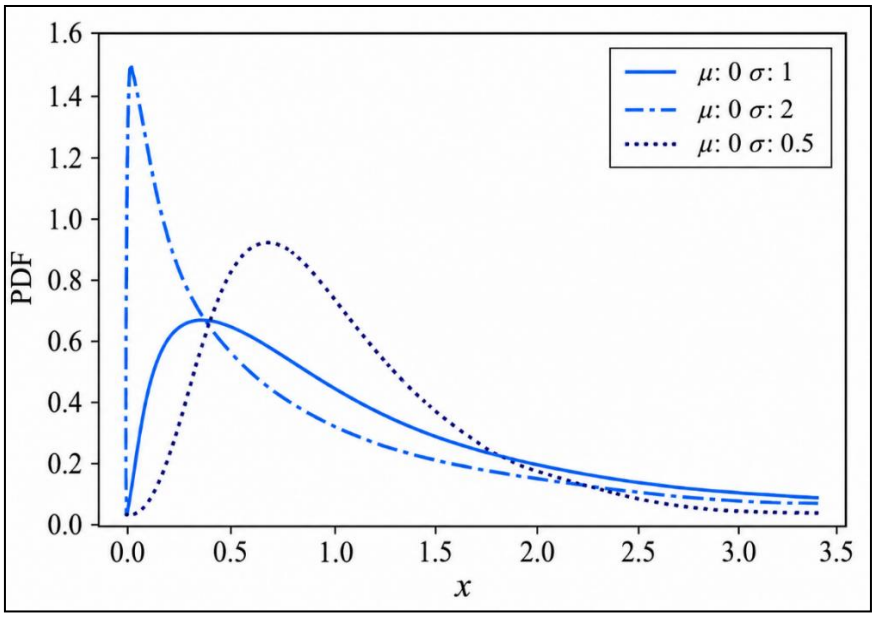


Рис. 2.40 Приклади функцій щільності, що описують логнормальний закон розподілу при різних значеннях параметрів

При використанні нормального розподілу для опису моделі реального об’єкту чи процесу приймалося припущення, що всі невраховані моделлю випадкові впливи можна вважати такими, які різнонаправлено впливають на результуючий параметр і незалежні між собою. При використанні логнормального розподілу вважається, що вплив вказаних факторів залежить від того, які дії відбулися раніше. Найпростіший приклад – аналіз результатів процесу удару одного механічного об’єкту об другий. Якщо розглядати серію таких ударів, кінцевою метою якої є руйнація одного з об’єктів, то логічно припустити, що вплив на об'єкт 10-ого удару буде більш відчутним, ніж вплив першого або другого. На цьому припущенні, наприклад, побудована теорія втоми (руйнації) матеріалів. У вказаному досліді кількість ударів, що знадобиться для руйнації об’єкту може бути описана логнормальним законом розподілу. Такі показники, як час руйнування (виходу з ладу) виробу або площа розповсюдження корозії теж можуть бути промодельовані з використанням логнормального закону. У фізиці логнормальний розподіл з високим ступенем адекватності описує розподіл частот частинок за їхніми розмірами (або масами) у процесах випадкового дроблення, фрагментації чи росту. Неочікувано, але за логнормальним законом розподіляється вартість квадратного метра нерухомості. Розмір градин у граді, тривалість партій в шахах і час збору кубика Рубіка; довжина інертних придатків (волосся, кігті, нігті, зуби) біологічних зразків у напрямку росту; приріст людини по роках; кількість опадів, чисельність видів рослин, що проживають на певній ділянці; глобальний розподіл доходів; відсоток жиру в організмі – всі ці явища, процеси, об’єкти можуть бути описані з використання цього логнормального закону розподілу.

У дослідженні даних, що пов’язані з забезпеченням інформаційної безпеки логнормальному розподілу здебільшого відповідає розподілення розміру кадрів, що передаються в мережі або сумарний об'єм інформації, що передається за одиницю часу. Довжина коментарів, що розміщуються на форумах в Інтернеті, час, що витрачається користувачами на читання онлайн-статті (жарти, новини тощо) також відповідають логнормальному розподілу. Розподіл розміру загальнодоступних файлів аудіо та відео (типи *MIME*) відповідає логнормальному розподілу. Спільне в розподілі даних в цих прикладах - наявність так званого «важкого хвоста», тобто значень, що лежать далеко від середнього, і помітною асиметрії, як правило, правою.

Формально, функція щільності логнормального розподілу описується наступним чином:

$$f(x,\mu,\sigma)=\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2},\ x>0.$$

Факт розподілу значень набора X за логнормальним законом, позначають як $X \sim Log\left(\mu,\sigma^2\right)$. Якщо $X \sim Log\left(\mu,\sigma^2\right)$ то $Y=\ln\left(X\right) \sim N(\mu,\sigma^2)$ І навпаки, якщо $Y \sim N(\mu,\sigma^2)$, то $X=e^Y \sim Log\left(\mu,\sigma^2\right)$. Рис.2.41 наочно демонструє ці співвідношення між розподілами наборів X та Y.

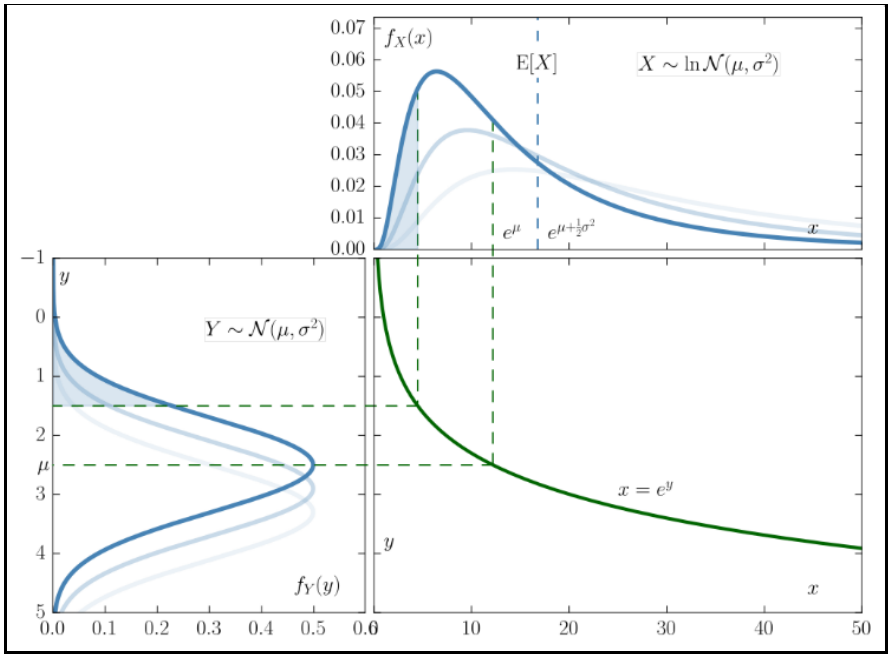


Рис.2.41 Співвідношення логнормального та нормального законів розподілу

У наведених визначеннях параметр μ є параметром розташування, а σ - параметром масштабу розподілу. *Важливо не плутати ці параметри із середнім значенням (на Рис. 2.41 позначеним як E[X]) або стандартним відхиленням самого логнормального розподілу.* Зв'язок між ними та звичними статистичними показниками виникає лише під час логарифмування:

1. Без перетворення: Для не перетворених даних X μ та σ - це просто математичні параметри, що визначають логнормальну функцію. Вони не є ні

середнім значенням, ні стандартним відхиленням, які обчислюються за стандартними процедурами для набору даних *X*.

2. Після перетворення: Якщо логнормальні дані перетворити за допомогою логарифмування, то μ буде відповідати середньому значенню, а σ - стандартному відхиленню цих вже перетвореного набору даних *Y*.

На Рис.2.41 середнє значення логнормального розподілу позначене як $E[X]$, і воно математично відрізняється від параметра μ.

Найефективніший спосіб аналізу даних з логнормальним розподілом полягає у застосуванні відомих методів, заснованих на нормальному розподілі, до логарифмічно перетворених даних, а потім, за потреби, у зворотному перетворенні результатів. Позатим, можна показати, що якщо працювати з традиційно отриманими значеннями середнього $\overline{x}$ (статистична оцінка $E[X]$) та стандартного відхилення s (статистична оцінка $D[X]$) для вибірки X, що розподілена за логнормальним законом, то:

$$\hat{\mu} = \ln\left(\frac{\overline{x}}{\sqrt{1+\left(\frac{s}{\overline{x}}\right)^2}}\right) \text{ та } \hat{\sigma} = \sqrt{\ln\left(1+\left(\frac{s}{\overline{x}}\right)^2\right)}$$

є статистичними оцінками параметрів μ та σ розподілу значень X.

2.4.3.2 Рівномірний розподіл

Рівномірний розподіл – це розподіл випадкових величин, за якого всі можливі результати або значення на визначеному відрізку мають однакову ймовірність. За межами цього відрізка ймовірність зустріти значення з такої вибірки дорівнює 0. Рівномірний розподіл буває двох типів: дискретний, де випадкова величина приймає скінченну кількість дискретних значень з однаковою ймовірністю (як при підкиданні монети або грального кубика), і безперервний, де ймовірність залежить від довжини інтервалу, а всі значення в межах певного інтервалу мають однакову щільність розподілу ймовірності.

Щільність розподілу ймовірності для ***безперервного рівномірного закону розподілу*** (англ. *Continuous Uniform Distribution*) даних визначається як:

$$f(x,a,b)=\begin{cases}\dfrac{1}{b-a}, \textit{якщо } a\le x\le b;\\ 0, \quad \textit{якщо } x>b \textit{ або } x<a.\end{cases}$$

Графік функції щільності розподілу для цієї функції наведений на рис. 2.42. Відповідна кумулятивні функція задається наступним чином:

$$F(x,a,b)=\begin{cases}0, \quad \textit{якщо } x<a;\\ \dfrac{x-a}{b-a}, \textit{ якщо } a\le x\le b;\\ 1, \quad \textit{якщо } x>b.\end{cases}$$

Зрозуміло, що зі збільшенням різниці між значеннями b та a висота прямокутника на графіку функції щільності зменшується (див. Рис.2.42). Це безпосередньо випливає з того факту, що площа під графіком функції (формальніше – інтеграл від функції) для будь якої функції щільності розподілу має дорівнювати 1. Для наборів даних, що розподілені за безперервним рівномірним законом, оцінка математичного очікування визначається як $\frac{(b+a)}{2}$, медіана дорівнює середньому, дисперсія - $\frac{(b-a)^2}{12}$.

Той факт, що набір даних підпорядковується безперервному рівномірному розподілу позначають як $X \square U(a,b)$. Якщо $a=0, b=1$, то такий розподіл називають *стандартним безперервним рівномірним розподілом*. Маючи генератор випадкових чисел, що отримані вибором з $U(0,1)$, легко побудувати генератор будь-якого безперервного рівномірного розподілу $Y \square U(a,b)$. А саме $Y=a+(b-a)X$.

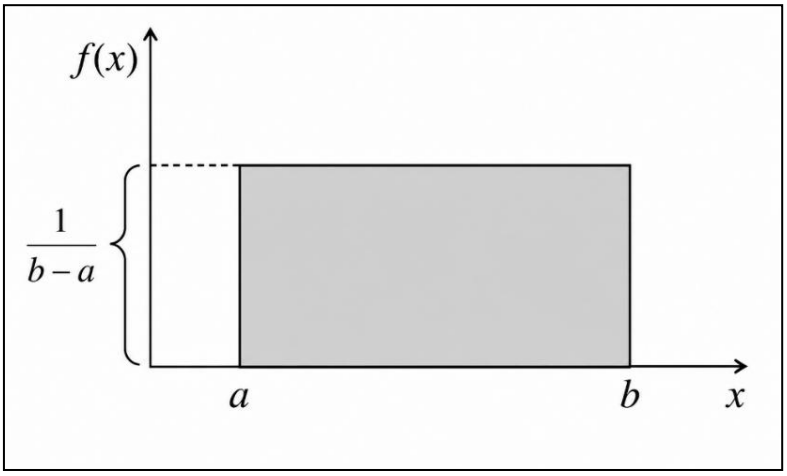


Рис 2.42 Функція щільності безперервного рівномірного розподілу

Безперервний рівномірний розподіл є універсальним інструментом для моделювання процесів, в яких немає підстав вважати одні значення імовірнішими за інші. Так, у кібербезпеці він забезпечує опис непередбачуваності (наприклад, випадковий вибір моментів перевірок), в IT - рівномірності розподілу ресурсів (випадковий розподіл навантаження між серверами), у телекомунікаціях - для моделювання випадкового часу появи пакета в межах тайм-слоту, процес розподілу вхідних запитів між серверами в кластері (наприклад, у хмарних обчисленнях) з метою забезпечення рівномірного завантаження серверів. У технічній діагностиці - для моделювання шумів і відхилень у допустимих межах, в енергетиці - для випадкового вибору моментів вимірювання параметрів мережі в межах інтервалу часу, у будівництві - для моделювання випадкових відхилень товщини шару матеріалу в межах допусків, розподіл насіння при сівбі, добрив при підживленні також моделюються як рівномірний розподіл по площі, відповідно води, зерна, препаратів тощо.

Дискретний рівномірний розподіл (англ. *Discrete Uniform Distribution*) - це розподіл ймовірностей, який описує ймовірність значень у випадку, коли кожне значення з скінченої множини їх можливих значень є однаково ймовірним. Розподіл характеризується постійною функцією маси ймовірності (нагадуємо, так називають аналог функції щільності ймовірності для дискретних функцій) у скінченному діапазоні значень. Найпростішим прикладом процесу, що описується цим розподілом є процес випадання граней грального кості. У прикладних галузях він описує процеси з випадковим вибором з фіксованого набору опцій. Наприклад, у моделях безпеки процес оцінки успішності атаки

(кількість портів, сканованих успішно, з фіксованої кількості) моделюється як дискретний рівномірний розподіл, щоб тестувати стійкість систем до різних сценаріїв. У технічній діагностиці в цей розподіл моделює ймовірності виникнення кожної окремої помилки при аналізі надійності системи.

Для даних, розподілених за дискретним рівномірним розподілом, функція маси (щільності) ймовірності може бути визначена як:

$$f(i,N)=\frac{1}{N},\ \forall i$$

Тобто, значення цієї функції залежить виключно від кількості елементів в сукупності даних, і не залежить від i. (Тут звичне позначення для значень набора що досліджується «x» замінено на «i» щоб підкреслити дискретний характер значень). Будемо вважати, що $i\in\{a,a+1,.....b-1,b\}$. Тоді можна записати:

$$f(i,a,b)=\frac{1}{b-a+1}$$

а кумулятивна функція розподілу визначається як:

$$F\left(i,a,b\right)=\frac{i-a+1}{b-a+1}$$

Для вибірок, що розподілені за дискретним рівномірним розподілом значення математичного очікування визначається як:

$$\mu=\frac{b+a}{2}$$

а дисперсія – як:

$$\sigma^2=\frac{\left(b-a+1\right)^2-1}{12}$$

Наприклад, позначивши сторони кубика відповідним числом від 1 до 6, можна підрахувати, що середнє значення, що отримується при «нескінченій» послідовності кидків кубика буде дорівнювати 3.5, а дисперсія – 2.917.

Спробуйте проінтерпретувати ці результати. На цьому прикладі ще раз підкреслимо, що випадкова величина ніколи не може приймати цього «очікуваного» значення для будь-якого окремого спостереження. Натомість, це те, що очікується як довгострокове середнє значення за багатьма (в ідеалі - нескінченої множини) спостереженнями.

За допомогою наведених вище формул можливо вирішувати і більш серйозні задача. Наприклад, т.з. задачу «німецьких танків», що постала під час другої світової війни. Тоді розвідка союзників намагалася провести статистичну оцінку німецького виробництва танків «Пантера». У той час звичайна розвідка оцінювала об'єм їх виробництва як 1000-1400 або навіть більше танків щомісяця, що спричиняло психологічний шок в керівництві союзників. З початком бойових дій союзні війська почали захоплювати деякі з цих танків та ретельно їх досліджували. Вони виявили, що серійні номери коробок передач були послідовними від 1 (перший побудований танк) до N (найновіший виготовлений). Німці не були б німцями, якби у них не було суворого порядку. Тобто, кожен окремий захоплений танк мав номер коробки передач з номерами від 1 до N, де N був параметром, для якого терміново потрібна була оцінка. Номери захоплених танків з статистичної точки зору можна розглядати, як ймовірнісну вибірку з генеральної сукупності (усіх виготовлених танків). Використовуючи властивості рівномірного розподілу виявили, що вибірка з захоплених танків мала максимальний серійний номер m. Якщо взяти всі номери з досліджених танків, то їх послідовність має також бути рівномірно розподілена на інтервалі від 1 до N. Таким чином можна оцінити різницю між усіма послідовними парами номерів захоплених танків. А відтак вважати, що найбільший номер захопленого танку плюс оцінка «відстані» між парами номерів і буде приблизною оцінкою кількості танків, що були випущені на поточний момент. Звичайно, в реальності все було набагато складніше, бо треба було врахувати і, наприклад, що танки, що належали одному підрозділу, мали близькі номери і ці дані могли вплинути на результат. І те, що багато танків часто відправлялися на ремонт, де трансмісія замінювалася, і багато інших факторів. Позатим, оцінки статистиків була в 4-6 разів менша, за оцінку розвідки! І науковцям вдалося довести, що їх оцінки точніше. Це дозволило скинути психологічний тягар і зрозуміти, що ворог менш потужний ніж вважалося. Стало більш зрозумілим як діяти з своїми наявними ресурсами, зокрема засобами протитанкового захисту тощо. А вже після війни,

коли до рук переможців потрапили всі документи німців щодо випуску танків, виявилося, що оцінка статистиків відрізнялася від дійсних цифр не більше, ніж на 10%. Це звісно, не *кібер*безпека, але це приклад того, як «нудна» наука може працювати у реальному житті, в тому числі і в галузі безпеки.

2.4.3.3. Експоненційний розподіл

Ще один неперервний розподіл, який необхідно дослідити - показниковий або ***експоненційний розподіл***, що моделює, зокрема, час між двома послідовними завершеннями (або настанням) однакових подій. Наприклад потік пакетів, які циркулюють в телекомунікаційній мережі, складається з великого числа незалежних потоків пакетів, що генеруються різними незалежними джерелами, які мають доступ до цієї мережі. В даному разі подія - це момент надходження пакету у мережу. З такої точки зору можуть вивчатися і багато інших об'єктів чи процесів, а саме потік викликів швидкої допомоги; потік крадіжок, що зареєстровані в певному районі; потік помилок друку на сторінках книги тощо. Цей розподіл також тісно пов'язаний з даними типу «*часу дожиття*». Розуміти цей термін слід досить широко. У медико-біологічних дослідженнях під ним може матися на увазі тривалість життя хворих при клінічних дослідженнях після проведення лікування, в техніці - тривалість безвідмовної роботи пристроїв, в психології - час, витрачений на виконання тестових завдань. Цьому закону також підпорядковується розподіл часу між запитами до веб-сайту. Ключовою властивістю процесів, що описуються цим законом розподілу є *«безпам'ятність»: імовірність настання події в найближчий проміжок часу не повинен залежити від того, скільки часу вже минуло від початку дослідження*.

Згадаємо розподіл Пуассона, про який було сказано раніше (див. розділ 2.4.1). Але якщо процес Пуассона використовує в якості параметру кількість подій, що відбуваються в певний проміжок часу, тобто є дискретним розподілом, то експоненційний закон моделює час між настанням двох подій, а цей параметр має безперервну природу, хоча і об'єкт дослідження і самі події можуть бути тими самими. А відтак і параметр, який використовується для опису функцій цих законів однаковий - інтенсивність потоку подій λ.

Позначимо $x-$ випадкову величину «проміжок часу між двома послідовними подіями». Щільність ймовірності експоненціального розподілу задається функцією:

$$f(x) = \lambda e^{-\lambda x}, \text{x} \geq 0$$

а кумулятивна функція розподілу:

$$F(x) = 1 - e^{-\lambda x}$$

Зрозуміти поведінку цих функцій при різних значеннях параметру λ можна з Рис.2.43.

Маточікування і дисперсія експоненціального розподілу визначаються як:

$$\mu = \frac{1}{\lambda} \text{ та } \sigma = \frac{1}{\lambda^2}$$

Маточіківання при цьому може інтерпретуватися як середній час між настанням подіями.

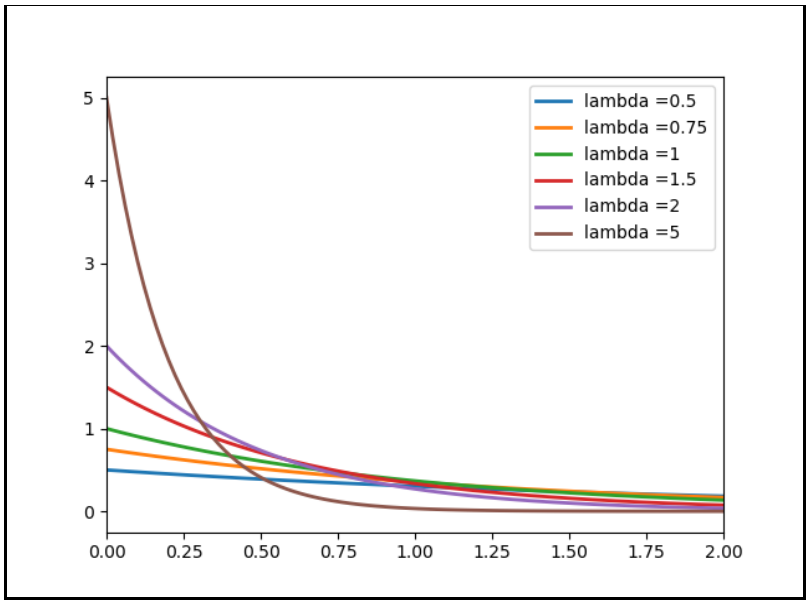


Рис. 2.43 Функція щільності експоненційного розподілу при різних значеннях параметру

2.4.4. «Штучні» закони розподілу

Розглянуті вище закони розподілу в той чи інший спосіб були пов'язані з деякими процесами реального світу. Окрім зазначених існують ще велика кількість таких теоретичних законів. Кожний з них може і мусить використовуватися лише до певного типу даних, а спроба їх використання поза цими межами призводить до помилкових результатів та висновків. Однак існують і активно застосовуються закони розподілу, знайти яким відповідність у реальному світі не так просто або навіть неможливо. Позатим важливість вивчення таких законів зумовлюється тим, що вони широко застосовуються не при первинному, безпосередньому, аналізі даних, а при поглибленому статистичному дослідженні. Наприклад - при спробі оцінити, як саме статистики, що отримані для вибірки, відповідають параметрам генеральної сукупності. Нагадаємо, параметри генеральної сукупності є незмінні величині та як правило нам невідомі. А вибіркові статистики – це результати алгоритмічної обробки наданих для дослідження даних. Вони гарантовано змінюються від вибірки до вибірки, навіть взятих з однієї й тієї ж генеральної сукупності, і їх можна розглядати як окремий випадковий набір даних. І виявляється, що ці статистики, як випадкові величини, теж підпорядковуються певним законам розподілу, але найчастіше – не найпростішим.

Розглянемо три таких розподіли, що використовуються найширше.

2.4.4.1. t-розподіл Стьюдента

t-розподіл Стьюдента – це розподіл, функції щільності якого виглядають майже ідентично кривій нормального розподілу, тільки трохи «нижчі» та мають трохи товстіші хвости. t-розподіл використовується, коли в наявності у дослідника вибірки невеликого розміру.

Історична довідка. t-розподіл був описаний у 1908 році англійцем, хіміком, статистиком, який працював в лабораторії на пивній броварні Гіннесса в м. Дублін (Ірландія). Хіміка звали Уільям Сілі Госсет (*William Sealy Gosset*) і він аналізував якість проб пива шляхом хімічного аналізу. Але оскільки його фірма забороняла своїм співробітниками розголошувати будь які дані щодо технологічного процесу, він змушений був публікувати такі результати – по суті одні з найвидатніших у всій історії статистики – під псевдонімом Стьюдент (*Student*), що означало просто «Студент». Так вони і ввійшли в сучасну наукову

термінологію. Постійно знімаючи проби, занотовуючи показники, аналізуючи результати, Госсет виявив, що отримані ним *розподіли реальних дослідних зразків* трохи більш «розмазані», *ніж це повинно бути* у звичайному стандартному нормальному розподілі. *І це не похибка дослідження, а базова властивість об'єктів будь якої природи.* В його експериментах хвости вибірок ставали більш «важкими». Це означає, якщо при розрахунках квантилів розподілів («площі під кривою») практичні результати будуть відрізнятися від розрахованих (передбачених). Госсет ввів, описав та дослідив новий закон розподілу (*t-розподіл Стьюдента*) та показав як його використовувати в реальних задачах. Це суттєво підвищило точність отриманих результатів. *По суті, t-розподіл дає можливість врахувати додаткову невизначеність, яка виникає, коли дисперсію генеральної сукупності замінюють її вибірковою оцінкою.* А на практиці це буває майже завжди.

Формально, функції щільності ймовірності t-розподілу визначається формулою:

$$f(t)=\frac{\Gamma\left(\frac{df+1}{2}\right)}{\sqrt{df\,\pi}\,\Gamma\left(\frac{df}{2}\right)}\left(1+\frac{t^2}{df}\right)$$

де $\Gamma(.)$-відома з курсу математики гама-функція,

df - кількість ступенів свободи (див. нижче).

Розподіл Стьюдента *симетричний відносно нуля*. Зі збільшенням розміру вибірки t-розподіл наближається до нормального розподілу і стає тотожнім з ним при нескінченій кількості зразків даних. Див. Рис. 2.44.

Використання розподілу Стьюдента *дає можливість отримувати більш точні результати саме в умовах обмеженої кількості наявних експериментальних даних.* (Детальніше про це буде описано в Розділі 3). Сьогодні t-розподіл є базовим інструментом у статистиці, особливо там, де даних небагато, але точність отриманих результатів критична. Для аналітики кіберсистем це означає можливість робити статистично обґрунтовані висновки навіть у ситуаціях обмеженості даних - що трапляється досить часто під час розслідувань атак, виявлення аномалій, збору зразків трафіку тощо.

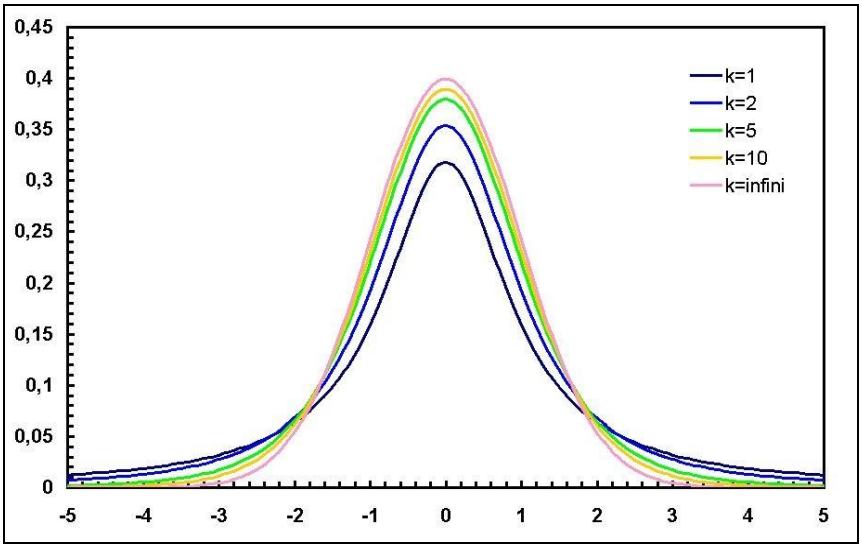


Рис.2.44 Функція щільності t-розподілу Стьюдента при різних значеннях параметру

t-розподіл не позбавлений і певних недоліків. Зокрема, він досить чутливий до відхилення даних від нормальності. Коли дані значно відхиляються від нормального розподілу, використання t-розподілу може призвести до неточностей у статистичних висновках. Викиди також можуть спотворювати результати. Позатим t-розподіл служить життєво важливим інструментом у статистиці, особливо під час оцінки значущості статистик малих за розмірами вибірок або вибірок з невідомими варіаціями. Хоча t-розподіл має дзвоноподібну та симетричну форму та характеристики близькі до нормального розподілу, він відрізняється більш товстими хвостами, що призводить до більшої ймовірності екстремальних значень. Розуміння його властивостей та застосувань є важливим для точного статистичного висновку в сценаріях, коли не виконуються припущення про нормальність та відоме стандартне відхилення популяції.

2.4.4.2. χ^2-розподіл

χ^2-розподіл (читається «хі-квадрат»), запропонований на початку XX сторіччя англійським вченим статистиком Карлом Пірсоном і *використовується при аналізі дисперсії даних вибірки*. Цей безперервний розподіл є розподілом суми квадратів стандартизованих нормальних величин. Формально, функція щільності χ^2-розподілу визначається за формулою:

$$f(x;k)=\frac{1}{2\ \Gamma\left(\frac{k}{2}\right)}x^{\left(\frac{k}{2}-1\right)}e^{-\frac{x}{2}};\quad x\geq 0$$

де $\Gamma(.)$ - гама-функція,

k - кількість ступенів свободи.

χ^2-розподіл насправді є серією (родиною) розподілів, форма яких змінюється залежно від кількості їх ступенів свободи. З її збільшенням розподіл стає більш симетричним і наближається до нормального розподілу. χ^2-розподіл визначається лише для невід'ємних значень, оскільки він базується на сумі квадратів стандартних нормальних змінних, які є невід'ємними по визначенню.

Пірсон показав, що якщо $X=\{x_i\}$ є вибіркою з нормально розподіленої генеральної сукупності, то величина:

$$\chi^2=x_1^2+x_2^2+\ldots+x_n^2$$

матиме χ^2 - розподіл з n ступенями свободи.

Для різних обсягів вибірки (точніше – для різних значень числа ступенів свободи - див. розділ 2.4.4.4) розподіл величини χ^2 буде асиметричним. При цьому, чим менша вибірка, тим сильніше проявляється асиметрія. Із збільшенням чисельності вибіркової сукупності асиметрія зменшується і χ^2 - розподіл на нескінченості наближається до нормального. Цікавою особливістю цього розподілу є те, n - єдиний параметр цього розподілу. Наочно характер такої зміни ілюструє графік на рис.2.45. Крім того, середнє значення даних, що підпорядковуються χ^2-розподілу, дорівнює кількості його ступенів свободи, а дисперсія - подвоєній кількості ступенів свободи.

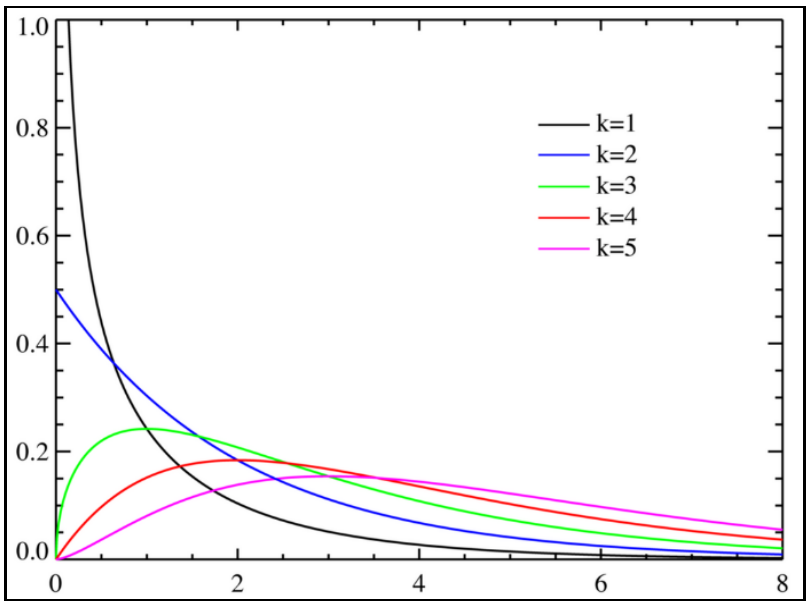


Рис.2.45 Функція щільності χ^2 - розподілу при різних значеннях параметру

χ^2-розподіл широко використовується в статистиці, науці про дані та машинному навчанні. Зокрема, він лежить в основі *χ^2-тестів* (або *χ^2-критеріїв* див. Розділ 3.3.5) для перевірки узгодженості між емпіричним і теоретичним розподілом, у χ^2-тестах незалежності для аналізу зв'язків між категоріальними змінними, а також у методах оцінювання дисперсії та перевірки гіпотез. У машинному навчанні χ^2-критерії застосовують для відбору ознак (англ.- *Feature Selection»*), коли потрібно оцінити силу зв'язку між змінними та цільовим класом. У сфері кібербезпеки цей розподіл допомагає виявляти статистично значущі відхилення в поведінці систем, аналізувати журнали доступу, розподіл кодів відповіді серверів чи структуру мережевого трафіку, що є ключовим для виявлення аномалій та потенційних атак.

2.4.4.3. *F*-розподіл

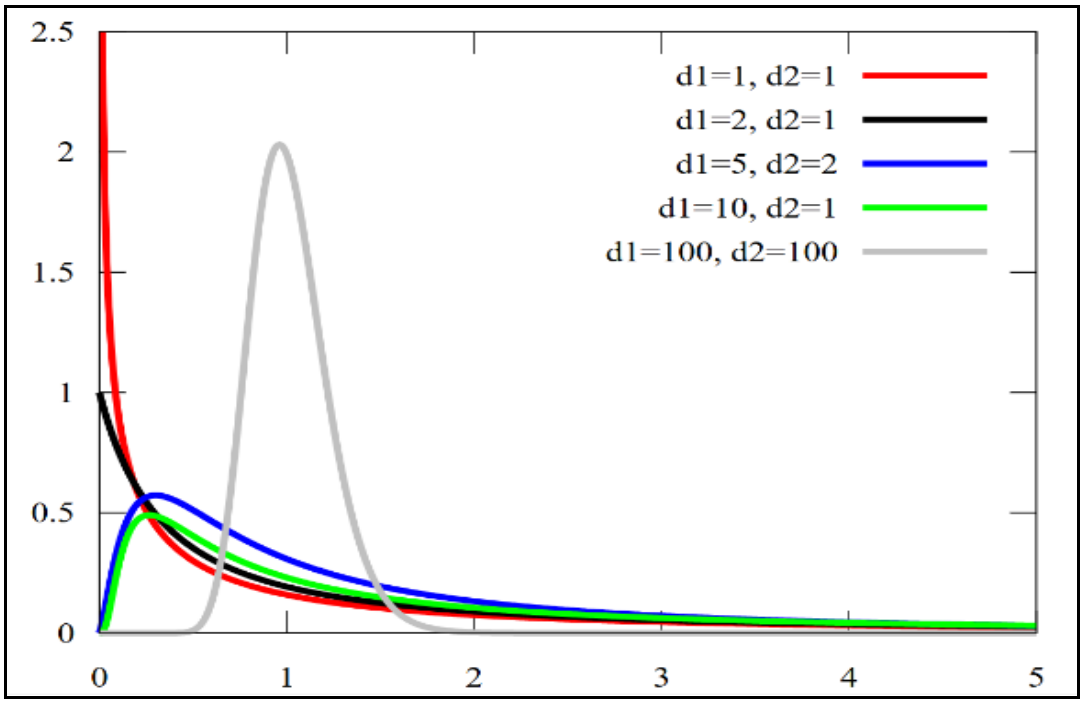
F-розподіл (або *розподіл Фішера* - названий на честь видатного англійського вченого, біолога, статистика, генетика, економіста, якого часто називають «батьком» чи «фундатором» сучасної статистики, Sir Ronald Aylmer Fisher, хоча сам розподіл був введений та вивчений іншим англійським дослідником Джорджем Снедекором - George Waddel Snedecor) -

двопараметрична родина неперервних розподілів, які виникають у статистиці під час порівняння двох дисперсій.

Нехай X_1 та X_2 - дві незалежні випадкові величини, що мають χ^2-розподіл (на практиці це означає, що вивчаються дисперсії деяких випадкових величин). Тоді розподіл випадкової величини $F = \frac{X_1/df_1}{X_2/df_2}$ підпорядковується **F-розподілу Фішера з ступенями свободи** (df_1, df_2)**.** Якщо дані підпорядковуються закону F-розподілу, то:

$$\mu = \frac{df_2}{df_2 - 2}$$

$$\sigma^2 = \frac{2df_2^2\left(df_2 + df_1 - 2\right)}{df_1\left(df_2 - 2\right)\left(d_4 - 4\right)}$$



2.46 Функція щільності розподілу Фішера при різних значеннях параметрів (ступенів свободи)

Цей закон, наприклад, у кібербезпеці може застосовуватися для аналізу стабільності поведінкових характеристик системи: зокрема, при порівнянні варіацій у часі відповіді сервера до і після підозрілої активності.

2.4.4.4. Ступені свободи

Декілька разів вище зустрічався термін «ступені свободи» про які треба сказати окремо. На рівні інтуїції, *ступені свободи* (англ. *Degrees of Dreedom, DF*) - це кількість незалежних величин у вибірці, які можна вільно змінювати при обчисленні певної статистики. Інакше кажучи, це число «вільних параметрів», що залишаються після врахування обмежень. Найпростіший приклад. Нехай треба створити множину з десяти чисел, сума яких повинна бути рівною довільному, але строго визначеному числу. В такому випадку можна вільно обрати дев'ять чисел. А десяте вже потрібно визначати через задане число та суму попередньо (вільно) обраних дев'яти чисел.

Формальне визначення. *Ступені свободи дорівнюють розміру вибірки мінус кількість накладених обмежень.* У статистиці це означає: при оцінюванні середнього ми втрачаємо 1 ступінь свободи (бо, як в наведеному прикладі, одне значення виявляється «прив'язане» до середнього); при оцінюванні кількох параметрів ми втрачаємо стільки ступенів свободи, скільки параметрів було оцінено. Так, для t-розподілу Стьюдента $df = n - 1$. Для χ^2-розподілу $df = n$. Для F-розподілу $df_1 = n_1;\ df_2 = n_2$.

Іншими словами, ступені свободи - це кількість незалежної інформації у вибірці, доступної для оцінки параметрів і проведення статистичних тестів. Ступені свободи показують, наскільки обмеженими чи надійними є статистичні висновки в залежності від кількості наданої інформації (розмірів наданих вибірок).

Ступені свободи також використовуються для визначення розподілів ймовірностей для перевірок різних статистичних гіпотез. Перевірки гіпотез часто використовують t-розподіл, F-розподіл та χ^2-розподіл для визначення статистичної значущості заданих статистик даних. За розширеним тлумаченням поняття «ступенів свободи» прикладами та особливостями його використанням радимо звернутися за посиланням [39].

2.4.5. Розподіли імовірностей та їх значення в аналізі реальних систем

Розподіли ймовірностей використовується в багатьох галузях математики інформатики при вирішені практичних завдань. У машинному навчанні вони допомагають робити прогнози та враховувати невизначеність. В обробці природної мови вони використовуються для моделювання частоти появи слів. У комп'ютерному зорі вони допомагають розуміти дані зображень та видаляти шум. При аналізі поведінки комунікаційних мереж розподіли, такі як пуассоновський, використовуються для аналізу потоків пакетів даних та для виявлення аномалій в цих процесах. Криптографія використовує випадкові числа на основі ймовірності спираючись на знання відповідних розподілів. Тестування програмного забезпечення та надійність також використовують розподіли для прогнозування помилок та збоїв. Приклади можна продовжувати нескінченно.

Розглянуті вище закони розподілу - лише невелика , хай навіть і найбільш широко вживана, частка від всіх законів, які використовувалися та використовуються в дослідженнях даних, що мають ймовірнісну природу. Так, в Вікіпедії, на сторінці [40] можна знайти згадку, короткий опис та посилання на більш ніж 200 законів розподілів. В спеціальний науковій літературі можна знайти опис тільки класичних і широко вживаних розподілів порядку 50–100 (нормальний, біноміальний, Пуассона, експоненційний, рівномірний, гамма, бета, логнормальний, Вейбулла, Парето, Коші, t-розподіл Стьюдента тощо). З урахуванням узагальнень, сімейств і спеціальних випадків - сотні. Разом із параметричними родинами та сучасними «сконструйованими» розподілами - тисячі. А у літературі з прикладної статистики та теорії ймовірностей постійно з’являються нові, за допомогою яких вдається описувати та вивчити нові та все більш складні випадкові процеси. Поглибити свої знання в цій галузі можна, наприклад, за посиланнями [41], [42].

А відтак постає питання - як маючи певну вибірку даних визначити, за яким саме законом розподілені дані? Від знаходження коректної відповіді на питання про те, який закон розподілу найбільш адекватно описує емпіричні дані, кардинально залежить якість та коректність результатів усього подальшого аналізу. Якщо для роботи з даними обрано «неправильний» (тобто такий, який погано відповідає реальному процесу) розподіл, статистичні оцінки параметрів, які будуть робитися на його основі, побудова довірчих інтервалів чи результати

перевірки різноманітних статистичних гіпотез можуть призвести до хибних висновків. У прикладному вимірі це означає, що аналітик ризикує не лише отримати спотворену картину процесу, але й прийняти неефективні або навіть небезпечні рішення. У сфері кібербезпеки, де аналіз часто спрямований на виявлення аномалій та загроз, помилка у виборі моделі розподілу може означати пропуск атаки чи, навпаки, надлишкову кількість хибних спрацювань. Таким чином, визначення найбільш адекватного розподілу для даних - це не формальність, а стратегічний крок, що забезпечує надійність усієї аналітичної системи. Підходи до пошуку відповіді на це питання буде надано в наступному розділі.

2.5. Оцінка щільності розподілу за даними вибірки

Багато поширених методів статистичного аналізу розроблялися виходячи з припущення про наявність апріорної інформації щодо закону розподілу імовірнісних даних - зокрема, про нормальність розподілу самих даних або про нормальність похибок у моделі. Подібні припущення є наріжним каменем таких класичних методів, як лінійний регресійний аналіз, t-критерій в перевірці гіпотез, кореляційний аналіз Пірсона та багато інших. Порушення цих припущень може суттєво знижувати надійність висновків, спотворювати оцінки параметрів і призводити до хибних результатів перевірки статистичних гіпотез.

Крім того, у багатьох прикладних задачах виникає необхідність знати розподіл окремих компонент моделі, що будується чи аналізується - залишків, прогнозованих значень, латентних змінних тощо. Зокрема, це важливо, наприклад, при моделюванні фінансових ризиків, де хвости розподілу визначають імовірність екстремальних подій; у задачах надійності технічних систем, де форма розподілу часу безвідмовної роботи безпосередньо впливає на прийняття інженерних рішень; у біомедичних дослідженнях, де розподіл клінічних показників може суттєво відрізнятися від нормального через неоднорідність популяцій тощо.

Однак на практиці істинний закон розподілу генеральної сукупності, як правило, є невідомим. Дослідник має у своєму розпорядженні лише скінченну вибірку спостережень, на основі якої необхідно відновити форму розподілу з достатнім ступенем точності. Зокрема, у кібербезпеці істинний закон розподілу випадкової величини значень параметру, щодо якого виконується моніторинг, майже ніколи не буває відомим. Натомість мається лише вибірка спостережень:

журнали подій, мережевий трафік, часові інтервали між атаками тощо. Так виникає фундаментальне завдання побудови, наскільки можливо універсального інструменту, здатного оцінювати (відновлювати) невідому функцію щільності імовірності безпосередньо з наявних даних вибірки - часто навіть без апріорних припущень щодо її параметричної форми. Вирішенню цього завдання присвячений даний розділ.

Нехай у нас є вибірка спостережень. Задача **оцінки щільності розподілу** (англ. *Density Estimation*) полягає в побудові за даними спостереженнями $\{x_1, x_2, ..., x_n\}$ оцінки $\hat{f}_n(x)$ невідомої функції щільності $f(x)$, що задовольняє наступним вимогам:

– оцінка сама є функцією щільності:

$$\hat{f}_n(x) \geq 0, \forall x \in \square, \int_{-\infty}^{+\infty} \hat{f}_n(x) dx = 1$$

– узгодженість оцінки та функції, яка оцінюється:

$$\hat{f}_n(x) \to f(x), n \to \infty$$

Існує два головних класи методів оцінки функції щільності - параметричні та непараметричні (див. розділ 3.2.6). Параметричні методи передбачають, що функція розподілу $f(x)$ належить до заздалегідь відомого параметричного сімейства розподілів (нормального, рівномірного, експоненційного тощо). Переваги такого підходу полягають в простоті інтерпретації результатів та обчислювальній ефективність самої процедури - але виключно за умови правильного припущення щодо сімейства розподілів. При порушенні цієї умови отримані результати катастрофічно втрачають адекватність. Параметричні методи оцінки в свою чергу розрізняються за критеріями, які вони використовують (метод максимальної правдоподібності, метод моментів, байєсівське оцінювання тощо); за метриками, які застосовуються; а також за узгодженістю з даними за формальними статистичними критеріями - χ^2 Пірсона, Колмогорова–Смирнова, Андерсона–Дарлінга тощо (див. розділи 3.3.5, 3.3.6, 3.4).

Непараметричні методи здатні працювати без попередніх припущень щодо аналітичної форми $f(x)$. Це робить задачу принципово складнішою, однак значно ширшою за сферою застосування. Серед непараметричних методів оцінки функції розподілу найбільш відомі методи, що основані на аналізі гістограм, та методи так званої *ядерної оцінки щільності*.

Використання гістограм як дискретного аналога функції щільності розподілу розглядалося у розділі 2.2.3. Нагадаємо, гістограма - це проста форма представлення даних, у якій визначаються інтервали та підраховується кількість точок даних у кожному інтервалі.

Підкреслимо, значення по осі Y на графіку функції щільності розподілу - це не ймовірність конкретної точки як така (вона для неперервної величини дорівнює 0), а саме *щільність ймовірності*. Безперервна пряма містить незліченно нескінченну кількість точок. Якщо кожна точка мала б ненульову імовірність, то сума імовірностей усіх точок була б нескінченною, що суперечить аксіомі повної ймовірності простору елементарних подій, яка вимагає, щоб така сума дорівнювала 1. Але щільність не імовірність, а *питома імовірність на одиницю виміру*. Тобто, *f(x)* показує, з якою інтенсивністю концентрується імовірність у околі точки *x*, але не саму імовірність у цій точці. Тому і виникає ситуація, яка часто ускладнює розуміння новачка - значення функції *f(x)* в точці може бути більшим за 1, при тому, що площа під графіком функції щільності ймовірності завжди дорівнює 1. Переходячи до скінченого набору точок та гістограм вибірки, виходить, що гістограми показує не самі ймовірності, а їх розподіл по інтервалах, і ймовірність отримується лише як *площа під стовпчиками (тобто, з врахуванням ширину інтервалів)*, а не їх висота сама по собі.

Основна проблема з гістограмами полягає в тому, що вибір довжини інтервалів може вплинути на кінцевий результат. Гістограма здатна забезпечити високу якість *оцінок* щільності ймовірності, але лише у разі вибірок великого об’єму. Як уже згадувалося раніше, розділити коротку послідовність на велику кількість інтервалі неможливо, а гістограма, що складається лише з 2-3 стовпців, не може допомогти скласти уявлення про закон розподілу щільності ймовірності такої послідовності.

Гістограми мають і інші недоліки, зокрема - розривність (тобто, оцінка є ступінчастою функцією незалежно від гладкості істинної щільності); залежність

від точки початку відліку інтервалів (зміщення меж між інтервалами змінює форму гістограми, часто досить суттєво).

Згладжування гістограми - це сукупність методів, що усувають ці недоліки шляхом отримання *безперервної та диференційованої оцінки щільності*. Найпростішим різновидом згладжування гістограм може бути ***полігонне згладжування частот*** (англ. *Frequency Polygon*) метод лінійної інтерполяції через центри інтервалів (див. рис.2.47):

$$\hat{f}_{FP} = \hat{f}_j + \frac{x - m_j}{m_{j+1} - m_j}\left(\hat{f}_{j+1} - \hat{f}_j\right), \quad x \in \left[m_j, m_{j+1}\right)$$

де m_j - значення, що відповідає середині j-ого інтервалу;

$\hat{f}_j$ - висота стовпчика гістограми для j-ого інтервалу.

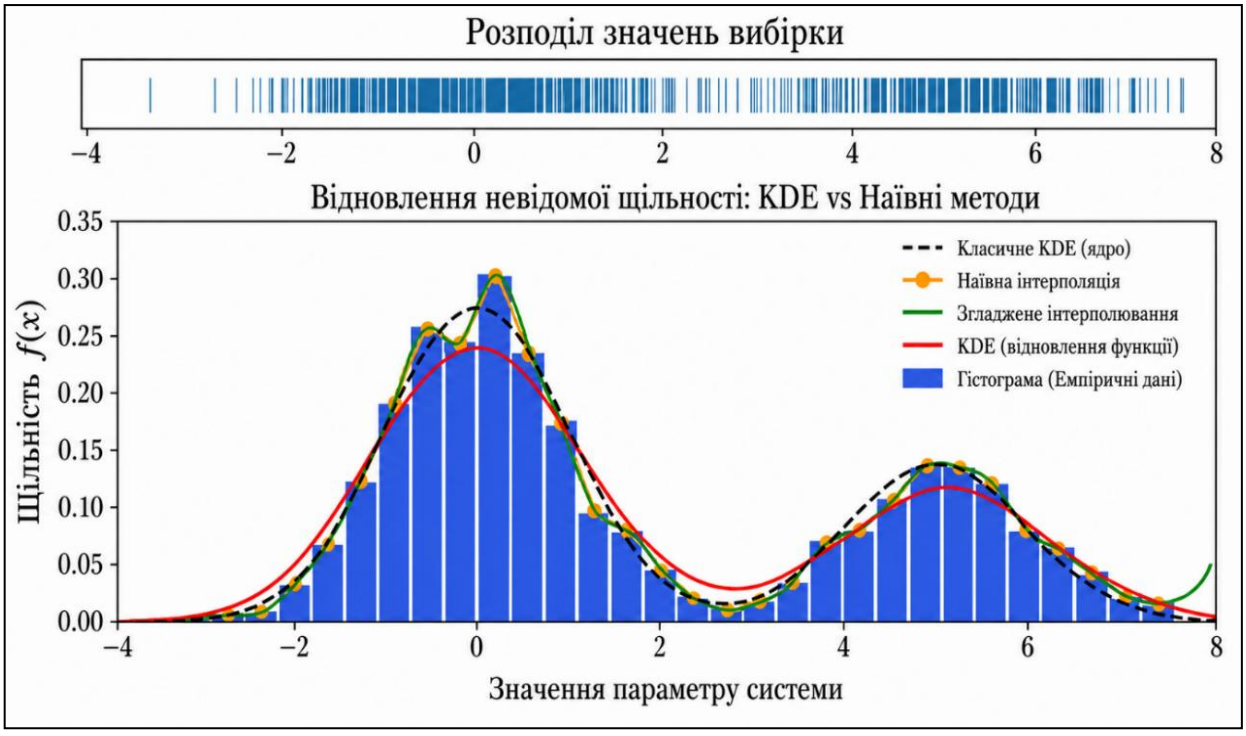


Рис.2.47 Оцінка функції щільності генеральної сукупності за даними вибірки

Такий метод забезпечує неперервність, але не диференційованість функції-оцінки $\hat{f}_n(x)$, досить просту реалізацію та інтерпретацію результату. Проте він схильний до ефекту перенавчання, в тому числі на хвостах розподілу, коли оцінка

відтворює випадкові коливання значень гістограм, та приховує суттєві шаблони поведінки об’єкту спостереження.

Згладжування гістограми сплайн-функціями пропонує компроміс між простотою гістограм та гнучкістю ядерного оцінювання (*KDE*) (див. далі) забезпечуючи більш гладку в порівнянні з лінійною інтерполяцією, неперервну оцінку щільності, обчислювальну ефективність та інтерпретованість за рахунок проходження через контрольні точки (вузли) гістограми. Проте і цьому методу притаманні слабкі сторони, зокрема осциляції сплайна між вузлами, зниження адекватності (точності) на хвостах вибірки (де можуть бути сконцентрована суттєва інформація про аномалії – або їх відсутності - вибірки), відчутна залежність від суб’єктивного вибору параметра згладжування.

Проте найбільш перспективним на сьогоднішній день вважається метод ядерної оцінки щільності ймовірності

Ядерне оцінювання щільності (англ. *Kernel Density Estimation, KDE*) є непараметричним методом оцінювання функції щільності ймовірності випадкової величини. Основна ідея ядерного згладжування є досить простою. Його легко уявити у вигляді ковзного вікна, всередині якого відбувається усереднення зважених значень послідовності. При цьому як правило вживається зважене усереднення, а коефіцієнти згладжування задають «форму» ядра, яка може бути прямокутною, експоненційною або іншою:

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

де $K(.)$ - функція ядра;

n - кількість елементів у вибірці;

h - ширина ковзного вікна, фактично - кількість значень вибірки, які беруться до уваги при визначення значення $\hat{f}_h(x)$ в кожній точці вибірки. Див. Рис.2.48.

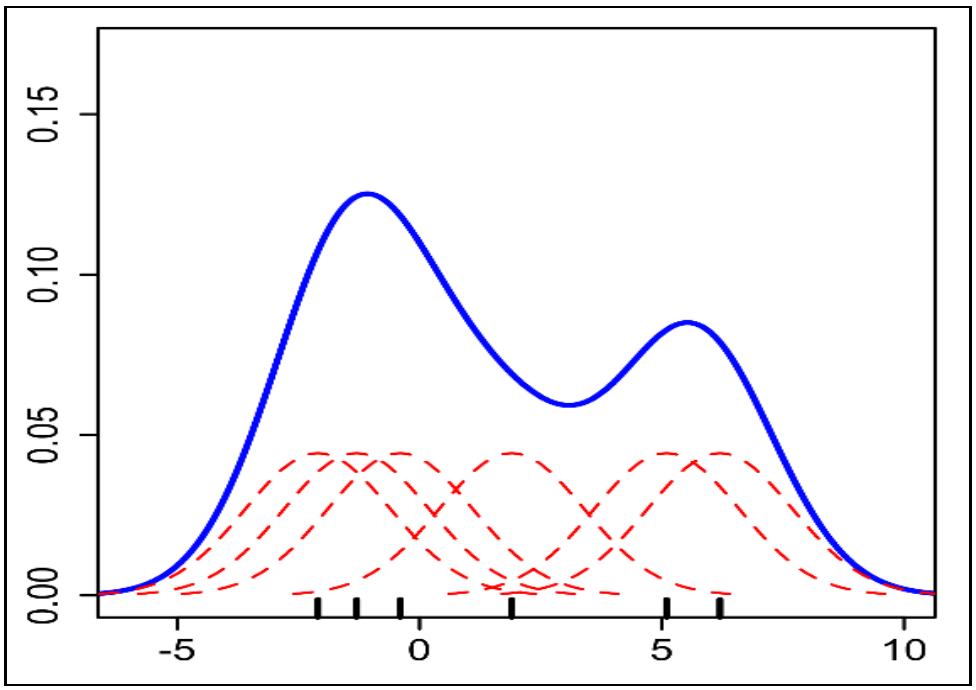


Рис. 2.48 Ядерна оцінка щільності (Приклад)

Функція ядра - невід'ємна ($K(\zeta) \geq 0, \forall \zeta$), зазвичай симетрична ($K(\zeta) = K(-\zeta)$) функція, що визначає форму кривої, яка використовується для генерації PDF. На відміну від гістограми, яка підраховує значення вибірки, які потрапили до кожного з інтервалів, при використанні ядерного оцінювання щільності *значення оцінки визначаються для кожної точки вибірки окремо*. Для цього для кожної точки підсумовують зважені значення для найближчих точок вибірки, тобто таких, які входять в вікно з центром в точці x та віддалених від неї не далі, аніж на $h/2$. Отримане значення приймають за значення оцінки $\hat{f}_h(x)$.

Окрім властивостей невід'ємності ($K(\zeta) \geq 0, \forall \zeta$), функція ядра має відповідати умовам

– нормованності:

$$\int_{-\infty}^{+\infty} K(\zeta) d\zeta = 1$$

– скінченності:

$$\int |K(\zeta)| d\zeta < \infty$$

– скінченності дисперсії:

$$\int \zeta^2 * K(\zeta) d\zeta < \infty$$

а також гладкості (наприклад, неперервність або диференційованість) - для отримання гладкої оцінки щільності, компактності (ядро дорівнює нулю поза певним інтервалом) - для обчислювальної ефективності, максимум у нулі - найбільший внесок в обчислення значення оцінки у точці мають найближчі до неї точки вибірки. Тобто, функція ядра в *KDE* - це спеціально підібрана функція, яка поводиться як щільність ймовірності і забезпечує коректне, гладке та статистично обґрунтоване наближення істинного розподілу даних. На рис.2.49 показані графіки, що дають візуальне уявлення про найбільш широко вживані функцій ядра, які використовуються в методі *KDE*.

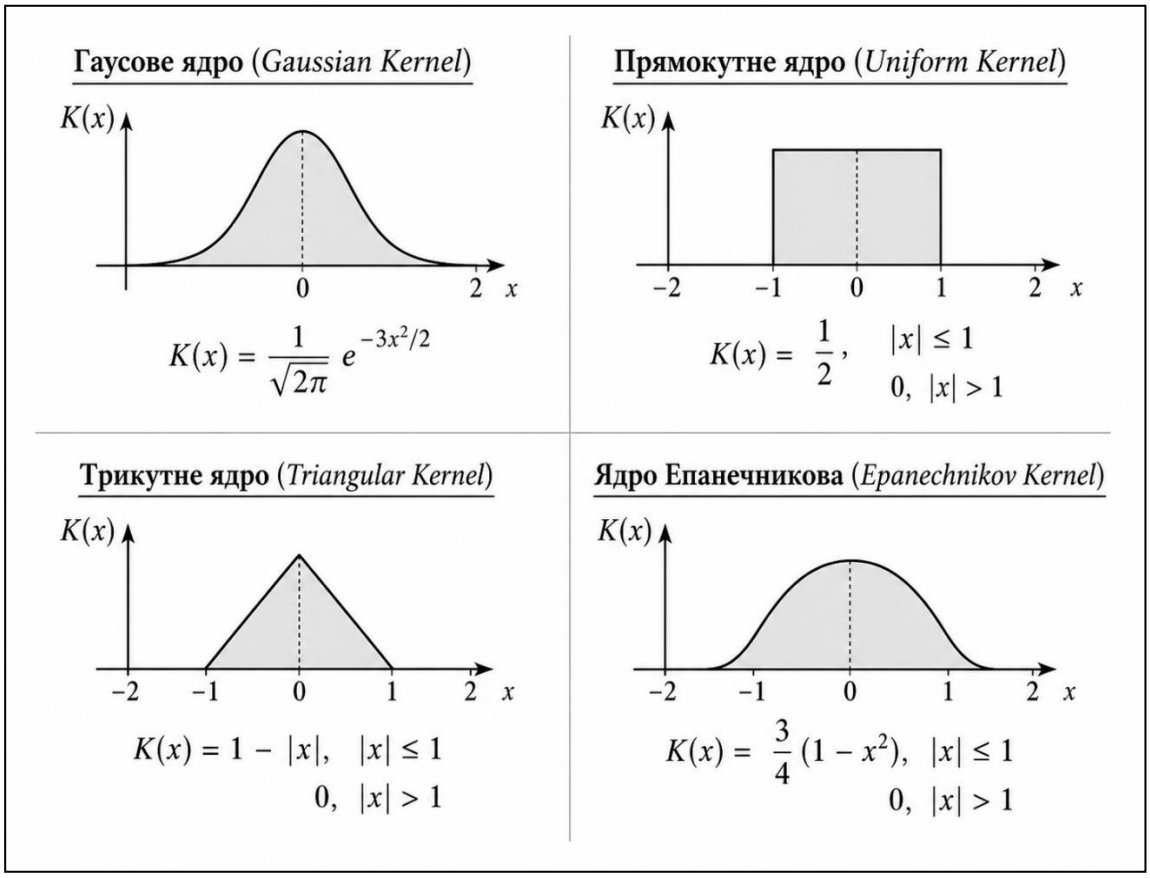


Рис.2.49 Деякі функції ядер, що використовуються при оцінці щільності методом ядерної оцінки

На практиці, позатим що вибір ядра має дуже значний вплив на якість оцінки - вибір значення h є *принципово важливішим*. Занадто мале значення h

призводить до появи множини окремих піків функції, що відповідають кожному елементу вибірки, створюючи зайвий «шум». В той час як великі значення – до надмірного згладжування функції оцінки. Занадто велике значення h взагалі перетворює будь-який розподіл на унімодальний (одногорбий), приховуючи важливі властивості даних, такі як мультимодальність (див. Розділ 2.3.1). Все це може мати негативні наслідки в практиці застосування методу *KDE*. Наприклад, «перегладжування» може «сховати» невеликий сплеск трафіку, який є ознакою стелс-сканування мережі, тоді як «недогладжування» призведе до помилкової ідентифікації звичайної варіативності як атаки (про помилки I та II типу – див.3.2.4).

На практиці, вибір параметру h виконується за допомогою емпіричних правил, серед яких:

– правило Скотта: $h = \left(\frac{4\hat{\sigma}^2}{3n} \right)^{\frac{1}{5}} \approx 1.06\hat{\sigma} n^{-\frac{1}{5}}$, що є оптимальним при використанні гаусового ядра;

– правило Сілвермана: $h = 0.9 * \min(\hat{\sigma}^2, \frac{IQR}{1.35}) * n^{-\frac{1}{5}}$;

де IQR - міжквартильний розмах (див. Розділ 2.3.2);

$\hat{\sigma}$ - оцінка середньоквадратичного відхилення вибірки.

У багатьох випадках, наприклад - при відхиленні форми оцінюваної щільності від нормальної, врахування міжквартильного розмаху за правилом Сілвермана дозволяє скоригувати у меншу сторону значення величини h, що робить запропоновану оцінку досить універсальною. Тому дане емпіричне правило має сенс використовувати як початкову базову оцінку значення h, тим більше що при цьому не доводиться проводити громіздких обчислень.

В більш складних випадках використовують і інші методи, зокрема з використанням методів крос-валідації або попереднього аналізу особливостей вибірки. Також за необхідності може використовуватися *адаптивне ядерне згладжування*, тобто таке згладжування, при якому значення h є функцією від значень вибірки:

$$\hat{f}_{ad}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h(x)} K\left(\frac{x - x_i}{h(x)} \right)$$

У екосистемі *Python* існує кілька поширених реалізацій до побудови ядерної оцінки щільності, які відрізняються рівнем абстракції, гнучкістю та сферою застосування.

Бібліотека *SciPy,* функція *scipy.stats.gaussian_kde()*. Це один із базових і найпростіших інструментів для *KDE*. Реалізує гаусове ядро та автоматичний вибір параметра згладжування (за правилом Сільвермана або Скотта). Добре підходить для швидкого статистичного аналізу, але має обмежену гнучкість (важко змінювати ядро чи використовувати складні моделі).

Бібліотека *Scikit-learn,* функція *sklearn.neighbors.KernelDensity()* реалізує більш універсальний і гнучкий підхід. Підтримує різні типи ядер (gaussian, tophat, epanechnikov тощо), дозволяє явно керувати параметром h і використовує високоефективні обчислювальні алгоритми. Часто застосовується у задачах машинного навчання, зокрема для оцінки правдоподібності та виявлення аномалій.

Бібліотека, функція *seaborn.kdeplot*. Високорівнева бібліотека *Seaborn* для візуалізації, яка будує *KDE* "під капотом" (зазвичай використовуючи при цьому засоби *SciPy*). Основна перевага - простота та гарна графічна подача результатів. Підходить для розвідувального аналізу даних, але не дає повного контролю над алгоритмом.

Бібліотека *Statsmodels*, функції *statsmodels.nonparametric.kde. KDEUnivariate()*, та *statsmodels.nonparametric.kde.KDEMultivariate().* Бібліотека, орієнтована на статистичний аналіз. Дає ширші можливості для роботи з *KDE*, включаючи різні ядра, методи вибору h і підтримку багатовимірних даних. Часто використовується у наукових дослідженнях.

Після побудови ядерної оцінки щільності основним завданням стає її практичне використання в подальшому аналізі. Отриману оцінку застосовують для обчислення ймовірностей подій, виявлення аномалій (області з низькою щільністю), побудови порогових критеріїв прийняття рішень та формування ознак для моделей машинного навчання. Важливим етапом є також перевірка адекватності оцінки - шляхом візуального аналізу, порівняння з альтернативними методами або використання процедур валідації. У практичних задачах, зокрема в кібербезпеці, *KDE* часто інтегрується *IDS* системи, де вона виступає як

компонент оцінки нормальної поведінки або розподілу даних, на основі якого приймаються подальші рішення.

Більш поглиблений опис методів оцінки щільності, зокрема – приклади програмної реалізації та застосування цих методів, можна знайти в роботах [43], [44], [45], [46].

Контрольні запитання.

1. Для випадкової величини знайти ймовірність того, що випадкове значення, отримане з цієї генеральної сукупності, буде лежати в діапазоні від 12 до 18.
2. У чому головна відмінність між t-розподілом Стьюдента та нормальним розподілом, і чому t-розподіл застосовують для невеликих вибірок?
3. Як визначається число ступенів свободи у:
   - t-розподілі для вибірки з n елементів?
   - χ^2-розподілі, якщо сума квадратів складається з k незалежних нормальних величин?
4. Маємо 5 чисел, сума яких зафіксована і дорівнює 100. Скільки ступенів свободи залишиться для вільного вибору цих чисел?
5. У чому полягає основна відмінність між гістограмою та ядерною оцінкою щільності (*KDE*).
6. Яку роль відіграє ширина ковзного вікна при використанні методу *KDE*?
7. Як можна використати оцінку функції щільності для виявлення аномалій у даних?

3. ПЕРЕВІРКА СТАТИСТИЧНИХ ГІПОТЕЗ. ПРИНЦИПИ, МЕТОДИ ТА АЛГОРИТМИ

3.1. Основна проблема теорії перевірки гіпотез. Методика її вирішення

Перевірка гіпотез, мабуть, є найзаплутанішою темою, з якою початківці зіткаються при вивченні статистичних основ машинного навчання. Але це також одна з найважливіших сходинок, по якій потрібно піднятися на шляху до глибокого розуміння самого машинного навчання. За темою «Перевірка статистичних гіпотез» можна натрапити на такі визначення, як: «…метод статистичного висновку з використанням даних наукового дослідження» (Вікіпедія). Або «перевірка гіпотез - це використання статистики для визначення ймовірності того, що дана гіпотеза є істинною». (Вольфрам). «Перевірка гіпотез - це правило для вибору між двома конкуруючими гіпотезами» [47] . Все це, звісно правильно, але от навряд чи читач може зрозуміти, про що йдеться в цих визначеннях. А йдеться про одну з фундаментальних концепцій статистики.

В найзагальнішому розумінні перевірка гіпотез - це в першу чергу *формалізований процес прийняття рішень на основі даних щодо властивостей цих самих даних, в разі коли йдеться про ймовірності дані*. Це статистичний інструмент, який дозволяє робити обґрунтовані висновки про певні аспекти реальності, маючи обмежений набір спостережень. В рамках парадигми перевірки гіпотез *дослідник формулює припущення (гіпотези) щодо певних параметрів або характеристик даних, а потім використовує статистичні методи, щоб визначити,* ***чи достатньо доказів у наших даних, щоб відкинути одне з цих припущень на користь іншого.***

У *комп'ютерних науках* фахівці дуже часто стикаються з задачами, для обгунтованого рішення яких необхідно застосувати саме методи перевірки гіпотез. Наприклад - потрібно порівняти ефективність різних алгоритмів (чи дійсно новий алгоритм сортування швидший за попередній?); провести A/B тестування нових функцій програмного забезпечення (чи покращує нова кнопка в екранному інтерфейсі конверсію?); оцінити продуктивність систем (чи вплинула зміна конфігурації на час відгуку?) тощо. Перевірка гіпотез надає чіткий, формалізований фреймворк для виконання таких порівнянь. *У кіберзахисті* це є основою в тому числі і для *автоматизації* багатьох критично важливих завдань. Наприклад, у системах виявлення вторгнень перевіряється

гіпотеза: «Чи поточна мережева активність, за якою можна слідкувати, аналізуючи поточні значення її параметрів, є нормальною, чи вона вказує на атаку?». При аналізі шкідливого програмного забезпечення можливо перевіряти гіпотезу: «Чи демонструє ця програма поведінку, характерну для відомого типу шпигунського ПЗ?». Фільтрація спаму, аналіз аномалій у логах, оцінка ефективності нових методів шифрування – все це сфери, де перевірка гіпотез відіграє ключову роль.

Отже, **статистичної гіпотезою** *називають наукове припущення про ті чи інші параметри статистичних законів розподілу випадкових величин, які можуть бути перевірені на основі аналізу даних наданої вибірки.*

Відомо, що отримати для дослідження дані загальної (генеральної) сукупності здебільшого виявляється неможливим. У такому випадку висновки роблять на основі аналізу даних вибірки, зокрема визначаються деякі оцінки (вибіркові статистики) параметрів загальної сукупності (див. Розділ 2.3.3). Ці припущення або твердження є гіпотезами, а ***перевірка гіпотези - це процес перевірки існування доказів для відхилення цієї гіпотези***.

Прикладом гіпотези можуть бути, наприклад, така: «математичне очікування (або будь-який інший параметр) нашої генеральної сукупності дорівнює μ». Якщо в ході вимірів отримано певне значення середнього $\overline{X}$, то ця гіпотеза стає еквівалентною питанню: «*з якою ймовірністю можливо* отримати саме таке значення $\overline{X}$, якщо б наша генеральна сукупність підпорядковувалася заданому закону розподілу з відомим математичним очікуванням μ»? Схожий приклад, який теж вже згадувався раніше, виникає при спробі дати відповідь на питання «чи рівні між собою деякі параметри двох генеральних сукупностей даних якщо для дослідження надані лише вибірки з цих генеральних сукупностей»? Наприклад - ті самі математичні очікування, які – як відомо - не можуть бути безпосередньо виміряні через недоступність повної інформації про всі значення генеральної сукупності.

В основі перевірки статистичних гіпотез лежить ідея, відома ще як мінімум з часів Сократа - *доказ від протилежного*. Спочатку допускають, що гіпотеза, яка була сформульована, вірна. Потім аналізують, а що відбувається в дійсності? Чи не вдасться знайти події або дані, які суперечать нашій гіпотезі? Якщо такі артефакти знаходяться, то гіпотезу треба відхилити як «помилкову», а якщо-ж ні

– то відхилити гіпотезу не можна. Зверніть увагу – *не довести гіпотезу як «вірну», а* ***НЕ відхилити*** *її.*

Невеликий філософський відступ: Може здатися, що дослідники змушені робити одну з класичних логічних помилок, яка навіть має старовинну латинську назву «ad ignorantiam». Це аргументація істинності деякого твердження, що ґрунтується на відсутності доказу його хибності. Наприклад, «снігова людина існує, оскільки ніхто не довів протилежного». Виявлення різниці між науковою гіпотезою та подібними хитрощами становить предмет цілої галузі філософії: *методології наукового пізнання*. Один із її яскравих результатів - критерій фальсифікованості, висунутий філософом Карлом Поппером ще у першій половині XX століття. Критерій покликаний відокремлювати наукове знання від ненаукового і на перший погляд здається парадоксальним: *Теорія чи гіпотеза може вважатися науковою, тільки якщо існує, нехай навіть гіпотетично, спосіб її спростувати.*

В статистичному аналізі це виглядає так: висувається припущення, яке називається ***нульовою гіпотезою*** H_0 (наприклад, що математичне очікування даних генеральної сукупності дорівнює М). Потім намагаються на підставі наявних спостережень (за даними вибірки) цю гіпотезу підтвердити або спростувати. *Ухвалення нульової гіпотези означає, що дані спостережень* ***не суперечать*** *припущенню* (в прикладах, що наводилися вище - відсутність суперечностей між *емпіричними статистиками* та *теоретичними параметрами* генеральної сукупності, або між параметрами функцій розподілів різних генеральних сукупностей тощо), *але* ***не доводять*** *справедливість самого припущення. Відкидання гіпотези* означає, що емпіричні дані суперечать висунутій нульовій гіпотезі та вірна інша, ***альтернативна гіпотеза***, що зазвичай позначається як H_a.

Зауваження: Замість виразу «гіпотеза не відхиляється» доволі часто говорять «гіпотеза приймається». В цілому, цей вислів також прийнятний, якщо його розуміти правильно, тобто якщо вважати, що *приймається саме гіпотеза* (одне з можливих пояснень), *а не конкретне твердження*.

Отже, підсумовуючи. *Нульова гіпотеза не вимагає доказів.* Саме так вона має бути сформульована. Альтернативна гіпотеза заперечує нульову та є припущенням, справедливість якого потрібно довести. Процедура перевірки гіпотези дає один з двох можливих результатів, а саме – прийняття або

відкидання нульової гіпотези. Якщо нульова гіпотеза відкидається, то альтернативна гіпотеза вважається істинною. Якщо нульова гіпотеза не відкидається, то альтернативна вважається недоведеною. *Але недоведеність альтернативної гіпотези не означає, що нульова гіпотеза правильна*.

3.2. Статистичний підхід до задач перевірки гіпотез

3.2.1. Рівень значущості як інструмент перевірки гіпотез

Розглянемо приклад, що допоможе краще зрозуміти логічні ланцюжки, які лежать в основі процедури перевірки гіпотез.

Уявіть, що ви граєте в гру «орел та решка». Ви поставили на орел, ваш візаві - на решку. Випадає решка. Імовірність такої події - 0.5. Другий раз випадає решка - ймовірність двох поспіль решок - 0.25. Трьох - 0.125, ... шістьох - 0.015625, сімох - 0.007813. Теоретично це можливо, проте досить малоймовірно. Фальшива монета? Якщо так - то коли саме ви можете починати вважати, що послідовність стала малоймовірною? Або після скількох кидків і випадіння решки поспіль можна вважати монету фальшивою? Щоб відповісти на це питання більш-менш обґрунтовано треба заздалегідь домовитися, якусь ж ймовірність вважати такою, з якою ще можна «змиритися». Або навпаки – скільки доказів (випадків) треба надати, у конкретній вибірці, перш ніж буде відкинута нульова гіпотеза? Іншими словами – треба встановити поріг, який визначає, наскільки переконливими мають бути дані, щоб вважати результат статистично значущим. Таке узгоджене значення називають ***рівнем значущості*** (англ. *Significance Level*, інший термін - ***довірчий рівень***) α.

В наведеному прикладі була висунута гіпотеза H_0 - «монета нормальна» та альтернативна до неї гіпотеза H_a - «монета фальшива». Потім був проведений відповідний експеримент з послідовного кидання монети. Далі результати цього експерименту порівняли з обраним значенням α, який можна розглядати як деякий «кордон», який розділяє результати за умови, якщо справедливою являється та чи інша гіпотеза. І вже на основі цього порівняння прийняли відповідні рішення (про фальшивість монети).

Аналізуючи все, що було сказано вище, можна дійти висновку, що ніколи неможливо бути впевнені в справедливості гіпотези на 100%. (Адже теоретично можливо і 100, і 1000 послідовних випадінь «орла» в згаданій грі). Тобто ступінь впевненості (або протиріччя) визначається обраним *рівнем значущості*, і логіка

визначення цього рівня підказує, що він в свою чергу залежить від того, як далеко значення вибіркового параметру відхиляється від значення, яке очікується теоретично, *тобто від того, яким це значення мало-б бути* ***за умови справедливості нульової гіпотези***. Вважається, що якщо нульова гіпотеза справедлива, відхилення можливо будь яке, але ***ймовірність відхилення зменшується з зростанням цього відхилення***. Якщо ймовірність досить мала (*менше обраного рівня значущості*), то наявність протиріччя висунутій нульовій гіпотезі приймається доведеним (*не забуваючи про можливу помилку*), а отже, сама гіпотеза H_0 -відкинута.

Можливо задати питання, а чому в цьому ланцюжку логічних висновків переходять від «відхилення значення параметру» до «ймовірності вказаного значення відхилення»? Щоб пояснити це, розглянемо ще один простий приклад. Нехай перевіряється гіпотеза про те, що жінки витрачають більше часу на розмови по телефону, ніж чоловіки. Зрозуміло, ми не можемо провести дослідження серед усіх жінок та чоловіків на планеті Земля та маємо обмежитися певною вибіркою. Припустимо, що в дослідженні брали участь 35 чоловіків і 32 жінки. Середній час розмови склав 28 хв. в день у чоловіків і 31 хв. в день у жінок. На перший погляд (31хв.>28хв.) відмінності виявлені, і ці результати підтверджують гіпотезу. Однак такий результат може бути отриманий і випадково, навіть якщо в генеральної сукупності відмінностей немає - за рахунок випадковості при формуванні групи тих, хто приймав участь в дослідженні, часу проведення експерименту, пори року, трансляції футболу по телебаченню під час проведення самого дослідження або безлічі якихось інших випадкових факторів. Однак такий саме результат може бути отриманий і тоді, коли такі відмінності насправді, об'єктивно існують. Тому закономірним стає запитання: чи достатньо отриманого значення величини відмінності в середніх значеннях для того, щоб стверджувати, що взагалі всі жінки світу в середньому спілкуються телефоном довше, ніж всі чоловіки? Яка *ймовірність, що це так*? В статистиці це питання формулюється так: «*Чи є різниця отриманих в досліді і очікуваних за висунутою гіпотезою значень параметрів генеральних сукупностей статистично значущою*»?

3.2.2. Рівень значущості α та параметр *p_value*

У процесі перевірки статистичних гіпотез важливу роль відіграють два тісно пов'язані, але різні поняття - рівень значущості α (який був введений у

попередньому розділі) та рівень спостережної значущості, який в професійній літературі зазвичай називають ***«p_value»*** (або, значно рідше, *p-значенням*). *Їх правильне розуміння є ключовим для інтерпретації результатів статистичного тестування.*

Рівень значущості α - це *заздалегідь встановлена межа ймовірності допустимої помилки* (уточнення - мова йде про помилку першого роду, див. розділ 3.2.4), тобто *це ризик відхилення нульової гіпотези в разі, якщо в дійсності вона істинна*. Іншими словами, розуміючи, що при відхиленні нульової гіпотезу у нас завжди лишається деякий шанс помилитися (пам'ятаєте, що «орел» «чесної» монети теоретично може випасти і 100 разів підрід?), рівень значущості можна трактувати як ризик, на який дослідник погоджується наперед, приймаючи відповідне рішення. Так, говорячи що значення α приймається на рівні 0.05, мають на увазі що ми готові прийняти не більше 5% ***ризику помилково відхилити істинну*** H_0.

Конкретне значення α вибирається з *позастатистичних міркувань*. Здебільшого - на основі традицій дослідження, що склалися в тій чи іншій предметній галузі. Зазвичай використовують деякі стандартні значення: 0.05; 0.01; 0.005; 0.001 тощо. Так наприклад, в психології, соціології, в економіці часто приймають значення 0.1. В технічних, інженерних дослідженнях – 0.05. В медицині, наприклад, раніше теж вживався такий показник, але років 20 назад результати дослідження стали сприйматися професійною медичною спільнотою тільки при використанні значення 0.04. А останні роки в галузі йде дискусія про можливе подальше зниження цього рівня. В галузях, де важлива висока точність прийняття рішень та надзвичайно висока «ціна» помилки першого роду - фінансові, технічні, екологічні або для безпеки людей – приймають значення 0.01. В будь якому разі треба пам'ятати, що це значення обирається *обов'язково наперед, до початку дослідження, і по-суті - довільно*.

Якщо значення α задане, то питання прийняття чи відхилення гіпотези зводиться до визначення значень параметру X (див. рис.3.1) , яке відсікає такі «хвости» розподілу цієї величини, ймовірність потрапляння в які менші за α. Треба звернути увагу, що якщо значення X не обмежені ані зверху, ані знизу, говорять про *двосторонні гіпотези*. Для них задане значення α треба рахувати для обох хвостів розподілу. Тобто площина кожної з двох зон (хвостів) має бути

$\frac{\alpha}{2}$ (що для двох хвостів розподілу разом дає, зрозуміло, значення α). Однак бувають випадки, коли точно відомо, що значення X може бути тільки - наприклад - більше за певну величину. В такому випадку можливість отримати значення, що лежить у лівому хвості вибірки відсутня. А точка критичного значення має відсікати площину α тільки в правому хвості розподілу значень. Аналогічна, але симетрична, ситуація виникає і якщо величина X обмежена зверху. Тоді вся площина α відхилення гіпотези відсікається зліва. В зазначених випадках говорять про *односторонні гіпотези*.

Отже, задача перевірки гіпотези починається з визначення таких ***критичних значень*** $X_{лів.}, X_{прав.}$(англ. *Critical Value*) - або одного значення в випадку односторонніх критеріїв, які відсікають задану величину $\frac{\alpha}{2}$ (або α - для одностороннього критерію) площини під функцією щільності розподілу. Ще раз підкреслимо, що *ці значення залежать від закону розподілу та є функцією від заданого (обраного) значення* α. Після того, як значення $X_{лів.}, X_{прав.}$ знайдені, та після того, як в результаті проведеного експериментального дослідження (вимірювання) отримані експериментальні (емпіричні) значення параметру X, *задача перевірки гіпотези зводиться до визначення того, в який саме інтервал* - відхилення або прийняття H_0 гіпотези - *це значення X потрапляє*.

Викладена вище методика на сьогоднішній день досить часто замінюється більш сучасною, що здобула поширення завдяки широкому впровадженню комп'ютерної техніки в практику дослідження. Справа в тому, що, порівнюючи отримані значення з обраним на основі рівня значущості критичним значенням (значеннями - для двостороннього критерію), ми не бачимо «силу доказу». Адже значення X може потрапити в область, відповідну 5% рівню значущості, а може і в область 1% значущості (тобто відхилитися ще далі). В обох випадках H_0 гіпотеза відхиляється, але впевненість, з якою це робиться, зовсім не однакова. Одна справа - «швидше за все» (як при 5% -му рівні), а інше - «напевно» (як при 1% рівні). Тому сучасний підхід до перевірки гіпотези в більшості випадків базується на реально отриманому рівні *спостережної значущості*, який називають ***p_value***.

p_value є імовірністю отримати такі або ще більші (або менші, в разі лівосторонньої гіпотези) *відхилення значення статистики за умови істинності*

гіпотези H_0. А отримані при цьому відхилення були зумовлені випадковими чинниками. Іншими словами, ***p_value*** **показує, яка ймовірність помилитися та відхилити істину гіпотезу H_0 на основі отриманих значень статистики** в той час, як ця гіпотеза є вірною.

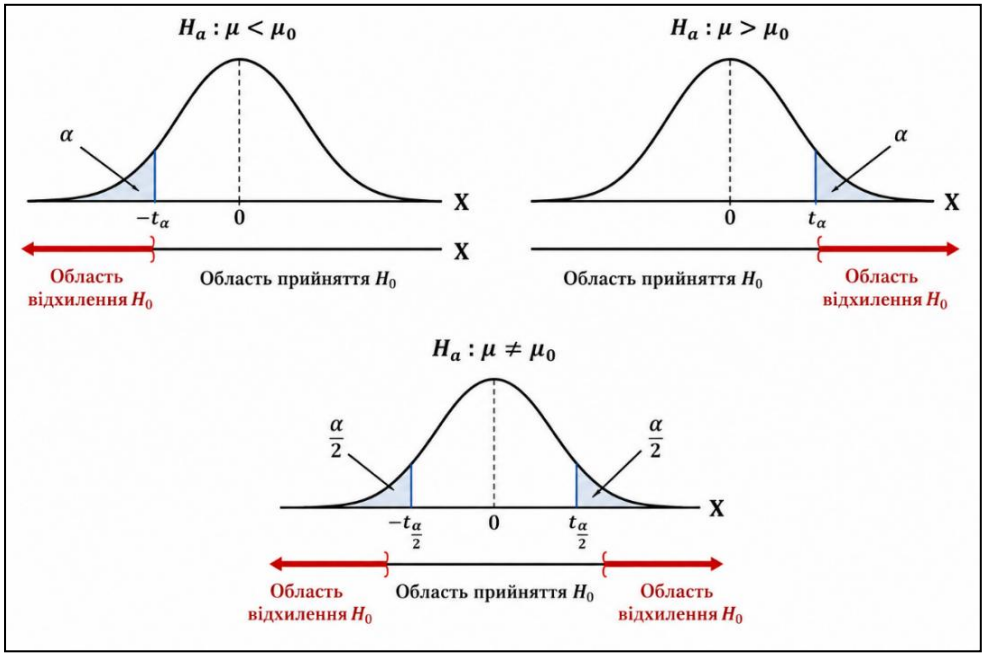


Рис 3.1 Області прийняття та відхилення нульової гіпотези та рівень значущості α для односторонніх та двосторонніх гіпотез

3.2.3. Аналіз висновків при перевірці гіпотез. Процес валідації результатів аналізу

У процесі перевірки гіпотези дослідник ніколи не отримує абсолютної, 100% впевненості в результатах аналізу. ***Завжди існує ймовірність помилки.*** Це не просто сприйняти навіть на побутовому рівні. Але саме розуміння природи таких помилок, їх невідворотності, а також *вміння оцінити ймовірності їх отримання* допомагає коректно інтерпретувати результати тестів і максимально уникати хибних висновків.

Важливо розуміти, що у реальних умовах проведення експериментів або аналізу даних істинність чи хибність нульової гіпотези ніколи не є відомою наперед. Неможливо достеменно знати, чи насправді існує різниця між групами об'єктів, що досліджуються. Чи ефект, що спостерігається є реальним, а не випадковим. Статистичний аналіз – в даному разі аналіз гіпотез - лише дає підстави для прийняття рішення на основі вибіркової інформації, але не гарантує

абсолютної правильності цього рішення. Будь-який тест, критерій чи алгоритм перевірки гіпотез - це інструмент *із певним рівнем надійності*, а не "детектор істини".

Щоб хоч якось оцінити, наскільки можна довіряти отриманим результатам (саме результатам, а не теоретичним конструкціям), необхідно провести *попередню перевірку* (*валідацію*) обраного методу дослідження для даних, наданих для аналізу. ***Валідація*** (англ. *Validation*) - це процес оцінювання точності алгоритму чи методу на контрольних (ще кажуть – *тестових*) даних, для яких істинний стан речей відомий (в нашому випадку - де відомо, істинна чи хибна гіпотеза H_0). *Мета валідації* - перевірити, наскільки правильно алгоритм приймає рішення, перш ніж використовувати його в подальшому на реальних даних. Крім того, її мета - переконатися, що результати не є випадковими і можуть бути відтворені на даних вибірок, відмінних від тої, на якій відбувалася побудова алгоритму (*тренування моделі*).

Іншими, більш простішими словами. В ході дослідження отримуються певні висновки на основі множини даних *тренувальної вибірки*. Наступним кроком виконується валідація, тобто перевіряється, на скільки висновки виявляться коректними, якщо їх застосувати до даних *тестувальної вибірки*. Якщо обраний показник якості результатів сприймається як достатньо високий, робиться припущення, що і при роботі з іншими даними (на рис 3.2 позначено червоним кольором), які отримуються від об’єкту спостереження, висновки будуть не менш коректними, аніж при роботі з тестовими даними. Якщо-ж при валідації виявиться, що на тестових даних отримуються результати суттєво (або кажуть – *статистично значимо*) гірші, аніж це було на тренувальних даних, це означає, що побудований алгоритм прийняття/відхилення гіпотези скоріш за все не буде коректно працювати і на інших, відмінних від тестових, даних. А отже використання його результатів для прийняття подальших рішень та формування висновків - неможливий.

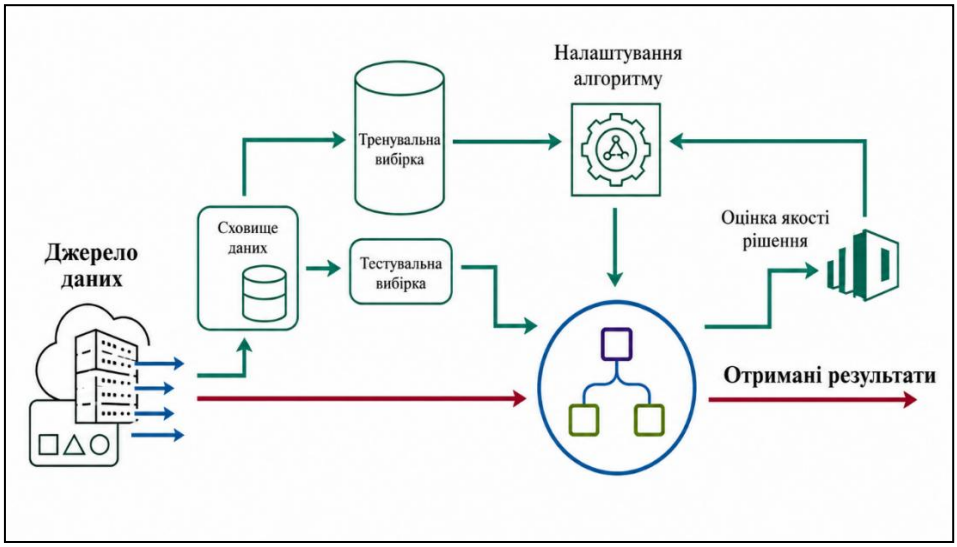


Рис. 3.2 Тренування, валідація та використання алгоритму (моделі)

Важливо: валідація на етапі дослідження є єдиним практичним способом перевірити, чи справді статистичний тест або алгоритм приймає достовірні рішення. Завдяки валідації можливо оцінити реальну ефективність статистичного підходу який використовується, і зрозуміти, чи можна - і головне *наскільки* можна - довіряти висновкам, отриманим при виконанні процедур аналізу. Це дає можливість обґрунтовувати статистичні висновки не тільки теоретичними побудовами чи математичними формулами, а й результатами їх емпіричної перевірки, реально оцінювати, наскільки результати досліджень можна вважати достовірними.

Отже, практична реалізація валідації передбачає розбиття наявних даних на декілька неперетинних підмножин, які виконують різні ролі в процесі аналізу. Найпростішою і водночас найпоширенішою схемою є поділ вибірки на:

- *тренувальну (навчальну)* вибірку - використовується для побудови моделі, оцінювання параметрів статистичного тесту або налаштування алгоритму прийняття рішень;
- *тестову* (контрольну) вибірку - використовується виключно для оцінки якості отриманих результатів; вона не повинна впливати на процес побудови моделі.

Принциповим є те, що тестові дані мають бути «невидимими» для алгоритму на етапі навчання. Якщо метод або модель використовували інформацію з тестової вибірки під час навчання, *то така перевірка вже не є*

коректною валідацією, а отримані оцінки якості - завищеними та методологічно некоректними.

У найпростішому випадку початкову вибірку ділять *випадковим чином*, наприклад, у співвідношенні 70–80% спостережень - тренувальна вибірка, 20–30% - тестова вибірка. Формального методу визначення такого співвідношення не існує. Але треба розуміти, що зменшуючи долю тренувальних даних ми вірогідно погіршуємо точність отриманої моделі, в той час, як зменшуючи долю тестувальних даних - погіршуємо точність оцінки якості моделі. Таким чином обране розбиття – результат компромісу між вказаними чинниками.

Випадковий характер розбиття важливий для того, щоб обидві вибірки були репрезентативними щодо генеральної сукупності. У випадках, коли дані мають складну структуру (наприклад, класи з різною частотою або часову залежність), застосовуються спеціальні схеми розбиття. Зокрема при роботі з часовими рядами для тестової вибірки мають відбиратися дані, що найпізніше були отримані в процесі вимірювання значень.

Процес валідації зазвичай має такий узагальнений вигляд:

1. *Побудова методу або моделі на тренувальних даних*. На цьому етапі визначаються параметри статистичного критерію, пороги прийняття рішень, оцінюються коефіцієнти моделей, налаштовуються гіперпараметри алгоритмів тощо. Усі рішення приймаються виключно на основі тренувальної вибірки.
2. *Фіксація побудованого алгоритму.* Після навчання модель вважається «замороженою»: її параметри більше не змінюються. Це важливий методологічний крок, що відокремлює етап побудови методу від етапу його перевірки.
3. Оцінка якості на тестових даних. Побудований алгоритм застосовується до тестової вибірки. Важливо, що для даних цієї вибірки нам відомі «правильні» результати, тобто дані, що отримані від джерела інформації в процесі його функціонування. В ідеальному випадку хотілося-б, щоб побудований алгоритм при тих самих вхідних даних давав результати, максимально наближені до вказаних «правильних» (або тестових) значень. Обчислюються показники якості, які можуть варіюватися в залежності від задачі, яка ставиться в дослідженні. Зокрема, в задачах регресії це можуть бути прямі статистичні метрики, як то стандартна квадратична помилка, середня

абсолютна помилка, коефіцієнт детермінації тощо. В задачах класифікації та інших алгоритмах - частота помилок першого та другого роду, точність, повнота, ймовірність правильного рішення, стабільність оцінок тощо - залежно від задачі перевірки гіпотез або типу моделі.

4. *Порівняння результатів на тренувальних і тестових даних.* Якщо якість на тренувальних даних істотно вища, ніж на тестових, це є ознакою т.з явища *перенавчання* (англ. *Overfitting*) - алгоритм добре «пояснює» конкретні дані, але погано узагальнює закономірності. Якщо ж якість на обох вибірках (тестовій та тренувальній) подібна, можна вважати, що метод є стабільним і має прийнятну узагальнюючу здатність.
5. *Прийняття рішення щодо застосовності методу.* За результатами валідації робиться висновок, чи можна використовувати побудований алгоритм для аналізу нових даних, та з яким рівнем довіри слід інтерпретувати його результати.

У багатьох практичних задачах одноразового поділу на тренувальну та тестову вибірки може бути недостатньо. У таких випадках застосовують більш стійкі процедури, зокрема багаторазове випадкове розбиття даних у різних його різновидах (т.з. *Крос-валідація*). Сенс цих підходів полягає в тому, щоб оцінити не «випадкову удачу» алгоритму на одному конкретному поділі даних, а його усереднену поведінку на різних можливих вибірках з тієї ж генеральної сукупності.

Валідація грає фундаментальну роль в усіх без винятку задачах, які будуть в подальшому розглядатися в рамках даного курсу. Фактично, валідація є мостом між теоретичною статистикою та реальними даними, де помилки неминучі, а надійність методів повинна підтверджуватися не лише формулами, а й експериментально. Тому розуміння ідей, що закладені в валідаційний процес і в обробку його результатів, вкрай важливі для будь-якого аналізу даних, від перевірки найпростіших статистичних гіпотез, методів аналізу часових рядів до побудови найскладніших сучасних алгоритмів машинного навчання та нейромережевих алгоритмів.

3.2.4. Помилки при перевірці гіпотез

Як було сказано вище, для виконання процедури валідації отриманих результатів дослідник має мати в своєму розпорядженні дані (наприклад, отримані в минулому в результаті моніторингу об’єкта спостереження), про які

точно відомо, чи гіпотеза, що досліджується, справдилася, чи виявилася хибною. При перевірці статистичної гіпотези можливі наступні варіанти комбінацій між реальністю та результатами, які отримуються в результаті дослідження, а саме:

- гіпотеза H_0 в дійсності істина, в результаті аналізу прийняте рішення про її підтвердження (тобто гіпотеза H_0 не відхиляється);
- гіпотеза H_0 в дійсності хибна, в результаті аналізу прийняте рішення про її відхилення (тобто приймається гіпотеза H_a);
- гіпотеза H_0 в дійсності істина, але результаті аналізу прийняте рішення про її відхилення (тобто помилково приймається гіпотеза H_a);
- гіпотеза H_0 в дійсності хибна, але в результаті аналізу прийняте рішення про її прийняття (тобто гіпотеза H_0 помилково не відхиляється).

В процесі валідації підраховується кількість експериментів, в результаті якого отримані ті чи інші результати, перераховані вище. Зрозуміло, що найкращим би виявився результат, коли два останні варіанти зовсім відсутні. Нажаль в реальних задачах така ситуація не зустрічається майже ніколи – і через ймовірнісну природу самих даних і через певні особливості алгоритмів їх аналізу, які є у розпорядженні дослідника.

Але абсолютні значення кількості правильних чи хибних рішень, отриманих в процесі валідації ще не дають повного уявлення про якість отриманих результатів. Реальне значення *має відношення випадків, в яких був зафіксований той чи інший* (можливий з чотирьох згаданих) *результат до загальної кількості випадків, які аналізувалися в процесі валідації*. З цими показниками якості і працюють далі.

Перший та другий з зазначених варіантів означають, що процедура перевірки гіпотез дала правильні рішення. Їх називають ще *істинне позитивне рішення* (англ. *True Positive, (TP)*) та *істинне негативне рішення* (англ. *True Negative, (TN)*) відповідно. Щоб не заплутатися - перше слово («істина», «True») підкреслює, що тест показав правильний результат, а друге слово – яке саме рішення по відношенню до гіпотези H_0 при цьому було прийнято («*позитивне» або «негативне»,* «Positive» або «Negative»).

Третій та четвертий з перелічених вище варіантів означають, що результат перевірки гіпотези привів до рішення, яке не відповідає реальності. Вони навіть мають свої спеціальні назви.

Ситуація, коли *відхиляється істинна нульова гіпотеза* носить назву **помилки I роду**. Прикладами таких ситуацій є, зокрема, формування висновків про наявну атаку, що були зроблені антивірусною системою у разі, коли такої атаки насправді не відбувалося (зверніть увагу, що нульова гіпотеза в даному разі передбачає відсутність атаки). Або система біометричного сканування, яка перевіряє права доступу, помилково ідентифікує легітимного користувача як стороннього, що призводить до блокування його доступу. В деяких прикладних галузях ціна такої помилки може бути надзвичайно високою. Наприклад, засудження невинної людини в криміналістиці. Помилку I роду ще часто називають *помилковою тривогою, помилкою хибного спрацьовування*, чи *хибного виявлення* (англ. *False Positive, (FP)*).

Помилка II роду виникає, коли *хибна нульова гіпотеза в результаті аналізу не відхиляється*. Тобто реальний ефект або відмінність існує, але застосований інструмент (метод, критерій, тест) не зміг його виявити. Наприклад - система кіберзахисту не виявила реальну атаку (відхилила H_0 гіпотезу про відсутність атаки, коли в дійсності атака мала місце), біометричній системі не вдається розпізнати незареєстрованого користувача, і вона вважає його легітимним. У кримінальному процесі - виправдання винного злочинця. Похибки другого роду є істотною проблемою в медичному тестуванні, бо дають пацієнтові і лікареві помилкове переконання, що захворювання відсутнє, тоді як насправді воно є. У термінології теорії прийняття рішень, технічної діагностики чи аналізу сигналів *"помилка пропущеного спрацьовування"* (англ. *Missed Detection Error*) - це ситуація, коли система не спрацьовує, хоча подія або сигнал справді відбулися. Також часто вживається термін *помилка пропущеного виявлення* (англ. *False Negative, (FN*)).

Інформацію, отриману в ході процесу валідації, часто зводять в ***матрицю невідповідностей*** (англ. *Confusion Matrix*). Див. Рис.3.3.

(*Застереження*: іноді в аналогічних таблицях міняють місцями рядки та стовпчики, або порядок самих стовпчиків та рядків. Тому завжди треба уважно дивитися, що і де позначено саме в тій таблиці, з якою вам доведеться працювати).

		Істинна гіпотеза	
		H_0	H_1
Результат застосування критерію	H_0	H_0 правильно прийнята	H_0 неправильно прийнята (Похибка *другого* роду)
	H_1	H_0 неправильно знехтувана (Похибка *першого* роду)	H_0 правильно знехтувана

Рис. 3.3 Матриця невідповідностей процесу валідації алгоритму перевірки статистичної гіпотези

Два типи помилок є конкуруючими: зменшення ймовірності одного типу помилки, як правило, збільшує ймовірність іншого. Чим чутливіша система, тим більше небезпечних (небажаних, незвичних, аномальних) подій вона детектує і, отже, запобігає. Але при підвищенні чутливості неминуче зростає і ймовірність помилкових спрацьовувань. Тому *занадто* чутливо налаштована система захисту може перетворитися на свою протилежність і привести до того, що побічна шкода від неї перевищуватиме користь. Якщо деяка система виявлення вторгнень генерує, наприклад, 100 попереджень на день, з яких лише 10 - реальні загрози, решта - помилкові спрацювання, і якщо при цьому зменшити рівень чутливості, кількість попереджень зменшиться до 20, але система пропустить частину реальних атак (зросте β, зменшиться потужність критерію – див. далі). Якщо навпаки, підвищити чутливість, можна буде виявити (майже) всі атаки, але аналітики центру забезпечення безпеки будуть перевантажені перевіркою сотень хибних сигналів. Отже, *ступінь чутливості системи повинен бути компромісом між імовірністю похибок першого і другого роду, що спирається на прикладний аналіз систем.* Де саме знаходиться точка балансу, залежить від оцінки ризиків обох типів помилок. Часто для ухвалення рішення використовується деяке порогове значення, яке може варіюватися з метою зробити тест суворішим або, навпаки, м'якшим. Таким пороговим значенням є рівень значущості, яким задаються при перевірці статистичних гіпотез. Наприклад, у випадку детектора

металу підвищення чутливості приладу приведе до збільшення ризику похибки першого роду (помилкова тривога), а пониження чутливості - до збільшення ризику похибки другого роду (пропуск забороненого предмету). У кібербезпеці оптимальний компроміс між помилками I та II роду часто залежить від «ціни» (ресурсів, потужностей, задіяних спеціалістів тощо) кожної помилки. Для військових, енергетичних чи фінансових систем частіше жертвують збільшенням α (хибних тривог), щоб мінімізувати β (ризик пропуску атаки).

Ймовірність похибки першого роду при перевірці статистичних гіпотез і є відомим вже нам рівнем значущості α. Ймовірності помилки другого роду, отримане при проведенні процедури валідації, позначають як β. Величина $1-\beta$ носить назву потужності критерію. Чим вище потужність, тим менше імовірність зробити помилку другого роду. Ці показники можна отримати за результатами аналізу процедури валідації:

- $\alpha = \frac{FP}{FP+TN}$, і представляє собою оцінку ймовірності зробити помилку спрацювання, коли гіпотеза H0 насправді вірна;
- $1-\beta = \frac{TP}{TP+FN}$, і представляє собою оцінку імовірності правильного виявлення істинної альтернативної гіпотези.

Для інформації. Це ще не всі показники якості процедур перевірки гіпотез, які використовуються на практиці. Багато з таких показників як правило застосовуються при аналізі окремих, спеціальних видів гіпотез, які виникають в ході досліджень. Ми будемо з ними знайомитися по ходу вивчення цих різновидів аналізу, тут лише приведемо деякі з них, щоб ви мали змогу з ними попередньо ознайомитися:

- *Точність (англ. Accuracy)* = (TP + TN) / (TP + TN + FP + FN) - загальна частка правильно прийнятих рішень;
- *Повнота (*або *Чутливість*; англ. *Recall, Sensitivity*) = TP/(TP+FN) - наскільки добре система виявляє реальні події;
- *Прецизійність (*англ. *Precision)* = TP/(TP+FP) - наскільки результати спрацювань є правильними (часто цей показник також називають «точністю», що вносить плутанину в розуміння отриманих результатів - будьте уважними);

– *Специфічність (*англ. *Specificity)* = TN/(TN+FP)TN / (TN + FP)TN/(TN+FP) - здатність правильно визначати відсутність події.

Досить часто виникає питання - «а навіщо так багато метрик»? Давайте поглянемо на дві з них – *Recall* та *Precision*. В якості прикладу візмемо систему виявлення шахрайських транзакцій у банку. В даному разі показник *Recall* (*повнота*) відповідає на запитання: із усіх реальних шахрайських транзакцій, який відсоток моделі вдалося виявити? Якщо зі 100 шахрайських модель знайшла 80 і пропустила 20 – *recall = 0.80*. Інші 20 пройшли повз, і банк втратив гроші.

Precision (*точність*) відповідає на інше питання: з усіх транзакцій, які модель визначила як шахрайські, який відсоток справді такими є? Якщо модель помітила 100 транзакцій як шахрайські, з яких 40 є реально шахрайські, а 60 - нормальні, які модель заблокувала помилково - *Precision = 0.40*. 60 клієнтів отримали дзвінок від банку чи заблоковану картку без причини.

Між *Recall* і *Precision* завжди існує конфлікт. Якщо основною задачею є зловити більше шахраїв – потрібно досягати якомога більш високого значення показника *Recall*. Модель буде агресивнішою, щодо аномалій, але водночас почне помилково блокувати більше нормальних транзакцій (*Precision* впаде). Якщо поставлена задача зменшити кількості помилкових блокувань (тобто отримати високе значення *Precision*), модель стане обережнішою і почне пропускати шахраїв (значення *Recall* впаде).

Цей компроміс визначається виключно з точки зору використання в прикладній системі. Для того ж спам-фільтра хибне спрацьовування (важливий лист у спамі) гірше, ніж пропущений спам, тобто показник *Precision* важливіший. Для медичного скринінгу пропущений діагноз може коштувати життю, і показник *Recall* стає критичним.

Досить повний, нескладний та цікавий вступ до основ теорії перевірки гіпотез можна знайти за посиланням [48].

3.2.5. Критерії узгодженості, однорідності та значущості

Спочатку про терміни. У статистиці «критерій» і «тест» - практично означають одне й те саме, різні назви для одного інструмента. Можна вважати, що *критерій - це певне правило прийняття рішення*, а *тест - це практична процедура застосування цього правила до даних*. Наприклад, уявімо, що є правило: «якщо різниця між середніми достатньо велика - вважаємо, що ефект є». Само це правило - статистичний критерій. А коли дослідник бере реальні дані,

рахуєте статистику, *p_value* і їх основі приймає рішення - виконується статистичний тест.

У процесі статистичного аналізу дослідник часто стикається з необхідністю оцінити, наскільки добре дослідні дані відповідають певним теоретичним уявленням або чи є суттєві відмінності між групами даних. Для цього використовують різноманітні статистичні тести, або часто говорять – критерії, кожен з яких відповідає певному типу гіпотез та має свою логіку застосування.

Серед найбільш поширених у практиці виділяють:

- *критерії узгодженості*,
- *критерії однорідності*,
- *критерії значущості* (або *критерій порівняння*).

Ці критерії розв'язують різні задачі, хоч і спираються на один загальний математичний апарат - теорію перевірки статистичних гіпотез.

Критерії узгодженості (англ. *Goodness-of-Fit Tests*) *застосовують для перевірки того, наскільки добре експериментальні дані узгоджуються з певним теоретичним законом розподілу*. Інакше кажучи, а ході аналізу перевіряється, *чи можна вважати, що дана вибірка походить із генеральної сукупності з певним розподілом.* Типові приклади прикладних задач, що зводяться до використання критерію узгодженості:

- чи відповідає розподіл типів мережевих пакетів (HTTP, HTTPS, DNS, FTP, тощо) у поточний момент часу нормальному (типовому) розподілу, який ми спостерігали у минулому? Відхилення може свідчити про DDoS-атаку, сканування, приховане перенесення даних тощо;
- в криптографії дуже важливо, щоб послідовності, згенеровані генераторами псевдовипадкових чисел були максимально наближені до рівномірного закону розподілу з заданими характеристиками;
- чи спроби невірної автентифікації (введення невірного пароля) розподілені за логнормальним законом? Аномально велика кількість невдалих спроб для одного чи кількох облікових записів може свідчити про атаку підбору паролів (brute-force) або розпилення паролів (password spraying);
- компанія, яка виробляє товар у чотирьох різних кольорах (А, Б, В, Г), з метою збільшення прибутків хоче перевірити, чи є попит на ці кольори

рівномірним, як передбачає їхня виробнича стратегія, чи певний колір користується значно більшою чи меншою популярністю;

– чи описує кількість звернень у службу підтримки розподіл Пуассона?

Цей перелік можна продовжувати практично нескінченно.

При використанні критеріїв узгодженості нульова гіпотеза формулюється наступним чином:

$$H_0 : F(x) = F_o(x)$$

тобто, емпіричний розподіл збігається з теоретичним.

Альтернативна гіпотеза при цьому:

$$H_a : F(x) \neq F_o(x)$$

Найпоширеніші критерії узгодженості, що застосовуються на практиці:

– χ^2-критерій Пірсона
– Критерій Колмогорова-Смирнова (K-S test)
– Критерій Шапіро-Уілка
– Критерій Андерсона-Дарлінга

Критерії однорідності використовують для перевірки того, *чи можна вважати, що декілька вибірок отримані з однієї й тієї самої генеральної сукупності, тобто чи є вони однорідними.* На відміну від критерію узгодженості, який порівнює спостережувані дані з теоретичним розподілом, критерій однорідності порівнює спостережувані дані між двома або більше емпіричними групами.

Типові реальні приклади, при аналізі яких доцільно використовувати критерії однорідності:

– чи однаковим був розподіл запитів у різні години доби?
– чи схожі профілі активності користувачів різних сегментів аудиторії?

– чи є розподіл типів виявлених вразливостей (або успішних атак) однаковим у внутрішній мережі, де працює посилений захист, і у демілітаризованій зоні, де захист є стандартним?
– чи є розподіл методів вторгнення, використаних зловмисниками, однаковим протягом "нормального" кварталу та протягом кварталу, коли відбулася велика геополітична подія, що потенційно змінила їхню тактику?
– чи можна вважати, що два алгоритми класифікації дають результати з однаковим розподілом помилок?
– тощо.

В даному разі нульова гіпотеза формулюється як:

$$H_0 : F_1(x) = F_2(x)$$

а альтернативна як:

$$H_a : F_1(x) \neq F_2(x)$$

Найбільш вживаними критеріями однорідності є:

– χ^2-критерій для таблиць спряженості (Contingency Tables);
– критерій Колмогорова-Смирнова для двох вибірок;
– непараметричні критерії Манна-Уітні, Крускала-Уолліса (для медіан).

Критерії значущості (критерії порівняння параметрів) - Це критерії, які використовують *для перевірки гіпотез про рівність параметрів розподілу*: середніх, дисперсій, часток тощо. На відміну від критеріїв узгодженості та однорідності, тут увага зосереджена не на законі розподілу в цілому, а на конкретних числових параметрах.

Типові приклади задач, де такі критерії застосовуються:

– чи змінився середній час відгуку сервера після оптимізації?
– чи є середнє навантаження на процесор однаковим для серверів, захищених проксі-сервером А, і серверів, захищених проксі-сервером Б?
– чи відрізняється середній бал студентів двох груп?
– чи змінилась частка позитивних відповідей після редизайну сайту?

– чи є мінливість (дисперсія) часу затримки мережевого пакету однаковою для двох різних провайдерів (Провайдер X та Провайдер Y)?

Формулювання гіпотез в даному разі змінюються, відповідно до того, про який з параметрів вибірок та популяцій йдеться в дослідженні. Детальніше ми про це поговоримо, коли будемо розглядати таки тести.

Поширені критерії значущості.

Для середніх:

– t-критерій Стьюдента (одновибірковий, двовибірковий, парний)
– t-критерій Велча (нерівні дисперсії)
– z-критерій (великі вибірки)

Для дисперсій:

– F-критерій Фішера (дві дисперсії)
– Тест Бартлетта
– Тест Левене

Для часток:

– z-тест для однієї частки
– z-тест для двох часток
– χ^2-критерій для таблиць 2×2

Критерії узгодженості, однорідності та значущості є фундаментальними інструментами в статистиці. *Їхнє грамотне застосування дозволяє оцінити відповідність даних певним теоретичним припущенням; перевірити, чи належать вибірки до однієї генеральної сукупності; визначити, чи є відмінності між параметрами груп статистично значущими.* Розуміння цих критеріїв є необхідною умовою для правильного формулювання та перевірки статистичних гіпотез, а також для адекватного інтерпретування результатів у наукових дослідженнях, аналітиці, машинному навчанні, та прикладній інженерії.

3.2.6. Параметричні та непараметричні критерії.

Вище вже з’являлися терміни «*параметричний*» або «*непараметричний*» тест. Статистичні критерії (тести) поділяються на параметричні та непараметричні залежно від того, які припущення робляться щодо розподілу даних та типу вимірювань. Обидва класи тестів застосовуються для перевірки гіпотез, але підходять для різних типів досліджень та різної якості даних.

Розуміння відмінностей між цими групами тестів дозволяє правильно обирати методи аналізу та уникати помилкових висновків.

Параметричні критерії - це статистичні методи, що базуються на припущеннях про параметри генеральної сукупності, передусім про форму розподілу (зазвичай - нормальність). Вони оцінюють характеристики, як-от середнє значення, дисперсію, кореляцію. Передумовою використання параметричних тестів, зокрема, часто є нормальність, або хоча-б відомість закону розподілу даних у вибірці або у вибірках, для деяких тестів - рівність дисперсій у різних групах. Дані для таких тестів мають бути представлені кількісними змінними (тобто виміряні у інтервальній шкалі або у шкалі відношень). Коли ці припущення виконуються, параметричні тести мають високу ефективність і статистичну потужність. Але для цих тестів притаманні і відповідні недоліки, зокрема - чутливість до відхилень від заданого закону розподілу; неможливість використання для порядкових або номінальних даних, щонайменше - без додаткових перетворень; неефективність при роботі з малими вибірками.

Непараметричні критерії - це статистичні методи, що не потребують припущень про конкретну форму розподілу даних. Вони часто використовуються до даних, що виміряні у рангових, порядкових шкалах, або до даних, що не є розподіленими за нормальним законом. Такі тести працюють переважно з порядком величин, а не з їх чисельними значеннями, що робить їх універсальнішими для «проблемних» вибірок. Непараметричні тести менш чутливі до викидів в даних, зберігають працездатність при роботі з вибірками малих розмірів, тощо. Їх недоліками є менша статистична потужність порівняно з параметричними тестами; складніше інтерпретування результатів у термінах параметрів популяції; через перехід до рангового представлення інформації, тобто через зниження потужності шкали вимірів деяка інформація про дані безповоротно втрачається.

Параметричні та непараметричні тести є інструментами статистичного аналізу, що доповнюють одні одного. Параметричні методи забезпечують точні й ефективні оцінки в тих випадках, коли дані мають «добрі» статистичні властивості, тоді як непараметричні методи дозволяють працювати з нестандартними, малими або ранговими вибірками. Правильний вибір тесту гарантує достовірність висновків і підвищує надійність статистичного аналізу. Далі ми познайомимося з прикладами таких тестів і застосуємо їх до перевірки

різноманітних статистичних гіпотез.

3.3. Практичні задачі з перевірки статистичних гіпотез. Деякі статистичні тести та їх застосування

Існують досить багато різновидів тестів (критеріїв), що реалізують процедури перевірки гіпотез. *Вибір конкретного тесту залежить від поставленої задачі дослідження; від шкали вимірів даних, що складають вибірки які досліджуються; від постановки статистичної задачі; від статистик, які перевіряються, тощо.* Нижче будуть розглянуті найбільш вживані в практичних завданнях тести.

3.3.1. Критерії аналізу середніх значень вибірок та математичних очікувань генеральних сукупностей

Нагадаємо. У статистиці та машинному навчанні *параметр* – це один з описів, ознак популяції (генеральної сукупності), тоді як *статистика* стосується описів відносно невеликої, але доступної для безпосереднього вивчення, частини популяції - вибірки. Див. розділи 2.3.3 та 2.3.4.

Сказане вище математично можна виразити через такі поняття, як математичне очікування μ, середнє значення вибірки $\bar{x}$ та гіпотези. Поки будемо розглядати випадки, коли вид закону розподілу значень генеральної сукупності відомий. Для прикладу, нехай це буде нормальний закон розподілу. Тоді задача формулюється наступним чином: «чи можливо отримати значення вибіркового середнього $\bar{x}$, якщо генеральна сукупність підпорядковується нормальному закону розподілу з параметрами математичного очікування, що дорівнює μ_0, та дисперсією, що дорівнює σ_0»? При цьому нульова гіпотеза та гіпотеза альтернативна до неї, формально записується наступним чином:

$$H_0 : \mu = \mu_0$$
$$H_a : \mu \neq \mu_0$$

В разі односторонніх тестів (див. Розділ 3.2.2) , альтернативна гіпотеза приймає вигляд:

$$H_a : \mu < \mu_0 \text{ або } H_a : \mu > \mu_0 .$$

Маючи вибірку, можливо визначити вибіркове середнє $\overline{x}$. (див. Розділ 2.3.4.2). Зрозуміло, що отримане таким чином значення $\overline{x}$ може відрізнятися від реального значення μ. Зрозуміло, що проводячи серію аналогічних експериментів, обираючи в кожному з них різні вибірки з однієї і тієї ж генеральної сукупності, кожного разу будуть отримуватися відповіді, що будуть відрізнятися між собою. Це добре видно з рис. 3.4. Але – на скільки відрізнятися?

А якщо в експерименті отримали лише одне значення $\overline{x}$, то що в такому разі можна сказати про μ? Або формально: якщо нульова гіпотеза $H_0 : \mu = \mu_0$ справедлива, то яка ймовірність того, що ми отримаємо саме значення $\overline{x}$?

3.3.1.1. z-критерій

z-критерій (або *z-тест*) - це один із найпоширеніших методів статистичної перевірки гіпотез щодо *середнього значення генеральної сукупності, коли відома дисперсія (стандартне відхилення) сукупності*. z-критерій ґрунтується на припущенні про те, що досліджувана змінна має *нормальний розподіл*, а також що вибірка є досить великою, щоб можна було використати висновки *центральної граничної теореми*. Як відомо з курсу теорії ймовірності, ці висновки стверджують, що *при проведенні нескінченної послідовності дослідів*, та побудови з такої сукупності вибірок набору *вибіркових середніх*, *середнє значення набору вибіркових середніх прямує до математичного очікування генеральної сукупності.*

За визначенням, z-критерій аналізує статистику z:

$$z_p = \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

де n - об'єм вибірки. Позначка «р» при z введена, щоб підкреслити, що за цією формулою можливо отримати розраховане (тобто - експериментальне, емпіричне) значення цієї статистики. З іншої сторони, по своїй суті z є квантилем стандартного нормального розподілу, а відтак може бути розрахований при обраному значенні рівня значущості α (який - ще раз підкреслимо - є позастатистичною величиною і має задаватися до початку процедури аналізу).

Якщо отримане значення z_p є більшим по абсолютній величині за критичне значення при обраному рівні значущості α, то це означає, що значення вибіркового середнього $\overline{x}$ «надто далеко» відхилися від μ_0.

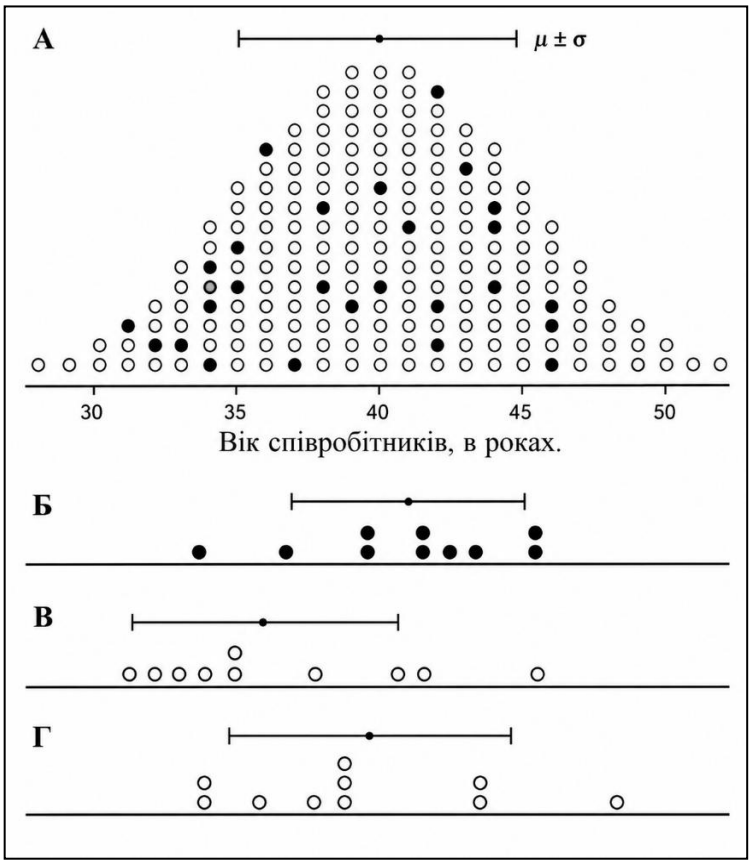


Рис 3.4 Генеральна сукупність, вибірки та середні значення вибірок

Таким чином *критичні значення* статистики $z_{кр}$ задають інтервал, в якому знаходиться $(1-\alpha)$ всіх її (статистики) можливих значень:

$$\left[-z_{1-\frac{\alpha}{2}}, +z_{1-\frac{\alpha}{2}}\right]$$

Зверніть увагу, що інтервал тим ширший, чим менше значення α, а також чим більше значення n. Вихід значення z_p за межі цього інтервалу вказує на обґрунтованість відхилення гіпотези H_0 (ще раз нагадаємо - з ймовірністю помилитися, що дорівнює α або менше). В іншому випадку підстав для відхилення гіпотези H_0 не знайдено.

У уважного читача може виникнути питання. Чому при стандартизації

нормального набору даних, що розглядалася раніше (розділ 2.4.2.2), використовується значення стандартного відхилення розподілу (σ), а в z-тесті – значення стандартного відхилення, яке ще й додатково ділиться на корінь квадратний з об'єму вибірки? Справа в тім, що при нормалізації ми працюємо з кожним значенням даних з наданого набору окремо. При цьому метою є їх переведення у стандартну шкалу. Для такого перетворення використовується формула:

$$z_i = \frac{x_i - \mu}{\sigma},\ x_i \in X$$

Отже, в даному разі просто визначається, на скільки стандартних відхилень певне значення x_i відрізняється від математичного очікування генеральної сукупності. Кожне значення x_i має свою варіацію, яку визначає σ - стандартне відхилення значень популяції.

У z-тесті аналізуються вже не окремі значення, а оцінюється *середнє значення вибірки середніх*, тобто статистик, *кожна з яких* характеризує усю вибірку. Але середнє вибірки не коливається так сильно, як окремі значення (це добре помітно навіть з рис. 3.4.). Центральна гранична теорема як раз і говорить, що середні значення, які отримані на різних вибірках однієї і тієї самої генеральної сукупності, будуть розподілені за нормальним законом розподілу з значенням математичного очікування, що співпадає з математичним очікуванням самої генеральної сукупності, і з дисперсією, що дорівнює $\sigma/\sqrt{n}$. Останню величину ще називають *стандартною похибкою середніх* (англ. *Standard Error of the Mean*). Тобто z-тест перевіряє, наскільки отримане в експерименті *вибіркове середнє* відрізняється від гіпотетичного математичного очікування μ_0, в вимірах *стандартної похибки середніх*, а не стандартного відхилення окремих значень популяції. Отже, провіряються не «наскільки одне спостереження відрізняється від математичного очікування», а «*наскільки середнє вибірки відхиляється від математичного очікування генеральної сукупності*», причому - що важливо - з урахуванням розміру вибірки, що надана для дослідження.

z-критерій є базовим інструментом для інших статистичних критеріїв (t-тести, χ^2-тести), але має певні обмеження та недоліки. Він менш точний (менш потужний) при малих розмірах вибірок або при навіть незначних відхилень

значень від нормального закону розподілу. Він потребує знання дисперсії генеральної сукупності. Принагідно зазначимо, що не існує універсального визначення розміру, за якого вибірка вважається достатньо великою, щоб допустити використання z-критерію. *Типове емпіричне правило: розмір вибірки має становити не менше 50-100 спостережень, а бажано - більше. І чим більше, тим краще.*

Екосистема Python, зокрема бібліотеки *scipy* та *statsmodels*, містить низку готових інструментів, які значно спрощують застосування z-критерію на практиці. Неведений нижче приклад показує, як за допомогою цих засобів може проводитися відповідний аналіз.

```
from statsmodels.stats.proportion import proportions_ztest
# x - кількість успішних подій
# n - загальна кількість спроб
x = 58
n = 100
p0 = 0.5
z_stat, p_value = proportions_ztest(count=x, nobs=n, value=p0)
print("z-статистика:", z_stat)
print("p_value:", p_value)
```

Отримані значення z-статистики та *p_value* дозволяють зробити висновок щодо прийняття або відхилення нульової гіпотези (в коді - не показано).

3.3.1.2. Одновибірковий t-критерій Стьюдента

Одновибірковий t-критерій Стьюдента, як і z-критерій, належать до параметричних методів перевірки статистичних гіпотез про математичне очікування генеральної сукупності. *Обидва критерії використовуються для оцінки того, чи відрізняється вибіркове середнє від відомого теоретичного значення, проте вони мають різні умови застосування.*

При визначенні статистики z-критерію потрібно мати в розпорядженні значення дисперсії генеральної сукупності. В деяких випадках така інформація може бути отримана, наприклад, з теоретичних міркувань, або при можливості попереднього проведення дуже великої (в теорії - нескінченої) кількості експериментів. Але так буває далеко не завжди. А от чи можна замінити значення

дисперсії генеральної сукупності на її вибіркову оцінку – питання не очевидне.

Як було сказано вище, вперше це питання було поставлено Уільямом Госсетом, на прикладі хімічних аналізів проб пива. Формально, статистика ***t-критерію Стьюдента*** визначається за виразом:

$$t_{\alpha,df} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

яка відрізняється від визначення статистики z-критерію лише заміною популяційної дисперсії σ на її вибіркову оцінку s. Але отримане значення порівнюється не з квантилями нормального закону розподілу, а з квантилями *t-розподілу Стьюдента*. Нагадаємо (див. розділ 2.4.4.1) цей розподіл залежить від двох параметрів – рівня значущості α та *кількості ступенів свободи* df, яке в умовах нашої задачі приймає значення *n-1*. (Що запобігти непорозумінь, відразу зазначимо, що в інших випадках нам доведеться зустрічатися і з іншими значеннями цього параметру). Розподіл має ширші хвости, ніж нормальний розподіл, що враховує додаткову невизначеність, пов'язану з оцінкою дисперсії. При малих значеннях *n* пік функції щільності t-розподілу нижчий, ніж у нормального. Отже, чим менша вибірка, тим "плоскішим" є розподіл, і тим вищим стають пороги критичних значень для відхилення нульової гіпотези. Використання t-критерію для малих вибірок дозволяє зменшити ризик помилки першого роду, який би виник при некоректному застосуванні z-критерію. Незважаючи на те, що при зростанні обсягу вибірки різниця в результатах при застосуванні обох критеріїв поступово зникає, і обидва методи дають однакові висновки, на практиці популярність застосування t-критерію лишається набагато вищою, аніж популярність застосування z-критерію.

Вище (розділ 3.2) ми ознайомилися з методологією прийняття або відхилення нульової гіпотези на основі критичних значень розподілу. По суті $t_{\alpha,df}$ є квантилем *t*-розподілу Стьюдента, а відтак може бути вирахувано, при заданих значеннях рівня значущості α (з усіма застереженнями в разі використання двосторонніх тестів). Безпосередньо з визначення випливає:

$$\mu_0 = \bar{x} \pm \frac{s}{\sqrt{n}} t_{\alpha,df}$$

Іншими словами, якщо нам відомі вибіркове середнє, дисперсія вибірки, кількість елементів у вибірці, то значення математичного очікування генеральної сукупності лежить у межах ***довірчого інтервалу***:

$$\bar{x} - \frac{s}{\sqrt{n}} t_{\alpha,df} \le \mu \le \bar{x} + \frac{s}{\sqrt{n}} t_{\alpha,df}$$

Якщо цей довірчий інтервал включає значення μ_0*, то це означає, що наявні дані* $(\bar{x}, s, n, \alpha)$ *не заперечують* ***можливість отримання середнього*** $\bar{x}$ *з значень, що вибрані з* ***генеральної сукупності з математичним очікуванням*** μ_0**.**

Отже, для отримання відповіді на питання, чи можна вважати, що маточікування генеральної сукупності дорівнює μ_0 необхідно:

1. Визначити вибіркове середнє та середньоквадратичне відхилення.
2. Визначити кордони довірчого інтервалу.
3. Перевірити, чи включає цей довірчий інтервал значення μ_0. Якщо так – ми відхилити нашу нульову гіпотезу неможливо. Якщо ні - нульова гіпотеза відхиляється (пам'ятаючи про заданий рівень значущості та ймовірність відповідної помилки).

Інформація про t-критерій Стьюдента, від його історії до його застосування, про його різновиди та приклади їх використання дуже широко розповсюдження в мережі інтернет. Для прикладу можу рекомендувати, наприклад, посилання [47] .

Як зазначалося раніше, часто виявляється необхідним не тільки прийняти чи відхилити нульову гіпотезу, але й надати інформацію, яка може певним чином трактуватися як впевненість в коректності результатів статистичного аналізу. Такою величиною є *p_value*, яка представляє собою квантиль розподілу, що відсікає деяку долю площі під кривою функції щільності розподілу. Визначення значення *p_value* досить не проста обчислювальна задача, але сьогодні, з впровадженням комп'ютерної техніки, її вирішення не являє складності.

Більшість сучасних пакетів для виконання інженерних, економічних, технічних досліджень – включаючи MS Excel, MatLab, SPSS, тощо – дають можливість отримати результати обчислень *p_value* через вбудовані функції. Не виняток і екосистема Python, в якій такі функції доступні в бібліотеці *scipy.stats*.

Формально, *p_value - це статистика, яка вимірює ймовірність отримання результату, такого самого або ще більш рідкісного аніж результат, що отриманий в експерименті, за умови, що нульова гіпотеза є істинною. (див. розділ 3.2.2).* Невелике значення величини *p_value* свідчить про те, що спостережуваний ефект (наприклад - відхилення значення маточікування від теоретичного значення) є ***статистично значущим***, але це не означає, що він обов'язково має місце. *Чим менше значення p_value, тим менше ймовірність помилитися, відхиливши нульову гіпотезу***.** Відповідно, при збільшенні значення *p_value* ймовірність помилитися при відхиленні нульової гіпотези зростає і з певного моменту стає неприйнятною для користувача. Звертаємо увагу, що в наведеному визначенні рівень значущості теж присутній, бо так чи інакше рішення має формулюватися в термінах «прийняти-відхилити». Значення α в такому разі і відіграє роль порогу, при перетині якого приймається відповідне рішення ($\alpha > p\_value$). Але присутність у результатах дослідження не тільки самого рішення, але й додаткового значення *p_value* (що теоретично може змінюватися в діапазоні від 0 до 1) надає користувачу значно більше інформації для обґрунтуванні своїх подальших дій, рішень, висновків. В деяких прикладних областях, зокрема в медичних дослідженнях, сьогодні рішення без вказівки відповідного значення *p_value* взагалі не приймається до розгляду.

Таким чином ми познайомилися з двома способами прийняття рішення щодо нульової гіпотези – на основі значення статистики, яка лежить в основі відповідного тесту, та на основі значення *p_value,* яке визначається як квантиль тестової статистики, який відповідає значенню статистики.

3.3.1.3. Реалізація одновибіркового t-тесту в бібліотеці SciPy

У *Python* одновибірковий t-критерій реалізовано у вигляді функції *ttest_1samp()* з бібліотеки *scipy.stats*. Ця функція приймає два вхідні параметри - безпосередньо вибірку, яка досліджується та очікуване значення математичного очікування генеральної сукупності. В якості результату функція повертає розраховане значення t-статистики та *p_value*.

```
from scipy.stats import ttest_1samp
t_stat, p_value = ttest_1samp(sample, mu0)
print("t-статистика:", t_stat)
print("p_value:", p_value)
```

Украй корисним доповненням до t-тесту є визначення довірчого інтервалу для середнього. Його легко обчислити самостійно:

```
from scipy.stats import t
n = len(sample)
x_bar = sample.mean()
s = sample.std(ddof=1)
alpha = 0.05
t_crit = t.ppf(1 - alpha/2, df=n-1)
ci_low = x_bar - t_crit * s / np.sqrt(n)
ci_high = x_bar + t_crit * s / np.sqrt(n)
print("Довірчий інтервал:", ci_low, ci_high)
```

У бібліотеці *scipy.stats* реалізовано обчислення довірчого інтервалу через об'єкти розподілів. Так, метод *scipy.stats.interval(alpha,df)* - повертає пару точок, що відповідають обраному значенню рівня значущості α. Зверніть увагу, що в даному разі параметр *alpha* задається «навпаки», тобто у вигляді 1-α. Крім того, в якості другого параметрі задається кількість ступеней свободи:

```
import scipy.stats as sc
alpha=0.05
sc.t.interval(1-alpha,df= n-1)
```

Інший спосіб отримати аналогічні результати:

```
sc.t.ppf (1-alpha/2, df= n-1)
```

Отримати значення *p-value* для розрахованого фактичного значення

критерію можна за допомогою методу *.cdf($t_{факт}$,df)*. Зверніть увагу на те, як цю функцію треба використати:

p_vl=1-sc.t.cdf(t_stat, df= n-1)
p_vl =sc.t.cdf(-t_stat, df= n-1)

Інший спосіб - використати спеціальну функцію:

*p_value= sc.t.sf(abs(tfact1), df= n-1) * 2*

3.3.1.4. Двовибіркові критерії. Залежні та незалежні вибірки

При перевірці гіпотез розрізняють одновибіркові та двовибіркові критерії. Їх *не слід плутати* з одно- та двосторонніми критеріями, які розглядалися раніше. В ***одновибірковому критерії*** статистика вибірки використовується для визначення істиності гіпотези щодо наперед відомого значення параметру розподілу (тести узгодженості). Прикладами одновибіркових тестів були розглянуті вище z-тест, та t-тест Стьюдента.

У ***двовибірковому критерії*** гіпотеза формується стосовно порівняння значення параметрів генеральних сукупностей, з яких були отримані вибірки.

В свою чергу, двовибіркові статистичні тести поділяються на тести з залежними вибірками і тести з незалежними вибірками.

Незалежними називають такі *вибірки*, у яких *спостереження однієї групи не впливають на спостереження іншої.* Типові ситуації: порівняння успішності двох різних класів учнів; порівняння продуктивності двох різних алгоритмів; дослідження відмінностей між двома видами об'єктів чи результатів експерименту від його умов.

Вибірки називають ***залежними***, якщо *між їх елементами існує природна або експериментальна відповідність*. Така ситуація виникає: у повторних вимірюваннях одного й того самого об'єкта; при дослідженні ефективності дії («до» та «після»); при спарених спостереженнях (наприклад, підбір «пари» за певним критерієм).

В прикладах, які наведені в таблиці 3.1, тест ліворуч справді має дві вибірки. В даному випадку - деякий параметр різних груп людей – ті, що пройшли тренування і ті, що його не проходили. В тесті праворуч порівнюється показники одних і тих самих людей, до та після того, як вони пройшли навчання.

Отже в першому випадку ми шукаємо відповідь на питання «*чи відрізняються треновані та нетреновані люди за обраним показником*». В другому випадку питання може бути сформульовано так: «*чи дійсно тренування впливає на зміну параметру, за яким ведеться спостереження*».

Якщо розглядати тести, які визначають наявність/відсутність відмінності між середніми значеннями вибірок, гіпотези будуть формулюватися наступним чином.

В разі тесту з незалежними вибірками:

$H_0 : \mu_1 = \mu_2$

$H_a : \mu_1 \neq \mu_2$ (варіант для двосторонніх гіпотез).

$H_a : \mu_1 > \mu_2$ або $H_a : \mu_1 < \mu_2$ (варіанти для односторонніх гіпотез).

де μ_1 та μ_2 - математичне очікування відповідних популяцій.

Таблиця 3.1

Порівняльна таблиця для двовибіркового t-тесту та парного t-тесту

Двовибірковий t-тест				Парний t-тест			
Співпрбітники, що не проходили тренування	Оцінка	Співпрбітники, що пройшли тренування	Оцінка	Співпрбітники	Оцінка до тренування	Оцінка після тренування	Різниця
І. Бондар	73	А. Шевчук	75	К. Олійник	79	82	3
Т. Гнатюк	81	Ю. Дяченко	79	П. Яременко	84	83	-1
М. Левченко	77	В. Коваль	85	Т. Кравець	82	85	3
П. Марчук	75	Ф. Лисенко	87	Р. Діденко	83	85	2
Р. Кравченко	80	М. Білик	63	Д. Ющук	87	92	5
Б. Сапожник	71	—	—	—	—		

Конкретний тест з множини можливих обирається дослідником в залежності від властивостей даних, які досліджуються:

- t-критерій Стьюдента для незалежних вибірок: коли дані приблизно нормально розподілені та дисперсії рівні або близькі;
- t-критерій з нерівними дисперсіями (критерій Уелча): коли дисперсії різні;
- непараметричний критерій Вілкоксона-Манна-Уітні: коли припущення нормальності розподілу не виконуються або дані є ранговими;

– ……будь який інший тест, який відповідає особливостям даних типу тестів (в дійсності їх є досить велика кількість і сам вибір найбільш адекватного тесту – окрема серйозна задача будь-якого дослідження).

В разі тесту з залежними вибірками нульова гіпотеза найчастіше стосується середнього значення різниць:

$H_0 : \mu_d = 0$

$H_a : \mu_d \neq 0$ в разі двосторонніх гіпотез або відповідні аналоги в разі використання односторонніх гіпотез.

де μ_d - середнє значення різниці між парними спостереженнями.

Основні методи перевірки гіпотез, які застосовуються при роботі з такими даними:

– Парний t-критерій Стьюдента - використовується для кількісних даних із приблизно нормальним розподілом різниць;
– Критерій Вілкоксона для парних вибірок - непараметричний тест, який не потребує нормальності;
– Знаковий критерій - якщо дані лише відображають напрямок змін, але не їх величину;
– …… будь який інший тест, який відповідає особливостям даних типу тестів.

3.3.1.5. Двовибірковий t-критерій Стьюдента для незалежних вибірок

Двовибірковий t-критерій (або t-тест) Стьюдента (англ. *Independent Samples Test*) *для незалежних вибірок* є одним із найбільш поширених статистичних методів, який використовується для порівняння *середніх значень двох різних, не пов'язаних між собою груп (незалежних вибірок)*.

Основна мета цього тесту - визначити, чи існує статистично значуща різниця між математичними очікуваннями двох генеральних сукупностей μ_1 та μ_2, припускаючи, що в наявності лише по одній випадковій вибірці з кожної з цих популяцій. Самі значення μ_1 та μ_2 нам невідомі. Все що має дослідник - два різні набори даних, статистичні характеристики яких можна визначити. Наприклад, це можуть бути дані про зріст жителів різних країн. Чи дані трафіку,

який спостерігався минулого і цього тижня. Чи кількість обстрілів, що відбувалися в листопаді 2024 та в листопаді 2025 років. Чи однаковий процент людей, що захворіли на COVID в молодшій та старших вікових групах тощо. Питання прикладного спеціаліста – «*чи однакові дані в цих двох вибірках*»? Математичне формулювання, до якого зводимо задачу – «*чи з однієї і тієї самої генеральної сукупності ці дані отримані*»? Якщо з однієї - то можливо припустити, що статистично значущої різниці між вибірки не повинно бути. У випадку, якщо статистичний тест не дасть можливості підтвердити нульову гіпотезу однаковості вибірки, приймається висновок про те, що дані дійсно різняться між собою.

Пошук загальної відповіді на поставлене питання досить складне і вирішується з застосуванням деякого *критерію однорідності*. Звузимо загальну задачу і почнемо з спроби замінити загальне питання більш простим, а саме - *чи рівні математичні очікування тих генеральних сукупностей, з яких ми отримали свої вибірки.* Тоді виникає можливість застосування критерію значущості.

Розрахунок t-статистики для даного тесту залежить від результатів попередньої перевірки умови про рівність дисперсій (див. Розділ 3.3.3). Існує декілька різновидів t-критерія Стьюдента в залежності від інформації про об'єми вибірок та їх дисперсії. В найбільш загальному випадку, коли кількість елементів вибірок можуть різнитися, як і дисперсія в них, використовують модифікацію критерію, яку називають *t-критерієм Уелча* (на честь британського статистика Bernard Lewis Welch). До речі, саме цей варіант реалізований у вигляді функції *ttest_ind()* в бібліотеці scipy.stats – див.далі). Статистика за цим критерієм обчислюється за формулою:

$$t_p = \frac{\left|\overline{X}_1 - \overline{X}_2\right|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

де $\overline{x}_1$, $\overline{x}_2$ - середні значення вибірок, s_1, s_2 - оцінка дисперсії кожної з вибірок, n_1, n_2 - об'єм вибірок. Доведено, що ця статистика розподілена за t-розподілом Стьюдента. При цьому кількість ступенів свободи, як параметр цього розподілу, визначається досить складно:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{(n_1 - 1)} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{(n_2 - 1)}}$$

Довірчий інтервал для різниці маточікувань генеральних сукупностей, з яких отримані вибірки, визначається за допомогою розподілу Стьюдента:

$$\overline{X}_1 - \overline{X}_2 - t_{\frac{\alpha}{2},df}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \le \mu_1 - \mu_2 \le \overline{X}_1 - \overline{X}_2 + t_{\frac{\alpha}{2},df}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} .$$

Нагадаємо, що в відповідності до критерію, гіпотеза H_0 приймається в разі, якщо $|t_p| > t_{\frac{\alpha}{2},df}$ В разі, якщо $|t_p| \le t_{\frac{\alpha}{2},df}$ підстав для відхилення нульової гіпотези немає. Якщо розраховується *p_value*, то гіпотеза H_0 відхиляється у випадку, коли *p_value*$< \alpha$.

В *scipy.stats* існує спеціальна функція, що використовується для визначення двовибіркового t-тесту для незалежних вибірок:

statistic, p_vl= sc.ttest_ind(sample1, sample2, equal_var=”False”)

Перші два параметри цього методу приймають значення обох вибірок, що порівнюються. Третій параметр, *equal_var*, використовується для того, щоб вказати, що дисперсії вибірок не однакові. Якщо вказати значенням цього параметру «*True*» функція самостійно виконає оцінку дисперсії кожної з вибірок. Це може підвищити потужність методу але ускладнить обчислення.

3.3.1.6. Двовибірковий t-критерій Стьюдента для залежних вибірок

Двовибірковий t-критерій Стьюдента для залежних (*пов’язаних, парних*) ***вибірок*** є статистичним методом, що *використовується для порівняння середніх*

значень двох взаємопов'язаних вимірювань. Взаємозв'язок означає, що кожному значенню в першій вибірці відповідає конкретне значення у другій. Основна мета тесту - з'ясувати, чи відрізняється *середня різниця між двома вимірюваннями* статистично значуще від нуля. Двовибірковий t-критерій Стьюдента для залежних вибірок - важливий інструмент аналізу експериментальних даних. Він дає можливість виявляти навіть незначні зміни всередині групи, що робить його незамінним у психологічних, медичних, педагогічних та соціальних дослідженнях. Знання умов застосування та правил інтерпретації результатів забезпечує коректні і надійні висновки.

За процедурою застосування даного тесту виконуються наступні дії:

1. Обчислюються значення $d_i = x_{1,i} - x_{2,i}$
2. Визначаються середня різниця:

$$\bar{d} = \frac{1}{n}\sum_{k=1}^{n} d_i$$

3. Розраховується стандартне відхилення різниць:

$$s_d = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}\left(d_k - \bar{d}\right)^2}$$

4. Обчислюється t-статистика за формулою:

$$t_p = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

Ступені свободи при цьому визначаються як $df = n - 1$

Прийняття рішення відбувається так само, як і в попередніх процедурах, шляхом порівняння отриманого при розрахунках значеннях t_p та значення квантиля $t_{\frac{\alpha}{2},df}$, або порівняння значення *p_value* та рівня значущості α.

У *scipy.stats* при реалізація даного критерію використовується функція *.ttest_rel()*.

3.3.2. Перевірка статистичної гіпотези про частки (долі) у загальній сукупності

У практиці статистичного аналізу часто виникає потреба зробити висновок не відносно математичного очікування генеральної сукупності, а щодо ***частки*** **(**або ***долі*** **)** елементів у генеральній сукупності, які мають певну властивість. Наприклад, оцінити частку користувачів, задоволених якістю сервісу; частку дефектної продукції; частку студентів, які склали іспит; частку населення, що підтримує певного кандидата тощо. У контексті інформаційної безпеки, мережевого адміністрування, програмування та технічної діагностики частка певних ознак у великій генеральній сукупності часто є критичним показником якості, надійності або безпеки системи. Типові питання можуть бути сформульовані так: Чи є частка успішних фішингових атак P меншою за визначений поріг? Чи відповідає частка браку в партії виробів прийнятому стандарту якості? Чи перевищує частка пройдених тестів необхідний рівень покриття коду? Чи зросла частка неавторизованих пакетів порівняно з нормальним фоном? Тощо.

В загальному випадку під «*часткою*» розуміють *відношення кількості певних подій* (наприклад, збоїв, успішних атак, помилок коду) *до загальної кількості спроб або об'єктів.*

Оскільки ніколи неможливо охопити усі існуючи об’єкти, випадки, значення тощо, як завжди будемо розглядати вибірку з деякої генеральної сукупності, робити певні висновки та потім переносити отримані результати на всю генеральну сукупність. Зрозуміло, що при цьому кожен крок дослідження має бути статистично обґрунтованим, що і надає необхідний рівень впевненості в коректності отриманих кінцевих результатів.

Нехай за результатами вибірки обсягом n отримано вибіркову частку:

$$\hat{p} = \frac{x}{n}$$

де x - кількість об'єктів у вибірці, які мають цікаву для нас ознаку. Також часто використовується термін «*кількість успіхів*». Іноді це може приводити до певних непорозумінь. Наприклад, якщо ми розраховуємо частку мережевих сесій, в яких були виявлені загрози інформаційної безпеки, «кількістю успіхів» буде саме

кількість таких сесій з загрозою. Звучить дещо дивно.

Нульова гіпотеза формулюється як:

$$H_0 : P = P_0$$

тобто, істинна частка генеральної сукупності P дорівнює певному, заздалегідь визначеному значенню P_0.

Альтернативна гіпотеза при цьому може мати наступні визначення:

двостороння: $H_a : P \neq P_0$;

одностороння: $H_a : P > P_0$ або $H_a : P < P_0$.

Тест ґрунтується на *центральній граничній теоремі*, тому для його коректності потрібна наявність великого об'єму експериментальних даних. Але, як і зазвичай, точного визначення, що таке «великий об'єм даних» в даному випадку не існує. В різних джерелах наводять різні емпіричні умови, наприклад $nP_0 > 10$ та $n(1-P_0) > 10$, але одне загальне правило лишається незмінним – *чим більше надано даних тим точнішими можуть бути результати*.

При справедливості нульової гіпотезі вибіркова частка $\hat{p}$ має приблизно нормальний закон розподілу з математичним очікуванням P_0 та дисперсією:

$$\sigma_{\hat{p}} = \sqrt{\frac{P_0(1-P_0)}{n}}$$

Тому тестова статистика набуває вигляду:

$$z_p = \frac{\hat{p} - P_0}{\frac{P_0(1-P_0)}{n}}$$

Після обчислення z-статистики потрібно порівняти її значення з критичним значенням $z_{кр}$, яке залежить від обраного рівня значущості α. Якщо $|z_p| > |z_{кр}|$, або якщо *p_value*$<\alpha$, нульова гіпотеза відкидається. Це означає, що є достатньо

даних для того, щоб стверджувати, що істинна частка P відрізняється від P_0. Якщо $|z_p| \le |z_{кр}|$, або якщо $p\_value \ge \alpha$ нульова гіпотеза не відкидається. Це означає, що немає достатньо доказів, щоб стверджувати, що P відрізняється від P_0.

Розглянемо практичний приклад. У відділі кібербезпеки компанії встановлено норматив: при нормальній роботі мережі не більше $P_0 = 0.08$ усіх зафіксованих мережевих подій протягом доби класифікуються як потенційні атаки (сканування портів, підозрілі пакети, аномальні запити, DDoS-події низької інтенсивності тощо). Після підключення нового публічного сервісу до мережі адміністратор SOC хоче перевірити, чи не зросла частота ворожих активностей. За минулу добу SIEM-система зареєструвала n=2000 подій, з яких x=190 були віднесені до категорії потенційних атак. Вибіркова частка при цьому становить:

$$\hat{p} = \frac{190}{2000} = 0.095$$

Перевіримо нульову гіпотезу $H_0 = 0.08$ проти альтернативної *односторонньої* гіпотези $H_a > 0.08$, тобто перевіряємо, чи збільшилася частка атак після зміни мережевої архітектури. В даному випадку розрахунок статистики дає $z_p \approx 2.47$. Критичне значення для правостороннього тесту при $\alpha = 0.05$ дорівнює $z_{1-\alpha} = 1.645$. Оскільки 2.47>1.645 нульову гіпотезу відхиляємо. Це означає, що частка підозрілих мережевих подій статистично значуще зросла. Для аналітиків SOC це сигнал про потенційне:

- підвищення інтенсивності сканувань або підготовчих дій зловмисника,
- початок нової хвилі DDoS-атаки,
- появу вразливого вузла після оновлення інфраструктури,
- потребу у посиленому моніторингу або додаткових правилах кореляції в SIEM.

і потребує негайного запровадження відповідних протоколів захисту.

У екосистемі Python порівняння двох сукупностей за часткою ознаки проводять за допомогою функції *proportions_ztest()* бібліотеки

statsmodels.stats.proportion:

```
from statsmodels.stats.proportion import proportions_ztest
x = [45, 52]   # успішні випадки в обох вибірках
n = [100, 120] # обсяги вибірок
z_stat, p_value = proportions_ztest(count=x, nobs=n)
print("z-статистика:", z_stat)
print("p_value:", p_value)
```

3.3.3. Перевірка гіпотези про рівність дисперсій за допомогою F-критерію Фішера

У статистичному аналізі важливо не лише оцінити середні значення двох вибірок, а й зрозуміти, чи відрізняється ступінь варіативності цих вибірок. Рівність дисперсій є ключовою умовою для багатьох параметричних тестів (наприклад, t-тесту для незалежних вибірок). Для перевірки такої умови використовується ***F-критерій Фішера***, також відомий як *F-тест на рівність дисперсій*. Метод був запропонований *Рональдом А. Фішером* (*Ronald Aylmer Fisher*)- английским статистиком, биологом-эволюционистом, генетиком, одним із засновників сучасної математичної статистики - і широко застосовується в медичних дослідженнях, соціальних науках, інженерії та сучасному машинному навчанні.

Мета F-тесту - перевірити нульову гіпотезу:

$$H_0 : \sigma_1 = \sigma_2$$

проти альтернативної:

$$H_a : \sigma_1 \neq \sigma_2$$

Або в односторонньому варіанті:

$$H_a : \sigma_1 > \sigma_2 \text{ або } H_a : \sigma_1 < \sigma_2$$

Тест ґрунтується на тому, що відношення двох незалежних оцінок дисперсій в разі приналежності даних вибірки одній генеральній сукупності має розподіл Фішера.

F-критерій є чутливим до порушень припущень, тому важливо

дотримуватися умов:

- обидві вибірки походять з нормально розподілених генеральних сукупностей;
- вибірки є незалежними.

Параметричний характер задачі: ознака виміряна в одній з сильних (числових) шкал.

Порушення нормальності може суттєво спотворювати результат F-тесту, тому інколи рекомендують альтернативні непараметричні методи (наприклад, критерій Левене).

Нагадаємо, що *вибіркові дисперсії* - тобто отримані в експерименті *статистичні оцінки дисперсії генеральної сукупності* - розраховуються за наступною формулою:

$$s_\gamma^2 = \frac{1}{n_\gamma - 1}\sum_{i=1}^{n_\gamma}\left(x_i - \overline{x}_\gamma\right)^2$$

де $\gamma = \{1,2\}$.

Статистика F-критерій Фішера визначається як:

$$F = \frac{s_{\max}^2}{s_{\min}^2}$$

де $s_{\max}^2 = \max\left(s_1^2, s_2^2\right), s_{\min}^2 = \min\left(s_1^2, s_2^2\right)$.

Використання найбільшої з двох дисперсії у чисельнику робить значення F ≥ 1, що спрощує порівняння з критичним значенням F-розподілу. Значення кількості ступенів свободи визначається за формулою:

$$df_1 = n_{max} - 1, df_2 = n_{min} - 1$$

Таким чином, алгоритм аналізу гіпотез на критерієм Фішера полягає в виконанні наступної послідовності дій:

1. Обчислити значення F-статистики F_p.

2. Визначити критичне значення F-розподілу:

$$F_{кр} = F_{1-\frac{\alpha}{2},df_1,df_2}$$

3. Якщо $F_p > F_{кр}$ то нульову гіпотезу відхиляємо: відмінність між дисперсіями є статистично значущою. Проте найпоширенішою на сьогоднішній день практикою є використання *p_value*, в знаходженні якого допомагають сучасні прикладні пакети статистичних розрахунків.

До інтерпретації результатів тесту - як і до результатів інших статистичних тестів - слід підходити дуже акуратно. Ми знаємо, що «прийняти H_0» в даному разі не означає, що дисперсії рівні, а лише вказує на відсутність статистичних доказів їх відмінності. *F-тест* може хибно сигналізувати про різницю, якщо розподіл не є нормальним. Для великих вибірок навіть невелика невідповідність дисперсій може стати статистично значущою. Не рекомендується також використовувати критерій Фішера якщо є підозра на наявність викидів у вибірках.

У Python немає окремої функції «готового F-тесту для дисперсій», проте всі необхідні елементи легко доступні: обчислення дисперсії (*numpy.var*), квантилі F-розподілу (*scipy.stats.f*), *p_value* через відповідні функції F-розподілу.

Окрім критерія Фішера існують і інші критерії, що використовуються для аналізу однорідності дисперсій вибірок, як то критерій Левене (стійкий до відхилень від нормальності), критерій Брауна–Форсайта, критерій Бартлетта. В пакеті *scipy.stats* реалізовано багато з них, однак їх використання вимагає розуміння особливостей, недоліків та обмежень кожного з них.

Для поглиблення розуміння теорії перевірки гіпотез існує дуже багато можливостей та джерел. Наведемо лише деякі з них [48], [49], [50], [51].

3.3.4. Непараметричні методи перевірки гіпотез. Ранговий критерій Вілкоксона-Манна-Уітні

Ранговий критерій Вілкоксона-Манна-Уітні (англ.*Wilcoxon–Mann–Whitney Test, WMW*) представляє собою непараметричний аналог t-критерію Стьюдента.

Цей метод виявлення відмінностей між вибірками був запропонований *Френком Вілкоксоном (F. Wilcoxon)*. У 1947 році він був істотно перероблений і розширений *Х. Б. Манном (H. B. Mann)* і *Д. Р. Уітні (D. R. Whitney)*, по іменах яких сьогодні зазвичай і називається.

(Засторога: Існує ще один тест, що носить назву одновибірковий тест Вілкоксона для знакової різниці (Wilcoxon Signed-Rank Test або Wilcoxon T-Test)- не треба змішувати та путати ці тести, однак майте на увазі, що в деяких джерелах тест Вілкоксона-Манна- Уітні іноді називають просто тестом Манна- Уітні, або U-тестом Манна- Уітні).

У сфері кібербезпеки, інших галузях практичної діяльності дослідник рідко має справу з «ідеальними» даними. Так, аналізуючи мережевий трафік, час реакції системи виявлення вторгнень (IDS) або ефективність різних алгоритмів шифрування, часто доводиться працювати з вибірками, які не підпорядковуються нормальному розподілу. Наявність викидів, асиметрія даних та малий обсяг вибірок роблять класичні параметричні методи (наприклад, t-критерій Стьюдента) непридатними або такими, що дають хибні результати. Ранговий критерій Вілкоксона-Манна-Уітні - непараметричний статистичний інструмент (собою непараметричний аналог t-критерію Стьюдента), який дозволяє порівнювати дві незалежні вибірки, не вимагаючи припущень про нормальність розподілу даних. Замість самих значень він оперує їхніми рангами, що робить його стійким до аномалій - критично важлива властивість при роботі з реальними логами безпеки або телеметрією.

(Засторога: Існує ще один тест, що носить назву одновибірковий тест Вілкоксона для знакової різниці (Wilcoxon Signed-Rank Test або Wilcoxon T-Test)- не треба змішувати та путати ці тести, однак майте на увазі, що в деяких джерелах тест Вілкоксона-Манна- Уітні іноді називають просто тестом Манна- Уітні, або U-тестом Манна- Уітні).

Ідея методу базується на тому очевидному припущенні, що якщо вибірки, які досліджуються, взяті з однієї генеральної сукупності даних (тобто підпорядковуються одному закону розподілу), то елементи як першої, так і другої вибірок *в спільному ряду об'єднаної вибірки* будуть перемішані рівномірно. При цьому гіпотези формулюється наступним чином:

$$H_0 : P(y_{1,*} < y_{2,*}) = 0.5$$
$$H_a : P(y_{1,*} < y_{2,*}) \neq 0.5$$

Іншими словами, якщо взяти по одному довільному елементу з кожної з вибірок, то ймовірність того, який з них буде більшим - однакові (50%). Зрозуміло, що якщо це не так, тобто статистично значимо частіше будуть виявлятися елементи однієї з цих вибірок, то можна говорити, що значення вибірок «зсунуті» між собою. Це стає зрозумілим з Рис. 3.5.

За своєю суттю критерій Вілкоксона-Манна-Уітні оцінює, наскільки тісно перемішані значення двох вибірок у спільному ранжованому ряду: чим менше вони перекриваються (тобто чим частіше значення однієї вибірки перевищують значення іншої), тим сильніше підстави вважати, що вибірки походять із різних – щонайменше з різними математичними очікуваннями - генеральних сукупностей. Однак простий перетин інтервалів зміни значень вибірок, який здається що може слугувати відповіддю на поставлене завдання, в дійсності не призводить до коректного рішення. З рис 3.6 видно, що навіть якщо області значення перекриваються майже повністю, розподіли суттєво відрізняються один від одного. Отже потрібно запропонувати міру, яка б враховувала обидва фактору – і перетин області значень, в безпосередню форму функції щільності розподілу.

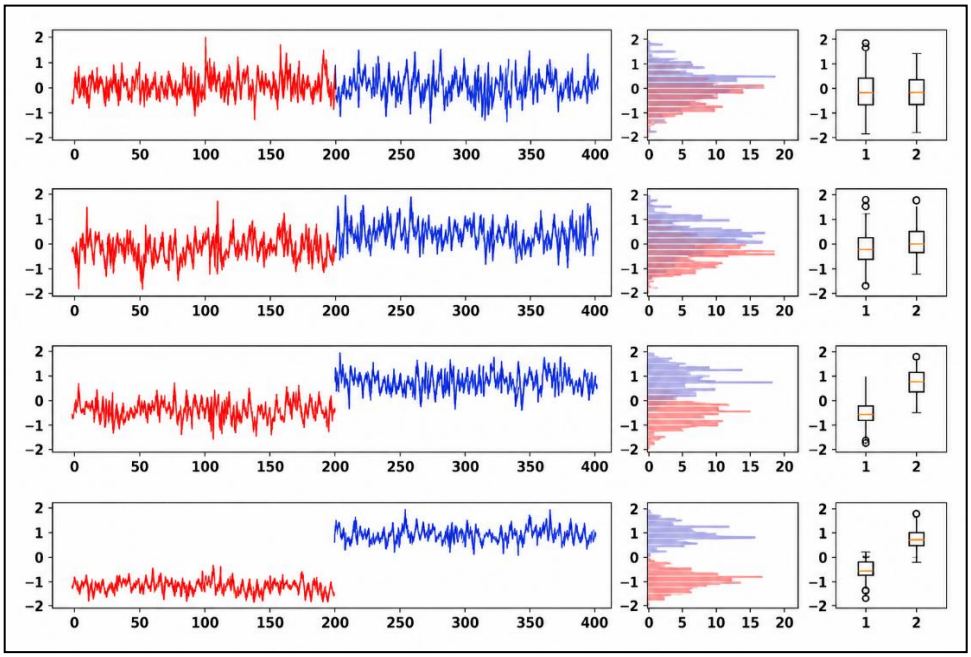


Рис.3.5 Вибірки та взаєморозташування їх гістограм при наявності зсуву в розподілах.

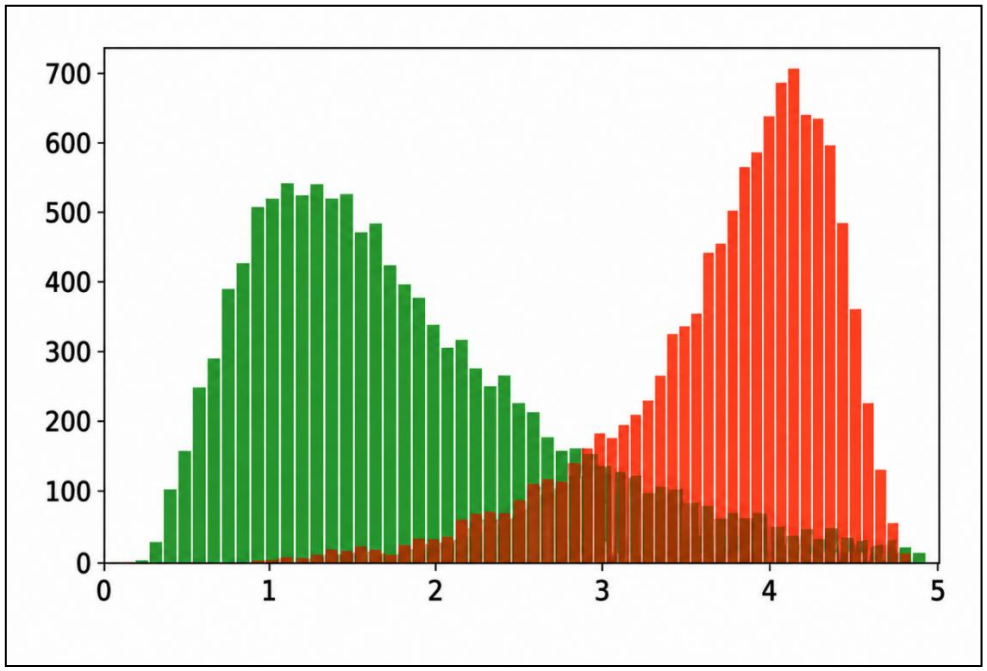


Рис.3.6 Приклад вибірок з рівними областями можливих значень та різними розподілами.

Якщо нульова гіпотеза істинна (вибірки з одного розподілу), то при об'єднанні та сортуванні всіх значень ранги обох груп будуть розподілені майже рівномірно - значення будуть «перемішані» як карти в колоді. Суми рангів для кожної вибірки будуть приблизно однаковими. Якщо істинна альтернативна гіпотеза (одна вибірка систематично «більша»), то значення однієї групи зосередяться в одному кінці ранжованого ряду, а іншої - в іншому. Це призводить до суттєвої різниці в сумах рангів, що й фіксує статистика критерію.

Існують два варіанти розрахунку статистики критерію, що відрізняються обчислювальною складністю.

Для випадків, коли вибірки малі ($n_1, n_2 < 20$), доречніше використовувати метод розрахунку цього критерію (*який, до речі і запропонували Манн та Уітні. Подекуди цей різновид називають U -критерієм Манна-Уітні, хоча в дійсності це два окремих випадки одного методу*), і який полягає в безпосередньому порівнянні всіх можливих пар елементів двох вибірок між собою:

$$U = \sum_{y_i \in Y_1} \sum_{y_j \in Y_2} h_{i,j}$$

$$h_{i,j} = \begin{cases} 1, \textit{ якщо } y_i < y_j; \\ \dfrac{1}{2}, \textit{якщо } y_i = y_j; \\ 0, \textit{ якщо } y_i > y_j. \end{cases}$$

По суті за цією формулою визначається кількість пар з вибірок, для яких $y_i < y_j$.

Після того, як значення критерію U отримано, його порівнюють з критичним значенням $U_{кр}$ для обраного рівня значущості або обчислюють *p_value*, щоб прийняти рішення щодо нульової гіпотези: якщо $p < \alpha$ (або $U \le U_{кр}$), нульову гіпотезу відхиляють, роблячи висновок про статистично значущу різницю між вибірками; в іншому разі - відмінності вважаються випадковими.

Критичні значення $U_{кр}$ можна знайти в відповідних, наперед складених таблицях, в яких розраховані значення критичних інтервалів статистики U в залежності від значень n_1, n_2 та обраного значення α. Таблиці створюють, користуючись методами комбінаторики, що можливо при невеликих об'ємах вибірки.

Нехай загальний обсяг об'єднаної вибірки $N = n_1 + n_2$. Отже, маємо N рангів (від 1 до N). Загальна кількість способів розподілити ці ранги між двома групами (обрати n_1 рангів для першої групи з N можливих) визначається біноміальним коефіцієнтом:

$$C_N^{n_1} = \frac{N!}{n_1! * n_2!}$$

Для кожної можливої комбінації можна розрахувати значення U, а далі - підрахувати, скільки разів зустрічається кожне можливе значення U (від 0 до $n_1 * n_2$). Ймовірність того, що U прийме певне конкретне значення k дорівнює

$$P(U=k)=\frac{\Omega(k)}{C_N^{n_1}}$$

де $\Omega(n_1,k)$ - кількість таких підмножин рангів розміру n_1, сума елементів яких (або відповідна статистика U) дорівнює конкретному значенню k. (Отримання аналітичного виразу для $\Omega(n_1,k)$ можна знайти в підручниках комбінаторики, хоча він і не зовсім тривіальний). Знаючи цей розподіл і задавши рівень значущості можна визначити значення кордонів критичного інтервалу.

Критичне значення $U^{\min}(n_1,n_2,\alpha)$ для заданого рівня значущості α - це максимальне значення U для якого кумулятивна ймовірність (сума ймовірностей всіх значень менших або рівних йому) не перевищує α.

В зазначених таблицях часто вказують лише нижні значення кордону інтервалу $U^{\min}(n_{1,},n_2,\alpha)$, а враховуючи, що загальна кількість пар значень, які порівнюються, дорівнює $n_1 * n_2$, верхній кордон може бути розрахований як:

$$U^{\max}(n_{1,},n_2,\alpha)=n_{1,} * n_2 - U^{\min}(n_{1,},n_2,\alpha).$$

Таблиці для визначення критичних значень для статистики U можна знайти в будь якому підручнику з матстатистики, або в мережі інтернет, наприклад - за посиланням [52].

Отже, інтервал прийняття H_0 гіпотези визначаються як:

$$[U^{\min}(n_{1,},n_2,\alpha),U^{\max}(n_{1,},n_2,\alpha)]$$

Якщо отримане значення U виходить за зазначені межі інтервалу, то приймається статистична значимість відмінностей між рівнями ознаки в аналізованих вибірках (приймається альтернативна гіпотеза). Достовірність відмінностей тим вище, чим менше значення U.

Алгоритм точного підрахунку $\Omega(n_1,k)$ має обчислювальну складність $O(N * n_1 * max\_sum)$, де max_sum - це максимально можлива сума рангів, яку може набрати перша вибірка розміру n_1 з загальної кількості рангів N. Така

складність робить пряме обчислення цього значення практично неможливим. Крім того, визначення *p_value* отриманої статистики також ґрунтується на методах комбінаторики, а отже при великих вибірках також є неприпустимо складним. (*Просто для довідки. Кількість комбінацій, які треба перебрати при розгляді ймовірності отримання результату може збільшуватися дуже швидко. Наприклад, при 2 вибірках розміром 15 елементів кожна, всього існує 155117520 можливих комбінацій розподілів елементів по об'єднаному ряду, перебор яких навіть на сучасних комп'ютерах не дуже швидкий. Заради повноти картини додаймо, що при сильно незбалансованих вибірках ситуація дещо покращується. Так, якщо в одній вибірці присутні 35 елементів, а в іншій 5, то існує лише(!) 658 008 можливих комбінацій. Не дуже обнадійливо).*

Тому коли вибірки є досить великими ($n_1, n_2 \geq 20$) розрахунок статистики критерію визначають іншим шляхом (*Цей варіант критерію був запропонований Вілкоксоном і подекуди називається «ранговим критерієм Вілкоксона*). Для цього дві вибірки об'єднуються та впорядковуються разом. Стратегія полягає в тому, щоб визначити, чи значення з двох вибірок випадковим чином змішані в порядку рангу (якщо зсуву нема), чи вони згруповані на протилежних кінцях при об'єднаного ряду (якщо зсув є). (Див. Рис.3.5). Для цього в об'єднаній вибірці визначають суму рангів, які мають в ній елементи кожної з вибірок:

$$R_1 = \sum_{y_i \in Y_1} r_i \text{ та } R_2 = \sum_{y_i \in Y_2} r_i$$

де r_i - ранг елементу y_i в об'єднаній вибірці. В разі рівних значень у кількох елементів, кожній з них приписується середнє арифметичне послідовних значень рангів. (Наприклад, для ряду значень [3, 5, 7, 7, 8] приписані ранги будуть виглядати так: [1, 2, 3.5, 3.5, 5]).

Визначимо, в яких кордонах $\left[R^{\min}, R^{\max}\right]$ може коливатися значення R для кожного ряду.

Якщо ряд (хай для визначеності це буде ряд 1, з кількістю елементів в ньому n_1) займає підряд найменші позиції в об'єднаному ряду (На картинці вище цей випадок відображений на останньому малюнку). Тоді для нього значення $R_1^{\min}$ буде визначатися відомою формулою суми послідовності чисел:

$$R_1^{\min} = \frac{n_1(n_1+1)}{2}.$$

Якщо елементи вибірки займають всі найбільші позиції загального рейтингу, то максимальним значенням суми рангів $R_1^{\max}$ буде

$$R_1^{\max} = n_1 * n_2 + \frac{n_1(n_1+1)}{2}.$$

При отриманні останнього співвідношення використовують той факт, що сума членів арифметичної прогресії дорівнює $\frac{(a_1 + a_n) * n}{2}$; де a_1 - перший член, в нашому випадку $a_1 = n_2 + 1$; a_n - останній член, в нашому випадку $a_n = n_1 + n_2$, де n – кількість членів в ряду, в нашому випадку $n = n_1$.

Введемо статистики:

$$U_1 = R_1^{\max} - R_1$$
$$U_2 = R_2^{\max} - R_2$$

які за суттю визначають, наскільки відрізняються розраховані значення сум рядів від максимально можливих для кожної з вибірок з врахуванням кількості елементів в кожній з них. Саме ці співвідношення і дають змогу говорити про те, що обидва розглянуті варіанти розрахунку статистики критерія лінійно пов’язані між собою.

Зрозуміло, що беззастережно до способу розрахунку виконується співвідношення: $0 \le U_i \le n_1 * n_2$, а також, що $U_1 + U_2 = n_1 * n_2$. Останнє співвідношення корисно використовувати для перевірки правильності обчислень. Крім того, останнє співвідношення дає можливість в якості «міри ступеня перемішування» елементів вибірок в критерії Вілкоксона-Манна-Уітні використовувати будь-яку з статистик U_1 чи U_2. Зазвичай беруть ту статистику, значення якої найменше: $U = \min(U_1, U_2)$.

При рівномірному перемішуванні елементів вибіркові значення статистики U будуть близькими до їх середнього по діапазону зміни значень, а саме $M(U) = \frac{n_1 * n_2}{2}$.

Можна також показати, що дисперсія цієї статистики є

$$D(U) = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

Як було вже сказано, при малих значеннях n_1, n_2, параметр *p_value* може бути вирахуваний методами комбінаторного аналізу. Коли ж обидві вибірки є великими ($n_1, n_2 \geq 20$) і нульова гіпотеза вірна, *розподіл U-статистики має тенденцію наближатися до нормального розподілу*. В цьому випадку гіпотези можуть бути оцінені з використанням стандартного *z-перетворення* (нагадаю, що z-перетворення, яким ми вже користувалися багато разів, це перетворення типу $z_i = \frac{x_i - \mu}{\sigma}$, див. розділ 2.4.2.2) і, відповідно, стандартного нормального розподілу з нульовим маточікуванням та дисперсією, що дорівнює 1:

$$Z(U) = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

Отже, розрахувавши цю величину можна в подальшому скористатися вказаним розподілом для визначення кордонів критичного інтервалу та/або *p_value* даного критерію.

В бібліотеці *scipy.stats* реалізація методу Вілкоксона-Манні-Уітні виконується за допомогою функції

scipy.stats.mannwhitneyu(arr1, arr2)

яка приймає в якості параметрів значення двох вибірок, та повертає кортеж з розрахованого значення статистики U та ввідповідного значення *p_value*.

Цей метод окрім інших параметрів, які описані в документації [53] , містить параметр *method*, який може примати значення «*auto*» (використовується по замовчуванню), «a*symptotic*» та «*exact*». Використання значення «*asymptotic*» призводить до використання описаних вище z-перетворення та статистики стандартизованого нормального розподілу для визначення *p_value*. При використанні «*exact*» використовуються точний розподіл, що визначається з використанням комбінаторики. Нарешті використання значення «*auto*» призводить до самостійного визначення функцією варіанту обчислення *p_value* виходячи з кількості елементів вибірок, наданих на вхід функції.

Таким чином, критерій Вілкоксона-Манна-Уітні - це простий, досить надійний і потужний спосіб порівняти дві групи даних, коли вони мають «незручний» розподіл, викиди або малий обсяг. Він використовується в найрізноманітніших прикладних задачах: від A/B-тестування інтерфейсів та бенчмаркінгу продуктивності баз даних до аналізу часу відгуку серверів, обробки біомедичних сигналів або оцінки результатів юзабіліті-тестувань. Головне - пам'ятати, що *метод працює з рангами, а не кількісними значеннями*, і дає змогу робити висновки на основі *p_value*, навіть коли дані далекі від ідеальної статистики.

Додаткову інформацію про критерій Вілкоксона-Манна-Уітні можливо знайти, наприклад, за посиланнями [54].

3.3.5. χ^2-статистика Пірсона ті її використання при перевірці гіпотез

У системі статистичного аналізу особливе місце посідають *непараметричні методи* (див. розділ 3.2.6), які дозволяють робити висновки про властивості популяції без жорстких припущень щодо нормальності розподілу ознак. Одним із найбільш універсальних та часто вживаних інструментів цього класу є ***критерій χ2 Пірсона*** (англ. *Pearson's Chi-Squared Test)* Він базується на порівнянні фактично отриманих даних (*емпіричних* або *спостережуваних*) із *очікуваними* (зокрема - *теоретичними*), тобто тими, які можна було б очікувати теоретично за умови істинності певної гіпотези. На відміну від параметричних тестів, *критерій χ2 аналізує* не середні значення чи дисперсії, а *частоти появи*

певних ознак у різних групах. В узагальненому вигляді статистика χ^2-критерію визначається формулою:

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(O_i - E_i\right)^2}{E_i}$$

де O_i - частота появи значення і-ї категорії, що була зафіксована в дослідженні;

E_i - очікувана частота тієї ж категорії, обчислена відповідно до обраної нульової гіпотези;

k- кількість категорій.

Якщо нульова гіпотеза істинна і виконуються умови застосування критерію, то статистика χ^2 має χ^2-розподіл з певною кількістю ступенів свободи, що дає змогу оцінити статистичну значущість відхилень між спостережуваними та очікуваними частотами. Чим більшим є отримане значення статистики χ^2, тим суттєвішою є розбіжність між фактами та теорією, і тим менш імовірною є випадковість виявлених відмінностей.

(*Термінологічне застереження.* Величини та логічніше було б називати "кількістю" або "абсолютною частотою", бо вони мають розмірність «кількість спостережень», не нормовані на обсяг вибірки, і їх значення не лежать у межах інтервалу [0,1]. Тобто з точки зору «суворої» термінології O та E - це «абсолютні кількості», а не частоти. Проте термін "частота" є загальноприйнятим, традиційним скороченням у статистичній науці в контексті таблиць спряженості (див. далі) для позначення саме кількості випадків у категорії. Зверніть на це увагу, бо іноді це може призводити до певних непорозумінь).

Особлива увага приділяється обсягу вибірки та величині очікуваних частот. Рекомендується, щоб загальна кількість спостережень була не меншою за 20, а в ідеалі - перевищувала 50. Критично важливою є умова «правила п’ятірки»: частоти O_i та E_i для кожного i повинні бути не меншими за 5. Якщо очікувані числа дуже малі (менше 5), апроксимація розподілом хі-квадрат стає неточною. У таких випадках слід об’єднувати суміжні категорії або використовувати альтернативні методи.

Узагальнена постановка χ^2-критерію Пірсона охоплює кілька важливих статистичних задач, зокрема:

- з'ясування наявності статистично значущого зв'язку між двома категоріальними змінними (критерій незалежності);
- перевірку узгодженості емпіричного розподілу з теоретичним (критерій узгодженості);
- перевірку однорідності розподілів у кількох вибірках (критерій однорідності).

У всіх цих випадках використовується одна й та сама статистична конструкція, яка базується на *порівнянні спостережуваних і очікуваних частот та оцінює величину їх відхилення від передбачень нульової гіпотези*.

3.3.5.1. χ^2-критерій незалежності Пірсона

χ^2-критерій незалежності Пірсона (англ. *Pearson's Chi-Squared Test for Independence*) використовується для перевірки нульової гіпотези H_0 про те, що між двома *категоріальними (номінальними)* змінними немає статистично значущого зв'язку (тобто вони є незалежними). χ^2-критерій широко застосовується в соціальних дослідженнях, медицині, маркетингу, аналізі інформаційних систем, кібербезпеці та інших галузях, де результати спостережень можуть бути подані *у вигляді частот виникнення тих чи інших подій*.

При застосуванні χ^2-критерію незалежності формулюються такі гіпотези:

- Нульова гіпотеза H_0: досліджувані змінні є статистично незалежними;
- Альтернативна гіпотеза H_a: між змінними існує статистична залежність.

Вхідними даними для застосування критерію є ***таблиця спряженості*** (англ. *Contingency Table*; інший можливий переклад - *таблиця зв'язаності*). Таблицею спряженості називається прямокутна таблиця розміру $r \times c$, де r - кількість можливих категорій (значень, що виміряні в номінальній шкалі, див. розділ 2.1.3.1) першої змінної; c - кількість можливих категорій другої змінної. Кожен об'єкт сукупності потрапляє в одну з клітин цієї таблиці відповідно до того, до якої категорії він відноситься по кожній з двох ознак. Після заповнення

таблиці кожен її елемент $O_{i,j}$ відображає кількість спостережень, *що одночасно належать до і-ї категорії першої змінної та j-ї категорії другої змінної.* Наприклад, число людей, що належать конкретній соціальній групі і при цьому хворіють деякою хворобою.

Як приклад, розглянемо залежність між типом інциденту інформаційної безпеки та результатом його обробки в SOC. В таблиці 3.2 наведені дані, що були отримані в результаті спостережень.

Таблиця 3.2 Таблиця спряженості для спостережень

Тип інциденту \ Результат	Підтверджений	Хибне спрацювання	Разом
Фішинг	45	15	**60**
Шкідливе ПЗ	70	10	**80**
Несанкціонований доступ	55	5	**60**
Разом	**170**	**30**	**200**

В наведеному прикладі побудова такої таблиці спряженості в змістовному аспекті дозволяє:

- виявити можливу залежність між типом інциденту та якістю детекції;
- оцінити, для яких категорій подій частіше виникають хибні спрацювання;
- сформулювати нульову гіпотезу про незалежність змінних і перевірити її.

Так, з даних таблиці прикладу можна побачити, що для фішингових подій частка хибних спрацювань є вищою, ніж для інцидентів типу несанкціонованого доступу. В загальному випадку кількість рядків та стовпчиків таблиці спряженості можуть бути довільними. Наприклад, якщо цікавляться питанням залежності стану погоди в послідовні дні, то кількість рядків і стовпчиків таблиці буде дорівнювати кількості станів погоди, які здатні бути розпізнані при спостереженні («*спекотно*» , «*сонячно*», «*мінлива хмарність*», «*хмарно*», «*дуже хмарно*», «*дощить*», «*злива*», «*ураган*» тощо). Якщо цікавляться подією

«*складання заліку*» та умовою «*курсант самостійно виконував домашні завдання*», таблиця буде мати по два рядки та стовпчика.

Як згадувалося вище, другий необхідний компонент для розрахунку χ^2-критерію є набір значень, які б мали мати значення елементів таблиці спряженості, якби зв'язку між значеннями ознак не було. Тобто, ***якби*** *нульова гіпотеза про незалежність була правдивою*.

Якщо змінні дійсно незалежні одна від одної, очікувана частота спільного потрапляння в клітинку з координатами (i, j) дорівнює добутку:

$$E_{i,j} = \frac{P(X_i) \times P(Y_j)}{N}$$

де $P(X_i)$ - кількість зафіксованих випадків появи значення X_i серед всіх значень вибірки (в таблиці спряженості - останній стовпчик), $P(Y_j)$ - кількість зафіксованих випадків появи значення Y_j серед всіх значень вибірки (в таблиці спряженості - останній рядок), N - загальний об'єм вибірки (в таблиці спряженості - права нижня комірка таблиці). Для прикладу аналізу залежності між типом інциденту інформаційної безпеки та результатом його обробки в результаті отримуємо значення, наведені в таблиці 3.3.

При аналізі таблиць спряженості статистика χ^2 - це міра загальної різниці між спостережуваними та очікуваними частотами у всіх клітинках таблиці. Чим більша ця різниця, тим сильніші докази проти гіпотези H_0. Формула для обчислення статистики χ^2 *в даній постановці* виглядає наступним чином:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(O_{i,j} - E_{i,j}\right)^2}{E_{i,j}}$$

Чим більше значення χ^2, тим сильніше відхилення від гіпотези незалежності. Число ступенів свободи для χ^2-критерію незалежності визначається за формулою:

$$df = (r-1)(c-1)$$

Таблиця 3.3 Таблиця спряженості для очікуваних значень

Тип інциденту \ Результат	Підтверджений	Хибне спрацювання	Разом
Фітинг	51	9	**60**
Шкідливе ПЗ	68	12	**80**
Несанкціонований доступ	51	9	**60**
Разом	**170**	**30**	**200**

Після обчислення статистики критерію можливі два підходи щодо формулювання висновків відносно гіпотези H_0:

– За критичним значенням:

Якщо $\chi^2_p > \chi^2_{кр}$ нульову гіпотезу відхиляють.

– За значенням p_value:

Якщо $p\_value < \alpha$ нульову гіпотезу відхиляють.

Критерій незалежності *Пірсона дозволяє лише встановити наявність або відсутність зв'язку між ознаками, проте не вимірює цей зв'язок, тобто не є мірою зв'язку* (як, наприклад, кореляція, див. Розділ 4). Статистика χ^2 прямо залежить від обсягу вибірки. Для оцінки сили зв'язку (розміру ефекту) беззастережно до цього показника розраховують додаткові коефіцієнти. Серед них найбільш вживаним є *коефіцієнт V Крамера* (англ: *Cramer's V*):

$$V = \sqrt{\frac{\chi^2}{\mathrm{N}^*\min(r-1, c-1)}}$$

На відміну від самого критерію χ^2, який лише відповідає на запитання «Чи є зв'язок?» (так/ні), коефіцієнт *V* показує, на скільки сильний цей зв'язок (в шкалі 0 до 1, де 0 – це повна незалежність, а 1 – повний детермінований зв'язок, тобто, одна змінна однозначно визначає іншу).

Зауваження. Інтерпретація коефіцієнту V залежить від предметної області. У соціальних науках V=0.3 може вважатися сильним ефектом, тоді як у технічних системах (наприклад, детекція аномалій) часто очікують значень, що перевищують 0.7 для надійного прогнозу.

3.3.5.2. Інші випадки використання χ^2-статистики Пірсона.

χ^2-статистика Пірсона є універсальною, але залежно від того, *що саме порівнюється*, вона використовується:

- як критерій незалежності між двома номінальними змінними (див. розділ 3.3.5.1);
- як критерій узгодженості між теоретичним та емпіричним розподілом;
- як критерій однорідності.

Критерій узгодженості χ^2 Пірсона (англ. *Pearson's Chi-Squared Goodness-of-Fit Test*) - це статистичний інструмент, що дозволяє відповісти на запитання: «Чи можна вважати, що емпіричні дані походять із певного теоретичного, наперед очікуваного розподілу?». Система гіпотез формулюється наступним чином:

H_0: Емпіричний розподіл даних не відрізняється від теоретичного (наприклад, дані підпорядковуються рівномірному або нормальному розподілу).

H_a: Емпіричний розподіл статистично значуще відрізняється від теоретичного.

На відміну від критерію незалежності, де порівнюються дві змінні між собою, тут виконується порівняння спостережуваної вибірки з очікуваною моделлю. Це фундаментальний метод для валідації статистичних припущень, на яких будуються багато алгоритмів машинного навчання та системи детекції аномалій.

Для обчислення статистики критерію необхідно визначити, з яким саме ймовірносним розподілом буде відбуватися порівняння емпіричних даних (рівномірний, Пуассона, нормальний тощо). Якщо дані неперервні, множина можливих значень розбивається на рівномірні інтервали (біни). Очікувані частоти розраховуються як:

$$E_i = Np_i$$

де p_i - ймовірність потрапляння в інтервал *i* згідно з теорією.

Емпіричні значення O_i визначаються як дані відповідної гістограми - *частина даних вибірки, що досліджується, які потрапили в інтервал i* .

У задачах кібербезпеки цей критерій часто використовується як фільтр "першого ешелону". Наприклад, перед запуском «важкої» та ресурсоємної моделі машинного навчання можна швидко відсіяти підозрілі потоки даних, які не проходять тест на узгодженість з профілем "нормального" (не аномального) трафіку.

На рис.3.7 показано використання χ^2-статистики *як критерію узгодженості* на прикладі порівняння теоретичного закону розподілу і гістограми, отриманої на основі емпіричних даних. З малюнку стає зрозумілим, що чим менше відрізняються емпіричні і теоретичні частоти (площини під кривою теоретичної функції щільності розподілу та площі відповідного прямокутника гістограми), тим менше буде величина χ^2-статистики, і, отже, вона буде характеризувати близькість емпіричного, тобто того, який ми досліджуємо, та теоретичного розподілів.

Як ***критерій однорідності*** (англ. *Test of Homogeneity*) χ^2-статистика може використовується для перевірки гіпотези про те, що кілька вибірок мають однаковий розподіл значень, інакше кажучи, перевіряється, чи є групи однорідними за певною номінальною ознакою. Говорячи мовою статистики, чи можна вважати, що вибірки отримані з однієї генеральними сукупності.

Математично цей різновид критерію ідентичний тесту на незалежність, що був розглянутий раніше. Проте він застосовується, коли дані надходять у вигляді кількох незалежних вибірок. Прикладом задач, в якій доцільно застосувати цей критерій, може бути задача перевірки, чи однакова частота виникнення певного захворювання у різних етнічних групах; задача порівняння розподілу параметрів мережевого трафіку при різних типах атак; перевірка того, чи однаково розподіляються типи помилок або вразливостей, виявлених у різних версіях програмного забезпечення; оцінка того, чи є розподіл дефектів продукції однаковим для різних виробничих ліній або постачальників; порівняння розподілу доходів населення у різних регіонах або містах тощо. При аналізу задач такого типу в таблиці спряженості кожен рядок відповідає одній вибірці (точніше – розподілу значень цієї вибірки), а стовпчики – категоріям, або інтервалам, можливих значень ознаки.

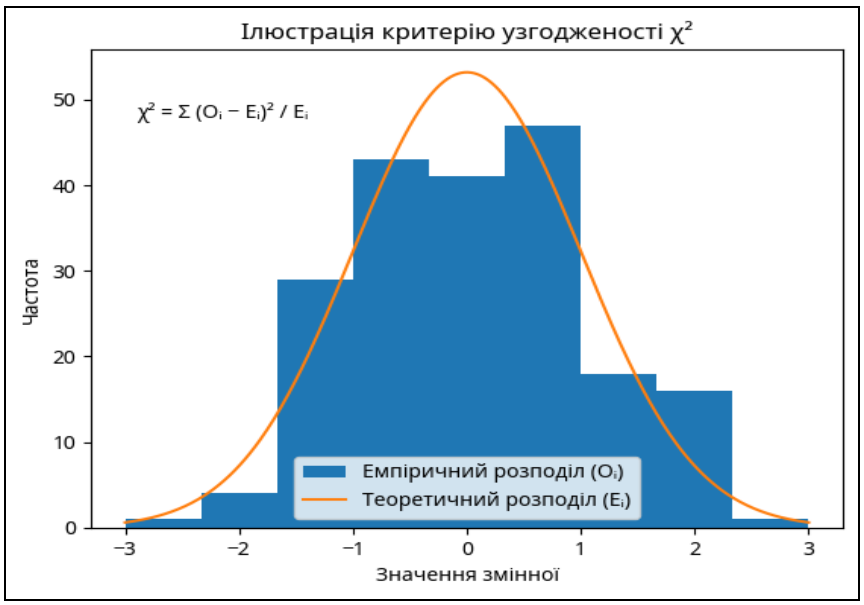


Рис. 3.7 Приклад порівняння емпіричного та теоретичного розподілу за допомогою гістограми.

Різниця між розглянутим вище критерієм незалежності на основі χ^2-статистики і критерієм однорідності на основі χ^2-статистика полягає *не в обчисленнях, а в інтерпретації даних.* На практиці це означає, що для кожної клітинки таблиці спряженості визначаються спостережувані частоти та відповідні очікувані частоти, які обчислюються за умови справедливості нульової гіпотези про однаковість розподілів. Далі за стандартною формулою обчислюється значення χ^2-статистики, яке порівнюється з критичним значенням для заданого рівня значущості та відповідної кількості ступенів свободи. Якщо отримане значення статистики перевищує критичне, нульову гіпотезу про однорідність вибірок відхиляють. У протилежному випадку підстав стверджувати, що розподіли у вибірках відрізняються, немає.

3.3.6. Критерій Колмогорова-Смирнова

Існують два різновиди критеріїв Колмогорова-Смирнова, тісно пов’язані один з іншим, але застосовані при вирішенні різних задач.

Одновибірковий тест Колмогорова-Смирнова (відноситься до тестів перевірки *гіпотез узгодженості*) призначений для визначення того, що вибірка має відповідний ймовірнісний розподіл. В основі критерію лежить статистика

Колмогорова-Смирнова, яка є *оцінкою відстані між емпіричною кумулятивною функцією розподілу та кумулятивною функцією теоретичного розподілу*.

Тест також може бути застосований у разі наявності двох вибірок (***двовибірковий критерій Колмогорова-Смирнова*** *однорідності вибірок*). В цій постановці *емпіричний розподіл однієї вибірки порівнюється з емпіричним розподілом іншої вибірки та вирішується, мають вибірки однаковий чи різний розподіл*. При цьому знання типу розподілу та його параметрів не вимагається. Тест Колмогорова-Смирнова з двома вибірками є одним із найбільш корисних і загальних непараметричних методів порівняння двох вибірок, оскільки він не залежить від розподілів вибірок, і чутливий одночасно до відмінностей як у положенні, так і у формі кумулятивних емпіричних функцій розподілу двох вибірок. Цей тест часто застосовується, зокрема, при аналізі мережевого трафіку інформаційно-комунікаційних мереж, в яких багато параметрів (кількість пакетів в одиницю часу, кількість з'єднань в одиницю часу тощо) виявляють ознаки локальної стаціонарності при функції щільності розподілів, *що не описується відомими теоретичними законами розподілу(!)*. Водночас, поява в розподілах значень часових рядів змін є проявом нехарактерної мережевої активності, зокрема, появи нової складової компоненти трафіку (з'єднання по протоколу *p2p*, появи потока візуальних даних), проявом спроби мережевої атаки, несанкціонованого витоку інформації, технічної несправності обладнання мережі тощо. Вияв таких ситуації є важливою складовою в системах аудиту та забезпечення безпеки комп'ютерних систем.

За визначенням, усі кумулятивні функції розподілу ймовірності завжди мають мінімальне значення 0 та максимальне значення 1, а також – знову-ж таки за визначенням - є неспадаючими функціями, тобто для них справедливе відношення:

$$F(x_i) <= F(x_j), \forall i, j : i < j.$$

Тест Колмогорова-Смирнова визначає, як відрізняються між собою дві кумулятивні функції розподілів $F_1(x)$ та $F_2(x)$, причому функція $F_1(x)$ - це в обох варіантах тесту емпірична функція розподілу вибірки (значень часового ряду), а функція $F_2(x)$ для одновибіркового варіанту тесту є теоретичною функцію відомого (включно з параметрами) ймовірнісного розподілу, а в випадку двовибіркого тесту - емпіричною функцією другої вибірки. Тест вимірює дуже

просту статистику D, яка визначається як *максимальне абсолютне значення різниці між двома кумулятивними функціями розподілу ймовірності*:

$$D = \max_x \left| F_1(x) - F_2(x) \right|$$

Тобто, аналізуються відмінності в значеннях кумулятивних функцій розподілу при однакових значеннях x. Абсолютне значення використано для того, щоб порядок операторів в формулі не впливав на знак результату, оскільки аналізується сам факт розбіжності та його величина, а не те, яка з функцій в даній точці більша або менша за іншу.

Згідно з алгоритмом критерію перевіряється нульова гіпотеза $H_0 : F_1(x) = F_2(x), \forall x$, тобто обидві вибірки належать до однієї генеральної сукупності проти альтернативної гіпотези $H_a : F_1(x) \neq F_2(x)$.

На Рис 3.8 показана побудова вказаної статистики в разі одновибіркого та двовибіркового тестів. Вважається, що чим менше значення D, тим більше схожі між собою дві функції розподілу.

В теорії ймовірності доведена теорема (так звана теорема Колмогорова [22], с.296), згідно якої закон розподілу статистики:

$$\lambda_C = \sqrt{n_1} D$$

у випадку одновибіркового критерію, та статистики:

$$\lambda_C = \sqrt{\frac{n_1 * n_2}{n_1 + n_2}} * D$$

у випадку двовибіркового критерію, при $n_1, n_2 \to \infty$, (де n_1 та n_2 - кількість елементів в кожній з вибірок), не залежать від виду функцій $F_1(x)$ та $F_2(x)$, і підпорядковуються розподілу Колмогорова:

$$\lambda(x) = \sum_{k=-\infty}^{k=+\infty} (-1)^k * e^{-2k^2 x^2}$$

Останній вираз при обчисленнях добре апроксимується формулою:

$$\tilde{\lambda}(\alpha)=\sqrt{-\frac{1}{2}\ln\frac{\alpha}{2}}$$

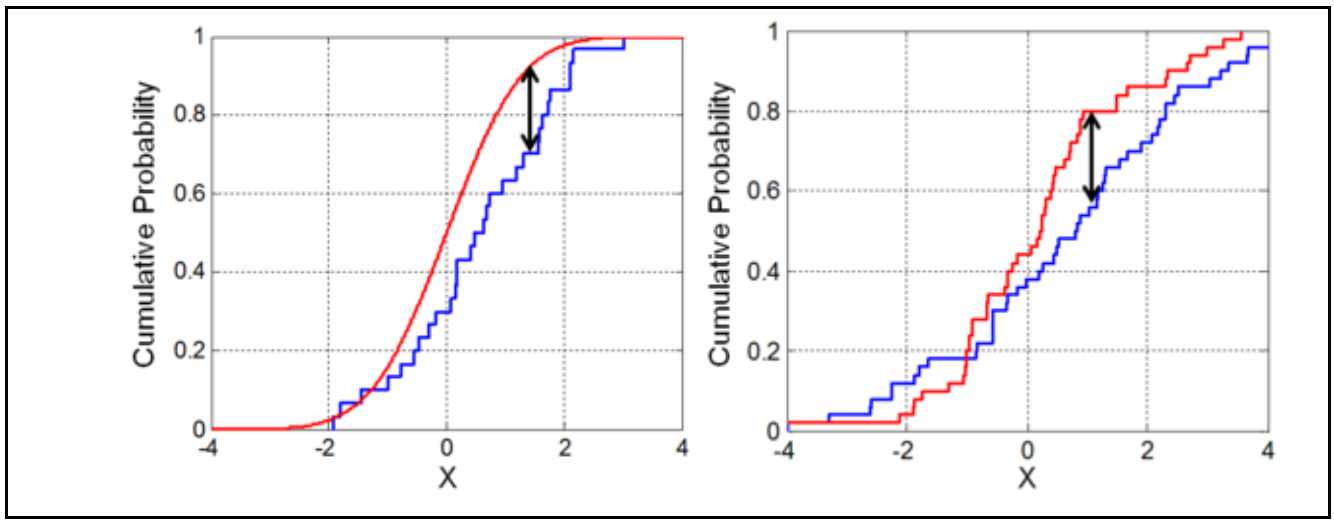


Рис. 3.8 Визначення статистики одновибіркового (зліва) та двовибіркового(справа) критерію Колмогорова-Смирнова

Довідково: візуально функція щільності розподілу Колмогорова виглядає приблизно так, як показано на Рис. 3.9.

Далі рішення щодо гіпотез приймається за звичною схемою: якщо λ_c перевищує квантиль розподілу Колмогорова $\lambda(\alpha)$ для заданого рівня значущості α, то гіпотеза H_0 (про однорідність вибірок) відхиляється. В іншому випадку ця гіпотеза приймається на рівні значущості α.

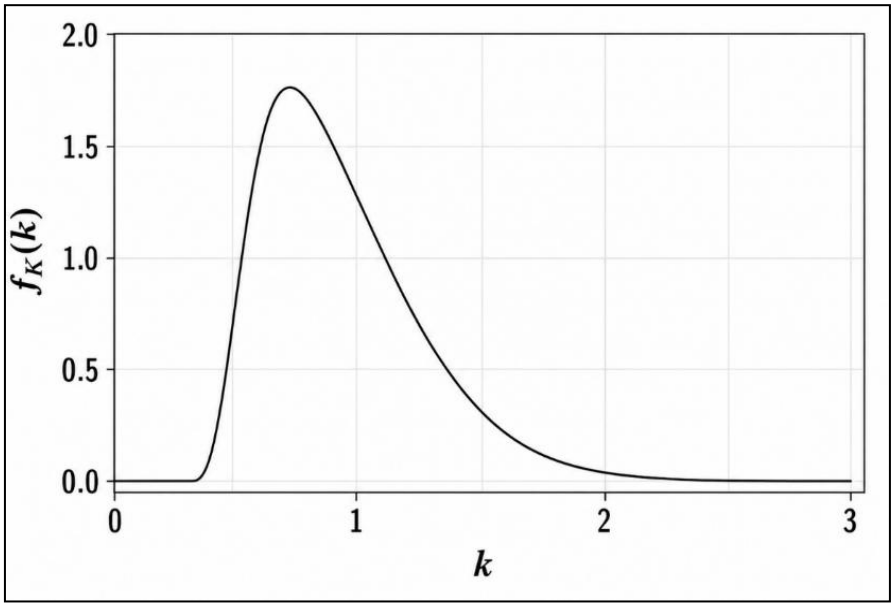


Рис. 3.9. Приклад функції щільності розподілу статистики λ

Довідково: Існує і що один різновид критерію, а саме тест Колмогорова-Смирнова-Лиллиефорса, який застосовується в разі використання одновибіркового тесту, але у випадку, коли теоретичний розподіл нам відомий, а його параметри ні, і вони мають бути визначені по самій вибірці. Використанню цього методу заважає той факт, що критичні значення для розподілу статистики визначаються виключно за таблицями, які в свою чергу створювалися на основі проведення статистичних експериментів, а не теоретичних висновків.

Корисну інформацію про критерій Колмогорова-Смирнова, його математичну базу, приклади застосування, сфери використання, реалізацію в середовищі Python тощо можна знайти, зокрема, за посиланням [55].

3.4. Спеціальні тести для перевірки гіпотези про підпорядкування даних нормальному закону розподілу

Перевірка *гіпотези про підпорядкування даних нормальному закону розподілу* (іноді вживають більш компактний вираз *«гіпотеза нормальності розподілу»*) є важливим етапом статистичного аналізу систем, оскільки багато класичних методів моделювання та статистичного висновку базуються на припущенні про нормальний характер розподілу величин, які досліджуються. *Нормальний (гаусовський)* розподіл (див. розділ 2.4.2.1) займає центральне місце у теорії ймовірностей і математичній статистиці завдяки *центральній граничній*

теоремі, згідно з якою *сума великої кількості незалежних випадкових величин, незалежно від їх індивідуальних розподілів, прямує до нормального розподілу*. Це пояснює, чому багато реальних процесів у складних системах — від часу відгуку серверів до похибок вимірювань сенсорів — демонструють нормальний або близький до нормального розподіл. Коректність застосування багатьох статистичних інструментів, зокрема *t-критерію Стьюдента* (див. Розділ 3.3.1), *F-критерію Фішера* (див. Розділ 3.3.3), регресійного аналізу, деякі методи кластеризації тощо, безпосередньо залежить від виконання припущення про нормальність. Порушення цього припущення може призвести до некоректних висновків: збільшення ймовірності *помилок першого роду* (хибне відхилення правильної гіпотези), зниження потужності критеріїв (неспроможність виявити справжні закономірності) та неадекватних прогнозів моделі. Саме тому *сучасна практика системного аналізу вимагає обов'язкової попередньої діагностики розподілу даних.* У випадку виявлення значних відхилень від нормальності дослідник має вибір: застосувати відповідні трансформації даних (логарифмування, степеневі перетворення), використати робастні методи, які менш чутливі до порушення нормальності, або звернутися до непараметричних альтернатив, які не вимагають припущень про конкретний тип розподілу. Таким чином, *аналіз відповідності емпіричних даних нормальному закону розподілу є* не формальною процедурою, а *фундаментальним кроком, що визначає вибір адекватного математичного апарату для подальшого моделювання системи та забезпечує достовірність отриманих результатів і обґрунтованість прийнятих на їх основі рішень.*

Розглянуті в розділах критерій Колмогорова–Смірнова (див. Розділ 3.3.6) що базується на максимальному відхиленні між емпіричною та теоретичною функціями розподілу, та χ^2-критерій узгодженості (див. Розділ 3.3.5), який порівнює спостережувані та очікувані частоти в інтервалах групування, хоча і мають широке коло можливостей, можуть використовуватися для перевірки відповідності будь-якому теоретичному розподілу, та часто використовуються зокрема і для перевірки узгодженості емпіричних даних з нормальним законом розподілу. Однак ця універсальність має певну ціну: жоден з згаданих критеріїв не є оптимально налаштованим спеціально для виявлення відхилень від нормальності.

3.4.1. Критерій Шапіро-Уілка

У 1965 році статистики *Семюел Шапіро* (*Samuel Sanford Shapiro*) та *Мартін Уілк* (Martin *Wilk*) запропонували альтернативний підхід — критерій, який спеціально розроблений виключно для перевірки гіпотези про нормальний розподіл і тому *здатен забезпечити значно вищу потужність* при виявленні відхилень від нормальності, особливо для вибірок малого та середнього обсягу (n < 50). На відміну від критеріїв, що порівнюють функції розподілу або частоти, ***критерій Шапіро-Уілка*** (англ. *Shapiro-Wilk test*) базується на оригінальній ідеї оцінювання, наскільки добре *лінійна комбінація порядкових статистик вибірки апроксимує вибіркову дисперсію*, використовуючи той факт, що для нормально розподілених даних існує специфічне співвідношення між цими величинами. Завдяки своїй високій чутливості до різних типів відхилень від нормальності — асиметрії, ексцесу, наявності викидів — *критерій Шапіро-Вілка сьогодні вважається одним з найнадійніших тестів нормальності* і широко застосовується як у наукових дослідженнях, так і в практичному аналізі даних, будучи реалізованим у більшості сучасних статистичних пакетах.

Система гіпотез критерію Шапіро-Уілка формулюється наступним чином:

H_0: вибірка походить із нормально розподіленої генеральної сукупності;

H_a: вибірка не підпорядковується нормальному розподілу.

Критерій Шапіро–Уілка базується на аналізі впорядкованої вибірки та порівнянні її структури з очікуваною структурою нормально розподілених даних. Нехай задано вибірку обсягу n $X = \{x_1, x_2, ..., x_n\}$. Впорядкуємо надані значення за зростанням, отримавши ряд $\{x_{(1)}, x_{(2)}, ..., x_{(n)}\}$, такий що $x_{(i)} \le x_{(i+1)}$ для $\forall i \in \{0, ..., n-1\}$. (*Зверніть увагу і не путайте* x_i - значення з наданої вибірки та $x_{(i)}$ - елемент, що посідає *i*-те місце у впорядкованому ряду значень).

Статистика критерію має вигляд:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

де a_i - вагові коефіцієнти, що визначаються на основі очікуваних значень порядкових статистик нормального розподілу. Теоретично a_i будують на основі математичного очікування та коваріаційної матриці порядкових статистик для стандартного нормального розподілу. Тобто ці коефіцієнти залежать не від самих конкретних значень вибірки, а лише від того, скільки у вибірці наявно елементів. В практичному застосуванні значення a_i не обчислюють "вручну" з даних, а беруть з таблиць або отримують програмно за алгоритмом, який базується на властивостях порядкових статистик нормального розподілу.

Чисельник формули оцінює лінійну комбінацію впорядкованих даних, яка була б оптимальною для нормального розподілу. Знаменник — це звичайна вибіркова дисперсія. Якщо дані дійсно розподілені за нормальним законом, відношення буде близьким до 1. Чим більше W відрізняється від 1 (у бік зменшення), тим сильніше дані відхиляються від нормальності.

На відміну від багатьох інших статистичних критеріїв, для критерію Шапіро–Уілка *p_value* не обчислюється за простою аналітичною формулою. Це пов'язано зі складною природою розподілу статистики W. Тому на практиці *p_value* визначається за допомогою табличних значень, чисельних апроксимацій або спеціалізованих алгоритмів, реалізованих у статистичних пакетах.

Критерій Шапіро-Уілка є одним з поширеніших статистичних тестів для перевірки гіпотези про нормальність розподілу даних, особливо ефективним для вибірок малого та середнього обсягу ($3 \leq n \leq 5000$). Перевагами критерію Шапіро-Уілка є його *високі статистична потужність* (низька ймовірність відкидання хибної нульової гіпотези), та *чутливість* (частка фактичних позитивних випадків, які були правильно ідентифіковані тестом – див. розділ 3.2.4). Тест добре працює навіть при невеликій кількості спостережень (від 3–5 елементів). При його використанні відсутня необхідність попереднього групування даних. На відміну від χ^2-критерію, або критерію Колмогорова-Смирнова, тест Шапіро-Уілка працює безпосередньо з «сирими» даними. В той же час треба чітко розуміти, що тест Шапіро–Уілка призначений виключно для перевірки

нормальності розподілу і не є універсальним щодо інших законів розподілу. Для дуже великих вибірок (n >2000) тест стає надто чутливим тобто майже будь-яка невелика *асиметрія*, *важкі хвости* або малі викиди можуть дати відхилення гіпотези нормальності. Це означає, що тест може бути «занадто суворим» і занадто часто відхиляти нульову гіпотезу навіть тоді, коли вона справедлива.

Як бачимо, тест Шапіро-Уілка хоч і показує високу якість при коректному застосуванні, але досить складний і багато в чому спирається на емпіричні, підібрані та апроксимуючі формули. Тому «ручне» його обчислення сьогодні вже не є актуальним. Використання критерію можливо через реалізації у більшості пакетів статистичної обробки. Зокрема, а екосистемі *Python* засоби роботи з даним критерієм зосереджені у функції *scipy.stats.shapiro() ,* яка одразу повертає і саме значення статистики *W* і пов'язане з ним *p_value*.

3.4.2. Критерій Харке-Бера

Критерій Харке-Бера (англ. *Jarque-Bera Test*) є статистичним тестом перевірки гіпотези про нормальність розподілу даних, запропонованим економістами *Карлосом Харке (Carlos Jarque)* та *Аніллом Бера (Anil K. Bera)* у 1980 році. На відміну від критерію Шапіро-Уілка, який базується на порядкових статистиках, або критерію Колмогорова-Смірнова, що використовує функції розподілу, критерій Харке-Бера *ґрунтується на аналізі вибіркових характеристик асиметрії та ексцесу* - двох ключових моментних характеристик розподілу, які для нормального закону мають чітко визначені теоретичні значення (див. Розділ 2.3.8).

Цей критерій особливо популярний в економетриці та фінансовій статистиці, оскільки асиметрія та ексцес мають пряму економічну інтерпретацію: асиметрія вказує на нерівномірність ризиків (переважання позитивних чи негативних відхилень), а ексцес характеризує частоту екстремальних подій («товстих хвостів» розподілу).

Відомо, що для нормального розподілу характерні наступні властивості моментних характеристик:

- коефіцієнт асиметрії (skewness): S = 0 (симетричний розподіл);
- коефіцієнт ексцесу (kurtosis): K = 3.

Будь-які *значущі відхилення вибіркових оцінок цих характеристик від теоретичних значень свідчать про порушення нормальності розподілу*.

Критерій Харке-Бера застосовує наступну систему гіпотез:

H_0: вибірка походить з нормально розподіленої генеральної сукупності (S = 0 і K = 3).

H_a: вибірка не походить з нормально розподіленої генеральної сукупності.

Для вибірки обсягу n $X = \{x_1, x_2, ..., x_n\}$ обчислюються наступні характеристики:

– вибірковий коефіцієнт асиметрії:

$$S = \frac{\hat{\mu}_3}{\hat{\sigma}_3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{\frac{3}{2}}}$$

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

$$\hat{\sigma}_3 = s^3$$

де $\hat{\mu}_3$ - оцінка третього центрального моменту.

– вибірковий коефіцієнт ексцесу:

$$K = \frac{\hat{\mu}_4}{\hat{\sigma}_4} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2}$$

$$\hat{\sigma}_4 = s^4$$

де $\hat{\mu}_4$ — оцінка четвертого центрального моменту.

Статистика критерію Харке-Бера обчислюється за формулою:

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right)$$

Перший доданок, S^2, характеризує квадрат відхилення асиметрії від нуля. Другий доданок, $(K-3)^2/4$, характеризує квадрат відхилення ексцесу від нормального значення «3». Множник $n/6$ забезпечує правильну нормалізацію.

Якщо дані походять з генеральної сукупності з нормальним закноном розподілу, при достатньо великому обсязі вибірки (*n > 30*), *статистика JB асимптотично має розподіл* χ^2 *з двома ступенями свободи*, що відповідають двом характеристикам, які досліджуються: асиметрії та ексцесу. Нульова гіпотеза про нормальність закону розподілу — це спільна гіпотеза про те, що асиметрія дорівнює нулю, а ексцес дорівнює 3. Як показує визначення JB, будь-яке відхилення від нормального закону розподілу збільшує статистику JB.

Статистика Харке-Бера дозволяє не лише перевірити загальну гіпотезу про нормальність, а й діагностувати конкретний характер відхилення:

- якщо основний внесок у *JB* вносить перший доданок *($S^2 >> (K-3)^2/4$)*, це вказує на асиметрію розподілу:
  - *S > 0* - правостороння асиметрія (довгий правий хвіст)
  - *S < 0* - лівостороння асиметрія (довгий лівий хвіст)
- якщо переважає другий доданок *($(K-3)^2/4 >> S^2$)*, проблема в ексцесі:
  - K > 3 -лептокуртичний розподіл (гостровершинний, товсті хвости)
  - K < 3 - платикуртичний розподіл (плосковершинний, тонкі хвости)
- *приблизно рівний внесок* свідчить про комбіноване порушення і асиметрії, і ексцесу.

Перевагами критерію Харке-Бера є:

- простота обчислень - використовує лише базові моментні характеристики, які легко розраховуються;
- інтерпретованість - асиметрія та ексцес мають чітку статистичну та практичну інтерпретацію;
- діагностична цінність - дозволяє ідентифікувати конкретний тип відхилення від нормальності;
- популярність в якості стандартного інструменту для перевірки залишків регресійних моделей;
- обчислювальна ефективність для великих масивів даних.

Водночас критерію притаманні і певні недоліки та обмеження:

- критерій є асимптотичним, тому для малих вибірок (n < 30) може давати неточні результати;

- низька потужність для малих вибірок - суттєво поступається критерію Шапіро-Уілка при n < 100;
- чутливість до викидів - моменти третього та четвертого порядків дуже чутливі до екстремальних значень, що може спотворювати результати;
- обмежена діагностика - розглядає лише два аспекти розподілу (асиметрію та ексцес), пропускаючи інші можливі відхилення;
- надмірна чутливість при великих n - для дуже великих вибірок (n > 5000) може виявляти статистично значущі, але практично незначні відхилення.

У бібліотеці *scipy.stats* критерій Харке-Бера реалізований у функції *scipy.stats.jarque_bera()* , яка обчислює статистику *JB* та відповідний *p_value* для масиву даних. Функція повертає кортеж (*statistic, pvalue*), де *statistic* - значення *JB*-статистики, *p_value* - двостороннє значення *p_value* для перевірки гіпотези про нормальність.

Критерій Харке-Бера, незважаючи на свої обмеження, залишається популярним інструментом експрес-діагностики нормальності, особливо в економетричних застосунках, завдяки простоті реалізації, швидкості обчислень та інформативності результатів щодо конкретного характеру відхилень від нормального розподілу.

3.5. Короткий огляд використання засобів екосистеми Python при аналізі статистичних гіпотез

Екосистема *Python* надає потужний інструментарій для статистичного аналізу та перевірки гіпотез. Основні бібліотеки - *numpy, scipy* (зокрема модуль *scipy.stats*) та *scikit-learn* - містять широкий спектр функцій для проведення параметричних і непараметричних тестів, обчислення описових статистик та аналізу розподілів даних.

Основні описові статистики

Бібліотека ***numpy*** хоча безпосередньо і не містить реалізацій статистичних тестів, проте надає набір базових функції для роботи з числовими масивами та

обчислення базових статистичних характеристик, що часто використовуються як вхідні величини при формулюванні статистичних гіпотез.

- *numpy.mean()* - обчислює середнє арифметичне значення масиву даних. Підтримує обчислення вздовж заданої осі для багатовимірних масивів.
- *numpy.median()* - визначає медіану вибірки, тобто значення, яке ділить упорядкований набір даних навпіл.
- *numpy.var()* - розраховує дисперсію вибірки. За замовчуванням обчислює незміщену оцінку (ділення на n-1), але може використовувати і зміщену (ділення на n).
- *numpy.std()* - обчислює стандартне відхилення як квадратний корінь з дисперсії.
- Крім того, модуль *numpy.random* містить генератори випадкових величин із різних теоретичних розподілів, що може застосовуватися для моделювання вибірок або проведення імітаційних експериментів.

Основні інструменти для перевірки статистичних гіпотез зосереджені у модулі ***scipy.stats*.**

- *scipy.stats.describe()*- повертає комплексну описову статистику для масиву даних, включаючи кількість елементів, мінімум, максимум, середнє, дисперсію, асиметрію та ексцес.
- *scipy.stats.sem()*- обчислює стандартну помилку середнього (стандартне відхилення вибірки, поділене на квадратний корінь з розміру вибірки), що є важливим показником для оцінювання точності вибіркового середнього.
- scipy.stats.*variation()*- визначає коефіцієнт варіації як відношення стандартного відхилення до середнього значення.

z-критерій та тести для середніх значень

Для перевірки гіпотез про параметри нормально розподілених вибірок з відомою дисперсією використовується z-критерій, однак у стандартних бібліотеках *Python* його пряма реалізація відсутня. Натомість застосовуються функції для обчислення z-статистики та критичних значень:

- *scipy.stats.norm.ppf()*- обчислює квантилі стандартного нормального розподілу, які використовуються для визначення критичних значень при заданому рівні значущості.

– *scipy.stats.norm.cdf()* та *scipy.stats.norm.sf()* кумулятивна функція розподілу та «функція виживання» ($SF(x) = 1 - CDF(x)$), що дозволяють обчислювати p_value для z-статистики.

Для практичного використання z-тесту можна також звернутися до бібліотеки *statsmodels*:

– *statsmodels.stats.weightstats.ztest()* — виконує одновибірковий або двовибірковий z-тест для перевірки гіпотези про середнє значення.

Критерій Стьюдента (t-тест)

Модуль ***scipy.stats*** містить повний набір функцій для різних варіантів t-критерію:

– *scipy.stats.ttest_1samp()* - одновибірковий t-тест для перевірки гіпотези про рівність середнього значення вибірки заданому теоретичному значенню. Повертає значення t-статистики та двосторонній *p_value*.

– *scipy.stats.ttest_ind()* - двовибірковий t-тест для незалежних вибірок. Перевіряє гіпотезу про рівність середніх двох незалежних вибірок. Параметр *equal_var* дозволяє вибрати між варіантом з рівними дисперсіями (класичний тест Стьюдента) та варіантом з нерівними дисперсіями (тест Уелча).

– *scipy.stats.ttest_rel()* - парний t-тест для залежних вибірок. Застосовується для аналізу парних спостережень, наприклад, вимірювань до і після впливу на одних і тих самих об'єктах.

Для обчислення критичних значень t-розподілу використовуються:

– *scipy.stats.t.ppf()* — квантильна функція t-розподілу Стьюдента для визначення критичних значень при заданій кількості ступенів свободи.

– *scipy.stats.t.cdf()* та *scipy.stats.t.sf()* - функції для обчислення *p_value* на основі t-статистики.

Критерій Фішера (F-тест)

Для перевірки гіпотез про рівність дисперсій двох нормально розподілених вибірок пов’язаних з F-критерієм, використовуються об’єкти класу *scipy.stats.f*,

що дозволяють обчислювати щільність, функцію розподілу, критичні значення та квантилі відповідного теоретичного закону.

- *scipy.stats.f.cdf()* та *scipy.stats.f.sf()* - кумулятивні функції розподілу Фішера, що використовуються для обчислення *p_value* F-статистики.
- *scipy.stats.f.ppf()* - квантильна функція F-розподілу для визначення критичних значень.

Хоча прямої функції для F-тесту порівняння дисперсій у *scipy.stats* немає, F-статистику можна обчислити як відношення двох вибіркових дисперсій, а *p_value* отримати за допомогою функцій розподілу.

Критерій Вілкоксона-Манна-Уітні

- *scipy.stats.mannwhitneyu()* — непараметричний тест для порівняння двох незалежних вибірок. Перевіряє гіпотезу про те, що розподіли двох вибірок однакові проти альтернативи, що одна вибірка має тенденцію до більших (або менших) значень. Параметр *alternative* дозволяє задати тип альтернативної гіпотези: двостороння (*'two-sided'*), односторонні (*'less', 'greater'*).

Знаковий критерій Вілкоксона для парних вибірок

- *scipy.stats.wilcoxon()* — непараметричний аналог парного t-тесту. Використовується для перевірки гіпотези про медіану різниць парних спостережень.

χ^2-квадрат критерій

χ^2-квадрат критерій має кілька варіантів застосування, для кожного з яких у *scipy.stats* є відповідна функція:

- *scipy.stats.chisquare()* - одновибірковий критерій χ^2 для перевірки узгодженості. Порівнює спостережувані частоти з очікуваними частотами за певним теоретичним розподілом.
- *scipy.stats.chi2_contingency()* - критерій χ^2 для таблиць спряженості (*Contingency Tables*). Використовується для перевірки незалежності двох категоріальних змінних. Вона також може використовуватися у задачах перевірки однорідності кількох вибірок, коли дані подані у

вигляді частотних таблиць. Функція повертає значення χ^2-статистики, *p_value*, кількість ступенів свободи та матрицю очікуваних частот.

Для роботи з розподілом χ^2 доступні:

- *scipy.stats.chi2.cdf()* та *scipy.stats.chi2.sf()* - функції для обчислення p-значень.
- *scipy.stats.chi2.ppf()*- квантильна функція для визначення критичних значень.

Критерій Колмогорова-Смирнова

Критерій Колмогорова-Смірнова використовується для перевірки гіпотез про відповідність емпіричного розподілу теоретичному або для порівняння двох емпіричних розподілів.

- *scipy.stats.kstest()* - одновибірковий критерій Колмогорова-Смірнова. Перевіряє гіпотезу про те, що вибірка походить із заданого теоретичного розподілу. Другий аргумент може бути назвою розподілу ('norm', 'uniform' тощо) або функцією кумулятивного розподілу.
- *scipy.stats.ks_2samp()* - двовибірковий критерій Колмогорова-Смірнова для порівняння двох незалежних вибірок. Перевіряє гіпотезу про те, що обидві вибірки походять з одного й того ж розподілу.
- *scipy.stats.ksone()* та *scipy.stats.kstwo()* - об'єкти розподілів для одновибіркового та двовибіркового варіантів, що містять методи для обчислення критичних значень та *p_value*.

Тести нормальності закону розподілу

Крім розглянутих, бібліотека *scipy.stats* містить великий набір інших функцій, що реалізують деякі спеціальні критерії, для перевірки найбільш популярних типів гіпотез. Так, для перевірки гіпотези про нормальність розподілу даних доступні спеціалізовані критерії:

- *scipy.stats.shapiro()* - критерій Шапіро-Уілка, один з найпотужніших тестів нормальності для невеликих вибірок ($n < 50$).
- *scipy.stats.normaltest()* - критерій д'Агостіно-Пірсона, що базується на асиметрії та ексцесі вибірки.
- *scipy.stats.jarque_bera()* - критерій Харке-Бера , альтернативний тест нормальності на основі асиметрії та ексцесу.

Контрольні запитання

1. Що таке статистична гіпотеза?
2. У чому різниця між нульовою та альтернативною гіпотезами?
3. Що означає перевірка статистичних гіпотез?
4. Які етапи процедури перевірки гіпотези?
5. Що таке рівень значущості і як його обирають?
6. Поясніть поняття *p_value* простими словами.
7. Що таке статистичний критерій (тест)?
8. У чому полягає різниця між односторонньою та двосторонньою перевіркою гіпотез?
9. Що таке область прийняття та область відхилення гіпотези?
10. Які існують помилки першого та другого роду?
11. Як пов’язані між собою рівень значущості та ймовірність помилки першого роду?
12. Що таке z-критерій і чим він відрізняється від t-критерію Стьюдента? Коли доцільно використовувати кожен з них?
13. У яких випадках застосовується t-критерій Стьюдента?
14. У яких випадках застосовується одновибірковий t-критерій Стьюдента? Які припущення необхідні для його коректного використання?
15. У яких ситуаціях доцільно використовувати непараметричні критерії замість параметричних? Назвіть основні переваги непараметричних методів.
16. Як інтерпретувати результат перевірки гіпотези на практиці?

4. ВИЯВЛЕННЯ ВЗАЄМОЗАЛЕЖНОСТЕЙ В ДАНИХ

4.1. Взаємозалежність між параметрами об’єктів. Прояви, аналіз, використання та значення для аналізу систем

У процесі дослідження будь-яких складних систем — технічних, соціальних, біологічних або інформаційних — важливо не лише вміти збирати дані про окремі *параметри* (*ознаки*), а й дослідити, чи *ці параметри взаємопов’язані між собою і – в разі отримання ствердної відповіді - як саме вони пов’язані.* У сучасному аналізі даних виявлення взаємозалежностей є одним із ключових кроків, що передує моделюванню, прогнозуванню, оптимізації та прийняттю рішень.

Параметри об’єктів рідко існують ізольовано: зміна одного фактору часто впливає на інші. Уявіть, що ви вибираєте на базарі яблуко. Його можна описати багатьма ознаками: ступень почервоніння, вага, кількість черв'яків тощо. Але як споживачів вас цікавить смак, який вимірюється «в папугах» (тобто, в невідомо яких і точно ніяк не стандартизованих «одиницях виміру»), припустимо - за п'ятибальною шкалою. Ви не можете безпосередньо надкусити яблуко та «виміряти» його смак. Розуміючи, що між ознакою «смак» і іншими згаданими ознаками існує певний взаємозв’язок ви можете спробувати «*передбачити*» значення «смаку» за значеннями інших параметрів. Наприклад, з життєвого досвіду вам відомо, що смак з пристойною точністю дорівнює

*5 * «червоність» + 2 * вага-7 * кількість черв'яків.*

Гарна підказка. Але щоб нею скористуватися спершу треба впевнитися, що такий зв’язок існує у дійсності, а не тільки у вашій уяві, а вже в разі отримання позитивної відповіді на це питання - будувати певну модель (тобто - формальний вираз), що пов’яже значення відповідних ознак. Зрозуміло, що в реальних задачах все буде набагато серйозніше аніж в нашому іграшковому прикладі. Наприклад, якщо звернути увагу на те, в яких шкалах виміряні параметри об’єкта дослідження, які обрані для аналізу, то стає зрозумілим, що взаємозалежність в парі «червоність»-«смак» і в парі «кількість черв’яків»-«смак» має і визначатися і інтерпретуватися зовсім по-різному. Тим більше формально додавати такі показники, та ще й попередньо помноживши їх на якість коефіцієнти, без огляду на застосовані шкали вимірів - виглядає м’яко кажучи «непрофесійно».

Взаємозалежність між ознаками одного об'єкту, або між параметрами різних – але пов'язаних в рамках деякої системи – об'єктів можна знайти майже всюди. Наприклад:

- рівень продуктивності мережевого обладнання може залежати від рівня навантаження мережі та часом затримки пакетів (останнє, в свою чергу, може визначатися технічними параметрами обладнання);
- успішність курсантів може бути пов'язана із часом, який вони витрачають на навчання;
- температура процесора в комп'ютері впливає на частоту його роботи та споживану потужність;
- середня температура дня наступного досить сильно залежить від середньої температури дня попереднього;
- кількість клієнтських звернень може залежати від маркетингової активності компанії;
- тощо.

У кожній з цих ситуацій дослідника цікавить не просто сукупність значень, а *структура зв'язків* між ними: чи є вони випадковими, закономірними? Сильними чи слабкими? Лінійними або нелінійними? В яких шкалах виміряні ознаки?

Для чого такі знання потрібні?

Наприклад, кожну людину можна формально описати, вказавши в якості ознак опису її вік і місце народження, рівень освіти і річний дохід, стать, соціальну приналежність тощо. Питання полягає в тому, чи можна за значеннями однієї ознаки зробити певні висновки – хоча б приблизні (ймовірнісні) – про значення іншої, або ж знання про значення одного параметру нічого не додає до знання про значення іншого (тобто ці ознаки проявляються незалежно одна від одної)? Відповіді на такі питання можуть мати значну практичну цінність. Наприклад, якщо буде встановлено, що ознаки "професія" і "політичні переконання" незалежні, то соціологічні опитування за передбаченням результатів виборів можна проводити без урахування професії опитуваних, що може зменшити вартість дослідження не впливаючи на точність отриманих результатів.

У статистиці під ***взаємозалежністю*** зазвичай розуміють наявність певного *систематичного зв'язку* між параметрами. *Такий зв'язок може проявлятися: у*

спільних тенденціях змін значень параметрів; у синхронності процесів, що досліджуються; у передбачуваному характері впливу одного параметра на інший тощо. Знання характеру зв’язків між параметрами дозволяє:

- проводити діагностику об’єктів і систем (наприклад, знаходити вузькі місця в мережевій інфраструктурі);
- будувати прогностичні моделі, де одні змінні використовуються для передбачення інших;
- виявляти аномалії, оскільки порушення звичних залежностей часто сигналізує про відмови або атаки;
- оптимізувати управління, змінюючи параметри, що найбільше впливають на результат;
- спрощувати моделі, виключаючи параметри, які не несуть корисної інформації.

У машинному навчанні аналіз взаємозалежностей є базою для ***відбору ознак*** (англ. *Feature Selection*), зменшення розмірності, побудови класифікаторів і регресійних моделей.

В деяких випадках ознаки можуть бути пов'язані жорстко: якщо професія - гірник або сталевар, то стать, (майже) безсумнівно, чоловіча. Тим самим за деякими значеннями ознаки «професія» можна «дізнатися» значення ознаки «стать». Інша крайність - відсутність зв'язку: якщо очі сірі, то яка професія у цієї людини? Але не все так однозначно. Спробуйте самостійно дати відповідь на питання - чи пов’язаний колір очей з ймовірністю захворіти певною хворобою? Чи пов’язана IP-адреса відправника пакету з ймовірністю того, що трафік, який він генерує, є спамом?

Дослідника в подібних завданнях в першу чергу цікавить, *наскільки точно можна передбачити значення однієї ознаки за значенням іншого*. Але рішенню цієї задачі має передувати рішення більш простої - а саме перевірки того, чи існує взагалі будь-який зв'язок між ознаками? Таким чином, в термінах теорії перевірки статистичних гіпотез виникає і вимагає перевірки наступна ***нульова гіпотеза:*** *прояви (значення вимірів) однієї ознаки, які ми отримали в результаті експеременту, незалежні від проявів іншої ознаки.*

4.2. Види залежностей між параметрами об’єктів: функціональна та кореляційна

У статистиці й прикладному аналізі даних розрізняють принаймні два принципово різні види зв’язків:

- функціональна залежність;
- кореляційна залежність.

Хоча обидві види описують вплив одних параметрів на інші, їх природа суттєво відрізняється.

Якщо значенням однієї величини x відповідає цілком певне значення y іншої, то кажуть, що між цими величинами має місце **функціональна залежність:**

$$y = f(x)$$

В даному випадку *існує точна формула або закон, що пов’язують значення змінних.* Змінна y в такому випадку визначається абсолютно однозначно, а у відношенні між параметрами відсутня будь яка випадковість. Наприклад, при фіксованих умовах, при відсутності зовнішніх впливів час затримки в мережевій лінії може бути точно розрахований як функція відстані та швидкості поширення сигналу.

Кореляційна залежність — *це статистичний зв’язок, який не є строго детермінованим*. Зміна значень однієї ознаки лише певним чином впливає на зміни значень іншої, але не визначає її однозначно. В прикладному аспекті це може бути пов’язане з тим, що на значення змінної y впливають не тільки значення x, але і інші чинники, які не враховуються в законі (моделі, формулі чи алгоритмі) що пов’язують x та y. Кореляція зазвичай показує чи є зв'язок суттєвим або випадковим; наскільки сильно змінні змінюються разом; у якому напрямку (прямий, обернений зв’язок).

Так, кількість мережевих атак може корелювати з часом доби, активністю користувачів, подіями в інтернет-просторі. Проте жодна із цих ознак та їх сукупність в цілому не можуть визначити (передбачити) точну кількість атак в певний момент часу. Час який був витрачений на навчання корелює з успішністю курсанта, але не визначає її точно та однозначно. Кількість помилок у

програмуванні може корелювати з рівнем досвіду розробника. Проте всі ці зв’язки також не є суворими: на реальний результат впливає багато факторів, що не враховані в моделі. В складних системах, де неможливо точно описати закони функціонування об’єктів, сама система надто складна, а впливів на її поведінку може бути дуже багато. Тому часто саме аналіз кореляції стає головним, а іноді і єдиним можливим інструментом дослідження.

Дуже важливе зауваження. ***Кореляція не означає причинності.*** І це є ключовою логічною пасткою у статистиці, яка загрожує дослідникам-початківцям. Навіть доведена статистично наявність кореляції між двома змінними не доводить, що одна з них *причинно* впливає на іншу. ***Наявність кореляції лише свідчить про те, що між змінними існує певний статистичний зв’язок*** — вони змінюються разом частіше, ніж випадково. ***Однак цей зв’язок може виникати з різних причин,*** і далеко не завжди він означає існування прямої залежності однієї змінної від іншої.

Є кілька типових сценаріїв, що пояснюють, чому так відбувається. Зокрема т.з. ***хибна кореляція*** відповідає ситуації, коли дві змінні корелюють лише випадково або через зовнішній фактор, але немає жодного реального прямого зв'язку між ними. Наприклад, кількість DDoS-атак у світовій мережі інтернет і популярність запиту у Google про напругу у відносинах між певними країнами можуть корелювати між собою, але це не означає, що саме ці пошукові запити викликають атаки. Багато «дивних» хибних кореляцій можна знайти у відкритих джерелах. Наприклад, кореляція між кількістю людей, що втопилися у басейнах, і кількістю фільмів за участі Ніколаса Кейджа, що були зняті у відповідному році, або кореляція між кількістю захворювань на кір у світі і кількості зареєстрованих шлюбів (хто зацікавився - рекомендуємо зазирнути за посиланням [56] де можна знайти ще багато прикладів досить дивної хибної кореляції). Відомим прикладом хибної кореляції є залежність між людською народжуваністю та кількістю пар лелек у різних регіонах Європи: хоча між цими двома величинами існує відповідність (тобто чим більше лелек гніздиться біля оселі, тим більше дітей з'являється на світ), проте немає певного причинно-наслідкового зв'язку (хибний висновок, що дітей приносить лелека). Кореляція між новонародженими та парами птахів пояснюється через той факт, що зазвичай лелеки селяться в сільській місцевості, де переважають багатодітні сім'ї.

Часто дві змінні корелюють лише тому, що на них впливає третя, ***непомічена змінна.*** Наприклад, зростання затримки в мережі (latency) та збільшення кількості тайм-аутів у додатку часто корелюють між собою, але справжня причина може бути в тому, що виникло перевантаження маршрутизатора, наявна DDoS-атака, вийшов з ладу мережевий комутатор, у провайдерів верхнього рівня відбувся збій тощо. В даному разі латентний фактор впливає і на затримку, і на тайм-аути. Іншим прикладом такої уявної кореляції є зв'язок між зарплатою батьків і успішністю учнів. В дійсності цей зв'язок не має причинно-наслідкового характеру. Справжньою спільною причиною є рівень освіти батьків. Висока освіта батьків одночасно призводить до вищих зарплат (через кращу кваліфікацію) і до вищої успішності дітей (через формування сприятливого навчального середовища, допомогу з навчанням та вищі очікування). Отже, обидві ознаки (зарплата і успішність) є наслідками третьої, прихованої змінної — освіти батьків.

Наведемо ще декілька прикладів, що показують як широке розповсюдженя, так і складність інтерпретації результатів аналізу взаємозалежності.

При роботі з даними спостерігається позитивна кореляція між зростанням продажів морозива та збільшенням кількості випадків утоплення. Спільна причина: літня спека. Коли температура висока, люди частіше купують морозиво і частіше плавають (що збільшує ризик утоплення). Але наявність такої кореляції зовсім не означає, що чим більше будуть продавати морозива, там більше буде кількість потопальників (або навпаки).

З віком у дітей збільшується і розмір ноги, і словниковий запас. Спільна причина: вік. З часом діти дорослішають, і обидві ознаки далі зростають незалежно одна від одної.

Кількість пожежних на місці пожежі та завдана шкода: дані демонструють, що чим більше пожежних приїжджає на пожежу, тим більша шкода від вогню. Спільна причина: серйозність (масштаб) пожежі. Велика пожежа вимагає більше пожежних і завдає більше шкоди.

Одужання після прийому гомеопатичних засобів ще не доводить, що воно настало в результаті цього прийому: ми можемо випустити з уваги невідомий третій фактор. Наприклад, деякий час вважалося, що прийняття певних ліків підвищує шанс на одужання при серцевих хворобах. Але згодом помітили (зрозуміли), що оскільки ліки, щодо яких проводилося дослідження

ефективності, були дуже коштовні, приймати їх могли дозволити собі більш заможні люди. А для них характерні і більш якісний медичний догляд і більш якісне харчування, і взагалі, більша увага до свого здоров'я - саме ці обставини виявилися справжніми причинами зниженого ризику коронарної недостатності. Ну а кореляція може бути і зовсім випадковістю.

В 90-х роках минулого сторіччя проводилася програма передбачення ймовірності смерті від пневмонії [57]. Мета – госпіталізувати лише людей із підвищеним ризиком, інших – лікувати амбулаторно, суттєво знижуючи витрати на лікування. При цьому результати впевнено показували, що наявність у людини астми – дуже важкої, подекуди смертельної хвороби – дуже сильно негативно корелює (тобто начебто знижує) ризик смерті від пневмонії – хто хворіє на астму менше помирає від пневмонії. Але в дійсності причиною кореляції цих факторів була наявність регулярної лікарської допомоги, що отримують люди з астмою, їх майже постійне перебування під особливим медичним наглядом. При захворюванні на пневмонію це дозволяло раніше її виявити та відповідно зменшити ймовірність ускладнень, в точу числі - летальних.

Логічна помилка «кореляція означає причинність» - *одна з найчастіших пасток у дослідницькій аналітиці*, економіці, соціології, інформаційних технологіях і навіть у машинному навчанні. Вміння виявляти такі «*кореляції через третю змінну*» - важливий навичок спеціаліста за аналізу даних. Так, в галузі кібербезпеки плутанина між кореляцію і причинністю, може:

- неправильно визначити джерело атаки,
- помилково конфігурувати систему виявлення вторгнень (IDS),
- зробити некоректні прогнози навантаження,
- зламати роботу ML-моделі через хибні ознаки,
- прийняти неправильні управлінські рішення,
- тощо.

Щоб відчути складність вивчення причинно-наслідкових зв’язків спробуйте самі відповісти на цікаве питання - чи є зв’язок між зростом людини та довжиною її волосся? Несподівана відповідь – так, такий зв’язок (що найменше - в нашому соціально-культурному середовищі) існує. Виходимо з того, що зріст чоловіків, як правило, вищий за зріст жінок, а от довжина волосся – навпаки. Отже – зв’язок (зворотній) є, але він пов’язує не безпосередньо

значення однієї ознаки з значеннями іншої, а через третій параметр, в даному прикладі - через стать об'єктів дослідження.

4.3. Методи оцінки кореляційної залежності. Коефіцієнт кореляції Пірсона

Математичної мірою кореляції двох випадкових величин служить коефіцієнт кореляції. ***Коефіцієнт кореляції*** (англ. *Correlation Coefficient*) або парний коефіцієнт кореляції - в теорії ймовірностей і статистиці - показник зв'язку між змінами випадкових величин. Через значення коефіцієнта кореляції намагаються виразити, чи існує статистично достовірний зв'язок між двома (або кількома) змінними в одній або декількох вибірках. Наприклад, взаємозв'язок між ростом і вагою дітей, між заробітною платою та результатами виконання тесту *IQ*, між стажем роботи і продуктивністю праці, між кількістю відвідувачів сайту та завантаженістю каналу зв'язку. У загальному випадку ***кореляційний аналіз*** *дозволяє оцінити напрямок, щільність та силу статистичного зв'язку між змінними*, не роблячи при цьому припущень про причинно-наслідковий характер цього зв'язку.

Найпоширенішою мірою *лінійної кореляції* між двома кількісними змінними є ***коефіцієнт кореляції Пірсона*** (англ. *Pearson Correlation Coefficient*). Нехай існують дві множини даних $X = \{x_n\}$, $Y = \{y_n\}$, що представляють значення ознак об'єкту (чи об'єктів), які досліджуються. Можна підрахувати середні значення $\bar{x}$ для об'єктів з множини *X* та для $\bar{y}$ об'єктів з множини *Y*. (див. Розділи 2.3.4.2 та 2.3.7) Маючи ці значення можливо описати ті самі властивості параметрів об'єктів, але в іншій системі координат - а саме в значеннях відхилення від уявної точки з координатами $(\bar{x}, \bar{y})$. Тобто, у значеннях $(x_i - \bar{x})$ та $(y_i - \bar{y})$. Рис.4.1 наглядно пояснює цей перехід.

Для кожної точки можливо знайти значення спільного відхилення $(x_i - \bar{x}) * (y_i - \bar{y})$ та середнє значення суми таких відхилень:

$$S_{XY} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) * (y_i - \bar{y})}{n - 1}$$

Для двох випадкових ймовірнісних замінних цей показник носить назву ***коваріація*** (англ. *Covariance*). По суті це математичне очікування добутків відхилень випадкових величин.

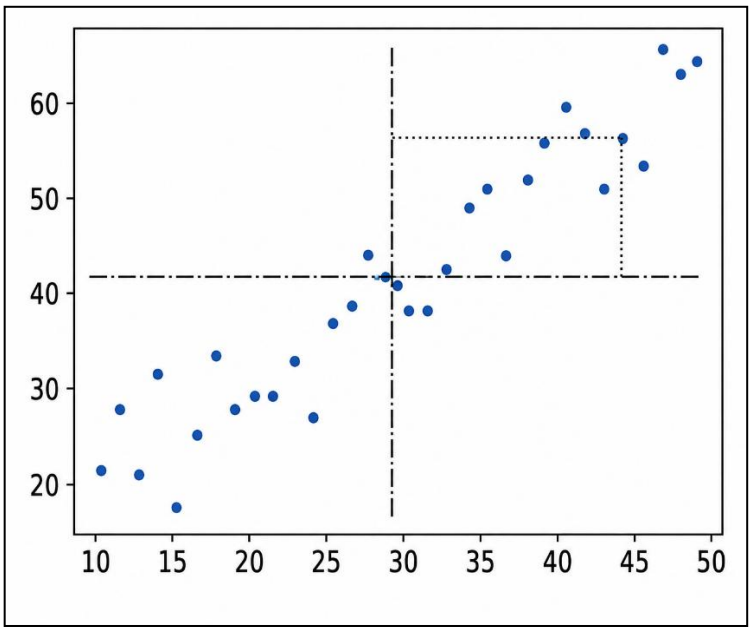


Рис. 4.1 До питання визначення коефіцієнту кореляції Пірсона

Якщо значення x_i вище середнього (позитивне відхилення) і значення y_i теж вище середнього (позитивне відхилення), добуток буде позитивним. Якщо обидва значення нижче середнього (негативні відхилення), добуток також буде позитивним. Сума цих добутків (коваріація) для всіх об'єктів, що досліджується, буде *позитивною*. Якщо x_i вище середнього, а y_i нижче або навпаки (тобто множники мають різні знаки), всі складові суми будуть негативним, а отже і сама сума *буде негативною*. Іншими словами, цей показник вказує, чи однонаправлено змінюються обидві ознаки в порівнянням з значеннями $\overline{x}$ та $\overline{y}$.

Проблема коваріації полягає в тому, що її значення залежить від одиниць вимірювання (наприклад, гривні чи тисячі гривень). Щоб отримати *стандартизований показник, значення якого завжди знаходиться в інтервалі між -1 та +1*, потрібно поділити значення коваріації на добуток стандартних відхилень для вибірок *X* та *Y*:

$$r_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right) * \left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 * \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}$$

Цей показник і є формальним визначенням *коефіцієнта кореляції Пірсона*.

Ще раз наголосимо, що цей показник вимірює *виключно* лінійну кореляцію між двома наборами даних. *Спроба його застосування в інших випадках може призвести до хибних результатів, що спотворюють реальний стан речей.*

Коефіцієнт кореляції Пірсона розглядається як вибіркова статистика (тобто - оцінка) невідомого нам (через відсутність інформації про значення параметрів закону розподілу) коефіцієнту кореляції генеральної сукупності:

$$\rho_{XY} = \frac{\mathrm{cov}\left(X, Y\right)}{\sigma_X^2, \sigma_Y^2}$$

Тут $\mathrm{cov}(X, Y)$ - теоретичне (і невідоме) значення параметру коваріації у генеральній сукупності.

В залежності від особливостей залежності, що досліджується, коефіцієнт кореляції Пірсона може приймати різні значення. Див. Рис.4.2.

По-перше, зауважимо – коефіцієнт кореляції приймає значення в потрібному нам діапазоні, між -1 та +1. *Чим ближчий значення коефіцієнту до +1 або до -1, тим сильніший зв'язок між двома величинами.* Значення +1 або -1 інформують про чисту функціональну залежність між значеннями: $Y=aX+b$. Звертаємо увагу – про кут нахилу (на графіку), або знак коефіцієнта «а» в даний момент не йдеться. Єдине, про що іде мова на даному етапі дослідження - *як тісно точки, що репрезентують об'єкти дослідження, групуються навколо теоретичних значень функції.*

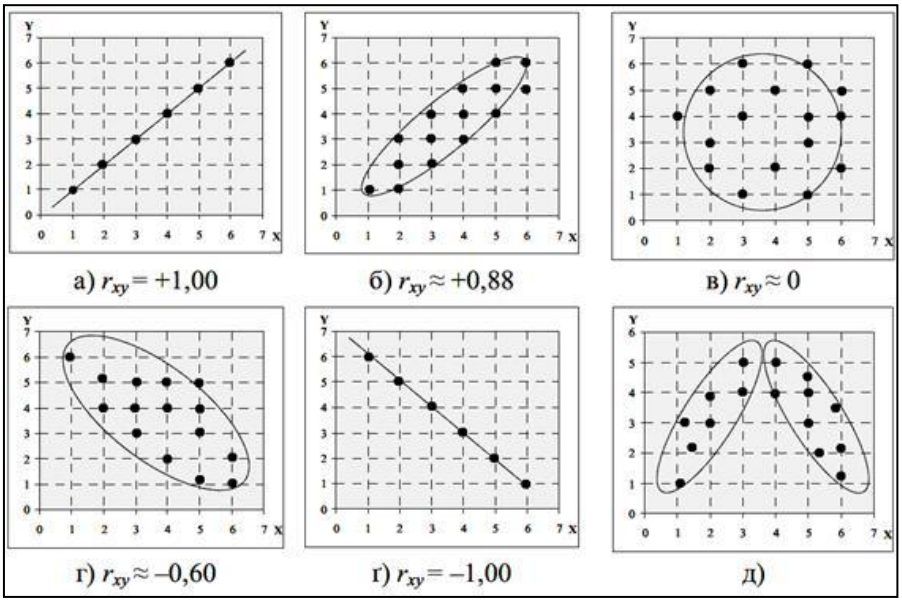


Рис.4.2 Зміни значення коефіцієнта кореляції в залежності від особливостей залежностей значень даних

Якщо значення коефіцієнту кореляції позитивне, то говорять про *позитивний або прямий зв'язок* між величинами, тобто при збільшенні однієї величини зростає друга (не функціонально, звісно, а ймовірнісно). І навпаки, якщо значення коефіцієнту кореляції негативне, то говорять про *негативний зв'язок*, тобто з ростом значень одного параметру спостерігається зменшення другого. Наприклад, якщо для пари змінних «кількість відвідувачів сайту» та «завантаженість каналу зв'язку» отримано значення r=0.82, це свідчить про сильний прямий лінійний зв'язок: зі зростанням кількості відвідувачів, як правило, зростає і навантаження на канал. Проте це не означає, що кожне збільшення відвідувачів автоматично спричиняє перевантаження — на результат можуть впливати й інші фактори, зокрема тип контенту або механізми кешування.

Наведемо також деякі приклади того, що називають «*негативною кореляцією*»:

- Вік автомобіля та його вартість: коли автомобіль старіє, вартість його перепродажу зазвичай зменшується.
- Зовнішня температура та витрати на опалення: у міру підвищення температури на вулиці витрати на опалення будинку, як правило, зменшуються.

– Час вивчення та кількість помилок у тесті: у міру збільшення часу, витраченого на навчання, кількість помилок у тесті зазвичай зменшується.
– Швидкість руху та час у дорозі: зі збільшенням швидкості руху час, необхідний для досягнення пункту призначення, зазвичай зменшується.
– Інтенсивність фізичних вправ і частота серцевих скорочень у стані спокою: у міру збільшення фізичних навантажень людини частота серцевих скорочень у стані спокою має тенденцію до зниження.

Термін «*негативна кореляція*» звучить дещо підозріло, ніби кажучи, що між двома змінними не існує зв’язку. Однак це не так. *Кореляція як така вказує, що знання значення однієї змінної несе деяку інформацію і про значення іншої.* Це стосується і негативної кореляції, яка просто вказує не те, що значення X вище середнього відповідає значенню Y нижче середнього, і навпаки. На відміну від цього, коли зв’язок між двома змінними дійсно не існує, збільшення значення однієї змінної не додає інформації до знання про значення іншої, яка лишатися випадковою (як зріст людини ніяк не пов’язаний з її успіхами у навчанні). Справжня *відсутність зв’язку має нульовий коефіцієнт кореляції* — ані негативний, ані позитивний.

Щоб визначити, чи існує концептуально негативна кореляція, запитайте себе: «Що трапитися, якщо значення змінної X зросте?» Якщо очікується, що змінна Y зменшиться, це зворотний або негативний зв’язок. *Чим ближче значення коефіцієнту кореляції до 0, тим зв’язок між значеннями слабший.* Нульова кореляція буде мати місце за відсутності зв'язку між змінними. Це відповідає хмаровидній формі діаграми розсіювання.

Дуже важливе зауваження - нульове значення коефіцієнту кореляції може свідчити лише про відсутність лінійної залежності, а не взагалі про відсутність будь-якого статистичного зв'язку.

Рис 4.3 демонструє ситуації, коли коефіцієнт кореляції між змінними дорівнює нулю, але причини такого значення - абсолютно різні.

Таким чином коефіцієнт кореляції використовується як певний «термометр», який показує «нуль» в разі незалежності змінних, плюс одиницю в разі прямої лінійної залежності змінних і мінус одиницю в разі зворотного лінійної залежності змінних. Значення коефіцієнта, що знаходяться між нулем і одиницею розуміються (хоча з формальної, математичної точки зору це не є

обгрунтованим – див далі.) так: чим ближче значення коефіцієнта кореляції до нуля, тим слабкіше залежність, чим ближче до (плюс або мінус) одиниці - тим залежність сильніша.

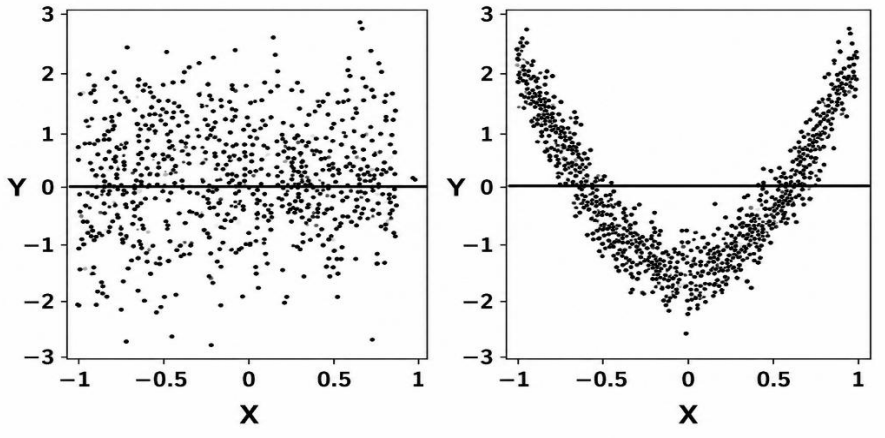


Рис. 4.3 Приклади даних, що мають нульовий коефіцієнт кореляції

Іноді отримані значення коефіцієнту переводять в деяку якісну, здебільшого – рангову, шкалу. Таке перетворення в даному разі носить суто довільний характер, є чисто умовним. Ранги можуть змінюватися залежно від галузі дослідження та слабо придатні для прийняття технічних рішень або для застосування в програмних застосунках.

Треба пам’ятати, що незважаючи на свою простоту, коефіцієнт кореляції Пірсона вимагає дотримання (та попередню перевірку) *умов застосування*, а саме:

- змінні є кількісними (використовуються інтервальна або відносна шкали вимірювання);
- зв’язок між змінними є приблизно лінійним;
- відсутні суттєві викиди, які можуть спотворити результат;
- бажано, щоб розподіли змінних були близькі до нормальних.

Порушення цих умов може призвести до хибних висновків щодо сили або навіть наявності зв’язку.

Як вже зазначалося, вибірковий коефіцієнт лінійної кореляції Пірсона, як і всі вибіркові характеристики, є випадковою величиною і при повторенні вимірювань може приймати дещо відмінні значення. Тому при статистичному аналізі важливо не лише обчислити саме значення r_{XY}, але й оцінити його

статистичну значущість та побудувати довірчий інтервал для параметра ρ - істинного коефіцієнта кореляції між значеннями ознак, які досліджуються в генеральній сукупності. Зрозуміло, цей показник (ρ) ми можемо тільки оцінити, спираючись на дані, що отримані в експерименті. А отже, для аналізу кореляційної залежності необхідно застосовувати відповідні інструменти теорії перевірки гіпотез (див. Розділ 3). В даному випадку вирішується наступне питання: чи може вибірковий коефіцієнт кореляції, який був отриманий для наданих двох вибірок даних, *випадково відрізнятися від нуля*, в той час, як в дійсності (в генеральній сукупності) випадкові змінні X і Y - некорельовані? Тобто - довести або відхилити на основі отриманих значень вибіркової статистики r нульову гіпотезу $H_0 : \rho = 0$. Альтернативна гіпотеза $H_a : \rho \neq 0$. Таким чином *працюючи з вибіркою мало встановити, що вибірковий коефіцієнт кореляцій малий або великий (близький або далекий за абсолютною величиною до значення 0), необхідно ще й впевнитися чи достатня його величина для обґрунтованого висновку* (прийняття або відхилення нульової гіпотези про відсутність кореляційної зв'язку).

Для цього можна застосувати вже відому нам t-статистику, але не до самого вибіркового коефіцієнту кореляції, а до певного його перетворення:

$$t_p = |r| \sqrt{\frac{n-2}{1-r^2}}$$

Подальший аналіз повністю відповідає фреймворку, що був описаний в Розділі 3. Отримане значення статистики t_p порівнюють з значеннями квантилів t-розподілу Стьюдента, при кількості ступенів свободи df = n – 2. Якщо $|t_p| > t_{кр(\alpha,df)}$ то нульову гіпотезу відкидають, і вважають, що вибірковий коефіцієнт кореляції значимо відрізняється від нуля, а отже - X і Y корельовані між собою.

Такий підхід досить простий як з математичної точки зору, так і з точки зору програмної реалізації. Суттєвий його недолік полягає в тому, що при $|r| < 0.5$ він втрачає точність. Тому його переважно застосовують для перевірки гіпотези про нульову кореляцію, але рідше для побудови довірчих інтервалів.

Оскільки розподіл коефіцієнта кореляції є несиметричним, пряма побудова довірчого інтервалу для r доволі складна. Тому на практиці широко застосовується ***перетворення Фішера*** (англ. *Fisher's* z-*Transformation*). Вводиться нова змінна

$$z = \frac{1}{2} ln\left(\frac{1+r}{1-r}\right) = arctanh(r)$$

яка для достатньо великих вибірок має наближений до нормального розподіл з математичним очікуванням

$$z = \frac{1}{2} ln\left(\frac{1+\rho}{1-\rho}\right)$$

та стандартним відхиленням

$$\sigma_z = \frac{1}{\sqrt{n-3}}$$

Тоді кордони - нижній $I_l(z)$ та верхній $I_u(z)$ - довірчого інтервалу для величини z визначаються за стандартною формулою нормального розподілу:

$$I_l(z) = z - z_{\frac{\alpha}{2}} \frac{1}{\sqrt{n-3}}$$

$$I_u(z) = z + z_{\frac{\alpha}{2}} \frac{1}{\sqrt{n-3}}$$

де $z_{\frac{\alpha}{2}}$ — квантиль стандартного нормального розподілу.

Після цього межі інтервалу перетворюють назад до шкали коефіцієнта кореляції:

$$I_l(r) = \frac{e^{2I_l(z)} - 1}{e^{2I_l(z)} + 1} = tanh(I_l(z))$$

$$I_u(r) = \frac{e^{2I_u(z)} - 1}{e^{2I_u(z)} + 1} = tanh(I_u(z))$$

Якщо отриманий за вказаною процедурою довірчий інтервал містить значення «0», то на заданому рівні довіри не можна стверджувати про наявність лінійного зв'язку між змінними. Якщо ж нуль не належить довірчому інтервалу, це узгоджується з результатом перевірки гіпотези про відсутність кореляції і свідчить про наявність статистично значущого зв'язку між замінними.

p-value у контексті кореляції Пірсона відповідає на запитання: «Яка ймовірність отримати такий самий або сильніший коефіцієнт кореляції за умови, що насправді кореляції в генеральній сукупності немає?» Таким чином, аналіз коефіцієнта кореляції Пірсона зазвичай включає два взаємодоповнюючі етапи: перевірку статистичної гіпотези щодо значущості кореляції за допомогою *p_value* та побудову довірчого інтервалу, який дозволяє оцінити можливий діапазон значень істинного коефіцієнта кореляції у генеральній сукупності.

У сучасному статистичному аналізі довірчі інтервали для коефіцієнта кореляції Пірсона майже завжди будують за допомогою z-перетворення Фішера, оскільки цей метод забезпечує кращу статистичну точність і коректніше відображає властивості розподілу оцінки кореляції. Однак треба зауважити, що при роботі з малими за об'ємом вибірками дослідник має бути дуже обережним щодо результатів, оскільки точність, що забезпечується методом z-перетворенням теж стає доволі обмеженою.

Насамкінець приведемо умови, виконання яких вимагається для коректного застосування коефіцієнта кореляції Пірсона:

- Лінійність зв'язку: Метод виявляє лише лінійні залежності. Нелінійні зв'язки можуть мати значення r близькі до 0 навіть при сильному взаємозв'язку.
- Нормальність розподілу: Обидві змінні повинні мати приблизно нормальний розподіл. Особливо це важливо для малих ($n < 30$) вибірок.
- Гомоскедастичність: Дисперсія залишків має бути постійною на всіх діапазонах значень змінних.

– Відсутність викидів: Окремі екстремальні значення можуть суттєво спотворити значення r.
– Неперервні дані: Змінні мають бути виміряні у інтервальній шкалі або у шкалі відносин.

4.4. Визначення коефіцієнта кореляції Пірсона засобами екосистеми Python

У практичному аналізі даних обчислення коефіцієнта кореляції Пірсона, перевірка його статистичної значущості та визначення довірчих інтервалів зазвичай виконуються за допомогою спеціалізованих програмних засобів. У середовищі Python такі можливості надають бібліотеки *numpy, scipy, pandas*, а також деякі інструменти з пакету *scikit-learn*. Вони реалізують як обчислення самого коефіцієнта кореляції, так і процедури статистичного тестування.

Найпростіший спосіб обчислення коефіцієнта кореляції Пірсона реалізовано у бібліотеці *numpy*. Функція *numpy.corrcoef()* обчислює матрицю коефіцієнтів кореляції між змінними, представленими у вигляді масивів даних. У випадку двох змінних відповідний елемент цієї матриці є оцінкою коефіцієнта кореляції Пірсона r. Дана функція виконує лише обчислення коефіцієнта кореляції і не надає інформації щодо статистичної значущості або довірчих інтервалів.

У бібліотеці *pandas*, яка широко використовується для обробки табличних даних, аналогічні обчислення можуть бути виконані за допомогою методів *Series.corr()* або *DataFrame.corr()*. За замовчуванням ці методи також використовують формулу коефіцієнта кореляції Пірсона та дозволяють швидко визначати кореляційні зв’язки між числовими ознаками у наборі даних.

Більш широкі можливості статистичного аналізу надає модуль *scipy.stats*. Функція *scipy.stats.pearsonr()* реалізує обчислення коефіцієнта кореляції Пірсона разом із перевіркою гіпотези про відсутність лінійного зв’язку між змінними. Результатом виконання цієї функції є два значення: оцінка коефіцієнта кореляції r та відповідне значення *p_value*. Останнє визначається на основі t-статистики, що має розподіл Стьюдента з *n-2* ступенями свободи, і використовується для перевірки нульової гіпотези $H_0 : \rho = 0$.

Функція *scipy.stats.pearsonr()* в якості результату повертає об’єкт, який окрім зазначених вище атрибутів має спеціальний метод *confidence_interval()*

дозволяє безпосередньо визначати довірчий інтервал для коефіцієнта кореляції. При цьому використовується перетворення Фішера, яке переводить коефіцієнт кореляції у змінну з наближено нормальним розподілом, після чого обчислюється стандартний довірчий інтервал і виконується зворотне перетворення до шкали коефіцієнта кореляції. При розрахунках за замовчуванням використовується рівень довіри 0.95, але користувач має змогу задати довільне, потрібне йому значення.

4.5. Коефіцієнт кореляції Спірмена

Коефіцієнт рангової кореляції запропонований Чарльзом Спірменом (англ. *Spearman's Rank Correlation Coefficient*), відноситься до *непараметричних показників зв'язку між змінними, виміряними в ранговій шкалі*. При розрахунку цього коефіцієнта не робиться ніяких припущень про характер розподілу ознак у генеральній сукупності. Цей коефіцієнт визначає ступінь тісноти зв'язку порядкових (або рангових) величин.

Нагадаємо, *ранги - це порядкові номери елементів в їх ранжованому ряду*. Звертаємо увагу – не конкретне виміряне значення (оцінка), а саме номер, що даний вимір зайняв у впорядкованому списку наданих значень. Наприклад, в ряду {1, 4, 8, 8, 12} ранг числа 4 дорівнює 2, а ранг числа 12 дорівнює 5. Ранжувати ознаки можливо в або від менших значень ознаки до більших, або навпаки.

Якщо проранжувати одні і ті самі об’єкти за двома ознаками, зв'язок між якими вивчається (оцінка та військове звання, місце в університетських спортивних змаганнях та рік навчання учасника тощо), повний збіг рангів означає максимально тісний прямий зв'язок, а повна протилежність рангів - максимально тісний зворотний (негативний) зв'язок. Чим ближче значення до нуля, тим менший кореляційний зв’язкок між наданими параметрами об’єкту.

Дамо формальне визначення. Припустимо, що N об'єктів можуть бути впорядковані як за ознакою X, так і за ознакою Y. Нехай $R_i(x)$ - ранг i-го об’єкту за ознакою X $(i = 1, N)$, а $R_i(y)$ - за ознакою Y. Приймемо за міру розбіжності значень для i-ого об’єкту величину:

$$d_i = R_i(x) - R_i(y)$$

Щоб запобігти ефекту «компенсації» позитивних та негативних значень розбіжностей між двома рядами будемо розглядати величину:

$$D(X,Y)=\sum_{i=1}^{N} d_i^2$$

Для того, щоб отримана статистика була подібна до коефіцієнту кореляції, потрібно щоб виконувалися обидві з двох вимог:

- коефіцієнт кореляції приймав значення +1, якщо всі ранги співпадають,
- коефіцієнт кореляції приймав значення -1, якщо рангові ряди мають строго зворотню направленість (наприклад, $R_i(x)=\{1,2,3,4,5\}; R_j(x)=\{5,4,3,2,1)\}$).

Будемо будувати наш коефіцієнт рангової кореляції таким чином, щоб він задовольняв вказаним вище вимогам, у вигляді $1-f*D(X,Y)$. Зазначимо, що в випадку, якщо всі об'єкти мають однакові ранги по першому та по другому параметру, то тоді $d_i^2=0$ для кожного *i*, і незалежно від значення f коефіцієнт стає рівним 1, що відповідає першій з вказаних вище вимог.

Вибір *f* так, щоб задовольнити другу вимогу більш складний. Доведення можна знайти, наприклад, в [58]. Зокрема, в роботі показано, що якщо покласти *f*=6, тобто прийняти коефіцієнт кореляції рівним:

$$r_s=1-\frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2-1\right)}$$

то обидві зазначені вище умови будуть виконуватися. Саме ця статистика носить назву *коефіцієнт (рангової) кореляції Спірмена*. Значення цієї статистики лежать в інтервалі +1 і -1 і він, як і коефіцієнт кореляції Пірсона, може бути позитивним та негативним, характеризуючи і «силу», і спрямованість зв'язку між двома ознаками, виміряними у ранговій шкалі.

Цікаво зазначити, що розглянутий у Розділі 4.4 *коефіцієнт кореляції Пірсона є узагальненням коефіцієнту кореляції Спірмена на випадок сильних шкал.* Особливо це помітно при великих вибірках (>50). Тобто, для таких значень можливо використовувати для оцінки критичних значень можна брати статистику:

$$|t| = \frac{|r_s|}{\sqrt{1 - r_s^2}} \sqrt{n-2}$$

Якщо розраховане значення *t-статистики* менше табличного при заданому числі ступенів свободи, статистична значимість взаємозв'язку відсутня. Якщо більше - кореляційний зв'язок вважається статистично значущим.

Якщо при розрахунку коефіцієнту кореляції Спірмана деякі значення виявляються однаковими, то вони отримають однакові ранги. В такому випадку говорять про «зв'язки», або про «*пов'язані ранги*». В такому випадку процедурі розрахунку коефіцієнту кореляції має передувати *заміна таких пов'язаних рангів на середнє арифметичне їх порядкових номерів*. Однак заміна середнім значенням – не абсолютно коректний підхід і насправді треба використовувати більш складні формули. Більш точне значення при наявності пов'язаних рангів дає формула дає спеціальна модифікація статистики, що враховує можливі повтори рангів. Нехай у нас виявилось *J* груп однакових значень рангів по ознаці *X* та *K* груп однакових рангів по ознаці *Y*. В такому разі рекомендується використовувати формулу:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2 + A + B}{n\left(n^2 - 1\right)}$$

$$A = \frac{1}{12}\sum_{j}^{J}\left(A_j^3 - A_j\right)$$

$$B = \frac{1}{12}\sum_{k}^{K}\left(B_k^3 - B_k\right)$$

де A_j-кількість однакових рангів в групі j (по ознаці X);

B_k-кількість однакових рангів в групі k (по ознаці Y).

Позатим зазначимо, що чим більше є даних що повторюються, тим менш точні значення дає статистика Спірмана.

Цікавою властивістю коефіцієнта кореляції Спірман є те, що він допускає завдання вхідних даних як в безперервній шкалі, так і в порядкової. Відомо, що якщо параметричні методи, що вимагають нормального розподілу даних, застосувати до даних з іншим типом розподілу, це приведе до помилкового

висновку (Про говорилося в розділі, що був присвячений використанню параметричних та непараметричних методів статистики – див. розділ 3.2.6). Навпаки, непараметричні методи допустимо застосовувати і в разі нормального розподілу, хоча в такому випадку чутливість цих методів буде трохи нижче чутливості параметричних методів. Що стосується коефіцієнта рангової кореляції Спірмена, то він в таких випадках теж програє коефіцієнту кореляції Пірсона, хоча і досить незначно.

Треба взяти до уваги, що попри спорідненість, коефіцієнти кореляції Пірсона та Спірмена по різному використовують інформацію про тип шкали даних. Так, значення коефіцієнту кореляції Спірмена дорівнює 1, коли дві змінні монотонно пов'язані між собою, навіть у випадку, коли це відношення не є лінійним. З іншого боку, коефіцієнт кореляції Пірсона в такому разі дасть результат, що відрізняється від 1 (див. рис. 4.4). З іншої сторони, кореляція Спірмена менш чутлива, ніж кореляція Пірсона, до сильних викидів, які знаходяться в хвостах обох вибірок. Це має місце тому, що коефіцієнт кореляції Спірмена обмежує викид значенням його рангу.

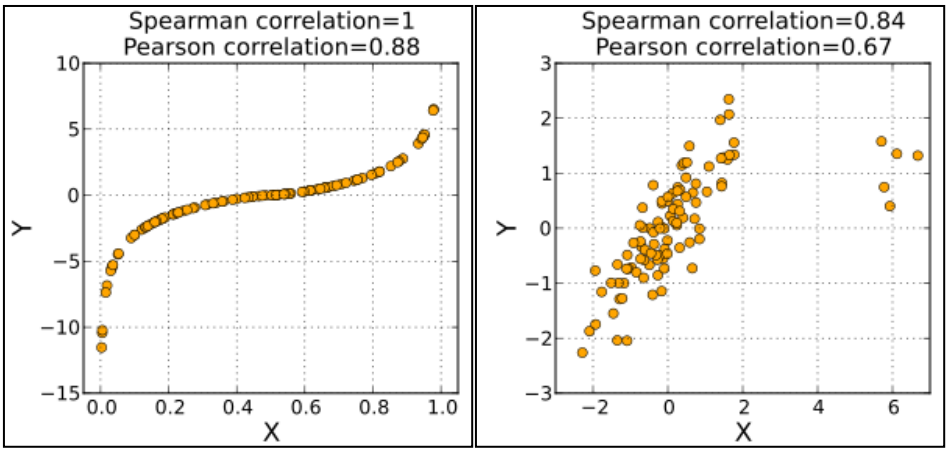


Рис. 4.4 Значення коефіцієнтів кореляції Пірсона та Спірмена в залежності від особливостей залежностей значень даних

В пакетах екосистеми Python існують готові реалізації для обрахунку коефіцієнту кореляції Спірмена. Так, в пакеті scipy.stats цей метод називається *scipy.stats.spearmanr()*, і повертає він об’єкт з двох елементів – самого коефіцієнту кореляції та *p_value.*

rs,pv=sps.spearmanr(arr1, arr2)
scipy.stats.spearmanr(arr1, arr2).correlation
scipy.stats.spearmanr(arr1, arr2).pvalue

В пакеті *Pandas* для розрахунку коефіцієнту кореляції Спірмена використовується той самий метод *corr()*, що і в випадку кореляції Пірсона, але при цьому треба явно задавати параметр *method*:

exampl1['Параметр X1'].corr(exampl1['Параметр X2'],method='spearman')

Розрахунок критичного значення довірчого інтервалу коефіцієнту кореляції Спірмена виконується або безпосередньо:

*t_r=(abs(rs)/math.sqrt(1-pow(rs,2)))*math.sqrt(len(exampl1)-2)*

або за допомогою відповідних методів бібліотек, наприклад:

scipy.stats.t.interval(1-0.05,df=len(exampl1)-2)

4.6. Виявлення залежності між даними, які виміряні в номінальних шкалах

У практиці статистичного аналізу та аналізу даних часто зустрічаються змінні, які *не мають числового змісту*, а лише позначають належність об'єктів до певних категорій. Такі змінні вимірюються в номінальній шкалі. На відміну від кількісних або порядкових шкал, номінальні шкали не допускають операцій порівняння типу «більше–менше» чи обчислення середніх значень. Проте між номінальними змінними також можуть існувати суттєві статистичні залежності, які потребують спеціальних методів виявлення та аналізу.

Оскільки номінальні дані не мають числового змісту, класичний кореляційний коефіцієнт Пірсона для них непридатний, хочаб через те, що для даних, які виміряні в номінальній шкалі, неможливо обчислити середні значення. В даному разі аналіз має базуватися на:

- порівнянні частот появи категорій;
- побудові таблиць спряженості;

– оцінці відхилення емпіричних частот від очікуваних.

Основне запитання аналізу звучить наступним чином: *Чи залежить розподіл однієї номінальної змінної від значень іншої?*

Існують різні критерії, що дозволяють виявляти взаємозв'язок номінальних ознак. Найбільш популярним з них є критерій, що ґрунтується на аналізі ***таблиць спряженості*** (англ. *Contingency Tables*; інший можливий переклад - таблиці зв'язаності).

Основним статистичним методом аналізу залежності між змінними є χ^2-*критерій незалежності Пірсона* (див. розділ 3.3.5.1). Але, як відомо, χ^2-критерій дозволяє лише перевірити наявність або відсутність залежності між ознаками, але не оцінює її силу, тобто не є мірою зв'язку (як, наприклад, кореляція). Крім того, значення статистики χ^2- ще й змінюється в довільному діапазоні, а також залежить від об'єму самої вибірки N. Для дослідження вказаних питань, зокрема, щоб отримати саме міру взаємодії, Пірсон запропонував використовувати критерій зв'язаності, що також носить його ім'я (***коефіцієнт взаємної спряженості Пірсона***):

$$K_p = \sqrt{\frac{\chi^2}{N + \chi^2}}$$

Легко бачити, що $K_p = 0$ в разі відсутності зв'язку (бо за таких умов $\chi^2 = 0$). Чим більше зв'язок між ознаками, тим більше коефіцієнт взаємної зв'язаності Пірсона наближається до 1. Як і інші коефіцієнти, що описуються нижче, критерій зв'язності Пірсона є сенс використовувати, якщо в результаті попереднього аналізу виявиться, що нема умов для відхилення нульової гіпотеза відносно χ^2.

Окрім коефіцієнт взаємної спряженості Пірсона, в якості міри сили корельованості використовується описаний в розділі 3.3.5.1 *Коефіцієнт взаємної спряженості Крамера* (V):

$$V = \sqrt{\frac{\chi^2}{\mathrm{N}*\min(r-1, c-1)}}$$

де r, c-кількість рядків та стовпчиків у таблиці відповідно. Можна легко побачити, що можливі значення цього коефіцієнту лежать в межах від 0 до 1.

Ще один аналогічний ***Коефіцієнт взаємної спряженості Чупрова*** обчислюється як відношення статистики χ^2 до значення цієї статистики, яке є максимального можливим для даної таблиці. Відповідно, за цим визначенням він може приймати значення від 0 до 1. При цьому 0 означає відсутність зв'язку, а 1 повну, функціональну взаємозалежність ознак:

$$K_C = \sqrt{\frac{\chi^2}{N\sqrt{(r-1)(s-1)}}}$$

Треба розуміти, що незважаючи на відсутність числового змісту, номінальні дані можуть містити важливу і корисну інформацію про структуру та закономірності досліджуваних процесів, особливо у соціальних дослідженнях, кібербезпеці, маркетингу та аналізі інформаційних систем. Тому для окремих випадках, що найчастіше зустрічаються на практиці, були запропоновані інші коефіцієнти спряжності. Так ***коефіцієнт асоціації Юла*** (англ. *Yule's Q*) використовується для оцінки сили та напряму зв'язку між двома *дихотомічними номінальними ознаками*. На відміну від критеріїв узгодженості або незалежності, він не перевіряє гіпотезу, а кількісно описує асоціацію. Таким чином: якщо χ^2-критерій відповідає на питання «чи існує зв'язок?», то коефіцієнт Юла — на питання «наскільки сильний і в якому напрямі діє цей зв'язок?». Дихотомічність шкал виміру означає, що а дані подано (або можна привести) до формату таблиці спряженості розмірності 2×2. Коефіцієнт асоціації Юла визначається як:

$$Q = \frac{ad - bc}{ad + bc}$$

де в термінах, що застосовувалися при розгляді таблиці спряженості (див. розділ 3.3.5.1) , можна записати: $a = O_{1,1}; b = O_{1,2}; c = O_{2,1}; d = O_{2,2}$.

Властивості коефіцієнта Юла:

- $Q \in [-1,1]$;
- $Q=0$ — відсутність асоціації;

– Q>0 — пряма асоціація;
– Q<0 — обернена асоціація;
– |Q|→1 — сильна асоціація.

Треба ще раз підкреслити, що коефіцієнт Юла не є критерієм узгодженості; він не має власного *p_value*; застосовується після застосування χ^2-критерію або біноміального аналізу для *інтерпретації сили і напрямку зв'язку*. Тому коефіцієнт Юла не замінює χ^2, а доповнює його.

4.7. Виявлення залежності між даними, які виміряні в різнотипних шкалах

У реальних прикладних задачах аналізу часто виникає необхідність дослідження залежностей між змінними, вимірянми в різних шкалах вимірювання. Наприклад, кількісні показники можуть співвідноситися з якісними характеристиками, порядкові оцінки — з числовими вимірюваннями, а номінальні змінні — з неперервними параметрами. У таких випадках класичні методи кореляційного аналізу, як-от кореляція Пірсона (для двох кількісних змінних) або χ^2-критерій (для двох категоріальних змінних) не завжди можуть бути застосовними, і вибір статистичного інструменту має ретельно враховувати тип шкал кожної змінної.

Найпоширеніші комбінації різнотипних шкал:
– номінальна – кількісна;
– порядкова (рангова) – кількісна;
– дихотомічна – кількісна (окремий випадок номінальної);
– номінальна – порядкова.

Наприклад, аналіз різниці середнього часу відгуку сервера для різних типів операційної системи потребує аналізу даних, що виміряні в номінальній та кількісній шкалах. Дослідження зв'язку між рівнем критичності інциденту («низький», «середній», «високий») та часом його усунення вимагають одночасної роботи з порядковою та кількісною змінними. Якщо номінальна змінна має лише два значення (наприклад, «атака»/«не атака»), застосовується точково-бісеріальний коефіцієнт кореляції(див. розділ 4.8), який є окремим випадком кореляції Пірсона. Пошук відповіді на питання: «чи пов'язаний тип інциденту інформаційної безпеки з рівнем його критичності?» передбачає роботу з даними, що виміряні в номінальній та порядковій шкалами тощо.

Ситуація одночасної наявності кількісної та категоріальної змінних - найпоширеніший випадок, коли порівнюють середні значення кількісної змінної у групах, визначених категоріальною змінною. Інструментом аналізу в такій постановці задачі може слугувати *t-критерій Стьюдента* (див. Розділ 3.3.1.2). В даному разі він використовується для перевірки, чи відрізняються *середні значення кількісної змінної між двома групами* (наприклад, середній рівень доходу чоловіків та жінок). Якщо одна змінна кількісна, а інша — порядкова (наприклад, рівень задоволеності: «Низький», «Середній», «Високий»), то використання непараметричних кореляційних методів є більш надійним. Зокрема, *рангова кореляція Спірмена* (див.розділ 4.5) вимірює силу та напрямок монотонного зв'язку (не обов'язково лінійного) між двома змінними. Він вимагає попереднього перетворення кількісної змінної на ранги і подальшого обчислення кореляції між рангами обох змінних. У випадку використання порядкової та номінальної шкал аналіз ґрунтується на порівнянні розподілів порядкової змінної між категоріями номінальної з використанням *χ^2-критерію незалежності* (див.розділ 3.3.5.1) або *коефіцієнта асоціації* для порядкових даних. Як бачимо, вибір адекватного інструменту є досить складною задачею, вимагає гнучкого підходу та відповідного обґрунтування. Правильне поєднання шкал і статистичних інструментів дозволяє отримати коректні та інтерпретовані результати навіть у складних прикладних задачах аналізу даних.

4.8. Точково-бісеріальний та бісеріальний коефіцієнти кореляції

У статистичному аналізі даних часто виникає потреба оцінити зв’язок між кількісною змінною та *дихотомічною змінною*, яка виміряна в номінальній шкалі та має природний (реальний) бінарний характер. Для таких ситуацій застосовується ***точково-бісеріальний коефіцієнт кореляції*** (англ. *Point-Biserial Correlation Coefficient*). Точково-бісеріальний коефіцієнт є спеціальним випадком коефіцієнта кореляції Пірсона і *дозволяє кількісно оцінити силу та напрям лінійного зв’язку між змінними різного типу*.

Точково-бісеріальний коефіцієнт застосовується, коли виконуються такі умови. По-перше, перша змінна - кількісна (інтервальна або шкала відношень); друга змінна - дихотомічна номінальна. По-друге, дихотомічна змінна має два значення, наприклад: «так/ні», «успіх/неуспіх»; спостереження між собою незалежні. Прикладам таких даних є пари «факт виявлення вторгнення (так/ні)» та «обсяг мережевого трафіку»; «наявність сертифіката безпеки» та «середній час

реагуванням на інцидент»; «складений/не складений іспит» і «кількість отриманих балів» тощо;

Точково-бісеріальний коефіцієнт обчислюється як:

$$r_{pb} = \frac{\overline{x}_1 - \overline{x}_0}{s_x} * \sqrt{pq} = \frac{\overline{x}_1 - \overline{x}_0}{s_x} * \sqrt{\frac{n_0 n_1}{N}}$$

де $\overline{x}_1$ — середнє значення кількісної змінної для групи з кодом 1;

$\overline{x}_0$ — середнє значення для групи з кодом 0;

s_x — середньоквадратичне відхилення кількісної змінної;

p — частка одиниць серед усіх значень Y;

$q = 1 - p$;

n_0 - кількість об’єктів з кодом 0 у вибірці;

n_1 - кількість об’єктів з кодом 1 у вибірці;

$N = n_0 + n_1$.

Значення r_{pb} належать інтервалу [-1,1], знак коефіцієнта визначає напрям зв’язку; значення, близькі до 0 свідчать про слабкий або відсутній зв’язок; великі за модулем значення вказують на сильний зв’язок.

Можна показати, що точково-бісеріальний коефіцієнт математично пов’язаний з t-критерієм для незалежних вибірок:

$$t = r_{pb} \sqrt{\frac{n-2}{1-r_{pb}^2}}$$

Це означає, що перевірка значущості r_{pb} еквівалентна перевірці різниці середніх. А також те, що коефіцієнт можна розглядати як міру сили взаємозв’язку між змінними. Основна перевага r_{pb} полягає в його математичній простоті та прямій інтерпретації: *він безпосередньо вимірює, наскільки розподіл кількісної змінної відрізняється залежно від двох категорій бінарної змінної.* Це найбільш коректний коефіцієнт для оцінки зв'язку, коли дихотомічна змінна є природною.

У Python, звернувшись до методу *scipy.stats.pointbiserialr* (x, y), можна отримати як саме значення точкового бісеріального коефіцієнту кореляції, так і

його p_value для вибірок, що передаються у функцію в якості її вхідних параметрів.

Точково-бісеріальний коефіцієнт кореляції є ефективним інструментом аналізу зв'язку між кількісними та дихотомічними номінальними змінними. Він поєднує інтерпретаційну наочність з математичною строгістю та широко застосовується в сучасному прикладному аналізі даних.

В деяких реальних випадках дані дихотомічної шкали по своїй природі є не природніми (яким є групи «чоловіки»-«жінки» або «хворий»-«здоровий»), *а штучно приведені до такої шкали шляхом загрублення даних початково виміряних в більш сильних шкалах.* Наприклад, відбулася *попередня змінна показника* "успішність" в балах від 0 до 100, на показник «заліковано»/«не заліковано» (тобто, відбулося *огрублення шкали виміру*). Або один з показників – «середня швидкість вхідного трафіку, Мбіт/с» кількісний, виміряний в шкалі відношень, а інший – «спрацювання детектора DDoS» дихотомічний, *і є результатом є результатом порогового поділу прихованої неперервної величини (інтенсивності атаки)*. Умовою для отримання коректних результатів для такого варіанту представлення даних є розподіл кількісної змінної, який має бути наближено нормальним. *Якщо дихотомічна змінна не є наслідком розбиття неперервної, застосування точково-бісеріального коефіцієнта є некоректним.* Якщо ні - має застосовуватися ***бісеріальний коефіцієнт кореляції***. (англ. *Biserial Correlation Coefficient*).

Бісеріальний коефіцієнт визначається як:

$$r_b = \frac{\overline{x}_1 - \overline{x}_0}{s_x} * \frac{pq}{y}$$

де y — значення ординати стандартного нормального розподілу в точці поділу.

Значення бісеріального коефіцієнта кореляції лежить в інтервалі [-1,+1], знак коефіцієнта вказує на напрям зв'язку; ступінь наближення до +1 або -1 - силу взаємозв'язку, значення, наближені до 0, вказують на слабкий або відсутній зв'язок.

Контрольні запитання

1. Що означає взаємозалежність між вибірками?
2. Що таке кореляція і як вона інтерпретується?
3. Як визначити силу та напрямок кореляційного зв’язку?
4. Які значення може приймати коефіцієнт кореляції і що вони означають?
5. У чому сутність коефіцієнта кореляції Пірсона і коли його застосовують?
6. Чим відрізняється коефіцієнт Спірмена від коефіцієнта Пірсона?
7. Що таке бісеріальний коефіцієнт кореляції і коли його використовують?
8. Чому важливо враховувати тип даних (кількісні, порядкові, номінальні, дихотомічні) при виборі коефіцієнта кореляції?

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Панкратова Н. Д. Системний аналіз. Теорія та застосування : підручник. НАНУ, НТУУ “КПІ”, ІПСА НАНУ. Київ : Наук. думка, 2018. 347 с. URL: https://library.kpi.kharkov.ua/files/new_postupleniya/sistan.pdf (дата звернення: 15.04.2026)
2. Ситник Г.П., Комаха Л.Г., Рудик А.О. Основи теорії систем та системного аналізу: навчальний посібник . ТОВ «Академпрес». Київ, 2024. 160 с. URL: https://ipacs.knu.ua/pages/dop/391/files/16db659a-6764-4c03-86b2-f517aec290b6.pdf (дата звернення: 15.04.2026)
3. Бродський Ю.Б. Системний аналіз та теорія прийняття рішень: навч. посіб. в 3-х частинах.Частина 1: Системологія [Електронний ресурс]. Житомир: Державнийуніверситет «Житомирська політехніка», 2022. 92 с. URL: https://learn.ztu.edu.ua/mod/resource/view.php?id=178901 (дата звернення: 15.04.2026)
4. Прокопенко Т. О. Теорія систем і системний аналіз : навч. посіб. [Електронний ресурс]. М-во освіти і науки України,Черкас. держ. технол. ун-т. Черкаси : ЧДТУ, 2019. 139 с. URL: https://elib.chdtu.edu.ua/e-books/3215 URL: https://elib.chdtu.edu.ua/e-books/3215 (дата звернення: 15.04.2026)
5. Системний аналіз та прийняття рішень в інформаційній безпеці : підручник./ В . Л . Бурячок , С . В . Толюпа , А . О . Аносов , В . А . Козачок , Н . В . Лукова - Чуйко. К .: ДУТ , 2015. 345 с . URL: https://duikt.edu.ua/uploads/l_1242_54311567.pdf (дата звернення: 15.04.2026)
6. Якименко Ю.М., Савченко В.А., Легомінова С.В. Системнийаналіз інформаційної безпеки: сучасні методи управління:підручник. Київ: Державний університет телекомунікацій, 2022.308 с. URL:https://duikt.edu.ua/uploads/l_2230_88161692.pdf (дата звернення: 15.04.2026)
7. Сілін Є.С., Кадубовський О.А. Основи кібербезпеки : навчальний посібник [Електронний ресурс]. Дніпро, 2023. 200 с. URL: https://ddpu.edu.ua/fmk/publications/manuals/manual_18.pdf (дата звернення: 15.04.2026)
8. Tolboom R. Computer Systems Security: Planning for Success. New Jersey Institute of Technology, Open Textbook Library. 2023. 118p. URL: https://open.umn.edu/opentextbooks/formats/3421 (дата звернення: 15.04.2026)

9. Якименко Ю. М., Мужанова Т. М., Легомінова С. В. Системний аналіз технічних систем забезпечення інформаційної безпеки. Кібербезпека:освіта, наука, техніка. N 4(12), 2021, с.36-50. DOI: 10.28925/2663-4023.2021.12.3650 (дата звернення: 15.04.2026)
10. Пановик У. П., Ткачук Р. Л. Кібербезпека через призму системного аналізу. COMPUTER TECHNOLOGIES OF PRINTING 2023/1(49) с.197-208. URL: https://sci.ldubgd.edu.ua/jspui/handle/123456789/12405 (дата звернення: 15.04.2026)
11. Приступа Б.В., Герасимюк Н.В., Рожковський Я.В. Використання штучного інтелекту в кібербезпеці. Інформатика та математичні методи в моделюванні, 2025, Том 15, N 1. с.115-124.
URL: http://immm.op.edu.ua/files/archive/n1_v15_2025/immm_n1_v15_2025.pdf (дата звернення: 15.04.2026)
12. Mohamed N. Artificial intelligence and machine learning in cybersecurity: a deep dive into state-of-the-art techniques and future paradigms. Knowl Inf Syst 67, 6969–7055 (2025). URL: https://doi.org/10.1007/s10115-025-02429-y (дата звернення: 15.04.2026)
13. Методи штучного інтелекту в кібербезпеці [Електронний ресурс] : навч. посіб. для здобувачів спец. 125 «Кібербезпека» / КПІ ім. Ігоря Сікорського ; уклад.: І.В. Стьопочкіна, О.М. Новіков. Київ : КПІ ім. Ігоря Сікорського, 2022. 82 с. URL: https://ela.kpi.ua/server/api/core/bitstreams/5b3026af-3e10-4297-a250-3aac4f805437/content (дата звернення: 15.04.2026)
14. Лисенко Н. О., Мазуренко В. Б., Федорович А. І., Астахов Д. С., Стаценко В. І. Огляд математичних методів у системах виявлення та попередження кіберзагроз. Актуальні проблеми автоматизації та інформаційних технологій. Том 25. 2021.- с.91-102. DOI: 10.15421/432110 (дата звернення: 15.04.2026)
15. Hero A., Kar S., Moura J., Neil J., Poor H.V., Turcotte M.,Xi B. Statistics and Data Science for Cybersecurity [Електронний ресурс]: Harvard Data Science Review(HDSR), Issue 5.1, Winter 2023. DOI: 10.1162/99608f92.a42024d0 URL: https://hdsr.mitpress.mit.edu/pub/koyzu1te/release/1 (дата звернення: 15.04.2026)
16. Юрчак І., Хіч А., Оксентюк В. Розуміння великих мовних моделей: майбутнє штучного інтелекту. Computer design systems. Theory and Practice. Vol. 6, No. 2, 2024. с.51-60. URL: https://science.lpnu.ua/sites/default/files/journal-paper/2024/oct/36178/626.pdf (дата звернення: 15.04.2026)

17. Мотало В. Аналіз шкал вимірювань. Вимірювальна техніка та метрологія, No 76, 2015 р. - с.21-35.
18. Franco V.R.,Heene M., Domingue B., Jiménez M. [Електронний ресурс] An uncomplicated introduction to measurement theory. 47p. December 2023 DOI: 10.31234/osf.io/cm4he URL:https://www.researchgate.net/publication/379545245 (дата звертання 20.04.2026)
19. Scales of Measurement in Business Statistics. [Електронний ресурс]. URL: https://www.geeksforgeeks.org/data-science/scales-of-measurement-in-business-statistics/ (дата звертання 20.04.2026)
20. Types of data and the scales of measurement [Електронний ресурс]. URL: https://studyonline.unsw.edu.au/blog/types-of-data (дата звертання 20.04.2026)
21. Ціделко В.Д., Яремчук Н.А., Гальовська М.В. Репрезентативна теорія вимірювань в сучасній метрології. нформаційні системи, механіка та керування : науково-технічний збірник. 2008. Вип.1. с.48-57. URL:https://ela.kpi.ua/server/api/core/bitstreams/48b4ecfd-2a03-4501-a001-08182972e67d/content (дата звертання 20.04.2026)
22. Гнєденко Б. В. Курс теорії ймовірностей : підручник / Б. В. Гнєденко – К.:Видавничо-поліграфічний центр "Київський університет", 2010.– 464 с.
23. Медведєв М.Г., Пащенко І.О. Теорія ймовірностей та математична статистика : підручник. Київ : Видавництво Ліра-К, 2024. -536 с.
24. Горбачук, В. М. Теорія ймовірностей та математична статистика [Електронний ресурс] : підручник для здобувачів ступеня бакалавра за технічними та економічними спеціальностями – Київ : КПІ ім. Ігоря Сікорського, 2023. – 351 с. URL:https://ela.kpi.ua/handle/123456789/52357
25. Statistics - A Full University Course on Data Science Basics. [Електронний ресурс] URL: https://www.youtube.com/watch?v=xxpc-HPKN28 (тата звертання 25.04.2026)
26. Шарадкін Д.М., Субач І.Ю., Микитюк А.В. Інструментальні засоби Python для моделювання та системного аналізу часових рядів при вирішенні задач кіберзахисту інформаційно-комунікаційних систем: навч. пос. / Шарадкін Д.М., Субач І.Ю., Микитюк А.В.; ІСЗЗІ КПІ ім. Ігоря Сікорського. – Київ : КПІ ім. Ігоря Сікорського, 2023.- 139 с. (https://ela.kpi.ua/handle/123456789/52426)
27. Knuth, K.H. “Optimal Data-Based Binning for Histograms”. arXiv:0605197, 2013, 30p. URL: https://arxiv.org/abs/physics/0605197

28. Coefficient of Skewness. [Електронний ресурс] URL: https://www.geeksforgeeks.org/maths/coefficient-of-skewness/ (дата звертання 25.04.2026)
29. Vila R., Balakrishnan N., & Saulo H. An upper bound and a characterization for Gini's mean difference based on correlated random variables. Statistics & Probability Letters, 207, 2024. 11з. https://doi.org/10.48550/arXiv.2301.07229 (дата звератння 25.04.2026)
30. Hewage, S. A Nonparametric K-sample Test for Variability Based on Gini's Mean Difference. J Stat Theory Appl 24, 2025. p334–353. https://doi.org/10.1007/s44199-025-00112-3 (дата звертання 25.04.2026)
31. Comparing Standard Deviation and Median Absolute Deviation (MAD) Metrics. [Електронний ресурс]. URL: https://www.numberanalytics.com/blog/comparing-standard-deviation-and-mad-metrics
32. Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median, Journal of Experimental Social Psychology, Volume 49, Issue 4, July 2013, Pages 764-766. https://orbilu.uni.lu/bitstream/10993/59204/1/Leys_MAD_final-libre.pdf (дата звертання 25.04.2026)
33. Поправка Бесселя. [Електронний ресурс]. URL: https://uk.wikipedia.org/wiki/Поправка_Бесселя (дата звернення 25.04.2026)
34. Agrawal A. Bessel's Correction: Why Do We Divide by n−1 Instead of n in Sample Variance? [Електронний ресурс]. URL: https://towardsdatascience.com/bessels-correction-why-do-we-divide-by-n-1-instead-of-n-in-sample-variance-30b074503bd9 (дата звернення 25.04.2026)
35. https://baguzin.ru/wp/srednekvadratichnoe-otklonenie-i-srednee-absolyutnoe-otklonenie/ (дата звернення 25.04.2026)
36. Taleb N.N. Statistical Consequences of Fat Tails: Real World Preasymptotics, Epistemology, and Applications (Technical Incerto). Stem Academic Press. 2020.-441p.
37. Poisson Distribution. [Електронний ресурс] URL:https://www.geeksforgeeks.org/maths/poisson-distribution/ (дата звернення 25.04.2026)

38. Data Transformation: Standardization vs Normalization. [Електронний ресурс]. URL: https://www.kdnuggets.com/2020/04/data-transformation-standardization-normalization.html (дата звернення 28.04.2026)
39. Degrees of Freedom. [Електронний ресурс]. URL: https://www.geeksforgeeks.org/maths/degrees-of-freedom/ (дата звернення 28.04.2026)
40. List of probability distributions. [Електронний ресурс]. URL: https://en.wikipedia.org/wiki/List_of_probability_distributions (дата звернення 28.04.2026)
41. Non Normal Distribution. [Електронний ресурс]. URL: https://www.statisticshowto.com/probability-and-statistics/non-normal-distributions/ (дата звернення 28.04.2026)
42. Основні закони розподілу дискретних випадкових величин. [Електронний ресурс]. URL: https://new.mmf.lnu.edu.ua/wp-content/uploads/2020/04/Lektsiia-7.pdf (дата звернення 28.04.2026)
43. The Math Behind Kernel Density Estimation Exploring the foundations, concepts, and math of kernel density estimation. [Електронний ресурс]. URL: https://towardsdatascience.com/the-math-behind-kernel-density-estimation-5deca75cba38/ (дата звернення 30.04.2026)
44. The importance of kernel density estimation bandwidth. [Електронний ресурс]. URL: https://aakinshin.net/posts/kde-bw/ (дата звернення 30.04.2026)
45. Density Estimation. [Електронний ресурс]. URL: https://scikit-learn.org/stable/modules/density.html (дата звернення 30.04.2026)
46. Kernel Density Estimation (KDE) Plots. [Електронний ресурс]. URL: https://www.analyticsvidhya.com/blog/2025/05/kde-plot/ (дата звернення 30.04.2026)
47. Wasserman, Larry. All of Statistics: A Concise Course in Statistical Inference. New York: Springer, 2004. (Springer Texts in Statistics).- 458p. [Електронний ресурс]. URL: https://egrcc.github.io/docs/math/all-of-statistics.pdf. (дата звернення 30.04.2026)
48. Hypothesis Testing. [Електронний ресурс]. URL: https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/ (дата звернення 02.05.2026)

49. T-Test (Student's T-Test): Definition and Examples. [Електронний ресурс]. URL: https://www.statisticshowto.com/probability-and-statistics/t-test/ (дата звернення 02.05.2026)
50. Hypothesis Testing with Python: T-Test, Z-Test, and P-Value. [Електронний ресурс]. URL: https://medium.com/@techwithpraisejames/hypothesis-testing-with-python-t-test-z-test-and-p-values-code-examples-fa274dc58c36 (дата звернення 05.05.2026)
51. P-Value: Comprehensive Guide to Understand, Apply and Interpretation. [Електронний ресурс]. URL:https://www.geeksforgeeks.org/machine-learning/p-value/ (дата звернення 05.05.2026)
52. Mann-Whitney Table. [Електронний ресурс]. URL: https://real-statistics.com/statistics-tables/mann-whitney-table/ (дата звернення 05.05.2026)
53. SciPy documentation. Mannwhitneyu. [Електронний ресурс]. URL: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html (дата звернення 05.05.2026)
54. Mann–Whitney U test. [Електронний ресурс]. URL:https://en.wikipedia.org/wiki/Mann–Whitney_U_test (дата звернення 05.05.2026)
55. Kolmogorov-Smirnov Test [KS Test]: When and Where to Use [Електронний ресурс]. URL: https://dataaspirant.com/kolmogorov-smirnov-test/ (дата звернення 05.05.2026)
56. Spurious correlations. Correlation is not causation. [Електронний ресурс]. URL: https://www.tylervigen.com/spurious-correlations (дата звернення 12.05.2026)
57. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M. and Elhadad, N. Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-Day Readmission. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, 10-13 August 2015, 1721-1730. DOI: 10.1145/2783258.2788613 URL: https://www.microsoft.com/en-us/research/wp-content/uploads/2017/06/KDD2015FinalDraftIntelligibleModels4HealthCare_igt143e-caruanaA.pdf (дата звернення 12.05.2026)
58. Derivation of Spearman's rank correlation coefficient and example calculation using python [Електронний ресурс]. URL: https://laid-back-scientist.com/en/derivation-of-spearmans-rank-correlation-coefficient-and-example-calculation-using-python (дата звернення 12.05.2026)

ПРЕДМЕТНИЙ ПОКАЖЧИК