Міністерство освіти і науки України

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

І. К. Рисцов

Бізнес-статистика

Конспект лекцій

Рекомендовано Методичною радою КПІ ім. Ігоря Сікорського
як навчальний посібник для здобувачів ступеня бакалаврів
за освітньою програмою «Економічна аналітика»
спеціальності 051 «Економіка»

Електронне мережеве навчальне видання

Київ
КПІ ім. Ігоря Сікорського
2025

УДК [519.2+311.2+330.4]
С10

Автор: *Рисцов Ігор Костянтинович*, д-р фіз.-мат наук, доц.

Рецензент *Кузьмін А, В.*, канд. фіз.-мат. наук, доц., Київський національний університет ім. Тараса Шевченка

Відповідальний редактор *Бояринова К.О.*, д-р екон. наук, проф.

Гриф надано Методичною радою КПІ ім. Ігоря Сікорського
(протокол № 5 від 06.03.2025 р.)
за поданням вченої ради факультету менеджменту та маркетингу
(протокол № 6 від 30.01.2025 р.)

Рисцов І.К.

С10 **Бізнес-статистика**. [Електронний ресурс] : конспект лекцій : навч. посіб. для здобувачів ступеня бакалавра за освіт. програмою «Економічна аналітика» спец. 051 Економіка / І. К. Рисцов; КПІ ім. Ігоря Сікорського. – Електрон. текст. дані (1 файл). – Київ : КПІ ім. Ігоря Сікорського, 2025. – 49 с.

В конспекті лекцій «Бізнес-статистика» викладено основи математичної статистики та її застосування в аналізі та управління бізнес-процесами. Посібник призначений для здобувачів ступеня бакалавра за спеціальністю 051 Економіка, але він буде також корисним для всіх студентів факультету менеджменту та маркетингу.

УДК [519.2+311.2+330.4]

Реєстр. № 24/25-318. Обсяг 3,1 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

1. Вступ у бізнес-статистику

Бізнес це сукупність ділових процесів, спрямованих на задоволення потреб споживачів, з одного боку, і на отримання прибутку, з іншого. Суть бізнесу полягає у з'єднанні матеріальних, фінансових, трудових та інформаційних ресурсів з метою виробництва товарів і надання послуг, призначених для продажу іншим економічним агентам. При довгостроковому з'єднання цих ресурсів, виникає організація під назвою «*підприємство*», в рамках якої виконуються бізнес-процеси.

На сучасному етапі розвитку економічної думки деякі вчені пропонують розглядати підприємство з динамічної точки зору, як поєднання бізнес-процесів, замість традиційної статичної точки зору, як сукупність його підрозділів. Основними підставами для формування цієї точки зору є те, що на підприємстві треба управляти процесами (а не підрозділами) для контролю за використанням ресурсів та часом виконання робіт. Існують різні підходи до визначення сутності бізнес-процесу, які можна знайти в літературі. Підсумовуючи різні підходи, можна дати наступне визначення.

Бізнес-процеси – це сукупність видів діяльності підприємства, результатом яких є створення комерційного продукту.

Бізнес-статистика досліджує методи аналізу інформації з метою прийняття ефективних управлінських рішень. В сучасних умовах роль статистики в бізнесі ще більше зростає, коли підприємці вимушені приймати управлінські рішення в умовах нестаціонарного середовища, іноді не маючи повної та достовірної інформації не лише стосовно зовнішнього середовища, а й щодо внутрішніх бізнес-процесів. Саме статистичні методи допомагають здійснювати аналіз доступної інформації і розробляти на її основі обґрунтовані тактичні та стратегічні рішення. Отже, можна дати наступне визначення.

Бізнес-статистика це наука, яка вивчає *методи статистичного збору та обробки даних* для дослідження ринкового, комерційного та господарського стану суб'єктів підприємницької діяльності. На рис. 1.1 показано місце статистиці серед економіко-математичних методів.

Рис. 1.1. Складові економічної аналітики.

1.1. Статистичний аналіз бізнес-процесів

Аналіз бізнес-процесів – це систематична оцінка інформації з точки зору покращення процесу. Приводом для проведення такого аналізу, зазвичай, є конкурентне становище підприємства на ринку. Аналіз існуючих бізнес-процесів може прояснити поточну ситуацію та вказати шляхи для її поліпшення.

Будь-який процес можна охарактеризувати за такими критеріями:

- *Результативність процесу*: досягає процес необхідних результатів чи ні. Наприклад, якщо в результаті процесу «Виготовлення сметани», вийшла сметана – процес є результативним.
- *Ефективність* – здатність процесу забезпечувати результативність при мінімальному використанні ресурсів і відсутності непродуктивних витрат, тобто ефективність показує скільки ресурсів витрачає процес на отримання результату. Якщо відомо, що на приготування 1 кг сметани має піти 5 л молока, а було витрачено 8 л, то процес не є ефективним.

- *Визначеність процесу* означає ступінь відповідності результату процесу його документальному описанню. Якщо сметана була виготовлена згідно певного стандарту, то процес є визначеним.
- *Повторюваність* є однією з найважливіших характеристик процесу. Вона показує чи можна в результаті даного процесу отримувати однакові результати кожного разу, чи ні. Якщо підприємство постійно видає різну (за смаковими та іншими характеристиками) сметану одного виду, то щось не так з процесом.
- *Здатність до адаптації* – це характеристика гнучкості процесу, що означає здібність змінюватися залежно від умов, тобто здатність відповідати майбутнім змінам вимог споживачів або задовольняти певні поточні потреби деяких із них.
- *Тривалість процесу* – час, необхідний для його виконання, тобто проміжок часу між початком процесу і його завершенням.
- *Вартість* – це сукупність усіх витрат на виконання процесу один раз. Для цього необхідно підрахувати скільки молока витрачається на виготовлення, наприклад, 1 т сметани, скільки коштує робочий час працівників, які її виготовляють, скільки коштує використання устаткування тощо.

Для кожної мети підприємства повинен бути сформульований один або декілька *ключових показників ефективності* (КПЕ), або *ключових показників результативності* (КПР), що дозволяють здійснювати моніторинг їх дотримання. Показники, можна поділити на таки групи:

1. *Показники вхідних ресурсів*. Вони повинні відображати всі параметри і характеристики (якісні та кількісні), вхідних матеріальних та інформаційних ресурсів.
2. *Показники ефективності процесів*. Вони відображають ступінь виконання запланованої діяльності та досягнення запланованих результатів, а також зв'язок досягнутих результатів та витрачених ресурсів.
3. *Показники результативності процесів*. Ці показники характеризують готовність споживача придбати за певну ціну результат бізнес-процесу. Одна з основних цілей бізнеса це отримання прибутку для власників бізнесу. Тому типовими показниками результативності можуть бути сума очікуваного прибутку або рівень рентабельності.
4. *Показники задоволеності клієнта*. Спочатку необхідно відібрати декілька параметрів, які підприємство вважає важливими з точки зору клієнта та оцінити їх за обраною шкалою. Зазвичай оцінюють ставлення покупців до продукту або послуги, сервіс, привабливість реклами і ціни на товари порівняно з конкурентами. Після визначення декількох оптимальних кількісних і якісних показників задоволеності клієнта їх необхідно об'єднати в один єдиний критерій – індекс задоволеності клієнта (customer satisfaction index, CSI). Цей індекс є одним з видів підсумкових показників, з яким керівництво підприємства може знайомитися з певною періодичністю, наприклад, один раз на місяць. Даний індекс буде показувати наскільки добре підприємство задовольняє потреби своїх клієнтів.

Виділяють такі основні етапи статистичного аналізу бізнес-процесів:

1) планування дослідження (аналізу);
2) попередній аналіз інформації;
3) оцінка (отримання на основі даних числове значення невідомої величини);
4) перевірка гіпотез (дані використовуються для прийняття рішення про відповідність висунутого припущення дійсності).

Прикладами оцінки невідомих величин у бізнесі можуть бути: обсяг продаж у наступному місяці, реакція населення м. Київ на новий продукт, рівень браку у виробничому процесі і т. д.

Прикладами гіпотез, які можна перевірити з використанням статистичної інформації являються:

– «середній мешканець м. Київ в наступному місяці планує витратити на придбання нашого продукту не менше 500 грн.»;

– «засіб марки “Х” ефективно відбілює»;
– «нове медичне обладнання безпечне та ефективне».

1.2. Інформаційні ресурси

У більшості сфер економічної діяльності прийняття управлінських рішень базується на використанні й аналізі наявних ресурсів (фінансових, трудових, матеріальних тощо). Але для управління усіма ресурсами підприємства необхідна *інформація*, тому інформацію можна розглядати в якості одного з основних ресурсів підприємства. Саме інформація (або дані) сьогодні стають одним із найважливіших видів ресурсів, а згодом їх значимість буде тільки зростати. Одне із свідчень цього феномену полягає в тому, що дані стають *товаром*, сукупна вартість якого на ринку вже порівнялась з вартістю традиційних ресурсів.

1.2.1. Класифікація наборів даних

Зазвичай, набір даних містить інформацію по кожному з окремих об’єктів, які називаються *елементами* (або індивідами). У якості таких об’єктів може виступати що завгодно, що представляє інтерес для вивчення, наприклад, люди, домогосподарства, міста, прилади і т. д. Для кожного з об’єктів реєструють значення однакових ознак. Ознака, яку реєструють (наприклад, вартість), називається *змінною* (або атрибутом). Існують такі основні способи *класифікації наборів даних*:

– за кількістю ознак (змінних);
– за типом вимірювання;
– впорядкований чи невпорядкований (наприклад за часом).

За кількістю змінних розрізняють: *одновимірний*, *двовимірний* і *багатовимірний* набір даних. *Одновимірний* набір даних (одна змінна) містить інформацію тільки про одну ознаку, що зареєстрована для кожного об’єкта. Такий набір даних зазвичай аналізують на типове (середнє) значення та характеристику мінливості даних (дисперсію), а також виділяють специфічні особливості або проблеми в даних. Типовий приклад одновимірного набору даних наведений в таблиці 1.1.

Таблиця 1.1.

Фінансовий результат діяльності підприємств України в 1-3 кварталі 2017 р.

Галузь	Результат (млн. грн.)
Оптова та роздрібна торгівля, ремонт транспортних засобів	31890,0
Транспорт, складське господарство, поштова та кур’єрська діяльність	11178,5
Інформація та телекомунікації	11003,6
Фінансова та страхова діяльність	49330,3

Двовимірний набір даних (дві змінні) містить дві ознаки, значення яких реєструються для кожного об’єкта. Двовимірні дані дозволяють вивчити зв’язок між двома змінними та передбачити значення однієї змінної на основі значення іншої.

Багатовимірний набір даних містить три або більше ознак, значення яких реєструються для кожного об’єкта. Багатовимірні дані дають можливість вивчити зв’язок попарно між змінними та передбачити значення однієї змінної на основі значення інших.

Друга характеристика набору даних, *за типом вимірювання*, розрізняє: *категорії* та *числа*. Якщо змінна містить інформацію про те, до якої з декількох нечислових категорій належить об’єкт, то вона називається *якісною змінною* або *категорією*. Існують два типи якісних змінних (категорій): *порядкові* (ординальні) та *номінальні*.

Якщо якісні дані можна природним чином і змістовно впорядкувати, виділити перший, другий, третій, найкращий, найгірший показник, то мова йде про порядкові дані, для яких під час статистичного аналізу розраховують *медіану*. У якості прикладу можна назвати посади (президент, віце-президент, начальник відділу тощо), кваліфікації робітників (розряди), результати опитування та інше.

Якщо порядок у даних відсутній, то мова йде про *номінальну* якісну змінну. Наприклад, номінальними змінними є стать людини, віросповідання, назва відділу (або

його номер). При статистичному аналізі номінальних якісних змінних зазвичай визначають *моду*. Незважаючи на те, що часто значення якісної змінної можна записати за допомогою чисел, наприклад номер відділу або номер дому, така змінна все ж залишається якісною, оскільки ці числа не мають будь-якої змістовної інтерпретації.

Значення змінних, які реєструються як *числа*, що мають змістовний сенс, називають *кількісними даними*. Вони можуть бути *дискретними* або *неперервними* (безперервними). *Дискретна* – це така змінна, яка може набувати обмежену кількість значень з деякого діапазону. Прикладами такої змінної є кількість комп'ютерів на підприємстві, число клієнтів, які звернулися на підприємство за певний проміжок часу, число випадків відмов обладнання протягом доби і т. д. *Неперервною* вважають змінну, яка не є дискретною. Вона може приймати значення з деякого проміжку. Прикладами неперервної змінної можуть бути температура, щільність, прибуток підприємства і т. д. Дискретні змінні зазвичай описуються *цілими* числами, а неперервні *дійсними* числами (десятковим дробом).

Залежно від того, чи є набір даних часовим рядом чи ні, розрізняють: *часові ряди* (динамічні ряди) і *дані одного часового зрізу*. Якщо послідовність запису даних має змістовний сенс (хронологію), то відповідний набір даних є *часовим рядом* (рис. 1.1).

Рис. 1.1. Динаміка індексу цін виробників промислової продукції.

Якщо послідовність запису даних не важлива, то відповідний набір містить дані про *однин часовий зріз*. Це означають лише те, що немає ніякого впорядкування у часі, а є лише інформація по деяких об'єктах у певний момент часу (свого роду «моментальний знімок»). Наприклад, курс долара в деяких обмінних пунктах міста на сьогоднішню дату і т. д.

1.2.2. Джерела даних

Важливим питанням дослідження даних виступають *джерела їх походження*. Вибір цих джерел здійснюється підприємством залежно від їх надійності, доступності, вартості та важливості для економічної діяльності. Якщо підприємство здійснює цей збір самостійно (або доручає іншим особам за відповідним договором), то в результаті отримують *первинні дані*. Дані, що зібрані іншими підприємствами або людьми для власних цілей, називають *вторинними*.

Первинними даними для підприємства можуть слугувати:

- *безпосереднє спостереження*, яке здійснюється шляхом реєстрації об'єктів та їх ознак на основі підрахунку, обмірювання, зняття показників приладів, реєстрації цін і обсягу проданої продукції, інвентаризації залишків товарно-матеріальних цінностей на складах підприємства тощо;
- *документальний облік* – спосіб спостереження, при якому використовують різні документи первинного обліку підприємств. Наприклад, документальне дослідження

постачальника: якості його сировини, рівня браку, динаміки цін на продукцію. Таке дослідження зазвичай проводить відділ маркетингу або технічного забезпечення. Безпосереднє спостереження та документальний облік дають найбільш достовірні дані;

– *опитування*, в ході якого джерелом даних є відомості, які дають опитувані особи. Наприклад, дані опитування відділу маркетингу серед споживачів продукції підприємства з приводу розширення асортименту.

Основними *перевагами* первинних даних є їх унікальний, ексклюзивний характер, недоступність конкурентам та іншим стороннім особам, можливість отримання якісної інформації, що характеризує ставлення покупців до товарів ті їх якості, цін, самого підприємства, можливість виявити мотиви поведінки покупців тощо.

У той же час, збір первинних даних має низку суттєвих *недоліків*, як, наприклад, великі витрати фінансових і трудових ресурсів, відносно тривалий характер отримання та обробки інформації, можлива неточність результатів, що виникає у випадку порушення правил збору первинної інформації. Тому до збору первинних даних прибігають тоді, коли необхідно отримати відомості для вирішення конкретного завдання, впевнитися у повноті, актуальності та достовірності даних партнера, або отримати недоступну конкурентам інформацію з метою отримання конкурентних переваг.

Розглянемо основні *джерела вторинних даних*. Зазвичай вони включають як внутрішні, так і зовнішні публікації та документи, які дозволяють отримати різноманітну інформацію: про стан розвитку галузі, про динаміку цін, про рух робочої сили, напрямки науково-дослідної діяльності та нових технології, нові джерела енергії та загальний стан її споживання тощо.

Усі *зовнішні джерела вторинних даних* можна розділити на чотири основні групи (таблиця 1.2).

Таблиця 1.2.

Класифікація джерел даних

Тип джерела	Приклад	Переваги	Недоліки	Основне призначення
1	2	3	4	5
Офіційні видання та документи	- економічна та технічна література; - річні звіти органі-зацій і підприємств; - засоби масової інформації.	високий ступінь достовірності та доступності, невеликі витрати.	неповнота інформації, дані можуть бути використані конкурентами, старіння даних.	- аналіз ринку; - пошук даних для розуміння первинної інформації
Неофіційні джерела	- спілкування з постачальниками, клієнтами, тор-говим персоналом.	можливість отримання ексклюзивних даних.	неструктурованість даних; трудоміст-кість контактів; зайві дані.	збір даних про клієнтів, партнерів, конкурентів.
Специфічні джерела інформації	придбання товарів конкурентів або використання їхніх послуг; відвідуван-ня виробництва на підприємствах.	відносна доступність даних; конкретна спрямованість інформації.	неповнота даних; зайві дані; висока трудомісткість; високий рівень фінансових витрат.	збір даних про переваги конкурентних товарів, послуг, технологій.
Комерційні джерела даних спеціалізо-ваних фірм	- дані про ринки; - бази даних; - відомості про споживачів.	висока якість, регулярність оновлення, вартість менша первинних даних.	- дані можуть бути у конкурентів; неможливість впливати на склад дослідження.	відстеження зміни цінностей споживачів, сегментація ринку.

В якості *внутрішніх джерел вторинних даних* розглядають власні документи та звіти попередніх досліджень, наприклад, дані про запаси продукції, інформація про клієнтів та обсяги їх закупівель, сезонний та географічний розподіл продажів, дані про прибутки та збитки підприємства в цілому та його окремих підрозділів тощо. Ці дані є достовірними, важливими та ексклюзивними. Однак, попередні дані можуть не враховувати всі потреби маркетингових відділів підприємства; вони можуть бути дубльованими, а доступ до них може бути обмежений внаслідок внутрішніх причин. Основним призначенням внутрішніх джерел даних є оцінка прийнятих маркетингових рішень, рівня обслуговування, якості продукції, ретроспективний аналіз ринку тощо.

Усі джерела даних можна також розділити на два великих класи: *документальні* та *електронні*. Серед документальних даних відмітимо *наукові журнали*, які складають близько 70% усіх наукових документів, а близько 80% фахівців різних рівнів вважають науковий журнал основним джерелом науково-технічної інформації. Але тепер наукові журнали зазвичай доступні і в електронній формі.

Тому особливої уваги заслуговують електронні дані, що знаходяться в мережі Інтернет або у специфічних базах. На кінець 2020 р. в мережі знаходилося біля 1,3 млрд. сайтів, що майже в 2 рази більше, ніж 5 років тому. Протягом дня тільки пошукова система Google робить біля 6 млн. запитів, у рамках того ж дня офіційно продається більше 600 тис. комп'ютерів і 3,5 млн. смартфонів. Все це свідчить про зростаючу важливість електронних даних.

Стосовно банків та баз даних звернемо увагу, що тільки в Україні для підприємців працює більше 140 відкритих джерел цього типу. До них відносяться:

- єдиний державний реєстр юридичних осіб та фізичних осіб-підприємців (irc.gov.ua/ua/Poshuk-v-YeDR);
- пошук виданих ліцензій (irc.gov.ua/ua/Poshuk-v-YeLR);
- інформація з фондового ринку України (smida.gov.ua/db);
- перевірка дипломів (osvita.net/checkdoc);
- база даних компанії You Control (youcontrol.com.ua/).

2. Основи математичної статистики

Математична статистика є основним інструментом бізнес-статистики. *Математична статистика* - це розділ математики, що вивчає методи збору та обробки даних, одержаних в результаті спостережень над випадковими явищами, з метою формування обґрунтованих практичних висновків.

Основними здачами математичної статистики являються:

1. Оцінка параметрів розподілу імовірностей значень випадкової величини на основі отриманої вибірки даних;
2. Визначення довірчих інтервалів для параметрів розподілу імовірностей генеральної сукупності даних;
3. Перевірка гіпотез про вигляд розподілу імовірностей або величини його параметрів;
4. Виявлення закономірності або залежності між випадковими величинами на основі вибіркових даних.

Таким чином, апарат математичної статистики зв'язаний із випадковими явищами, але на відміну від теорії імовірностей методи математичної статистики дозволяють охарактеризувати випадкове явище по його обмеженій вибірці. Водночас методи математичної статистики широко застосовуються для опрацювання статистичних даних, що не мають ймовірної природи, тому вони широко застосовуються в багатьох галузях людської діяльності: політиці, економіці, фінансах, медицині, військовій справі та інше.

2.1. Основні поняття

Означення 2.1. Множина значень $x_1, x_2, \cdots, x_n$ випадкової величини X, що одержані в результате n спостережень називається *вибіркою обсягу* n. Множина всіх значень величини X називається *генеральною сукупністю*.

Для подальшого зафіксуємо довільну вибірку чисел із генеральної сукупності:

$$x_1, x_2, \cdots, x_n \ , \tag{2.1}$$

наприклад, 8,6,1, 7,3, 6,3,7,9, 8, 3,1. Для вирішення багатьох статистичних задач треба знати розподіл ймовірностей випадкової величини X. Для цього вибірку (2.1) треба впорядкувати $y_1 \le y_2 \le \cdots \le y_n$. В нашому випадку впорядкована вибірка має вигляд 1,1,3,3,3, 6,6,7,7, 8,8,9. *Різні значення* величини X, які зустрічаються у ряді (2.1) називаються *варіантами*, а сукупність варіантів розташованих у зростаючому порядку називається *варіаційним рядом*:

$$v_1 < v_2 < \cdots < v_k \ , \tag{2.2}$$

де $1 \le k \le n$. В нашому випадку варіаційний ряд буде мати вигляд 1,3,6,7,8,9 і $k = 6$.

Якщо варіанти (2.2) зустрічаються в ряду (2.1) відповідно $n_1, n_2, \cdots, n_k$ раз, то числа n_i називаються *частотами варіант*, а відношення $\omega_i = n_i/n, \ 1 \le i \le k$, - *відносними частотами*. В нашому випадку частотами будуть $n_1 = 2, n_2 = 3, \ n_3 = 2, \ n_4 = 2, \ n_5 = 2, \ n_6 = 1$ Частоти і відносні частоти задовольнлять умовам нормування:

$$\sum_{i=1}^{k} n_i = n \ , \qquad \sum_{i=1}^{k} \omega_i = 1 \ . \tag{2.3}$$

Таким чином, відносні частоти наближають *значення імовірностей варіант* (2.2). У нашому випадку відносні частоти будуть такими $\omega_1 = 2/12 = 1/6, \ \omega_2 = 3/12 = 1/4, \ \omega_3 = \omega_4 = \omega_5 = 1/6, \ \omega_6 = 1/12.$

Таблиця, в першому рядку якої розташований варіаційний ряд (2.2), а в другому відповідні частоти n_i (або відносні частоти ω_i) називається *дискретним статистичним розподілом*.

Таблиця 2.1.

Варіанти	v_1	v_2	$\cdots$	v_k
Частоти	ω_1	ω_2	$\cdots$	ω_k

У нашому випадку дискретний розподіл буде мати вигляд:

Варіанти	1	3	6	7	8	9
Частоти	1/6	1/4	1/6	1/6	1/6	1/12

Якщо кількість варіант у вибірці дуже велика або досліджуються значення неперервної випадкової величини X, то будується *інтервальний статистичний розподіл*. Відрізок значень $[x_{min}, x_{max}]$ ділитися на m рівних інтервалів $[a_i, a_{i+1})$, $1 \le i \le m$, де m дорівнює цілої частини числа $\sqrt{n}$ або $\log_2 n$. Нехай n_i^*, $1 \le i \le m$, суми частот тих варіант, які потрапили до інтервалу $[a_i, a_{i+1})$, тоді інтервальний розподіл записується у вигляді таблиці, де в першому рядку вказуються інтервали, а в другому - відносні частоти $\omega_i^* = n_i^*/n$.

Таблиця 2.2.

Інтервали	$[a_1, a_2)$,	$[a_2, a_3)$	⋯	$[a_m, a_{m+1}]$
Частоти	ω_1^*	ω_2^*	⋯	ω_m^*

Таким чином, інтервальний розподіл наближає *щільність розподілу* випадкової величини X. Розмір інтервалів (або крок) $h = a_{i+1} - a_i$, $1 \le i \le m$, обчислюється за формулою:

$$h = \frac{x_{max} - x_{min}}{m} \quad . \tag{2.4}$$

Частота найбільшого значення x_{max} відноситься до останнього інтервалу. Якщо значення варіанти потрапило в точності на межу двох інтервалів (малоймовірна подія), то його частота або ділиться порівну між цими інтервалами, або відноситься цілком до правого інтервалу. Інколи для зручності подальшої обробки результатів спостережень від інтервального розподілу переходять до дискретного розподілу, замінюючи кожен інтервал його серединою:

$$y_i = \frac{a_i + a_{i+1}}{2}, \qquad 1 \le i \le m. \tag{2.5}$$

2.2. Полігони та гістограми

Для створення наочного уявлення про статистичні розподіли застосовуються *полігони* і *гістограми*.

Означення 2.2. Полігоном частот (або відносних частот) називається ламана лінія, яка будується за таблицею 2.1 і відрізки якої сполучають на площині точки з координатами (v_i, n_i) (або (v_i, ω_i)), $1 \le i \le k$.

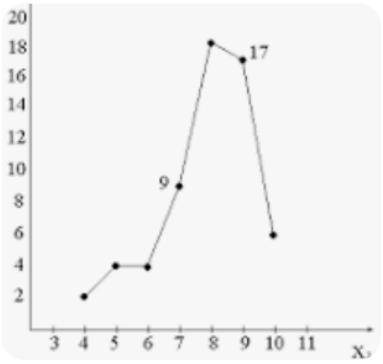


Рис. 2.1. Полігон частот.

Гістограма застосовується для графічного зображення інтервального статистичного розподілу (таблиця 2.2). Для побудови гістограми частот (або відносних частот) на осі абсцис відкладаються частинні інтервали довжини h і на цих відрізках будуються прямокутники висотою n_i^*/h (або $n_i^*/(hn)$), $1 \le i \le m$.

Оскільки висоти прямокутників є щільністю частот на відповідних частинних інтервалах, то при достатньо великих обсягах вибірки n (і малих h) гістограма може бути

достатньо близьким статистичним аналогом щільності імовірності випадкової величини X у генеральної сукупності (рис. 2.2).

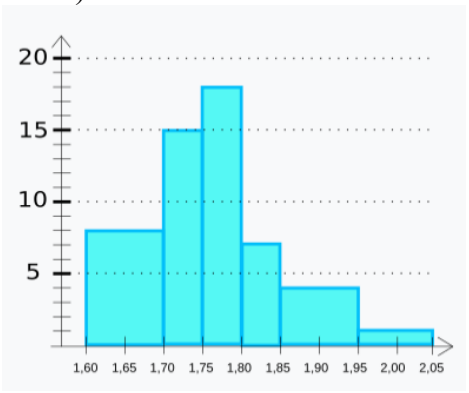


Рис. 2.2. Гістограма.

Полігон зазвичай використовується для графічного зображення дискретного статистичного розподілу, але його можна застосувати також разом з гістограмою для графічного зображення інтервального розподілу або при переході від інтервального розподілу до дискретного (формула 2.5). У цьому випадку за абсциси точок беруться центри частинних інтервалів (рис. 2.3).

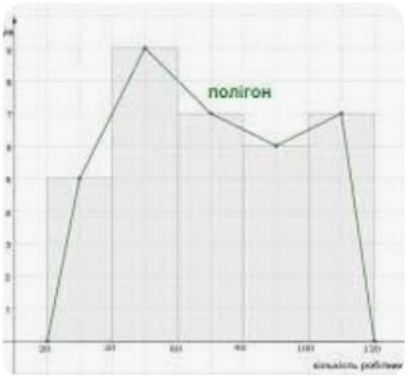


Рис. 2.3. Полігон і гістограма.

2.3. Емпірична функція розподілу

Нехай заданий статистичний розподіл частот величині X отриманий на основі n спостережень і x – довільне число. Позначимо як n_x – сумарну частоту всіх варіант в ряду (2.2), які менше за x. Тоді відносна частота події $X < x$ дорівнює n_x/n і вона є функцією від x. Ця функція знаходиться емпіричним шляхом (в результате спостережень) і тому називається емпіричною.

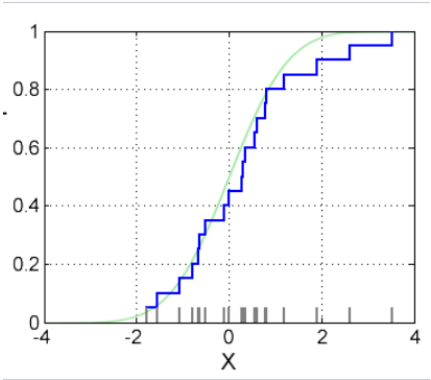
Означення 2.3. Інтегральна функція розподілу $F(x) = P(X < x)$ генеральної сукупності називається *теоретичною функцією розподілу*. *Емпірична функція розподілу* отримана на основі n спостережень визначається формулою:

$$F^*(x) = \frac{n_x}{n} \quad , \tag{2.6}$$

Таким чином, функція $F(x)$ визначає імовірність події $X < x$, а функція $F^*(x)$ – відносну частоту цієї події. При збільшенні числа випробувань емпірична функція буде наближати теоретичну (рис. 2.4). Звідси випливає практична корисність емпіричної функції. Можна сказати, що функція $F^*(x)$ є статистичним аналогом функції $F(x)$. Такий висновок підтверджується також тим, що властивості функції $F^*(x)$ аналогічні властивостям функції $F(x)$, а саме виконуються таки умови:

$0 \le F^*(x) \le 1$, $F^*(x) \le F^*(y)$ для $x < y$, $F^*(x_{min}) = 0$, $F^*(x) = 1$ для $x > x_{max}$.

Легко помітити, що ці властивості аналогічні властивостям функції розподілу дискретної випадкової величини.



2.4. Емпіричний розподіл (синя лінія), теоретичний розподіл (зелена лінія).

2.4. Числові характеристики вибірки

Одним з важливих засобів обробки даних є обчислення їх числових характеристик. Найбільш важливі з них: середнє значення, вибіркова дисперсія, вибіркове середньоквадратичне відхилення.

Вибіркове середнє. Нехай $x_1, x_2, \cdots, x_n$ – вибірка із генеральної сукупності і $v_1 < v_2 < \cdots < v_k$ її варіаційний ряд. Вибірковим середнім називають таку величину:

$$\overline{x_в} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{j=1}^{k} v_j n_j = \sum_{j=1}^{k} v_j \omega_j \ . \qquad (2.7)$$

Таким чином, вибіркове середнє є *статистичним аналогом математичного сподівання*.

Для інтервального розподілу, заданого таблицею 2.2, середнє значення обчислюється як середнє значення відповідного дискретного розподілу за формулою:

$$\overline{x_в} = \frac{y_1 n_1^* + \cdots + y_m n_m^*}{n} = \frac{1}{n}\sum_{i=1}^{m} y_i n_i^* = \sum_{i=1}^{m} y_i \omega_i^* \ , \qquad (2.8)$$

де y_i – середини відповідних інтервалів, а ω_i^* - відносні частоти.

Розмахом вибірки називається число:

$$R = x_{max} - x_{min}. \qquad (2.9)$$

Медіана. Крім середнього аріфметичного в статистичному аналізі застосовують порядкові (або структурні) середні медіану і моду. *Медіаною* Me упорядкованою вибірки $x_1 \le x_2 \le \cdots \le x_n$ називається її порядкове середнє, тобто значення її *серединного елемента*. Для вибірки непарної довжини $n = 2q - 1, \ q \ge 1,$ медіана дорівнює серединному елементу x_q, а для вибірки парної довжини $n = 2q, \ q \ge 1,$ – півсумі двох серединних елементів x_q і x_{q+1}:

$$Me = \begin{cases} x_q \ , & \text{при } n = 2q - 1 \\ \dfrac{x_q + x_{q+1}}{2}, & \text{при } n = 2q \end{cases} . \qquad (2.10)$$

Наприклад, для вибірки $1, 3, 3, 6, 7, 8, 9$ маємо $n = 7, \ q = 4, \ (7 = 2 \cdot 4 - 1)$, тому медіаною буде четверте число, тобто число 6. Для вибірки $1, 2, 3, 4, 5, 6, 8, 9$ маємо $n = 8, \ q = 4$, тому медіаною буде пів-сума двох чисел серединних чисел, тобто число $(4 + 5)/2 = 4{,}5$.

Медіана варіаційного ряду $v_1 < v_2 < \cdots < v_k$ визначається аналогічно, але замість довжини n у формулу (2.10) треба підставити довжину k. Зауважимо, що при цьому втрачається інформація о частотах варіант в ряду (2.1).

Нагадаємо, що в теорії імовірності медіана $Me(X)$ безперервної випадкової величини X з функцією розподілу $F(x)$ визначається умовою $F\big(Me(X)\big) = 1/2$. Тому для обчислення медіани інтервального статистичного розподілу (таблиця 2.2) спочатку визначається медіанний інтервал $[a_j, a_{j+1})$, де знаходиться медіана (серединний елемент) ряду $x_1 \le x_2 \le \cdots \le x_n$. Потім обчислюється сумарна частота $s_{j-1}^* = n_1^* + \cdots + n_{j-1}^*$ і тоді медіана обчислюється за формулою:

$$Me = a_j + h\frac{n - 2 \cdot s^*_{j-1}}{2 \cdot n^*_j}, \qquad (2.11)$$

де a_j – початок медіанного інтервалу, h - його довжина, n – об'єм виборки, n^*_j - частота медіанного інтервалу.

Приклад 2.1. Нехай штрафи за порушення правил дорожнього руху визначаються такою таблицею.

Таблиця 2.3.

Розмір штрафу грн.	Число штрафів	Сумарна частота
До 100 грн.	4	4
100 – 200	20	24
200 – 300	26	50
300 – 400	15	65
400 – 500	8	73
500 – 600	3	76
600 – 700	2	78
700 і більше	2	80

У даному випадку $n = 80$, тому за формулою (2.10) медіана ряду (2.1) є пів-сумою 40 і 41 членів, які знаходяться у третьому інтервалі (там знаходяться члени від 25 до 50). Таким чином, у даному випадку третій інтервал буде медіанним $j = 3$ і $a_3 = 200$. Тоді за формулою (2.11) маємо:

$$Me = a_3 + h\frac{n - 2 \cdot s^*_{j-1}}{2 \cdot n^*_j} = 200 + 100\frac{80 - 2 \cdot 24}{2 \cdot 26} = 261{,}5\ .$$

Таким чином, половина штрафів менше 261,5 грн., а половина – більше.

Мода. *Мода вибірки* (2.1) це значення, що трапляється найчастіше в сукупності спостережень. Наприклад, для вибірки $1, 3, 3, 6, 7, 8, 9$ мода дорівнює числу 3. Іноді трапляється більше, ніж одна мода (наприклад: 2, 6, 6, 6, 8, 9, 9, 9, 10). У такому випадку, можна сказати, що сукупність спостережень полі-модальна (або мультимодальна). Як правило це вказує на те, що набір даних не підпорядковується нормальному закону.

Мода, як середня величина, вживається також для даних, що мають *нечислову* природу. Серед перелічених кольорів автомобілів - «білий», «чорний», «синій металік», «білий», «синій металік», «білий» - мода дорівнюватиме значенню «білий». За експертної оцінки з її допомогою визначають найпопулярніші типи продукту, що враховується при прогнозі продажів чи плануванні їх виробництва.

Модою Mo вибірки для дискретного статистичного розподілу (таблиця 2.1) називається варіанта з найбільшою частотою (відносною частотою). Нагадаємо, що в теорії імовірності модою $Mo(X)$ безперервної випадкової величини X називається значення з найбільшою щільністю імовірності. Тому для обчислення моди інтервального статистичного розподілу (таблиця 2.2) спочатку визначається модальний інтервал $[a_j, a_{j+1})$, з найбільшою частотою n^*_j, а потім мода обчислюється за формулою:

$$Mo = a_j + h\frac{n^*_j - n^*_{j-1}}{(n^*_j - n^*_{j-1}) + (n^*_j - n^*_{j+1})}, \qquad (2.12)$$

де a_j – початок модального інтервалу, h - його довжина, n^*_j - частота модального інтервалу, n^*_{j-1} і n^*_{j+1} - частоти відповідно попереднього і наступного за модальним інтервалами.

Приклад 2.2. Нехай штрафи за порушення правил дорожнього руху визначаються таблицею 2.3. Модальним інтервалом тут також буде третій інтервал з максимальною частотою 26. Знайдемо моду інтервального розподілу за формулою (2.12):

$$Mo = a_3 + h\frac{n^*_3 - n^*_2}{(n^*_3 - n^*_2) + (n^*_3 - n^*_4)} = 200 + 100\frac{26 - 20}{(26 - 20) + (26 - 15)} = 235{,}3\ .$$

Таким чином, найчастіше розмір штрафу становить 235,3 грн.

Вибіркова дисперсія. Дисперсія вибірки $x_1, x_2, \cdots, x_n$ обчислюється як середнє значення квадратичних відхилень від середнього значення за формулою:

$$D_{\text{в}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x_{\text{в}}})^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \overline{x_{\text{в}}}^2 \ , \tag{2.13}$$

де $\overline{x_{\text{в}}}$ – вибіркове середнє. Якщо для вибірки складений варіаційний ряд (2.2) і дискретний розподіл (таблиця 2.1), то вибіркова дисперсія (2.13) буде мати вигляд:

$$D_{\text{в}} = \left(\sum_{j=1}^{k} v_j^2 \omega_j\right) - \overline{x_{\text{в}}}^2 \ , \tag{2.14}$$

де ω_j – відносні частоти розподілу.

Якщо дані представлені у вигляді інтервального ряду, то треба перейти до дискретного розподілу узяв за точки середини інтервалів за формулою (2.5), а потім скористатися формулою для дисперсії:

$$D_{\text{в}} = \left(\sum_{i=1}^{m} y_i^2 \omega_i^*\right) - \overline{x_{\text{в}}}^2 \ , \tag{2.15}$$

Оскільки вибірка даних була обрана випадковим чином, обидві величини $\overline{x_{\text{в}}}$ і $D_{\text{в}}$ є випадковими величинами. Їх математичні сподівання можна розрахувати, якщо усереднити послідовність всіх значень вибірки розміром n із генеральної сукупності. Таким чином можна показати, що має місце рівність:

$$E(D_{\text{в}}) = \frac{n-1}{n} D(X) \ , \tag{2.16}$$

де $D(X)$ – дисперсія генеральної сукупності. Якщо розглядати вибіркову дисперсію $D_{\text{в}}$ як оцінку дисперсії $D(X)$, то вона буде зміщена на коефіцієнт $(n-1)/n$. З цієї причини, дисперсію $D_{\text{в}}$ називається *зміщеною дисперсією вибірки*.

Із формули (2.15) випливає, що введенням поправки $n/(n-1)$, можна із дисперсії $D_{\text{в}}$ отримати *виправлену* (або незміщену*) дисперсію для вибірки* $S_{\text{в}}$, яка більш точно наближає $D(X)$:

$$S_{\text{в}} = \frac{n}{n-1} D_{\text{в}} = \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x_{\text{в}}})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x_{\text{в}}})^2. \tag{2.17}$$

Застосування дільника $n-1$ замість n називають *поправкою Бесселя*. Обидві вибіркові дисперсії $D_{\text{в}}$ і $S_{\text{в}}$ застосовують на практиці і при великих n вони майже рівні. При $n > 30$ різниця між коефіцієнтами $1/(n-1)$ і $1/n$ несуттєва, тому використовують дисперсію $D_{\text{в}}$, оскільки її простіше обчислювати. При $n \leq 30$ більш доцільно використовувати виправлену вибіркову дисперсію $S_{\text{в}}$.

Корінь квадратний з вибіркової дисперсії $D_{\text{в}}$ називається *вибірковим середнім квадратичним відхиленням* $\sigma_{\text{в}} = \sqrt{D_{\text{в}}}$. Корінь квадратний із виправленої вибіркової дисперсії $S_{\text{в}}$ називається *виправленим середньо-квадратичним відхиленням* $s_{\text{в}} = \sqrt{S_{\text{в}}}$.

Зауважимо, що в Excel функція ДИСП.В обчислює виправлену вибіркову дисперсію, а функція ДИСП.Г – зміщену вибіркову дисперсію (в довіднику вона називається дисперсією генеральної сукупності). Аналогічно, функція СТАНДОТКЛОН.В обчислює виправлене середньо-квадратичним відхилення, а функція СТАНДОТКЛОН.Г вибіркове середньо-квадратичним відхилення (зміщене відхилення).

3. Довірчі інтервали

3.1. Квантилі

В теорії імовірності медіаною безперервної випадкової величини X з функцією розподілу $F(x)$ називається розвязок рівняння $F(x) = 1/2$ і тому її позначають також як $x_{1/2}$. Поняття медіани можна узагальнити таким чином.

Означення 3.1. Нехай $0 < \alpha < 1$, тоді (теоретичною) *α-квантиллю* x_α називається розв'язок рівняння $F(x) = \alpha$, тобто $F(x_\alpha) = \alpha$.

Число x_α називають також квантилем рівня α. Якщо функція розподілу $F(x)$ неперервна і строго монотонна, то вона має обернену функцію $F^{-1}(x)$ і тому квантиль визначається однозначно для будь якого $\alpha \in (0,1)$ за формулою $x_\alpha = F^{-1}(\alpha)$ (рис. 3.1).

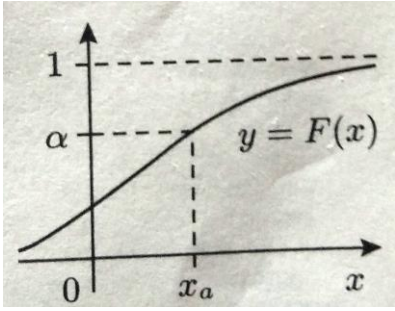


Рис. 3.1. Квантиль рівня α.

Деякі квантилі отримали спеціальні назви. Квантиль $x_{1/4} = x_{0.25}$ називається *першою* (чи нижньою) *квартиллю* (від латинського quarta). Квантиль рівня 0,5 збігається з медіаною $x_{1/2} = x_{0.5}$ і другою квартиллю. Квантиль $x_{3/4} = x_{0.75}$ називається *третьою* (або верхньою) квартиллю. Інтер-квартильним розмахом називається різниця між третьою та першою квартиллю $x_{0.75} - x_{0.25}$. Це число характеризує розкид розподілу випадкової величини та є аналогом дисперсії. Наприклад, для стандартного нормального розподілу $N(0,1)$ маємо $x_{0.25} = -0{,}674$, $x_{0.5} = 0$, $x_{0.75} = 0{,}674$, $x_{0.75} - x_{0.25} = 1{,}348$.

3.2. Розподіли від нормально розподілених величин

Означення 3.2. Нехай випадкові величини $X_1, \cdots, X_n$ незалежні і розподілені за стандартним нормальним законом $N(0,1)$, тоді розподіл випадкової величини:

$$R_n^2 = X_1^2 + \cdots + X_n^2 , \qquad (3.1)$$

називають розподілом *хі-квадрат* (або розподілом *Пірсона*) з n ступенями свободи.

Коротко випадкову величину R_n^2, яка розподілена за законом хі-квадрат з n ступенями свободи позначають таким чином $R_n^2 \sim \chi_n^2$. Розподіл хі-квадрат є одним з найважливіших у статистиці, зокрема він використовується у критеріях хі-квадрат, наприклад у критерії Пірсона. Розподіл імовірностей функції хі-квадрат виражається через гамма-функцію (рис. 3.2). Її математичне очікування дорівнює $E(R_n^2) = n$, а дисперсія дорівнює $D(R_n^2) = 2n$.

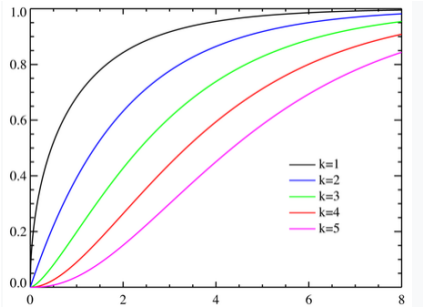


Рис. 3.2. Функція розподілу імовірностей хі-квадрат (тут n позначено як k).

Означення 3.3. Нехай випадкові величини X і R_n^2 незалежні і розподілені згідно законам $N(0,1)$ і χ_n^2 відповідно, тоді розподіл випадкової величини:

$$T_n = \frac{X}{\sqrt{R_n^2/n}}\ , \tag{3.2}$$

називають t_n-*розподілом* (або розподілом *Стьюдента*) з n ступенями свободи (вільності).

Коротко випадкову величину T_n, яка розподілена за t_n-розподілом з n ступенями свободи позначають таким чином $T_n \sim t_n$. Розподіл Стьюдента також грає важливу роль у статистиці, оскільки на ньому основане обчислення довірчих інтервалів для випадкових величин. Математичне очікування випадкової величини T_n дорівнює нулю $E(T_n) = 0$ в силу парності щільності розподілу, а дисперсія дорівнює $D(T_n) = n/(n-2)$ при $n > 2$ (нескінченна при $n = 2$). На рисунку 3.2 показано щільність t-розподілу при різних n (тут n позначено як k).

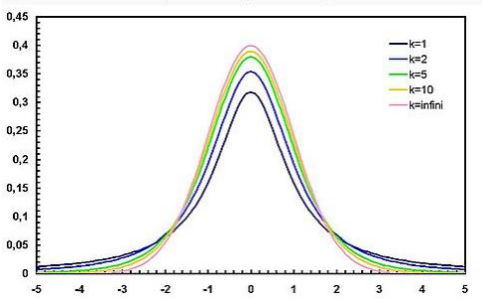


Рис. 3.2. Щільність розподілу Стьюдента.

3.3. Інтервальні оцінки

Точкові оцінки параметрів розподілу не дають можливості зробити висновків про їхню точність та надійність, оскільки вони є випадковими величинами. Щоб мати уявлення про точність оцінки, у математичної статистиці використовують довірчі інтервали. Іншими словами, замість того, щоб наближати невідомий параметр θ «точковою» оцінкою $\hat{\theta}$ можна локалізувати його за допомогою випадкового інтервалу $(\hat{\theta}_1, \hat{\theta}_2)$, якій покриває θ з імовірністю близькою до одиниці.

Означення 3.4. Нехай $0 < \alpha < 1$, тоді *довірчим інтервалом* (або *інтервальною оцінкою*) невідомого параметра θ з *коефіцієнтом довіри* $1 - \alpha$ називається випадковий інтервал $(\hat{\theta}_1, \hat{\theta}_2)$, який покриває значення θ із заданою імовірністю $1 - \alpha$.

Статистики $\hat{\theta}_1$ і $\hat{\theta}_2$ є функціями вибіркового вектору $\xi = (x_1, x_2, \cdots, x_n)$, всі компоненти якого розподілені за законом $F_\theta(x)$, тому із означення 3.4 випливає, що для довірчого інтервалу повинна виконуватися така умова:

$$P\left(\hat{\theta}_1(\xi) < \theta < \hat{\theta}_2(\xi)\right) \geq 1 - \alpha\ . \tag{3.3}$$

Статистики $\hat{\theta}_1$ і $\hat{\theta}_2$ називаються відповідно нижньою та верхньою межами довірчого інтервалу, а імовірність $1 - \alpha$ називається також довірчою імовірністю.

На практиці зазвичай вважають, що $\alpha = 0{,}05$ або ще менше, і говорять о довірчому інтервалі з коефіцієнтом довіри 95%. Як правило, довжина довірчого інтервалу збільшується при збільшенні коефіцієнта довіри $1 - \alpha$ і збігається до нуля при $n \to \infty$.

3.3.1. Інтервали для нормальних випадкових величин

Приклад 3.1. Припустимо, що випадкова величина X розподілена за нормальним законом $N(\theta, \sigma^2)$, причому параметр θ невідомий, а параметр σ відомий. Знайдемо інтервальну оцінку параметру зсуву θ на основі вибірки $\xi = (x_1, x_2, \cdots, x_n)$, де всі елементи x_i незалежні і розподілені за законом $N(\theta, \sigma^2)$. Цю модель застосовують до даних, отриманих при незалежних вимірах величини θ за допомогою приладу, що має відому середню похибку σ.

Нехай $\Phi(x)$ – інтеграл Лапласа, тобто функція розподілу для закону $N(0{,}1)$, який розглядався на другому практикумі (формула (2.12)). Для $0 < \alpha < 1$ позначимо як x_α α-квантиль стандартного нормального закону $N(0{,}1)$, тобто число, для якого виконується рівність $\Phi(x_\alpha) = \alpha$. В подальшому α буде розглядатися як маленьке число $\alpha \approx 0$, яке

дорівнює імовірності похибки. Тому приведемо деякі значення квантиль стандартного розподілу для малих α, які можна знайти у таблицях посібників з теорії імовірності.

Таблиця 3.1.

α	0,05	10^{-2}	10^{-3}
$x_{1-\alpha/2}$	$x_{0,975} = 1,96$	$x_{0,995} = 2,58$	$x_{0,9995} = 3,29$

Відомо, що вибіркове середнє $\bar{x}_{\text{в}}$ буде незміщеною оцінкою математичного сподівання (практикум 4), тому його можна взяти за основу довірчого інтервалу. Крім того відомо, що вибіркове середнє $\bar{x}_{\text{в}}$ для нормального розподілу $N(\theta, \sigma^2)$, буде також нормально розподіленою випадковою величиною за законом $N(\theta, \sigma^2/n)$, тобто її дисперсія буде в n разів менше. Тоді випадкова величина $y = \sqrt{n}(\bar{x}_{\text{в}} - \theta)/\sigma$ буде розподілена за стандартним нормальним законом $y \sim N(0,1)$ (практикум 2 формула (2.13)). Щоб довірчий інтервал мав імовірність $1 - \alpha$ треба щоб виконувались нерівності:

$$x_{\alpha/2} < \sqrt{n}(\bar{x}_{\text{в}} - \theta)/\sigma < x_{1-\alpha/2} \, . \tag{3.4}$$

Дійсно в цьому випадку маємо:

$$P(x_{\alpha/2} < y < x_{1-\alpha/2}) = \Phi\left(x_{1-\alpha/2}\right) - \Phi\left(x_{\alpha/2}\right) = 1 - \alpha/2 - \alpha/2 = 1 - \alpha \, .$$

Із верхньої нерівності (3.4) отримуємо нижню межу, а із нижньої нерівності верхню межу довірчого інтервалу з коефіцієнтом довіри $1 - \alpha$ для математичного очікування нормального закону, отриманих на основі n-вибірки:

$$\bar{x}_{\text{в}} - x_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \theta < \bar{x}_{\text{в}} - x_{\alpha/2}\frac{\sigma}{\sqrt{n}} \, . \tag{3.5}$$

В силу парності стандартного закону $N(0,1)$ та із рівності $\Phi(-x) = 1 - \Phi(x)$ (практикум 2 формула (2.14)) отримуємо $\Phi\left(-x_{\alpha/2}\right) = 1 - \Phi\left(x_{\alpha/2}\right) = 1 - \alpha/2 = \Phi\left(x_{1-\alpha/2}\right)$. Тому $-x_{\alpha/2} = x_{1-\alpha/2}$ (в силу неперервності і строгої монотонності $\Phi(x)$) і тоді інтервал (3.5) можна записати в більш симетричної формі відносно вибіркового середнього:

$$\bar{x}_{\text{в}} - x_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \theta < \bar{x}_{\text{в}} + x_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \, . \tag{3.6}$$

Таким чином, із таблиці 3.1 видно, що з імовірністю 95% значення параметра θ знаходиться в інтервалі:

$$\bar{x}_{\text{в}} - 1,96\frac{\sigma}{\sqrt{n}} < \theta < \bar{x}_{\text{в}} + 1,96\frac{\sigma}{\sqrt{n}} \, . \tag{3.7}$$

Якщо округлити 1,96 до 2, то ми отримуємо інтервал $\bar{x}_{\text{в}} \pm 2\sigma/\sqrt{n}$, якій називається *правилом двох сигм* (для простоти запам'ятовування).

Щоб простіше запам'ятати формулу (3.6) можна також покласти $\beta = 1 - \alpha/2$ і записати її у вигляді:

$$\bar{x}_{\text{в}} - x_{\beta}\frac{\sigma}{\sqrt{n}} < \theta < \bar{x}_{\text{в}} + x_{\beta}\frac{\sigma}{\sqrt{n}} \, , \tag{3.8}$$

але треба пам'ятати, що $\beta \approx 1$ точніше $(1/2 < \beta < 1)$ та коефіцієнт довіри цього інтервалу дорівнює $\beta - \alpha/2 = 2\beta - 1$. Наприклад, при $\beta = 0,975$ коефіцієнт довіри дорівнює 0,95. □

Якщо значення середньо-квадратичне відхилення σ невідомо, то його в формулі (3.8) заміняють на спроможну оцінку $\hat{\sigma} = \sigma_{\text{в}} = \sqrt{D_{\text{в}}}$, де $\sigma_{\text{в}}$ - вибіркове середнє квадратичне відхилення (лекція 2). Таким чином із формули (3.7) отримують *наближений* 95%-й інтервал $\bar{x}_{\text{в}} \pm 2\sigma_{\text{в}}/\sqrt{n}$. Однак виявилося, що цей випадок вимагає окремого теоретичного аналізу.

Приклад 3.2. Нехай $X \sim N(\theta_1, \theta_2^2)$, де обидва параметри θ_1 і θ_2 нормального розподілу невідомі. Знайдемо інтервальну оцінку для параметру зсуву θ_1 на основі вибірки $\xi = (x_1, x_2, \cdots, x_n)$, де всі елементи вибірки розподілені за законом $N(\theta_1, \theta_2^2)$. Для цього нам знадобиться наступна теорема, яку ми приведемо без доказу.

Теорема 3.3. Нехай $s_{\text{в}}$ виправлене середнє квадратичне відхилення вибірки $\xi = (x_1, x_2, \cdots, x_n)$, де $x_i \sim N(\theta_1, \theta_2^2)$, $1 \le i \le n$, тоді випадкова величина $z = \sqrt{n}(\bar{x}_{\text{в}} - \theta_1)/s_{\text{в}}$ розподілена за законом Стьюдента t_{n-1}.

Зауважимо, що випадкова величина $\sqrt{n}(\bar{x}_в - \theta_1)/s_в$ відрізняється від величини $\sqrt{n}(\bar{x}_в - \theta_1)/\sigma$, яка нам зустрілась у прикладі 3.1, тим, що відхилення σ замінено в неї на $s_в$. Першим теорему 3.3 відкрив у 1908 році ірландський статистик Уїльям Госсет. Він помітив, що при заміні σ на $s_в$ стандартний нормальний розподіл $N(0,1)$ змінюється на t_{n-1}. Це було наукове відкриття, тому що розподіл Стьюдента був невідомий у той час. Але за умовами контракту Госсет не міг публікувати статті за своїм прізвищем (статистика може бути комерційною таємницею). Тоді Госсет опублікував статтю під псевдонімом «Student». Так з'явився розподіл Стьюдента, хоча ніякого Стьюдента в природі не існувало.

Далі за допомогою теореми 3.3 діємо також як в прикладі 3.1, але замість стандартного нормального розподілу будемо використовувати розподіл Стьюдента t_{n-1}. Таким чином, отримуємо довірчий інтервал з коефіцієнтом довіри $1-\alpha$ для математичного очікування нормального закону, отриманого на основі n-вибірки при невідомому середнє квадратичному відхиленні:

$$\bar{x}_в - y_{1-\alpha/2}\frac{s_в}{\sqrt{n}} < \theta_1 < \bar{x}_в + y_{1-\alpha/2}\frac{s_в}{\sqrt{n}}\ , \tag{3.9}$$

де $y_{1-\alpha/2}$ – квантиль розподілу Стьюдента t_{n-1}, а $s_в = \sqrt{S_в}$ – виправлено вибіркове середнє квадратичне відхилення.

Порівняємо квантилі $x_{1-\alpha/2}$ закону $N(0,1)$ з відповідними квантилями $y_{1-\alpha/2}$ розподілу Стьюдента t_{n-1} при $n = 10$.

Таблиця 3.2.

α	0,05	10^{-2}	10^{-3}
$x_{1-\alpha/2}$	$x_{0,975} = 1,96$	$x_{0,995} = 2,58$	$x_{0,9995} = 3,29$
$y_{1-\alpha/2}$	$y_{0,975} = 2,26$	$y_{0,995} = 3,25$	$x_{0,9995} = 4,78$

Із таблиці 3.2 видно, що для вибірки розміру 10 довжина довірчого інтервалу для θ_1 з коефіцієнтом довіри 95% при заміни невідомого відхилення σ на його оцінку $s_в \approx \sigma$ збільшується приблизно в 2,26/1,96=1,15 раз.

Виправлене вибіркове середнє квадратичне відхилення $s_в = \sqrt{S_в}$ можна замінити у формулі (3.9) на вибіркове середнє квадратичне відхилення $\sigma_в = \sqrt{D_в}$, тоді в силу рівності $s_в/\sqrt{n} = \sigma_в/\sqrt{n-1}$ (лекція 2) отримуємо такий інтервал:

$$\bar{x}_в - y_{1-\alpha/2}\frac{\sigma_в}{\sqrt{n-1}} < \theta_1 < \bar{x}_в + y_{1-\alpha/2}\frac{\sigma_в}{\sqrt{n-1}}\ . \tag{3.10}$$

Тут також можна перейти до $\beta = 1-\alpha/2$ і отримати довірчий інтервал з коефіцієнтом довіри $2\beta - 1$, де $\beta \approx 1$ $(1/2 < \beta < 1)$:

$$\bar{x}_в - y_\beta\frac{\sigma_в}{\sqrt{n-1}} < \theta_1 < \bar{x}_в + y_\beta\frac{\sigma_в}{\sqrt{n-1}}\ . \tag{3.11}$$

Приклад 3.3. Припустимо, що випадкова величина X розподілена за нормальним законом $N(\mu, \theta^2)$, причому параметр зсуву μ відомий, а параметр масштабу θ невідомий. Знайдемо інтервальну оцінку параметра θ на основі вибірки $\xi = (x_1, x_2, \cdots, x_n)$, де всі елементи x_i незалежні і розподілені за законом $N(\mu, \sigma^2)$. Цю модель можна застосовувати для визначення середньо точності приладу шляхом багатократного вимірювання еталону.

Розглянемо випадкову величину D_n, яка дорівнює n-кратної вибіркової дисперсії:

$$D_n = nD_в = \sum_{i=1}^{n}(x_i - \bar{x_в})^2\ .$$

Тоді випадкова величина $R_n = D_n/\theta^2$ буде розподілена за законом хі-квадрат $R_n \sim \chi_n^2$:

$$R_n = \frac{D_n}{\theta^2} = \sum_{i=1}^{n}\left(\frac{x_i - \bar{x_в}}{\theta}\right)^2\ , \tag{3.12}$$

оскільки кожний доданок $(x_i - \bar{x_в})/\theta$, $1 \le i \le n$, розподілений за стандартним нормальним законом $N(0,1)$ (практикум 2 формула (2.13)).

Для $0 < p < 1$ позначимо p-квантиль розподілу χ_n^2 з n ступенями свободи як z_p і нехай α, як і раніше, буде імовірністю похибки. Тоді, щоб довірчий інтервал мав імовірність $1 - \alpha$ треба щоб виконувались такі умови:

$$1 - \alpha = P(z_{\alpha/2} < R_n < z_{1-\alpha/2}) = P(z_{\alpha/2} < D_n/\theta^2 < z_{1-\alpha/2}) . \quad (3.13)$$

Із нижньої нерівності $z_{\alpha/2} < D_n/\theta^2$ отримуємо верхню межу довірчого інтервалу $\hat{\theta}_2 = \sqrt{D_n/z_{\alpha/2}}$, а із верхньої нерівності $D_n/\theta^2 < z_{1-\alpha/2}$ отримуємо нижню межу довірчого інтервалу $\hat{\theta}_1 = \sqrt{D_n/z_{1-\alpha/2}}$. Таким чином, дістаємо такий довірчий інтервал:

$$\sqrt{\frac{D_n}{z_{1-\alpha/2}}} < \theta < \sqrt{\frac{D_n}{z_{\alpha/2}}} . \quad (3.14)$$

В загальному випадку, коли обидва параметри θ_1 і θ_2 нормального розподілу невідомі $X \sim N(\theta_1, \theta_2^2)$ нічого по суті не змінюється, оскільки було доведено, що випадкова величина $R_n = D_n/\theta_2^2 = n\sigma_в^2/\theta_2^2$ і в цьому випадку розподілена за законом хі-квадрат. Тому для параметру масштабу θ_2 ми отримуємо також довірчий інтервал (3.14), але записаний через вибіркове середньоквадратичне відхилення $\sigma_в$:

$$\frac{\sqrt{n}\,\sigma_в}{\sqrt{z_{1-\alpha/2}}} < \theta_2 < \frac{\sqrt{n}\,\sigma_в}{\sqrt{z_{\alpha/2}}} . \quad (3.15)$$

4. Статистичні гіпотези

4.1. Статистичні критерії

З теорією статистичного оцінювання тісно пов'язана перевірка статистичних гіпотез. Вона використовується щоразу, коли необхідний обґрунтований висновок про переваги того чи іншого способу про рівень прибутковості цінних паперів, про значущість математичної моделі та т. д.

Означення 4.1. *Статистичною гіпотезою* називається будь-яке припущення відносно виду або параметрів невідомого закону розподілу випадкової величини X.

Статистичну гіпотезу формулюють як припущення, з певним рівнем значущості, про властивості генеральної сукупності на основі оцінок вибіркової сукупності. Наприклад, статистичними гіпотезами являються:

a) Генеральна сукупність X розподілена за нормальним законом;
b) Дисперсії двох нормальних сукупностей X і Y рівні між собою.
c) Середнє значення сукупності X дорівнює a;
d) Середнє значення сукупності X менше a;

Статистичну гіпотезу прийнято позначати латинською літерою H (від лат. Hypothesis). Наприклад, гіпотези a) – d) в скороченому вигляді можна записати так:

$$H: X \sim N(\theta_1, \theta_2^2),\ H: D(X) = D(Y),\ H: E(X) = a,\ H: E(X) < a\ .$$

Гіпотезу, яку перевіряють, називають *основною* (або *нульовою*) гіпотезою і позначають як H_0. Поряд з нульовою гіпотезою розглядають також *альтернативну* (або конкуруючу) гіпотезу H_1, яка зазвичай є логічним запереченням нульової. Так альтернативною гіпотезою для гіпотези c) може бути гіпотеза $H_1: E(X) \neq a$.

Статистичні гіпотези перевіряються на основі статистичної вибірки даних вибірки $\xi_n = (x_1, x_2, \cdots, x_n)$. Для цього вибирають випадкову величину (тестову статистику) $T(\xi_n)$, розподіл якої відомий заздалегідь. Гіпотеза відхиляється тоді, коли отримується значення критерію $T(\xi_n)$, яке при істинності гіпотези *є малоймовірним*. Правило, за яким гіпотеза приймається або відхиляється називається с*татистичним критерієм* (або *тестом*).

Отже в загальному випадку задається мале число α (*рівень значимості*), яке дорівнює імовірності, з якою ми можемо відхилити вірну гіпотезу H (зазвичай $\alpha = 0{,}05$). Потім за вибірковим розподілом статистики $T(\xi_n)$ визначається найменше (*критичне*) значення тесту $\theta_{\text{кр}}$, яке задовольняє умові:

$$P(T(\xi_n) \geq \theta_{\text{кр}}) \leq \alpha\ . \qquad (4.1)$$

Якщо функція розподілу статистики $T(\xi_n)$ неперервна, то $\theta_{\text{кр}}$ дорівнює $(1-\alpha)$-квантилі $x_{1-\alpha}$ вибіркового розподілу. В цьому випадку статистичний критерій можна сформулювати таким чином: *гіпотеза H відхиляється, якщо* $T(\xi_n) \geq \theta_{\text{кр}}$ (відбулася малоймовірна подія) *і приймається в іншому випадку, коли* $T(\xi_n) < \theta_{\text{кр}}$.

Таким чином, множина можливих значень статистики $T(\xi_n)$ розбивається на дві підмножини, що не перетинаються: *критичну область* (область відхилення гіпотези) і область *допустимих значень* (область прийняття гіпотези). Розрізняють правобічну, лівобічну, і двобічну критичну область.

Правобічною називають критичну область, яка визначається умовою $T(\xi_n) \geq \theta_{\text{кр}}$, де $\theta_{\text{кр}} > 0$. Лівобічною називають критичну область, яка визначається умовою $T(\xi_n) \leq \theta_{\text{кр}}$, де $\theta_{\text{кр}} < 0$. Двобічною називають критичну область, яка визначається умовою $T(\xi_n) \leq \theta_1$ і $T(\xi_n) \geq \theta_2$, де $\theta_1 < \theta_2$. Зокрема, якщо $\theta_2 = -\theta_1 = \theta_{\text{кр}}$, то критична область визначається умовою $|T(\xi_n)| \geq \theta_{\text{кр}}$.

Приклад 4.1. Припустимо, хтось підкинув 10 разів монету і у 8 випадках вона впала гербом вгору. Чи можна вважати монету симетричною, тобто чи вірна гіпотеза $H: \theta = 1/2$, де θ – ймовірність випадання герба на цієї монеті.

Розв'язування. Будемо використовувати схему Бернуллі, тобто вважатимемо реалізацією експерименту вибірку $\xi_{10} = (x_1, x_2, \cdots, x_{10})$, де $x_i = 1$ (випадає герб) з

ймовірністю θ і $x_i = 0$ (випадає решка) з ймовірністю $1 - \theta$. Як статистику критерія тут можна взяти суму $T(\xi_{10}) = x_1 + x_2 + \cdots + x_{10}$ (кількість гербів). Тоді гіпотезі $H: \theta = 1/2$ суперечать значення критерія, які близькі до нуля і n.

Перевіримо для нашої вибірки гіпотезу $H: \theta = 1/2$ на рівні значимості $\alpha = 0{,}05$. Відомо, що схемі Бернуллі сума $x_1 + x_2 + \cdots + x_{10}$ має біноміальний розподіл:

$$P(T(\xi) \geq k) = \sum_{i=k}^{n} C_n^i \cdot \theta^i (1-\theta)^{n-i}\,. \qquad (4.2)$$

Для $\theta = 1/2$ права частина формули (4.2) при $k = 8$ рівна $(45 + 10 + 1)/1024 \approx 0{,}055$ і при $k = 9$ рівна $(10 + 1)/1024 \approx 0{,}011$. Тому для $\alpha = 0{,}05$ найменшим значенням $x_{0{,}95}$, яка задовольняє умові (4.1) буде число $x_{0{,}95} = 9$ (оскільки випадкова величина дискретна). Оскільки $T(\xi) = 8$ і $8 < 9$, то на заданому рівні значимості гіпотеза $H: \theta = 1/2$ *приймається*. З іншої сторони, оскільки $P(T(\xi) \geq 8) \approx 0{,}055$, що всього на $0{,}005$ більше ніж $\alpha = 0{,}05$, то вже при $\alpha = 0{,}06$ гіпотезу H треба *відхилити* при $T(\xi) = 8$. □

Цей приклад показує, що тут не можна впевнено прийняти або відхилити гіпотезу. Варто було б ще кілька разів підкинути монету, щоб дійти більш виваженого рішення. Нехай в результаті експерименту отримана вибірка даних $\xi_n = (x_1, x_2, \cdots, x_n)$, на якому критерій T приймає значення $T(\xi_n) = t_n$.

Означення 4.2. Імовірність $P(T(\xi) \geq t_n)$, яка обчислюється за усіма вибірками ξ довжини n називається *фактичним рівнем значимості*.

Фактичний рівень значимості це *найменший рівень значимості*, при якому гіпотеза H приймається при заданому виходу ξ_n. Приклад 4.1 показує, що обчислення фактичного рівня значимості нерідко дозволяє уникати категоричних (і при цьому помилкових) висновків, зроблених лише на основі порівняння значення t_n з критичним значеннями $x_{1-\alpha}$, знайденим для формально заданого числа α.

Під час перевірки гіпотези згідно одного з критеріїв можливі такі випадки (табл. 4.1).

Таблиця 4.1.

Гіпотеза *H*	**Приймається**	**Відхиляється**
Вірна	правильне рішення	помилка I роду
Невірна	помилка II роду	правильне рішення

Означення 4.3. Імовірність α, з якою можна допустити помилку I роду називається *рівнем значимості критерія*.

Імовірність, з якою можна допустити помилку II роду, тобто прийняти гіпотезу H, коли вона невірна, зазвичай позначається як β.

Означення 4.4. Імовірність $1 - \beta$, з якою відхилюється гіпотеза H, якщо вона невірна, називається *потужністю критерія*.

При перевірки статистичних гіпотез слід мати на увазі такий тезис, який стосується будь-яких експериментальних обґрунтувань:

- *Перевірка статистичної гіпотези не дає логічного доказу її вірності або невірності. Прийняття гіпотези слід розцінювати як досить правдоподібне твердження, яке не суперечить експерименту.*

Статистичний критерій називається *параметричним*, якщо він стосується окремих параметрів *відомого розподілу імовірностей* генеральної сукупності, на основі якого обчислюється розподіл самого критерію. Якщо розподіл генеральної сукупності невідомий, то такі критерії називаються *непараметричними*.

За своїм прикладним змістом статистичні гіпотези можна поділити на кілька основних типів:

- про значення параметрів;
- про закони розподілу;
- про рівність числових параметрів двох генеральних сукупностей.

4.2. Довірчі інтервали і прості гіпотези

Довірчі інтервали дозволяють перевірити деякі прості гіпотези, які стосуються числових значень параметрів розподілу. Одною із найпростіших статистичних гіпотез є гіпотеза $H: E(X) = \mu$ о рівності середнього значення нормальної генеральної сукупності X числу μ. Для розв'язання цієї задачі достатньо за вибіркою даних $\xi_n = (x_1, x_2, \cdots, x_n)$ і заданим рівнем значимості α обчислити довірчий інтервал:

$$I_n = \left(\bar{x}_{\text{в}} - y_{1-\alpha/2} \frac{\sigma_{\text{в}}}{\sqrt{n-1}} , \bar{x}_{\text{в}} + y_{1-\alpha/2} \frac{\sigma_{\text{в}}}{\sqrt{n-1}} \right).$$

де $\bar{x}_{\text{в}}$- вибіркове середнє, $y_{1-\alpha/2}$ – квантиль розподілу Стьюдента t_{n-1} з $n-1$ степеню свободи, $\sigma_{\text{в}}$ – вибіркове середнє квадратичне відхилення. Якщо $\mu \in I_n$, то гіпотеза H приймається, в іншому випадку відхиляється. Зауважимо, що тут критична область двобічна, оскільки альтернативна гіпотеза $E(X) \neq \mu$ може виконуватися як при малих, так і при великих значеннях μ.

4.3. Критерії згоди

Більш складні гіпотези стосуються невідомих законів розподілу випадкових величин. Розглянемо на приклад гіпотезу $H_0: X \sim N(\theta_1, \theta_2^2)$, де нам треба визначити нормальність розподілу та оцінити його параметри θ_1 і θ_2. Перевірка гіпотези про передбачуваний закон невідомого розподілу провадиться так само, як і перевірка гіпотези про параметри розподілу. т. е. з допомогою спеціально підібраної випадкової величини — критерію згоди.

Критерієм згоди (непараметричним критерієм) називають критерій перевірки гіпотези про передбачуваний закон невідомого розподілу. Є кілька критеріїв згоди: хі-квадрат Пірсона, критерій Колмогорова та інші. Обмежимося описом застосування критерію Пірсона до перевірки гіпотези про нормальний розподіл генеральної сукупності, хоча цей критерій *універсальний* і його аналогічним чином можна застосовувати і до інших розподілів (у цьому полягає його перевага).

4.3.1. Критерій Пірсона

В якості мери відхилення від нормального закону Карл Пірсон запропонував порівняти емпіричні (що спостерігаються) і теоретичні (обчислені в припущенні нормального розподілу) частоти в даних. Отже, нехай отримана вибірка $\xi_n = (x_1, x_2, \cdots, x_n)$ об'єму n і для неї складений варіаційний ряд $v_1 < v_2 < \cdots < v_k$ та емпіричний розподіл $(v_1, n_1), \cdots, (v_k, n_k)$, де n_i, $1 \leq i \leq k$, - частоти варіант, а $\omega_i = n_i/n$ – відносні частоти варіант, які будемо вважати емпіричними імовірностями. Припустимо, що за варіаційним рядом ми обчислили теоретичні частоти варіант в нормальному розподілі $m_i = np_i$, де p_i, $1 \leq i \leq k$, теоретичні імовірності варіант. Як обчислити імовірності p_i для нормального розподілу буде показане на практикумі. Таким чином, при заданому рівні значимості α необхідно перевірити нульову гіпотезу, у тому що генеральна сукупність розподілена нормальне.

В якості критерію згоди Пірсон запропонував таку статистику:

$$T^2(\xi_n) = \sum_{i=1}^{k} \frac{(n_i - m_i)^2}{m_i} = n \sum_{i=1}^{k} \frac{(\omega_i - p_i)^2}{p_i} \quad . \tag{4.3}$$

Зрозуміло, що чим менше розрізняються емпіричні та теоретичні частоти, тим менша величина критерію (4.3). і тим ближче емпіричний закон до теоретичного розподілу. Доведено, що при $n \to \infty$ закон розподілу випадкової величини $T^2(\xi_n)$ *незалежно від того, якому закону розподілу підпорядкована генеральна сукупність*, збігається до закону Пірсона хі-квадрат з q ступенями свободи. Тому цей критерій називають критерієм згоди «хі квадрат».

Число ступенів свободи знаходять по рівності $q = k - r - 1$, де k – число варіант (або інтервалів) вибірки і r – число невідомих параметрів розподілу. Зокрема для

нормального розподілу $r = 2$, тому в цьому випадку $q = k - 3$. Для закону Пуассона маємо $r = 1$, тому $q = k - 2$.

Критичне значення $\theta_{\text{кр}}$ визначається за таблицями розподілу хі-квадрат в залежності від рівня значимості α і числа ступенів свободи за формулою:

$$P(T^2(\xi_n) \geq \theta_{\text{кр}}(\alpha, q)) = \alpha \ . \tag{4.4}$$

Таким чином, нульова гіпотеза $H_0: X \sim N(\theta_1, \theta_2^2)$ *приймається, якщо* $P(T^2(\xi_n) < \theta_{\text{кр}}(\alpha, q))$ і *відхиляється, якщо* $P(T^2(\xi_n) \geq \theta_{\text{кр}}(\alpha, q)$.

Для спрощення обчислень перетворимо формулу (4.3) скориставшись рівностями $n_1 + \cdots + n_k = n = m_1 + \cdots + m_k$, оскільки $p_1 + \cdots + p_k = 1$.

$$\sum_{i=1}^{k} \frac{(n_i - m_i)^2}{m_i} = \sum_{i=1}^{k} \frac{n_i^2}{m_i} - n = n \sum_{i=1}^{k} \frac{\omega_i^2}{p_i} - n \ . \tag{4.5}$$

Пример 4.2. *Генетичні закони Менделя*. В експериментах з селекцією гороху Мендель спостерігав частоти різних видів насіння, одержуваних при схрещуванні рослин з круглим жовтим насінням і рослин з зморшкуватим зеленим насінням. У таблиці 4.2 наведено експериментальні дані та теоретичні ймовірності, що визначаються відповідно до законів Менделя. Перевірити гіпотезу H_0 про узгодженість емпіричних та теоретичних частот за допомогою критерію хі-квадрат.

Таблиця 4.2.

Тип насіння	Відносні частоти	Імовірності
Кругли жовті	$315/556$	$9/16$
Зморшкуваті жовті	$101/556$	$3/16$
Кругли зелені	$108/556$	$3/16$
Зморшкуваті зелені	$32/556$	$1/16$

Розв'язування. Тут $n = 556$, $k = 4$, $q = 3$, $n_1 = 315$, $n_2 = 101$, $n_3 = 108$, $n_4 = 32$, $m_1 = 556 \cdot 9/16 = 312{,}75$, $m_2 = m_3 = 556 \cdot 3/16 = 104{,}25$, $m_4 = 556/16 = 34{,}75$. За формулою (4.5) маємо:

$$T^2(\xi_{556}) = \frac{315^2}{312{,}75} + \frac{101^2}{104{,}25} + \frac{108^2}{104{,}25} + \frac{32^2}{34{,}75} - 556 \approx 0{,}47 \ .$$

За таблицею отримуємо, значення $T^2(\xi_{556})$ знаходимо між кватилями $z_{0{,}05} = 0{,}35$ і $z_{0{,}1} = 0{,}58$ (Excel дає значення ХИ2РАСП$(0{,}47; 3; 1) \approx 0{,}07$). Таким чином, спостерігаємо дуже гарну згоду з гіпотезою H_0, тому її треба прийняти вже на рівні 0,08. □

Інші приклади застосування критерія Пірсона будуть розглянути на практикумі.

5. Кореляція

Поняття *кореляції* та *регресії* виникли в середині XIX ст. завдяки роботам англійських статистиків Пірсона і Гальтона. Перший термін походить від латинського «corelatio» (співвідношення, взаємозв'язок). Другий термін (від лат. «regressio» - рух назад) запроваджено Гальтоном.

5.1. Детермінована і стохастична залежність

В багатьох прикладних задачах необхідно виявити залежність між двома ознаками X, Y одного і того ж об'єкту. Тоді X називають незалежною змінною або *факторною ознакою* (фактором), а Y – залежною змінною або *результативною ознакою*.

Якщо кожному значенню фактору X відповідає *одне і тільки одне значення* ознаки Y, то така залежність називається *функціональною* (або *детермінованою*) і позначається як $Y = f(X)$. При детермінованої залежності зв'язок між величинами настільки тісний, що знаючи значення однієї з них можна точно вказати значення іншої. Функціональні залежності часто зустрічаються в природознавстві, наприклад, закон Клапейрона-Менделєєва в фізиці записується у вигляді рівності $PV = RT$.

Економічні показники підлягають впливу багатьох випадкових факторів, тому значенню величині X зазвичай відповідають *декілька значень* величини Y. Однак в цьому випадку між змінними X і Y може бути *статистичний зв'язок*.

Означення 5.1. Випадкові величини X і Y називаються *статистично* (або *стохастичне*) *залежними*, якщо розподіл однієї з них залежіть від значення іншої.

Дослідження статистичного зв'язку між змінними являється складним і трудомістким процесом, оскільки він вимагає аналізу багатовимірних таблиць даних. Тому зазвичай досліджується частковий випадок статистичної залежності, якій називається *кореляційним зв'язком*.

Означення 5.2. Статистична залежність між ознаками X і Y, при якої кожному значенню факторної ознаки X відповідає умовне математичне очікування (середнє значення) ознаки Y називається *кореляційною залежністю* між X і Y.

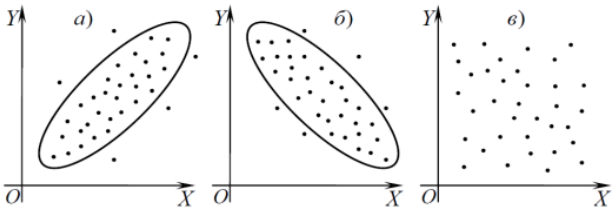
Іншими словами, при кореляційної залежності існує функціональна залежність виду:

$$E_x(Y) = f(x)\,, \qquad (5.1)$$

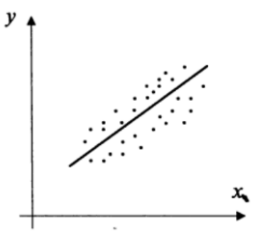
яка називається *рівнянням регресії*. Тут вважається, що функція $f(x)$ не є константою $f(x) \neq \text{const}$. Якщо $f(x) = \text{const}$, то говорять, що кореляційної залежності між X і Y *немає*.

Для знаходження рівняння регресії (5.1), взагалі кажучи, потрібно знати закон розподілу двовимірної випадкової величини (X, Y). На практиці ми маємо лише вибірку пар значень (x_i, y_i), $1 \leq i \leq n$, причому частіше всього зустричаються вибірки двох видів:

a) на площині точки створюють «облако» (кластер) еліптичного типу (рис. 5.1 а,б);
b) точки розташовані в околі деякої прямої або кривої (рис. 5.2).



5.1. Облака точок.



5.2. Зашумлена пряма.

Випадки а) і б) на рис. 5.1 свідчять о наявності кореляційної залежності між X та Y і дослідженню підлягає рівень цієї залежності. Рівномірний розподіл точок у виипадку в) на рис. 5.1 свідчить про *відсутність кореляційної залежності*. Випадок на рис. 5.2 відповідає «функціональної» залежності між X і Y, яка зіпсована *випадковим шумом*. Залежності першого типу вивчаються методами *кореляційного аналізу*, а другого – методами *регресійного аналізу*.

Показником, що вимірює стохастчний зв'язок між змінними X і Y є *коефіцієнт кореляції*, якій свідчить з певною імовірністю, наскільки кореляційний зв'язок між змінними близкий до *лінійної залежності*. Спочатку нагадаємо деяки теоретичні властивості коваріації і коефіцієнта кореляції з теорії імовірності.

5.2. Коваріація і коефіцієнт кореляції

Нехай задані дві випадкових величини (генеральні сукупності) X і Y, для яких відомі їх математичні очікування $E(X) = a,\ E(Y) = b$ і дисперсії $D(X),\ D(Y)$. Однак цих параметрів недостатньо щоб характеризувати двовимірну випадкову величину (X, Y), тому що вони не відображають ступеня залежності величин X і Y. Цю роль виконують *коваріація* та *коефіцієнт кореляції*.

Коваріація $cov(X, Y)$ (або кореляційний момент) випадкових величин X і Y визначається формулою:

$$cov(X, Y) = E[(X - a)(Y - b)]\ . \tag{5.2}$$

Іншими словами, коваріація це математичне очікування добутку двох центрованих випадкових величин $X - a$ і $Y - b$. Наприклад, для дискретних випадкових величин $X = (x_1, \cdots, x_m)$ і $Y = (y_1, \cdots, y_n)$ маємо:

$$cov(X, Y) = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_i - a)\big(y_j - b\big)p_{ij}\ \ , \tag{5.3}$$

де p_{ij} – імовірність події $X = x_i$ і $Y = y_j$.

Із формули (5.2) видно, що коваріація випадкової величини X сама з собою дорівнює її дисперсії $cov(X, X) = D(X)$. Таким чином, коваріація характеризує як ступень *залежності* величин X і Y, так і розкид випадкової величини (X, Y) навколо точки (a, b). Про це, зокрема, свідчать властивості коваріації.

(1) Коваріація двох *незалежних* випадкових величини X і Y дорівнює нулю $cov(X, Y) = 0$. Дійсно, в цьому випадку $p_{ij} = p_i p_j$ і з формули (5.3) випливає, що кореляція буде добутком двох центральних моментів першого порядку, кожний з яких дорівнює нулю:

$$cov(X, Y) = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_i - a)\big(y_j - b\big)p_{ij} = \sum_{i=1}^{m}(x_i - a)\,p_i \sum_{j=1}^{n}\big(y_j - b\big)\,p_j = 0\ .$$

(2) Має місце рівність:

$$E(XY) = E(X)E(Y) + cov(X, Y)\ . \tag{5.4}$$

Із формули (5.2) і властивостей математичного сподівання маємо:

$$cov(X, Y) = E[(XY - Xb - aY + ab)] = E(XY) - E(X)b - aE(X) + ab = E(XY) - ab.$$

(3) Має місце нерівність, яка випливає із нерівності Коші-Буняковського:

$$|cov(X, Y)| \le\ \sigma(X)\sigma(Y)\ , \tag{5.5}$$

де $\sigma(X)$ і $\sigma(Y)$ середні квадратичні відхилення величин X і Y.

За допомогою коваріації можна отримати додаткові властивості математичного очікування і дисперсії. Наприклад, із двох перших властивостей коваріації випливає, що для незалежних випадкових величин X і Y виконується рівність $E(XY) = E(X)E(Y)$. Крім того, має місце така рівність:

$$D(X + Y) = D(X) + D(Y) + 2cov(X, Y)\ . \tag{5.6}$$

Дійсно, маємо $E(X + Y) = E(X) + E(Y) = a + b$. Звідси отримуємо:

$$D(X + Y) = E(X - a + Y - b)^2 = E(X - a)^2 + 2E[(X - a)(Y - b)] + E(Y - b)^2 = \\ = D(X) +\ 2cov(X, Y) + D(Y)\ .$$

Із рівності (5.6) і першої властивості коваріації слідує, що для незалежних випадкових величин X і Y виконується рівність $D(X+Y)=D(X)+D(Y)$.

Коваріація, як зазначалося, характеризує як ступінь залежності двох випадкових величин, так їх розкид, розсіяння. Крім того, вона величина розмірна, її розмірність визначається добутком розмірностей випадкових величин. Це затрудняє використання коваріації як оцінки ступеня залежності випадкових величин. Тому для цієї оцінки використовують безрозмірний коефіцієнт кореляції.

Означення 5.3. *Коефіцієнтом кореляції* $corr(X,Y)$ двох випадкових величин X і Y називається відношення їх коваріації до добутку їх середніх квадратичних відхилень:

$$corr(X,Y)=\frac{cov(X,Y)}{\sigma(X)\sigma(Y)}\ . \qquad (5.7)$$

Відмітимо властивості коефіцієнта кореляції, які в основному випливають із властивостей коваріації:

1) Із нерівності (5.5) випливає, що коефіцієнт кореляції приймає значення з інтервалу $[-1,1]$, тобто $-1\le corr(X,Y)\le 1$.
2) Якщо випадкові величини X і Y *незалежні*, то з першої властивості коваріації слідує, що $corr(X,Y)=0$, У зв'язку з цим випадкові величини X і Y називаються *некорельованими*, якщо їх коефіцієнт кореляції дорівнює нулю. Таким чином, із незалежності випадкових величин випливає, що вони некорельовані. Зворотне твердження, взагалі кажучи, невірне, оскільки існують залежні некорельовані випадкові величини.
3) Якщо коефіцієнт кореляції двох випадкових величин X і Y дорівнює (за абсолютною величиною) одиниці, то між цими випадковими величинами *існує лінійна функціональна залежність*:

$$Y=b-\frac{\sigma(Y)}{\sigma(X)}(X-a),\text{ при } corr(X,Y)=-1,\ Y=b+\frac{\sigma(Y)}{\sigma(X)}(X-a),\text{ при } corr(X,Y)=1.$$

5.3. Вибірковий коефіцієнт кореляції

Нехай $\xi_n=(x_1,\cdots,x_n)$.і $\eta_n=(y_1,\cdots,y_n)$ - вибірки даних із генеральних сукупностей X і Y відповідно, тоді позначимо як $\overline{\xi_n}$ і $\overline{\eta_n}$ їх середні значення. Позначимо також середнє значення вибірки $(x_1y_1,\cdots,x_ny_n)$ як $\overline{\xi_n\eta_n}$. Тоді *вибірковим коефіцієнтом кореляції* $r(\xi_n,\eta_n)$ називається таке число:

$$r(\xi_n,\eta_n)=\frac{\overline{\xi_n\eta_n}-\overline{\xi_n}\,\overline{\eta_n}}{\sigma(\xi_n)\sigma(\eta_n)}\ , \qquad (5.8)$$

де $\sigma(\xi_n)$ і $\sigma(\eta_n)$ - середні квадратичні відхилення вибірок ξ_n і η_n. Із формули (5.4) видно, що чисельник у формулі (5.8) є вибірковим аналогом коваріації. Таким чином, вибірковий коефіцієнт кореляції визначається формулою, яка аналогічна формулі (5.7), тільки коваріація і відхилення змінені на їх вибіркові аналоги.

Вибірковий коефіцієнт кореляції є *показником тісноти лінійної залежності* між випадковими величинами, як і його теоретичний аналог $corr(X,Y)$. Наприклад, на рис. 5.3 показані дві кореляційні залежності величини Y від X. Зрозуміло, що у випадку а) залежність між змінними менш тісна і коефіцієнт кореляції має бути меншим, ніж у випадку б), тому що точки кореляційного поля а) далі відстоюють від прямої, ніж точки поля б).

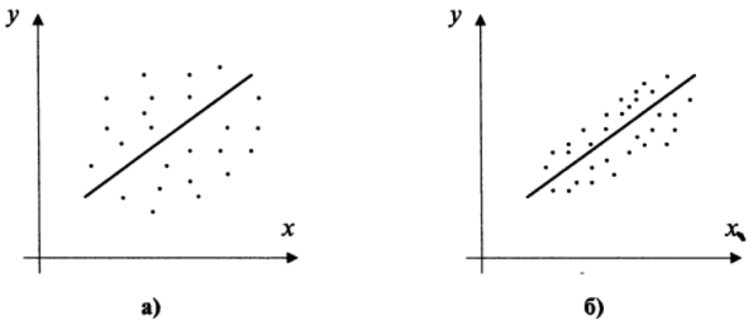


Рис. 5.3. Кореляційні залежності.

Введемо позначення $S_x = x_1 + \cdots + x_n$, $S_y = y_1 + \cdots + y_n$, $S_{xy} = x_1y_1 + \cdots + x_ny_n$, $S_{2x} = x_1^2 + \cdots + x_n^2$, і $S_{2y} = y_1^2 + \cdots + y_n^2$. Помножимо і чисельник і знаменник формули (5.8) на n^2, тоді отримуємо таку формулу:

$$r(\xi_n, \eta_n) = \frac{nS_{xy} - S_xS_y}{\sqrt{nS_{2x} - S_x^2}\sqrt{nS_{2y} - S_y^2}}. \qquad (5.9)$$

Ця формула найбільш зручна для практичних розрахунків, так як по ній коефіцієнт кореляції знаходиться безпосередньо з даних спостережень і на його величині не позначається округлення даних, пов'язані з розрахунком середніх та відхилень від них.

Приклад 5.1. Знайти коефіцієнт кореляції між продуктивністю праці Y та енергоозброєністю праці X (в розрахунку на одного працюючого) для 14 підприємств регіону за даними приведеними в таблиці.

Таблиця 5.1.

x_i	2,8	2,2	3,0	3,5	3,2	3,7	4,0	4,8	6,0	5,4	5,2	5,4	6,0	9,0
y_i	6,7	6,9	7,2	7,3	8,4	8,8	9,1	9,8	10,6	10,7	11,1	11,8	12,1	12,4

Розв'язування. Обчислимо суми $S_x = 64{,}2$, $S_y = 132{,}9$, $S_{xy} = 650{,}99$, $S_{2x} = 335{,}26$, $S_{2y} = 1313{,}95$. За формулою (5.9) маємо:

$$r(\xi_{14}, \eta_{14}) = \frac{14 \cdot 650{,}99 - 64{,}2 \cdot 132{,}9}{\sqrt{14 \cdot 335{,}26 - 64{,}2^2}\sqrt{14 \cdot 1313{,}95 - 132{,}9^2}} \approx 0{,}898\,.$$

Це говоріть о тісному зв'язку між змінними. □

Зазначимо основні властивості коефіцієнта кореляції (при достатньо великому обсязі вибірки n), аналогічні властивостям коефіцієнта кореляції двох випадкових величин $corr(X, Y)$.

1. Вибірковий коефіцієнт кореляції приймає значення з інтервалу $[-1{,}1]$, тобто $-1 \le r(\xi_n, \eta_n) \le 1$. Залежно від того, наскільки близько $|r(\xi_n, \eta_n)|$ наближається до 1, розрізняють зв'язок слабкий, помірний, помітний, достатньо тісний, тісний і дуже тісний, тобто, чим ближче $|r(\xi_n, \eta_n)|$ до 1, то тим тісніше зв'язок.
2. Якщо до всіх значень у вибірці додати одне і теж число, або помножити їх на одне і теж ненульове число, то вибірковий коефіцієнт кореляції не зміниться.
3. При $r(\xi_n, \eta_n) = \pm 1$ кореляційний зв'язок перетворюється на *лінійну функціональну залежність*. При цьому всі точки (x_i, y_i) розташовуються на одній прямій.

На рис. 5.4 показані випадки, коли модуль коефіцієнта кореляції близький до одиниці.

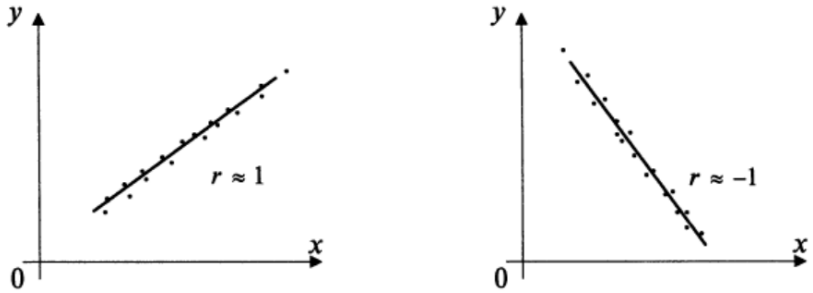


Рис. 5.4. Модуль коефіцієнта кореляції близький до одиниці.

При $r(\xi_n, \eta_n) = 0$ лінійний кореляційний зв'язок *відсутній*. При цьому лінія регресії (5.1) перетворюється в пряму, яка паралельна осі абсцис (рис. 5.5). Рівність $r(\xi_n, \eta_n) = 0$ говорить лише про відсутність *лінійної кореляційної залежності*, але не взагалі про відсутність кореляційної, а тим більше статистичної залежності.

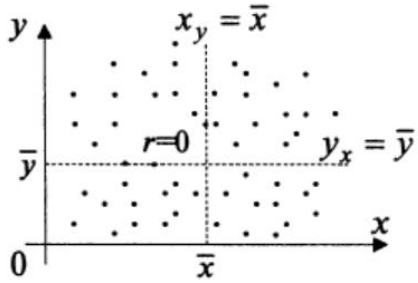


Рис. 5.5. Відсутність лінійної кореляційної залежності.

5.4. Кореляційний аналіз

Основне завдання кореляційного аналізу полягає у *виявленні зв'язку між випадковими змінними шляхом точкової та інтервальної оцінок* різних коефіцієнтів кореляції. Вибірковий коефіцієнт кореляції $r(\xi_n, \eta_n)$ є *оцінкою генерального коефіцієнта кореляції* $corr(X, Y)$, тим більш точною, чим більше обсяг вибірки n. Зазначені вище властивості, строго кажучи, справедливі лише для $corr(X, Y)$. Однак при досить великому n їх можна поширити і на вибірковий коефіцієнт $r(\xi_n, \eta_n)$. Зауважимо також, що на відміну від генерального коефіцієнта $corr(X, Y)$, вибірковий коефіцієнт $r(\xi_n, \eta_n)$ являється *випадковою величиною*.

Нехай обчислене значення $r_n = r(\xi_n, \eta_n) \neq 0$. Виникає питання, чи це справді існує лінійний кореляційний зв'язок між змінними X і Y в генеральній сукупності або це є наслідком випадковості відбору значень у вибірці (тобто при іншому відборі можливо, наприклад, $r_n = 0$ або коефіцієнт змінить свій знак). Зазвичай у цих випадках перевіряється гіпотеза H_0 про відсутність лінійної кореляційної залежності між змінними у генеральній сукупності, тобто H_0: $corr(X, Y) = 0$.

Доведено, що статистика:

$$t_n = \frac{r_n\sqrt{n-2}}{\sqrt{1-r_n^2}}\ , \tag{5.10}$$

має розподіл Стьюдента з $k = n - 2$ ступенями свободи. Гіпотеза H_0 відхиляється, якщо вибірковий коефіцієнт кореляції r_n суттєво відрізняється від нуля. Для цього обчислюється статистика t_n, задається рівень надійності $\gamma = 1 - \alpha$ і визначається квантиль розподілу Стьюдента $x_{\gamma,k}$ (за таблицями або на комп'ютері). Якщо виконується нерівність:

$$|t_n| > x_{\gamma,k}\ , \tag{5.11}$$

то гіпотеза H_0 відхиляється, в іншому випадку, коли $|t_n| \leq x_{\gamma,k}$, - приймається.

Приклад 5.2. Перевірити з надійністю $\gamma = 0{,}95$ вибірковий коефіцієнт кореляції $r = 0{,}74$ між змінними X і Y, який був отриманий на вибірці розміром $n = 50$.

Розв'язування. Обчислимо статистику критерія (5.10):

$$t_{50} = \frac{0{,}74\sqrt{48}}{\sqrt{1-0{,}74^2}} \approx 7{,}62\ .$$

Знаходимо по таблицям критичне значення статистики $x_{0{,}95;48} = 2{,}01$. Оскільки $t_{50} > x_{0{,}95;48}$, то нульова гіпотеза H_0: $corr(X, Y) = 0$ відхиляється, тобто генеральний коефіцієнт кореляції значимо відрізняється від нуля. □

Для значущого вибіркового коефіцієнта кореляції r_n, який задовольняє нерівність (5.11), доцільно знайти довірчий інтервал (інтервальну оцінку), яка із заданою надійністю $\gamma = 1 - \alpha$ містить (точніше, «накриває») невідомий *генеральний коефіцієнт кореляції* $\rho = corr(X, Y)$. Для побудови такого інтервалу необхідно знати вибірковий розподіл коефіцієнта кореляції r_n, який при $\rho \neq 0$, несиметричний і дуже повільно (зі зростанням n) збігається до нормального розподілу. Тому використовують спеціально підібрані функції від r_n, які сходяться до добре вивчених розподілів. Найчастіше для підбору функції застосовують z-перетворення Фішера:

$$z = \frac{1}{2}\ln\left(\frac{1+r_n}{1-r_n}\right). \quad (5.12)$$

Розподіл випадкової величини z вже при невеликих n наближається до нормального з математичним очікуванням:

$$E(z) = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) + \frac{\rho}{2(n-1)}\ . \quad (5.13)$$

і дисперсією $D(z) = 1/(n-3)$. Тоді ми можемо спочатку знайти довірчий інтервал для $E(z)$. Дійсно, випадкова величина $(z - E(z))/\sqrt{D(z)}$ розподілена за стандартним нормальним законом (практикум 2 формула 2.13), значення z і дисперсія $D(z)$ нам відомі, тому довірчий інтервал з надійністю $\gamma = 1 - \alpha$ для невідомого очікування $E(z)$ буде мати вигляд (лекція 3):

$$z - \frac{x_{(1+\gamma)/2}}{\sqrt{n-3}} \le E(z) \le z + \frac{x_{(1+\gamma)/2}}{\sqrt{n-3}}\ , \quad (5.14)$$

де $x_{(1+\gamma)/2}$ – квантиль стандартного нормального розподілу.

Позначимо як z_1 і z_2 нижню і верхню межу довірчого інтервалу (5.14). Тепер нам треба повернутися від z до ρ, тобто знайти обернену функцію для функції (5.13). Це складна функція, для якої існують спеціальна таблиця, тому для спрощення обчислень користуються оберненою функцією для функції (5.12).якою є гіперболічний тангенс:

$$\mathrm{th(z)} = \frac{e^z - e^{-z}}{e^z + e^{-z}}\ .$$

Таким чином, отримуємо довірчий інтервал для генерального кореляційного коефіцієнта:

$$\mathrm{th}(z_1) \le corr(X,Y) \le \mathrm{th}(z_2)\ . \quad (5.15)$$

Приклад 5.3. Знайти з надійністю γ =0,95 інтервальну оцінку (довірчий інтервал) для генерального коефіцієнта кореляції $corr(X,Y)$ між добовим виробленням продукції Y та величиною основних виробничих фондів X, якщо вибірковий коефіцієнт кореляції $r = 0{,}74$ і $n = 50$.

Розв'язання. Оскільки коефіцієнт кореляції $corr(X,Y)$ значимий (див. приклад 5.2), то побудуємо довірчий інтервал для $corr(X,Y)$, застосовуючи z-перетворення Фішера. За формулою (5.12) маємо:

$$z = \frac{1}{2}\ln\left(\frac{1+0{,}74}{1-0{,}74}\right) \approx 0{,}95\ .$$

Далі маємо $(1+\gamma)/2 = 0{,}975$, звідси за таблицями отримуємо $x_{0{,}975} \approx 1{,}96$. Тоді за формулою (5.14) дістаємо довірчий інтервал для $E(z)$:

$$0{,}95 - \frac{1{,}96}{\sqrt{47}} \le E(z) \le 0{,}95 + \frac{1{,}96}{\sqrt{47}}\ ,$$

або $0{,}664 \le E(z) \le 1{,}236$. Звідси за формулою (5.15) отримуємо довірчий інтервал для коефіцієнта кореляції $\mathrm{th}(0{,}664) \le corr(X,Y) \le \mathrm{th}(1{,}236)$ або $0{,}581 \le corr(X,Y) \le 0{,}844$.

6. Регресія

У практиці економічних досліджень дуже часто наявні дані не можна вважати вибіркою з нормальної сукупності, наприклад, коли *одна з змінних, що розглядаються, не є випадковою*. У цих випадках намагаються визначити криву, яка дає найкраще (в деякому сенсі) наближення до вихідних даних. Наприклад, якщо ми хочемо передбачити ціну на товар у наступному кварталі, то ми намагаємося це зробити на основі даних, отриманих у поточному кварталі. Змінна часу в цьому разі не є випадковою величиною. Відповідні методи наближення отримали назву *регресійного аналізу*. Таким чином, основною задачею регресійного аналізу *є встановлення форми залежності між змінними*, статистична оцінка функції регресії, прогноз значень залежної змінної.

Точніше, до регресійного аналізу відносяться задачі виявлення спотвореної випадковим "шумом" функціональної залежності:

$$y = f(x_1, \cdots, x_m) + \varepsilon(x_1, \cdots, x_m)\ , \qquad (6.1)$$

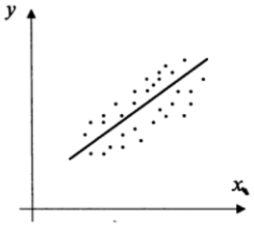
де y – залежна змінна, $x_1, \cdots, x_m$ – незалежні (невипадкові) змінні, $\varepsilon(x_1, \cdots, x_m)$ – випадкова змінна, яка називається *збуренням*.

Вихідними даними служить таблиця експериментально отриманих "зашумлених" значень змінної y на різних наборах значень змінних $x_1, \cdots, x_m$. Основною метою зазвичай є якомога точніший прогноз значень змінної y для заданих значень *предикторних* (від англійського Predict - передбачати) змінних $x_1, \cdots, x_m$. Підкреслимо, що значення предикторних змінних в регресійному аналізу вважаються *невипадковими*, але регресія працює і в тому випадку, коли вони випадкові, наприклад, в кореляційному аналізі, де також використовується регресія. Зауважимо також, що для предикторних змінних $x_1, \cdots, x_m$ в літературі зустрічаються назви – фактори, факторні признаки, регресори і т. д., а для змінної y – вихідна змінна, відгук, змінна, що пояснюється, результативна ознака тощо.

6.1. Парна лінійна регресія (підгонка прямої)

Проілюструємо осовні ідеї регресії на простому прикладі підгонки прямої під хмару експериментальних точок (рис. 6.1), отриманих відповідно до *парної лінійної моделі*:

$$y_i = ax_i + b + \varepsilon_i\ , \qquad 1 \le i \le n\ . \qquad (6.2)$$



6.1. Зашумлена пряма.

Тут коефіцієнти прямої a і b – невідомі параметри, $x_i, 1 \le i \le n$, - (невипадкові) значення незалежної змінної, ε_i, $1 \le i \le n$, – випадкові величини, які моделюють випадкову змінну в рівнянні (6.1). Зауважимо одразу, якщо значення незалежної змінної x отримані випадковим чином, як в кореляційному аналізі, то змінюється прикладний сенс регресії (оцінка коефіцієнта кореляції), але математичний спосіб розвязування задачі залишається тем же самим.

Спочатку відмитимо, властивості випадкових величин ε_i (залишків), які називаються *основними передумовами регресійного аналізу*, (аксіомами регресії):

(А1) Випадкові величини ε_i, $1 \le i \le n$, в рівнянні (6.2) являються незалежними (некорельованими), однаково розподіленими, з нулевим математичним очікуваням:

$$E(\varepsilon_i) = 0,\ \ 1 \le i \le n\ ,\ \ E(\varepsilon_i \varepsilon_j) = 0\ ,\ \ 1 \le i < j \le n\ . \qquad (6.3)$$

(А2) Збурення ε_i, $1 \le i \le n$, є нормально розподіленими випадковими величинами.

Для отримання рівняння регресії достатньо першої аксиоми. Друга аксіома необхідна для оцінки *точності* рівняння регресії і його параметрів. Якщо аксіома А1 не

виконується, а таке на практиці трапляється, можна отримати помилкові результати. Для находження оцінок для коефіцієнтів a і b в рівнянні (6.2) використовується *метод найменших квадратів* (МНК).

6.2. Метод найменших квадратів

Природною умовою підгонки пробної прямої $y = ax + b$ служить близькість до нуля всіх залишків $\delta_i(a,b) = y_i - ax_i - b$ (рис. 6.2).

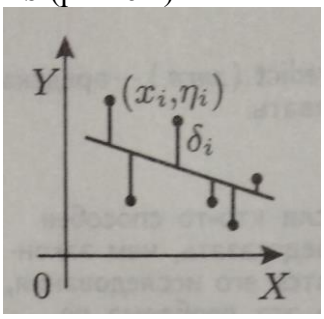


Рис. 6.2. Залишки.

Загальну міру близькості до нуля всіх залишків можна вибирати по-різному (наприклад як $\max|\delta_i|$), але найпростіші формули для оцінок $\hat{a}$ і $\hat{b}$ коефіцієнтів регресії виходять, якщо в такий спосіб взяти суму квадратів всіх залишків:

$$F(a,b) = \sum_{i=1}^{n} \delta_i^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2 \ . \qquad (6.4)$$

Позначимо як $\bar{x}$ і $\bar{y}$ – середні значення величин x_i і y_i $1 \le i \le n$, відповідно. Мінімум функції $F(a,b)$ досягається у точці $(\hat{a},\hat{b})$, де обидві частинні похідні $\partial F/\partial a$ і $\partial F/\partial b$ дорівнюють нулю (теорема Ферма, необхідна умова екстремуму). Звідси заключаємо, що мінімум функції $F(a,b)$ досягається у точці $(\hat{a},\hat{b})$, яка визначається формулами:

$$\hat{a} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) / \sum_{i=1}^{n} (x_i - \bar{x})^2 \ , \qquad (6.5)$$

$$\hat{b} = \bar{y} - \hat{a} \cdot \bar{x} \ . \qquad (6.6)$$

Кожен, хто вивчає статистику, повинен раз у житті виконати всі необхідні обчислення та отримати самостійно формули (6.5) та (6.6). Для цього треба з рівняння $\partial F/\partial b = 0$ отримати співвідношення (6.6) та підставити його у функцію F. Потім із рівняння $\partial F(a, \bar{y} - a \cdot \bar{x})/\partial a = 0$ отримати рівність (6.5).

Із формули (6.6) випливає, що МНК-пряма $y = \hat{a} \cdot x + \hat{b}$ завжди проходить через середню точку $(\bar{x},\bar{y})$. Можна трошки спростити формулу (6.5), якщо поділити чисельник і знаменник у цій формулі на n і помітити, що чисельник перетворюється на вибіркову "коваріацію" векторів $x = (x_1, \cdots, x_n)$ і $y = (y_1, \cdots, y_n)$, а знаменник на вибіркову "дисперсію" вектору x. Тому можна записати таку рівність:

$$\hat{a} = \frac{cov(x,y)}{\sigma(x)^2} \ . \qquad (6.7)$$

Хоча ця формула не має імовірнісного сенсу в регресійному аналізі, але її легко запам'ятати. Коефіцієнт $\hat{a}$ показує, наскільки змінюється y при зміні x на одиницю. Навпаки, у кореляційному аналізі ця формула пов'язує коефіцієнт $\hat{a}$ з вибірковим кореляційним коефіцієнтом $r(x,y)$, що випливає з формули (5.8) (лекція 5):

$$\hat{a} = \frac{cov(x,y)}{\sigma(x)^2} = \frac{cov(x,y)}{\sigma(x)\sigma(y)} \frac{\sigma(y)}{\sigma(x)} = r(x,y) \frac{\sigma(y)}{\sigma(x)} \ . \qquad (6.8)$$

Метод найменших квадратів був вперше опублікований у 1805 р. французьким математиком Лежандром. Великий німецький математик Гаус стверджував, що він використав МНК ще до 1803 р. і між ними виникла боротьба за пріоритет. Імовірно вони відчули, що МНК стане основним методом статистичного аналізу.

6.3. Нелінійна регресія

Співвідношення між соціально-економічними явищами та процесами не завжди можна відобразити лінійними функціями, оскільки при цьому можуть з'явитися невиправдано великі помилки. У таких випадках використовують *нелінійну регресію*. Найчастіше зустрічаються такі види рівнянь нелінійної регресії:

- *поліноміальне* $y = a_0 + a_1 x + \cdots + a_m x^m$;
- *гіперболічне* $y = a_0 + a_1/x$;
- *показникове* (степеневе) $y = ax^b$.

Приклад 6.1. Припустимо, що треба дослідити залежність урожайності зернових культур від кількості опадів за даними, які приведені в таблиці:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
25	27	30	35	36	38	39	41	42	45	46	47	50	52	53
23	24	27	27	32	31	33	35	34	32	29	28	25	24	25

Тут в другому рядку показана кількість опадів, в третьому рядку – урожайність.

Розв'язування. Побудуємо кореляційне поле для цієї задачі (рис. 6.3), із якого становиться очевидно, що в даному випадку найбільш підходящою є квадратична залежність $y = a_0 + a_1 x + a_2 x^2$.

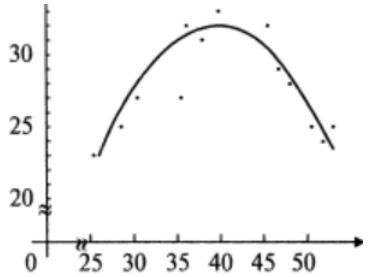


Рис. 6.3. Квадратична залежність.

Коефіцієнти регресії a_0, a_1, a_2 можна знайти за методом найменших квадратів. Для цього спочатку складаємо функцію залишків:

$$F(a_0, a_1, a_2) = \sum_{i=1}^{15} \delta_i^2 = \sum_{i=1}^{15} (y_i - a_0 - a_1 x_i - a_2 x_i^2)^2 .$$

Потім обчислюємо частинні похідні $\partial F/\partial a_0$, $\partial F/\partial a_1$, $\partial F/\partial a_2$ і прирівнюємо їх до нуля. В результаті отримуємо систему із трьох лінійних рівнянь з трьома невідомими:

$$\begin{cases} 15a_0 + a_1 \sum_{i=1}^{15} x_i + a_2 \sum_{i=1}^{15} x_i^2 = \sum_{i=1}^{15} y_i , \\ a_0 \sum_{i=1}^{15} x_i + a_1 \sum_{i=1}^{15} x_i^2 + a_2 \sum_{i=1}^{15} x_i^3 = \sum_{i=1}^{15} x_i y_i , \\ a_0 \sum_{i=1}^{15} x_i^2 + a_1 \sum_{i=1}^{15} x_i^3 + a_2 \sum_{i=1}^{15} x_i^4 = \sum_{i=1}^{15} x_i^2 y_i . \end{cases} \tag{6.9}$$

Студентам пропонується самим розв'язати цю систему, а ми перейдемо до більш загального способу розв'язання цієї задачі.

Зауважимо, що поліноміальне і гіперболічне рівняння *лінійні за невідомими змінними*, а показникове становиться лінійним після логарифмування. Іншими словами, вся нелінійність залишається за предикторними змінними. Тому замість того, щоб розв'язувати нелінійну задачу від однієї змінної можна розв'язати *лінійну задачу від декількох змінних*.

6.4. Множинна лінійна регресія

Припустимо, що (з точністю до випадкових помилок) змінна y є лінійною функцією предикторних змінних $x_1, \cdots, x_m$ з невідомими коефіцієнтами $y = a_1x_1 + \cdots + a_mx_m$. Позначимо як y_i, $1 \le i \le n$, i-е спостереження змінної y, що відповідає заданим значенням $x_{i1}, \cdots, x_{im}$, $1 \le i \le n$, предикторних змінних. Тоді *модель множинної лінійної регресії* матиме такий вигляд:

$$y_i = a_1x_{i1} + \cdots + a_mx_{im} + \varepsilon_i\,, \qquad 1 \le i \le n\,, \tag{6.10}$$

де випадкові величини ε_i, $1 \le i \le n$, задовольняють умовам (6.3). Тут будемо вважати, що $m \le n$, тобто число спостережень не менше числа предикторів. Для зручності обчислень перейдемо до векторних і матричних позначень. Нехай $Y = (y_1, \cdots, y_n)$ – вектор-стовпець значень залежної змінної, $a = (a_1, \cdots, a_m)$ – вектор-стовпець невідомих, $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)$ – випадковий вектор-стовпець, а i-им рядком матриці X являється вектор $(x_{i1}, \cdots, x_{im})$, $1 \le i \le n$. Таким чином, матриця X має розмір $n \times m$ і називається *матрицею експерименту*. Тоді лінійна модель (6.10) в матричної формі приймає такий вигляд:

$$Y = X \cdot a + \varepsilon\,. \tag{6.11}$$

Припустимо, що для моделі (6.11) виконується така умова:

(А3) Стовпці матриці X лінійні незалежні.

Оскільки $m \le n$, то із умови (А3) випливає, що ранг матриці X дорівнює m. Умову (А3) зазвичай можна виконати збільшив число спостережень.

Оцінимо параметри $a = (a_1, \cdots, a_m)$ методом найменших квадратів, мінімізуючи за вектором a функцію залишків:

$$F(a) = \sum_{i=1}^{n} (y_i - a_1x_{i1} - \cdots - a_mx_{im})^2 = |Y - X \cdot a|^2\,. \tag{6.12}$$

Таким чином, функція залишків дорівнює квадрату норми вектору $Y - X \cdot a$. Точка мінімуму $\hat{a}$ функції $F(a)$ називається *МНК-оцінкою*, вектор $\hat{\delta} = Y - X \cdot \hat{a}$ – *вектором залишків*, а значення функції в точці мінімуму $F(\hat{a})$ – *остаточною сумою квадратів* RSS (від англійського Residual Sum of Squares).

Далі для знаходження оцінки $\hat{a}$ скористаємося властивостями евклідова лінійного простору $\mathbb{R}^n$ замість диференціювання функції $F(a)$. Розглянемо в просторі $\mathbb{R}^n$ підпростір $L(X)$ породжений стовпцями $x_1, \cdots, x_m$ матриці X:

$$L(X) = \{c_1x_1 + \cdots + c_mx_m : c_i \in \mathbb{R}\,, 1 \le i \le m\}\,. \tag{6.13}$$

Оскільки $X \cdot a = a_1x_1 + \cdots + a_mx_m$, то із формули (6.13) очевидно слідує, що вектор $X \cdot a$ буде пробігати вісь підпростір $L(X)$, коли вектор a пробігає простір $\mathbb{R}^m$.

Таким чином, нам треба в підпросторі $L(X)$ знайти вектор Xa такий, щоб довжина (норма) вектору залишків $Y - Xa$ була мінімальною. Як відомо із геометрії, таким вектором буде *ортогональна проекція* вектору Y на підпростір $L(X)$. Іншими словами, вектор залишків $Y - Xa$ повинен бути перпендикулярним підпростору $L(X)$, а значить і кожному із векторів x_i, $1 \le i \le m$, які його породжують. Це означає, що скалярний добуток кожного вектору-рядка x_i^T на вектор $Y - Xa$ дорівнює нулю $x_i^T(Y - Xa) = 0$, $1 \le i \le m$. Звідси ми отримуємо таке співвідношення:

$$X^T(Y - X \cdot a) = 0 \;\text{ або }\; X^TX \cdot a = X^TY\,, \tag{6.14}$$

де X^T – матриця транспонована до матриці експерименту X.

Систему лінійних рівнянь $X^TX \cdot a = X^TY$ відносно параметрів $a = (a_1, \cdots, a_m)$ можна було отримати і диференціюванням функції $F(a)$, але більш складним чином. Цю систему можна розв'язати методом Гауса або Холецького, але існує і інший шлях розв'язання цієї системи. Зауважимо, що матриця $B = X^TX$ є симетричною додатньо-означеною матрицею (всі власні значення додатні) і тому вона має обернену матрицю B^{-1}. Помножимо матричне рівняння $B \cdot a = X^TY$ зліва на матрицю B^{-1}, тоді отримуємо єдиний розв'язок системи (6.14) у явному вигляді:

$$\hat{a} = B^{-1}X^TY\,. \tag{6.15}$$

Приклад 6.2 (Підгонка поліному). Розв'яжемо задачу із прикладу 6.1 методом множинної лінійної регресії.

Розв'язування. Оскільки регресійне рівняння є параболою $y = a_0 + a_1x + a_2x^2$ з вільним членом a_0, то введемо фіктивну змінну x_0, значення якої во всіх спостереженнях (випробуваннях) буде дорівнювати одиниці $x_0 = 1$. Покладемо також $x_1 = x$ і $x_2 = x^2$, тоді регресійна модель приймає лінійний вигляд $y = a_0x_0 + a_1x_1 + a_2x_2$. Обчислимо матрицю експерименту X розміру 15×3 і представимо її для зручності у транспонованому вигляді:

1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
25	27	30	35	36	38	39	41	42	45	46	47	50	52	53
625	729	900	1225	1296	1444	1521	1681	1764	2025	2116	2209	2500	2704	2809

Аналогічно представимо в транспонованому виді вектор Y:

23	24	27	27	32	31	33	35	34	32	29	28	25	24	25

Тепер множимо за допомогою Excel матрицю X^T на матрицю X і отримуємо матрицю B розміру 3×3:

$$B = X^TX = \begin{pmatrix} 15 & 606 & 25548 \\ 606 & 25548 & 1115808 \\ 25548 & 1115808 & 50158200 \end{pmatrix}.$$

Потім інвертуємо матрицю B за допомогою Excel:

$$B^{-1} = \begin{pmatrix} 30{,}23924 & -1{,}56884 & 0{,}019498 \\ -1{,}56884 & 0{,}08277 & -0{,}00104 \\ 0{,}019498 & -0{,}00104 & 0{,}000013 \end{pmatrix}.$$

Множимо матрицю X^T на вектор-стовпець Y і отримуємо вектор-стовпець X^TY:

$$X^TY = \begin{pmatrix} 429 \\ 17731 \\ 730123 \end{pmatrix}$$

Множимо матрицю B^{-1} на вектор-стовпець X^TY і отримуємо фінальний результат:

$$\hat{a} \approx \begin{pmatrix} -43{,}93 \\ 3{,}83 \\ -0{,}05 \end{pmatrix}.$$

Таким чином, $a_0 = -43{,}93$, $a_1 = 3{,}83$, $a_2 = -0{,}05$ і рівняння регресії має вигляд:

$$y = -43{,}93 + 3{,}83x - 0{,}05x^2 \text{ . } \square$$

7. Дисперсійний аналіз

7.1. Показники варіації

Основою будь-якого статистичного дослідження є формування *якісно однорідної сукупності даних*. Тому серед статистичних методів особливе місце займає дисперсійний аналіз, одним з основних завдань якого є *вимірювання однорідності сукупності даних*. При цьому якщо досліджувана сукупність *розбита на групи* за певною ознакою, то можна визначити дисперсію як у цілому по сукупності, так і *в кожній групі окремо*, крім того можна визначити *між-групову дисперсію*.

Нагадаємо, що характеристикою розсіювання (варіації) значень кількісної ознаки у вибірки $X = (x_1, x_2, \cdots, x_n)$ генеральній сукупності навколо свого середнього значення є вибіркова дисперсія, яка розраховується за формулою:

$$D_{\text{в}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x_{\text{в}}})^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \overline{x_{\text{в}}}^2 \ , \qquad (7.1)$$

де $\overline{x_{\text{в}}}$ – вибіркове середнє.

Припустимо, що вибірка X розділена на k груп $X_1, \cdots, X_k$ даних, тоді, розглядаючи кожну групу як самостійну сукупність даних можна визначити дисперсію в групі, яка характеризує розсіювання значень кількісної ознаки навколо групової середньої величини:

$$D_i = \frac{1}{n_i}\sum_{j=1}^{n_i}(x_{ij} - m_i)^2 \ , \qquad 1 \le i \le k \ , \qquad (7.2)$$

де n_i – число елементів в групі X_i, $1 \le i \le k$, m_i – середнє значення в групі X_i, $1 \le i \le k$, x_{ij}, $1 \le j \le n_i$ – елементи групи X_i, $1 \le i \le k$. Як завжди виконуються умови нормування:

$$\sum_{i=1}^{k} n_i = n \ , \qquad \sum_{i=1}^{k} \omega_i = 1 \ , \qquad (7.3)$$

де $\omega_i = n_i/n$, - частка (або частота) групи X_i, $1 \le i \le k$. Звідси отримаємо співвідношення для групових середніх і вибіркового середнього $m = \overline{x_{\text{в}}}$:

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{k} m_i n_i = \sum_{i=1}^{k} m_i \omega_i \ . \qquad (7.4)$$

Узагальнюючою мірою *внутрішньо-групової варіації є середня з групових дисперсій*, яка розраховується за такою формулою:

$$D_{\text{вг}} = \sum_{i=1}^{k} D_i \omega_i \ . \qquad (7.5)$$

Мірою варіації групових середніх навколо загальної середньої є *між-групова дисперсія*. Вона показує на скільки відрізняються між собою групові середні і розраховується за формулою:

$$D_{\text{мг}} = \sum_{i=1}^{k} (m_i - m)^2 \omega_i \ . \qquad (7.6)$$

Теорема 7.1. Якщо вибірка даних складається із декількох груп, то вибіркова дисперсія дорівнює сумі середньої з внутрішньо-групових і між-групової дисперсій:

$$D_{\text{в}} = D_{\text{вг}} + D_{\text{мг}} \ . \qquad (7.7)$$

Доведення. Спочатку розкриємо між-групову дисперсію за допомогою формул (7.3), (7.4) і (7.6):

$$D_{\text{мг}} = \sum_{i=1}^{k} (m_i^2 - 2m_i m + m^2)\omega_i = \sum_{i=1}^{k} m_i^2 \omega_i - 2m^2 + m^2 = \sum_{i=1}^{k} m_i^2 \omega_i - m^2 \ .$$

Аналогічно розкриваємо середню з внутрішньо-групових дисперсій за формулами (7.2), (7.5) і співвідношенням $\omega_i/n_i = 1/n$:

$$D_{\text{вг}} = \sum_{i=1}^{k} \frac{\omega_i}{n_i} \sum_{j=1}^{n_i} (x_{ij} - m_i)^2 = \sum_{i=1}^{k} \frac{\omega_i}{n_i} \sum_{j=1}^{n_i} (x_{ij}^2 - 2x_{ij}m_i + m_i^2) = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 \right) - \sum_{i=1}^{k} m_i^2 \omega_i \, .$$

Складаємо ці рівності та із формули (7.1) одержуємо:

$$D_{\text{вг}} + D_{\text{мг}} = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 \right) - m^2 = D_{\text{в}} \, .$$

Таким чином, теорему доведено. □

Формула (7.7) називається *правилом додавання дисперсій* і грає важливу роль у статистичному аналізі. Можна сказати, що у формулі (7.7) прихована *основна ідея дисперсійного аналізу*, різниця лише в тому, яким чином та з якою метою дані розбиваються на групи. Відношення між-групової дисперсії до вибіркової характеризує частку варіації результативної ознаки, яка обумовлена впливом групувальної ознаки. Це відношення називається *емпіричним коефіцієнтом детермінації*:

$$\eta^2 = \frac{D_{\text{мг}}}{D_{\text{в}}} \, . \tag{7.8}$$

Для оцінки тісноти зв'язку між групувальною і результативною ознаками розраховується *емпіричне кореляційне відношення*:

$$\eta = \sqrt{\frac{D_{\text{мг}}}{D_{\text{в}}}} \, . \tag{7.9}$$

Емпіричне кореляційне відношення може набувати значень від 0 до 1. Чим більше його значення наближається до 1, тим більш тісним є зв'язок між групувальною і результативною ознаками.

Приклад 7.2. 24 студенти групи розподілені на дві підгрупи за кількістю пропущених занять із статистики: студенти, які пропустили менше, ніж 10 % загальної кількості годин, і всі інші, як показано в таблиці 7.1. Потрібно визначити, чи впливає наявність або відсутність пропусків занять на рейтинг студентів.

Таблиця 7.1.

Групи студентів за кількістю пропусків занять, %	Кількість студентів, чол. (f_i)	Середній бал ($\overline{X}_i$)	Варіація балів (σ_i)
До 10	18	78	11
10 і більше	6	46	16
Усього	24	-	-

Розв'язування. В наших позначеннях маємо $n_1 = 18, n_2 = 6, n = 24, m_1 = 78, m_2 = 78, \sqrt{D_1} = 11, \sqrt{D_2} = 16$. Нам необхідно визначити емпіричне кореляційне відношення, тобто необхідно обчислити між-групову та вибіркову дисперсії. Зауважимо, що ми не можемо обчислити дисперсію за формулою (7.1), оскільки нам невідомі рейтинги студентів в групах, але за допомогою формули (7.7) це можна зробити.

Спочатку за формулою (7.4) обчислимо загальне середнє значення за всією групою:

$$m = \frac{78 \cdot 18 + 46 \cdot 6}{24} = 70 \, .$$

Потім визначаємо між-групову дисперсію за формулою (7.5):

$$D_{\text{мг}} = \frac{(78 - 70)^2 \cdot 18 + (46 - 70)^2 \cdot 6}{24} = 192 \, .$$

Тепер визначимо середню з внутрішньо-групових дисперсій за формулою (7.4):

$$D_{\text{вг}} = \frac{11^2 \cdot 18 + 16^2 \cdot 6}{24} = 154{,}75 \, .$$

За формулою (7.7) визначаємо загальну (вибіркову) дисперсію $D_{\text{в}} = 154{,}75 + 192 = 346{,}75$. Далі визначаємо емпіричний коефіцієнт детермінації $\eta^2 = 192/346{,}75 \approx 0{,}554$.

Отже, 55,4% загальної варіації рейтингу студентів пояснюється пропусками занять. Емпіричне кореляційне відношення складає $\eta = \sqrt{0{,}554} \approx 0{,}744$. Оскільки значення η близьке до 1, то між кількістю пропусків занять та рівнем успішності студенів (балами їх рейтингу) існує достатньо тісний зв'язок. □

7.2. Одно-факторний дисперсійний аналіз

Дисперсійний аналіз визначається як статистичний метод, призначений для оцінки впливу різних факторів на результат експерименту.

Дисперсійний аналіз розробив англійський математик і статистик Рональд Фішер у 1918 р. для обробки результатів агрономічних дослідів щодо виявлення умов отримання максимального врожаю різних сортів сільськогосподарських культур. Дисперсійний аналіз широко застосовується в багатьох сферах діяльності: при дослідженні рівня життя населення, в медицині, хімії, тощо. Зокрема, дисперсійний аналіз використовується для перевірки наявності відмінностей доходів чи витрат різних груп населення, наприклад, він дозволяє встановити, чи є суттєвими регіональні відмінності в середніх доходах, чи ні.

За кількістю факторів, вплив яких досліджується, розрізняють *одно-факторний* та *багатофакторний* дисперсійний аналіз. Ми коротко розглянемо лише одно-факторний дисперсійний аналіз.

Отже одно-факторна дисперсійна модель має вигляд:

$$x_{ij} = \mu + F_i + \varepsilon_{ij} \ , \qquad (7.10)$$

де x_{ij}, $1 \le i \le m$, $1 \le j \le n$ - – значення ознаки, що отримане на i-ому рівні фактору і при j-ому вимірі на цьому рівні, μ - середнє значення ознаки, F_i – вплив фактору на i-ому рівні, ε_{ij} - випадковий вплив інших факторів. Під рівнем фактору розуміється деяка його міра або стан, наприклад, кількість добрив, що вносяться, або номер партії деталей і т. п. Таким чином, при проведенні дисперсійного аналізу досліджувана сукупність даних природним чином розбивається на однорідні групи, за рівнями фактору. Зауважимо, що вплив фактору зазвичай вважається детермінованим, але він може бути і випадковим (друга модель).

Основні передумови дисперсійного аналізу аналогічні передумовам регресійного аналізу:

(А4)Випадкові величини ε_{ij} являються незалежними, однаково розподіленими, з нулевим математичним очікуваням:

$$E(\varepsilon_{ij}) = 0, \quad D(\varepsilon_{ij}) = \sigma^2 \ , \ \ 1 \le i \le m, 1 \le j \le n \ . \qquad (7.11)$$

(А5)Збурення ε_{ij}, $1 \le i \le m, 1 \le j \le n$, є нормально розподіленими випадковими величинами із класу $N(0, \sigma^2)$.

В одно-факторному дисперсійному аналізі статистичні гіпотези формулюють так:

- нульова гіпотеза (Н0) - усі середні значення у групах рівні між собою;
- альтернативна гіпотеза (Н1) - не всі середні рівні між собою, принаймні, існує хоча б дві групи, де вибіркові середні величини відрізняються між собою.

Для перевірки гіпотези щодо суттєвості відмінностей між середніми величинами у вибірках, що сформовані за факторною ознакою, використовується *F-критерій Фішера* (точніше Фішера-Снедекора, але ми будемо використовувати тільки одне прізвище).

Розглянемо, наприклад, задачу, коли необхідно з'ясувати, чи є суттєва різниця між партіями виробів за деяким показником якості, тобто, треба перевірити вплив на якість товару одного фактору – партії виробів.

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}$$

Рис. 7.1. Матриця спостережень.

Нехай є m партій виробів. З кожної партії виберемо відповідно n_i, $1 \le i \le m$, виробів (для простоти вважаємо, що $n_1 = \cdots = n_m = n$). Значення показника якості цих виробів представимо у вигляді матриці спостережень (рис. 7.1), де в першому рядку розташована якість виробів з першої партії і т. д.

Необхідно перевірити суттєвість впливу партій виробів з їхню якість. Будемо вважати, що елементи рядків матриці спостережень X — це чисельні значення (реалізації) випадкових величин X_i, $1 \le i \le m$, що виражають якість виробів в i-ой партії і мають нормальний закон розподілу з математичними очікуваннями відповідно a_i, $1 \le i \le m$ і однаковими дисперсіями σ^2. Тоді наша задача зводиться до перевірки методами дисперсійного аналізу нульової гіпотези H_0: $a_1 = \cdots = a_m$.

Спочатку обчислимо середні значення в групах (рядках матриці X):

$$\mu_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij} \ , \qquad 1 \le i \le m \ , \tag{7.12}$$

і визначимо загальне середнє значення:

$$\mu = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij} = \frac{1}{m}\sum_{j=1}^{n} \mu_i \quad . \tag{7.13}$$

Обчислимо суму квадратів відхилень спостережень x_{ij} від середнього значення μ і розділимо її на доданки:

$$Q = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-\mu)^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}(\mu_i-\mu)^2 + \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-\mu_i)^2 + 2\sum_{i=1}^{m}\sum_{j=1}^{n}(\mu_i-\mu)(x_{ij}-\mu_i) \ ,$$

або $Q = Q_1 + Q_2 + Q_3$. Останній доданок Q_3 дорівнює нулю:

$$Q_3 = 2\sum_{i=1}^{m}\sum_{j=1}^{n}(\mu_i-\mu)(x_{ij}-\mu_i) = 2\sum_{i=1}^{m}(\mu_i-\mu)\sum_{j=1}^{n}(x_{ij}-\mu_i) = 0 \ ,$$

так як сума відхилень значень в рядку матриці X від її середньої дорівнює нулю в силу формули (7.12). Зауважимо, що перший доданок можна записати у вигляді:

$$Q_1 = \sum_{i=1}^{m}\sum_{j=1}^{n}(\mu_i-\mu)^2 = \sum_{j=1}^{n}\sum_{i=1}^{m}(\mu_i-\mu)^2 = n\sum_{i=1}^{m}(\mu_i-\mu)^2 \ , \tag{7.14}$$

оскільки внутрішня сума не залежить від індексу j.

В результаті отримуємо рівність:

$$Q = Q_1 + Q_2 \ , \tag{7.15}$$

де Q – *загальна або повна* сума квадратів відхилень від загальної середньої, Q_1 – сума квадратів відхилень групових середніх від загальної середньої, яка відображає *між-групову* (*факторну*) суму квадратів відхилень, Q_2 – сума квадратів відхилень спостережень від групових середніх або *внутрішньо-групова* (*остаточна*) сума квадратів відхилень. Якщо поділити рівність (7.15) на число спостережень, то ми отримуємо рівність (7.7). Тому, як вже зазначалося, в рівності (7.15) заключено вся суть методу аналізу дисперсій.

У дисперсійному аналізі аналізуються не самі суми квадратів відхилень Q_1 і Q_2, а так звані *середні квадрати* D_1 і D_2, які є незміщеними оцінками відповідних дисперсій і які виходять діленням сум квадратів відхилень Q_1 і Q_2 на відповідне число ступенів свободи. Нагадаємо, що число ступенів свободи визначається як загальна кількість спостережень мінус число рівнянь, що їх зв'язують. Тому для середнього квадрата D_1, що є *незміщеною оцінкою між-групової дисперсії,* число ступенів свободи дорівнює $k = m - 1$, так як при його розрахунку використовуються m групових середніх, пов'язаних між собою одним рівнянням (7.13). Для середнього квадрату D_2, що є *незміщеною оцінкою внутрішньо-групової дисперсії* число ступенів свободи дорівнює $l = mn - m$, оскільки всі mn спостережень пов’язані з їх груповими середніми m рівняннями (7.12). Тому маємо:

$$D_1 = \frac{Q_1}{m-1}, \quad D_2 = \frac{Q_2}{mn-m} \ . \tag{7.16}$$

На основі оцінок D_1 і D_2 обчислюється *F-критерій*:

$$F = \frac{D_1}{D_2} \ . \tag{7.17}$$

Гіпотеза H_0 *відхиляється* , якщо обчислене значення статистики F більше критичного:

$$F > F_{\gamma,k,l} \ , \tag{7.18}$$

де $F_{\gamma,k,l}$ – квантиль розподілу Фішера, визначена з надійністю γ (значимістю $\alpha = 1-\gamma$) при числі ступенів свободи $k = m-1$ і $l = mn-m$. Якщо $F \le F_{\alpha,k,l}$, то гіпотеза H_0 приймається. У нашої задачі спростування гіпотези H_0 означає наявність істотних відмінностей в якості виробів в різних партіях на заданому рівні значимості.

Приклад 7.3. Є чотири партії сировини для текстильної промисловості. З кожної партії відібрано по п'ять зразків та проведено випробування на визначення величини розривного навантаження. Результати випробувань наведено в таблиці 7.2. Необхідно визначити з надійністю 0,95 чи суттєво впливає сировина на величину розривного навантаження.

Таблиця 7.2.

Номер	Навантаження				
1	200	140	170	145	165
2	190	150	210	150	150
3	230	190	200	190	200
4	150	170	150	170	180

Розв'язування. Тут ми маємо $m = 4$, $n = 5$. Обчислимо спочатку середні значення в рядках за формулою (7.12), тоді отримуємо $\mu_1 = 164$, $\mu_2 = 170$, $\mu_3 = 202$, $\mu_4 = 164$. Далі визначимо загальне середнє значення за формулою (7.13) $\mu = 175$. Обчислимо суми квадратів Q_1 і Q_2 за формулами на попередньому аркуші (за допомогою Excel), тоді отримуємо $Q_1 = 4980$ і $Q_2 = 7270$. Звідси дістаємо за формулою (7.16) $D_1 = 4980/3 = 1660$ і $D_2 = 7270/16 \approx 454{,}38$. Статистика F, яку обчислюємо за формулою (7.17), дорівнює $F = 1660/454{,}38 \approx 3{,}65$. Критичне значення квантилі $F_{0,95,3,16}$ розподілу Фішера знаходимо за таблицями або за допомогою програми Excel $F_{0,95,3,16} = 3{,}24$. Оскільки $3{,}65 > 3{,}24$, то гіпотеза H_0 відхиляється. Таким чином, різниця між партіями сировини суттєво впливає на величину розривного навантаження. Це можна також спостерігати в таблиці 7.2, оскільки середнє $\mu_3 = 202$ суттєво перевищує інші середні. □

8. Кластеризація даних

В задачі *кластеризації* треба за вибіркою *багатовимірних даних* $X = (x_1, x_2, \cdots, x_n)$ розділити їх на непересічні компактні групи (які називаються *кластерами*), так щоб кожен кластер складався з «близьких» в якомусь сенсі даних, а дані у різних кластерах суттєво відрізнялися друг від друга. Таким чином, задачу кластеризації можна розглядати як розбиття даних на *однорідні сукупності даних*, і тому *задача кластеризації передує* всім іншим статистичним методам обробки даних таким, як кореляція і регресія.

8.1. Нормування і відстані

Розбиття даних на класи може залежати від вибору одиниць виміру (масштабів шкал). Наведемо невеликий приклад.

Приклад 8.1. Студенти групи вирішили порівняти свою вагу (x) та зріст (y). За цими даними склали діаграму розсіювання, яка показана на рис. 8.1а, де вага вимірювалася в кілограмах, а зріст в сантиметрах. Тут видно, що дівчата (A) досить чітко відокремлюються від юнаків (B). На рис. 8.1б шкала на осі абсцис стиснута вдвічі. При цьому кластери змінилися і тут можна говорити про високих юнаків (D) і всіх інших студентів (C).

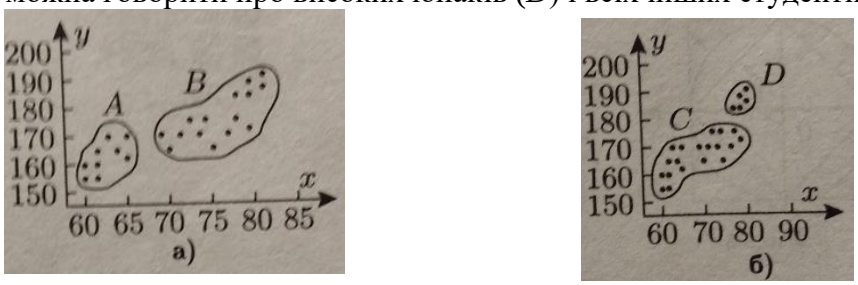


Рис. 8.1. Зміна шкали.

Таким чином, попередньо необхідно нормувати показники та перетворити їх на безрозмірні величини. Нагадаємо основні методи нормування числової вибірки $z = (z_1, z_2, \cdots, z_n)$:

(N1). Лінійне перетворення на інтервал $[0,1]$:

$$z_i' = \frac{z_i - z_{min}}{z_{max} - z_{min}} \ , \qquad 1 \le i \le n \ , \qquad (8.1)$$

(N2). Нелінійне перетворення на інтервал:

$$z_i' = \frac{z_i - \bar{z}}{\sigma(z)} \ , \qquad 1 \le i \le n \ , \qquad (8.2)$$

де $\bar{z}$ – вибіркове середнє, $\sigma(z)$ – вибіркове середнє квадратичне відхилення.

Крім нормування, вирішальний вплив на результат кластеризації надає *вибір міри близькості* між векторами. Найбільш відомими *метриками* $d(x, y)$ між m-вимірними векторами $x = (x_1, \cdots, x_m)$ і $y = (y_1, \cdots, y_m)$ є такі відстані:

(D1). *Манхетеньська відстань* (метрика міста):

$$d(x, y) = \sum_{i=1}^{m} |x_i - y_i| \ . \qquad (8.3)$$

Ця метрика при $m = 2$ дорівнює найкоротший відстані між точками x і y прямокутної решітки, яку складають авеню і вулиці в центральному районі Нью-Йорка. Для бінарних векторів x і y ця відстань дорівнює кількості їх нерівних координат (метрика Хеминга).

(D2). *Евклідова метрика*:

$$d(x, y) = \left(\sum_{i=1}^{m} (x_i - y_i)^2 \right)^{1/2} \ . \qquad (8.4)$$

(D3). *Метрика Чебишева*:

$$d(x, y) = \max_{1 \le i \le m} |x_i - y_i| \ . \qquad (8.5)$$

8.2. Задача кластеризації

Розв'язання задачі кластеризації принципово неоднозначне, і тому є кілька причин. По-перше, не існує найкращого критерію якості розбиття даних. По-друге, число кластерів, як правило, невідомо заздалегідь і встановлюється в ході кластеризації відповідно до деякого суб'єктивного критерію. По-третє, результат кластеризації істотно залежить від метрики, вибір якої, як правило, також суб'єктивний та визначається дослідником.

Цілі кластеризації також можуть бути різними залежно від особливостей конкретної прикладної задачі:

- Спростити подальшу обробку даних, розбити множину даних X на групи подібних (однорідних) векторів, щоб працювати з кожною групою окремо (задачі кореляції, регресії, прогнозування).
- Скоротити обсяг даних, що зберігаються, залишивши по одному представнику від кожного кластера (задача стиснення даних).
- Виділити нетипові об'єкти, які не підходять до жодного кластера (задача одно-класової класифікації).
- Побудувати ієрархію об'єктів (задача таксономії).

У першому випадку кількість кластерів намагаються зробити менше. У другому випадку важливіше забезпечити високий ступінь подібності об'єктів усередині кожного кластера, а кластерів може бути скільки завгодно. У третьому випадку найбільший інтерес представляють окремі об'єкти (вектори), які не вписуються в жодний із кластерів.

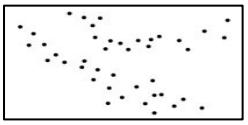
У всіх цих випадках може застосовуватися ієрархічна кластеризація, коли великі кластери дробляться на дрібніші, ті в свою чергу дробляться ще дрібніше, і т. д. Такі задачі називаються *задачами таксономії* (taxonomy). Результатом таксономії є не просте розбиття множини об'єктів на кластери, а деревовидна ієрархічна структура. Об'єкт характеризується перерахуванням всіх кластерів, яким він належить, від великого до дрібного.

Класичним прикладом таксономії з урахуванням подібності є систематизація живих істот, запропонованих Карлом Ліннеєм в середині XVIII століття. У сучасної біології ця ієрархія має близько 30 рівнів, 7 з них вважаються основними: царство, тип, клас, загін, сімейство, рід, вид. Таксономії будуються у багатьох галузях знання, щоб упорядкувати інформацію про велику кількість об'єктів.

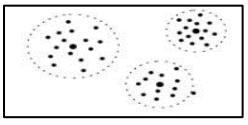
Типи кластерних структур. Складемо реєстр різних типів кластерних структур, які можуть виникати на практиці.

(С1) Кластери типу *ядра* (*згущення*), коли внутрішньо-кластерні відстані, як правило, менше між-кластерних.

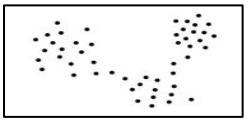
(С2) Кластери типу *стрічки* (*слабке згущення*). Для будь-якого об'єкта знайдеться близький до нього об'єкт того ж кластера, в той же час існують об'єкти одного кластери, які не є близькими.



(С3) Кластери *з центром*. У кожному кластері знайдеться об'єкт (вектор), такий, що майже всі вектори кластера лежать усередині кулі з центром у цьому векторі.



(С4) Кластери можуть з'єднуватися перемичками, що затрудняє роботу багатьох алгоритмів кластеризації.



8.3. Евристичні алгоритми кластеризації

Різні алгоритми кластеризації можуть бути більш менш успішними на цих типах кластерів. Прості алгоритми, як правило, вузько спеціалізовані і дають адекватні результати тільки для одного або двох типах. Однак створення алгоритму, що успішно працює у всіх ситуаціях без винятку, видається важкою і навряд чи розв'язною задачею.

Важливий клас алгоритмів кластеризації заснований на поданні вибірки $X = (x_1, x_2, \cdots, x_n)$ у вигляді неорієнтованого графу $G(X)$. Вершини графа v_i, $1 \le i \le n$, відповідають векторам вибірки x_i, $1 \le i \le n$, а ребру (v_i, v_j) відповідає відстань $d(x_i, x_j)$ між парою векторів x_i і x_j (зважений граф). Достоїнством графічних алгоритмів кластеризації є наочність, відносна простота реалізації, можливість вносити різні вдосконалення, спираючись на прості геометричні міркування.

8.3.1. Алгоритм виділення зв'язних компонентів графа

Задається параметр $\tau > 0$ (поріг) і у графі $G = (V, E)$ залишаються тільки ребра (v_i, v_j), для яких $d(x_i, x_j) \le \tau$, тобто $E = \{(v_i, v_j) : d(x_i, x_j) \le \tau\}$. Таким чином, з'єднаними залишаються тільки близькі пари векторів. Ідея алгоритму полягає у тому, щоб підібрати таке значення порогу τ, для якого граф розділиться на кілька зв'язних компонентів $V_1, \cdots, V_k$, які і будуть кластерами.

Зв'язним компонентом графа $G = (V, E)$ називається підмножина його вершин W, в якої будь-які дві її вершини можна з'єднати шляхом, що повністю лежить у цієї підмножині. Для пошуку зв'язних компонентів можна використовувати стандартні алгоритми пошуку в ширину або пошуку в глибину. Але зв'язані компоненти графа можна знайти також через його відношення досяжності $T_G = \{(v_i, v_j) : v_i \to v_j\}$, коли із вершини v_i існує шлях у вершину v_j в графі G. Відношення досяжності T_G для неорієнтованого графу буде відношенням еквівалентності, а його класи і будуть компонентами зв'язності графа G.

Обчислення відношення досяжності T_G вершин у графі називають також задачею про транзитивне замикання відношення (або двійкової матриці), оскільки відношення досяжності є транзитивним замиканням відношення суміжності (суміжні вершини на діаграмі графа). Для вирішення цієї задачі існує елегантний алгоритм Воршела (Warshall), що складається по суті з одного потрійного циклу (один оператор!). Наступна процедура 8.2 написана на Пітоне реалізує цей алгоритм. На вхід алгоритму надходить двійкова симетрична квадратна матриця A розміру $n \times n$, яка представляє матрицю суміжності неорієнтованого графа, тобто $A[i,j] = 1$, якщо $d(x_i, x_j) \le \tau$ і $A[i,j] = 0$ в іншому випадку. Нагадаємо, що в мові Пітон квадратна матриця представляється списком списків рівної довжини. Вершини графу нумеруються числами $0 \le v \le n-1$. В процедурі 8.2 використовуються логічні операції кон'юнкції (and) і диз'юнкції (or).

Процедура 8.2. Алгоритм Воршела обчислення транзитивного замикання двійкової матриці A (матриці суміжності автомату).

```
def Warshall(A):
    n = len(A);                                  # число станів автомату
    T = A.copy();                                # копіювання матриці
    V = [i for i in range(n)]                    # множина станів автомату
    for i in V:
        T[i][i] = 1                              # рефлексивне замикання матриці
    for k in range(1, n + 1):                    # основний цикл, номер ітерації
        for i in V:                              # цикл по парам вершин
            for j in V:                          # цикл по парам вершин
                T[i][j] = T[i][j] or (T[i][k - 1] and T[k - 1][j])  # перерахунок шляху
    return T
```

Зауважимо, що алгоритм Воршела має складність порядку $O(n^3)$ бітових операцій. Обґрунтування алгоритму Воршела буде надано на практичних заняттях.

Наступна процедура 8.3 дозволяє отримати список класів відношення еквівалентності за його матрицею. У даному випадку вхідною матрицею є матриця T відношення досяжності графу, яку ми отримали у процедурі 8.2, тому класами еквівалентності будуть зв'язані компоненти графу.

Процедура 8.3. Процедура виділення класів еквівалентності за двійковою матрицею відношення еквівалентності.

```
def E_classes(T):
    n = len(T)                      # Number of elements
    r = 0                           # Number of classes
    CC = []                         # List of classes (components)
    S = []                          # List of elements
    for i in range(n) :             # Main cycle
      if i not in S :               # Check element
        S.append(i)                 # New element
        CC.append([i])              # New class
        for j in range(i + 1, n):
          if T[i][j] == 1 :         # Check equivalent element
            S.append(j)             # Add element to the list
            CC[r].append(j)         # Add element to the class
        r = r + 1                   # Increase the number of classes
    print(CC)                       # Print the list of classes
    return CC
```

Для підбору параметру τ можна побудувати *гістограму розподілу попарних відстаней* $d(x_i, x_j)$. На рис. 8.2 показана така гістограма, де висота прямокутника над інтервалом Δ_i пропорційна кількості відстаней $d(x_i, x_j)$. У задачах з гарною кластерною структурою ця гістограма має два піки: при $d(x_i, x_j) \approx d_{int}$ (типова внутрішньо-класова відстань) і при $d(x_i, x_j) \approx d_{out}$ (типова між-класова відстань). Тоді параметр τ можна задати як точку мінімуму між цими піками $\tau = (d_{int} + d_{out})/2$.

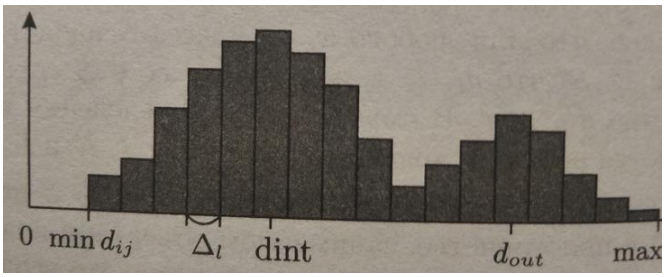


Рис. 8.2. Гістограма розподілу попарних відстаней.

Зазначимо два недоліки цього алгоритму.

- Обмежена застосовність. Алгоритм виділення зв'язкових компонент найбільш підходить для виділення кластерів типу згущень або стрічок. Наявність розрідженого фону або «вузьких перемичок» між кластерами призводить до неадекватної кластеризації.
- Погана керованість числом кластерів. Для застосування кластеризації зручніше задавати не параметр τ, а число кластерів або деякий поріг «чіткості кластеризації». Керувати числом кластерів за допомогою параметра τ досить важко. Доводиться багаторазово вирішувати задачу за різних τ, що негативно впливає на тимчасові витрати.

8.3.2. Алгоритм побудови мінімального кістякового дерева

Нехай $G = (V, E)$ - неорієнтований зв'язний граф, де V - множина вершин, а E - множина ребер. Граф називається зв'язним, якщо будь-які дві його вершини пов'язані деяким шляхом, тобто граф складається із одної компоненти зв'язності. Деревом називається зв'язний граф без циклів. Із теорії графів відомо, що дерево з n вершинами містить $n - 1$ ребро. Остовним (кістяковим) деревом зв'язного графа $G = (V, E)$ називається його підграф (V, T), який є деревом і який містить усі вершини графа.

Припустимо тепер, що на ребрах графа задано вагову функцію $d: E \to R^+$, у нашому випадку це функція відстаней між вершинами $d(v_i, v_j)$, що перетворює його на зважений

граф $G = (V, E, d)$. Вагою кістякового дерева (V, T) зваженого графу G називається сума ваг усіх ребр, що входять до цього дерева.

*Задача MST***.** За заданим зв'язним зваженим неорієнтованим графом $G = (V, E, d)$ потрібно знайти його *мінімальне остовне дерево* (Minimal Spanning Tree), тобто кістякове дерево, що має мінімальну вагу.

Ідея алгоритму кластеризації тут аналогічна попередньої і полягає у тому, щоб побудувати MST, а потім видалити із нього $k-1$ самих довгих ребер, що призведе до створення k компонент зв'язності графа $V_1, \cdots, V_k$, які і будуть кластерами. Причому тут параметр k часто являється наперед заданим параметром алгоритму. Зауважимо, що видалення будь-якого ребра в ациклічному графі збільшує на одиницю число його компонент зв'язності. На рисунку 8.3 показано зважений граф і його мінімальне кістякове дерево.

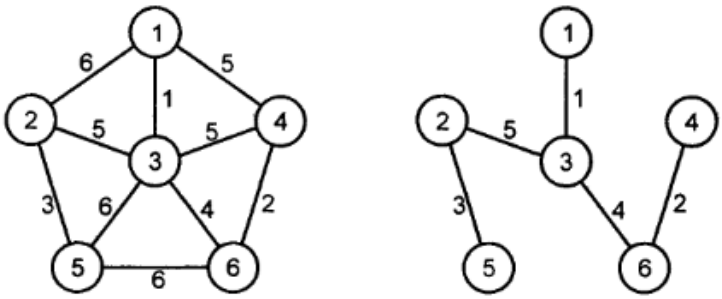


Рис. 8.3. Зважений граф (а) і його кістякове дерево (б).

У задачі MST може бути декілька розв'язків, причому більш, менш очевидно, що завжди існує хоча б один розв'язок, наприклад для цього достатньо перебрати всі під-графи з $n-1$ ребром. Існує два алгоритми побудови MST, а саме, алгоритм Пріма та алгоритм Крускала, але для кластеризації даних використовують в основному алгоритм Пріма як більш простий в реалізації.

Ідея алгоритму Пріма досить проста. Нехай задано зв'язний зважений граф $G = (V, E, d)$, де $V = \{0,1, \cdots, n-1\}$. У нашому випадку це буде повний граф з n вершинами, оскільки між будь-якими вершинами існує відстань. У цьому алгоритмі будується дерево T, яке "росте" із одної вершини до множини V. Спочатку беремо розбиття $U = \{0\}$ і $W = V \backslash \{0\}$ і знаходимо ребро найменшої ваги $\{0, w\}$ де $w \in W$. Якщо таких ребер кілька, то беремо будь-яке з них, наприклад, перше за порядком. Потім це ребро приєднується до дерева T, а вершина w додається до U і видаляється з W. Потім знову знаходимо мінімальне ребро, яке з'єднує U і W і т. д. Цей процес повторюється доти, доки не виконається рівність $U = V$. Алгоритм Пріма ґрунтується на такій властивості мінімальних кістякових дерев, яку ми приведемо без доказу.

Теорема 8.5. Нехай $G = (V, E, d)$ – зв'язний зважений граф, $\{U, W\}$ – розбиття його вершин на дві частини та $e = \{u, w\}$ – ребро мінімальної вартості таке, що $u \in U$ і $w \in W$, тоді в графі G існує мінімальне остовне дерево, що містить ребро e.

Пряма реалізація алгоритму Пріма вимагає перебору всіх ребер із множини $U \times W$, що веде до алгоритму з кубічною складністю $O(n^3)$. Але невелика попередня підготовка дає змогу знизити час роботи алгоритму до $O(n^2)$.

Введемо два списки $\mathrm{Closest}[w]$ і $\mathrm{LowCost}[w]$, $0 \le w \le n-1$. У списку Closest для кожної вершини w з W вказується номер вершини $u \in U$ такий, що ребро $\{u, w\}$ має мінімальну вартість серед усіх ребер графа з множини $U \times \{w\}$. Число $\mathrm{LowCost}[w]$ дорівнює вазі ребра $\{\mathrm{Closest}[w], w\}$ для всіх $w \in W$. На кожному циклу алгоритму переглядається список W і в ньому знаходиться вершина w з мінімальним значенням $\mathrm{LowCost}[w]$. Тоді ребро $\{\mathrm{Closest}[w], w\}$ буде мати мінімальну вартість серед усіх ребер в розбиті $\{U, W\}$. Після цього ребро $\{\mathrm{Closest}[w], w\}$ приєднується до вже побудованого лісу MST, а вершина w видаляється з W (і неявно приєднується до множини U). Після цього змінюються (перераховуються) списки Closest і LowCost.

Процедура 8.5 реалізує алгоритм Прима. У цій процедурі граф G задано симетричною матрицею суміжності $C[i,j]$ розміру $n \times n$, що містить ваги всіх ребер графа. У нашому випадку $C[i,j]$ дорівнює відстані між вершинами v_i і v_j. Діагональні елементи матриці $C[i,i]$ у нашому випадку будуть дорівнювати нулю.

Процедура 8.5. Алгоритм Пріма побудови мінімального кістякового дерева.

```
def Prim(C) :
  n = len(C)                                  # число вершин графа
  MST = []; Tc = 0; U = [0]                   # початкова вершина дерева
  W = [i for i in range(1, n)]                # список відкритих вершин
  Closest = [0 for i in range(n)]             # початкова близька вершина
  LowCost = [C[0][i] for i in range(n)]       # вага близького ребра
  while (len(W) > 0) :                        # основний цикл
    Y = [LowCost[i] for i in W]               # список вартостей ребер
    w = W[Y.index(min(Y))]                    # номер мінімальної вершини
    u = Closest[w]                            # близька вершина к w
    MST.append((u, w))                        # додати ребро до дерева
    Tc = Tc + C[u][w]                         # додати вартість ребра
    W.remove(w)                               # закрити вершину w
    for i in W :                              # зміна вартості
      if C[w][i] < LowCost[i] :               # перевірка вартості
        LowCost[i] = C[w][i]                  # нова вартість
        Closest[i] = w                        # нова близька вершина
  return Tc, MST
```

Цей алгоритм, як і попередній, має обмежену застосовність. Наявність розрідженого фону чи перемичок призводить до неадекватної кластеризації.

9. Показники ефективності діяльності підприємства

9.1. Сутність показників ефективності

Одним із найбільш важливих аспектів діяльності підприємства виступає дослідження *його ефективності*. Цей аналіз є необхідним не тільки керівнику підприємства для оцінки фінансового стану, а й цілому колу інших осіб, які беруть безпосередню участь у господарській практиці:

- керівникам маркетингових відділів, які на основі отриманих даних розробляють стратегію просування товару на ринок;
- інвесторам, яким необхідно прийняти рішення про формування портфелю цінних паперів підприємства;
- аудиторам, яким необхідно перевірити звітність і діяльність підприємства для надання відповідних рекомендацій по веденню бухгалтерського обліку;
- кредиторам, які приймають рішення про видачу кредитів підприємству для впевненості, що їх кошти будуть повернуті з відсотками.

Економічний ефект передбачає будь-який корисний результат, виражений у вартісній оцінці. Зазвичай у якості корисного результату виступають прибуток або економія витрат. Економічний ефект – величина абсолютна, що залежить від масштабів виробництва. *Економічна ефективність* – це співвідношення між результатами господарської діяльності і витратами ресурсів та праці. Таким чином, економічна ефективність – величина відносна, що отримується в результаті зіставлення ефекту з витратами.

Зазвичай аналізуються обидва показники, що характеризують успішність економічної діяльності, оскільки окремо ці показники не можуть дати повної оцінки діяльності підприємства. Наприклад, може бути досягнутий значний економічний ефект при відносно низькій економічній ефективності. І навпаки, виробництво може характеризуватися високим рівнем ефективності при невеликій величині економічного ефекту.

Ефективність діяльності підприємства визначається за допомогою низки економічних показників:

- показники використання праці (робочої сили);
- показники використання основних і оборотних засобів;
- показники рентабельності;
- показники енергоємності;
- показники екологічності. В умовах забруднення навколишнього середовища є одним із важливих показників неефективності промислового підприємства.

9.2. Система показників ефективності господарської діяльності

До основних показників, які використовуються для оцінки *ефективності праці* на підприємстві відноситься виробіток:

$$q = \frac{Q}{T} \ , \tag{9.1}$$

де Q – обсяг виготовленої продукції за одиницю часу T (година, день і т. д.). Виробіток характеризує кількість продукції, виготовленої за одиницю робочого часу (або на одного працівника). Використовується також обернений до показника (9.1) показник, який називається трудомісткістю:

$$t = \frac{T}{Q} \ . \tag{9.2}$$

Трудомісткість характеризує витрати робочого часу на виробництво одиниці продукції. Цей показник має низку переваг перед показником виробітку, оскільки встановлює пряму залежність між обсягом виробництва і трудовими витратами. Залежно від складу витрат, що включаються в трудомісткість продукції, розрізняють наступні її види:

- технологічна трудомісткість (витрати праці основних робітників);
- трудомісткість обслуговування виробництва (витрати праці допоміжних робітників);

- виробнича трудомісткість (витрати праці основних і допоміжних робітників);
- трудомісткість управління виробництвом (витрати праці керівників, фахівців і службовців);
- повна трудомісткість (витрати праці всього промислово-виробничого персоналу).

Наступною групою є показники оцінки ефективності використання *основних засобів*. Основні засоби є частиною засобів виробництва підприємства, що беруть участь у процесі виробництва протягом тривалого часу, зберігаючи при цьому свою форму, їх вартість переноситься на продукцію поступово, по мірі використання. Ефективність використання основних засобів підприємства характеризується показниками завантаження, які відображають ступінь використання виробничої потужності і обсягами продукції, що випускається. У фінансовому аналізі найбільш часто застосовуються показники оцінки ефективності використання машин і устаткування, як найбільш активної частини основних засобів підприємства. Одним із основних показників тут є загальна фондовіддача:

$$\Phi_{з} = \frac{Q}{\text{ОВФ}} \ , \tag{9.3}$$

де Q – загальний обсяг виготовленої (або реалізованої) продукції протягом року, а ОВФ – середньорічна вартість основних виробничих фондів. Цей показник характеризує скільки гривень виготовленої (або реалізованої) продукції припадає в середньому на одну гривню основних виробничих засобів. Він не має загальноприйнятого «нормального» значення, оскільки сильно залежить від галузевих особливостей. Використовується також обернений до показника (9.3) показник, який називається фондоємністю продукції:

$$\Phi_{є} = \frac{\text{ОВФ}}{Q} \ . \tag{9.4}$$

Цей показник характеризує скільки гривень основних виробничих засобів припадає в середньому на одну гривню виготовленої (або реалізованої) продукції. Фондоємність показує ступінь використання основного капіталу у виробництві продукції. Фондоозброєність праці обчислюється за формулою:

$$\Phi_{о} = \frac{\text{ОВФ}}{N} \ , \tag{9.5}$$

де N - – середня чисельність робітників (працівників). Цей показник характеризує вартість основних засобів, що припадає в середньому на одного робітника (працівника) підприємства, тобто відображає ступінь забезпеченості персоналу основними засобами виробництва. Зростання фондоозброєності праці відображає заміщення живої праці технікою, ліквідацію ручних процесів, підвищення ступеня механізації та комплексної автоматизації виробництва.

Наступна група це *узагальнюючі показники*, які характеризують ефективність господарської діяльності в цілому. До цієї групи відносять такі показники:

$$h = \frac{Qz}{Qp} \ , \tag{9.6}$$

де Qz - витрати на виробництво і реалізацію продукції, Qp - – обсяг (вартість) реалізованої продукції. Іншими словами, цей показник характеризує скільки гривень витрат припадає у середньому на одну гривню реалізованої продукції. Показник *рентабельності продукції* обчислюється за формулою:

$$R_{\text{п}} = \frac{\text{П}}{Qz} \ , \tag{9.7}$$

де П – прибуток від реалізованої продукції, Qz - витрати на виробництво і реалізацію продукції. Цей показник характеризує скільки гривень прибутку отримує в середньому підприємство з кожної гривні, витраченої на виробництво і реалізацію продукції. Його розраховують по підприємству в цілому і по окремих підрозділах чи видах продукції. Негативна динаміка показника свідчить про необхідність перегляду цін або посилення контролю за витратами на виробництво та реалізацію продукції.

Рентабельність всього виробництва (загальна) обчислюється за формулою:

$$R_{\text{п}} = \frac{\text{БПр}}{\text{ОВФ} + \text{ОВЗ}} \ , \tag{9.8}$$

де БПр - балансовий прибуток (за період часу), ОВФ – середня вартість основних виробничих фондів, ОВЗ - – середня вартість оборотних виробничих засобів. Цей показник характеризує прибутковість виробничої діяльності за певний період часу. Рентабельність виробництва співставляє величину отриманого прибутку і розміру коштів, які дозволили його отримати, показує суму прибутку в розрахунку на одну гривню витрачених виробничих засобів. Чим менше засобів використано для отримання певної суми прибутку, тим вище рентабельність виробництва, а отже, вище ефективність діяльності підприємства.

Рентабельність продажів (загальна) обчислюється за формулою:

$$R_{\text{пр}} = \frac{\text{П}}{Q} \ , \tag{9.9}$$

де П – прибуток від реалізованої продукції, Q – загальний обсяг реалізованої продукції (за певний період часу). Цей показник характеризує суму прибутку, яку отримує підприємство з однієї гривні реалізованої продукції. Він дозволяє охарактеризувати найголовніше для підприємства – реалізацію основної продукції. Знаючи рентабельність продажів, підприємство може контролювати цінову політику і витрати. Варто зауважити, що різні підприємства виробляють товари за допомогою різних стратегій і технік, що викликає відмінність рівнів рентабельності.

Показники рентабельності є найважливішими показниками ефективності діяльності підприємства. Будучи загальними показниками, вони найбільш повно і всебічно характеризують ефективність його діяльності в цілому.