Міністерство освіти і науки України
Національний університет водного
господарства та природокористування

П.М. Грицюк, О.П. Остапчук

АНАЛІЗ ДАНИХ

Навчальний посібник

Рівне 2008

УДК 519.23 : 681.3 (075.8)
ББК 22.172.517.8 : 32.973 я7
Г 92

Затверджено вченою радою Національного університету водного господарства та природокористування (Протокол № 3 від 28 березня 2008 р.)

Рецензенти:
Джунь Й.В., д-р. ф.-м. наук, професор, Міжнародний економіко-гуманітарний університет ім.акад. С.Дем'янчука
Вітлінський В.В., д-р. ек. наук, професор, Київський національний економічний університет ім. В.Гетьмана

Грицюк П.М., Остапчук О.П.
Г 92 Аналіз даних: Навчальний посібник.– Рівне: НУВГП, 2008. – 218 с.

Навчальний посібник містить основні відомості з сучасних методів аналізу даних, зокрема: методів побудови описової статистики, перевірки гіпотез, дисперсійного, кореляційного, факторного і регресійного аналізу, методів класифікації даних. Посібник призначено для студентів напряму 6.040301 – "Прикладна математика". Його також можуть використовувати студенти інших напрямів підготовки при вивченні методів статистичного аналізу даних та їх застосуванні у прикладних дослідженнях.

УДК 519.23 : 681.3 (075.8)
ББК 22.172.517.8:32.973 я7

ЗМІСТ

1. Основні поняття та основні задачі аналізу даних 5
1.1. Мета і завдання дисципліни. Етапи аналізу даних 5
1.2. Класифікація ознак за шкалами вимірювання 6
1.3. Основні методи аналізу даних 9
1.4. Генеральна сукупність і вибірка 13
1.5. Варіаційна статистика 14
1.6. Описова статистика 22
1.7. Завдання до розділу 1 31
2. Перевірка статистичних гіпотез 41
2.1. Параметричні тести 44
2.2. Множинні параметричні порівняння 47
2.3. Непараметричні тести 47
2.4. Визначення моделей розподілу емпіричних даних 55
2.5. Завдання до розділу 2 58
3. Дисперсійний аналіз 62
3.1. Однофакторний аналіз 62
3.2. Двофакторний аналіз 73
3.3. Функції розподілу, які найчастіше використовують при розрахунку критеріїв 81
3.4. Завдання до розділу 3 87
4. Кореляційний аналіз 95
Завдання до розділу 4 121
5. Регресійний аналіз 128
5.1. Загальна характеристика методів та задач регресійного аналізу

5.2. Парна лінійна регресія 128
5.3. Поліноміальні моделі 136
5.4. Множинна лінійна регресія 144
5.5. Завдання до розділу 5 148
6. Факторний аналіз 151
6.1. Метод головних факторів 154
6.2. Метод максимуму правдоподібності 157
6.3. Центроїдний метод 166
6.4. Додаткові зауваження 167
6.5. Завдання до розділу 6 168
7. Задачі та методи класифікації даних 169
7.1. Кластерний аналіз 171
7.2. Класифікація з навчанням 175
7.3. Завдання до розділу 7 193
Додаток 205
Література 207
216

1. ОСНОВНІ ПОНЯТТЯ ТА ОСНОВНІ ЗАДАЧІ АНАЛІЗУ ДАНИХ

1.1. Мета і завдання дисципліни. Етапи аналізу даних

Аналіз даних – це дисципліна, що розробляє математичні методи збору, систематизації і комп'ютерної обробки статистичних даних з метою їх зручної інтерпретації і отримання наукових та практичних висновків. Аналіз даних – це математична обробка експериментальних даних з використанням статистичних методів за допомогою комп'ютера. Аналіз даних використовує методи математичної статистики.

Завданням аналізу даних є: провести класифікацію даних, знайти закономірності і залежності між змінними.

Мета курсу - створити засоби, які дозволяють охопити зміст таблиць даних значного обсягу через їх представлення (бажано візуальні), що будуть добре зрозумілі для користувача.

Галузі застосування – економіка, маркетинг, промисловість, екологія, медицина, фармакологія та ін.

Етапи аналізу даних.

Основні етапи статистичного аналізу даних:

- Початковий аналіз досліджуваної системи.
- Складання плану для збору вихідної інформації.
- Збір початкових даних та їх підготовка для введення в комп'ютер.
- Попередня обробка даних.

- Вибір основних методів і алгоритмів обробки даних. Складання детального плану чисельного аналізу зібраних даних.
- Реалізація чисельного аналізу даних за допомогою комп'ютера.
- Проведення підсумків досліджень.

Необхідною умовою ефективного застосування математико-статистичних методів аналізу даних є попередній якісний змістовний аналіз об'єктів і процесів. Саме якісний аналіз визначає постановку завдання, відокремлює предмет дослідження, визначає засоби дослідження, зокрема адекватні завданням кількісні методи, використання яких поглиблює, робить конкретнішим наше знання. Успішність застосування будь-якого методу аналізу даних залежить від відповідності аналізованих даних його вихідним припущенням. Методи, придатні для одного типу даних, можуть приводити до серйозних помилок при їх використанні для даних інших типів.

1.2. Класифікація ознак за шкалами вимірювання

Кількісні методи можуть бути застосовані в дослідженні лише після того, як емпіричні дані переведені на мову чисел. Передумовою і початком застосування кількісних методів дослідження є вимір. Під виміром ми розуміємо процедуру, внаслідок якої встановлюється відповідність між властивостями об'єкта і властивостями відповідних їм чисел. Набір властивостей об'єкта і чисел, що відповідають їм, називають шкалою. Найчастіше використовують три типи шкал: номінальні, порядкові і кількісні (метричні). Першим етапом аналізу будь-яких даних є визначення їх типу, тобто класифікація за відповідною шкалою.

- Номінальні ознаки (ознаки з невпорядкованими станами, класифікаційні ознаки) − це дані, що вимірюються в номінальній шкалі (класифікаційній, шкалі найменувань). Найменування класів можуть бути виражені за допомогою чисел, але ці числа можуть використовуватися лише для відповіді на питання: належать два об'єкти до одного класу чи ні. З погляду автоматизації аналізу даних і застосування стандартних алгоритмів доцільно обирати такі позначення класів: 0, 1, 2, …
- Порядкові ознаки (ознаки з упорядкованими станами, ординальні ознаки) − це дані, що вимірюються в порядкових шкалах. Ці дані можуть порівнюватися між собою у певному відношенні: "більше − менше", "легше − важче" тощо. Числа, якими позначають класи, можуть застосовуватися і для порівняння ступеня виразності класифікаційної ознаки, але відстані між класами при цьому будуть не визначені.
- Кількісні (числові, варіаційні) ознаки − це ознаки, які вимірюють у кількісних (інтервальних, відносних, циклічних та абсолютних) шкалах вимірювань. Інтервальні змінні дозволяють не тільки впорядкувати вимірювання, але й чисельно їх виразити та встановити різницю між ними. Відносні змінні мають властивості інтервальних змінних. Крім того вони мають визначену точку абсолютного нуля (значення, від якого ведеться відлік інших значень).

Дані, отримані у шкалах вищих рангів, можуть приводитися до шкал нижчих рангів. Наприклад, дані, що виміряні у шкалі відношень, можна привести до інтервальної шкали. Такі перетворення називають

зниженням шкали. Необхідність у них зазвичай виникає при обробці даних, що виміряні у шкалах різного типу. Зворотну операцію − перетворення даних, що виміряні у нижчих шкалах, до вищих − вважають некоректною. Зниження шкали веде до втрати частини наявної інформації про досліджувані ознаки. Важливими типами класифікації є поділ ознак за дискретністю або неперервністю теоретичної функції розподілу, законом розподілу тощо.

Дії, що можуть виконуватися з числовими характеристиками даних, залежать від шкали вимірювань. В узагальненому вигляді характеристики основних типів даних наведено в табл. 1.1.

Таблиця 1.1

Характеристики основних типів даних

Шкала вимірювань	Визначальні відношення	Еквівалентні перетворення	Допустимі операції над даними	
			(первинна обробка)	(вторинна обробка)
Номінальна	Еквівалентність	Перестановки найменувань	Обчислення символу Кронекера δ_{ij}	Обчислення відносних частот та операції над ними
Порядкова	Еквівалентність, перевага	Монотонні (такі, що не змінюють порядку)	Обчислення δ_{ij} та рангів R_i	Обчислення відносних частот та квантилів, операції над ними
Інтервальна	Еквівалентність, перевага, збереження відношення інтервалів	Лінійне перетворення $y = ax + b$, $a > 0, b \in R$	Обчислення δ_{ij}, рангів R_i, та інтервалів (різниць між даними)	Арифметичні дії над інтервалами
Циклічна	Еквівалентність, перевага, збере-	Зсув $y = a + nb$,	Обчислення δ_{ij}, рангів	Арифметичні дії над інтер-

	ження відношення інтервалів, періодичність	$b = const$, $n = 0,1,2,...$	R_i, та інтервалів (різниць між даними)	валами
Відношень	Еквівалентність, перевага, збереження відношення інтервалів, збереження відношення двох значень	Розтяг $y = ax, a > 0$	Усі арифметичні операції	Будь-яка потрібна обробка
Абсолютна	Еквівалентність, перевага, збереження відношення інтервалів, збереження відношення двох значень, абсолютна й безрозмірна одиниця, абсолютний нуль	Не існує (шкала є унікальною)	Усі арифметичні операції, використання як показника степеня, основи та аргумента логарифма	Будь-яка потрібна обробка

1.3. Основні методи аналізу даних

У більшості задач аналізу даних ми маємо справу з величинами, які є випадковими. **Випадковою** називають величину, яка в результаті певного випробування приймає одне значення, яке наперед невідоме і залежить від випадкових причин. Найчастіше доводиться мати справу з числовими випадковими величинами. Числова випадкова величина може бути дискретною і неперервною.

Дискретною називають числову випадкову величину, яка приймає окремі, ізольовані можливі значення з певними імовірностями. **Неперервною** називають числову випадкову величину, яка може приймати всі значення з деякого конечного або безконечного проміжку. Якщо в результаті дослідження зареєстровано лише одне число, то відповідна випадкова величина називається одномірною або

скалярною. Якщо ж результатом кожного дослідження є декілька чисел, які характеризують різні властивості об'єкта, випадкова величина називається багатомірною або ж **векторною**. Значення векторної випадкової величини можна розділити на незалежні (факторні) змінні і залежні (результуючі) змінні. Залежні змінні можна розглядати як відгук системи (об'єкта) на зовнішні впливи, виражені незалежними змінними. Найпростішою задачею є задача, коли оцінюється вплив одного незалежного фактора на досліджувану змінну.

Основними методами аналізу даних є наступні методи.

- **Кореляційний аналіз**. Його суть полягає у визначенні ступеня тісноти зв'язку між двома і більше випадковими величинами (факторами). Кореляційний аналіз є корисним на попередньому етапі обробки даних.
- **Дисперсійний аналіз**. Це група методів для обробки даних, які залежать від якісних факторів. Ці методи оцінюють суттєвість впливу факторів на результати спостережень. Дисперсійний аналіз використовують на попередньому етапі аналізу даних при перевірці однорідності даних, адекватності моделі регресії і т.п.
- **Регресійний аналіз**. Методи регресійного аналізу дозволяють встановити структуру і параметри моделі, яка зв'язує кількісні показники: результуючу і факторну змінні. Даний вид аналізу дозволяє розв'язати головне завдання експерименту.
- **Факторний аналіз**. Його суть полягає в тому, що фактори, які "лежать на поверхні" у випадку сильної кореляції між собою можуть бути замінені іншими "внутрішніми" факторами, які

важко, або неможливо виміряти, але які визначають поведінку поверхневих факторів і, тим самим, результуючої змінної. Факторний аналіз є джерелом виникнення різноманітних гіпотез, які намагаються перевірити в ході експериментів.

Найчастіше для аналізу даних використовується кореляційний та регресійний аналіз. Основними задачами кореляційного аналізу є :

- Вимірювання ступеня зв'язності двох чи більше явищ (тіснота або сила зв'язку).
- Відбір факторів, які виявляють найбільш істотний вплив на результативну ознаку на підставі вимірювання ступеня зв'язності явищ. Найбільш важливими є фактори, які найсильніше корелюють з досліджуваною величиною. Свідомо змінюючи впливаючі фактори можна добитися бажаних змін в результативній ознаці.
- Виявлення невідомих причинних зв'язків. При цьому слід мати на увазі, що кореляція не є прямим підтвердженям причинного зв'язку між явищами, а лише вказує на ймовірність такого зв'язку. Механізм зв'язків виясняється за допомогою логічно – професійних міркувань.

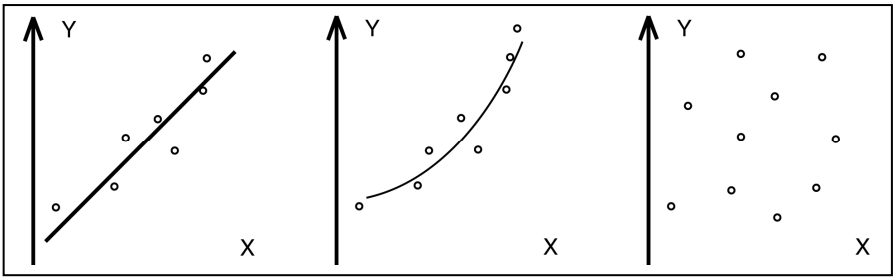

Рис. 1.1а. Залежність спожитої електроенергії від об'єму виробництва

Рис. 1.1б. Залежність між прибутковим податком і зарплатою

Рис. 1.1в. Залежність продуктивності праці від стажу роботи

Основними задачами регресійного аналізу є :

- Встановлення форми залежності. Відрізняють додатну лінійну, додатну нелінійну, від'ємну лінійну та від'ємну нелінійну залежності.
- Визначення виду функції регресії. Як видно з рис.1.1 при кореляційному зв'язку кожному значенню визначаючого фактора відповідає розподіл значень результуючої змінної. Нам важливо не лише вказати загальну тенденцію зміни залежної змінної але і вияснити, як би впливали на залежну змінну головні фактори – причини, якби другорядні фактори не мінялись. Для цього визначають функцію регресії у вигляді математичного рівняння. Процес знаходження функції регресії називається вирівнюванням окремих значень залежної змінної. Побудова регресії і встановлення впливу визначаючих факторів на результуючу змінну – друга задача регресійного аналізу.
- Оцінка невідомих значень залежної змінної. Знаючи функцію регресії можна відтворити значення залежної змінної всередині інтервалу заданих значень впливаючих факторів – задача інтерполяції, або ж оцінити протікання процесу поза заданим інтервалом – задача екстраполяції. Ці задачі розв'язуються шляхом підстановки у рівняння регресії значень впливаючих факторів. Тому регресійний аналіз є корисним інструментом, при плануванні виробництва і прогнозуванні динаміки економічних показників.

1.4. Генеральна сукупність і вибірка

Виконуючи статистичний аналіз даних необхідно відрізняти два поняття – генеральна сукупність і вибіркова сукупність (вибірка).

Сукупність всіх значень деякої ознаки досліджуваних об'єктів утворює **генеральну сукупність** (ГС).

Повне дослідження ГС у більшості випадків практично неможливо, оскільки це вимагає багато часу, матеріальних затрат, а інколи вимагає знищення об'єкта. В таких випадках використовують вибірковий метод, суть якого полягає в тому, що спостерігається не вся ГС, а лише її частина, на основі дослідження якої роблять висновки про властивості ГС. Статистично досліджувана частина ГС називається вибірковою сукупністю або **вибіркою**.

Теоретичною основою використання вибіркового методу є теорема Чебишева, яка стверджує, що з імовірністю близькою до одиниці можна говорити що при достатньо великому обсязі вибірки і обмеженій дисперсії (розсіянні) різниця між вибірковими середніми і середнім ГС буде як завгодно малою.

Одним з актуальних питань, від успішного розв'язання якого залежить достовірність отриманих висновків в результаті статичної обробки, є питання репрезентативності вибірки, тобто повноти і адекватності представлення нею властивостей ГС.

Вибірка називається **репрезентативною**, якщо її імовірнісні характеристики співпадають, або наближаються до відповідних характеристик ГС в межах заданої допустимої похибки.

Для отримання репрезентативної вибірки необхідно:

- забезпечити кожному елементу ГС однакову імовірність потрапити у вибірку;
- відібрати таку кількість елементів ГС при якій забезпечується потрібна точність характеристик ГС.

Визначення оптимального об'єму вибірки є досить складною проблемою. В математичній статистиці доведено, що мінімально необхідна кількість повторної випадкової вибірки визначається виразом

$$n_{\min} = t^2 * \sigma^2 / (\Delta x)^2 . \qquad (1.1)$$

Тут $\Delta\tilde{o}-$ задана допустима абсолютна похибка визначення середнього арифметичного значення; σ - вибіркове середнє квадратичне відхилення; t – коефіцієнт довіри, який залежить від того, з якою довірчою імовірністю потрібно гарантувати результати вибіркового дослідження. Значення σ і Δ вводять в іменованих одиницях, тобто вказують одиниці їх вимірювання, які для цих двох величин мають бути однаковими. Для обчислення t використовується функція статистичного пакету Excel СТЬЮДРАСПОБР $(1-\alpha, n-1)$. Тут α − рівень значущості, n - обсяг вибірки.

1.5. Варіаційна статистика

Варіаційною статистикою називають обчислення числових та функціональних характеристик емпіричного розподілу. Зростаючий (неспадний) числовий ряд даних називається **варіаційним рядом** або **емпіричним розподілом**. Якщо ми маємо справу із дискретним розподілом, то кожному значенню випадкової величини (варіанті) ставиться у відповідність її частота. Якщо досліджувана ознака є непе-

рервною, варіаційний ряд слід розбити на ряд інтервалів – інтервальний варіаційний ряд. Перелік часткових інтервалів і відповідних їм частот називається інтервальним статистичним розподілом.

Для практичної побудови інтервального варіаційного ряду необхідно визначити кількість класів (груп, інтервалів) *k*. Іноді застосовують емпіричні правила, згідно з якими k обирають у межах 10 − 20 або 9 − 15. При цьому для симетричних розподілів рекомендується обирати непарні значення k. Але найчастіше для визначення кількості класів застосовують **правило Стержесса**:

$$k = 1 + 1.447 \ln n \qquad (1.2)$$

де n − кількість елементів сукупності. Необхідно об'єднувати класи, якщо кількість спостережень у них менше ніж чотири або п'ять. Це ускладнює наступну обробку, але є необхідним при застосуванні деяких методів, зокрема критерію χ^2.

Після визначення кількості класів обирають величини інтервалів. Зазвичай їх беруть рівними, тоді $d = \frac{R}{k}$, де d − величина інтервалу групування, R − розкид вибірки. Для нерівних інтервалів і-й класовий інтервал $d_i = x_i - x_{i-1}$, де x_i, x_{i-1} − межі інтервалу. При класифікації за якісною або порядковою ознакою поняття величини і меж класового інтервалу не мають сенсу.

Теоретичною функцією розподілу випадкової величини x називають функцію дійсного аргументу, що задається як $F(x) = P(X \le x)$.

Емпіричною функцією розподілу називають функцію $F_n(x)$, яка кожному значенню x приводить у відповідність частку подій $X \le x$:

$F_n(x) = \frac{n_x}{n}$, де n_x − кількість елементів вибірки, що є меншими за x (нагромаджені, або кумулятивні абсолютні частоти), n − загальна кількість елементів вибірки. Величини $\frac{n_x}{n}$ називають **нагромадженими** (**кумулятивними**) **відносними частотами**, або **інтенсивностями**. Вони можуть подаватися у частках або у відсотках. Теоретична і емпірична функції розподілу виявляють такі властивості:

- якщо $x_1 < x_2$, то $F(x_1) \le F(x_2)$;
- $\lim\limits_{x \to -\infty} F(x) = 0, \quad \lim\limits_{x \to +\infty} F(x) = 1$.

Теоретична функція розподілу, крім того, є неперервною зліва при кожному x.

Корінь X_p рівняння

$$P(X \le X_q) = F(X_q) = q \tag{1.3}$$

називають **вибірковим квантилем порядку *q*** функції розподілу $F(x)$. Вибірковий квантиль порядку 1/2 називають **вибірковою медіаною**; квантилі порядку 1/4, 1/2 та 3/4 − **вибірковими квартилями**; квантилі порядку 10 %, 20 %, …, 90 % − **децилями**, квантилі порядку 1%, 2%, …, 99% − **процентилями**.

Слід зазначити, що для негаусових вибірок квантилі розподілу є більш робастними мірами, ніж середнє та стандартне відхилення. Ве-

личину $X_{0.75} - X_{0.25}$ називають **розмахом розподілу відносно центрального значення**. Часто її подають у нормованому вигляді:

$$\frac{X_{0.75} - X_{0.25}}{X_{0.5}}.$$

Розрізняють дискретні та неперервні розподіли випадкових величин. Дискретним притаманні такі властивості:

- $P(X \le r) = \sum_{i=0}^{r} p_i$;
- $P(r < X \le s) = \sum_{i=r+1}^{r} p_i$;
- $\sum_{i=0}^{\infty} p_i = 1$.

Приклад дискретного статистичного розподілу представлений у таблиці 1.2

Таблиця 1.2

Дискретний статистичний розподіл

x_i	-6	-4	-2	2	4	6
n_i	5	10	15	20	40	10
W_i	0.05	0.1	0.15	0.2	0.4	0.1
F(x)	0.05	0.15	0.30	0.50	0.90	1.00

Для дискретного розподілу графік абсолютних частот доцільно зображувати як точковий або лінійчатий графік, а для неперервного − як стовпчикову діаграму. Графіком $F_n(x)$ для дискретної випадкової величини є східчаста лінія, зображена на рис.1.2.

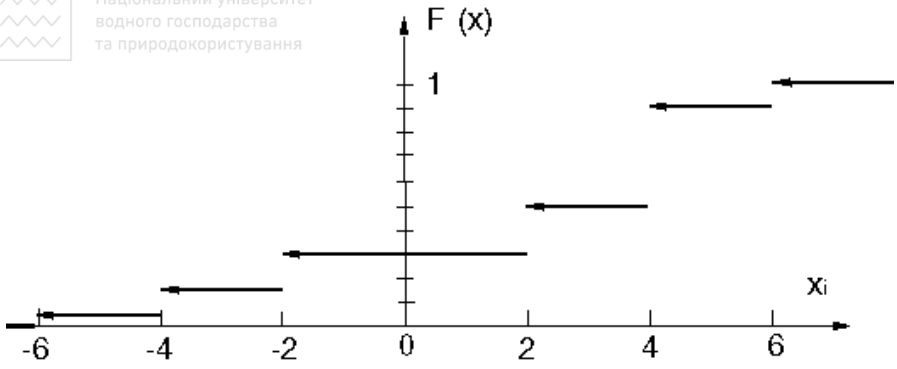

Рис.1.2. Графік емпіричної функції розподілу (кумуляти)

Іншим способом зображення емпіричного закону розподілу дискретної випадкової величини є полігон частот. Ламана лінія, відрізки якої сполучають точки (x_i, n_i) називається **полігоном частот**. Тут x_i - значення варіанти, n_i - відповідна частота. Графік полігона частот має вигляд, зображений на рис.1.3.

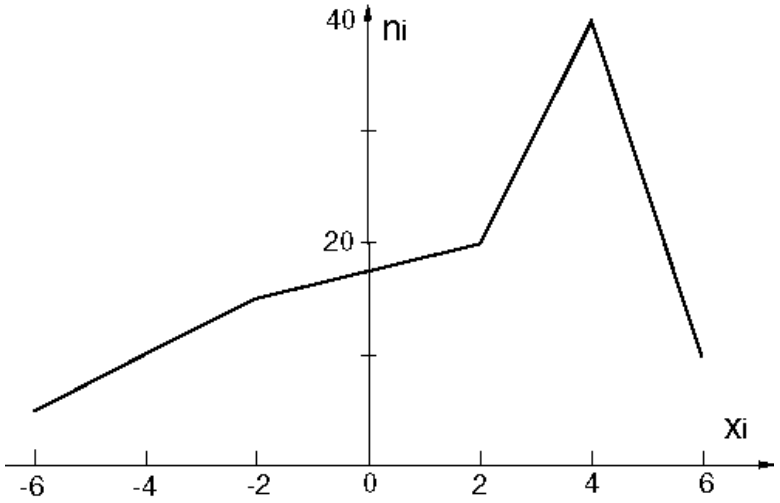

Рис.1.3. Полігон частот

Для неперервних випадкових величин аналогічні властивості можна записати у такій спосіб:

- $P(X \le a) = F(x) = \int_{-\infty}^{a} f(x)dx$;
- $P(a < X \le b) = \int_{a}^{b} f(x)dx = F(b) - F(a)$;
- $\int_{-\infty}^{+\infty} f(x)dx = 1$, де $f(x)$ – щільність імовірності.

Частотами (**абсолютними**, або **груповими**) розподілу ν_i називають кількості елементів вибірки, що попали до і-го класу. **Відносними частотами**, або **частостями,** називають величини $f_i = \frac{\nu_i}{n}$, де n – загальна кількість елементів вибірки. За великих n вони наближаються до імовірностей реалізації відповідних значень параметрів (подій).

Приклад інтервального статистичного розподілу представлений у таблиці 1.3. Графіком $F_n(x)$ (гістограмою частот) для даного розподілу є стовпчикова діаграма, зображена на рис.1.4.

Таблиця 1.3

Інтервальний статистичний розподіл

d = 8	0-8	8-16	16-24	24-32	32-40	40-48
n_i	10	15	20	25	20	10
W_i	0.1	0.15	0.2	0.25	0.2	0.1

При побудові емпіричної функції F(x) за основу береться припущення, що ознака на кожному частинному інтервалі має рівномірну щільність імовірностей. Тому графіком кумуляти для випадку непере-

рвної випадкової величини є ламана лінія, яка зростає і наближається до одиниці (рис.1.5).

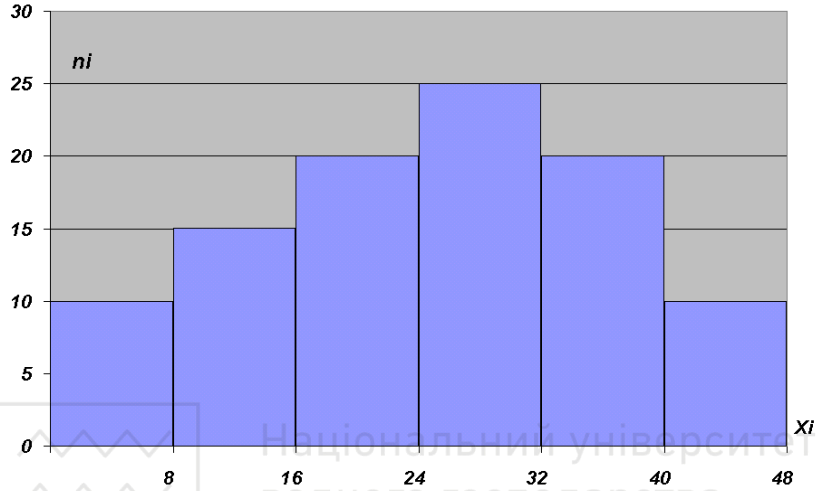

Рис.1.4. Гістограма частот інтервального статистичного розподілу

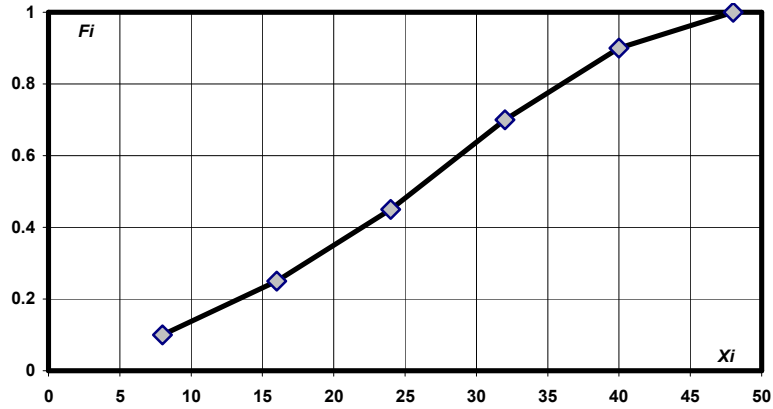

Рис.1.5. Кумулята для випадку неперервного розподілу

У практиці для побудови емпіричних функцій розподілу застосовують два способи.

1. Задають класові інтервали і розносять дані вихідної вибірки за класами. Потім будують масив відносних частот, послідовне сумування елементів якого дає масив функції розподілу. Графічно він зображується як незгасаючий східчастий графік, за віссю абсцис якого відкладені середини класових інтервалів, а за віссю ординат − значення частот. Для використання такого способу необхідно, щоб вихідний обсяг даних був достатньо великим (не менше ніж 100 елементів). Інакше побудована функція може неправильно відображати характер розподілу.

2. Вихідну вибірку впорядковують за зростанням. Потім будують графік, за віссю абсцис якого відкладають значення елементів вибірки, а за віссю ординат − відношення їх номерів до загальної кількості елементів вибірки. Для малих вибірок такий спосіб є єдино придатним для отримання функції розподілу, яку можна використовувати у наступних розрахунках.

Рангом спостереження називають номер, який відповідне спостереження отримає після впорядкування даних за певним правилом. Якщо кілька спостережень мають рівні значення ознаки, за якою здійснюється ранжування, то їм, як правило, присвоюють середні ранги. Але ця процедура не є коректною, оскільки ранги є величинами, що вимірюють у порядковій шкалі, в якій операції сумування та ділення не визначені. Для кількісних даних ранжування знижує вихідну шкалу до порядкової і, відповідно, призводить до певної втрати інформації.

Поправку на об'єднання рангів у загальному випадку обчислюють за формулою:

$$T = \sum_{i=1}^{g} t_i (t_i^2 - 1), \tag{1.4}$$

де g − кількість зв'язок (груп збігів), t_i − кількість збігів в і-й зв'язці.

1.6. Описова статистика

Описова статистика − це набір основних статистичних показників емпіричної вибірки. Стандартні методи їх **розрахунку**, як правило, розроблені виходячи із припущення, що розподіл є нормальним. Якщо він істотно відрізняється від нормального, необхідно використовувати інші методи та формули. Тому процедура аналізу емпіричної вибірки завжди має починатися з перевірки закону розподілу на нормальність.

Центр статистичного розподілу характеризують його середнє значення, медіана та мода. Для дискретних випадкових величин, що виміряні у кількісних шкалах, **середнє арифметичне значення** обчислюють за виразом:

$$\mu = \sum_{i=0}^{\infty} x_i p_i \,, \tag{1.5}$$

де x_i − значення випадкової величини, розподіленої на інтервалі $(-\infty;+\infty)$; p_i - імовірності реалізації x_i. Для аналогічних неперервних величин середнє значення обчислюють за формулою:

$$\mu = \int_{-\infty}^{+\infty} x f(x) dx \tag{1.6}$$

де $f(x)$ − щільність розподілу випадкової величини x.

Вибіркове середнє та його середнє квадратичне відхилення обчислюють за формулами:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i , \quad S_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{1.7}$$

де $x_i (i = 1,2,...,n)$ – значення результатів спостережень, n – обсяг вибірки, σ – вибіркове середнє квадратичне відхилення. Якщо вихідні дані подано як частоти розподілу випадкової величини за інтервалами, вибіркове середнє **обчислюють** за формулою:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} b_i \nu_i \tag{1.8}$$

де $b_i (i = 1,2,...,k)$ – середини інтервалів, ν_i – емпіричні частоти, k – кількість інтервалів.

Застосовують також інші види середніх величин:

- **середнє гармонічне**:

$$h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \tag{1.9}$$

- **середнє геометричне** (**середнє пропорційне**):

$$g = \sqrt[n]{\prod_{i=1}^{n} x_i} \tag{1.10}$$

- **степеневе середнє**:

$$\omega_\alpha = \sqrt[\alpha]{\frac{1}{n}\sum_{i=1}^{n} x_i^\alpha} , \quad \alpha > 0 . \tag{1.11}$$

Слід зазначити, що середнє геометричне та степеневі середні недоцільно застосовувати у випадках, коли ознака x_i може набувати від'ємних значень, оскільки функції, що стоять у правій частині наведених вище виразів, у цьому разі будуть визначеними лише для непарних n та α. При побудові розрахункових алгоритмів необхідно враховувати, що для вибірок великого обсягу добуток, який використовується у формулі середнього геометричного, може перевищити граничне значення, допустиме для певного типу даних. У цьому разі розрахункову формулу необхідно перетворити. Застосування формули (1.9) для обчислення середнього гармонічного у випадку, коли його величина є близькою до нуля, а досліджувана ознака може набувати як додатних так і від'ємних значень, може призводити до накопичення похибки округлення і суттєвого зсуву розрахункового середнього від істинного значення.

Для розглянутих величин за умови α > 1 виконується нерівність $h \le g \le \bar{x} \le \omega_\alpha$.

Довірчий інтервал для математичного сподівання при двобічній гіпотезі за невідомої дисперсії у припущенні нормального закону розподілу визначається співвідношенням:

$$\mu \in \left[\bar{x} - t_{n-1,\alpha} - \frac{S_{\bar{x}}}{\sqrt{n-1}};\ \bar{x} + t_{n-1,\alpha} - \frac{S_{\bar{x}}}{\sqrt{n-1}} \right] \qquad (1.12)$$

де α − рівень значущості, $t_{n-1,\alpha}$ − значення оберненої функції t-розподілу. Якщо M елементам сукупності загальним обсягом N вла-

стива певна якісна ознака (значення 1), а іншим вона не властива (значення 0), то середнє значення сукупності $\mu = \frac{M}{N}$.

Вибіркова медіана є числовою характеристикою неперервно розподіленої випадкової величини, яка визначається умовою, що з імовірністю 0,5 випадкова величина може набувати значення як більші за медіану, так і менші за неї, тобто:

$$\int_{-\infty}^{m} f(x)dx = \int_{m}^{+\infty} f(x)dx = \frac{1}{2} . \qquad (1.13)$$

Для дискретно розподіленої випадкової величини медіаною вважають таке ціле число *m*, що

$$\sum_{i=0}^{m-1} p_i < \frac{1}{2}; \quad \sum_{i=0}^{m} p_i > \frac{1}{2} . \qquad (1.14)$$

Вона може бути визначена як розв'язок рівняння

$$F_n(x) = \frac{1}{2}, \qquad (1.15)$$

де $F_n(x)$ – емпірична функція розподілу випадкової величини. Похибка медіани $s_{me} \approx \sigma\sqrt{\frac{\pi}{2n}}$.

Для інтервального варіаційного ряду вибіркову медіану визначають як варіанту з порядковим номером $\frac{n+1}{2}$ для непарного *n* при нумерації з одиниці. Для парного *n* порядковий номер медіанної варіанти не визначають, а медіану беруть рівною середньому арифметичному двох середніх варіант:

$$me = \begin{cases} X_m,\ n = 2m+1; \\ \frac{1}{2}\left[X_{m-1} + X_m\right],\ n = 2m; \end{cases} \tag{1.16}$$

де X_m − елементи варіаційного ряду, n − його довжина.

Модою називають точку максимуму емпіричної функції розподілу.

Дисперсія характеризує ступінь відхилення елементів сукупності від середнього в одиницях вимірювання відповідної ознаки. Для ознак, що визначаються у кількісних шкалах, дисперсію розраховують за формулами:

- для неперервного випадку: $D = \sigma^2 = \int\limits_{-\infty}^{+\infty} x^2 f(x)dx - \mu^2$ (1.17)
- для дискретного випадку: $D = \sigma^2 = \sum\limits_{i=0}^{\infty} x_i^2 p_i - \mu^2$. (1.18)

Якщо середнє значення сукупності $\bar{x}$ є відомим, то

$$D = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 . \tag{1.19}$$

Якщо середнє значення оцінюється за самою вибіркою, то для розрахунку дисперсії використовують скореговану формулу:

$$D = \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{\infty}\left(x_i - \bar{x}\right)^2 . \tag{1.20}$$

Похибку дисперсії визначають за формулою $S_{\alpha^2} = \sigma^2\sqrt{\frac{2}{n}}$.

Середнє квадратичне (стандартне) відхилення:

$$s = \sigma = \sqrt{D} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \tag{1.21}$$

або, якщо середнє значення $\bar{x}$ відоме з незалежних оцінок:

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \tag{1.22}$$

Середнє квадратичне відхилення для стандартного відхилення для нормально розподілених даних: $S_s = \frac{\sigma}{\sqrt{2n}}$.

Якщо вихідні дані задано у вигляді частот розподілу, дисперсію можна оцінити за формулою:

$$\sigma^2 = \frac{1}{n-1}\left[\sum_{i=1}^{k}\nu_i b_i^2 - \frac{1}{n}\left(\sum_{i=1}^{k}\nu_i b_i\right)^2\right] \tag{1.23}$$

де $b_i\left(i = 1,2,...,k\right)$ − середини класових інтервалів, ν_i − частоти, k − кількість класових інтервалів. Ця формула за певних умов дає завищену оцінку дисперсії. Для її корегування уводять поправку Шеппарда і визначають уточнене значення за формулою:

$$s'^2 = \sigma^2 - \frac{d^2}{12}, \tag{1.24}$$

де d − інтервал між групами, який за рівних відстаней між групами збігається з величиною класового інтервалу. Довірчий інтервал для дисперсії у випадку двобічної гіпотези:

$$\sigma^2 \in \left[\frac{ns^2}{\chi^2_{n-1,\alpha/2}}; \frac{ns^2}{\chi^2_{n-1,1-\alpha/2}}\right] \tag{1.25}$$

де χ^2_{n-1} − значення оберненої функції χ^2-розподілу, α − рівень значущості. Для якісних ознак стандартне відхилення можна обчислити за формулою:

$$\sqrt{p(1-p)} \qquad (1.26)$$

де p − частка відповідної ознаки.

Середнє відхилення також є кількісною характеристикою розсіяння даних. На відміну від середнього квадратичного відхилення, воно є менш чутливим до форми розподілу. Його обчислюють за формулою:

$$d = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|. \qquad (1.27)$$

Середня різниця Джині характеризує розкид даних одне відносно одного й не залежить від будь-якого центрального значення (середнього, медіани тощо). Її розраховують за формулою:

$$g = \frac{1}{n(n-1)}\sum_{\substack{i,j=1 \\ i\neq j}}^{n}\left|x_i - x_j\right|. \qquad (1.28)$$

Величину $C_V = \frac{s}{\bar{x}}$ називають **коефіцієнтом варіації** вибірки. Під розмахом вибірки розуміють величину

$$R = x_{\max} - x_{\min}. \qquad (1.29)$$

Вихідні дані іноді доцільно подавати у **стандартизованому вигляді**: $z = \frac{x - \bar{x}}{s}$.

Асиметрією (**вибірковим коефіцієнтом скісності**) називають міру відхилення симетричного розподілу відносно максимальної ординати. Для симетричного розподілу вона дорівнює нулю. Від'ємні значення відповідають розширенню лівої гілки щільності розподілу, додатні − розширенню правої гілки. Асиметрію A_s розраховують як основний

момент третього порядку, а її стандартне відхилення s_{As} – за формулою $s_{As} = \sqrt{\frac{6}{n+3}}$. Останню формулу часто застосовують як критерій відхилення розподілу від нормальності.

Для кількісної оцінки ступеня відхилення емпіричної кривої розподілу від теоретичної також застосовують **показник ексцесу** (**вибірковий коефіцієнт гостроверхості**). Нормальному розподілу відповідає нульове значення показника ексцесу. Від'ємні значення свідчать про більш пологу, додатні – про більш гостру вершину максимуму розподілу. Показник ексцесу E та його стандартне відхилення s_E визначають за формулами:

$$E = r_4 - 3, \quad s_E = 2s_{As} = 2\sqrt{\frac{6}{n+3}} \qquad (1.30)$$

де r_4 – основний момент четвертого порядку.

Показник точності експерименту є величиною похибки середнього значення у відсотках від самого середнього. Показник та його стандартне відхилення розраховують за формулами:

$$P = \frac{S_{\bar{x}}}{\bar{x}}100\%, \quad s_P = P\sqrt{\frac{1}{2n} + \left(\frac{P}{100}\right)^2} \quad . \qquad (1.31)$$

Ступінь точності зазвичай вважають задовільним, якщо значення показника не перевищує 5 %. Він може бути підвищений шляхом збільшення кількості повторних експериментів або ж підвищенням точності вимірювання значень досліджуваної ознаки.

Моментами розподілу називають середні значення відхилень даних:

- від середнього значення $\bar{x}$ (**центральні моменти** μ_k);
- від довільного числа С (**умовні моменти** m_k);
- від нуля (**початкові моменти** b_k).

Порядок моменту дорівнює степеню *k*, до якого підносять відповідні відхилення. У практиці зазвичай обмежуються моментами перших чотирьох порядків. За вибірковими даними моменти розраховують, використовуючи такі формули:

$$\mu_k = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^k}{n}, \quad m_k = \frac{\sum_{i=1}^{n}(x_i - C)^k}{n}, \quad b_k = \frac{\sum_{i=1}^{n}(x_i)^k}{n}. \qquad (1.32)$$

Початковий момент першого порядку b_1 є середнім арифметичним, а центральний момент другого порядку μ_k – зміщеною оцінкою дисперсії.

Величини $r_k = \frac{\mu_k}{s^k}$ називають **основними моментами порядку *k***.

Основний момент порядку 3 є коефіцієнтом асиметрії

$$A = \frac{\mu_3}{s^3}. \qquad (1.33)$$

Якщо варіанти симетрично розміщенні відносно $\bar{x}$, $A = 0$. При $A < 0$ переважають варіанти $x_i < \bar{x}$, при $A > 0$ переважають варіанти $x_i > \bar{x}$ (від'ємна і додатна асиметрія, або правостороння і лівостороння).

Основний момент порядку 4 використовується для розрахунку показника ексцесу

$$E = \frac{\mu_4}{s^4} - 3. \qquad (1.34)$$

Ексцес оцінює крутизну розподілу у порівнянні з нормальним законом розподілу. Для нормального розподілу $E = 0$. Якщо $E < 0$ розподіл називається плосковершинним , $E > 0$ – гостровершинним. Якщо кількість елементів сукупності перевищує 500, застосовують поправки Шеппарда до початкових і центральних моментів на довжину інтервалу. При цьому має виконуватися умова наближеності розподілу до симетричного.

1.7. ЗАВДАННЯ ДО РОЗДІЛУ 1

За вибірками А і В розв'язати наступні підзадачі:

1) побудувати варіаційний ряд для вибірки А та інтервальний варіаційний ряд для вибірки В;
2) обчислити відносні частоти (n_i / n) і накопичені частоти;
3) побудувати графіки варіаційного ряду (полігон та гістограму);
4) скласти емпіричну функцію розподілу;
5) побудувати графік емпіричної функції розподілу;
6) обчислити числові характеристики варіаційного ряду:
   - середнє арифметичне $\overline{x}$;
   - дисперсію $\overline{S}^2$;
   - стандартне відхилення $\overline{S}$;
   - моду M_0 , медіану M_e ;
   - асиметрію A , ексцес E .

$$\overline{X} = \frac{\sum_{i=1}^{m} n_i \frac{x_i - c}{k}}{n} * k + c \ ,$$

де x_i - варіанти випадкових величин ,

n_i - відповідні частоти ,

m - кількість варіантів ,

n - обсяг вибірки ,

k - крок таблиці (інтервал між сусідніми варіантами) ,

c - довільне число (для простоти беруть варіанту , що має максимальну частоту).

Для інтервального варіаційного ряду в даних формулах використовується значення середини інтервалу замість x_i .

Зауваження **:** формулу використовують, коли варіаційний ряд має постійний крок k .

Для змінного кроку використовують формули :

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{m} x_i n_i \ , \qquad \overline{S^2} = \frac{\sum_{i=1}^{m} (\frac{x_i - c}{k})^2 n_i}{n} k^2 - \left(\overline{x} - c\right)^2$$

Для інтервального варіаційного ряду медіана і мода визначаються так :

$$M_0 = x_0 + k \frac{n_i - n_{i-1}}{\left(n_i - n_{i-1}\right) + \left(n_i - n_{i+1}\right)} \ ,$$

де

x_0 - початок модального інтервалу, тобто інтервалу, що має максимальну частоту ;

k - довжина модального інтервалу ;

n_i - частота модального інтервалу ;

n_{i-1}, n_{i+1} - частоти інтервалів перед і після модального відповідно.

$$M_e = x_0 + k \frac{n/2 - T_{i-1}}{n_i} \quad ,$$

де x_0 - початок медіанного інтервалу, тобто інтервалу , в якому міститься серединний елемент ;

k - довжина медіанного інтервалу ;

n - об'єм вибірки ;

T_{i-1} - сума частот інтервалів, які передують медіальному ;

n_i - частота медіанного інтервалу.

Емпіричну функцію розподілу знаходять за допомогою накопичених частот :

$$F^*(x) = \sum_{x_i < x} \frac{n_i}{n}, \quad \text{тобто} \quad F^*(x) = \begin{cases} 0, x \le x_1 \\ \frac{n_1}{n}, x_1 < x \le x_2 \\ \frac{n_1 + n_2}{n}, x_2 < x \le x_3 \\ \dots \\ \sum_{i=1}^{m-1} \frac{n_i}{n}, x_{m-1} < x \le x_m \\ 1, x > x_m \end{cases}$$

Завдання зручно оформити у вигляді таблиці:

***- для вибірки А*:**

x_i	n_i	n_i/n	Накопичені частоти	$\frac{x_i - c}{k}$	$\frac{x_i - c}{k} n_i$	$\left(\frac{x_i - c}{k}\right)^2$	$\left(\frac{x_i - c}{k}\right)^2 n_i$
…	…	…	…	…	…	…	…
Σ	?				?		?

- для вибірки В:

Інтервал	Середина інтервалу	n_i	n_i/n	Накпичені Частоти	$\frac{x_i - c}{k}$	$\frac{x_i - c}{k} n_i$	$\left(\frac{x_i - c}{k}\right)^2$	$\left(\frac{x_i - c}{k}\right)^2 n_i$
…	…	…	…	…	…	…	…	…
Σ		?				?		?

Варіант 1

Вибірка А1

0	4	2	0	5	1	1	3	0	2	2	4	3	2	3	3	0	4	5	1
3	1	5	2	0	2	2	3	2	2	2	6	2	1	3	1	3	1	5	4
5	5	3	2	2	0	2	1	1	3	2	3	5	3	5	2	5	2	1	1
2	3	4	3	2	3	2	4	2											

N=69: Початок першого інтервалу: 0: Довжина інтервалу: 1.

Вибірка В1

135	133	124	132	104	152	134	130	129	120	122	124
117	123	123	129	121	122	125	131	147	124	137	112
126	128	111	129	115	147	131	132	137	119	125	120
129	125	123	127	132	118	133	132	132	134	131	120
135	132	125	132	108	114	121	133	133	135	131	125
114	115	122	131	125	132	120	126	115	117	118	118
132	134	127	127	124	135	128	127	115	144	129	120
137	127	125	116	132	120	117	127	118	109	127	122
120	135	116	118	133	136	125	126	119	126	129	127
129	124	127	132	126	131	127	130	126	124	135	127
124	123	123	130	132	143	122	139	120	134	108	132
121	111	123	140	137	120	125	131	118	120	120	136
129	127	116	138	128	133	122	131	128	140	138	134
120	120	109	137	111	115	117	130	113	126	115	124
125	118	115	128	123	129	128	120	115	134	118	135
134											

N=181: Початок першого інтервалу: 102; Довжина інтервалу: 4.

Варіант 2

Вибірка А2

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

N=66; Початок першого інтервалу: 0; Довжина інтервалу: 1.

Вибірка В2

95	96	103	89	72	105	85	85	91	101	82	91
80	85	91	87	101	94	98	85	82	94	86	72
89	83	100	86	85	95	95	83	87	92	92	79
93	88	77	92	92	103	85	90	83	86	104	104
85	85	80	95	91	93	70	83	93	95	95	78
111	95	94	84	64	87	85	87	87	81	82	97
101	86	89	80	88	85	93	79	95	90	107	93

96	83	88	91	95	94	88	80	96	93	77	71
88	97	90	86	93	91	98	95	83	84	91	99
109	80	95	87	89	85	87	72	77	90	97	87
95	91	88	91	81	88	78	75	80	97	95	83
91	78	87	92	103	77	101	66	71	90	105	76
97	75	95	88	84	96	79	89	94	100	87	100
92	100	79	96	104	84	89	82	93	92	85	80
104	87	90	85	89	83	84	98	81	97	86	81
96	82	102	73	100	81	86	84	86	88	90	94
81	99	100	81	95	88	90	87	97	90	100	94
88	85	95	74	85	88	78	97	74			

N=213; Початок першого інтервалу: 62; Довжина інтервалу: 4.

Варіант 3

Вибірка А3

3	5	6	8	4	5	4	7	2	7	7	3	7	4	4
5	4	4	5	2	4	8	8	4	6	5	9	4	0	4
4	4	9	3	3	2	1	5	2	5	5	3	4	4	7
8	9	11	4	5	2	5	7	6	1	2	5	6	3	1
2	6	7	3	3	2	5	4	8	2	6	5	9	5	5
2	8	3	6	4	6	6	8	7	3	3	7	3		

N=88; Початок першого інтервалу: 0; Довжина інтервалу: 1.

Вибірка В3

71	62	43	80	70	44	42	25	48	55	58	44	74	55	56
49	54	63	60	57	70	52	74	65	61	60	72	69	68	47
30	62	81	56	55	38	68	55	74	50	29	35	55	52	27
58	50	62	80	49	68	68	81	66	64	41	45	48	68	79
56	82	76	84	47	44	72	58	58	80	61	55	66	36	69
44	88	88	73	39	70	70	35	51	69	50	59	35	43	71
54	65	85	63	59	52	88	64	60	61	31	64	48	49	50
41	62	42	76	81	76	70	76	75	53	66	87	74	61	68
73	44	61	53	46	69	71	48	63	73	56	65	53	77	39
83	45	55	77	61	42	72	49	52	67	62	68	72	46	76
67	53	70	76	56	62	38	59	53	50	76	52	73	34	51
60	61													

N=167; Початок першого інтервалу: 23; Довжина інтервалу: 5.

Варіант 4

Вибірка А4

3	3	1	0	0	3	3	5	3	0	0	4	1	5	1	6	5	4	7	4
5	3	3	0	2	3	1	4	1	2	4	3	4	5	4	0	5	6	6	3
5	4	1	3	3	6	3	1	1	5	2	3	5	3	3	4	1	5	6	1
3	3	3	5	6	1	2	1	3	4										

N=70; Початок першого інтервалу: 0; Довжина інтервалу: 1.

Вибірка В4

58	78	84	62	63	100	55	90	102	70	66	89
71	92	71	93	83	42	110	110	56	96	95	87
88	102	104	88	64	96	92	67	78	95	71	105
50	66	73	76	100	72	86	46	102	95	98	84
82	46	60	94	109	93	79	74	62	97	94	91
81	71	89	78	85	80	93	64	65	109	89	55
103	98	108	68	65	71	82	70	84	73	65	79
99	81	92	76	82	95	75	45	94	81	84	68
77	90	103	119	57	102	100	83	68	69	68	81
83	69	90	99	69	85	84	70	80	117	76	104
78	114	79	70	56	62	73	71	77	98	86	82
54	62	82	103	91	61	93	68	109	96	67	110
84	82	56	78	80	88	66	78	65	50	88	72
94	92	89	109	69	58	75	72	101	92	75	77
85	76	85	84	68	74	78	87	69	75	61	53
70	106	68	81	61	64	100	73	74	57	63	102
96	80										

N=194; Початок першого інтервалу: 39; Довжина інтервалу: 6.

Варіант 5

Вибірка А5

4	10	7	6	3	7	8	7	4	7	10	7	3	9	3
1	5	8	10	11	6	5	7	6	3	8	4	3	8	4
10	6	8	7	8	7	7	7	4	6	7	10	4	4	0
5	4	4	8	5	5	10	7	3	8	5	6	6	6	3
5	7	8	5	7	10	9	10	8	2	3	6	9		

N=73; Початок першого інтервалу: 0; Довжина інтервалу: 1.

Вибірка В5

324	296	313	323	312	321	322	301	337	322	329	307
301	328	312	318	327	315	319	317	309	334	323	340
326	322	314	335	313	322	319	325	312	300	323	335
339	326	298	298	337	322	303	314	315	310	316	321
312	315	331	322	321	336	328	315	338	318	327	323
325	314	297	303	322	314	317	330	318	320	312	333
332	319	325	319	301	305	316	308	318	335	327	321
332	288	322	334	295	318	329	305	310	304	326	319
317	316	316	307	309	309	328	317	317	322	316	304
303	350	309	327	345	329	338	311	316	324	310	306
308	302	315	314	343	320	304	310	345	312	330	324
308	326	313	320	328	309	306	306	308	324	312	309
324	321	313	330	330	315	320	313	302	295	337	346
327	320	307	305	323	331	345	315	318	331	322	315
304	324	317	322	312	314	308	303	333	321	312	323
317	288	317	327	292	316	322	319	313	328	313	309
329	313	334	314	320	301	329	319	332	316	300	300
304	306	314	323	318	337	325	321	322	288	313	314
307	329	302	300	316	321	315	323	331	318	334	316
328	294	288	312	312	315	321	332	319			

N=237; Початок першого інтервалу: 285; Довжина інтервалу: 7.

Варіант 6

Вибірка А6

2	2	1	3	4	2	1	1	3	3	4	3	2	4	2	1	4	3	1	4
0	4	2	3	4	3	7	1	3	3	3	4	3	2	1	2	3	3	1	5
3	0	2	1	2	3	0	0	3	6	2	4	3	4	2	4	1	2	0	3
1	0	0	2																

N=64; Початок першого інтервалу: 0; Довжина інтервалу: 1.

Вибірка В6

61	59	60	50	58	71	57	61	55	75	68	65	63	68	60
66	52	70	69	62	58	56	54	65	61	67	64	58	61	64
71	60	51	54	57	56	55	57	65	56	61	49	67	64	59
65	63	72	67	54	53	58	69	63	66	55	57	68	63	61
55	69	54	64	54	61	66	65	57	60	72	62	68	61	62

52	62	55	70	72	64	71	54	58	71	66	65	66	62	68
60	64	63	61	60	64	65	68	64	66	69	53	57	59	62
60	63	65	60	66	68	66	64	64	67	62	55	65	62	60
55	65	56	57	72	53	62	68	63	57	55	68	59	61	63
62	63	62	59	67	56	65	67	56	69	63	53	55	67	61
54	68	59	63	67	57	64	68	76	64	64				

N=161; Початок першого інтервалу: 48; Довжина інтервалу: 3

Варіант 7

Вибірка А7

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

N=80; Початок першого інтервалу: 1; Довжина інтервалу: 1.

Вибірка В7

78	85	52	53	62	56	58	68	98	58	94	84	57	68	64
61	66	62	53	89	66	54	62	57	64	66	35	53	73	57
62	54	75	52	59	72	54	66	46	44	57	63	86	63	61
59	54	83	53	71	64	60	48	77	47	51	54	60	67	85
54	66	64	82	78	70	88	61	63	77	41	62	69	60	64
64	66	80	71	53	99	58	63	43	56	51	70	73	76	73
60	58	59	67	53	56	74	71	86	30	55	58	67	76	69
73	85	50	63	50	74	78	60	78	68	72	65	87	62	72
51	68	65	64	72	72	70	70	78	50	56	66	73	67	60
65	59	64	58	71	76	51	52	67	71	61	73	45	82	64
63	53	76	58	58	77	68	67	60	69	64	53	60	79	79
80	53	83	51	46	63	74	45	73	70	92	79	82	73	64
69	56	48	64	72	62	67	49	58	73	52	64	67	57	40
70	64	75	78	59	51	86	74	72	43	53	65	53	98	64
66	54	70	81	47	68	85	93	70	51	71	87	56	63	49
79	46	54	49	63	96	61	82	61						

N=235; Початок першого інтервалу: 28; Довжина інтервалу: 5.

Варіант 8

Вибірка А8

2	1	2	3	1	1	0	2	2	4	3	3	0	3	0	3	2	3	1	2
2	3	0	2	3	0	2	3	3	4	4	1	4	0	0	1	2	4	4	3
0	0	0	2	2	3	2	1	0	0	0	3	1	0	1	2	1	2	2	4
6	2	0	0	1	0	3	0	0	3	1	3	4	2	3	3	2	0	4	

N=79; Початок першого інтервалу: 0; Довжина інтервалу: 1.

Вибірка В8

56	76	65	76	76	62	79	48	62	50	47	80	67	87	78
55	67	51	73	75	61	88	46	57	65	60	72	28	75	51
69	68	65	34	77	63	57	61	42	85	49	41	62	63	80
62	55	45	56	66	92	60	43	52	80	68	70	76	82	55
42	87	81	67	65	81	90	38	58	60	79	79	50	64	70
58	77	73	54	58	77	86	52	61	42	70	93	54	65	51
53	64	65	76	88	59	62	67	62	90	88	69	61	81	65
72	58	68	94	54	58	58	81	57	70	71	78	52	93	89
57	68	70	58	72	57	62	63	87	61	91	57	57	66	68
40	63	86	48	75	66	83	64	55	75	65	67	54	70	44
51	86	67	58	73	71	46	86	68	79	50	58	66	69	61
64	78	78	60	46	71	71	74	79	65	61	62	84	53	67
83	43	64	67	50	60	83	61	83	67	67	58	46	73	58
47	76	81	72	66	83	73	71	70	60	68	52	51	63	63
75	61	80	51	63	62	46								

N=217; Початок першого інтервалу: 26; Довжина інтервалу: 5.

Питання для самоконтролю

1. Назвіть основні етапи аналізу даних.
2. Як поділяються ознаки за шкалами вимірювання?
3. Які основні методи аналізу даних?
4. Що називають генеральною сукупністю і вибіркою?
5. Яка вибірка називається репрезентативною?
6. Що називається варіаційним рядом або емпіричним розподілом?
7. Дайте визначення теоретичної та емпіричної функцій розподілу.
8. Що таке ранг спостережень?
9. Що таке описова статистика?
10. Назвіть основні описові статистики. Як вони обчислюються?

2. ПЕРЕВІРКА СТАТИСТИЧНИХ ГІПОТЕЗ

Розглянуті в цьому розділі методи застосовують при порівнянні двох вибірок. При більшій кількості вибірок використовують методи дисперсійного аналізу.

Під **статистичною гіпотезою** розуміють будь-яке твердження щодо генеральної сукупності, яке перевіряється на основі вибірки. Статистичні гіпотези висловлюють як відносно законів розподілу, так і відносно параметрів розподілу.

Для перевірки основної гіпотези H_0 весь вибірковий простір ділять на дві області: **критичну** *(w)* та **область прийняття** *(W − w)*. Якщо вибіркова точка потрапляє до першої області, гіпотезу відхиляють. У противному разі її приймають. Якщо розподіл імовірностей спостережень, що відповідає нульовій гіпотезі H_0 є відомим, то w визначають так, щоб при виконанні H_0 імовірність її відхилення була рівною заздалегідь заданій величині (**рівню значущості**) α.

$$P\left(x \in w \middle| H_0\right) = \alpha \qquad (2.1)$$

Замість рівня значущості можна використовувати також **довірчий рівень** $p = 1 - \alpha$**.**

Розглядають два типи похибок:

- похибкою першого роду є помилкове відхилення правильної нульової гіпотези H_0. Рівень значущості α є імовірністю такої похибки;
- похибкою другого роду є помилкове прийняття помилкової нульової гіпотези H_0.

Зменшення імовірності похибки першого роду водночас призводить до підвищення імовірності похибки другого роду β. Тому додатково уводять поняття **потужності критерію** $1-\beta$, яка є імовірністю відхилення помилкової нульової гіпотези:

$$P\{x \in (W-w)|H_1\} = 1-\beta \quad або \quad P\{x \in w|H_1\} = 1-\beta \qquad (2.2)$$

Потужність критерію можна підвищити, збільшуючи обсяг вибірки. У практиці рекомендується обирати більш потужні критерії, якщо їх застосування є обґрунтованим.

Загальна методика отримання висновків при перевірці гіпотез передбачає обрання рівня значущості (зазвичай 0,01; 0,05 або 0,1) та його наступне порівняння з розрахованим значенням певного критерію. У багатьох випадках значення критерію може бути наближено визначено за відповідними таблицями, або розраховується після застосування нормальної або іншої апроксимації вихідної статистики.

При перевірці гіпотез рекомендується застосовувати різні методи, призначені для одного й того самого типу аналізу та однакових типів даних. Причинами розбіжності отримуваних при цьому результатів зазвичай є:

- помилки при введенні даних;
- непридатність окремих методик для типу аналізованих даних;
- алгоритмічні помилки у програмах, що використовуються для аналізу.

Залежно від наявності або відсутності можливості визначення напряму розбіжності порівнюваних вибірок, розрізняють **однобічні** та **двобічні критерії**. Перші застосовують, якщо наявні дані дають змогу

вказати такий напрям, наприклад зробити висновок, що значення порівнюваної ознаки для одної вибірки є вищим, ніж у іншої. Двобічні критерії дають можливість зробити висновок лише про різницю вибірок за порівнюваною ознакою. Відповідно до цього говорять про однобічні й двобічні гіпотези.

Для двобічних критеріїв рівень значущості є удвічі більшим, ніж для відповідних однобічних. При використанні однобічних критеріїв рекомендується попередньо розраховувати двобічні. Якщо за двобічним критерієм різниці між вибірками немає, то наступне порівняння за однобічним є необґрунтованим.

Дані реальних експериментів можуть бути подані **незалежними** або **спряженими** вибірками. Для незалежних вибірок критерії допомагають виявити статистичну значущість різниці, що спостерігається. Прикладами незалежних вибірок є:

- мешканці двох різних населених пунктів (при демографічних дослідженнях);
- дві партії однотипної продукції, виготовлені різними працівниками на різному обладнанні (при розробці технології виробництва).

Критерії, що застосовують до вибірок з попарно спряженими даними, називають **парними**. Прикладами спряжених вибірок є:

- дані опитування громадської думки до і після певної суспільно значущої події;
- дві партії однотипної продукції, виготовлені одними й тими самими працівниками на одному й тому самому обладнанні до і після внесення певних змін до технології.

2.1. Параметричні тести

Критерії та тести, що застосовують для порівняння вибірок ділять на дві групи: параметричні й непараметричні. Особливістю параметричних критеріїв є припущення, що розподіл ознаки в генеральній сукупності підпорядковується нормальному закону. Нормальність вибірки має бути доведеною до застосування будь-якого з параметричних тестів. Як правило, параметричні критерії є потужнішими за непараметричні. Застосування непараметричних критеріїв у випадках, коли допустиме використання параметричних, приводить до збільшення імовірності прийняття помилкової нульової гіпотези, тобто похибки другого роду.

Для порівняння середніх значень вибірок застосовують **t - критерій Стьюдента**. Розглядають дві незалежні вибірки з генеральних сукупностей, що мають відомі (рівні або нерівні), чи невідомі але рівні дисперсії. Нульова гіпотеза полягає у рівності середніх.

У разі відомих дисперсій значення критерію обчислюють за формулою

$$t = \frac{\left|\overline{x_1} - \overline{x_2}\right|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad (2.3)$$

де n_1, n_2 − кількість елементів у вибірках. Для застосування критерію наявність повної таблиці даних не є обов'язковою. Відповідна статистика має стандартний нормальний розподіл.

За невідомих рівних дисперсій дисперсії заміняють вибірковим середнім квадратичним відхиленням, яке розраховують за формулами:

$$s^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}, \tag{2.4}$$

якщо середні вибірок оцінюють за самими вибірками, або:

$$s^2 = \frac{s_1^2 n_1 + s_2^2 n_2}{n_1 + n_2 - 2}, \tag{2.5}$$

якщо вони були відомими. Формула для розрахунку значення критерію набуває при цьому вигляду:

$$t = \frac{\left|\overline{x_1} - \overline{x_2}\right|}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{2.6}$$

Відповідна статистика має розподіл Стьюдента з $n_1 - n_2 - 2$ степенями вільності.

При аналізі спряжених вибірок їх порівняння здійснюють з метою визначення наявності ефекту від певного фактора, наприклад, впливу змін у технології на якість виробленої продукції. Вимога рівності дисперсій при цьому не висувається. Нульова гіпотеза полягає у відсутності різниці між середніми. Значення критерію розраховують за формулою:

$$t = \frac{\sum_{i=1}^{n} \delta_i}{\sqrt{\frac{n\sum_{i=1}^{n} \delta_i^2 - \left(\sum_{i=1}^{n} \delta_i\right)^2}{n-1}}}, \tag{2.7}$$

де n – кількість елементів у кожній із вибірок, $\delta_i = x_i - y_i$, x_i та y_i - відповідні значення елементів першої та другої вибірок. Іноді цей критерій називають одновибірковим критерієм Стьюдента. Відповідна

статистика має розподіл Стьюдента з кількістю степенів вільності $n-1$.

Якщо дисперсія або їх відношення є невідомими і припущення про рівність дисперсій є необґрунтованою, то виникає так звана **проблема Беренса-Фішера,** яка полягає у перевірці вказаної вище нульової гіпотези за таких умов. Одним з підходів до її вирішення є застосування **критерію Уелча.** Його значення розраховують за формулою:

$$d = \frac{\left|\overline{x_1} - \overline{x_2}\right|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{2.8}$$

де s_1^2, s_2^2 - розраховані за вибірками оцінки дисперсії. Статистика цього критерію приблизно є такою, як для розподілу Стюдента з кількістю степенів вільності

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} \tag{2.9}$$

F-критерій Фішера використовують для порівняння дисперсій двох вибірок. Його розраховують за формулою:

$$F = \frac{s_1^2}{s_2^2} \tag{2.10}$$

де s_1^2, s_2^2 - значення оцінок більшої та меншої дисперсій, відповідно. Числа степенів вільності для пошуку критичного значення обираються рівними $n_1 - 1$ та $n_2 - 1$. Гіпотеза про рівність дисперсій порівнюва-

них сукупностей відхиляється, якщо обчислення перевищує табличне при заданому довірчому рівні.

2.2. Множинні параметричні порівняння

Параметричні тести можна застосовувати також при **множинних порівняннях**, тобто при порівнянні двох груп вибірок одна з одною. Кожну групу задають подібно до того, як задають параметри масивів даних у методах двофакторного дисперсійного аналізу. При множинних порівняннях використовують багатовимірні узагальнення тестів, що були розглянуті вище.

Значення критерію Стьюдента розраховують за формулою:

$$t = \frac{\left|\overline{x_A} - \overline{x_B}\right|}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} \qquad (2.11)$$

де σ_A^2, σ_B^2 – відомі внутрішньогрупові дисперсії, n_A *та* n_B – чисельності груп. Для *m* груп рівної чисельності статистика має t-розподіл з кількістю степенів вільності $m(n-1)$.

2.3. Непараметричні тести

У багатьох випадках емпіричні дані не задовольняють вимогу нормального розподілу. Тому для їх аналізу некоректно застосовувати параметричні тести. Серед непараметричних тестів важливе місце займають так звані **робастні методи**, що виявляють слабку чутливість до відхилень від стандартних умов і можуть використовуватися в широкому діапазоні реальних умов.

Для незалежних вибірок можна застосовувати **критерій рандомізації компонент Фішера**. Він розроблений для аналізу вибірок малого обсягу. Нульова гіпотеза полягає в належності двох досліджуваних вибірок генеральним сукупностям з однаковими серед-німи. Нехай є дві вибірки: $x_i = (i = 1,2,...,n_x)$ та $y_i = (i = 1,2,...,n_y)$, де n_x, n_y – кількість елементів у них. Методика тесту основана на переборі всіх комбінацій даних. Знаходимо величину:

$$S = \min\left\{\sum_{i=1}^{n_x} x_i; \sum_{j=1}^{n_y} y_j\right\} \tag{2.12}$$

Кількість сприятливих результатів визначають за формулою:

$$N = 2\sum_{i=1}^{C_n^m} n_i,\ n_i = \begin{bmatrix} 0\ (s_i < S) \\ 1\ (s_i \geq S) \end{bmatrix}, \tag{2.13}$$

де n_i – оцінка і-го результату, C_n^m – загальна кількість результатів, $n = n_x + n_y$ – чисельність об'єднаної вибірки, m – чисельність вибірки, що відповідає мінімальному значенню $s_j = \sum_{j=1}^{m} z_{ij}$ $(i = 1,2,...,C_n^m)$, z_{ij} – масив сполучень із об'єднаної вибірки, який будується подібно до розглянутої нижче процедури побудови матриці можливих результатів для зв'язаних вибірок. Однобічне значення p розраховують як $p = \frac{N}{C_n^m}$. Його порівнюють із заданим рівнем значущості α. Нульову гіпотезу відхиляють, якщо $p < \alpha$ або $p > 1 - \alpha$. Для застосування однобічного критерію обсяг вибірки має бути не нижчим ніж 5 або 6

при рівнях значущості 0,01 та 0,05 відповідно. Для двобічних критеріїв при тих самих рівнях значущості мінімальні обсяги вибірки дорівнюють, відповідно, 7 та 8. За великих обсягів вибірки час обчислень швидко зростає і більш доцільно використовувати інші критерії.

Нульову гіпотезу про рівність середніх двох зв'язаних сукупностей можна перевірити за аналогічним методом. Основою є перебір можливих результатів, побудованих з різницевих оцінок. Нехай дані вибірки $x_i, y_i \ \left(i = 1,2,...,n\right)$ де n − кількість пар експериментальних значень. Значення різницевих оцінок визначимо за формулою:

$$s_j = \sum_{j=1}^{n} a_{ij} \left| x_j - y_j \right| \ \left(i = 1,2,...,2^n\right) \qquad (2.14)$$

де $a_{ij} \left(i = 1,2,...,2^n; j = 1,...,n\right)$ − елементи матриці можливих результатів, побудованої згідно з методикою побудови повного ортогонального плану експерименту. Вона є матрицею з 2^n рядків та n стовпців. При цьому j-й стовпець містить величини +1 та −1, що чергуються з кроком 2^{j-1}. Для $n = 3$ такий план має вигляд:

$$A = \begin{pmatrix} +1 & +1 & +1 \\ -1 & +1 & +1 \\ +1 & -1 & +1 \\ -1 & -1 & +1 \\ +1 & +1 & -1 \\ -1 & +1 & -1 \\ +1 & -1 & -1 \\ -1 & -1 & -1 \end{pmatrix}$$

Сума масиву різницевих оцінок $S = \sum_{i=1}^{n} s_i$. Кількість сприятливих результатів:

$$N = \sum_{i=1}^{2^n} n_i, n_i = \begin{bmatrix} 0 \ (s_i < S) \\ 1 \ (s_i \geq S) \end{bmatrix}. \qquad (2.15)$$

Однобічне p - значення розраховують за формулою $p = \frac{N}{2^n}$ і порівнюють із заданим значенням довірчого рівня p^*. Якщо $p > p^*$, то нульову гіпотезу приймають на рівні значущості $1 - p^*$. Як і у попередньому випадку, критерій застосовують для відносно малих вибірок. Їх мінімальний допустимий обсяг є таким самим, як і для незв'язаних вибірок.

W-критерій Вілкоксона (**критерій рангових сум**) застосовують для порівняння двох незалежних сукупностей за їх центральною тенденцією, тобто за центрами емпіричних функцій розподілу. Сукупності можуть мати як однакові, так і різні чисельності. Критерій оперує не чисельними значеннями даних, а їх рангами − місцями у впорядкованих за спаданням або зростанням рядах даних.

Процедура обчислення значення критерію є близькою до обчислення критерію рандомізації компонент. Різниця полягає в тому, що замість даних використовують їх ранги. Ранжирування порівнюваних вибірок здійснюють сумісно. З погляду простоти реалізації найбільш зручним є об'єднання вибірок, їх сортування, ранжирування та наступне рознесення рангів на місця даних, що відповідають їм в обох вибірках. Якщо деякі значення збігаються, то відповідним спостережен-

ням призначають середній ранг. Обчислення статистики критерію здійснюють за формулою:

$$W = \min\left\{\sum_{i=1}^{n_1} R_i, \sum_{i=1}^{n_2} S_i\right\} \tag{2.16}$$

де R_i − ранги вибірки, що має найменшу, а S_i − вибірки, яка має найбільшу суму рангів. Статистика $W^* = \left|\frac{W - \mu_W}{\sigma_W}\right|$, де $\mu_W = \frac{n_1(N+1)}{2}$ − математичне сподівання, $\sigma_W^2 = \frac{n_1 n_2 (N+1)}{12}$ − дисперсія, $N = n_1 + n_2$, наближається до стандартного нормального розподілу. Нульову гіпотезу відхиляють на рівні значущості α, якщо $W^* > z_{1-\alpha/2}$ для двобічної гіпотези. Якщо отримане значення α перевищує 0,02, то вводять поправку на неперервність і вважають, що нове значення найменшої суми рангів дорівнює W + 0,5.

U-критерій Манна − Вітні призначений для перевірки нульової гіпотези про однаковість розподілу досліджуваних сукупностей або для перевірки рівності окремих параметрів цих розподілів, наприклад, середніх значень. Спостереження мають бути непарними. Обчислення здійснюють за формулами:

$$\begin{aligned} U_1 &= n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \\ U_2 &= n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 \\ U &= \max\{U_1, U_2\} \end{aligned} \tag{2.17}$$

де R_1, R_2 − суми рангів вибірок, n_1, n_2 − кількість елементів у них.

Модифікована статистика $\frac{U-\mu_U}{\sigma_U}$, де $\mu_U=\frac{n_1 n_2}{2}$ − математичне сподівання, $\sigma_U^2=\frac{n_1 n_2(N+1)}{12}$ − дисперсія, $N=n_1+n_2$, має стандартний нормальний розподіл. Результати обчислення за цим критерієм збігаються з даними, отримуваними за W-критерієм Вілкоксона. На цьому критерії базується **багатовимірний тест Джонкхіра−Терпстра**.

Т-критерій Вілкоксона (**одновибірковий**, **знаковий ранговий критерій Вілкоксона**) застосовують для порівняння вибірок з попарно спряженими значеннями. Перевіряють нульову гіпотезу про симетричність розподілу різниць спряжених значень відносно нуля. Методика розрахунку є близькою до розрахунку значення W-критерію Вілкоксона. Але в цьому разі оперують модулями різниць відповідних значень. Масив різниць ранжують. Потім рангам надають знаки різниць і обчислюють суму додатних W^+ рангів, яку порівнюють з критичним значенням. Статистика $\frac{W^*-\mu_{W^*}}{\sigma_{W^*}}$, де $\mu_{W^*}=\frac{n_1(N+1)}{4}$ − математичне сподівання, $\sigma_{W^*}^2=\frac{N(N+1)(2N+1)}{24}$ − дисперсія, N − чисельність кожного ряду, має стандартний нормальний розподіл.

Зупинимося докладніше на двох проблемах, що виникають при застосуванні критерію Вілкоксона та подібних йому критеріїв. Першою з них є проблема рівних рангів яка існує лише при застосуванні нормальної, або іншої апроксимації критерію і не має місця при точному обчисленні відповідних статистик. Для W-критерію Вілкоксона у

цьому разі необхідно врахувати поправку до дисперсії та розраховувати її за формулою:

$$\sigma_W^2 = \frac{n_1 n_2}{12}\left[N + 1 - \frac{T}{N(N-1)}\right], \qquad (2.18)$$

де T − поправка на об'єднання рангу, формулу для обчислення якої надано у розділі 1, $N = n_1 + n_2$. Для U-критерію Манна − Вітні модифікований вираз для дисперсії має аналогічний вигляд. Для Т-критерію Вілкоксона скореговану дисперсію визначають за формулою:

$$\sigma_{W*}^2 = \frac{2N(N+1)(2N+1) - T}{48}, \qquad (2.19)$$

де N − чисельність кожного ряду.

Друга проблема виникає тільки для Т-критерію Вілкоксона і полягає в тому, що наявність збігів одночасно в обох аналізованих вибірках призводить до нульових різниць. На сьогодні ця проблема залишається неповністю дослідженою. Один з підходів до її вирішення полягає у викреслюванні рівних спряжених значень із порівнюваних вибірок. При цьому обсяги вибірок зменшуються на кількість викреслених значень.

Критерій χ^2 використовують для перевірки нульової гіпотези про однаковість розподілу досліджуваних випадкових величин. Він має широке застосування у дисперсійному аналізі та інших методах аналізу даних. Цей критерій оперує не первинними даними, а їх розподілом за класами. Тому необхідно враховувати вимогу щодо мінімальних обсягів вибірок та кількостей класів. За різними оцінками мінімальна допустима кількість класів знаходиться у межах 4−7, а кількість

елементів у вибірках − у межах 20−40. Якщо аналізовані дані вимірювали в кількісних або порядкових шкалах, то при порівнянні вибірок однакового обсягу значення критерію обчислюють за формулою:

$$\chi^2 = \sum_{i=1}^{N} \frac{(f_i - g_i)^2}{f_i + g_i}, \qquad (2.20)$$

де f_i, g_i $(i = 1,2,...,n)$ − частоти розподілів порівнюваних вибірок, n − чисельність кожного з масивів частот.

За тих самих умов для вибірок різного обсягу значення критерію обчислюють за формулою:

$$\chi^2 = \frac{1}{n_1 n_2} \sum_{i=1}^{N} \frac{(n_2 f_i - n_1 g_i)^2}{f_i + g_i}, \qquad (2.21)$$

де n_1, n_2 − чисельність порівнюваних масивів.

Критерій χ^2 можна застосовувати також і для порівняння вибірок значень номінальних ознак. У цьому разі аналізують дані, подані у вигляді таблиці спряженості ознак. Елементами таблиці є числа, рівні кількості елементів певної вибірки, для яких ступінь прояву деякої ознаки відповідає деякому класу. Кожний рядок таблиці характеризує розподіл елементів відповідної вибірки за класами, а кожний стовпець − наповненість певного класу в різних вибірках. Значення критерію розраховують за формулою:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(a_{ij} - e_{ij})^2}{e_{ij}}, \qquad (2.22)$$

де a_{ij} $(i = 1,2,...,r; j = 1,2,...,c)$ − елементи таблиці спряженості ознак, r − кількість вибірок (стовпців), c − кількість класів (рядків), e_{ij} − очі-

кувані величини, що відповідають значенням a_{ij}. Їх обчислюють як добуток і-го вектора-стовпця на j-й вектор-рядок, ділений на суму елементів усієї таблиці $\sum_{i=1}^{r}\sum_{j=1}^{c} a_{ij}$. Кількість степенів вільності при обчисленні Р-значення статистики χ^2 беруть рівною $(r-1)(c-1)$.

2.4. Визначення моделей розподілу емпіричних даних

У практиці часто виникає проблема перевірки відповідності емпіричного розподілу деякому заданому теоретичному. При цьому розрізняють прості та складні гіпотези. Якщо гіпотеза стверджує що із l параметрів розподілу k мають задані значення, то гіпотезу вважають простою, коли $k = l$, і складною, у випадку $k < l$. Різницю $l - k$ називають кількістю степенів вільності гіпотези, а k − кількістю накладених обмежень.

У практиці важливу роль відіграє перевірка розподілу на нормальність. Найпростіші способи її реалізації основані на розрахунку значень коефіцієнтів асиметрії та ексцесу. Вважають, що розподіл є близьким до нормального, якщо $|A|, |E| < 0{,}1$, і сильно відрізняється від такого, коли $|A|, |E| > 0{,}5$.

Для перевірки відповідності емпіричного розподілу теоретичному застосовують критерії Колмогорова, ω^2-критерій Мізеса, критерій χ^2, критерії Ястремського, Бернштейна та інші.

Критерій згоди Колмогорова (**статистика Колмогорова − Смирнова**) має вигляд:

$$D_n = \sup_{-\infty<x<+\infty} |F_n(x) - F(x)|, \qquad (2.23)$$

де D_n − максимальна різниця між відносними частотами, n − обсяг ряду частот, $F_n(x)$ − емпірична функція розподілу, $F(x)$ − теоретична функція розподілу, x − інтервальний варіаційний ряд, побудований за вихідною вибіркою. Цю формулу застосовують, якщо емпіричну функцію розподілу будують за масивом частот. У разі, коли її побудова здійснюється безпосередньо за вихідною вибіркою (при цьому чисельність вихідної вибірки та масиву функції розподілу збігаються), то для розрахунку критерію при двобічній гіпотезі застосовують формули:

$$D_n = \max_{1\le m\le n}\left\{D_n^{(1)}, D_n^{(2)}\right\}$$
$$D_n^{(1)} = \max_{1\le m\le n}\left\{\frac{m}{n} - F(x_m)\right\} \quad D_n^{(2)} = \max_{1\le m\le n}\left\{F(x_m) - \frac{m-1}{n}\right\} \qquad (2.24)$$

а при однобічній – $D_n = D_n^{(1)}$.

Функція розподілу *Dn* є однією й тією самою для всіх неперервних розподілів, а функція розподілу величини $D_n\sqrt{n}$ за великих n збігається до статистики Колмогорова. Обмеженнями для застосування даного критерію є:

- вимога неперервності теоретичної функції розподілу;
- необхідність незалежної, тобто здійсненої не за самою вибіркою, оцінки параметрів розподілу (проста гіпотеза).

Для складних гіпотез уводять модифіковані статистики Колмогорова, але їх функції розподілу є різними для різних типів неперервних розподілів.

Критерій χ^2 як критерій згоди застосовують для порівняння емпіричної та теоретичної функцій розподілу. Його розраховують за формулою:

$$\chi^2 = \sum_{i=1}^{k} \frac{(\nu_i - np'_i)^2}{np'_i} \qquad (2.25)$$

де ν_i − абсолютні частоти для k класів, p'_i − теоретичні ймовірності обраного розподілу (параметри теоретичного розподілу розраховують за емпіричною вибіркою або задають), n − загальна кількість спостережень (для неперервного розподілу цю величину треба помножити на довжину класового інтервалу *d)*. Кількість степенів вільності беруть рівною $k - r - l$, де r − кількість параметрів теоретичного розподілу. Зокрема, при обчисленні параметрів теоретичного розподілу за інтервальним варіаційним рядом кількість степенів вільності беруть рівною $k - 2$ для біноміального і $k - 3$ для нормального розподілу.

Загальна схема застосування критерію χ^2 є такою. Спочатку будують емпіричну функцію розподілу. Потім визначають r параметрів, які необхідні для побудови теоретичної функції розподілу, розраховують її та визначають значення критерію χ^2.

Розглянуті вище критерії можуть застосовуватися для вибірок достатньо великого обсягу. Для перевірки нормальності вибірок обсягом від 3 до 50 значень можна використовувати **W-критерій Шапіро − Вілка**, побудований на регресії порядкових статистик. Обчислення здійснюють за формулами:

$$W = \frac{b^2}{S^2}; \quad S^2 = \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2; \quad b = \sum_{i=1}^{k} a_{n-i+1} \left(x_{n-i+1} - x_i\right) \qquad (2.26)$$

де x_i $(i=1,2,...,n)$ – ранжований ряд, n – обсяг вибірки, параметр k беруть рівним $n/2$ для парних і $(n-1)/2$ для непарних n, a_{n-i-1} $(i=1,2,...,k; n=3,4,...,50)$ – константи. Гіпотезу про нормальний розподіл приймають, якщо значення критерію перевищує критичну для заданого довірчого рівня величину.

Перевірку нормальності розподілу можна здійснити також за результатами аналізу процентилей. Розподіл можна вважати близьким до нормального, якщо значення окремих процентилей є близькими до наведених нижче величин:

- 2,5 % процентиль $\approx \mu - 2\sigma$;
- 16 % процентиль $\approx \mu - \sigma$;
- 50 % процентиль $\approx \mu$;
- 84 % процентиль $\approx \mu + \sigma$;
- 97,5 % процентиль $\approx \mu + 2\sigma$.

2.5. ЗАВДАННЯ ДО РОЗДІЛУ 2

Завдання 1

За рядками F_1 і F_2 при рівні значимості α перевірити гіпотезу про рівність дисперсій $H_0 : \sigma_1^2 = \sigma_2^2$ при альтернативній гіпотезі $H_a : \sigma_1^2 > \sigma_2^2$.

$\alpha = \begin{cases} 0{,}05, & V - парне \\ 0{,}01, & V - непарне \end{cases}$, де V – номер варіанту.

Варіант 1

F1	43	51	44	47	34
F2	52	12	40	38	37

Варіант 2

F1	41	41	34	38	50	45
F2	49	46	41	41	47	47

Варіант 3

F1	110	92	113	110
F2	88	94	100	99

Варіант 4

F1	-18	-12	-18	-20	-21	-19	-18
F2	-26	-23	-16	-13	-22	-28	-18

Варіант 5

F1	49	43	47	43	46	45
F2	45	46	46	44	43	45

Завдання 2

За рядками X_1, X_2 вибірки С при рівні значимості α перевірити гіпотезу про рівність середніх значень.

$\alpha = \begin{cases} 0{,}05, & V - парне \\ 0{,}01, & V - непарне \end{cases}$, де V – номер варіанту.

Теоретичні відомості

$H_0 : \bar{x}_1 = \bar{x}_2$,

$H_a : \bar{x}_1 > \bar{x}_2$,

$\nu = n_1 + n_2 - 2$ - кількість степенів вільності.

Відповідна статистика для перевірки гіпотези про рівність середніх значень з використанням критерію Стьюдента обчислюється так:

$$t = \frac{\overline{x_1} - \overline{x_2} - \delta_0}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \quad \delta = \mu_1 - \mu_2, \quad \delta_0 = 0.$$

Варіант 1

Вибірка С

X1	67	68	70	76	80	87	75	79	79	73	86	78	79	67	79
X2	82	70	83	80	76	81	80	76	70	79	74	77	65	80	79

Варіант 2

Вибірка С

X1	48	40	52	50	39	47	38	46	47	44	45	44	53	52	45	42
X2	61	42	55	52	44	41	47	43	55	43	49	42	31	40	47	43

Варіант 3

Вибірка С

X1	46	55	57	55	51	62	43	64	56	65	56	51	58	42	46	54
X2	59	57	48	41	57	50	64	43	57	54	50	59	48	45	51	40

Варіант 4

Вибірка С

X1	67	57	82	62	57	83	69	86	66	79	82	73	65	55	61	64
X2	75	68	65	80	71	61	77	69	64	68	52	59	63	56	62	67

Варіант 5

Вибірка С

X1	31	28	30	23	25	25	25	27	31	25	28	25	30	28	28	31
X2	23	28	26	26	28	27	25	26	27	28	27	29	24	23	27	27

Завдання 3

$$\alpha = \begin{cases} 0{,}1, & k = 0 \\ 0{,}05, & k = 1 \\ 0{,}02, & k = 2 \\ 0{,}01, & k = 3 \end{cases}$$

За вибіркою В (див. стор. 35-40) при рівні значимості α перевірити гіпотезу про нормальний розподіл відповідної генеральної сукупності.

Тут k - залишок $V/4$, V - номер варіанту, $\nu = r - 3$ -кількість степенів вільності, r - кількість об'єднаних інтервалів. Обчислення зручно оформити у вигляді наступної таблиці:

Інтервал	x_i	n_i	$\frac{x_i - \bar{x}}{S^2}$	$\Phi\left(\frac{x_i - \bar{x}}{S^2}\right)$	P_i	m_i
…		…			…	…
$\sum$		?			?	?

продовження таблиці

$n_i - m_i$	$(n_i - m_i)^2$	$\frac{(n_i - m_i)^2}{m_i}$
…	…	…
		?

Питання для самоконтролю

1. Що таке статистична гіпотеза? Види статистичних гіпотез.
2. Що таке рівень значущості?
3. Які два типи похибок розглядають?
4. Які критерії існують для перевірки гіпотез?
5. Які критерії належать до параметричних, до непараметричних?
6. Назвіть критерії згоди.

3. ДИСПЕРСІЙНИЙ АНАЛІЗ

Дисперсійний аналіз був створений для статистичної обробки агрономічних даних. В наш час його використовують для обробки економічної та соціальної інформації.

Дисперсійний аналіз - це сукупність статистичних методів, призначених для перевірки гіпотез про зв'язок між певною ознакою та досліджуваними факторами, які не мають кількісного опису, а також встановлення ступеня впливу факторів та їх взаємодії. **Факторами** називають контрольовані чинники, що впливають на кінцевий результат. **Рівнем фактора**, або **способом обробки**, називають значення, що характеризують конкретний прояв даного фактора. Ці значення зазвичай подають у номінальній або порядковій шкалі вимірювань. Значення вимірюваної ознаки називають **відгуком**. Для здійснення дисперсійного аналізу необхідно, щоб результати експериментів були незалежними нормально розподіленими випадковими змінними з однаковою дисперсією. Результати експериментів можна класифікувати та аналізувати за одним деяким фактором - однофакторний аналіз, за двома факторами - двофакторний аналіз тощо.

3.1. Однофакторний аналіз

Основною метою однофакторного аналізу зазвичай є оцінка сили впливу конкретного фактора на досліджуваний відгук. Іншою метою може бути порівняння двох або декількох факторів з метою визначення різниці їх впливу на відгук, яку часто називають **контрастом** факторів. Попереднім етапом є перевірка нульової гіпотези H_0 про відсутність будь-якого впливу досліджуваного фактора, тобто гіпотези про

те, що зміни значень ознаки у порівнюваних вибірках є випадковими, і всі дані належать до однієї генеральної сукупності.

Якщо нульову гіпотезу відкидають, то наступним етапом є кількісна оцінка впливу досліджуваного фактора і побудова довірчих інтервалів для отриманих характеристик. У разі, коли нульова гіпотеза не може бути відкинутою, зазвичай її приймають і роблять висновок про відсутність впливу. Але, якщо є підстави вважати, що такий вплив має бути присутнім (наприклад, це може випливати з теоретичних уявлень про об'єкт дослідження), то необхідно перевірити наявність інших факторів, що можуть впливати на відгук.

При **однофакторному дисперсійному аналізі** вихідні дані подають у вигляді таблиць, в яких кількість стовпчиків дорівнює кількості рівнів фактора, а кількість значень у кожному стовпчику – кількості спостережень при відповідному рівні фактора (табл.3.1). Для різних рівнів фактора кількість спостережень може бути різною. Припускають, що результати спостережень для різних рівнів є вибірками з нормально розподілених сукупностей, середні значення та дисперсії яких є однаковими і не залежать від рівнів.

Таблиця 3.1

Форма таблиці спостережень при проведенні однофакторного дисперсійного аналізу

Результати вимірювань	Рівні фактора			
	1	*2*	…	*k*
1	x_{11}	X_{12}	…	x_{1k}
2	x_{21}	X_{22}	…	x_{2k}
…	…	…	…	…
n_i	x_{ni1}	x_{ni2}	…	x_{nik}

В основі дисперсійного аналізу лежить правило розкладу загальної дисперсії на дисперсію обумовлену впливом фактора і залишкову дисперсію, обумовлену впливом випадкових величин.

$$\sigma^2 = \sigma_1^2 + \sigma_2^2$$

Варіацію ознаки x можна пояснити двома причинами: 1. систематичний вплив, який піддається регулюванню (рівень фактора) 2. випадковий вплив, який не регулюється. Щоб оцінити достовірність різниці між груповими середніми, необхідно виміряти міжгрупову та внутрігрупову варіацію. Завданням аналізу є перевірка нульової гіпотези H_0 про рівність середніх значень сукупностей, що розглядаються. Сутність методу полягає у порівнянні двох оцінок для дисперсії. Першу з них (залишкова дисперсія) розраховують за формулою:

$$\sigma_1^2 = \frac{1}{N-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij} - \left\langle x_j \right\rangle\right)^2, \quad (3.1)$$

де $\left\langle x_j \right\rangle = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ij}$; $j = 1,...,k$; $N = \sum_{j=1}^{k} n_j$ − загальна чисельність спостережень, k − кількість вибірок, n_j *(i = 1, 2, ..., k)* − кількість елементів в j - й вибірці, $\left\langle x_j \right\rangle$ − середнє значення j - ої вибірки. Ця дисперсія є мірою варіації всередині вибірок і не пов'язана з припущенням про рівність середніх значень, тому $\sigma^2 \approx \sigma_1^2$ незалежно від справедливості нульової гіпотези.

Другу оцінку (дисперсія фактора) отримують за формулою:

$$\sigma_2^2 = \frac{1}{k-1}\sum_{j=1}^{k} n_j\left(\left\langle x_j \right\rangle - \bar{x}\right)^2, \quad (3.2)$$

де $\bar{x} = \frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}$ – загальне середнє. Вона характеризує варіацію між вибірками. При справедливості нульової гіпотези $\sigma^2 \approx \sigma_2^2$, а при її порушенні величина σ_2^2 стає тим більшою, чим сильніший вплив зовнішнього фактора.

Значення критерію розраховують за формулою:

$$F = \frac{\sigma_2^2}{\sigma_1^2} = \frac{(N-k)\sum_{j=1}^{k} n_j \left(\langle x_j \rangle - \bar{x}\right)^2}{(k-1)\sum_{j=1}^{k}\sum_{i=1}^{n_j} \left(x_{ij} - \langle x_j \rangle\right)^2} \qquad (3.3)$$

Ця величина має F-розподіл Фішера з параметрами $k-1$ та $N-k$. Нульову гіпотезу відхиляють, якщо імовірність *P(F>F*)* є достатньо малою. Тут F − значення, розраховане за емпіричними даними за формулою (3.3). Критичне значення параметра *F** можна знайти у таблиці розподілу Фішера, або ж використати функцію пакету Excel FРАСПОБР(α, k-1, N-k).

Непараметричним аналогом однофакторного дисперсійного аналізу є **ранговий однофакторний аналіз Краскела − Уолліса**, призначений для перевірки нульової гіпотези про рівність ефектів впливу на досліджувані вибірки з невідомими, але рівними середніми. При цьому кількість вибірок має бути більшою ніж дві. Рангові методи, у тому числі й метод Краскела − Уолліса, не передбачають нормальності розподілу результатів спостережень і можуть застосовуватися як для кількісних даних з невідомим законом розподілу, так і для порядкових ознак.

У табл. 3.1 замість спостережень заносять їх ранги r_{ij}, отримані шляхом впорядкування за зростанням усієї сукупності спостережень x_{ij}. Для кожного рівня фактора, тобто для кожного стовпця, розрахо-вують суму рангів $R_j = \sum_{i=1}^{n_j} r_{ij}$, або відповідні середні ранги $\langle R_j \rangle = \frac{1}{n_j} \sum_{i=1}^{n_j} r_{ij}$. Якщо між стовпцями немає систематичної різниці, то ці середні ранги будуть близькими до середнього рангу, розрахова-ного за усією сукупністю, який дорівнює $(N+1)/2$, де $N = \sum_{i=1}^{k} n_i$ – загальна кількість спостережень. Тому величини $\langle R_j \rangle - (N-1)/2$ мають бути відносно малими, якщо нульова гіпотеза є правильною. Обчислення критерію здійснюють за формулами:

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1), \qquad (3.4)$$

або

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j \left(\langle R_j \rangle - \frac{N+1}{2} \right)^2 \qquad (3.5)$$

За великих n_i статистика критерію асимптотично наближається до $\chi 2$ - розподілу з кількістю степенів вільності $k-1$. У цьому разі нульову гіпотезу відхиляють на рівні значущості α, якщо $H > \chi^2_{1-\alpha}$, де $\chi^2_{1-\alpha}$ – квантиль рівня $1-\alpha$ розподілу χ^2 з $k-1$ ступенем вільності. При $k = 2$ статистика Краскела – Уолліса стає еквівалентною з

W -статистикою Вілкоксона. Якщо серед спостережень є рівні значення, описану вище схему аналізу можна застосовувати як наближену. Надійність її висновків буде тим нижчою, чим більшою є кількість збігів. Для підвищення надійності можна використовувати середні ранги. При цьому в разі, коли вони не є цілими числами, їх не округляють. Якщо кількість збігів є великою, використовують модифіковану форму статистики Краскела − Уолліса:

$$H' = \frac{H}{1 - \left(\sum_{j=1}^{g} \frac{T_j}{N^3 - N} \right)}, \qquad (3.6)$$

де g − кількість груп спостережень, що збігаються, $T_j = \left(t_j^3 - t_j\right)$, t_j − кількість спостережень, що збігаються в j - ій групі.

Критерій Джонкхієра застосовують, тоді, коли заздалегідь відомо, що наявні групи результатів впорядковані за зростанням впливу досліджуваного фактора, який вимірюється в порядковій шкалі. Таблиця даних має такий самий вигляд, як і в попередньому випадку. Будемо вважати, що її перший стовпчик відповідає найменшому рівню фактора, другий − наступному за величиною, і так далі, останній стовпчик відповідає найбільшому рівню. За виконання таких припущень критерій Джонкхієра стосовно гіпотези про монотонний вплив фактора є більш потужним, ніж критерій Краскела − Уолліса.

Спочатку для кожної пари вибірок з номерами u, v $(1 \le u < v \le k)$, де k − кількість рівнів фактора, розраховують **статистику Манна − Вітні**:

$$U_{u,v} = \sum_{\substack{i=1,\ldots,n_u \\ j=1,\ldots,n_v}} \varphi\left(x_{iu}, y_{jv}\right) \tag{3.7}$$

де

$$\varphi\left(x_i, y_j\right) = \begin{cases} 1 & \left(x_i < y_j\right) \\ 0.5 & \left(x_i = y_j\right) \\ 0 & \left(x_i > y_j\right) \end{cases}. \tag{3.8}$$

Потім розраховують статистику Джонкхієра:

$$J = \sum_{1 \le u \le v \le k} U_{u,v} \,. \tag{3.9}$$

Великі значення J свідчать проти гіпотези про однорідність вибірок.

Для вибірок великого обсягу статистика Джонкхієра апроксимується нормальним розподілом з параметрами:

$$MJ = \frac{1}{4}\left(N^2 - \sum_{j=1}^{k} n_j^2\right);\ DJ = \frac{1}{72}\left[N^2(2N+3) - \sum_{j=1}^{k} n_j^2(2n_j+3)\right] \tag{3.10}$$

Свідченням проти гіпотези однорідності є великі, порівняно з відсотковими точками стандартного нормального розподілу, значення статистики $\dfrac{J - MJ}{\sqrt{DJ}}$.

М-критерій Бартлетта застосовують для перевірки нульової гіпотези про рівність дисперсій декількох нормальних генеральних сукупностей, з яких отримані досліджувані вибірки. Обчислення критерію здійснюють за формулами:

$$t = c\sum_{j=1}^{k}\left(n_j - 1\right)\ln\frac{s^2}{s_j^2}$$

$$c = \left[1 + \frac{1}{3(k+1)}\left(\sum_{j=1}^{k}\frac{1}{n_j - 1} - \frac{1}{N-k}\right)\right]^{-1} \qquad (3.11)$$

$$s^2 = \frac{1}{N-k}\sum_{j=1}^{k}\left(n_j - 1\right)s_j^2 ,$$

де N – загальна чисельність, k – кількість вибірок $(k > 2)$, n_j – чисельність j - ї вибірки, $s_j^2 = \frac{1}{n_j - 1}\sum_{j=1}^{n_j}\left(x_{ij} - \overline{x_j}\right)^2$ – дисперсія j - ї вибірки, $\overline{x_j}$ – середнє значення j -ї вибірки. За великих n_j статистика критерію асимптотично наближається до χ^2 - розподілу з кількістю степенів вільності $k-1$.

G-критерій Кокрена використовують для перевірки нульової гіпотези про рівність дисперсій k $(k \geq 2)$ нормальних генеральних сукупностей за незалежними вибірками рівного обсягу. Значення критерію обчислюють за формулою

$$G = \frac{\max\limits_{1\leq j\leq k}\sigma_j^2}{\sum\limits_{j=1}^{k}\sigma_j^2}, \qquad (3.12)$$

де σ_j^2 – дисперсія j-ї вибірки.

Розглянуті вище критерії дають змогу встановити різницю дисперсій сукупностей, але не дають можливості дати кількісну оцінку впливу фактора на досліджувану ознаку, а також встановити, для яких саме

сукупностей дисперсії є різними. Для встановлення кількісного впливу досліджуваного фактора часто застосовують **адитивну модель**, яка передбачає, що значення відгуку є сумою впливу фактора і незалежної від нього випадкової величини:

$$x_{ij} = a_j + \varepsilon_{ij} \ \ (j = 1,...,k; i = 1,...,n), \tag{3.13}$$

де a_j − невипадкові невідомі величини, що визначаються значеннями рівнів фактора, ε_{ij} − незалежні випадкові величини, які мають однаковий розподіл і відображають внутрішню мінливість, що не пов'язана із значеннями рівнів фактора. Модель (3.13) можна записати у вигляді:

$$x_{ij} = \mu + \tau_j + \varepsilon_{ij} \ \ (j = 1,...,k; i = 1,...,n), \tag{3.14}$$

де $\mu = \frac{1}{k}\sum_{j=1}^{k} a_j$ − середній рівень, $\tau_j = a_j - \mu$ - відхилення від середнього рівня при j-му значенні рівня фактора. У такій формі модель має на один невідомий параметр більше (середній рівень і k значень відхилень від нього), але кількість незалежних невідомих параметрів залишається рівною k, оскільки відхилення зв'язані співвідношенням

$$\sum_{j=1}^{k} \tau_j = 0 .$$

Розглянемо різницю відгуків для двох значень рівня фактора, яку часто називають **зсувом**. Як оцінку зсуву можна взяти **медіану Ходжеса − Лемана**:

$$z_{ij} = med\left\{x_{ui} - x_{vj}\right\} \ \ (u = 1,...,n_i; v = 1,...,n_j). \tag{3.15}$$

Вона має властивість: $z_{ij} = -z_{ji}$. Статистика z_{ij} може застосовуватися для оцінювання величини $\tau_i - \tau_j$. Її суттєвим недоліком є невиконання рівності: $z_{ij} = z_{ik} + z_{kj}$. Тому частіше використовують зважені скореговані оцінки зсуву (**оцінки Спетволя**):

$$W_{ij} = \overline{\Delta}_i - \overline{\Delta}_j, \qquad (3.16)$$

де величини

$$\overline{\Delta}_i = \frac{\sum_{u=1}^{k} n_u z_{iu}}{N} \quad (i=1,...,k) \qquad (3.17)$$

відображають зсув і-ї вибірки відносно всіх інших, усереднений з ваговими коефіцієнтами $n_1,...,n_k$. Оцінки Спетволя задовольняють співвідношення $W_{ij} = W_{ik} + W_{kj}$. Але вони мають інший недолік: оцінка зсуву двох вибірок одна відносно одної залежить від усіх інших вибірок.

Лінійним контрастом у моделі адитивного впливу фактора на відгук називають лінійну функцію середніх значень k незалежних нормальних вибірок з невідомими рівними дисперсіями $L = \sum_{j=1}^{k} c_j \tau_j$, де c_j – відомі сталі, які задовольняють вимогу $\sum_{j=1}^{k} c_j = 0$. Найпростішим прикладом лінійного контрасту є різниця $\tau_i - \tau_j$, де $c_i = 1,\ c_j = -1,$ $c_u = 0$ при усіх $u \neq i, j$. У разі, коли вихідні дані задовольняють умови застосовності параметричного критерію, за оцінки параметрів a_j

моделі (3.13) можна взяти внутрішньогрупові середні x_j, які у цьому випадку мають нормальний розподіл і є статистично незалежними від оцінки дисперсії σ_1^2 . Тому відношення

$$t = \frac{\langle x_j \rangle - a_j}{\sigma_1} \sqrt{n_j} \tag{3.18}$$

підпорядковується розподілу Стьюдента з $N - k$ степенями вільності.

Звідси для довірчого інтервалу значення a_j можна отримати вираз:

$$\left| \langle x_j \rangle - a_j \right| < \frac{\sigma_1}{\sqrt{n_j}} t_{1-\alpha}, \tag{3.19}$$

де $t_{1-\alpha}$ – квантиль рівня $1-\alpha$ відповідного розподілу Стьюдента. Лінійний контраст у цьому випадку дорівнює: $L = \sum_{j=1}^{k} c_j a_j$, а довірчий рівень для нього:

$$\left| L^* - L \right| = \sigma_1 t_{1-\alpha} \sqrt{\sum_{j=1}^{k} \frac{c_j}{n_j}}, \tag{3.20}$$

де $L^* = \sum_{j=1}^{k} c_j \langle x_j \rangle$.

Для порівняння вибірок, що належать до деякої множини даних, при умові, що дисперсії є різними, а середні значення однаковими, найчастіше застосовують **метод множинних порівнянь Шеффе**. Обчислення критерію при перевірці нульової гіпотези $L = L_0$ здійснюють за формулою:

$$t = \frac{\sum_{i=1}^{k} c_i \overline{x_i} - L_0}{\sqrt{M \sum_{i=1}^{k} \frac{c_i^2}{n_i}}}, \qquad (3.21)$$

де N − загальна кількість, n_i − кількість елементів в і - й вибірці, $\overline{x_i}$ − середнє значення для і-ї вибірки, $M = \frac{1}{N-k} \sum_{i=1}^{k} \sum_{j=1}^{n_j} \left(x_{ij} - \overline{x}_i \right)^2$ − середній квадратичний залишок. Розраховане значення критерію при порівнянні з критичним необхідно брати за модулем.

3.2. Двофакторний аналіз

Двофакторний аналіз застосовують для зв'язаних нормально розподілених вибірок. Дані подають у вигляді табл. 3.2, у стовпчиках якої наводять дані, що відповідають певному рівню першого фактора, а у рядках − дані, що відповідають рівням другого. Таблиця даних має розмірність $n \times k$, де n і k – кількість рівнів першого та другого факторів, відповідно.

Таблиця 3.2

Таблиця даних двофакторного аналізу

Рівні фактора А	Рівні фактора В			
	1	2	…	K
1	x_{11}	x_{12}	…	X_{1k}
2	x_{21}	x_{22}	…	X_{2k}
…	…	…	…	…
n	x_{n1}	x_{n2}	…	x_{nk}

Основною відмінністю від таблиці однофакторного дисперсійного аналізу є можлива неоднорідність даних у стовпцях, якщо вплив дру-

гого фактора є істотним. У практиці часто використовують і складніші таблиці двофакторного аналізу, зокрема такі, у яких кожна комірка містить набір даних (повторні вимірювання), що відповідають фіксованим значенням рівнів обох факторів.

Для опису даних табл.3.2 можна застосовувати адитивну модель, яка припускає, що значення відгуку є сумою внесків кожного із факторів b_i і t_j, а також незалежної від факторів випадкової компоненти ε_{ij}

$$x_{ij} = \mu + \beta_i + \tau_j + \varepsilon_{ij} \quad . \tag{3.22}$$

Тут $\mu = \frac{1}{kn}\sum_{j=1}^{k} x_{ij}$ – загальне середнє за всіма спостереженнями, а β_i і τ_j – відхилення від середнього, зумовлені факторами А і В, відповідно. У разі, коли випадкова компонента ε_{ij} підпорядковується нормальному розподілу з нульовим середнім і рівними для всіх і, j дисперсіями σ^2 застосовують двофакторний дисперсійний аналіз (дисперсійний аналіз за двома ознаками).

Основою проведення двофакторного дисперсійного аналізу служить розклад дисперсії за формулою

$$\sigma_o^2 = (n-1)\sigma_A^2 + (k-1)\sigma_B^2 + (n-1)(k-1)\sigma_1^2 \quad . \tag{3.23}$$

Тут σ_0^2 - оцінка загальної вибіркової дисперсії, яка враховує вплив факторів А,В і неврахованих залишкових факторів; σ_A^2 - оцінка дисперсії, пов'язаної з фактором А; σ_A^2 - оцінка дисперсії, пов'язаної з

фактором В; σ_1^2 - оцінка дисперсії, пов'язаної із впливом залишкових факторів; n - кількість рядків (фактор А); k - кількість рядків (фактор В). Нульова гіпотеза може полягати в рівності ефектів стовпчиків між собою $H_{01}: \tau_1 = \tau_2 = ... = \tau_k = 0$, або рівності ефектів рядків між собою $H_{02}: \beta_1 = \beta_2 = ... = \beta_n = 0$. У першому випадку припускають відсутність впливу фактора В, а у другому − фактора А. Як і у випадку однофакторного дисперсійного аналізу розраховують дві оцінки дисперсії. При перевірці гіпотези H_{01} першу з них розраховують за формулою:

$$\sigma_1^2 = \frac{1}{(n-1)(k-1)} \sum_{i=1}^{n} \sum_{j=1}^{k} \left(x_{ij} - \langle x_i \rangle - \langle x_j \rangle + \langle x \rangle \right)^2 \qquad (3.24)$$

де $\langle x \rangle = \frac{1}{nk} \sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij}$ − загальне середнє за всіма спостереженнями, $\langle x_i \rangle = \frac{1}{k} \sum_{j=1}^{k} x_{ij}$ − середнє за і-м рядком, $\langle x_j \rangle = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$ − середнє за j-м стовпчиком. Дисперсія σ_1^2 пов'язана із впливом неврахованих нами факторів (залишкова дисперсія). Оцінка дисперсії σ_1^2 є незміщеною і не залежить від справедливості нульової гіпотези. Другу оцінку розраховують за формулою:

$$\sigma_2^2 = \frac{n}{k-1} \sum_{j=1}^{k} \left(\langle x_j \rangle - \langle x \rangle \right)^2 . \qquad (3.25)$$

Вона є незміщеною лише за умови справедливості нульової гіпотези. Чим більшою є різниця між результатами дії фактора В, тим більшим є значення, розраховане за формулою (3.25). Для перевірки спра-

ведливості гіпотези H01 необхідно розрахувати відношення дисперсій:

$$F=\frac{\sigma_2^2}{\sigma_1^2}=\frac{n(n-1)(k-1)\sum_{j=1}^{k}\left(\langle x_j\rangle-\langle x\rangle\right)^2}{(k-1)\sum_{i=1}^{n}\sum_{j=1}^{k}\left(x_{ij}-\langle x_i\rangle-\langle x_j\rangle+\langle x\rangle\right)^2}. \qquad (3.26)$$

Воно має F-розподіл Фішера з кількостями степенів вільності $(k-1)$ і $(n-1)(k-1)$. Нульову гіпотезу приймають на рівні значимості α, якщо $F < F_{1-\alpha}$, де $F_{1-\alpha} - \alpha$ - квантиль F-розподілу з відповідними кількостями степенів вільності. Для перевірки нульової гіпотези H02 можна використовувати формулу (3.26), у якої необхідно попарно поміняти місцями величини n і k, а також індекси і та j.

Розглянемо **приклад**. В наступній таблиці 3.3 маємо вибіркові дані про розривне навантаження пряжі, виготовленої на різних станках A1, A2, A3 із різної сировини B1, B2. Необхідно при рівні значимості α=0.05 вияснити , чи впливають на якість пряжі тип станка і вид сировини.

Таблиця 3.3

Дисперсійний аналіз за двома ознаками

	B1	B2	Середнє	Дисперсія
A1	10	50	30	225
A2	20	60	40	25
A3	30	100	65	400
Середнє	20	70	45	650
Дисперсія	1875	1875	3750	

Кількість ступенів вільності: k = 3; n = 2. Залишкова дисперсія: $\sigma_1^2 = 150$. Дисперсія за фактором А: $\sigma_A^2 = 650$. Дисперсія за фактором В: $\sigma_B^2 = 3750$. Загальна вибіркова дисперсія: $\sigma_0^2 = 5350$. Відношення дисперсій для фактора А становить: $F_A = \frac{\sigma_A^2}{\sigma_1^2} = 4.33$. Відношення дисперсій для фактора В становить: $F_A = \frac{\sigma_B^2}{\sigma_1^2} = 25$. Критичне значення $F_{1-\alpha}$ при рівні значущості $\alpha = 0.95$ дорівнює $F_{1-\alpha} = 18.51$. Для розрахунку $F_{1-\alpha}$ можна використати функцію пакету MS Excel FРАСПОБР(1-α, n-1, (n-1)(k-1)).

Висновки. Оскільки $F_A < F_{1-\alpha}$, вплив фактора А є несуттєвим. Оскільки $F_B > F_{1-\alpha}$, вплив фактора В є суттєвим. Оскільки $\frac{\sigma_B^2}{\sigma_0^2} = \frac{3750}{5350} = 0.70$, це означає, що 70% варіації ознаки x викликано впливом фактора В.

Якщо припущення, необхідні для застосування двофакторного дисперсійного аналізу не виконуються, то використовують непараметричний ранговий критерій Фрідмена (Фрідмена, Кендалла та Сміта), який не залежить від типу розподілу. Припускають лише, що розподіл величин ε_{ij} є однаковим і неперервним, а самі вони незалежні одна від одної. При перевірці нульової гіпотези $H_{01}: \tau_1 = \tau_2 = ... = \tau_k = 0$ вихідні дані подають у формі прямокутної матриці, у якої n рядків ві-

дповідають рівням фактора В, а k стовпців − рівням фактора А. Кожна комірка таблиці (блок) може бути результатом вимірювань параметрів на одному об'єкті або на групі об’єктів при сталих значеннях рівнів обох факторів. У цьому разі відповідні дані подають як середні значення певного параметра за всіма вимірюваннями або об'єктами досліджуваної вибірки. Для застосування критерію в таблиці вихідних даних необхідно перейти від безпосередніх результатів вимірювань до їх рангів. Ранжирування здійснюють за кожним рядком окремо, тобто величини x_{ij} впорядковують для кожного фіксованого значення і, отримуючи при цьому k значень відповідних рангів rij. Це дає можливість усунути вплив фактора В, значення якого для кожного рядка є однаковим. Обчислення критерію здійснюють за формулою:

$$S = \left[\frac{12}{nk(k+1)} \sum_{j=1}^{k} \left(\sum_{i=1}^{n} r_{ij} \right)^2 \right] - 3n(k+1). \qquad (3.27)$$

Якщо необхідно перевірити нульову гіпотезу $H_{02}: \beta_1 = \beta_2 = ... = \beta_n = 0$, то вихідні дані необхідно ранжувати за стовпчиками і повторити описану вище процедуру із заміною *n* на *k* і навпаки.

При справедливості нульової гіпотези і $n \to \infty$ S-статистика Фрідмена асимптотично наближається до статистики χ^2 з $k-1$ степенем вільності. Тому нульову гіпотезу можна прийняти на рівні значущості α, якщо $S < \chi^2_{1-\alpha}(k-1)$.

Критерій Пейджа (L-критерій) призначений для перевірки нульових гіпотез

$H_{01}: \tau_1 = \tau_2 = ... = \tau_k = 0$, або $H_{02}: \beta_1 = \beta_2 = ... = \beta_n = 0$,

проти альтернатив

$$\tau_1 \le \tau_2 \le ... \le \tau_k \text{ або, відповідно, } \beta_1 \le \beta_2 \le ... \le \beta_n,$$

у яких принаймні одна із нерівностей є строгою. Для впорядкованих альтернатив він є потужнішим за критерій Фрідмена. Значення критерію обчислюють за формулами:

$$L_1 = \sum_{j=1}^{k} jr_j, \qquad L_2 = \sum_{i=1}^{n} ir_i \tag{3.28}$$

де $r_j = \sum_{i=1}^{n} r_{ij}$, $r_i = \sum_{j=1}^{k} r_{ij}$. Для великих вибірок застосовують апроксимацію статистики Пейджа:

$$L_1^* = \frac{L - nk(k+1)^2/4}{\sqrt{n(k^3-k)^2/144(k-1)}}; \; L_2^* = \frac{L - nk(n+1)^2/4}{\sqrt{k(n^3-n)^2/144(n-1)}}, \tag{3.29}$$

які за умови справедливості відповідних нульових гіпотез підпорядковуються стандартному нормальному розподілу. У разі, коли в рядках вихідної таблиці є однакові значення, необхідно використовувати середні ранги. При цьому точність висновків буде тим гіршою, чим більшою є кількість таких збігів.

Q-критерій Кокрена використовують у випадках, коли групи однорідних суб'єктів піддаються впливам, кількість яких перевищує два, і для яких можливі два варіанти відгуків − умовно негативний (0) та умовно позитивний (1). Нульова гіпотеза полягає в рівності ефектів впливу. Значення критерію розраховують за формулою:

$$Q = \frac{(c-1)\left(c\sum_{j=1}^{c} T_j^2 - \left(\sum_{j=1}^{c} T_j\right)^2\right)}{c\sum_{i=1}^{r} T_i - \sum_{i=1}^{r} T_1^2} \qquad (3.30)$$

де $T_j = \sum_{i=1}^{r} x_{ij}$ (j=1,2,...,c) − суми стовпців, $T_i = \sum_{j=1}^{c} x_{ij}$ (i=1,2,...,r) - суми рядків, c − кількість стовпців (вибірок), r − кількість рядків (параметрів). Довірчий рівень визначається функцією розподілу χ^2 з кількістю степенів вільності $c-1$. Дисперсійний аналіз за двома ознаками (двофакторний дисперсійний аналіз) дає можливість визначити існування ефектів обробки, але не дає змоги встановити, для яких саме стовпців існує цей ефект. При вирішенні цієї проблеми застосовують метод множинних порівнянь Шеффе для зв'язаних вибірок. Значення критерію розраховують за формулою:

$$t = \frac{\left(\sum_{i=1}^{r} c_i \overline{x_i}\right)^2}{\frac{(r-1)S}{c}\sum_{i=1}^{r} c_i^2} \qquad (3.31)$$

де c_i $(i = 1,2,...,r)$ − константи, $S = \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij} - \frac{T^2}{rc}$ − залишковий середній квадрат, $T = \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}$ − загальна сума, c − кількість стовпців (вибірок), r − кількість рядків (параметрів). Довірчий рівень визначається функцією розподілу Фішера з параметрами $(r-1)$ та

$(r-1)(c-1)$ при дослідженні ефекту рядків і $(c-1)$ та $(r-1)(c-1)$ при дослідженні ефекту стовпців.

3.3. Функції розподілу, які найчастіше використовують при розрахунку критеріїв

При обчисленні критеріїв за функціями розподілу застосовують такі методи:

- точне обчислення критичних значень;
- пряме інтегрування функцій щільності розподілу;
- розклад у ряд Тейлора або Маклорена з наступним інтегруванням;
- кускова апроксимація елементарними функціями з наступним інтегруванням;
- апроксимація за допомогою нейронної мережі.

Пряме інтегрування використовують лише для дискретних розподілів. Відповідні алгоритми зазвичай потребують великого розрахункового часу і в багатьох випадках доцільнішою є апроксимація статистик критеріїв стандартними розподілами. Пряме інтегрування можна здійснювати за формулами трапецій, Сімпсона, методами Монте-Карло, Гаусса тощо. Нехай потрібно обчислити певну функцію розподілу, яка в загальному випадку може бути подана як

$$P(x) = \int_{-\infty}^{x} f(t)dt \,. \tag{3.32}$$

При цьому слід враховувати такі властивості:

- для симетричних відносно x=0 розподілів

$$P(x) = \frac{1}{2} + \int_0^x f(t)dt\,; \qquad (3.33)$$

- функція f(t) досить швидко спадає при збільшенні $|t|$, що дає змогу досить точно вказати відрізок, на якому в межах заданої похибки значення P(x) істотно відрізняється від нуля або одиниці;
- функція f(t) є неперервною і диференційованою нескінченну кількість разів, тому вона може бути розкладена у ряд Тейлора або Маклорена; з погляду часу роботи програми для досягнення високої точності це буде найшвидшим алгоритмом обчислення. Кускова апроксимація при обчисленні неперервних розподілів не забезпечує високої точності. Але у разі, коли достатньо отримати 2−3 правильних цифри після десяткової коми, відповідні алгоритми істотно перевищують інші методи за швидкістю обчислень.

Апроксимація за допомогою нейронних мереж забезпечує точність до 2−4 знаків після десяткової коми і застосовується, якщо формули для щільності розподілів є невідомими або досить складними. Згідно з визначенням для функції розподілу *F*, яка відображує значення x в імовірність α, зворотну функцію, що відображує α в x, називають зворотною функцією розподілу. При перевірці гіпотез їх застосовують, коли йдеться про визначення довірчого рівня або рівня значущості. Зазвичай зворотний розподіл обчислюють за методом ділення навпіл, що забезпечує достатню для більшості практичних застосувань швидкість. Функція щільності нормального розподілу має вигляд

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-a)^2}{2\sigma^2} \right],\ x \in (-\infty; +\infty), \qquad (3.34)$$

де a – середнє арифметичне, σ^2 – дисперсія. Шляхом введення нормованого відхилення $t = \dfrac{x-a}{\sigma}$ її перетворюють до функції щільності стандартного нормального розподілу

$$f(t) = \frac{1}{\sqrt{2\pi}} \exp\left[-t^2/2\right]. \tag{3.35}$$

Функцію стандартного нормального розподілу (функцію Лапласа)

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-y^2/2\right] dy \tag{3.36}$$

зазвичай розраховують за допомогою апроксимаційних формул, або розкладом у ряд. Якщо необхідно отримати підвищену точність результату (5–6 знаків) застосовують пряме інтегрування за методом Сімпсона. Щільність Г-розподілу (гамма-розподілу) визначається формулою

$$f(x) = \begin{cases} \dfrac{\beta^{\alpha}}{\Gamma(\alpha)} \quad x^{\alpha-1} e^{-\beta x}, x \geq 0, (\alpha, \beta > 0) \\ 0, x < 0, \end{cases} \tag{3.37}$$

де $\Gamma(\alpha)$ – гамма-функція.

Гамма функцією Ейлера називають функцію виду:

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t} dt (x > 0). \tag{3.38}$$

Вона задовольняє співвідношення $\Gamma(x+1) = x\Gamma(x)$. $\Gamma(1) = 1$. Звідси для цілих n маємо: $\Gamma(n+1) = n!\,(n = 1,2,...)$. Значення

$\Gamma(1/2)=\sqrt{\pi}$ дає змогу отримати значення також для будь-якого напівцілого значення аргументу. Для розрахунку функції при інших значеннях аргумента (як і у випадку інших функцій розподілу) можна використати один з математичних пакетів, або ж пакет Microsoft Excel.

Розподіл χ^2 є окремим випадком Г-розподілу з параметрами $\alpha = \mathrm{n}/2$, $\beta = 1/2$ (n − кількість степенів вільності). Іншим окремим випадком є розподіл Ерланга. Функцію розподілу χ^2 обчислюють за формулою

$$F_n(x) = 1 - P_n(x), \tag{3.39}$$

де

$$P_n(x) = \frac{1}{2^{n/2}\Gamma(n/2)} \int_x^{\infty} y^{\frac{n}{2}-1} e^{\frac{y}{2}} dy \tag{3.40}$$

− інтеграл імовірностей.

Щільність В-розподілу (бета-розподілу) визначається формулою

$$f(x) = \begin{cases} \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, x \in (0,1),\ \alpha,\beta > 0; \\ 0, x \in (-\infty,0] \cup [1,+\infty). \end{cases} \tag{3.41}$$

Функцією В-розподілу є

$$I_x(a,b) = \frac{1}{B(a,b)} \int_0^x t^{a-1}(1-t)^{b-1} dt, \tag{3.42}$$

де $B(a,b)$ − B-функція Ейлера, $a = n_2/2, b = n_1/2$. Для вибірок великого обсягу розрахунки здійснюють за асимптотичними формулами. Функція щільності F-розподілу має вигляд

$$f(x)=\begin{cases}\dfrac{\Gamma\left(\dfrac{m}{2}+\dfrac{n}{2}\right)}{\Gamma(m/2)\Gamma(n/2)}\left(\dfrac{m}{n}\right)^{m/2}\dfrac{x^{m/2}-1}{\left(1+\dfrac{m}{n}x\right)^{\frac{m+n}{2}}}, x\in(0,+\infty)\\ 0, x\in(-\infty,0],\end{cases};\ (3.43)$$

де $m>0$ − кількість степенів вільності чисельника, $n>0$ − кількість степенів вільності знаменника. Зазвичай її розраховують через функцію B-розподілу або шляхом розкладання в ряд. Щільність t-розподілу Стьюдента обчислюють за формулою

$$f(x)=\frac{\Gamma\left(\dfrac{\alpha+1}{2}\right)}{\sqrt{\alpha\pi}\,\Gamma(\alpha/2)}\left(1+\frac{x^2}{\alpha}\right), x\in(-\infty,+\infty), \alpha>0. \quad (3.44)$$

Відповідна функція розподілу має вигляд:

$$S_n(t)=\frac{1}{\sqrt{\pi n}}\frac{\Gamma\left(\dfrac{n+1}{2}\right)}{\Gamma(n/2)}\int_{-\infty}^{t}\left(1+\frac{u^2}{n}\right)^{\frac{n+1}{2}}du. \quad (3.45)$$

При $n\to\infty$ розподіл Стьюдента наближається до нормального. Зазвичай вважають, що його апроксимація нормальним розподілом є прийнятною при $n>30$. При $n=1$ t-розподіл Стьюдента називають розподілом Коші. Розподіл критичних значень G-критерію Кокрена визначається за точною формулою

$$g_{k,v,1-\alpha}=\frac{F}{(k-1)+F}, \quad (3.46)$$

де k – кількість вибірок, ν − обсяг кожної з них, α − рівень значущості, F − значення оберненої функції F-розподілу для ν та $(k-1)\nu$ степенів вільності й довірчого рівня $(1-\frac{\alpha}{k})$. При $k>2$, $\nu > 10$ можна використовувати також апроксимаційну формулу

$$g_{k,\nu,1-\alpha} = \frac{2x}{\nu(2k-1)-2+x+\frac{(2-\nu)(2+\nu+x)+2x^2}{6(\nu(2k-1)-2)}}, \qquad (3.47)$$

де x − значення оберненої функції χ^2 − розподілу для ν степенів вільності і довірчого рівня $\left(1-\frac{\alpha}{k}\right)$.

Основним типом розподілів типу Колмогорова − Смірнова є λ-розподіл. Його критичне значення можна обчислювати за точною формулою

$$K(\lambda) = \begin{cases} \sum_{i=-\infty}^{+\infty} (-1)^i e^{-2i^2\lambda^2}, \lambda > 0; \\ 0, \lambda \le 0. \end{cases} \qquad (3.48)$$

Ряд, що стоїть у формулі, швидко збігається, і в більшості випадків достатньо обмежитися його 11 членами; для підвищення точності при малих значеннях λ кількість членів ряду можна збільшити до 31.

3.4. ЗАВДАННЯ ДО РОЗДІЛУ 3

Завдання 1. Виконати завдання згідно варіантів та зробити висновок (економічну інтерпретацію) про вплив фактора на досліджувану ознаку.

Варіант 1

У результаті проведення досліду з метою з'ясування впливу чорного пару на врожайність пшениці з ділянки в 9 га (3 га були під чорним паром; 3 га – під картоплею; 3 га під кормовими травами) дістали такі результати:

Фактор	Врожайність, ц/га
Чорний пар	26,6; 26,6; 30,6
Площа під картоплею	24,3; 25,2; 25,2
Площа під кормовими травами	26,6; 28,0; 31,0

За рівень значущості береться $\alpha = 0{,}01$.

Варіант 2

Експериментально досліджувався вплив на зносостійкість колінчатих валів технології їх виготовлення. Застосовувалися 4 технології виготовлення валів, отже фактор А має 4 рівні. Одержані результати наведено в таблиці:

Ступінь впливу фактора А	Кількість відпрацьованих місяців
A_1	9;8;10;12
A_2	10;12;11;8
A_3	8;16;10;18
A_4	9;18;10;8

При рівні значущості $\alpha = 0{,}01$ перевірити вплив технологій на зносостійкість валів.

Варіант 3

Для перевірки впливу методики навчання виробничим навикам на якість підготовки із випускників виробничо-технічного училища навмання вибирають 4 групи учнів, які після закінчення навчання за різними методиками тестують на кількість виготовлених однотипних деталей протягом робочої зміни. Результати тестування наведено в таблиці:

Ступінь впливу фактора А (методики)	Кількість виготовлених деталей за робочу зміну
A_1	60,80,75,80,85,70
A_2	75,66,85,80,70,80,90
A_3	60,80,65,60,86,75
A_4	95,85,100,80

При рівні значущості $\alpha = 0{,}05$ з'ясувати вплив методики навчання на якість підготовки учнів.

Варіант 4

Досліджується залежність урожайності пшениці від сорту пшениці, яких є 4. Результати дослідження наведено в таблиці:

Ступінь впливу фактора А (сорт пшениці)	Урожайність, ц/га
A_1	28,7;26,7;21,6;25,0;28,2
A_2	24,5;28,5;27,7;28,7;32,5
A_3	23,2;24,7;20,0;24,0;24,0
A_4	29,0;28,7;20,5;28,0;27,0

При рівні значущості $\alpha = 0{,}01$ з'ясувати вплив сортності пшениці на її врожайність.

Варіант 5

Стальні болти з різною добавкою компоненти А в сталі, з якої вони виготовлялися, були піддані випробуванням на міцність. Результати цих випробувань наведено в таблиці:

Ступінь впливу фактора А (відсоткова добавка)	Міцність, $кг / мм^2$
A_1	25;28;20;22
A_2	29;22;21;18
A_3	19;25;30;22
A_4	18;30;24;20

При рівні значущості $\alpha = 0{,}01$ з'ясувати вплив добавки компоненти на міцність болта.

Варіант 6

Електролампочки напругою 220 В виготовляються на трьох заводах із використанням різних технологій. З кожної партії, що надходила в науково – дослідний інститут від різних заводів, навмання брали 4 електролампочки і піддавали їх випробуванням на тривалість горіння.

Результати цих випробувань наведено в таблиці:

Ступінь впливу фактора А (технології виготовлення)	Тривалість горіння, год
A_1	90;85;105;110;95
A_2	80;110;115;90;105
A_3	75;120;110;90;85

При рівні значущості $\alpha = 0{,}01$ з'ясувати вплив технологій виготовлення на тривалість горіння лампочок.

Варіант 7

Рейтинг лівих партій, що вимірювався у навмання вибраних шести районах на Заході України, у центральній її частині і на Сході України дав такі результати:

Ступінь впливу фактора А	Рейтинг, %
A_1 (західні райони)	14,5;5,6;23,8;6,4;26,2;14,5
A_2 (центральні райони)	22,5;12,2;24,8;16,8;11,9;26,6
A_3 (східні райони)	13,4;20,8;30,8;20,8;6,4;12,3

При рівні значущості $\alpha = 0{,}001$ з'ясувати, чи впливає істотно регіон України на рейтинг лівих партій.

Варіант 8

На дослідних ділянках, кожна з яких має площу 6 га, досліджувалась залежність урожайності пшениці від внесення в ґрунт добрив A_1, A_2, A_3, A_4. Результати цих експериментів наведено в таблиці:

Ступінь впливу фактора А (тип добрива)	Урожайність, ц/га
A_1	25,6;36,2;22,8;30,2;32,5;28,4
A_2	28,5;40,6;42,8;36,4;22,4;29,6
A_3	24,4;38,6;48,4;50,2;28,4;22,8
A_4	29,5;52,8;24,2;22,8;56,2;48,4

При рівні значущості $\alpha = 0{,}01$ з'ясувати вплив типу добрива, що вноситься в ґрунт, на урожайність пшениці.

Завдання 2. При рівні значущості $\alpha = 0{,}05$ визначити вплив факторів А і В та їх сумісний вплив на досліджувану ознаку. Зробити відповідні висновки.

Варіант 1

Досліджується вплив на зносостійкість деталей таких факторів: А-матеріал для виготовлення деталей (застосували три види сталі) і В-технологія виготовлення деталей (дві технології). Результати експерименту наведено в таблиці:

Фактор В	Фактор А		
	A_1	A_2	A_3
B_1	10;7;8;6;12;8; 11;10;14;13	8;14;6;10;16; 14;13;12;11;15	15;12;11;9;8;13; 11;12;16;14
B_2	12;13;6;9;8;11; 10;10;13;17	11;12;12;16;13; 8;10;9;8;15	13;12;14;8;6;8; 16;12;14;16

Варіант 2

Досліджується вплив на врожайність ячменю таких факторів А: посів здійснюється після чорного пару - A_1 ; після коренеплодів – A_2; після колосових культур – A_3 ; В – сортність ячменю(три сорти). Результати досліджень наведено в таблиці :

Фактор В	Фактор А		
	A_1	A_2	A_3
B_1	34,2;30,6;36, 8;35;2,5;34,2; 33,4;36	42,5;40,4;44,6; 46,8;39,4;38,6; 45,8;49,3	44,2;46;45,6; 48;49,3;45,8; 42,3;40,8;41,4;40
B_2	32,5;30,4;39, 4;40,3;36,4; 38,9;42;	30,3;35,3;36,8; 40,5;28,4;33, 2;39,1;26,9;	40,3;45;46,8; 30,2;48,8;50,2; 39;38,5
B_3	33,3;34,8; 39,2;35;32,4; 34;39,8;40,8	30,4;36;40,5; 44,4;30,8;42,5; 46;33,5	32,3;29,8; 34,3;42;34,8; 31,6;40;29,6

Варіант 3

Досліджується вплив на міцність чавуну двох факторів : А – вміст кремнію в чавуні , а саме A_1 – 0,24%; A_2 – 0,42%; A_3 – 0,52%; В – температурний режим плавлення (два режими). Результати досліджень наведено в таблиці :

Фактор В	Фактор А		
	A_1	A_2	A_3
B_1	40,2; 40,8; 38,2; 39,6;42,4; 44,5; 40,1; 38,8	42,5; 43,4; 44,5; 46,4; 40,1; 36,5; 40,3; 41,8;38; 43,5	49,2; 50,2; 48,4; 50; 52,5; 38,4; 49,8; 50,4;51,8; 49
B_2	33,4; 36,5; 34,4; 40,2;42; 30,2; 31,8; 35,5; 34; 41,8	31,6; 33,4; 38,4; 35; 38,9; 29,5; 43; 28,4; 30,6; 32,9	29,3; 35,6; 36; 26,8; 38; 28,5; 30,6; 40,2; 33,3

Варіант 4

Досліджується вплив на врожайність кукурудзи двох факторів : А – внесення добрив у ґрунт (три добрива); В – глибина поливу землі (три глибини поливу). Результати досліджень наведено в таблиці :

Фактор В	Фактор А		
	A_1	A_2	A_3
B_1	30.2;30.8;31.6; 32;32.6;28.9;30.5; 32.6;33	28.4;29.9;30.6; 44.3;36.2;42.3; 28.2;26.5;34.3;26.5	40.2;42.3;42.7; 43.5;44;36.8;38.9; 45.3;46.2;45.4
B_2	44.2;42.8;43.7; 46.5;46.9;40.5;45.6; 38.4;32.5;44.6	42.4;43.5;40.6; 36.8;40;36.4; 38.5;43.2;34.6;39.8	42.3;43.4;45.2; 44;36.5;29.8;25.4; 43.2;45;46.8
B_3	40.2;36.4;36.9;41.8; 40.4;34.8;38.5; 38.6;42.4	38.5;33.4;30.2; 29.4;40.1;26.2; 25.4;44.1;30.6;34.5	43.2;44.5;39.5; 32.5;45;40.8;36.3; 43.5;47.8;49

Варіант 5

Досліджується вплив факторів А і В на число виготовлених втулок зі ста взятих, які відповідають нормам стандарту: А- використали дві технології виготовлення; В - заготовки надходили із трьох заводів. Результати досліджень наведено в таблиці:

Фактор В	Фактор А	
	A_1	A_2
B_1	90;88;90;96;98; 76;80;95;85;80	100;99;82;98;95;80; 96;95;99;191;89;90
B_2	79;88;92;76;80; 83;85;90;96;75	81;82;100;98;89;85; 96;98;75;97
B_3	82;78;75;79;80; 81;86;89;75;90	80;86;90;91;78;76; 75;82;73;82

Варіант 6

Досліджується вплив факторів А і В на продуктивність праці підприємства певної галузі промисловості: А- фондозабезпеченість (три рівні); В- рівень оплати праці робітникам (два рівні). Результати досліджень наведено в таблиці:

Фактор В	Фактор А		
	A_1	A_2	A_3
B_1	14,85;11,94;10,5; 12,3;15,62;13,2; 10,62;12,82;11,48; 13,5	6,42;5,23;4,96;5,6; 9,82;10,23;12,44; 16,5;5,41;6,32	7,82;9,63;12,92; 10,82;9,36;5,11; 13,52;14,2;8,96; 9,92
B_2	12,5;13,8;14,9; 12,6;10,85; 11,96;12,6;13,42; 16;17,2	10,2;10,85;12,34; 11,95;12,4;14,92; 9,86;9,62;8,36; 13,62	13,62;12,55;14,7; 13,25;14,66;8,35; 10,96;11,62;6,12; 15,66

Варіант 7

Експериментально досліджувався вплив на зносостійкість деталей факторів А і В: фактор А – тип сталі (три типи) ; фактор В – технологія виготовлення (дві технології).

Результати досліджень наведено в таблиці:

Фактор В	Фактор А		
	A_1	A_2	A_3
B_1	10;8;6;9;5; 12;5;8;10;11	8;12;12;10;11; 6;10;10;9;5	15;14;14;8;8; 13;10;11;9;6
B_2	12;9;9;6;6; 5;10;8;8;9	12;13;13;14; 15;8;9;10;11;11	13;13;10;5;5; 10;15;14;14;10

Питання для самоконтролю

7. Що таке дисперсійний аналіз?
8. Які задачі розв'язує дисперсійний аналіз?
9. Які види дисперсійного аналізу знаєте?
10. В чому полягають однофакторний та двофакторний дисперсійні аналізи?
11. Назвіть основну модель однофакторного дисперсійного аналізу.
12. В чому полягає ранговий однофакторний аналіз Краскела − Уолліса?
13. Які ще критерії відомо для перевірки гіпотез щодо моделей однофакторного дисперсійного аналізу?
14. Назвіть основну модель двофакторного дисперсійного аналізу.
15. Які критерії відомо для перевірки гіпотез щодо моделей двофакторного дисперсійного аналізу?
16. Які функції розподілу найчастіше використовують при розрахунку критеріїв?

4. КОРЕЛЯЦІЙНИЙ АНАЛІЗ

Кореляцією (**кореляційним зв'язком**) між випадковими величинами (ознаками) називають наявність статистичного або імовірнісного зв'язку між ними. При цьому закономірна зміна певних ознак призводить до закономірної зміни середніх значень інших, пов'язаних з ними ознак. **Кореляційним аналізом** називають сукупність методів виявлення кореляційного зв'язку. Наявність кореляційного зв'язку не означає існування причинно-наслідкового зв'язку між досліджуваними ознаками. Зокрема, вона може бути зумовлена наявністю інших невідомих факторів, що впливають на досліджувані ознаки.

Розглядають також протилежну гіпотезу про відсутність зв'язку між досліджуваними величинами. Нехай ознака А має r рівнів $A_1,A_2,..A_r$, а ознака В − s рівнів $B_1,B_2,...,B_s$. Ці ознаки вважають **незалежними**, якщо події "ознака А набуває значення A_i" та "ознака В набуває значення B_j" є незалежними для усіх можливих пар i, j, тобто:

$$P(A_i, B_j) = P(A_i)P(B_j). \quad (4.1)$$

Це можна сформулювати в інший спосіб: ознаки є незалежними, якщо значення ознаки А не впливає на ймовірності реалізації можливих значень ознаки В:

$$P(B_j / A_i) = P(B_j), \forall(A_i, B_j). \quad (4.2)$$

Кореляційний аналіз здійснюють на початковому етапі вирішення всіх трьох основних проблем статистичного аналізу даних. У проблемі статистичного аналізу залежностей він дає змогу встановити сам факт існування залежності та оцінити ступінь тісноти зв’язку між змінни-

ми. У проблемах класифікації даних і зменшення розмірності досліджуваного простору ознак за допомогою кореляційного аналізу отримують вихідну інформацію у вигляді коваріаційних та кореляційних матриць та інших характеристик парних порівнянь.

Загальна методика перевірки наявності зв'язку між ознаками передбачає три основних етапи: визначення типу даних; перевірка гіпотези про відсутність зв'язку і, при її відхиленні, оцінка сили зв'язку. Тип вихідних даних істотно впливає на вибір методів і критеріїв, які можна застосовувати на наступних етапах аналізу.

Універсальною характеристикою ступеня тісноти зв'язку між кількісними ознаками є **коефіцієнт детермінації**. Вибірковий коефіцієнт детермінації певної ознаки y по вектору незалежних ознак $X = \left(x^{(1)}, x^{(2)}, ..., x^{(p)}\right)$ можна розрахувати як

$$K_d\left(y; X\right) = 1 - \frac{s_\varepsilon^2}{s_y^2}, \tag{4.3}$$

де вибіркове значення **дисперсії ознаки** y обчислюють за формулою

$$S_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \ , \tag{4.4}$$

а вибіркове значення **дисперсії нев'язок** ε обчислюють за однією з таких формул:

$$S_\varepsilon^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}\left(X_i\right)\right)^2, \tag{4.5}$$

де $\hat{f}\left(X_i\right)$ є статистичною оцінкою невідомого значення функції регресії $f(X)$ у точці X_i, або

$$S_\varepsilon^2 = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{\nu_i}\left(y_{ij} - \bar{y}_{y^*}\right)^2, \qquad (4.6)$$

де ν_j − кількість даних, що потрапили до і-го інтервалу групування, y_{ji} −значення і-го спостереження досліджуваної ознаки, що потрапило до j-го інтервалу, $\bar{y}_{j^*} = \frac{\sum_{i=1}^{\nu_j} y_{ji}}{\nu_j}$ − її середнє значення за спостереженнями, які потрапили до j-го інтервалу. Формулу (4.5) застосовують у випадку, коли за результатами попереднього аналізу встановлено, що умовна дисперсія $D(\varepsilon|X) = \sigma_\varepsilon^2$=const , тобто не залежить від x. Формулу (4.6) використовують, якщо ця умова не виконується, а також у всіх випадках, коли обчислення здійснюють за згрупованими даними.

Величина коефіцієнта детермінації може змінюватися в межах від нуля до одиниці й відображає частку загальної дисперсії досліджуваної ознаки, яка зумовлена зміною функції регресії f(**X**). При цьому нульове значення коефіцієнта детермінації відповідає відсутності будь-якого зв'язку, а його рівність одиниці − наявності строго функціонального (однозначного) зв'язку.

Інші поширені характеристики ступеня тісноти зв'язку між ознаками можна розглядати як окремі випадки коефіцієнта детермінації, отримані для конкретних типів зв'язку.

Розрізняють парні та частинні кореляційні характеристики. Парні характеристики розраховують за результатами вимірювань тільки досліджуваної пари ознак. Тому вони не враховують опосередкованого або сумісного впливу інших ознак. Частинні характеристики є очищеними від випливу інших факторів, але для їх розрахунку необхідно

мати вихідну інформацію не тільки про досліджувані ознаки, але й про всі інші, вплив яких необхідно усунути.

Для кількісних ознак найширше застосовуються коефіцієнти Пірсона і Фехнера. Коефіцієнт кореляційного відношення Пірсона (парний коефіцієнт кореляції, **вибірковий коефіцієнт кореляції**, коефіцієнт Бравайса-Пірсона) вимірює ступінь лінійного кореляційного зв'язку між кількісними скалярними ознаками. Його розраховують за формулою:

$$r = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}. \qquad (4.7)$$

Його застосування як міри зв'язку є обґрунтованим лише за умови, що сумісний розподіл пари ознак є нормальним. Значення -1 та +1 відповідають строгій лінійній функціональній залежності, яка в першому випадку є спадною, а у другому – зростаючою. Для функціональної залежності y=const коефіцієнт кореляції, як видно з наведеної формули, є невизначеним, оскільки в цьому разі знаменник дорівнює нулю. Чим ближчим є значення коефіцієнта кореляції до -1 та +1, тим більш обґрунтованим є припущення про наявність лінійного зв'язку. Наближення його значення до нуля свідчить про відсутність лінійного зв'язку, але не є доказом відсутності статистичного зв'язку взагалі. На рис.4.1 показано дві кореляційні залежності. Для обох серій є очевидним існування статистичного зв'язку між параметрами *x* та *y*. Але коефіцієнти кореляції для них дорівнюють, відповідно, $r_1=0.995$ і $r_2=0.006$. Близькість коефіцієнта кореляції до нуля для другої серії свідчить не про відсутність зв'язку, а про його нелінійність. Для порі-

вняння, коефіцієнти детермінації для тих самих серій дорівнюють приблизно 0.98 та 1.00.

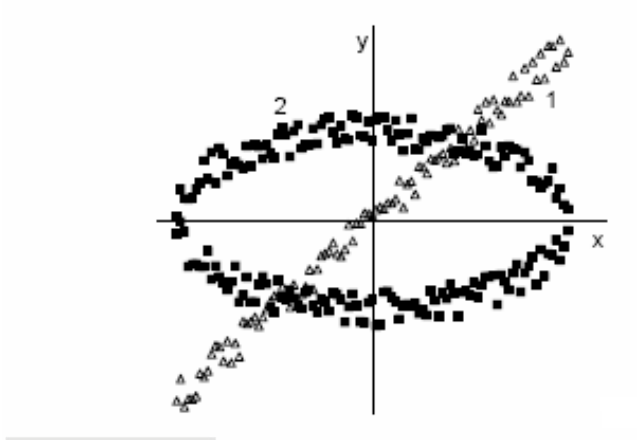

Рис.4.1. Порівняння ступеня кореляції лінійного та нелінійного зв'язку

При дослідженні багатовимірних сукупностей випадкових величин із коефіцієнтів кореляції, обчислених попарно між ними, можна побудувати квадратну симетричну **кореляційну матрицю** з одиницями на головній діагоналі. Вона є основним елементом при побудові багатьох алгоритмів багатовимірної статистики, наприклад у факторному аналізі. Довірчий інтервал вибіркової оцінки коефіцієнта кореляції для двовимірної нормальної генеральної сукупності.

$$\mathrm{r} \in \left[\tanh\left(z(r) - \frac{N_{\frac{1+p}{2}}}{\sqrt{n-3}} \right); \tanh\left(z(r) + \frac{N_{\frac{1+p}{2}}}{\sqrt{n-3}} \right) \right], \qquad (4.8)$$

де n-обсяг вибірки, $N_{\frac{1+p}{2}}$-квантиль нормального розподілу, p - стандартне значення довірчого рівня, z(r) - z-перетворення Фішера вибіркового коефіцієнта кореляції r.

Коефіцієнт кореляції можна застосувати для перевірки гіпотези про значущість. Для нормально розподілених вихідних даних величину вибіркового коефіцієнта кореляції вважають значуще відмінною від нуля, якщо виконується нерівність:

$$r^2 > \left[1 + (n-2)/t_\alpha^2\right]^{-1}, \qquad (4.9)$$

де t_α - критичне значення t-розподілу з $(n-2)$ степенями вільності.

Статистика $\sqrt{n-1}r$ має r - розподіл із щільністю:

$$\varphi_{r(n)}(r) = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-2}{2}\right)} \left(1 - r^2\right)^{\frac{n-4}{2}} \quad (-1 < r < 1). \qquad (4.10)$$

Для великих вибірок статистика $\sqrt{n-1}r$ наближається до стандартного розподілу.

Як уже зазначалося вище, близькість коефіцієнта кореляції до нуля в загальному випадку не є доказом незалежності ознак. Але можна довести, що в разі, коли сумісний розподіл випадкових величин (α, β) є нормальним, рівність r = 0 свідчить про статистичну незалежність α і β.

У разі, коли між двома наборами ознак існує нелінійний зв'язок, для оцінки ступеня його тісноти часто використовують **кореляційне відношення**. Це можливо, якщо щільність розміщення емпіричних точок на координатній площині дає можливість їх групування за однією із змінних і підрахунку часткових середніх значень другої змінної для кожного інтервалу. Тоді кореляційне відношення залежної змінної y за незалежною змінною x можна розрахувати за формулою:

$$\rho_{yx}^2 = \mathrm{S}_{\mathrm{y(x)}}^2 / \mathrm{S}_{\mathrm{y}}^2, \qquad (4.11)$$

де

$$\mathrm{S}_{\mathrm{y(x)}}^2 = \frac{1}{n}\sum_{j=1}^{s} \nu_i \left(\bar{y}_{j^*} - \bar{y}\right)^2; \quad S_y^2 = \frac{1}{n}\sum_{j=1}^{s}\sum_{i=1}^{\nu_i} \left(y_{ji} - \bar{y}\right)^2;$$

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{s} \nu_j \bar{y}_{j^*}; \quad \bar{y}_{j^*} = \left(\sum_{i=1}^{\nu_i} y_{ij}\right) / \nu_j,$$

n – обсяг вибірки, s – кількість інтервалів групування за віссю абсцис, ν_i - кількість точок, що потрапляють до і-го інтервалу.

Кореляційне відношення може змінюватися в інтервалі від нуля до одиниці. Із рівності $\rho_{yx} = 1$ випливає наявність строго функціонального зв'язку між досліджуваними ознаками, і, навпаки, однозначний функціональний зв'язок між ними свідчить про те, що $\rho_{yx} = 1$. За відсутності зв'язку $\rho_{yx} = 0$, і, навпаки, коли $\rho_{yx} = 0$, це означає , що для всіх інтервалів групування $\bar{y}_{j^*} = \bar{y}$, тобто часткові середні $\bar{y}_{j^*}$ не залежать від x.

На відміну від коефіцієнта кореляції, кореляційне відношення не є симетричним: у загальному випадку $\rho_{yx} \neq \rho_{xy}$. Більше того, можливі ситуації, коли один із цих коефіцієнтів дорівнює нулю, а другий – одиниці. Зокрема, це може спостерігатися для парних функцій за умови, що функція розподілу значень незалежної змінної є симетричною відносно нуля. Для даних, що наведені на рис.4.1, кореляційне відношення першої серії дорівнює приблизно 0,98 і в межах похибки обчи-

слень збігається з коефіцієнтами детермінації і кореляції. Для другої серії $\rho_{yx} \approx 0.63$ і $\rho_{xy} \approx 0.72$.

Можна довести, що кореляційне відношення збігається із модулем коефіцієнта кореляції між тими самими змінними за наявності лінійного зв'язку, а також за відсутності зв'язку. В інших випадках воно перевищує модуль коефіцієнта кореляції. Це дає можливість використовувати їх різницю як характеристику ступеня відхилення зв'язку від лінійності. Для цього розраховують величину:

$$v^2 = \frac{(n-k)\left(\rho_{yx}^2 - r^2\right)}{(k-2)\left(1-\rho_{yx}^2\right)}, \qquad (4.12)$$

де n – кількість емпіричних точок, k – кількість невідомих параметрів моделі. Ця величина приблизно підпорядковується F-розподілу з параметрами $s-2$ та $n-s$. Якщо розраховане за формулою (4.12) значення перевищує точку v_α^2 розподілу $F(s-2, n-s)$, то гіпотезу про лінійний зв'язок вважають необґрунтованою з імовірністю похибки α.

Коефіцієнт кореляції Фехнера розраховують за формулою:

$$r_F = \frac{C-H}{C+H} = \frac{2C-n}{n} = \frac{2C}{n} - 1, \qquad (4.13)$$

де С- кількість збігів знаків відхилень варіант від відповідних середніх, H – кількість знаків, що не збігаються. Значення коефіцієнта Фехнера можуть змінюватися в межах від -1 до +1. Як і в попередньому випадку, він показує наявність лінійного зв'язку: чим ближчим до одиниці за модулем є значення коефіцієнта, тим сильнішим є зв'язок. Малі значення абсолютної величини коефіцієнта свідчать про відсутність лінійного зв'язку, але цього недостатньо для твердження про

відсутність будь-якого зв'язку взагалі. Зокрема, для серій, що наведені на рис.4.1, значення коефіцієнта Фехнера дорівнюють відповідно, $r_{F1} = 0.941$ і $r_{F2} = -0.010$. Застосування для обчислення коефіцієнта лише кількості збігів або незбігів знаків відхилень від середніх значень можна розглядати як зведення первинної кількісної шкали до номінальної, що має призвести до втрати частини корисної інформації. Тому цей критерій застосовується досить рідко.

Коваріацією називають змішаний момент другого порядку. На відміну від інших показників, що характеризують наявність статистичного зв'язку, коваріація не є безрозмірною величиною. Її розраховують за формулою:

$$Cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}). \qquad (4.14)$$

Коваріація вибірки із самою собою є дисперсією. З наведеної формули можна отримати корисне співвідношення для коефіцієнта кореляції Пірсона

$$r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y},$$

де σ_X, σ_Y - середні квадратичні відхилення вибірок.

При аналізі багатовимірних вибірок часто застосовують коваріаційні матриці:

$$C = \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1n} \\ c_{21} & c_{22} & \dots & c_{2n} \\ \dots & \dots & \dots & \dots \\ c_{n1} & c_{n2} & \dots & c_{nn} \end{pmatrix}, \qquad (4.15)$$

де $c_{ij} = \mathrm{cov}(x_i, x_j)$. Діагональні елементи цієї матриці є дисперсіями $c_{ij} = \sigma^2(x_i)$ відповідних рядів спостережень. Коваріаційна матриця є симетричною, тобто $c_{ij} = c_{ji}$.

Для лінійно зв'язаних один з одним випадкових векторів X та Y має місце співвідношення:

$$C_y = TC_x T^T, \tag{4.16}$$

де

$$T = \begin{pmatrix} \frac{dy_1}{dx_1} & \frac{dy_1}{dx_2} & \cdots & \frac{dy_1}{dx_n} \\ \frac{dy_2}{dx_1} & \frac{dy_2}{dx_2} & \cdots & \frac{dy_2}{dx_n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{dy_m}{dx_1} & \frac{dy_m}{dx_2} & \cdots & \frac{dy_m}{dx_n} \end{pmatrix}.$$

Під **ранговою кореляцією** розуміють статистичний зв'язок між порядковими ознаками. Вихідні дані зазвичай подають у вигляді таблиці 4.1.

Таблиця.4.1

Таблиця оцінки рангової кореляції

Порядковий номер об'єкта	Порядковий номер досліджуваної ознаки						
	0	1	2	…	K	…	P
1	x_{10}	x_{11}	x_{12}	…	x_{1k}	…	x_{1p}
2	x_{20}	x_{21}	x_{22}	…	x_{2k}	…	x_{2p}
…	…	…	…	…	…	…	…
I	x_{i0}	x_{i1}	x_{i2}	…	x_{ik}	…	x_{ip}
…	…	…	…	…	…	…	…
n	x_{n0}	x_{n1}	x_{n2}	…	x_{nk}	…	x_{np}

У цієї таблиці елемент x_{ik} дає ранг і-го об'єкта за k-ю властивістю.

Завданнями аналізу в цьому разі можуть бути: вивчення структури досліджуваних об'єктів; перевірка сукупної узгодженості ознак та умовне ранжування об'єктів за ступенем тісноти зв'язку кожної з них з іншими ознаками; побудова єдиного групового впорядкування об'єктів (задача регресії на порядкових змінних).

У першому випадку кожну послідовність впорядкованих за k-ю ознакою n об'єктів подають як точку $X^{(k)} = \left(x_1^{(k)}, x_2^{(k)}, ..., x_n^{(k)}\right)$, $k = 0,1,..., p$ у n-вимірному просторі. Найхарактернішими типами структур є такі.

1. Аналізовані точки рівномірно розкидані по всій області їх можливих значень. Це означає відсутність будь-якого зв'язку між досліджуваними ознаками.

2. Частина точок утворює ядро (кластер) із точок, що розташовані близько одна до одної, а інші випадково розкидані навколо цього ядра. Це відповідає існуванню підмножини узгоджених ознак.

3. Аналізовані точки утворюють декілька кластерів, розташованих відносно далеко один одного. Це означає, що існує істотний статистичний зв'язок між ознаками, які належать до однієї і тієї самої підмножини, і не існує значущого зв'язку між ознаками, які належать до різних підмножин.

Прикладом завдання другого типу є визначення узгодженості думок групи експертів з наступним впорядкуванням їх за рівнем компетентності. Для цього розраховують коефіцієнти конкордації для різних сукупностей досліджуваних змінних.

Розв'язання завдань третього типу зводиться до побудови такого впорядкування, яке б у певному значенні було б найближчим до кожного з наданих впорядкувань досліджуваних ознак. Для цього часто застосовують середнє арифметичне або медіану наявних базових рангів. Це дозволяє розглядати цю задачу як задачу найкращого у певному розумінні відновлення невідомого ранжування за наявними емпіричними даними, що зумовлює можливість її розгляду як задачі регресії.

Коефіцієнт рангової кореляції Спірмена (показник кореляції рангів Спірмена, коефіцієнт кореляції рангів) використовують, якщо досліджується лінійний зв'язок між рядами даних, виміряними за порядковою шкалою. Його можна застосувати також і для кількісних даних, але, як правило, це буває недоцільним. У найпростішому випадку досліджувані об'єкти класифікуються за двома ознаками. Наприклад, ми можемо спочатку впорядкувати групу учнів за їх здібностями до математики, а потім – до іноземних мов. Місця, які і-й учень займе в обох списках, будуть його рангами r_i та s_i. Якщо досліджувані ознаки взаємопов'язані, то послідовність рангів $r_1,r_2,\ldots,r_n$ певною мірою впливає на послідовність рангів $s_1,s_2,\ldots,s_n$.

Ступінь близькості двох послідовностей відображає величина:

$$S_\rho = \sum_{i=1}^{n} \left(r_i - s_i\right)^2 . \qquad (4.17)$$

Якщо для нумерації об'єктів попередньо впорядкувати їх за однією із ознак, наприклад за зростанням рангів r_i, то формула (4.17) може бути записана так:

$$S_\rho = \sum_{k=1}^{n} \left(k - s_k\right)^2 . \qquad (4.18)$$

Величина S_ρ набуває найменшого можливого значення $S_\rho = 0$ тоді й тільки тоді , коли послідовності повністю збігатимуться. Найбільше можливе значення $S_\rho = \frac{1}{3}\left(n^3 - n\right)$ відповідає випадку, коли послідовності є повністю протилежними, тобто для будь-яких i, j з нерівності $r_i > r_j$ випливає $s_i < s_j$ і послідовності рангів першої ознаки $r_i = \{1,2,...,n\}$ відповідає послідовності рангів другої $s_i = \{n, n-1,...,1\}$. Величина S_p не може бути мірою зв'язку, оскільки на її значення впливає кількість пар варіант досліджуваних рядів n.

Як міру зв'язку можна використовувати **коефіцієнт рангової кореляції Спірмена**, значення якого розраховують за формулою:

$$\rho_s = 1 - \frac{6\left(S_\rho + B_x + B_y\right)}{n^3 - n}, \qquad (4.19)$$

де B_x, B_y - поправки на об'єднання рангів у відповідних рядах, які обчислюють за формулою:

$$B_i = \frac{1}{12} \sum_{i=1}^{m} n_i\left(n_i^2 - 1\right). \qquad (4.20)$$

де m – кількість груп об'єднаних рангів у вибірці, n_i – кількість рангів в і – й групі .

Значення коефіцієнта можуть змінюватися в межах від -1 до +1, при цьому -1 відповідає повній протилежності послідовностей рангів, а +1 – їх повному збігу. При $B_i = 0$ кореляція відсутня.

Розглянемо **приклад**. У наступній таблиці 4.2 наведені дані про підприємства регіону, які виставили свої акції на чековий аукціон. Використовуючи коефіцієнт Спірмена необхідно визначити залежність між величиною статутного капіталу X і кількістю виставлених акцій Y.

Таблиця 4.2

Розрахунок коефіцієнта рангової кореляції Спірмена

Номер підприємства	Уставний капітал X грн	Кількість виставлених акцій Y	Ранг r_i	Ранг s_i	$(r_i - s_i)^2$
1	410200	1563	1	1	0
2	295400	856	2	4	4
3	281300	815	3	5	4
4	262500	616	4	8	16
5	235000	682	5	6	1
6	226400	661	6	7	1
7	179500	495	7	9	4
8	175100	858	8	3	25
9	170000	467	9	10	1
10	160500	930	10	2	64
					$\Sigma = 120$

Згідно із формулою (4.19) маємо (об'єднання рангів відсутнє): $\rho_s = 1 - \frac{6 \cdot 120}{1000 - 10} = 0.27$. **Висновок**: зв'язок між статутним капіталом та кількістю виставлених на продаж акцій є слабким.

Показник рангової кореляції Спірмена можна застосувати як показник некорельованості вибірок. У цьому разі розраховують величину:

$$t_p = \sqrt{n-2}\,\frac{\rho_s}{\sqrt{1-\rho_s^2}}. \qquad (4.21)$$

Для великих вибірок (n>50) критичні значення мають розподіл Стьюдента з (n-2) степенями вільності. Статистика $\sqrt{n-1}\rho_s$ для великих вибірок наближається до стандартного нормального розподілу.

Інший спосіб вирахування коефіцієнта рангової кореляції був запропонований Кендаллом. Він так само обчислюється по рангах r_i та s_i. Елементи вибірки сортуються так, щоб послідовність рангів однієї вибірки r_i являла собою натуральний ряд 1,2,3,...n. Для кожного і-го рангу другої послідовності s_i визначаємо два числа :

p_i – кількість рангів другої послідовності, які слідують після s_i і є більшими за s_i;

q_i – кількість рангів, які слідують після s_i, але є меншими за s_i.

Коефіцієнт рангової кореляції Кендалла (коефіцієнт кореляції рангів, ранговий коефіцієнт кореляції), який розраховують за формулою:

$$\tau = \frac{K}{\sqrt{\left(\frac{n(n-1)}{2} - B_x\right)\left(\frac{n(n-1)}{2} - B_y\right)}}, \qquad (4.22)$$

де $K = P - Q$; $P = \sum p_i$; $Q = \sum q_i$; B_x, B_y –поправки на об'єднання рангів у відповідних рядах, які обчислюють за формулою:

$$B_i = \frac{1}{2}\sum_{i=1}^{m} n_i(n_i - 1), \qquad (4.23)$$

де m – кількість груп об'єднаних рангів у вибірці, n_i – кількість рангів в і-й групі.

У випадку, коли об'єднання рангів відсутні, формула (4.22) спрощується і набирає вигляду

$$\tau = \frac{2K}{n(n-1)}. \qquad (4.24)$$

Для прикладу, представленого вище (табл.4.2) ми маємо $P = 29; Q = 16$. Отже, К = 13, $\tau = \frac{2 \cdot 13}{10 \cdot 9} = 0.29$. Отримане значення є досить близьким до значення коефіцієнта Спірмена.

Коефіцієнт рангової кореляції Кендалла має ряд переваг порівняно з коефіцієнтом Спірмена. Основними з них є:

- кращий рівень вивченості його статистичних властивостей, зокрема його вибіркового розподілу;
- можливість його застосування для визначення частинної кореляції;
- більша зручність перерахунку при додаванні нових даних.

Типовою ситуацією, у якої зустрічається необхідність перевірки зв'язку між номінальними ознаками, є обробка результатів соціологічних досліджень, що можуть містити такі комбінації ознак, як освіта, стать, професія, підтримка певної політичної партії, регіон мешкання тощо.

При дослідженні зв'язків між категоризованими ознаками вихідні дані представляють у вигляді **таблиці спряженості** (табл.4.3). До ка-

тегоризованих зараховують номінальні ознаки, а також порядкові ознаки, для яких є відомим скінчений набір можливих градацій.

Таблиця 4.3

Таблиця спряженості категоризованих ознак

Рівні ознаки 1	Рівні ознаки 2				Разом
	1	2	…	r	
1	f_{11}	f_{12}	…	f_{1r}	n_1
2	f_{21}	f_{22}	…	f_{2r}	n_2
…	…	…	…	…	…
c	f_{c1}	f_{c2}	…	f_{cr}	n_c
Разом	m_1	m_2	…	m_r	S

Величини f_{ij} показують скільки разів зустрічалася комбінація ознак, за якої рівень першої має значення і, а рівень другої –значення j, m_j є сумами стовпців, n_i – сумами рядків. За даними таблиці можна оцінити значення імовірностей, що входять до формули (4.1):

$$p_{ij} = P\left(A_i B_j\right) = \frac{fi_j}{S}\,;\; p_i = P(A_i) = \sum_{j=1}^{r} p_{ij} = \frac{n_i}{S}\,;$$

$$p_j = P(B_j) = \sum_{i=1}^{c} pi_j = \frac{m_j}{S}\,. \qquad (4.25)$$

Звідси для незалежних ознак маємо:

$$f_{ij} \approx n_i m_j / S\,. \qquad (4.26)$$

Величини $\varphi_{ij} \approx n_i m_j / S$ називають **очікуваними частотами**. Нульову гіпотезу про відсутність зв'язку відхиляють, якщо різницю між ними і частотами, що спостерігаються, не можна пояснити випадковими чинниками. Як критерій можна використовувати величину:

$$X^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{\left(f_{ij}-\varphi_{ij}\right)^2}{\varphi_{ij}} = n\left[\sum_{i=1}^{c}\sum_{j=1}^{r}\frac{f_{ij}^2}{n_i m_j}-1\right], \qquad (4.27)$$

яка при достатньо великому обсязі вибірки наближається до розподілу χ^2 з кількістю степенів вільності *(r-1)(c-1)*. У практиці для можливості застосування критерію часто вважають достатнім, щоб усі значення f_{ij} були не меншими ніж п'ять. При збільшені кількості степенів вільності мінімальні значення f_{ij} можуть бути дещо меншими.

Існує велика кількість показників степеня тісноти статистичного зв'язку, призначених для категоризованих змінних, які не є універсальними, а відбивають окремі властивості такого зв'язку.

Коефіцієнт Крамера розраховують за формулою:

$$C = \left[\frac{\sum_{i=1}^{c}\sum_{j=1}^{r}\frac{f_{ij}^2}{n_i m_j}-1}{\min(c-1, r-1)}\right]^{1/2} \qquad (4.28)$$

Він змінюється в межах від нуля до одиниці. При цьому значення C=0 свідчить про статистичну незалежність аналізованих ознак, а значення C=1 – про можливість однозначного відтворення значень однієї ознаки за відомими значеннями другої. Дисперсію оцінки коефіцієнта Крамера можна отримати з виразу:

$$\sigma_{\hat{N}}^2 \approx \frac{1}{n\,\min(c-1, r-1)}. \qquad (4.29)$$

Її довірчий інтервал:

$$\left[C - u_{1-\alpha}\sigma_C; C + u_{1-\alpha}\sigma_C\right], \tag{4.30}$$

де u_q - q-квантиль стандартного нормального розподілу.

Поліхоричний коефіцієнт спряженості Чупрова призначений для дослідження кореляції номінальних ознак у таблиці спряженості $\mathrm{r} \times \mathrm{c}$. Його значення розраховують за формулою:

$$T = \frac{J-1}{\sqrt{(r-1)(c-1)}}\ ; J = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{f_{ij}^2}{n_i m_j}. \tag{4.31}$$

Існує велика кількість коефіцієнтів, що характеризують кореляцію між ознаками для випадку, коли кожна з двох ознак може мати лише два рівні, які частіше відповідають наявності та відсутності ознаки. У цьому разі таблиця спряженості має розмір 2×2 і її елементи позначають так: $a = f_{11}, b = f_{12}, c = f_{21}, d = f_{22}$.

Коефіцієнт (показник подібності) Жаккара обчислюють за формулою:

$$J = \frac{a}{a+b+c}, \tag{4.32}$$

Значення цього коефіцієнту можуть змінюватися в межах від нуля до одиниці.

Простий коефіцієнт зустрічальності (показник подібності Сокала та Мітченера) розраховують за формулою:

$$J = \frac{a+d}{n} = \frac{a+d}{a+b+c+d}. \tag{4.33}$$

Як і в попередньому випадку, значення коефіцієнта можуть змінюватися в межах від нуля до одиниці.

Показник подібності Рассела і Рао обчислюють як:

$$J = \frac{a}{n} = \frac{a}{a+b+c+d}. \tag{4.34}$$

Його значення також можуть змінюватися в межах від нуля до одиниці.

Коефіцієнт спряженості Бравайса (φ-коефіцієнт Пірсона, показник подібності Чупрова) розраховують за формулою:

$$C = \frac{ad - bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}. \tag{4.35}$$

Значення цього коефіцієнта може змінюватися в межах від -1 до +1.

Коефіцієнт асоціації Юла визначають із співвідношення

$$Q = \frac{ad - bc}{ad + bc} \tag{4.36}$$

Значення коефіцієнта змінюється в межах від -1 до +1.

Хеммінгова відстань (метрика Хеммінга) $H = a + d$ також може застосовуватися для визначення кореляції, але, як і коваріація, вона не є безрозмірною величиною і може набувати будь-яких невід'ємних значень (верхньою межею є загальна кількість спостережень n).

Коефіцієнт Гауера застосовують у разі, коли досліджувані ознаки виміряні в різних шкалах. Обчислення елементів матриці подібності здійснюють за формулою:

$$s_{ij} = \frac{\sum_{k=1}^{p} S_{ijk}}{\sum_{k=1}^{p} W_{ijk}}, \left(i = 1,...,n; j = 1,...,n\right), \tag{4.37}$$

де $s_{ijk}\ (i, j = 1,...,n; k = 1,..., p)$ - внесок ознаки у подібність об'єктів, W_{ijk} - вагова змінна ознаки, p – кількість ознак, що характеризують об'єкт, n – кількість об'єктів.

Для дихотомічних ознак алгоритм підрахунку внеску ознаки і визначення вагових коефіцієнтів збігається з коефіцієнтом Жаккара. Для порядкових ознак алгоритм підрахунку внеску ознаки збігається з Хеммінговою відстанню, узагальненою на порядкові змінні, а вагові коефіцієнти беруть рівними одиниці для всіх ознак. Для кількісних ознак:

$$S_{ijk} = 1 - \frac{\left| x_{ik} - x_{jk} \right|}{R_k}, \tag{4.38}$$

де x_{ik}, x_{jk} - значення k – ї змінної для об'єктів і та j, R_k – розкид k-ї ознаки, обчислений за всіма об'єктами.

Бісеріальний коефіцієнт кореляції призначений для дослідження кореляції в таблицях розміром $2 \times n$,які є дихотоміями за певною номінальною ознакою і класифікаціями за номінальною або порядковою ознакою, яка класифікується за q класами і може бути впорядкованою або невпорядкованою. Вихідний розподіл має бути двовимірним нормальним.

У разі класифікації за порядковою ознакою бісеріальний коефіцієнт:

$$r_b = \frac{(\bar{x}_1 - \bar{x}) n_1}{n \sigma_x z_k}, \tag{4.39}$$

де $\bar{x}_1$ - середнє за першим рядком, $\bar{x}$ - загальне середнє за всією таблицею, σ_x - вибіркове середнє квадратичне відхилення, n_1 - чисель-

ність першого рядка, n – загальна чисельність всіх вибірок, z_k - ордината щільності нормального розподілу в точці k, де k- розв'язок рівняння.

$$1 - F(k) = n_i / n . \tag{4.40}$$

Похибку бісеріального коефіцієнта можна визначити за формулою:

$$m_{r_b} = \frac{1 - r_b}{\sqrt{n}} . \tag{4.41}$$

Він має t- розподіл з кількістю степенів вільності (n-2).

У разі, коли одна із змінних дихотомізована, а інша – виміряна в кількісній шкалі, обчислюють **точково-бісеріальний коефіцієнт кореляції**, який визначають за формулою:

$$r_{pb} = \frac{\left|\bar{x}_p - \bar{x}\right|}{\sigma_x} \sqrt{\frac{n_p}{n_q}} , \tag{4.42}$$

де $\bar{x}_p$ - середнє варіант кількісної вибірки, які відповідають подіям верхнього(першого) рівня дихотомічної вибірки, $\bar{x}$ - середнє кількісної вибірки, σ_x - середнє квадратичне кількісної вибірки, n_p - кількість подій у верхній (з рівнем 1) групі, n_q - кількість подій у нижній (з рівнем 2) групі. При цьому передбачається, що дихотомічна змінна може набувати лише два значення: 1 (верхній рівень) та 0 (нижній рівень). З погляду теорії точково-бісеріальну кореляцію можна розглядати як окремий випадок кореляційного відношення Пірсона.

Величину точково-бісеріального коефіцієнта кореляції вважають відмінною від нуля на рівні значущості α, якщо виконується нерівність:

$$r_{pb}\sqrt{\frac{n-2}{1-r_{pb}^2}} \geq t_\alpha , \tag{4.43}$$

де t_α - критичне значення t- розподілу з *(n-2)* степенями вільності.

Про множинну кореляцію говорять, коли певна ознака залежить не від одного, а від декількох зовнішніх факторів.

У разі, коли досліджувані ознаки задовольняють багатовимірному нормальному розподілу, частинний коефіцієнт кореляції між двома ознаками при фіксованих значеннях інших ознак розраховують за формулою:

$$r_{ijX(i,j)} = -\frac{R_{ij}}{\sqrt{R_{ii}R_{jj}}}, \tag{4.44}$$

де R_{k1} - алгебраїчне доповнення для елемента r_{k1} у кореляційній таблиці. Для тривимірної ознаки звідси можна отримати:

$$r_{01(2)} = \frac{r_{01} - r_{02}r_{12}}{\sqrt{\left(1-r_{02}^2\right)\left(1-r_{12}^2\right)}} . \tag{4.45}$$

Частинні коефіцієнти кореляції порядку *k*, тобто такі, що не враховують опосередкований вплив *k* інших змінних, можна розрахувати за коефіцієнтами порядку $k-1$, використовуючи рекурентну формулу:

$$r_{01(2,3,\ldots,k+1)} = \frac{r_{01(2,\ldots,k)} - r_{0k+1(2,\ldots,k)}r_{1k+1(2,\ldots,k)}}{\sqrt{\left(1-r_{0k+1(2,\ldots,k)}^2\right)\left(1-r_{1k+1(2,\ldots,k)}^2\right)}} . \tag{4.46}$$

Тісноту зв'язку між змінними у випадку множинної регресії можна оцінити за допомогою **коефіцієнта множинної кореляції**

$$R = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \qquad (4.47)$$

де Y_i - значення змінної, взяті з кореляційної таблиці, y_i - відповідні значення, розраховані за формулою регресії.

Крім того застосовують **множинний коефіцієнт кореляції**, який є мірою лінійної кореляції між певною змінною y та сукупність величин $X_1, X_2, ..., X_n$ і визначається як звичайний парний коефіцієнт кореляції між y та множинною лінійною регресією за $X_1, ..., X_n$. Його можна розраховувати за формулою:

$$R_{yX}^2 = 1 \frac{|R|}{|R|_{00}}, \qquad (4.48)$$

де $|R|$ - визначник кореляційної матриці, або за частинними коефіцієнтами кореляції:

$$R_{yX}^2 = 1 - (1 - r_{01}^2)(1 - r_{02(1)}^2)(1 - r_{03(12)}^2)...(1 - r_{0n(1,2,...,n-1)}^2). \qquad (4.49)$$

При цьому припускають, що досліджувана сукупність підпорядковується багатовимірному нормальному закону.

Множинний коефіцієнт кореляції мажорує будь-який парний або частинний коефіцієнт кореляції, що характеризує статистичні зв'язки досліджуваної ознаки. Як видно з формули (4.49), додавання нових ознак не може зменшувати коефіцієнт множинної кореляції.

Для багатовимірних нормальних сукупностей виконується рівність:

$$K_d(y, X) = R^2_{yX} \ . \qquad (4.50)$$

Коефіцієнт конкордації призначений для дослідження узгодженості k ранжувань вибірки з n членів одного з одним. Зокрема, таке завдання виникає при вирішенні питання, наскільки добре узгоджуються між собою ранжування масиву даних розміру n, здійснені k різних експертами. Коефіцієнт обчислюють за формулою:

$$W = \frac{12\sum_{i=1}^{n}\left(\sum_{j=1}^{k} r_{ij} - \frac{k(n+1)}{2}\right)^2}{k^2 n\left(n^2 - 1\right)} \ , \qquad (4.51)$$

де $r_{ij} \quad (i = 1,2,...,n;\ j = 1,2,...,k)$ – масив рангових оцінок; n – розмір кожного масиву; k – кількість масивів. Значення коефіцієнта конкордації може змінюватися в межах від нуля до одиниці, при цьому він дорівнює одиниці лише за умови, що всі досліджувані ранжування збігаються. Коефіцієнт конкордації дорівнює нулю, якщо $k \geq 3$ і всі ранжування є випадковими впорядкуваннями вихідної вибірки.

Розглянемо **приклад.** На експертизу представлені n зразків, які необхідно ранжувати за деякою ознакою. В оцінці приймає участь m експертів. В результаті проведення експертної оцінки отримуємо наступну таблицю 4.4 (6 товарів, 3 експерти). Згідно із формулою (4.51) маємо $W = \frac{12 * 133}{9 \cdot 6 \cdot (36 - 1)} = 0.84$.

Висновок : оцінки експертів є добре узгодженими.

Замість експертів ми можемо розглядати m різних ознак товару чи явища і ранжувати зразки за цими ознаками, а потім пе-

ревірити, чи узгоджується ранжування за різними ознаками і наскільки.

Таблиця 4.4

Розрахунок коефіцієнта конкордації

Зразок	**Експерти**			$\sum r_{ij}$	$(\sum r_{ij} - \frac{k(n+1)}{2})^2$
	1	**2**	**3**		
1	1	2	1	4	42.25
2	2	1	3	6	20.25
3	4	4.5	3	11.5	1.00
4	5	4.5	6	15.5	25.00
5	3	3	3	9	2.25
6	6	6	5	17	42.25
Сума	21	21	21	63	133.00

Величина $(k-1)\frac{W}{1-W}$ має F - розподіл з кількостями степенів вільності $(n-1)$ та $((n-1)(k-1)-2)$. Великі значення функції F – розподілу свідчать про високий рівень узгодженості між ранжуваннями. При $n > 7$ величина $k(n-1)W$ має розподіл, близький до χ^2 з $n-1$ степенем вільності. Якщо

$$k(n-1)W > \chi_\alpha^2(n-1), \tag{4.52}$$

то гіпотезу про відсутність рангової кореляції можна відкинути при рівні значущості α.

ЗАВДАННЯ ДО РОЗДІЛУ 4

Завдання 1.

1. За даними таблиці :

а) побудувати графік залежності ;

б) знайти коефіцієнти парної лінійної кореляції та детермінації для незгрупованих даних .

2. За даними цієї ж таблиці :

а) побудувати кореляційну таблицю ;

б) знайти коефіцієнти парної лінійної кореляції та детермінації для згрупованих даних .

3. Порівняти значення обох коефіцієнтів і зробити висновок про тісноту кореляційного зв'язку .

Варіант 1

Залежність кількості проданих пар чоловічого взуття Y від його розміру X наведена в таблиці:

$Y = y_i$, шт...	10	25	68	136	152	162	170	180
$X = x_i$,	44	43	42	41	40	39	38	37

Варіант 2

Вимірювання температури в грудні, здійснені у двох містах, що умовно позначені А і В, наведено в таблиці:

Місто А $Y = y_i$, °C	-10,2	-11,5	-12,4	-12,8	-13,0	-13,5	-14,2	-14,6
Місто В $X = x_i$, °C	-20,2	-20,5	-21,4	-21,8	-22,0	-22,5	-22,8	-22,8

Продовження таблиці

Місто А $Y = y_i$, °C	-14,6	-15,7	-16,4	-17,2	-17,5	-18,2	-18,6	-18,9
Місто В $X = x_i$, °C	-23,2	-24,1	-24,5	-25,1	-25,8	-26,0	-26,5	-27,0

Варіант 3

Конденсатор було заряджено до повної напруги в певний момент часу t, після чого він починає розряджатися. Залежність напруги Y від часу розряджання X наведено в таблиці:

$Y = y_i$,	100	85	70	65	60	55	50	45	40	35	30	25	22	20
$X = x_i$,	0	1	2	3	4	5	6	7	8	9	10	11	12	13

Варіант 4

Залежність граничного навантаження на болт *Y* від його твердості *X* наведено в таблиці:

$Y=y_i$,умов.од.	12,96	13,44	13,60	13,95	14,50	14,98	15,48	15,96	16,50
$X=x_i$, умов.од.	64,8	65,4	68,4	69,2	70,5	74,5	76,8	78,5	80,0

Продовження таблиці

$Y=y_i$,умов.од.	10,1	10,3	10,45	10,9	11,20	11,35	11,9	12,45	12,58
$X=x_i$, умов.од.	50,0	50,2	52,8	53,5	54,0	56,8	58,8	59,5	60,5

Варіант 5

Залежність урожайності Y пшениці від кількості внесених добрив X наведено у таблиці

$Y=y_i$, ц\га	10	12	14	16	18	20	22	24	26	28	30	32	34
$X=x_i$, кг\га	10	30	40	50	60	70	80	90	100	110	120	130	140

Завдання 2

1. За двома довільними стовпцями таблиці :

а) побудувати графік залежності ;

б) знайти коефіцієнти парної нелінійної кореляції та детермінації для незгрупованих даних .

2. . За двома довільними стовпцями таблиці :

а) побудувати кореляційну таблицю ;

б) знайти коефіцієнти парної нелінійної кореляції та детермінації для згрупованих даних .

3. Порівняти значення обох коефіцієнтів і зробити висновок про тісноту кореляційного зв'язку .

Варіант 1

	Y	X	Z
1	508	2050	300
2	534	2060	320
3	519	2070	340
4	542	2100	350
5	524	2150	370
6	549	2210	410
7	534	2300	550
8	542	2350	530
9	531	2340	550
10	535	2450	490
11	507	2500	350
12	496	2600	330
13	485	2650	350
14	500	2700	410
15	486	2750	440

Варіант 2

	Y	X	Z
1	468	1200	600
2	496	1300	650
3	484	1400	630
4	528	1450	620
5	495	1500	610
6	543	1550	590
7	509	1600	580
8	565	1650	560
9	502	1630	570
10	568	1680	540
11	511	1710	520
12	575	1780	510
13	536	1810	500
14	557	1830	490
15	534	1850	430

Варіант 3

	Y	X	Z
1	486	2750	440
2	481	2850	460
3	464	2900	480
4	450	3000	510
5	467	2900	550
6	475	2850	560
7	484	2800	550
8	492	2750	540
9	500	2700	530
10	06	2650	550
11	514	2600	510
12	519	2550	530
13	521	2500	570
14	529	2450	520
15	534	2400	510

Варіант 4

	Y	X	Z
1	524	2150	370
2	549	2210	410
3	542	2350	530
4	535	2450	490
5	496	2600	330
6	500	2700	410
7	481	2850	460
8	450	3000	510
9	475	2850	560
10	492	2750	540
11	506	2660	550
12	519	2550	530
13	529	2450	520
14	537	2350	530
15	544	2250	500

Варіант 5

	Y	X	Z
1	557	1830	490
2	548	1740	420
3	550	1910	390
4	534	2060	320
5	542	2100	350
6	549	2210	410
7	542	2350	530
8	535	2450	490
9	496	2600	330
10	500	2700	410
11	481	2850	460
12	450	3000	510
13	475	2850	550
14	492	2750	540
15	506	2650	550

Завдання 3.

Знайти коефіцієнти множинної лінійної кореляції $R_{y\bullet x_1 x_2}$, $R_{x_1\bullet yx_2}$, $R_{x_2\bullet yx_1}$ та коефіцієнти частинної кореляції для трьох змінних за даними таблиці.

Варіант 1

	Y	X_1	X_2
1	14,85	60	30
2	11,94	48	19
3	8,03	39	8
4	7,11	28	18
5	9,50	45	9
6	9,40	37	23
7	11,60	58	15
8	8,14	27	17
9	15,62	58	28
10	11,12	47	116
11	7,34	38	7
12	10,58	44	15
13	7,37	23	25
14	10,63	57	8
15	10,63	38	24

Варіант 2

	Y	X_1	X_2
1	11,12	47	16
2	7,34	38	7
3	10,58	44	15
4	7,37	23	25
5	10,63	57	8
6	10,63	38	24
7	7,85	22	15
8	5,73	29	7
9	14,84	56	27
10	10,30	45	15
11	7,85	34	9
12	9,68	51	14
13	9,49	55	5
14	12,53	43	26
15	10,29	44	27

Варіант 3

	Y	X_1	X_2
1	5,73	29	7
2	7,85	34	9
3	12,53	43	26
4	12,28	33	24
5	7,47	53	13
6	5,23	26	12
7	12,16	32	23
8	6,86	51	8
9	11,02	43	22
10	7,77	29	9
11	10,62	37	12
12	7,40	49	5
13	10,55	57	11
14	12,30	46	15
15	7,83	29	21

Варіант 4

	Y	X_1	X_2
1	8,99	37	8
2	12,28	33	24
3	8,00	25	18
4	7,27	29	4
5	7,47	53	13
6	10,86	41	9
7	5,23	26	12
8	12,16	32	23
9	9,19	59	11
10	10,12	48	3
11	6,86	51	8
12	11,02	43	22
13	7,77	29	9
14	10,62	37	12
15	7,40	49	5

Варіант 5

	Y	X_1	X_2
1	10,58	44	15
2	7,37	23	25
3	10,63	38	24
4	7,85	22	15
5	5,73	29	7
6	14,84	56	27
7	10,30	45	15
8	9,68	51	14
9	9,49	55	5
10	12,53	43	26
11	10,29	44	27
12	12,28	33	24
13	8,00	25	18
14	7,27	29	4
15	7,47	53	13

Питання для самоконтролю

1. Що таке кореляція?
2. В чому суть кореляційного аналізу?
3. Як обчислюються коефіцієнти парної лінійної та множинної лінійної кореляції і детермінації?
4. Дайте визначення кореляційної матриці.
5. В чому полягає застосування індекса Фехнера? Як він обчислюється?
6. Для чого використовується кореляційне відношення?
7. Що розуміють під ранговою кореляцією?
8. Які коефіцієнти рангової кореляції відомо?
9. Що таке коваріація?
10. Що таке коефіцієнт конкордації?

5. РЕГРЕСІЙНИЙ АНАЛІЗ

Завданням дослідження складних систем і процесів часто є перевірка наявності і встановлення типу зв'язку між незалежними змінними x_i (предикторами, факторами), значення яких можуть змінюватися дослідником і мають певну, заздалегідь задану похибку, та залежною змінною (відгуком) y. Розв'язання таких завдань є предметом регресійного аналізу.

5.1. Загальна характеристика методів та задач регресійного аналізу

Класичний регресійний аналіз включає методи побудови математичних моделей досліджуваних систем, методи визначення параметрів цих моделей і перевірки їх адекватності. Він припускає, що регресія є лінійною комбінацією лінійно незалежних базисних функцій від факторів з невідомими коефіцієнтами (параметрами). Фактори і параметри є детермінованими, а відгуки − рівноточними (тобто мають однакові дисперсії) некорельованими випадковими величинами. Припускають також, що всі змінні вимірюються у неперервних числових шкалах.

Звичайна **процедура класичного регресійного аналізу** є такою.

Спочатку обирають гіпотетичну модель, тобто формулюють гіпотези про фактори, які суттєво впливають на досліджувану характеристику системи, і тип залежності відгуку від факторів. Потім за наявними емпіричними даними про залежність відгуку від факторів оцінюють параметри обраної моделі. Далі за статистичними критеріями перевіряють її адекватність.

При побудові регресійних моделей реальних систем і процесів вказані вище припущення виконуються не завжди. У більшості випадків їх невиконання призводить до некоректності використання процедур класичного регресійного аналізу і потребує застосування більш складних методів аналізу емпіричних даних.

Постулат про рівноточність і некорельованість відгуків не є обов'язковим. У разі його невиконання процедура побудови регресійної моделі певною мірою змінюється, але суттєво не ускладнюється. Більш складною проблемою є обрання моделі та її незалежних змінних. У класичному регресійному аналізі припускають, що набір факторів задається однозначно, всі суттєві змінні наявні у моделі, і немає ніяких альтернативних способів обрання факторів. На практиці це припущення не виконується. Тому виникає необхідність розробки формальних і неформальних процедур перетворення та порівняння моделей. Для пошуку оптимальних формальних перетворень використовують методи факторного та дискримінантного аналізу. На сьогодні розроблено комп'ютеризовані технології послідовної побудови регресійних моделей.

У класичній регресії фактори вважають детермінованими, тобто припускають, що дослідник має про них всю необхідну інформацію з абсолютною точністю. У практиці це припущення не виконується. Відмова від детермінованості незалежних змінних веде до побудови моделей кореляційного аналізу. Їх практичне використання обмежене переважно випадком однофакторних моделей (парною кореляцією). Це пов'язано зі складностями забезпечення та перевірки вимог до багатовимірних функцій розподілу. В окремих випадках можуть викори-

стовуватися компромісні методи конфлюентного аналізу, які припускають можливість нормально розподіленого та усіченого розкиду значень факторів. Якщо ця умова виконується, побудову моделі можна звести до багаторазового розв'язування регресійної задачі.

Відмова від припущення про детермінованість параметрів моделей у регресійному аналізі призводить до суттєвих ускладнень, оскільки порушує його статистичні основи. У деяких випадках можна вважати параметри випадковими величинами із заданим законом розподілу. Тоді як оцінки параметрів можна обрати їх умовні математичні сподівання для відгуків, що спостерігалися. Умовні розподіли та сподівання розраховують за узагальненою формулою Байєса. Тому відповідні методи називають **байєсівським регресійним аналізом**.

Регресійні моделі часто використовують для опису процесів, що розвиваються у часі. У певних випадках це призводить до необхідності переходу від випадкових величин відгуків до випадкових послідовностей, випадкових процесів або випадкових полів. Однією з поширених і найпростіших моделей такого типу є модель авторегресії. Вона припускає, що відгук залежить не тільки від факторів, але також і від часу. Якщо останню залежність можна виявити, то проблема зводиться до стандартної задачі побудови регресії для модифікованого відгуку. В інших випадках необхідно використовувати більш складні прийоми.

Процедури класичного регресійного аналізу припускають, що закон розподілу відгуків є нормальним. Проте на практиці найчастіше цей закон невідомий, або ж відомо, що він не є нормальним. В таких випадках є необхідним застосування **непараметричного регресійно-**

го аналізу, який не передбачає необхідності попереднього задання функції розподілу.

Важливою проблемою, яка виникає при оцінюванні параметрів регресійних моделей, є наявність грубих помилок серед набору аналізованих даних. Ці помилки можуть виникати внаслідок неправильних дій дослідника, збоїв у роботі апаратури, неконтрольованих короткотривалих сильних зовнішніх впливів на досліджувану систему тощо. У таких випадках використовують два підходи, що дають змогу зменшити вплив грубих помилок на результати аналізу. У першому з них розробляють критерії та алгоритми пошуку помилкових даних. Потім ці дані відкидають. У другому підході розробляють алгоритми аналізу, які є нечутливими до наявних помилкових даних (алгоритми **робастного оцінювання параметрів**). У таких алгоритмах, як критерій оптимальності, часто використовують мінімум суми модулів похибок, або мінімум максимального модуля похибки.

Одним з основних постулатів класичного регресійного аналізу є припущення, що найкращі оцінки параметрів можна одержати, використовуючи **метод найменших квадратів**. У практиці оцінки, одержані за допомогою цього методу, часто бувають недостатньо точними і містять великі похибки. Причиною цього може бути структура регресійної моделі. Якщо вона становить собою лінійну комбінацію експонент або є поліномом високого степеня, то це призводить до поганої обумовленості матриці системи нормальних рівнянь і нестійкості оцінок параметрів. Підвищення стійкості оцінок можна досягти шляхом відмови від вимоги їх незміщеності. Розвиток цього напряму дослі-

джень спричинив виникнення **гребеневого**, або **рідж-регресійного аналізу**.

Найчастіше **задачу побудови регресійної моделі** формулюють так. Необхідно знайти функцію $z(\alpha, X)$ заданого класу, для якої функціонал нев'язки є мінімальним

$$F(\alpha) = \sum_{i=1}^{n} \left(z_i(\alpha, X) - y_i\right)^2 \to \min . \qquad (5.1)$$

У цьому виразі $z_i(\alpha, X)$ – значення функції, яка апроксимує залежність в і-й точці, y_i – відповідне значення емпіричної залежності, α – вектор параметрів, які треба знайти, X – вектор незалежних змінних. Одержану функцію $z(\alpha, X)$ називають **регресійною моделлю**. Метод її пошуку, оснований на застосуванні критерію (5.1), називають **методом найменших квадратів**.

У випадку однієї незалежної змінної апроксимуючу функцію найчастіше шукають у вигляді полінома $z(x) = \sum_{j=0}^{M} \alpha_j x^j$, експоненціальної функції $z = \alpha e^x$, показникової функції $z = \alpha b^x$, степеневої функції $z = \alpha x^b$, тригонометричного ряду Фур'є тощо. У разі декількох незалежних змінних найчастіше використовують моделі, лінійні як за параметрами, так і за незалежними змінними $z = \alpha_0 + \sum_{i=1}^{p} \alpha_i x_i$, а також поліноміальні моделі, які є лінійними за параметрами, але нелінійними за незалежними змінними

$$z = \alpha_0 + \sum_{i=1}^{p} \alpha_i x_i + \sum_{\substack{i,j=1 \\ i \geq j}}^{p} \alpha_{ij} x_i x_j + \sum_{\substack{i,j,k=1 \\ i \geq j \\ j \geq k}}^{p} \alpha_{ijk} x_i x_j x_k + \ldots .$$

Останні відповідають розкладенню функції відгуку в ряд Тейлора. Можливе використання і інших видів апроксимаційних залежностей. Регресійні моделі називають **лінійними** або **нелінійними**, якщо вони є, відповідно, лінійними або нелінійними за параметрами. При цьому визначення "лінійна" часто опускають. Значення найвищого степеня предиктора в поліноміальних моделях називають порядком моделі. Наприклад,

$$z = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3, \qquad (5.2)$$

є лінійною моделлю третього порядку.

Вибір виду апроксимуючої функції є нетривіальним завданням. Спочатку рекомендують побудувати точковий графік емпіричних даних. Це дає можливість визначити наявність чи відсутність залежності між досліджуваними величинами, а також зробити певні припущення про тип такої залежності (рис.1, Розділ 1).

На рис.5.1 у якості ще одного прикладу подано результати, одержані окремими кандидатами на виборах Президента України у 1999 році. З наведених даних видно, що в першому випадку кореляція практично відсутня, а в другому можна говорити про наявність від'ємної кореляції.

Після того, як наявність кореляції між досліджуваними величинами встановлено, переходять до підбору функції, що апроксимує шукану залежність.

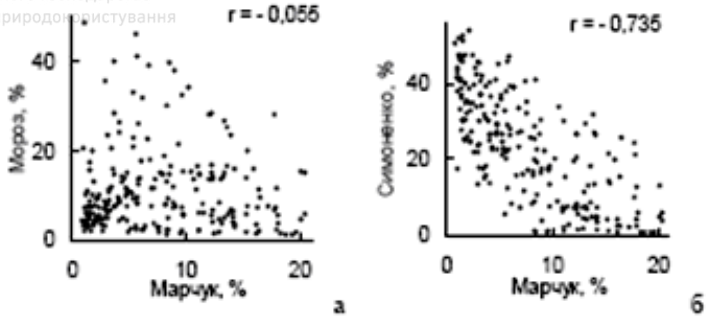

Рис.5.1. Кореляція між рівнями підтримки окремих кандидатів у І турі виборів президента України у 1999 році (r – коефіцієнт кореляції)

Важливою особливістю регресійних моделей є те, що їх не можна використовувати за межами тієї області значень вихідних параметрів, для якої була побудована ця модель. При побудові регресійних моделей типу полінома, тригонометричного ряду та деяких інших слід мати на увазі, що, збільшуючи кількість членів ряду, можна одержати скільки завгодно близьке до нуля значення функціонала (5.1). Проте це не завжди свідчить про якість апроксимації, оскільки цей функціонал не дає інформації про ступінь близькості апроксимуючої функції та емпіричної залежності у проміжках між відомими точками. Найпростішим критерієм, що дає можливість обмежити кількість членів ряду, який апроксимує емпіричну залежність, є порівняння відношення значення функціонала (5.1) до суми квадратів похибок емпіричних значень функції, а також величини, оберненої до цього відношення, із значенням критерію Фішера для відповідного числа степенів вільності. Якщо перше відношення перевищує значення критерію Фішера, то це свідчить про недостатню кількість членів ряду. Якщо, значення

критерію Фішера є меншим за друге відношення, можна зробити висновок про надмірно велику кількість членів ряду.

Іншою проблемою може бути наявність декількох екстремумів функціонала (5.1). У цьому разі необхідно мати на увазі, що більшість стандартних алгоритмів дає можливість знаходити локальні, а не глобальні екстремуми функціоналів, і результат мінімізації залежатиме від вибору початкових умов пошуку. Тому для визначення адекватності одержуваних моделей доцільно застосовувати аналіз залишків, тобто різниць між наявними емпіричними даними і відповідними точками модельної залежності. Для адекватної моделі ряд залишків має наближатися за властивостями до **білого шуму**, характерними рисами якого є рівність середнього арифметичного нулю і відсутність автокореляції.

Поліноміальні регресійні моделі, як правило, є формальними. Їх використовують для опису систем і процесів, теорію яких розроблено недостатньо. Більш цікавими для дослідників зазвичай є **змістовні** моделі, що відображають структуру та зв'язки у системах, сутність та механізми процесів. Якщо теоретичні основи досліджуваних систем і процесів достатньо розроблені, експериментальні дані потрібні для визначення окремих параметрів моделі.

Як правило, змістовні моделі бувають нелінійними за параметрами. Методологію їх дослідження розробляє **нелінійний регресійний аналіз**. Для задач, що розв'язуються у межах цього напряму, характерними є різні ускладнення. Зокрема, моделі зазвичай бувають багатовимірними. При цьому окремі відгуки можуть бути пов'язані один з одним. Сама регресійна модель часто задається у неявному вигляді і є

неаналітичним розв'язком певної системи алгебраїчних або диференціальних рівнянь. Нестійкість оцінок параметрів у нелінійних моделях різко зростає. Як правило, такі задачі мають кілька розв'язків або не мають розв'язків взагалі.

5.2. Парна лінійна регресія

Найпростішим для аналізу і найбільш дослідженим є випадок лінійної кореляційної залежності між досліджуваними величинами X та Y. Наявність такої кореляції можна перевірити, розрахувавши коефіцієнт кореляції:

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{X}\,\overline{Y}}{\sigma_x \sigma_y} = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y}, \tag{5.3}$$

де n – кількість пар відповідних значень (x_i, y_i), $\overline{X}, \overline{Y}$ – середні арифметичні величин X та Y, σ_x та σ_y – їх середньоквадратичні відхилення, $\mathrm{cov}(x, y)$ - коваріація величин X та Y

$$\sigma_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}, \; \sigma_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{Y}\right)^2},$$

$$\mathrm{cov}(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right). \tag{5.4}$$

Максимально можливе значення коефіцієнта кореляції $|r| = 1$ свідчить про наявність лінійного функціонального зв'язку між величинами x та y. Якщо $|r| > 0.5$, це є свідченням помітного лінійного зв'язку ве-

личин x та y. Наступним кроком є оптимальний вибір параметрів залежності

$$z(x) = \alpha_0 + \alpha_1 x. \tag{5.5}$$

Найчастіше для оцінки невідомих параметрів α_0 і α_1 використовують **метод найменших квадратів**. При цьому вважають, що найкращими значеннями параметрів α_0 і α_1 будуть ті, для яких сума квадратів відхилень емпіричних значень y_i від розрахункових значень $z(x_i)$ має мінімальне значення. Для знаходження таких значень треба розв'язати систему:

$$\begin{cases} \dfrac{\partial}{\partial \alpha_0} \displaystyle\sum_{i=1}^{n} \left[z(x_i) - y_i\right]^2 = \dfrac{\partial}{\partial \alpha_0} \sum_{i=1}^{n} \left[\alpha_1 x_i + \alpha_0 - y_i\right]^2 = 0, \\ \dfrac{\partial}{\partial \alpha_1} \displaystyle\sum_{i=1}^{n} \left[z(x_i) - y_i\right]^2 = \dfrac{\partial}{\partial \alpha_1} \sum_{i=1}^{n} \left[\alpha_1 x_i + \alpha_0 - y_i\right]^2 = 0, \end{cases} \tag{5.6}$$

Виконавши диференціювання у (5.6), отримаємо систему двох лінійних рівнянь, розв'язком якої будуть оцінки α_0^* і α_1^* коефіцієнтів лінійної залежності:

$$\begin{gathered} \alpha_1^* = \frac{n\sum\limits_{i=1}^{n} x_i y_i - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} y_i}{n\sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2}; \\ \alpha_0^* = \frac{\sum\limits_{i=1}^{n} y_i - \alpha_1 \sum\limits_{i=1}^{n} x_i}{n} = \frac{\sum\limits_{i=1}^{n} x_i^2 \sum\limits_{i=1}^{n} y_i - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} x_i y_i}{n\sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2}. \end{gathered} \tag{5.7}$$

Після деяких перетворень можна представити співвідношення (5.7) у іншому вигляді

$$\alpha_1^* = \frac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2} = \frac{\operatorname{cov}(x,y)}{\sigma_x^2} = \frac{\sigma_y}{\sigma_x} r. \qquad (5.8)$$

$$\alpha_0^* = \overline{Y} - \alpha_1^* \overline{X}\,.$$

Коефіцієнт α_1 є мірою впливу фактора x на досліджувану величину y. Оцінкою точності описання реальної залежності між y та x за допомогою рівняння лінійної регресії є середнє квадратичне відхилення регресійного значення $z(x_i)$ від фактичного значення y

$$\sigma_{çàë} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - z(x_i))^2} \qquad (5.9)$$

Величина $\sigma_{çàë}$ є мірою точності передбачення значень випадкової величини y, тому її називають стандартною похибкою передбачення. Чим менше значення $\sigma_{çàë}$, тим вірогідніший лінійний зв'язок між y та x.

Щоб перевірити **відповідність математичної моделі експериментальним даним**, використовують **критерій Фішера**. В ролі нульової гіпотези H_0 виступає твердження про те, що регресія в генеральній сукупності є лінійна. Для перевірки нульової гіпотези обчислюють величину

$$F = \frac{(n-2)(\sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2)}{(n-1)\sum_{i=1}^{n}(y_i - \alpha_1 x_i - \alpha_0)^2} \qquad (5.10)$$

Величина F має розподіл Фішера із ступенями вільності $v_1 = 1$ і $v_2 = n-2$. За таблицями розподілу Фішера визначаємо критичне значення $F_{kp}(\alpha, 1, n-2)$. Можна також використати стандартну функцію пакету Microsoft Excel FРАСПОБР$(\alpha, 1, n-2)$. Якщо $F \geq F_{kp}$, нульову гіпотезу відхиляють. Якщо $F < F_{kp}$, немає підстав для відхилення гіпотези H_0, тобто лінійна модель є адекватною до вихідних даних.

Перевірка значимості. Можлива ситуація, коли частина обчислених коефіцієнтів регресії не мають достатнього ступеня значимості, тобто абсолютні значення самих коефіцієнтів будуть меншими від їхньої стандартної похибки. Якщо коефіцієнт α_0 виявиться статистично не значимим, його слід виключити з рівняння регресії і воно набуде вигляду $z = \alpha_1 x$. Якщо коефіцієнт α_1 виявиться статистично не значимим, це означає, що величини x і y є незалежними і слід вибрати іншу залежність.

Для перевірки статистичної значимості коефіцієнтів лінійної регресії використовують наступний алгоритм.

1. Обчислюємо середнє квадратичне залишкове відхилення

$$\sigma_{çàë} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - z(x_i))^2} \qquad (5.11)$$

2. Обчислюємо стандартну помилку коефіцієнта α_0

$$\sigma_{\alpha_0} = \sigma_{çàë}\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad (5.12)$$

3. Обчислюємо стандартну помилку коефіцієнта α_1

$$\sigma_{\alpha_1} = \sigma_{çàë} \Big/ \sqrt{\sum (x_i - \bar{x})^2} \qquad (5.13)$$

4. Значимість коефіцієнтів регресії перевіряємо за допомогою t-критерію Стьюдента $t_i = \dfrac{\alpha_i}{\sigma_{\alpha_i}}$.

5. Формуємо нульову гіпотезу H_0 : $\alpha_i = 0$. Якщо ця гіпотеза вірна, величина t_i має розподіл Стьюдента із $k = n - 2$ ступенями вільності.

6. За табличкою розраховуємо критичне значення параметра t_{kp} (або ж СТЬЮДРАСПОБР (α,N-2)). Значення α, як правило, вибирають рівним $\alpha = 0.05$.

7. Якщо $|t_i| \geq |t_{kp}|$, нульова гіпотеза H_0 відхиляється, тобто коефіцієнт α_i є статистично значимим. Якщо $|t_i| < |t_{kp}|$ коефіцієнт α_i є статистично не значимим.

8. Визначаємо межі довірчих інтервалів для коефіцієнтів регресії

$$\alpha_i^{\min} = \alpha_i - t_{kp}\sigma_{\alpha_i}, \quad \alpha_i^{\max} = \alpha_i + t_{kp}\sigma_{\alpha_i}. \qquad (5.14)$$

Це означає, що з імовірністю $1 - \alpha$ відповідний коефіцієнт знаходиться в інтервалі $\alpha_i^{\min} \leq \alpha_i \leq \alpha_i^{\max}$.

Розглянемо **приклад** побудови лінійної регресійної моделі згідно з алгоритмом, описаним вище. У розрахунковій таблиці 5.1 подано емпіричні дані x_i, y_i, для яких треба побудувати регресійну модель, та їх комбінації, необхідні для використання формул (5.3) - (5.8).

Таблиця 5.1

Приклад розрахунку регресійної моделі

№ випробування	x_i	y_i	x_i^2	$x_i y_i$	z	Залишок
1	0	0.310	0	0.000	0.516	0.206
2	0.3	1.037	0.09	0.311	1.506	0.469
3	0.6	2.513	0.36	1.508	2.497	-0.016
4	0.9	3.843	0.81	3.459	3.487	-0.356
5	1.2	4.840	1.44	5.808	4.477	-0.363
6	1.5	6.020	2.25	9.030	5.467	-0.553
7	1.8	5.865	3.24	10.557	6.457	0.592
8	2.1	7.470	4.41	15.687	7.447	-0.023
9	2.4	8.889	5.76	21.334	8.438	-0.451
10	2.7	9.254	7.29	24.986	9.428	0.174
11	3	10.393	9	31.179	10.418	0.025
12	3.3	11.113	10.89	36.672	11.408	0.295
∑	19.8	71.547	45.54	160.530	71.547	0.000
$a_1^* =$	3,30055	$a_0^* =$	0.51633			

На рис.5.2 наведено емпіричні точки та графік лінійної модельної залежності (пряма лінія), побудованої за методом найменших квадратів. Як видно з рисунка, одержана модель задовільно описує наявні емпіричні дані.

Незважаючи на те що, зазвичай, реальні залежності відгуків від факторів є нелінійними, розглянутий випадок широко використовують у практиці побудови регресійних моделей. Це пов'язано з трьома основними причинами. По-перше, він є найбільш дослідженим. Зокрема,

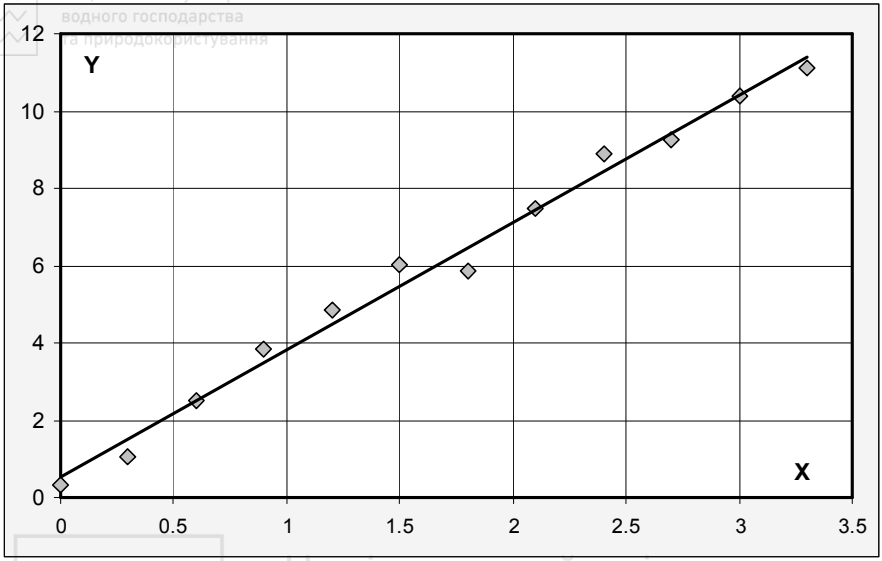

Рис.5.2. Побудова лінійної регресії за емпіричними даними

для нього досить повно розроблені процедури визначення статистичних характеристик одержуваних оцінок параметрів, перевірки адекватності моделей тощо. По-друге, у багатьох випадках складні залежності можна подати як набір лінійних (на малих відрізках зміни факторів) залежностей. По-третє, нелінійні залежності часто можна перетворити до лінійного вигляду шляхом заміни змінних (метод вирівнювання залежностей). Приклади такого перетворення наведено у табл. 5.2.

Перетворення нелінійних залежностей до лінійних є припустимим, якщо вихідні дані є точними. Але на практиці вони завжди вимірюються з деякою похибкою. Розглянемо модель

$$z = \alpha_0 x^{\alpha_1} + \varepsilon , \qquad (5.15)$$

Таблиця 5.2

Приклади лінеаризації нелінійних залежностей

Вихідна залежність	Лінеаризована залежність	Нові змінні
$z = \alpha_0 \exp(-\alpha_1 x)$	$\ln z = \ln \alpha_0 - \alpha_1 x$	$x, \ln z$
$z = \alpha_0 [1 - \exp(-\alpha_1 x$	$\ln \frac{\alpha_0}{\alpha_0 - z} = \alpha_1 x$	$x, \ln \frac{\alpha_0}{\alpha_0 - z}$
$z = \alpha_0 \exp(-\alpha_1 / x)$	$\ln z = \ln \alpha_0 - \alpha_1 / x$	$1/x, \ln z$
$z = \alpha_0 x^{\alpha_1}$	$\ln z = \ln \alpha_0 + \alpha_1 \ln x$	$\ln x, \ln z$
$z = \alpha_0 x + \alpha_1 x^2$	$z / x = \alpha_0 + \alpha_1 x$	$x, z / x$
$z = \alpha_0 \sin(\alpha_1 x)$	$\arcsin(z / \alpha_0) = \alpha_1 x$	$x, \arcsin(z / \alpha_0)$

де ε – похибка вимірювань. Її лінеаризована форма матиме вигляд:

$$\ln z = \ln \alpha_0 + \alpha_1 \ln x + \varepsilon', \tag{5.16}$$

де ε' є невідомою випадковою величиною. Тому використання як лінеаризованої форми виразу

$$\ln z = \ln \alpha_0 + \alpha_1 \ln x \tag{5.17}$$

буде коректним лише у разі, коли величина ε' є малою порівняно з іншими доданками правої частини (5.16).

5.3. Поліноміальні моделі

У багатьох випадках емпіричні залежності можна описати поліноміальними моделями вигляду

$$z = \sum_{i=1}^{q} \alpha_i x^i . \tag{5.18}$$

Згідно з принципом найменших квадратів

$$S(\alpha_0, \alpha_1, ..., \alpha_q) = \sum \left[y_i - \left(\alpha_0 + \alpha_1 x + ... + \alpha_q x^q \right) \right]^2 \to \min \tag{5.19}$$

Умова мінімуму має вигляд

$$\frac{\partial S}{\partial \alpha_0} = \frac{\partial S}{\partial \alpha_1} = ... = \frac{\partial S}{\partial \alpha_q} = 0. \tag{5.20}$$

Тоді оцінки параметрів можна одержати шляхом розв'язування системи нормальних рівнянь вигляду:

$$\begin{pmatrix} n & \sum x_i & \sum x_i^2 & ... & \sum x_i^q \\ \sum x_i & \sum x_i^2 & \sum x_i^3 & ... & \sum x_i^{q+1} \\ ... & ... & ... & ... & ... \\ \sum x_i^q & \sum x_i^{q+1} & \sum x_i^{q+2} & ... & \sum x_i^{2q} \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ ... \\ \alpha_q \end{pmatrix} = \begin{pmatrix} \sum Y_i \\ \sum Y_i x_i \\ ... \\ \sum Y_i x_i^q \end{pmatrix} \tag{5.21}$$

Зазвичай стовпці, що утворюють матрицю X, неортогональні. Це призводить до того, що за необхідності збільшення степеня полінома буває необхідним перераховувати оцінки всіх його коефіцієнтів. Тому для поліномів високих степенів більш раціональним є заміна вихідного рівняння (5.18) іншим:

$$z = \sum_{i=1}^{q} \alpha_i' \zeta_i , \tag{5.22}$$

де $\zeta_i = \zeta_i(x)$ є поліномами і-го степеня від x, які задовольняють умовам ортогональності:

$$\begin{cases} \sum_{j=1}^{n} \zeta_{ij} = 0, \quad i = 1,2,...,q; \\ \sum_{j=1}^{n} \zeta_{ij}\zeta_{i'j} = 0, \quad i \neq i', \end{cases} \tag{5.23}$$

де ζ_{ij} є і-м поліномом для точки x_j.

Квадратична форма для мінімізації за методом найменших квадратів має вигляд

$$Q = \sum_{j=1}^{n} \left(Y_j - \alpha'_0 - \alpha'_1\zeta_{1j} - ... - \alpha'_q\zeta_{qj}\right)^2. \tag{5.24}$$

Значення, що відповідають мінімуму, можна знайти, розв'язавши систему:

$$\begin{cases} \dfrac{\partial Q}{\partial \alpha'_0} = -2\sum_{j=1}^{n}\left(Y_j - \alpha'^*_0 - \alpha'^*_1\zeta_{1j} - ... - \alpha'^*_q\zeta_{qj}\right) = 0; \\ \dfrac{\partial Q}{\partial \alpha'_i} = -2(\sum_{j=1}^{n} Y_j\zeta_{ij} - \alpha'^*_0\sum_{j=1}^{n}\zeta_{ij} - \alpha'^*_1\sum_{j=1}^{n}\zeta_{1j}\zeta_{ij} - ... - \\ - \alpha'^*_i\sum_{j=1}^{n}\zeta^2_{ij} - ... - \alpha'^*_q\sum_{j=1}^{n}\zeta_{qj}\zeta_{ij}) = 0, \quad i = 1,2,...,q. \end{cases} \tag{5.25}$$

Звідси, використовуючи умови ортогональності (5.23), одержуємо:

$$\alpha'^*_0 = \overline{Y}; \quad \alpha'^*_i = \frac{\sum_{j=1}^{n} Y_j\zeta_{ij}}{\sum_{j=1}^{n}\zeta^2_{ij}}. \tag{5.26}$$

Використовуючи умови ортогональності, можна одержати явний вигляд поліномів для випадку, коли значення x змінюються з постійним кроком ω:

$$\begin{cases} \zeta_{0j} = 1; \\ \zeta_{1j} = v_j - \overline{v_j}; \\ \zeta_{2j} = \zeta_{1j}^2 - \left(n^2 - 1\right)/12, \end{cases} \tag{5.27}$$

де $v_j = \left(x_{j+1} - x_1\right)/\omega$. Поліноми вищих степенів одержують з рекурентної формули

$$\zeta_{r+1,j} = \zeta_{1j}\zeta_{rj} - \frac{r^2\left(n^2 - r^2\right)}{4\left(4r^2 - 1\right)}\zeta_{r-1,j}. \tag{5.28}$$

Розглянемо **приклад**. У табл.5.3 наведено результати вимірювання деяких величин x і y.

Таблиця 5.3

Результати вимірювання величин X та Y

J	1	2	3	4	5	6	7	8	9	10
Xj	0	10	20	30	40	50	60	70	80	90
Yj	23	29	41	60	79	88	83	61	33	27

Побудуємо модель досліджуваної залежності у вигляді полінома 5-го степеня:

$$z = \alpha_0' + \alpha_1'\zeta_1 + \alpha_2'\zeta_2 + \alpha_3'\zeta_3 + \alpha_4'\zeta_4 + \alpha_5'\zeta_5, \tag{5.29}$$

де $\zeta_i = \sum_{t=1}^{i} \beta_{i_t} x^t$ – поліноми i-го степеня, які задовольняють умовам ортогональності. Алгоритм розрахунку значень ζ_{ij} і коефіцієнтів α_i

реалізований у таблиці 5.4.

Таблиця 5.4

Розрахунок регресійного полінома 5-го степеня

j	x_j	y_j	y_j^*	$(y_j - y_j^*)^2$	v_j	ζ_{0j}	ζ_{1j}	ζ_{2j}	ζ_{3j}	ζ_{4j}	ζ_{5j}
1	0	23	22.99	0.00020	1	1	-4.5	12	-25.2	43.2	-60
2	10	29	29.00	0.00002	2	1	-3.5	4	8.4	-52.8	140
3	20	41	41.13	0.01584	3	1	-2.5	-2	21	-40.8	-10
4	30	60	59.86	0.01956	4	1	-1.5	-6	18.6	7.2	-110
5	40	79	78.70	0.09183	5	1	-0.5	-8	7.2	43.2	-60
6	50	88	88.64	0.41391	6	1	0.5	-8	-7.2	43.2	60
7	60	83	82.76	0.05653	7	1	1.5	-6	-18.6	7.2	110
8	70	61	60.71	0.08355	8	1	2.5	-2	-21	-40.8	10
9	80	33	33.28	0.07824	9	1	3.5	4	-8.4	-52.8	-140
10	90	27	26.93	0.00489	10	1	4.5	12	25.2	43.2	60
Σ				0.76457	5.5	10	82.5	528	3088.8	16474	78000

табл.5.4 (продовження)

Розрахунок регресійного полінома 5-го степеня

j	$y_j\zeta_{0j}$	$y_j\zeta_{1j}$	$y_j\zeta_{2j}$	$y_j\zeta_{3j}$	$y_j\zeta_{4j}$	$y_j\zeta_{5j}$
1	23	-103.5	276	-579.6	994	-1380
2	29	-101.5	116	243.6	-1531	4060
3	41	-102.5	-82	861	-1673	-410
4	60	-90	-360	1116	432	-6600
5	79	-39.5	-632	568.8	3413	-4740
6	88	44	-704	-633.6	3802	5280
7	83	124.5	-498	-1543.8	598	9130
8	61	152.5	-122	-1281	-2489	610
9	33	115.5	132	-277.2	-1742	-4620
10	27	121.5	324	680.4	1166	1620
Σ	524	121	-1550	-845.4	2968.8	2950
α'_i	52.40	1.467	-2.936	-0.274	0.180	0.038
$\sigma^2_{\alpha'_i}$	0.02549	0.003089	0.000483	0.000083	0.000015	0.000003

Легко перевірити, що для одержаних даних виконуються умови ортогональності, тобто суми значень ζ_{ij} у кожному рядку і суми за і добутків вигляду $\zeta_{ij}\zeta_{kj}(i \neq k)$ дорівнюють нулю. За даними таблиці розраховуємо коефіцієнти α_i' (передостанній рядок табл.5.4). Оцінками дисперсії цих коефіцієнтів є величини $\sigma^2_{\alpha'_i}$ (останній рядок табл.5.4)

$$\sigma^2_{\alpha_i'} = \frac{\sum_{i=1}^{n}\left(y_i - y_i^*\right)^2}{(n-q-1)\sum_{j=1}^{n}\zeta_{ij}^2}. \qquad (5.30)$$

Використовуючи формули (5.29) одержуємо оцінки Y_i^* значень досліджуваної величини Y у точках $x = x_j$ які є досить близькими до її емпіричних значень (стовпці 3 і 4 таблиці 5.4).

5.4. Множинна лінійна регресія

В дійсності кожне явище визначається дією не однієї причини, а багатьох. Розглянемо найпростіший випадок, коли шукана регресійна залежність є лінійною, а кількість визначальних факторів q. Рівняння множинної регресії має вигляд

$$z = \alpha_0 + \alpha_1 x_1 + \ldots + \alpha_q x_q . \qquad (5.31)$$

Згідно з принципом **найменших квадратів**

$$\sum\left[y_i - \left(\alpha_0 + \alpha_1 x_1 + \ldots + \alpha_q x_q\right)\right]^2 \to \min . \qquad (5.32)$$

Умовою мінімуму є

$$\frac{\partial S}{\partial \alpha_0} = \frac{\partial S}{\partial \alpha_1} = ... = \frac{\partial S}{\partial \alpha_q} = 0. \qquad (5.33)$$

Звідси отримуємо систему лінійних рівнянь:

$$\begin{cases} \alpha_0 n + \alpha_1 \sum x_{1i} + \alpha_2 \sum x_{2i} + ... + \alpha_q \sum x_{qi} = \sum y_i \\ \alpha_0 \sum x_{1i} + \alpha_1 \sum x_{1i}^2 + \alpha_2 \sum x_{1i} x_{2i} + ... + \alpha_q \sum x_{1i} x_{qi} = \sum y_i x_{1i} \\ ..... \\ \alpha_0 \sum x_{qi} + \alpha_1 \sum x_{qi} x_{1i} + \alpha_2 \sum x_{qi} x_{2i} + ... + \alpha_q \sum x_{qi}^2 = \sum y_i x_{qi} \end{cases}. \qquad (5.34)$$

Розв'язавши систему рівнянь (5.34) отримаємо шукані коефіцієнти регресії $\alpha_0, \alpha_1, ..., \alpha_q$. Тепер для вибірки обсягом n ми можемо записати систему n лінійних рівнянь.

$$\begin{cases} y_1 = \alpha_0 + \alpha_1 x_{11} + ... + \alpha_q x_{1q} + \varepsilon_1 \\ y_2 = \alpha_0 + \alpha_1 x_{21} + ... + \alpha_q x_{2q} + \varepsilon_2 \\ ..... \\ y_n = \alpha_0 + \alpha_1 x_{n1} + ... + \alpha_q x_{nq} + \varepsilon_n \end{cases}. \qquad (5.35)$$

Тут ε_i - залишок (нев'язка, неузгодження). Середнім значенням залишку ε_i є число 0. Середнє квадратичне відхилення залишку

$$\sigma_{çàë} = \sqrt{\frac{1}{n-2} \sum \left(y_i - z(x_i) \right)^2} \qquad (5.36)$$

є мірою точності передбачення значення y_i формулою

$$z_i = \alpha_0 + \alpha_1 x_{i1} + ... + \alpha_q x_{iq.} \qquad (5.37)$$

Продемонструємо **векторно-матричний підхід** до лінійної множинної регресії. У матричній формі система (5.35) запишеться так

$$\vec{y} = x\alpha + \vec{\varepsilon} \tag{5.38}$$

де

$$\vec{y} = \begin{pmatrix} y_1 \\ y_2 \\ .... \\ y_n \end{pmatrix}; \quad \alpha = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ .... \\ \alpha_n \end{pmatrix}; \quad \vec{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ .... \\ \varepsilon_n \end{pmatrix}. \tag{5.39}$$

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & ... & x_{1q} \\ 1 & x_{21} & x_{22} & ... & x_{2q} \\ 1 & x_{31} & x_{32} & ... & x_{3q} \\ ... & ... & ... & ... & ... \\ 1 & x_{n1} & x_{n2} & ... & x_{nq} \end{pmatrix} \tag{5.40}$$

Матрицю X розміром $n \times q$ називають регресійною, а елементи x_{ij}, цієї матриці – регресорами. Позначимо X^T матрицю, транспоновану до X. Розглянемо наступну матрицю В

$$B = \left(X^T X\right)^{-1} \tag{5.41}$$

Тоді вектор коефіцієнтів регресії шукають як розв'язок рівняння

$$\vec{\alpha} = \left(X^T X\right)^{-1} X^T \vec{y} = BX^T \vec{y}\,. \tag{5.42}$$

Для перевірки **статистичної значимості** коефіцієнтів множинної регресії застосовують наступний алгоритм:

1. Обчислити стандартні помилки коефіцієнтів

$$\sigma_{\alpha i} = b_{ii}\,, \tag{5.43}$$

де b_{ii} - діагональні елементи матриці (5.41).

2. Визначити критичне значення t –критерію Стьюдента

$$t_{kp} = t(\alpha, n-m-1). \quad (5.44)$$

3. Розрахувати t критерій Стьюдента

$$t_i = \frac{\alpha_i}{\sigma_{\alpha i}}. \quad (5.45)$$

4. Якщо $|t_i| > |t_{kp}|$, оцінка коефіцієнта α_i є статистично значимою. У противному разі відповідну змінну необхідно виключити з рівнямня регресії і розрахунок регресії виконати заново.
5. Межі довірчих інтервалів для коефіцієнтів регресії будуть наступними:

$$\alpha_i - t_{kp}\sigma_{\alpha i} \le \alpha_i \le \alpha_i + t_{kp}\sigma_{\alpha i}. \quad (5.46)$$

5.5. ЗАВДАННЯ ДО РОЗДІЛУ 5

Завдання 1. Лінійний регресійний аналіз

За даними таблиці для незгрупованих та для згрупованих даних:

1. знайти параметри регресійної прямої методом найменших квадратів;
2. побудувати регресійну пряму, смугу відхилення та довірчий інтервал.

Варіанти

1.

X	0,0	0,8	1,7	2,6	3,4	4,2	5,1	6,0
Y	4,0	4,8	5,8	6,9	3,0	9,2	10,4	11,8

2.

X	3,0	3,5	4,1	4,7	5,2	5,8	6,4
Y	4,9	5,9	7,0	8,1	9,3	10,5	11,7

3.

X	1,0	1,4	1,8	2,2	2,7	3,1	3,5	4,0
Y	3,0	4,0	5,1	6,4	7,8	9,2	10,8	12,4

4.

X	0,0	1,0	2,0	3,0	4,0	5,0	6,0	7,0
Y	3,0	4,0	5,1	6,4	7,8	9,2	10,8	12,4

5.

X	2,0	3,0	4,0	5,0	6,0	7,0	8,0	9,0
Y	4,1	6,4	8,7	11,2	13,8	16,4	19,2	22,0

6.

X	2,0	2,2	2,5	2,8	3,1	3,4	3,7
Y	1,4	1,7	1,9	2,2	2,4	2,7	3,0

7.

X	1,0	1,8	2,7	3,5	4,4	5,2	6,1
Y	0,3	0,7	1,2	1,7	2,3	3,0	3,7

Завдання 2. Нелінійний регресійний аналіз

За даними таблиці:

1. лінеаризувати задану функцію;
2. знайти параметри регресійної прямої методом найменших квадратів;
3. побудувати регресійну пряму, смугу відхилення та довірчий інтервал.

Варіанти:

$$y = a_0 \cdot 3^{a_1 x}$$

1.

x	1.00	1.11	1.22	1.32	1.44	1.54
y	6.00	6.76	7.64	8.51	9.69	10.85

2.

x	1.10	1.22	1.34	1.44	1.58	1.69
y	13.34	15.23	17.41	19.45	22.69	25.59

$$y = a_0 \cdot 5^{a_1 x}$$

3.

x	1.05	1.16	1.26	1.39	1.50	1.60
y	117.42	167.35	230.89	350.91	499.96	689.83

4.

x	0.10	0.24	0.36	0.46	0.57	0.78
y	2.72	4.31	6.33	8.79	12.49	19.85

$$y = a_0 \cdot e^{a_1 x}$$

5.

X	0.10	0.23	0.35	0.49	0.69	0.74
Y	4.40	5.01	5.66	6.49	7.29	8.39

6.

X	1.05	1.19	1.29	12.43	1.57	1.71
Y	16.31	21.61	26.40	34.89	46.22	61.15

$$y = a_0 \cdot a_1 \cdot \ln x$$

7.

X	1.52	1.62	1.77	1.91	2.03	2.21
Y	5.44	5.51	5.53	5.63	5.73	5.80

8.

X	1.00	1.11	1.23	1.35	1.54	1.67
Y	5.02	5.34	5.60	5.92	6.29	6.49

Питання для самоконтролю

11. В чому полягає суть регресійного аналізу?
12. Яким методом можна одержати найкращі оцінки параметрів регресійного аналізу?
13. В чому суть методу найменших квадратів, алгоритм методу?
14. Який критерій використовують, щоб перевірити відповідність математичної моделі експериментальним даним?
15. Який алгоритм використовують для перевірки статистичної значимості коефіцієнтів лінійної регресії?
16. В чому полягає лінеаризація нелінійних залежностей? Наведіть приклади.
17. Який вигляд поліноміальної моделі?
18. В яких випадках використовується лінійна множинна регресія?
19. В чому полягає векторно-матричний підхід до лінійної множинної регресії?

6. ФАКТОРНИЙ АНАЛІЗ

При дослідженні складних систем можливість безпосереднього вимірювання величин, що визначають їх властивості (**факторів**), часто буває відсутньою. Більше того, нерідко є невідомими кількість та природа цих факторів. Але можуть вимірюватися інші величини, що залежать від них. Якщо невідомий фактор впливає на декілька вимірюваних ознак, останні виявляють зв'язок, наприклад корельованість між собою. Тому загальна кількість факторів може бути значно меншою, ніж кількість вимірюваних змінних. Для виявлення таких факторів використовують факторний аналіз. Необхідність зменшення кількості факторів може бути зумовленою вимогами забезпечення збіжності алгоритмів, скорочення ресурсів пам'яті комп'ютера та часу, потрібного для їх обробки, бажанням візуалізувати отримані результати.

Формально задачу можна записати у такій спосіб. Є масив n-вимірних спостережень.

$$Y_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \dots \\ y_{in} \end{pmatrix}, \text{ i=1,2,…,p.}$$

Необхідно подати кожне із спостережень у вигляді нового вектора такого, що його розмірність r є істотно нижчою, ніж n.

$$L_i = \begin{pmatrix} l_{i1} \\ l_{i2} \\ \dots \\ l_{ir} \end{pmatrix},$$

Першим етапом факторного аналізу зазвичай є вибір нових ознак (факторів), які є лінійними комбінаціями старих і відображають переважну частку загальної мінливості вихідних даних. Кожен з факторів являє собою зважену суму групи тісно скорельованих між собою ознак

$$l_i = f_{i1}y_1 + f_{i2}y_2 + ... + f_{in}y_n \qquad (6.1)$$

Тут f_{ij} - ваги (факторні навантаження). Другим етапом є обертання факторів з метою спрощення їх інтерпретації.

Як правило, основним об'єктом дослідження методами факторного аналізу є кореляційна матриця, побудована із застосуванням коефіцієнта кореляції Пірсона для кількісних ознак. Основною вимогою до цієї матриці є її додатна напіввизначеність. Згідно з умовами Сильвестра для цього достатньо, щоб усі її головні мінори були невід'ємними. З додатної напіввизначеності кореляційної матриці випливає невід'ємність усіх її власних значень.

Методами факторного аналізу вирішують три основні групи проблем:

- пошук передбачуваних неявних закономірностей, що визначаються впливом зовнішніх або внутрішніх чинників на досліджуваний процес;
- виявлення та вивчення статистичного зв'язку ознак з факторами або головними компонентами;
- стискування інформації шляхом опису процесу за допомогою узагальнених факторів, або головних компонент, кількість яких є меншою за кількість обраних спочатку ознак (параметрів) але

достатньою для забезпечення відтворення кореляційної матриці з потрібною точністю.

У літературі вирізняють **R-техніку** факторного аналізу, яка передбачає розрахунок коефіцієнтів кореляції, що утворюють матрицю, між параметрами (ознаками), та **Q-техніку**, згідно з якою вивчають кореляцію між станами об'єктів, що описуються векторами параметрів. Основними методами факторного аналізу є методи головних факторів (компонент), максимальної правдоподібності та центроїдний. Усі вони ґрунтуються на припущенні, що досліджувана залежність є лінійною. Вихідні дані мають підпорядковуватися багатовимірному нормальному розподілу, але центроїдний метод є досить стійким до відхилень від такого закону.

Процес заміни одиниць на головній діагоналі кореляційної матриці так званими загальностями називають її **редукуванням**. За визначенням **загальність** – це сума квадратів факторних навантажень. Вона характеризує ту частку дисперсії, яка зумовлена загальними факторами (у припущенні, що повна дисперсія включає також частки, пов'язані із специфічними для даної змінної факторами, а також з похибками).

Метою факторного аналізу є зменшення кількості змінних та визначення структури взаємозв'язків між змінними (класифікація даних). З формального погляду метою факторного аналізу є одержання матриці факторного відображення. Її рядки є координатами векторів, що відповідають n змінним у r-вимірному факторному просторі. Близькість цих векторів свідчить про взаємну залежність змінних. Якщо кількість факторів перевищує одиницю, зазвичай здійснюють

обертання матриці факторного відображення для одержання більш простої структури.

Однією з проблем, що виникають при застосуванні факторного аналізу, є необхідність знаходження власних значень кореляційної матриці. Якщо вона є виродженою, ця задача може виявитися нерозв'язною. Для матриць високого порядку може відбуватися втрата значущості у процесі обчислень. У певних випадках проблему виродженості можна зняти виключенням лінійно залежних параметрів. Метод Якобі дає змогу визначити власні значення і для вироджених кореляційних матриць. Але при цьому частина їх, яка дорівнює різниці між порядком та рангом матриці, буде мати значення, що не перевищують обчислювальної похибки. Завдяки цьому метод головних компонент виявляється стійкішим до аналізу відповідних даних, ніж метод максимуму правдоподібності. Водночас він є гіршим за останній з погляду можливості отримання точної оцінки загальності й досягнення повного відтворення кореляційної матриці.

6.1. Метод головних факторів

Цей метод використовують для зменшення кількості змінних. У його основі лежить припущення, що не всі змінні, які вимірювалися при дослідженні системи, є незалежними. Тому можливо формування нових змінних, що достатньо повно відображають наявну інформацію.

Спочатку розглянемо один з варіантів цього методу, відомий як **метод головних компонент**, або **компонентний аналіз**. Цей метод за свою сутністю зводиться до вибору нової ортогональної системи координат у просторі спостережень. Як першу головну компоненту оби-

рають напрям, вздовж якого масив спостережень має найбільшу дисперсію. Якщо уявити набір даних у вигляді хмарки точок у просторі розмірності n, напрям головної компоненти співпадатиме з напрямком найбільшої витягнутості хмарки. Кожну наступну компоненту обирають з умови максимізації частки дисперсії, що залишилася, вздовж неї, доповненої умовою ортогональності усім раніше обраним компонентам. Із зростанням номера компоненти буде зменшуватися пов'язана з нею частка загальної дисперсії. Кількість компонент визначається значною мірою суб'єктивно, виходячи з розуміння того, яка величина загальної дисперсії відповідає випадковій мінливості, що відображає похибку вимірювань, вплив неконтрольованих випадкових чинників тощо.

Основну модель можна записати в матричному вигляді:

$$Y = LZ, \tag{6.2}$$

де Y − матриця стандартизованих вихідних даних розміру $\mathrm{p} \times \mathrm{n}$, Z − матриця факторного відображення розміру $\mathrm{r} \times \mathrm{n}$, L - матриця виду

$$L = \begin{pmatrix} l_{11} & \dots & l_{1r} \\ \dots & \dots & \dots \\ l_{p1} & \dots & l_{pr} \end{pmatrix}. \tag{6.3}$$

Матриця L − це матриця значень факторів розміру $\mathrm{p} \times \mathrm{r}$, стовпчики якої задовольняють умову ортогональності. Матрицю стандартизованих вихідних даних визначають з матриці вихідних даних X за формулою:

$$y_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j},\ i=1,2,\ldots,n;\ j=1,2,\ldots,p, \tag{6.4}$$

де x_{ij} − елемент матриці вихідних даних, $\overline{x_j}$ − середнє значення для j-го стовпчика, s_j − стандартне відхилення для j-го стовпчика. Якщо всі компоненти досліджуваного вектора X мають загальне змістовне значення і виміряні в одних і тих самих одиницях, то при переході до матриці стандартизованих вихідних даних можна обмежитися лише центруванням: $y_{ij} = x_{ij} - \overline{x}_j$.

Кореляційну матрицю розміром $p \times p$ обчислюють за формулою:

$$R = \frac{1}{p-1} YY^T. \tag{6.5}$$

На головній діагоналі кореляційної матриці R стоять значення, які дорівнюють одиниці.

Невідомими є матриці A і L. Згідно з **основною теоремою факторного аналізу**, матрицю A можна знайти з рівняння

$$R = ACA^T, \tag{6.6}$$

де C − кореляційна матриця, що відображає зв'язок між факторами. Якщо вона є одиничною, то фактори є ортогональними, в іншому разі − косокутними. Для матриці C виконується співвідношення:

$$\frac{1}{p-1} LL^T = C. \tag{6.7}$$

Далі ми будемо розглядати ортогональні фактори, для яких

$$R = AA^T . \quad (6.8)$$

Модель класичного факторного аналізу припускає, що є декілька загальних факторів та по одному характеристичному для кожної змінної фактору. Формула (6.2) є основною моделлю факторного аналізу для методу головних компонент. Кількість головних компонент завжди є меншою або рівною кількості змінних. Для методу головних факторів основну модель записують у вигляді:

$$Y = FL^+, \quad (6.9)$$

де F – повна факторна матриця, L^+ – матриця значень факторів (у тому числі характеристичних). Матриця Y має розмір $n \times p$, матриця F - $n \times (r + n)$, матриця L^+ – $(r \times n) \times p$. Матрицю F можна подати як

$$F = A + U \quad (6.10)$$

де A та U – відповідно, матриці навантажень загальних та характеристичних факторів, які мають розмірність F. Матриця A (точніше її ненульова частина розміром $n \times r$) є матрицею факторного відображення.

Повна дисперсія змінної складається із **загальності** h_i^2 і **характерності** u_i^2. Значення першої є меншим або рівним одиниці і відображає частину загальної дисперсії змінної, що припадає на головні фактори. Значення другої відображає ту частину дисперсії, яка визначається характеристичними факторами. При цьому:

$$u_i^2 + h_i^2 = 1. \quad (6.11)$$

Частина матриці U розміром $n \times r$ є нульовою, а її інша частина розміром $m \times m$ є діагональною матрицею, ненульові елементи якої є квадратними коренями з відповідних характерностей. Використовуючи основну теорему факторного аналізу і вважаючи фактори ортогональними, можна записати:

$$R = FF^T = R_h + U^2, \tag{6.12}$$

де $R_h = AA^T; U^2 = UU^T$.

На відміну від методу головних компонент, у методі головних факторів на головній діагоналі кореляційної матриці R_h необхідно поставити значення загальностей. Процес їх оцінювання називають редукцією кореляційної матриці. Найчастіше застосовують такі методи оцінювання загальностей:

- метод найбільшої кореляції;
- застосування коефіцієнта множинної кореляції; у цьому разі загальності розраховують як

$$h_i^2 = 1 - \frac{1}{\rho_{ii}}, \tag{6.13}$$

де ρ_{ii} – діагональні елементи матриці, що є оберненою до R_h;

- застосування середніх за стовпцем кореляційної матриці коефіцієнтів кореляції;
- метод тріад.

Для заміни діагональних елементів кореляційної матриці оцінками загальностей необхідно, щоб редукована матриця була матрицею Грама. Але при застосуванні коефіцієнтів множинної кореляції ця властивість часто втрачається.

Матрицю факторного відображення в методах головних факторів та головних компонент визначають за методом множників Лагранжа. Фактори є пропорційними власним векторам матриці R (або R_h, залежно від методу). Стандартну проблему власних значень можна записати у вигляді рівняння:

$$(\mathrm{R} - \lambda_k \mathrm{I}) = 0\,, \qquad (6.14)$$

де λ_k (k = 1,2, ..., n) − власні значення матриці R. Матриця R є дійсною та симетричною, тому для вирішення проблеми власних значень можна застосовувати ефективні і стійкі алгоритми. Спосіб вимірювання головних компонент заснований на використанні основної моделі (6.2), з якої випливає: $\left(A^T A\right)^{-1} A^T Y = \left(A^T A\right)^{-1} A^T AL$ або

$$L = \left(A^T A\right)^{-1} A^T Y\,. \qquad (6.15)$$

Застосування факторного аналізу проілюструємо наступним **прикладом.**

Таблиця 6.1

Матриця кореляцій між ознаками

	Відносини з Росією	Приватна власність	Звільнення цін	Російська мова	Банкрутство
Відносини з Росією	1.000	-0.274	-0.243	0.422	-0.208
Приватна власність	-0.274	1.000	0.434	-0.176	0.408
Звільнення цін	-0.243	0.434	1.000	-0.160	0.441
Російська мова	0.422	-0.176	-0.160	1.000	-0.088
Банкрутство	-0.208	0.408	0.441	-0.088	1.000

У 1998 році Київський міжнародний інститут соціології провів опитування серед населення України за анкетою, яка включала в себе 5 питань, які стосувалися економічної політики держави, відносин з Росією та відношення до російської мови. Результати опитування ілюструє кореляційна матриця (таблиця 6.1).

У результаті застосування факторного аналізу було виділено 2 фактори, які пояснюють 66% варіації вихідних ознак, і для кожного фактора розраховані факторні навантаження (тобто коефіцієнти f_{ji} у формулі 6.1). Зі змістовної точки зору матриця факторних навантажень (таблиця 6.2) є основним результатом факторного аналізу, що може безпосередньо інтерпретуватися.

Таблиця 6.2

Матриця факторних навантажень

Вихідні ознаки	Фактори	
	1	**2**
Відносини з Росією	-0.624	0.540
Приватна власність	0.732	0.238
Звільнення цін	0.730	0.310
Російська мова	-0.489	0.718
Банкрутство	0.682	0.422

Як бачимо, перший фактор має найбільші навантаження на ознаки 2 і 3. Чим вище ступінь підтримки введення приватної власності і ступінь усвідомлення необхідності звільнення цін – тим вище значення першого фактора. Можна інтерпретувати перший фактор як ступінь підтримки ринкових реформ: чим вище значення фактора – тим вище

підтримка реформ. Другий фактор має найвищі навантаження на ознаки, що характеризують відносини з Росією і ставлення до російської мови. Умовно можна назвати цей фактор ”проросійськістю”.

У розглянутому прикладі нам досить легко вдалося інтерпретувати два виділених нами фактори. Але при цьому ознака, що характеризує відносини з Росією, має вище навантаження на перший фактор, ніж на другий, і це не є логічним. Прояснити ситуацію допоможе **обертання факторів**. Для ілюстрації обертання перейдемо до графічної інтерпретації факторів. Кожна ознака може бути представлена як вектор у просторі факторів: на осі X відкладається навантаження першого фактора на дану ознаку, на осі Y – навантаження другого фактора (рис.6.1). При повороті системи координат співвідношення між ознаками (кути) не зміняться, не зміниться, також, і частка поясненої дисперсії. Що зміниться – це навантаження факторів на кожну з ознак. Мета обертання – домогтися простішої структури, тобто зробити так, щоб фактори легше інтерпретувалися. Один з основних алгоритмів обертання – **варімакс**. Цей метод мінімізує кількість змінних, які мають високі навантаження на фактори.

У таблиці 6.3 наведено матрицю факторних навантажень для нашого прикладу після обертання. Як бачимо, поділ на фактори став чіткішим, всі ознаки, що визначають інтерпретацію першого фактора, мають навантаження вище за 0.7 на перший фактор і нижче за 0.2 на другий фактор. Ознаки, які визначають інтерпретацію другого фактора, мають високі навантаження на другий фактор і низькі на перший. Графічна ілюстрація результату обертання представлена на рис.6.2.

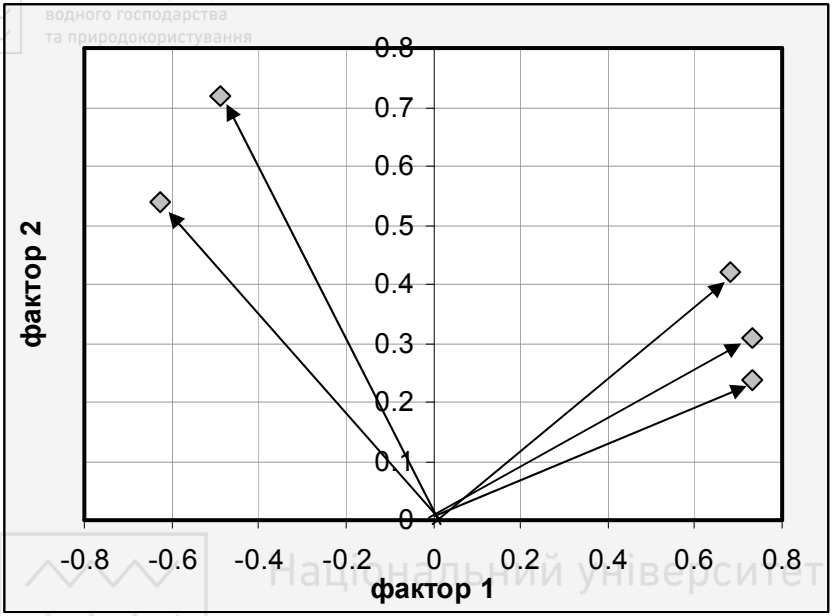

Рис.6.1. Вектори ознак у просторі факторів (до обертання).

Таблиця 6.3

Матриця факторних навантажень (після обертання)

Вихідні ознаки	Фактори	
	1	**2**
Відносини з Росією	-0.232	0.792
Приватна власність	0.782	-0.135
Звільнення цін	0.744	-0.198
Російська мова	-0.021	0.869
Банкрутство	0.802	-0.015

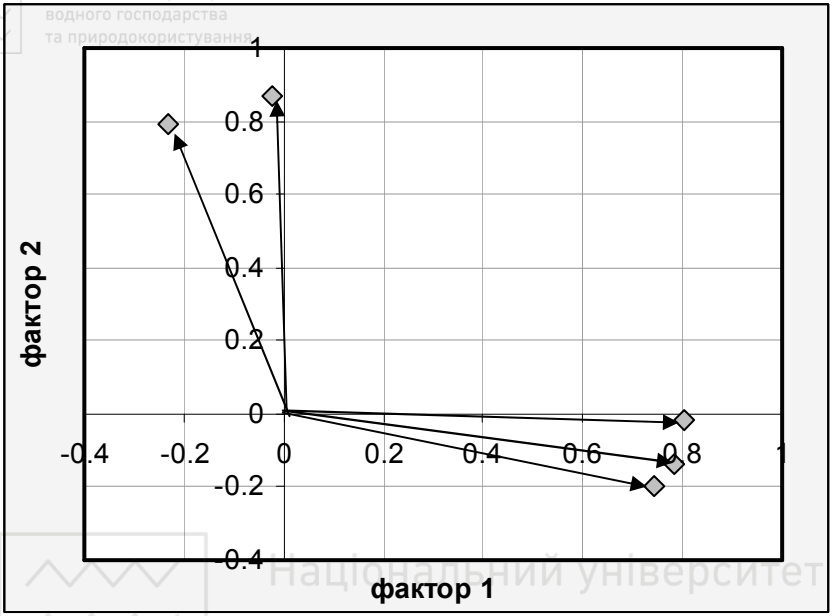

Рис.6.2. Вектори ознак у просторі факторів (після обертання)

6.2. Метод максимуму правдоподібності

У методі максимуму правдоподібності оцінювання загальностей до безпосереднього застосування алгоритму факторного аналізу не здійснюють. Їх визначають за результатами обчислень із умови повного відтворення кореляційної матриці. За будь-якої кількості факторів, що розглядаються, цей метод дає можливість відтворити її з точністю до похибки обчислень. Якщо кількість факторів дорівнює кількості параметрів, то оцінки загальностей будуть збігатися із загальностями нередукованої кореляційної матриці, тобто дорівнювати одиниці. Основним недоліком методу є його нестійкість при використанні окремих типів даних, зокрема даних, що містять лінійно залежні вектори. Це призводить до виродження матриці характерностей. У такому разі мо-

жна спробувати зняти проблему шляхом виключення із розгляду лінійно залежних параметрів або застосування методу головних факторів.

6.3. Центроїдний метод

У попередніх двох методах максимізується квадратичний критерій. На відміну від них, у центроїдному методі максимізують модульний критерій. З погляду змістовної інтерпретації ці критерії є еквівалентними. Перевагою методу, внаслідок його непараметричності, є відносна стійкість до відхилень від нормального розподілу.

Існує декілька схем обчислень, які відрізняються одна від одної операцією **відбиття**. Відбиттю підлягають змінні, які мають найбільшу кількість від'ємних значень. Перший варіант: спочатку вибирається змінна, що має найбільшу кількість від'ємних значень, потім наступна за кількістю і т.д. Другий варіант: спочатку вибирається змінна з номером стовпця, що має максимальну за модулем від'ємну суму значень, а потім інші змінні за тим же критерієм.

За Терстоуном, максимальна кількість факторів, які можуть бути однозначно визначені за наявності n змінних:

$$n = \frac{2p+1-\sqrt{8p+1}}{2}. \qquad (6.16)$$

Її можна оцінити також із співвідношення

$$p+n < (p-n)^2. \qquad (6.17)$$

6.4. Додаткові зауваження

Осі координат, що відповідають виділеним факторам, є ортогональними. Їх напрями встановлюють послідовно за критерієм мінімуму дисперсії, що залишається. Але така процедура ускладнює змістову інтерпретацію одержаних факторів. Тому на наступному етапі, обертаючи систему координат відносно її початку, отримують так звану найпростішу факторну структуру. Існує декілька критеріїв оцінки максимальної кількості факторів, що можуть аналізуватися. Критерії, що ґрунтуються на аналізі визначників вихідної та відтвореної кореляційної матриць, часто виявляються нестійкими. Критерії, які базуються на величині власних значень кореляційної матриці, у підсумку призводять до аналізу відсотка дисперсії, виділеної факторами. Усі загальні фактори, кількість яких дорівнює кількості параметрів, виділяють 100 % дисперсії. Якщо сума відсотків за факторами перевищує 100 %, це свідчить про отримання від'ємних власних значень і, відповідно, комплексних власних векторів, що може бути наслідком некоректної редукції вихідної кореляційної матриці. Доцільно здійснювати двоетапну процедуру аналізу. На першому етапі максимальну кількість факторів не задають. Після його проведення аналізують дисперсії, оцінюють приблизну кількість факторів і проводять повторний аналіз.

6.5. ЗАВДАННЯ ДО РОЗДІЛУ 6

Завдання 1. Використовуючи результати позачергових виборів до Верховної Ради України 2007 року виконати факторний аналіз даних засобами пакету Statistica (Statistics, Multivariate Exploratory Techniques, Factor Analysis). Розглянути 2 фактори. Розрахувати навантаження на фактори і дати їх змістовне трактування. Провести обертання методом варімаксу. Виконати графічну ілюстрацію. Перед виконанням аналізу у пакеті Statistica дані транспонувати.

		ПР	БЮТ	НУНС	КПУ	БЛ
1	АР Крим	60.98	6.95	8.30	7.62	3.91
2	Вінницька	12.55	49.95	18.59	4.96	3.14
3	Волинська	6.72	57.53	19.97	2.72	4.59
4	Дніпропетровська	48.15	20.93	6.33	7.62	5.10
5	Донецька	72.05	3.93	1.63	6.05	0.86
6	Житомирська	22.41	37.00	15.13	5.80	8.29
7	Закарпатська	19.76	28.84	31.10	1.77	6.00
8	Запорізька	55.45	14.66	4.72	8.30	5.45
9	Івано-Франківська	2.95	50.54	36.69	0.78	0.97
10	Київська	13.04	53.35	15.12	2.95	5.13
11	Кіровоградська	26.99	37.57	11.68	6.43	5.54
12	Луганська	73.53	5.10	1.73	8.48	2.40
13	Львівська	4.19	50.38	36.02	1.03	1.10
14	Миколаївська	54.40	16.61	5.83	7.18	4.53
15	Одеська	52.22	13.72	6.49	6.16	5.12
16	Полтавська	24.75	37.86	14.50	6.48	4.89
17	Рівненська	10.41	50.94	20.76	2.40	6.11
18	Сумська	15.69	44.45	20.75	5.79	3.33
19	Тернопільська	3.01	51.43	35.07	0.69	1.55
20	Харківська	49.61	16.36	8.11	8.28	4.55
21	Херсонська	43.23	23.06	9.07	9.09	3.66
22	Хмельницька	14.05	48.14	18.41	3.95	6.62
23	Черкаська	15.49	47.00	15.24	4.86	4.93
24	Чернівецька	16.79	46.16	20.32	2.29	2.54
25	Чернігівська	20.72	41.92	14.85	6.65	4.19

Завдання 2. Використовуючи дані про економічну діяльність підприємства ТОВ «Прометей А» виконати факторний аналіз даних засобами пакету Statistica. Розглянути 3 фактори. Розрахувати навантаження на фактори і дати їх змістовне трактування. Провести обертання методом варімаксу. Виконати графічну ілюстрацію.

		2001	2002	2003	2004	2005	2006
1	Коеф рівня економ стійкості	0.30	0.43	0.41	0.46	0.49	0.54
2	Коефіцієнт залишкової вартості ОВЗ	0.40	0.47	0.43	0.39	0.35	0.25
3	Коефіцієнт зносу ОВЗ	0.17	0.23	0.27	0.32	0.38	0.45
4	Коефіцієнт оновлення ОВЗ	0.08	0.11	0.16	0.19	0.21	0.20
5	Коефіцієнт працевіддачі	20.51	26.99	20.87	21.89	19.72	26.66
6	Фондовіддача	4.47	5.40	6.59	4.90	5.25	6.53
7	Коефіцієнт автономії	0.55	0.81	0.84	0.78	0.70	0.64
8	Коефіцієнт швидколіквідності	0.22	0.52	0.55	1.11	0.65	0.87
9	Коеф маневр власн капіталу	0.21	0.51	0.44	0.44	0.45	0.39
10	Коефіцієнт співвідношення позикових і власних коштів	0.11	0.05	0.13	0.22	0.36	0.34
11	Коеф оборотності власн кап	3.58	3.76	2.89	4.68	5.33	7.22
12	Коефіцієнт оборотності матеріальних оборот коштів	2.92	4.19	3.81	5.35	5.05	6.36
13	Фондовіддача необоротних активів	4.52	7.62	5.17	8.45	9.81	11.87
14	Коефіцієнт рентабельності власного капіталу	0.01	0.08	0.00	0.01	0.06	0.11
15	Рентаб основної діяльності	0.03	0.04	0.00	0.01	0.04	0.04
16	Рентабельність продажу	0.02	0.03	0.00	0.01	0.03	0.03
17	Коеф стійкості на ринку	0.03	2.27	3.13	2.44	5.19	5.63
18	Коеф стійкості постачальника	0.67	0.70	0.69	0.75	0.81	0.86

Питання для самоконтролю

20.В чому суть факторного аналізу?
21.Які основні проблеми розв'язують методами факторного аналізу?
22.В чому полягає метод головних компонент?
23.Які методи оцінювання загальностей використовують найчастіше?
24.Суть методу максимуму правдоподібності.
25.В чому полягає центроїдний метод?

7. ЗАДАЧІ ТА МЕТОДИ КЛАСИФІКАЦІЇ ДАНИХ

У загальному випадку **класифікацією** (**розпізнаванням образів**) називають поділ досліджуваної сукупності об'єктів на однорідні у певному розумінні групи (класи) або зарахування кожного із заданої множини об'єктів до деякого із заздалегідь відомих класів. При побудові методів класифікації зазвичай прагнуть мінімізувати імовірність неправильної класифікації. Для цього можна побудувати функцію втрат $c(j \mid i)$, що характеризує втрати від помилкового зарахування об'єкта і-го класу до j-го класу. При $i = j$ беруть $c(j|i)=0$, а при $i \neq j$ – $c(j|i)>0$. Якщо кількість таких помилок є m(j|i), то загальні втрати:

$$C_n = \sum_{i=1}^{k} \sum_{j=1}^{k} c(j|i) m(j|i), \qquad (7.1)$$

де n – кількість класифікованих об'єктів, k – кількість класів.

Величина C_n залежить від n. Для усунення такої залежності можна ввести питомі втрати (у розрахунку на один об'єкт) і перейти до границі при $n \to \infty$. Тоді:

$$C = \lim_{n\to\infty} (C_n / n) = \sum_{i=1}^{k} \pi_i \sum_{j=1}^{k} c(j|i) m(j|i) = \sum_{i=1}^{k} \pi_i C^{(i)}, \qquad (7.2)$$

де π_i – апріорна імовірність (питома вага) і-го класу, $\mathrm{P(j\,|\,i)}$ – імовірність помилкового зарахування об'єкта і-го класу до j-го класу, величина $C^{(i)}$ визначає середні втрати від неправильної класифікації об'єктів і-го класу. У багатьох випадках втрати є однаковими для будь-якої пари і та j, тобто:

$$c(j \mid i) = c_0 = const(i \neq j). \tag{7.3}$$

У цьому разі мінімізація функції втрат еквівалентна максимізації імовірності правильної класифікації, яка дорівнює $\sum_{i=1}^{k} \pi_i P(i \mid i)$. Тому при побудові процедур класифікації часто розв'язують задачу мінімізації імовірності неправильної класифікації

$$\frac{C}{c_0} = 1 - \sum_{i=1}^{k} \pi_i P(i|i). \tag{7.4}$$

У **методах розпізнавання образів без навчання** програмна система на основі визначених нею самою критеріїв здійснює класифікацію певних об'єктів (образів). У деяких випадках можуть бути задані окремі параметри, але розподіл об'єктів за класами на основі цих параметрів виконується автоматично.

Параметричні методи розпізнавання образів без навчання використовують при класифікації об'єктів $O_1, O_2, ..., O_n$, якщо апріорна інформація про класи може бути подана у вигляді суміші параметрично заданих одномодальних функцій щільності розподілу імовірностей $f_j(X, \Theta_j)$ j=1,...,k з невідомими значеннями векторних параметрів Θ_i. Функцію f(X) називають дискретною або неперервною **сумішшю ймовірнісних розподілів**, якщо її можна записати у вигляді, відповідно:

$$f(X) = \sum_{j=1}^{k} \pi_j f_j(X, \Theta(j)) \tag{7.5}$$

або

$$f(X) = \int f_{\omega}(X, \Theta(\omega))\pi(\omega)d\omega. \qquad (7.6)$$

У задачах класифікації зазвичай розглядають дискретні суміші.

Розв'язання задачі розщеплення суміші розподілів передбачає побудову статистичних оцінок для кількості компонентів суміші (класів) k, їх питомих ваг (апріорних імовірностей) π_j та функцій $f_j(X, \Theta(j))$ для кожного із компонентів за наявною вибіркою спостережень $X_1, X_2, ..., X_n$.

Основною ідеєю більшості методів розв'язання цієї задачі є прагнення зарахувати спостереження X_i до того класу, для якого функція правдоподібності буде максимальною. У найпростішому випадку із попередніх досліджень можуть бути відомими кількість класів, їх апріорні імовірності та параметричний вигляд функцій щільності імовірності $f_j(X, \Theta(j))$, але невідомі значення параметрів Θ_j. Якщо при цьому є вибірки, що навчають, то ми отримуємо задачу параметричного дискримінантного аналізу, яка більш докладно розглядається нижче. Якщо ж таких вибірок немає, то значення параметрів необхідно оцінити за наявною вибіркою спостережень за допомогою одного із статистичних методів оцінювання параметрів − максимальної правдоподібності, моментів тощо. Після отримання оцінок невідомих параметрів можна застосовувати схему параметричного дискримінантного аналізу. Аналогічний підхід використовують і в більш складних випадках, коли кількість класів та їх апріорні ймовірності є невідомими. У цьому разі їх також необхідно оцінити за наявною вибіркою.

У непараметричному випадку ми не маємо інформації про загальний вигляд функцій $f_j\left(X,\Theta\left(j\right)\right)$. Ми можемо знати лише окремі загальні відомості про них: компактність або обмеженість діапазонів змінювання компонент класифікованих багатовимірних спостережень, неперервність або гладкість відповідних законів розподілу ймовірностей тощо. Вихідні дані зазвичай подають у вигляді матриці спостережень, яка містить значення всіх ознак для кожного із досліджуваних об'єктів, або матриці подібності, що містить попарні відстані між класифікованими спостереженнями. Для формалізації задачі класифікації кожний об'єкт зручно інтерпретувати як точку в багатовимірному просторі ознак. Геометрична близькість точок у такому просторі відповідає близькості досліджуваних об'єктів з погляду досліджуваних властивостей. Залежно від мети дослідження задачу класифікації можна сформулювати як розбиття аналізованих об'єктів на певну кількість груп, всередині яких вони розташовані на порівняно малій відстані один від одного, або як виявлення природного розшарування сукупності, що вивчається, на окремі кластери. Другу задачу можна також сформулювати як визначення областей підвищеної густини точок, які відповідають наявним спостереженням. Перша задача завжди має розв'язок, а друга може не мати розв'язку, що відповідає відсутності природного розшарування досліджуваних об'єктів (наприклад, вони утворюють один кластер, або відповідні точки рівномірно заповнюють весь простір ознак).

7.1. Кластерний аналіз

Класичними **непараметричними методами** розпізнавання образів без навчання є методи **кластерного аналізу** (**таксономії**). З їх допомогою вирішують проблему такого розбиття (класифікації, кластеризації) множини об'єктів, за якого всі об'єкти, що належать до одного класу, є більш подібними один до одного, ніж до об'єктів інших класів. З формального погляду основне завдання методів кластерного аналізу можна сформулювати, як визначення класів еквівалентності та рознесення за ними досліджуваних об'єктів. Під **класом**, як правило, розуміють генеральну сукупність, що описується одномодальною функцією щільності імовірності f(**X**) або, у випадку дискретних ознак, − одномодальним полігоном імовірностей. Номери класів не мають змістовного значення й використовуються лише для того, щоб відрізняти їх один від одного.

Для формування кластерів застосовують міри подібності та відмінності даних, які можуть бути поділені на три основних види:

- **міри подібності (відмінності) типу "відстань"** (при їх застосуванні об'єкти вважають тим більше подібними один до одного, чим меншою є відстань між ними);
- **міри подібності типу "зв'язок"** (у цьому разі об'єкти вважають тим більше подібними, чим сильнішим є зв'язок між ними);
- **інформаційна статистика**.

Як міри відстані найчастіше використовують евклідову та манхеттенську відстані, супремум-норму, а також відстань Махаланобіса. Вони відображають усе різноманіття підходів до цієї проблеми. Евклідова метрика традиційно використовується як міра відстані. Ман-

хеттенська відстань є найбільш відомою з класу метрик Мінковського. Відстань Махаланобіса за допомогою дисперсійно-коваріаційної матриці пов'язана з кореляціями змінних і широко використовується у кластерному аналізі та інших методах аналізу даних. Вказані міри подібності можуть бути застосовані при реалізації методів ближнього зв'язку, середнього зв'язку Кінга, Ворда, k-середніх Мак-Квіна.

Як міри зв'язку для кількісних ознак можна обирати коефіцієнт кореляції Пірсона і дисперсію-коваріацію. Для порядкових ознак призначені коефіцієнти рангової кореляції Спірмена і Кендалла. Вони можуть бути перетворені до мір подібності типу "відстань" за допомогою формул:

$$d_{ij} = 1 - \rho s; \quad d_{ij} = 1 - \tau . \tag{7.7}$$

У цьому разі їх називають, відповідно, **відстанями Спірмена** і **Кендалла**.

Для дихотомічних ознак і ознак, що розміщуються в таблицях спряженості, використовують хеммінгову відстань, показник Жаккара, простий коефіцієнт зустрічальності, показник Рассела й Рао, коефіцієнт асоціації Юла, коефіцієнт спряженості Бравайса. Розглянуті показники (крім хеммінгової відстані) можна перетворити у відстані, віднімаючи обчислені значення від одиниці.

Для змішаних ознак користуються коефіцієнтом Гауера.

Перелічені міри зв'язку застосовують у методах ближнього зв'язку, кореляційних плеяд та максимального кореляційного шляху. Зазвичай за допомогою першого з цих методів класифікують об'єкти, а за допомогою двох інших − параметри. Але шляхом транспонування матриці вихідних даних можна легко змінити тип класифікації на проти-

лежний. Результати класифікації різними методами, як правило, принципово не відрізняються.

Вибір метрики може істотно впливати на результати аналізу. Тому для кожної конкретної задачі його необхідно здійснювати окремо. При цьому треба враховувати головні цілі дослідження, фізичну та статистичну природу вихідних даних, повноту апріорних відомостей про тип функцій розподілу ймовірності. Зокрема, якщо кластери можна інтерпретувати як нормальні генеральні сукупності з однією і тією самою коваріаційною матрицею, то доцільно обирати відстані типу Махаланобіса, окремими випадками якої є евклідова, зважена евклідова та хеммінгова відстані.

Дивергенція між двома сукупностями *i* та *j* (**інформаційна відстань Каллбека**) може бути розрахована за формулою:

$$J_{ij} = \frac{1}{2} tr\left[\left(C_i - C_j\right)\left(C_j^{-1} - C_i^{-1}\right)\right] + \frac{1}{2} tr\left[\left(C_i^{-1} + C_j^{-1}\right)\left(m_i - m_j\right)\left(m_i - m_j\right)'\right], \quad (7.8)$$

де C_i, C_j − дисперсійно-коваріаційні матриці для сукупностей і та j, m_i, m_j − вектори середніх для сукупностей і та j.

Міра Махаланобіса (**узагальнена евклідова відстань**) − це відстань від точки спостереження, яка належить до класу 2, до центра ваги спостережень класу 1 у багатовимірному просторі ознак:

$$d_{ij} = \sqrt{\left(X_i - m(X_j)\right)^T \Delta^T \Sigma^{-1} \Delta \left(X_i - m(X_j)\right)}. \quad (7.9)$$

де Δ − симетрична невід'ємно визначена матриця вагових коефіцієнтів, яку найчастіше обирають діагональною, а Σ − коваріаційна матриця генеральної сукупності, до якої належать спостереження. Іноді за-

мість міри Махаланобіса використовують її квадрат. Розглянемо деякі з інших мір відстані детальніше.

Евклідова відстань (**евклідова метрика**) є відомою із загальноматематичних дисциплін і визначається формулою:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}. \qquad (7.10)$$

Вона збігається із відстанню Махалонобіса у випадку, коли незалежні змінні є некорельованими. Її доцільно обирати, якщо:

- спостереження належать до генеральних сукупностей, які підпорядковуються багатовимірним нормальним законам, а компоненти вектора спостережень є незалежними і мають одну й ту саму дисперсію;
- компоненти вектора спостережень є однорідними з погляду змістової інтерпретації та однаково важливими для класифікації;
- простір ознак має розмірність 1, 2 або 3, і поняття близькості об'єктів у цьому просторі збігається із звичайною геометричною близькістю.

Поряд із евклідовою відстанню як міру близькості часто використовують її квадрат.

Зважену евклідову відстань розраховують за формулою:

$$d_{ij}^{*} = \sqrt{\sum_{k=1}^{p} \omega_k \left(x_{ik} - x_{jk}\right)^2}. \qquad (7.11)$$

де ω_k − невід'ємні вагові коефіцієнти, які є пропорційними ступеню важливості критерію з погляду класифікації. Зазвичай беруть

$0 \le \omega_k \le 1$. Визначення вагових коефіцієнтів за аналізованою вибіркою, як правило, є недоцільним, оскільки може призводити до істотних помилок. Тому рекомендують обирати вагові коефіцієнти за результатами експертних опитувань або інших незалежних попередніх досліджень.

Метрика Мінковського є узагальненням звичайної евклідової відстані:

$$d_{ij} = \sqrt[r]{\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^r}. \qquad (7.12)$$

У випадку r = 2 вона збігається з евклідовою метрикою.

У випадку r = 1 метрика Мінковського дає **манхеттенську відстань**:

$$d_{ij} = \sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|. \qquad (7.13)$$

При $r \to \infty$ метрика Мінковського збігається із **супремум-нормою**:

$$d_{ij} = \sup\left\{ \left| x_{ik} - x_{jk} \right| \right\}, k = 1,2,..., p. \qquad (7.14)$$

Хеммінгову відстань використовують для ознак, які можуть набувати лише значення 0 або 1 і розраховують за формулою:

$$d_{ij} = \sum_{s=1}^{p} \left| x_{is} - x_{js} \right|, \qquad (7.15)$$

тобто вона збігається із кількістю значень відповідних ознак, що не збігаються, у і-го та j-го об'єктів.

При конструюванні різноманітних процедур класифікації доцільно використовувати міри близькості кластерів один до одного. Найбільш

поширеними з них є відстані, що вимірюються за принципами найближчого, і найдальшого сусідів, середнього зв'язку та за центрами ваги.

Нехай S_i – і-й кластер, n_i – кількість об'єктів у ньому, $\overline{X}(i)$ – центр ваги і-го кластера, тобто середнє арифметичне векторних спостережень, що його утворюють, ρ_{lm} – відстань між класами ℓ і m. Тоді **відстань, що вимірюється за принципом найближчого сусіда**:

$$\rho_{lm}^{\min} = \min_{X_i \in S_l; X_j \in S_m} d_{ij}; \tag{7.16}$$

відстань, що вимірюється за принципом найдальшого сусіда:

$$\rho_{lm}^{\max} = \max_{X_i \in S_l; X_j \in S_m} d_{ij}; \tag{7.17}$$

відстань, що вимірюється за принципом середнього зв'язку (середнє арифметичне всіх можливих попарних відстаней між представниками класів, які розглядаються):

$$\rho_{lm}^{m} = \frac{1}{n_l n_m} \sum_{X_i \in S_l} \sum_{X_j \in S_m} d_{ij}; \tag{7.18}$$

відстань, що вимірюється за центрами ваги:

$$\rho_{lm} = d\left(\overline{X}_l, \overline{X}_m\right). \tag{7.19}$$

Існує також узагальнена (за О.М.Колмогоровим) формула розрахунку відстаней між класами (**узагальнена К - відстань**):

$$\rho_{lm}^{K} = \left[\frac{1}{n_l n_m} \sum_{X_i \in S_l} \sum_{X_j \in S_m} d_{ij}^{\tau} \right]^{1/\tau}. \qquad (7.20)$$

При $\tau \to -\infty$ вона переходить у формулу (7.16), при $\tau \to +\infty$ – у формулу (7.17), а при $\tau = 1$ – у формулу (7.18).

Якщо S_w – є новим класом, отриманим як об'єднання класів m і q, то його узагальнену відстань від класу S_l можна розрахувати за формулою:

$$\rho_{lw}^{K} = \left[\frac{n_m \left(\rho_{lm}^{K}\right)^{\tau} + n_q \left(\rho_{lq}^{K}\right)^{\tau}}{n_m + n_q} \right]^{1/\tau}. \qquad (7.21)$$

Для перерахунку відстані між класами використовують також **загальну формулу Ланса та Вільямса**

$$\rho_{lw} = a_m \rho_{lm} + a_q \rho_{lq} + b\rho_{mq} + c\left|\rho_{lm} - \rho_{lq}\right|, \qquad (7.22)$$

де a_m, a_q, b, c – параметри, що визначають спосіб розрахунку відстані між класами. Зокрема, для відстаней ближнього зв'язку $a_m = a_q = \frac{1}{2}, b = 0, c = -\frac{1}{2}$; для відстаней далекого зв'язку $a_m = a_q = c = \frac{1}{2}, b = 0$; для відстаней середнього зв'язку:

$$a_m = \frac{n_m}{n_m + n_q}; \; a_q = \frac{n_q}{n_m + n_q}; \; b = c = 0. \qquad (7.23)$$

Для розрахунку ступеня близькості класів використовують також розглянуті вище інформаційну відстань Каллбека (у разі, коли класи

можна розглядати як багатовимірні нормальні сукупності) та відстань Махаланобіса (якщо додатково відомо, що вони мають однакові коваріаційні матриці).

Порівняння різних способів розбиття досліджуваної сукупності об'єктів на класи здійснюють за допомогою **функціонала якості розбиття** Q(S). Найкращим вважають розбиття, при якому забезпечується екстремум цього функціонала. Не існує строгих методів обрання функціоналів якості.

У разі, коли кількість класів є заданою, за функціонали якості найчастіше обирають наступні.

Сума внутрішньокласових дисперсій:

$$Q_1(S) = \sum_{l=1}^{k} \sum_{X_i \in S_l} d^2\left(X_i, \overline{X}(l)\right). \tag{7.24}$$

Сума попарних внутрішньокласових відстаней між елементами:

$$Q_2(S) = \sum_{l=1}^{k} \sum_{X_i, X_j \in S_l} d_{ij}^2 \ . \tag{7.25}$$

Узагальнену внутрішньокласову дисперсію можна розраховувати як показник середньоарифметичної або середньогеометричної дисперсії, відповідно, за формулами:

$$\begin{aligned} Q_3(S) &= \det\left(\sum_{l=1}^{k} n_l \Sigma_l\right); \\ Q_4(S) &= \prod_{l=1}^{k} \left(\det \Sigma_l\right)^{n_l}, \end{aligned} \tag{7.26}$$

де елементи вибіркової коваріаційної матриці Σ класу S_l розраховують як:

$$\sigma_{qt}(l) = \frac{1}{n_l} \sum_{X_i \in S_l} \left(x_i^{(q)} - \bar{x}^{(q)}(l) \right) \left(x_i^{(t)} - \bar{x}^{(t)}(l) \right), \quad q,t = 1,2,...,p. \quad (7.27)$$

Такі функціонали доцільно застосовувати, якщо допускається можливість зосередженості розбитих на класи спостережень у просторі розмірності меншої ніж p.

За невідомої кількості класів функціонали якості розбиття зазвичай обирають у вигляді простої алгебраїчної комбінації (суми, різниці, добутку, частки) двох функціоналів, один з яких є незростаючою функцією кількості класів і характеризує внутрішньокласовий розкид спостережень, а другий − неспадною функцією кількості класів. Останній може характеризувати взаємну віддаленість (близькість) точок, втрати від надмірної деталізації вихідного масиву даних, концентрацію наявної структури точок тощо. У **схемі О.М. Колмогорова** для побудови такого функціонала використовують **міру концентрації точок**:

$$Z_\tau(S) = \left[\frac{1}{n} \sum_{i=1}^{n} \left(\frac{v(X_i)}{n} \right)^\tau \right]^{1/\tau}, \quad (7.28)$$

де $v(X_i)$ − кількість елементів у кластері, що містить точку X_i, а вибір параметра τ залежить від мети розбиття.

Такий функціонал відповідає середній мірі внутрішньокласового розсіювання $I_\tau^{(K)}(S)$. При визначенні слід взяти до уваги, що:

$$Z_{-\infty}(S) = \min_{1 \le i \le k}\left(\frac{n_i}{n}\right);$$

$$Z_{-1}(S) = \frac{1}{k};$$

$$\log Z_0(S) = \sum_{i=1}^{k} \frac{n_i}{n} \log \frac{n_i}{n};$$

$$Z_1(S) = \frac{1}{n} \sum_{j=1}^{n} \frac{v(X_i)}{n} = \frac{1}{n^2} \sum_{i=1}^{k} n_i^2;$$

$$Z_{\infty}(S) = \max_{1 \le i \le k}\left(\frac{n_i}{n}\right),$$

де k − кількість різних кластерів у розбитті S, n_i − кількість елементів в і-му кластері. Величина Z_0 є природною інформаційною мірою концентрації.

За будь-якого τ міра (7.28) має мінімальне значення 1/n при розбитті досліджуваної множини на n одноточкових кластерів і максимальне значення 1 при об'єднанні всіх вихідних даних до одного кластера. Як середню міру внутрішньокласового розсіювання можна використовувати величину:

$$I_{\tau}^{(K)}(S) = \left[\frac{1}{n} \sum_{i=1}^{n} \frac{1}{v(X_i)} \sum_{X_l \in S(X_i)} d_{il}^{\tau}\right]^{1/\tau} \qquad (7.29)$$

Тоді сумарне розсіювання характеризує величина $n(I_{\tau}^{(K)}(S))^{\tau}$ і оптимальною слід вважати таку кількість кластерів, за якої величина

$$\frac{\Delta\left[n(I_{\tau}^{(K)}(S))^{\tau}\right]}{\Delta Z_{\tau}(S)} \qquad (7.30)$$

набуває мінімального можливого значення.

Залежно від кількості вихідних спостережень вирізняють задачі класифікації невеликих за обсягом (до декількох десятків точок) ма-

сивів спостережень і задачі класифікації великих масивів. Такий поділ зумовлений різницею процедур, які доцільно використовувати при класифікації відповідних даних.

З погляду апріорної інформації про кількість кластерів виділяють такі типи задач:

- із заданою кількістю класів;
- з невідомою кількістю класів, яку треба оцінити;
- з невідомою кількістю класів, яку не потрібно оцінювати (таку задачу зазвичай формулюють як побудову ієрархічного дерева, або дендрограми вихідної сукупності.

Найбільш поширеними методами кластерного аналізу є:

- **ієрархічні методи** (ближнього зв'язку, середнього зв'язку Кінга, Ворда, далекого зв'язку);
- **ітеративні методи** групування (метод k-середніх Мак-Квіна);
- **алгоритми типу розрізування графа** (кореляційних плеяд Терентьєва, вроцлавська таксономія).

Ієрархічні (агломеративні та дивізимні) методи призначені переважно для побудови ієрархічних дерев відносно невеликих за обсягом сукупностей. Іноді їх використовують також для задач класифікації перших двох типів. У цьому разі реалізацію ієрархічного алгоритму продовжують до досягнення кількості класів, рівної заздалегідь заданому числу k, або до досягнення екстремуму одного з критеріїв якості розбиття.

Перевагами ієрархічних методів є можливість більш повного і тонкого аналізу структури досліджуваної сукупності порівняно з іншими методами, а також наочність подання результатів кластеризації. Їх ос-

новними недоліками є громіздкість обчислювальної процедури, яка пов'язана з перерахунком усієї матриці відстаней на кожному кроці, а також "скінченна неоптимальність" гранично оптимальних алгоритмів у багатьох випадках.

Метод ближнього зв'язку є найпростішим для розуміння з числа ієрархічних агломеративних методів кластерного аналізу. Процес класифікації у цьому разі починають з пошуку та об'єднання двох найближчих один до одного об'єктів у матриці подібності. На наступному етапі знаходять два наступні найближчі об'єкти й т.д. до повного вичерпання матриці подібності. Як правило, робота алгоритму закінчується, коли всі спостереження об'єднані до одного класу. Для виділення кластерів після закінчення кластеризації задають пороговий рівень подібності, на якому можна виділити більше, ніж один кластер.

Описана процедура не завжди приводить до утворення одного великого кластера на останньому етапі. Часто вона закінчується явним розділенням досліджуваних об'єктів на кластери.

У цьому методі два об'єкти потрапляють до одного й того самого кластера в тому випадку, коли існує ланцюжок близьких один до одного об'єктів, які їх з'єднують (**ланцюжковий ефект**). У процесі кластеризації можна явно простежити утворення таких ланцюжків. Для запобігання цьому ефекту можна задавати обмеження на максимальну відстань між елементами одного й того самого кластера. Після проведення кластеризації рекомендується візуалізувати результати шляхом побудови дендрограми, яка дає можливість отримати уявлення про загальну конфігурацію об'єктів. Кластери, одержувані за методом

ближнього зв'язку, не обов'язково бувають опуклими. Залежно від обставин, це можна розглядати і як перевагу, і як недолік методу.

Метод середнього зв'язку Кінга є подібним до методу ближнього зв'язку. Відмінність полягає в тому, що об'єднані до одного кластера об'єкти надалі вважають одним об'єктом з усередненими за кластером параметрами. При цьому новому об'єкту надають номер меншого з номерів об'єднуваних об'єктів, а об'єкти, що залишилися, перенумеровують. Таким чином їх загальна кількість зменшується на одиницю. Подальша процедура є подібною до попереднього методу. В іншому варіанті методу середнього зв'язку відстань між класами розраховують як середнє значення відстаней між усіма можливими парами представників цих класів. У процесі кластеризації також простежується формування ланцюжків об'єктів, що дає змогу задати пороговий рівень подібності, на якому можна виділити більше ніж один кластер. Часто процедура кластеризації закінчується явним розподілом об'єктів на кластери. Після закінчення кластеризації також доцільно здійснити візуалізацію результатів шляхом побудови дендрограми.

К-узагальнена ієрархічна процедура ґрунтується на тому, що перелічені відстані, а також відстань, що визначається за принципом найдальшого сусіда (застосовується в методі далекого зв'язку), є окремими випадками узагальненої відстані Колмогорова (7.20), яка і використовується у даному разі як міра близькості. Описані вище методи можна розглядати як окремі випадки К-узагальненої ієрархічної процедури.

Порогові ієрархічні процедури передбачають задання монотонної послідовності порогів $c_1, c_2, ..., c_t$. В агломеративних методах на пер-

шому кроці попарно об’єднуються елементи, відстані між якими не перевищують c_1. На другому об'єднують елементи або групи елементів, відстані між якими не перевищують c_2 і т.д. При достатньо великих значеннях c_t на останньому кроці всі елементи будуть об’єднані до одного загального кластера. Недоліком цих процедур є можливість перетину класів, тому вони бувають ефективними за умови, що ланцюжковий ефект слабко виражений, а вихідна сукупність природно ділиться на достатньо віддалені одне від одного скупчення точок у досліджуваному просторі ознак.

Метод Ворда є близьким до методу середнього зв’язку Кінга. Він відрізняється тим, що підставою для приєднання об’єкта до кластера є не близькість у значенні певної міри подібності, а мінімум дисперсії всередині кластера після поміщення до нього обраного об’єкта.

Паралельні ітеративні процедури передбачають одночасне використання всіх наявних спостережень, тому їх застосовують для розв’язання задач класифікації перших двох типів при порівняно малих обсягах досліджуваних сукупностей.

Послідовні ітераційні процедури на кожному кроці використовують лише невелику кількість спостережень, а також результат попереднього кроку кластеризації. Їх застосовують, як правило, для розв’язання перших двох типів задач кластеризації при великих обсягах досліджуваних сукупностей.

Прикладом послідовної ітеративної процедури є **метод k-середніх Мак-Квіна**. Розв’язується задача розбиття n об’єктів на k (k < n) однорідних у певному розумінні кластерів. На початковому етапі його реалізації вихідні точки впорядковують (можливо випадковим чином)

і перші k точок у подальшому розглядають як окремі кластери, яким надають одиничні вагові коефіцієнти. Потім беруть точку X_{k+1} і з'ясовують, до якого з наявних кластерів вона є найближчою. Цей кластер замінюють новим, який розташований у центрі ваги вихідного кластера і точки X_{k+1}. При цьому ваговий коефіцієнт отриманого кластера збільшують на одиницю порівняно із ваговим коефіцієнтом вихідного. Якщо точка X_{k+1} є рівновіддаленою від декількох кластерів, то її вміщують до кластера з найменшим номером або з найбільшим ваговим коефіцієнтом. Потім почергово приєднують до наявних кластерів точки, що залишилися. При достатньо великих обсягах досліджуваних вибірок центри ваги отримуваних кластерів згодом перестають змінюватися, тобто ітераційна процедура збігається до певної границі. Якщо ж вона не збігається за задану кількість кроків, то використовують один із таких прийомів. Перший передбачає, що після розгляду останньої точки X_n повертаються до точок X_1, X_2 і т.д. Другий підхід передбачає багаторазове повторне обрання вихідних кластерів. При цьому на кожному етапі як вихідні обирають точки, які найчастіше були найближчими до фінальних кластерів на попередніх етапах.

Особливістю методу є алгоритмічне гарантування того, що кожний з класифікованих об'єктів буде віднесений лише до одного з кластерів. У цьому методі немає особливої необхідності у візуалізації результатів. Але для наочності можна здійснити її за допомогою зображення просторових еліпсоїдів, що містять класифіковані об'єкти (якщо розмірність не перевищує три), або двовимірних зрізів простору. У багатьох випадках метод k-середніх дає змогу отримати розбиття, близьке до найкращого з погляду функціонала якості.

У разі, коли кількість класів є невідомою, необхідно задати дві константи: **міру грубості** φ та **міру точності** ψ. На нульовому кроці беруть довільне значення кількості класів k_0, вихідні точки впорядковують і розглядають k_0 перших точок як центри кластерів, яким надають одиничні вагові коефіцієнти. Потім здійснюють **огрубіння** вихідних кластерів. Для цього послідовно розраховують попарні відстані між ними і, якщо відстань між двома кластерами не перевищує φ, їх об'єднують до одного, який є їх зваженим середнім і має ваговий коефіцієнт, рівний сумі вагових коефіцієнтів вихідних кластерів. Після закінчення процедури ми отримуємо $k_0^* \le k_0$ кластерів.

Далі здійснюють послідовне рознесення точок, що залишилися, за кластерами. Для кожної точки визначають найближчий до неї кластер. Якщо відстань між ними не перевищує ψ, то відповідну точку приєднують до цього кластера за вищеописаною процедурою. У протилежному випадку її вважають центром нового кластера, якому надається одиничний ваговий коефіцієнт. Після рознесення усіх точок за кластерами повторюють процедуру огрубіння і переходять до чергового кроку ітерацій.

Обираючи різні значення констант φ та ψ, можна отримати різні розбиття вихідної сукупності. Вибір вважають задовільним, якщо результат класифікації є близьким до оптимального за оцінками експертів або з погляду функціонала якості.

Сутність методу **кореляційних плеяд** полягає у наступному. Візуально результати класифікації можна подати у вигляді кореляційного циліндра, розсіченого площинами, перпендикулярними до його осі.

Площини відповідають рівням від нуля до одиниці з кроком 0,1. На цих рівнях об'єднуються класифіковані параметри або об'єкти. Метод наближається до методу ближнього зв'язку із фіксованими рівнями об'єднання. Графічно результати зображують у вигляді кіл, які є зрізами кореляційного циліндра (плеяд). На них відмічають класифіковані об'єкти, зв'язки між якими вказують за допомогою хорд, що з'єднують відповідні точки кіл. Метод кореляційних плеяд є основою для багатьох порогових алгоритмів.

У **методі Вроцлавської таксономії** визначають пари чисел, які вказують порядок з'єднання попарно найближчих один до одного об'єктів (параметрів), що підлягають класифікації. Одержуваний незамкнений найкоротший шлях можна відобразити графічно у вигляді оптимального дерева (дендрита). Цей метод є подібним до методу ближнього зв'язку, але його алгоритм належить до алгоритмів розрізання графів. Якщо за міру подібності обрати коефіцієнт кореляції, ми отримаємо метод найбільшого кореляційного шляху.

Результати ієрархічних методів кластерного аналізу стають більш наочними, якщо їх подати у вигляді **дендрограми**. Типовий вигляд дендрограми наведено на рис. 7.1.

Пари об'єктів при побудові дендрограми з'єднують згідно із рівнем зв'язку, який відкладають за віссю ординат. Задаючи кількість кластерів, наприклад n = 3, знаходять, на якому рівні кількість перетинів горизонтальної лінії, яка відповідає рівню зв'язку, і вертикальних ліній, що відповідають об'єктам, дорівнює трьом. У нашому випадку такій кількості кластерів відповідає рівень зв'язку, що знаходиться приблизно в межах від 38 до 45 одиниць.

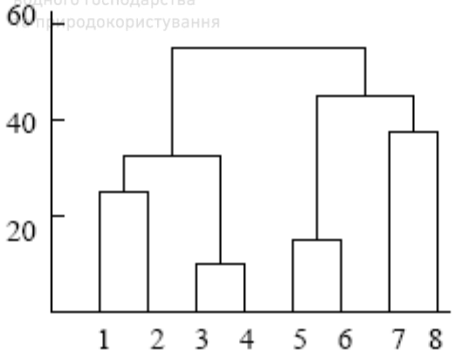

Рис. 7.1. Дендрограма – графічне представлення результатів ієрархічного кластерного аналізу

У методі вроцлавської таксономії результати розрахунків відображають (рис.7.2) у вигляді **графа** (**дендрита**). Розміщення на площині точок, що є зображеннями параметрів або об'єктів, і з'єднуючих їх відрізків, які зображують зв'язки, є довільним. При цьому рівні зв'язку відображають максимальні значення зв'язків відповідних параметрів (об'єктів) з іншими параметрами (об'єктами).

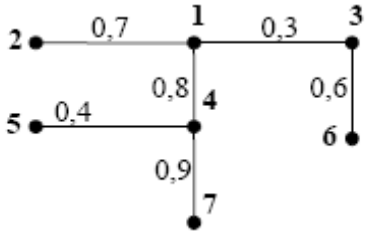

Рис. 7.2. Приклад графа, що відображає зв'язок між параметрами

Аналізуючи отриманий граф, можна зробити висновки про взаємозв'язки тих чи інших параметрів або груп параметрів, зокрема про

те, що для графа, поданого на рис. 7.2, параметри умовно розділяються на три групи: (1, 3, 6), (2) та (4, 5, 7).

Перед застосуванням процедур класифікації рекомендується здійснити дослідження наявних ознак з метою вибору найбільш інформативних із них і скорочення розмірності простору ознак. З цією метою буває доцільним розглянути компоненти $x^{(1)}, x^{(2)}, ..., x^{(p)}$ як об'єкти, що підлягають класифікації. Це дає змогу виявити групи компонентів, що відображають окремі властивості або групи властивостей досліджуваних об'єктів, і при подальшому аналізі враховувати лише по одному представнику з кожної такої групи.

7.2. Класифікація з навчанням

Методи розпізнавання образів з навчанням (із вчителем) призначені для віднесення некласифікованих об'єктів до заздалегідь описаних класів (кластери; навчаючі вибірки). Програму (пристрій) розпізнавання зазвичай називають **класифікатором**, а в разі, коли вона видає відповідь у вигляді дійсного числа або вектора дійсних чисел − **предиктором** (**локалізатором**).

Задачу побудови оптимальної процедури класифікації у цьому разі можна сформулювати таким чином. Є відомими p - вимірні спостереження $X_1, X_2, ..., X_n$ та функції щільності розподілу $f_1(X), f_2(X), ..., f_k(X)$, які задають k класів і можуть розглядатися як вибірки, що навчають. Спостереження, що підлягають класифікації, можна розглядати як вибірки з генеральної сукупності, яка описується сумішшю k одномодальних функцій розподілу (класів):

$$f(X) = \sum_{j=1}^{k} \pi_j f_j(X). \qquad (7.31)$$

Розв'язуючим правилом (дискримінантною функцією) називають функцію $\delta(X)$, що має такі властивості. Її значеннями можуть бути тільки додатні цілі числа $1,2,...,k$. При цьому ті спостереження X, для яких вона має значення j, зараховують до j-го класу S_j, тобто:

$$S_j = \{X : \delta(X) = j\}. \qquad (7.32)$$

Функцію $\delta(X)$ будують так, щоб об'єднання класів $S = \bigcup_{j=1}^{k} S_j$ охоплювало всі можливі значення аналізованої багатовимірної ознаки X, і для будь яких $i \neq j$ виконувалася умова $S_i \bigcap S_j = \varnothing$.

Розв'язуючі правила дають можливість зараховувати досліджувані об'єкти до заданих класів. Їх можна отримати у вигляді:

- імовірності діагнозу при заданому комплексі симптомів (метод Байєса);
- простих функцій, що класифікують (лінійний дискримінантний аналіз Фішера);
- дискримінантних функцій (канонічний дискримінантний аналіз);
- певних характеристик: групова кореляційна матриця, груповий вектор середніх та визначник коваріаційної матриці (лінійний дискримінантний аналіз);

- настроєної ваги синапсів і зміщень нейронів (нейронна мережа, що навчається).

Розв'язуюче правило називають оптимальним (байєсівським), якщо воно забезпечує мінімальні втрати серед усіх можливих процедур класифікації. Оптимальне розв'язуюче правило можна задати таким чином:

$$S_j^{(opt)} = \left(X : \sum_{\substack{i=1 \\ i \neq j}}^{k} \pi_i f_i(X) c(j \mid i) = \min_{1 \le l \le k} \sum_{\substack{i=1 \\ i \neq j}}^{k} \pi_i f_i(X) c(l \mid i) \right) \quad (7.33)$$

Це означає, що спостереження $X_v (v = 1,2,...,n)$ зараховують до j-го класу в разі, коли відповідні втрати будуть меншими порівняно із втратами від його зарахування до будь-якого іншого класу. Якщо $c(j \mid i) = c_0 = const,$ то спостереження X_v зараховують до j-го класу за умови:

$$\pi_i f_i(X_v) = \max_{1 \le l \le k} \pi_l f_l(X_v). \quad (7.34)$$

Це розв'язуюче правило можна сформулювати таким чином: спостереження X_v зараховують до класу j_0, якщо

$$\frac{f_{j_0}(X_v)}{f_j(X_v)} \ge \frac{\pi_j}{\pi_{j_0}} \quad (7.35)$$

для всіх $j = 1,2,...,k$.

Одним з основним методів розпізнавання образів з навчанням є **дискримінантний аналіз**. Він належить до класу лінійних методів, оскільки його модель є лінійною відносно дискримінантних функцій.

Користувач повинен задати певну кількість об'єктів і вказати їх належність до так званих груп, що навчають (класів, кластерів, популяцій). Тому застосуванню дискримінантного аналізу має передувати дослідження методами розпізнавання без навчання − кластерного аналізу, багатовимірного шкалування або емпіричної класифікації. Кластери можуть перетинатися, особливо в разі, коли навчання здійснюється за допомогою емпіричної класифікації.

Якщо встановлена належність окремих об'єктів до стандартно описаних груп, рекомендується утворювати з них нові кластери. Для навчання необхідно використовувати об'єкти (вибірки, що навчають), заздалегідь класифіковані тим чи іншим способом. Якість дискримінації визначається ймовірністю правильної класифікації. Зазвичай найкращі результати дає застосування методу *k*-середніх, який гарантовано будує кластери, що не перетинаються, а також методу ближнього зв'язку.

Під **інформативністю** параметрів зазвичай розуміють їх спроможність описувати об'єкт класифікації з достатньою для її здійснення точністю. Як правило, розпізнаванню образів з навчанням має передувати застосування дисперсійного, кореляційного або факторного аналізу чи деякого іншого методу з метою виділення інформативних параметрів, а також класифікація без навчання для виділення груп, що навчають.

Можливі ситуації, коли кількість параметрів є недостатньою для правильної з погляду дослідника класифікації, або, навпаки, є зайві параметри, які не є обов'язковими для класифікації, але призводять до отримання громіздких результатів, які важко інтерпретувати. В окре-

мих випадках об'єкти з вибірок, що навчають, після класифікації можуть бути віднесені не до тих кластерів, куди вони були поміщені на попередньому етапі. Особливо часто таке відбувається при застосуванні емпіричних класифікацій. У таких випадках необхідно виконати додаткове дослідження стосовно необхідності й достатності тих параметрів, за якими проводиться класифікація.

Для практичної реалізації оптимальних розв'язуючих правил (7.33, 7.34) необхідно знати апріорні імовірності π_j і функції щільності імовірності $f_j(X)$. Вони можуть бути відомими з теоретичних міркувань або попередніх досліджень. Якщо ж вони невідомі, то їх замінюють статистичними оцінками, одержуваними на основі наявних вибірок, що навчають.

Як оцінки апріорних імовірностей часто беруть величини

$$\pi_j = n_j / n_{sum}, \qquad (7.36)$$

де n_j – обсяг j-ї вибірки, а $n_{sum} = n_1 + n_2 + ... + n_k$ – сумарний обсяг вибірок, що навчають.

При оцінюванні функцій щільності імовірностей застосовують два підходи. У першому (**параметричний дискримінантний аналіз**) припускають, що всі класи характеризуються функціями щільності імовірності, які належать до однієї параметричної сім'ї $\{f(X,\Theta)\}$ і відрізняються лише значеннями векторного параметра Θ. У цьому разі відповідні значення параметра Θ_j оцінюють за спостереженнями, що належать до j-ї вибірки. У другому випадку (**непараметричний дискримінантний аналіз**) загальний вигляд функцій $f_j(X)$ є неві-

домим. Тоді необхідно використовувати спеціальні прийоми їх оцінювання, наприклад, будувати непараметричні оцінки гістограмного або ядерного типу.

Розглянемо більш докладно параметричний дискримінантний аналіз у випадку нормальних класів. Припустимо, що кожний j-й клас є p-вимірною нормальною сукупністю із вектором середніх значень $\mathbf{a}_j$ і коваріаційною матрицею Σ, яка є загальною для всіх класів. Тоді функції $f_j(X)$ доцільно задати у вигляді щільності p-вимірного нормального розподілу ймовірності:

$$\varphi(X, M, \Sigma) = \frac{1}{2\pi|\Sigma|^{1/2}} e^{-\frac{1}{2}(X-A)^T\Sigma^{-1}(X-A)}, \qquad (7.37)$$

де A − матриця, утворена векторами середніх значень, X − матриця значень ознак. Обидві матриці мають розмір $p \times k$.

Оцінки для векторів середніх значень $a_j = \left(a_j^{(1)}, ..., a_j^{(p)}\right)^T$ і елементів коваріаційної матриці, отримані методом найбільшої правдоподібності за вибірками, що навчають, мають вигляд:

$$a_j^{(l)} = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ji}^{(l)}; \qquad (7.38)$$

$$\sigma_{lq} = \frac{1}{n_{sum} - k} \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(x_{ji}^{(l)} - a_j^{(l)}\right)\left(x_{ji}^{q} - a_j^{q}\right); \qquad (7.39)$$

$$(l = 1, ..., p; j = 1, ..., k).$$

Якщо функції $f_j(X)$ визначаються формулою (7.37), то розв'язуюче правило (7.35) набуває вигляду:

$$\left[X_v - \frac{1}{2}\left(a_{j_0} + a_j\right)\right]^T \Sigma^{-1}\left(a_{j_0} + a_j\right) \geq \ln\frac{\pi_j}{\pi_{j_0}} \qquad (7.40)$$

для усіх $j = 1,2,...,k$.

Для *k = 2* апріорні імовірності $\pi_1 = \pi_2 = 0{,}5$. У цьому разі спостереження X_v зараховують до першого класу, якщо

$$\left[X_v - \frac{1}{2}\left(a_1 + a_2\right)\right]^T \Sigma^{-1}\left(a_1 + a_2\right) \geq 0\,, \qquad (7.41)$$

і до другого − у всіх інших випадках.

Одним з поширених методів розпізнавання образів є **метод Байєса**. Він дає можливість враховувати ознаки різної розмірності (фізичної природи) завдяки використанню рівнозначних безрозмірних характеристик ознак − частот зустрічальності ознак при різних станах.

Основою методу є **діагностична матриця**, у стовпчиках якої подають значення ймовірностей певної ознаки для різних класів, а у рядках − імовірності всіх ознак для окремих класів. У табл. 7.1 наведено

Таблиця 7.1

Приклад діагностичної матриці для багаторозрядних ознак

D	K_1			K_2		K_3			P(D)
	$P(K_{11})$	$P(K_{12})$	$P(K_{13})$	$P(K_{21})$	$P(K_{22})$	$P(K_{31})$	$P(K_{32})$	$P(K_{33})$	
D1									
D2									

приклад такої матриці для випадку двох класів і трьох ознак. При цьому перша і третя ознаки мають по три розряди, а друга − два розряди. Розрахунок імовірності віднесення об'єкта до класу D_i здійснюють за формулою:

$$P(D_i \mid K^*) = \frac{P(D_i)P(K^* \mid D_i)}{\sum_{s=1}^{n} P(D_s)P(K^* \mid D_s)}, \qquad (7.42)$$

де $K(K_1, K_2, ..., K_v)$ − ряд ν багаторозрядних ознак, K^* − його реалізація, $P(D_i \mid K^*)$ − імовірність віднесення об'єкта до класу D_i за умови, що комплекс ознак К набув реалізації K^*, $P(K^* \mid D_i)$ − імовірність появи комплексу ознак K^* у об'єкта, що належить до класу D_i, $P(D_i)$ − апріорна імовірність потрапляння до класу D_i, яка визначається за емпіричними даними, і − номер кластера.

Якщо комплекс ознак містить ν ознак, то

$$P(K^* \mid D_i) = P(K_1^* \mid D_i)P(K_2^* \mid K_1^* D_i)...P(K_v^* \mid K_1^* K_2^* ... K_{v-1}^* D_i). \qquad (7.43)$$

У багатьох випадках, навіть за наявності істотних кореляційних зв'язків, можна використовувати формулу Байєса для незалежних ознак. У цьому разі:

$$P(K^* \mid D_i) = \prod_{r=1}^{v} P(K_r^* \mid D_i). \qquad (7.44)$$

В основі **методу лінійного дискримінантного аналізу Фішера** лежить припущення, що класифікацію можна здійснити за допомогою лінійної комбінації дискримінантних (розрізняючих) змінних. Підгру-

нтям для зарахування об'єкта до певного кластера є максимальне значення функції, що класифікує, яка є лінійною комбінацією дискримінантних змінних X і може бути записана для k-го кластера у вигляді:

$$h_k = b_{k0} + \sum_{i=1}^{p} b_{ki} X_i \,, \tag{7.45}$$

де p − кількість дискримінантних змінних, b_{ki} − коефіцієнт для і-ї змінної k-го класу

$$b_{ki} = (n-g)\sum_{j=1}^{p} a_{ij} X_{jk} \,, \tag{7.46}$$

n − загальна кількість спостережень за усіма класами, a_{ij} − елементи матриці, яка є оберненою до матриці розкидів всередині класів і розраховується за формулою

$$w_{ij} = \sum_{k=1}^{j} \sum_{n=1}^{n_k} (X_{ikm} - X_{ik})(X_{jkm} - X_{jk}) \,, \tag{7.47}$$

g − кількість класів, n_k − кількість спостережень у k-му класі, X_{ikm} − значення m-го спостереження і-ї змінної у k-му класі, X_{ik} − середнє значення і-ї змінної у k-му класі.

Для використання методу необхідно виконання таких умов:

- кластери, серед яких здійснюють дискримінацію, підпорядковані нормальному багатовимірному розподілу;
- різниці між коваріаційними матрицями цих кластерів є статистично незначущими.

Останнє припущення спрощує обчислювальну процедуру. Але його необґрунтоване застосування може призвести до втрати найсуттєві-

ших індивідуальних характеристик кластерів, які мають істотне значення для дискримінації. Це припущення також дає змогу отримати розв'язок у випадку, коли кількість вибірок, що навчають у кластері є меншою, ніж кількість дискримінантних функцій, тобто коли лінійний дискримінантний аналіз не може бути використаний.

За якістю дискримінації (відсотком правильно класифікованих об'єктів) результати лінійного дискримінантного аналізу збігаються з результатами більш складного методу канонічного дискримінантного аналізу.

Канонічний дискримінантний аналіз ґрунтується на знаходженні канонічних дискримінантних функцій

$$f_{km} = u_0 + \sum_{i=1}^{p} u_i X_{ikm,} \tag{7.48}$$

де u_i - коефіцієнти, що визначають за формулою

$$u_i = v_i \sqrt{n-g}\,, u_0 = -\sum_{i=1}^{p} u_i \overline{X}_i\,, \tag{7.49}$$

$\overline{X}_i$ - середнє значення і-ї змінної за всіма класами, v_i - коефіцієнти, які розраховують, як компоненти власних векторів розв'язку узагальненої проблеми власних значень:

$$Bv = \lambda Wv, \tag{7.50}$$

тут В – міжгрупова сума квадратів відхилень, v - власний вектор. Інші позначення збігаються з тими, що ми використовували для лінійно-

го дискримінантного аналізу. Кількість дискримінантних функцій може бути меншою або рівною кількості параметрів об'єкта.

Матрицю В визначають як

$$B = T - W, \tag{7.51}$$

де Т – матриця сум квадратів і попарних добутків:

$$t_{ij} = \sum_{k=1}^{g} \sum_{m=1}^{n_k} (X_{imk} - \overline{X}_i)(X_{jmk} - \overline{X}_j). \tag{7.52}$$

Зарахування нових некласифікованих об'єктів до заданих кластерів здійснюється після обчислення дискримінантних функцій на основі евклідової метрики.

Недоліком методу лінійного дискримінантного аналізу Фішера є припущення про рівність коваріаційних матриць досліджуваних вибірок. У методі **лінійного дискримінантного аналізу (не Фішера)**, навпаки, припускають, що коваріаційні матриці різних вибірок є різними, що істотно ускладнює процедуру розрахунків. Відмова від припущення про статистичну нерозрізненість коваріаційних матриць для кластерів, що навчають, веде до необхідності того, щоб кількість вибірок, що навчають, у кластері була не меншою, ніж кількість дискримінантних функцій. Якщо ця умова не виконується, необхідно застосовувати лінійний дискримінантний аналіз Фішера або канонічний дискримінантний аналіз.

Підґрунтям для зарахування об'єкта до того чи іншого класу є найбільше за всіма класами значення функції щільності ймовірності для даного об'єкта. Якість розпізнавання для цього методу є приблизно на 5% вищою, ніж для двох попередніх.

Розпізнавання образів є необхідним попереднім етапом статистичної обробки багатовимірних даних. Це пов'язано із тим, що вплив одних і тих самих факторів на поведінку різних кластерів зазвичай є різним, а іноді й протилежним. Тому застосування методів кореляційно-регресійного аналізу до сукупності в цілому може призводити до істотних похибок і, як правило, не дає можливості дати змістовну інтерпретацію отриманих параметрів.

7.3. ЗАВДАННЯ ДО РОЗДІЛУ 7

Завдання 1. Використовуючи дані про середню врожайність озимої пшениці по областях України виконати кластерний аналіз даних засобами пакету Statistica (Statistics, Multivariate Exploratory Techniques, Cluster Analysis, Joining (tree clustering)). Побудувати ієрархічне дерево та виконати його змістовний текстовий аналіз. Перед виконанням аналізу у пакеті Statistica дані транспонувати.

Роки	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**
АР Крим	20.7	22.9	20	15.1	22.6	22.5	23.7
Вінницька	24.9	27.1	31.1	13.5	31.9	29	27.4
Волинська	23	23.2	28.8	25.3	32.9	28	22.9
Дніпропетровська	16.5	43	36.5	8	35.8	34.8	20.5
Донецька	14.9	38.1	31.9	10.7	32	31.3	25.4
Житомирська	21.9	22	26.7	14.2	27.2	23.1	25.3
Закарпатська	18.5	28.3	30.3	28.7	38.9	32.3	29.8
Запорізька	17.1	34.8	27.9	9.4	31	30.2	22.8
Івано-Франківська	22.3	21.1	27.2	21	30.5	24.1	23.4
Київська	26	29.7	32.1	22	36.2	34.8	29.8
Кіровоградська	18.2	41.9	38.2	7.5	37.7	32.8	21.2
Луганська	8.9	32.2	25.8	15.4	25	31.6	21.7
Львівська	21.9	21.9	27.1	23.6	29.2	24.4	24.9
Миколаївська	16.4	33.2	28.6	6	33.5	22.3	15.5
Одеська	19.5	34.3	31.1	6.4	34.7	24.1	18.7
Полтавська	12.2	33.1	36.2	10.9	32.3	32.6	29.4
Рівненська	25.1	22.9	31.9	19.5	30.9	25.3	22.1
Сумська	16.8	26.9	30.9	13.7	29.7	24.3	29.1
Тернопільська	22.3	18.5	27.4	18.2	28.6	23.6	23
Харківська	15.1	35.9	37.1	12.6	32	36.4	28.9
Херсонська	18.8	30	24.1	6.2	29.8	24.5	19.1
Хмельницька	27.3	20.7	29	17.3	29.8	22.5	19.7
Черкаська	24.7	37.3	34.9	13.5	36.9	36	30.4
Чернівецька	19.9	22.6	27.7	11.5	26.3	24.8	30.4
Чернігівська	15.2	18.7	23.9	15	29.6	25.2	32.3

Завдання 2. Використовуючи офіційні результати позачергових виборів до Верховної Ради України 2007 року (Завдання 1 до Розділу 6) виконати виконати кластерний аналіз даних засобами пакету Statistica (Statistics, Multivariate Exploratory Techniques, Cluster Analysis, Joining (tree clustering)). Побудувати ієрархічне дерево та виконати його змістовний текстовий аналіз. Перед виконанням аналізу у пакеті Statistica дані транспонувати.

Питання для самоконтролю

1. Що таке класифікація?
2. Коли використовують параметричні методи розпізнавання образів без навчання?
3. В чому суть методу кластерного аналізу?
4. Що таке клас?
5. Назвіть основні види міри подібності та відмінності даних. Наведіть приклади.
6. Суть схеми О.М. Колмогорова**.**
7. Які методи кластерного аналізу є найбільш поширеними?
8. Розкрийте суть методів: метод ближнього зв'язку, метод середнього зв'язку Кінга, К-узагальнена ієрархічна процедура, морогові ієрархічні процедури, метод Ворда, метод k-середніх Мак-Квіна, метод кореляційних плеяд, метод Вроцлавської таксономії.
9. Що таке розв'язуюче правило? У якому вигляді можна отримати розв'язуючі правила?
10. Що розуміють під дискримінантним аналізом?
11. Які різновиди дискримінантного аналізу існують і в яких випадках вони використовуються?

ДОДАТОК

Таблиця 1. Подвоєна нормована функція Лапласа (нормальний розподіл)

t	Φ(t)	t	Φ(t)	t	Φ(t)	t	Φ(t)
0	0.0000	0.34	0.2661	0.68	0.5035	1.02	0.6923
0.01	0.0080	0.35	0.2737	0.69	0.5098	1.03	0.6970
0.02	0.0160	0.36	0.2812	0.7	0.5161	1.04	0.7017
0.03	0.0239	0.37	0.2886	0.71	0.5223	1.05	0.7063
0.04	0.0319	0.38	0.2961	0.72	0.5285	1.06	0.7109
0.05	0.0399	0.39	0.3035	0.73	0.5346	1.07	0.7154
0.06	0.0478	0.4	0.3108	0.74	0.5407	1.08	0.7199
0.07	0.0558	0.41	0.3182	0.75	0.5467	1.09	0.7243
0.08	0.0638	0.42	0.3255	0.76	0.5527	1.1	0.7287
0.09	0.0717	0.43	0.3328	0.77	0.5587	1.11	0.7330
0.1	0.0797	0.44	0.3401	0.78	0.5646	1.12	0.7373
0.11	0.0876	0.45	0.3473	0.79	0.5705	1.13	0.7415
0.12	0.0955	0.46	0.3545	0.8	0.5763	1.14	0.7457
0.13	0.1034	0.47	0.3616	0.81	0.5821	1.15	0.7499
0.14	0.1113	0.48	0.3688	0.82	0.5878	1.16	0.7540
0.15	0.1192	0.49	0.3759	0.83	0.5935	1.17	0.7580
0.16	0.1271	0.5	0.3829	0.84	0.5991	1.18	0.7620
0.17	0.1350	0.51	0.3899	0.85	0.6047	1.19	0.7660
0.18	0.1428	0.52	0.3969	0.86	0.6102	1.2	0.7699
0.19	0.1507	0.53	0.4039	0.87	0.6157	1.21	0.7737
0.2	0.1585	0.54	0.4108	0.88	0.6211	1.22	0.7775
0.21	0.1663	0.55	0.4177	0.89	0.6265	1.23	0.7813
0.22	0.1741	0.56	0.4245	0.9	0.6319	1.24	0.7850
0.23	0.1819	0.57	0.4313	0.91	0.6372	1.25	0.7887
0.24	0.1897	0.58	0.4381	0.92	0.6424	1.26	0.7923
0.25	0.1974	0.59	0.4448	0.93	0.6476	1.27	0.7959
0.26	0.2051	0.6	0.4515	0.94	0.6528	1.28	0.7995
0.27	0.2128	0.61	0.4581	0.95	0.6579	1.29	0.8029
0.28	0.2205	0.62	0.4647	0.96	0.6629	1.3	0.8064
0.29	0.2282	0.63	0.4713	0.97	0.6680	1.31	0.8098
0.3	0.2358	0.64	0.4778	0.98	0.6729	1.32	0.8132
0.31	0.2434	0.65	0.4843	0.99	0.6778	1.33	0.8165
0.32	0.2510	0.66	0.4907	1	0.6827	1.34	0.8198
0.33	0.2586	0.67	0.4971	1.01	0.6875	1.35	0.8230

продовження табл. 1

t	Φ(t)	t	Φ(t)	t	Φ(t)	t	Φ(t)
1.36	0.8262	1.72	0.9146	2.08	0.9625	2.44	0.9853
1.37	0.8293	1.73	0.9164	2.09	0.9634	2.45	0.9857
1.38	0.8324	1.74	0.9181	2.1	0.9643	2.46	0.9861
1.39	0.8355	1.75	0.9199	2.11	0.9651	2.47	0.9865
1.4	0.8385	1.76	0.9216	2.12	0.9660	2.48	0.9869
1.41	0.8415	1.77	0.9233	2.13	0.9668	2.49	0.9872
1.42	0.8444	1.78	0.9249	2.14	0.9676	2.5	0.9876
1.43	0.8473	1.79	0.9265	2.15	0.9684	2.51	0.9879
1.44	0.8501	1.8	0.9281	2.16	0.9692	2.52	0.9883
1.45	0.8529	1.81	0.9297	2.17	0.9700	2.53	0.9886
1.46	0.8557	1.82	0.9312	2.18	0.9707	2.54	0.9889
1.47	0.8584	1.83	0.9328	2.19	0.9715	2.55	0.9892
1.48	0.8611	1.84	0.9342	2.2	0.9722	2.56	0.9895
1.49	0.8638	1.85	0.9357	2.21	0.9729	2.57	0.9898
1.5	0.8664	1.86	0.9371	2.22	0.9736	2.58	0.9901
1.51	0.8690	1.87	0.9385	2.23	0.9743	2.59	0.9904
1.52	0.8715	1.88	0.9399	2.24	0.9749	2.6	0.9907
1.53	0.8740	1.89	0.9412	2.25	0.9756	2.61	0.9909
1.54	0.8764	1.9	0.9426	2.26	0.9762	2.62	0.9912
1.55	0.8789	1.91	0.9439	2.27	0.9768	2.63	0.9915
1.56	0.8812	1.92	0.9451	2.28	0.9774	2.64	0.9917
1.57	0.8836	1.93	0.9464	2.29	0.9780	2.65	0.9920
1.58	0.8859	1.94	0.9476	2.3	0.9786	2.66	0.9922
1.59	0.8882	1.95	0.9488	2.31	0.9791	2.67	0.9924
1.6	0.8904	1.96	0.9500	2.32	0.9797	2.68	0.9926
1.61	0.8926	1.97	0.9512	2.33	0.9802	2.69	0.9929
1.62	0.8948	1.98	0.9523	2.34	0.9807	2.7	0.9931
1.63	0.8969	1.99	0.9534	2.35	0.9812	2.71	0.9933
1.64	0.8990	2	0.9545	2.36	0.9817	2.72	0.9935
1.65	0.9011	2.01	0.9556	2.37	0.9822	2.73	0.9937
1.66	0.9031	2.02	0.9566	2.38	0.9827	2.74	0.9939
1.67	0.9051	2.03	0.9576	2.39	0.9832	2.75	0.9940
1.68	0.9070	2.04	0.9586	2.4	0.9836	2.76	0.9942
1.69	0.9090	2.05	0.9596	2.41	0.9840	2.77	0.9944
1.7	0.9109	2.06	0.9606	2.42	0.9845	2.78	0.9946
1.71	0.9127	2.07	0.9615	2.43	0.9849	2.79	0.9947

закінчення табл. 1

T	Φ(t)	T	Φ(t)	t	Φ(t)
2.8	0.9949	3.16	0.9984	3.52	0.9996
2.81	0.9950	3.17	0.9985	3.53	0.9996
2.82	0.9952	3.18	0.9985	3.54	0.9996
2.83	0.9953	3.19	0.9986	3.55	0.9996
2.84	0.9955	3.2	0.9986	3.56	0.9996
2.85	0.9956	3.21	0.9987	3.57	0.9996
2.86	0.9958	3.22	0.9987	3.58	0.9997
2.87	0.9959	3.23	0.9988	3.59	0.9997
2.88	0.9960	3.24	0.9988	3.6	0.9997
2.89	0.9961	3.25	0.9988	3.61	0.9997
2.9	0.9963	3.26	0.9989	3.62	0.9997
2.91	0.9964	3.27	0.9989	3.63	0.9997
2.92	0.9965	3.28	0.9990	3.64	0.9997
2.93	0.9966	3.29	0.9990	3.65	0.9997
2.94	0.9967	3.3	0.9990	3.66	0.9997
2.95	0.9968	3.31	0.9991	3.67	0.9998
2.96	0.9969	3.32	0.9991	3.68	0.9998
2.97	0.9970	3.33	0.9991	3.69	0.9998
2.98	0.9971	3.34	0.9992	3.7	0.9998
2.99	0.9972	3.35	0.9992	3.71	0.9998
3	0.9973	3.36	0.9992	3.72	0.9998
3.01	0.9974	3.37	0.9992	3.73	0.9998
3.02	0.9975	3.38	0.9993	3.74	0.9998
3.03	0.9976	3.39	0.9993	3.75	0.9998
3.04	0.9976	3.4	0.9993	3.76	0.9998
3.05	0.9977	3.41	0.9994	3.77	0.9998
3.06	0.9978	3.42	0.9994	3.78	0.9998
3.07	0.9979	3.43	0.9994	3.79	0.9998
3.08	0.9979	3.44	0.9994	3.8	0.9999
3.09	0.9980	3.45	0.9994	3.81	0.9999
3.1	0.9981	3.46	0.9995	3.82	0.9999
3.11	0.9981	3.47	0.9995	3.83	0.9999
3.12	0.9982	3.48	0.9995	3.84	0.9999
3.13	0.9983	3.49	0.9995	3.85	0.9999
3.14	0.9983	3.5	0.9995	3.86	0.9999
3.15	0.9984	3.51	0.9996	3.87	0.9999

Таблиця 2. Значення величини χ_2^2 залежно від рівня значущості

$$P(\chi^2 > \chi_2^2)$$

k	5%	1%	0.1%
1	3.84	6.63	10.83
2	5.99	9.21	13.82
3	7.81	11.34	16.27
4	9.49	13.28	18.47
5	11.07	15.09	20.52
6	12.59	16.81	22.46
7	14.07	18.48	24.32
8	15.51	20.09	26.12
9	16.92	21.67	27.88
10	18.31	23.21	29.59
11	19.68	24.72	31.26
12	21.03	26.22	32.91
13	22.36	27.69	34.53
14	23.68	29.14	36.12
15	25.00	30.58	37.70
16	26.30	32.00	39.25
17	27.59	33.41	40.79
18	28.87	34.81	42.31
19	30.14	36.19	43.82
20	31.41	37.57	45.31
21	32.67	38.93	46.80
22	33.92	40.29	48.27
23	35.17	41.64	49.73
24	36.42	42.98	51.18
25	37.65	44.31	52.62
30	43.77	50.89	59.70
40	55.76	63.69	73.40
50	67.50	76.15	86.66
60	79.08	88.38	99.61
70	90.53	100.43	112.32
80	101.88	112.33	124.84
100	124.34	135.81	149.45
200	233.99	249.45	267.54
500	553.13	576.49	603.45
1000	1074.68	1106.97	1143.92

Таблиця 3. Критичні точки розподілу Стьюдента (t-розподілу)

k	5%	1%	0.1%
1	12.71	63.66	636.62
2	4.30	9.92	31.60
3	3.18	5.84	12.92
4	2.78	4.60	8.61
5	2.57	4.03	6.87
6	2.45	3.71	5.96
7	2.36	3.50	5.41
8	2.31	3.36	5.04
9	2.26	3.25	4.78
10	2.23	3.17	4.59
11	2.20	3.11	4.44
12	2.18	3.05	4.32
13	2.16	3.01	4.22
14	2.14	2.98	4.14
15	2.13	2.95	4.07
16	2.12	2.92	4.01
17	2.11	2.90	3.97
18	2.10	2.88	3.92
19	2.09	2.86	3.88
20	2.09	2.85	3.85
21	2.08	2.83	3.82
22	2.07	2.82	3.79
23	2.07	2.81	3.77
24	2.06	2.80	3.75
25	2.06	2.79	3.73
30	2.04	2.75	3.65
40	2.02	2.70	3.55
50	2.01	2.68	3.50
60	2.00	2.66	3.46
70	1.99	2.65	3.44
80	1.99	2.64	3.42
90	1.99	2.63	3.40
100	1.98	2.63	3.39
150	1.98	2.61	3.36
200	1.97	2.60	3.34
500	1.96	2.59	3.31

Таблиця 4а. Критичні точки розподілу Фішера (F-розподілу). Рівень значущості 0.05

k	1	2	3	4	5	7
1	161.4	199.5	215.7	224.6	230.2	236.8
2	18.5	19.0	19.2	19.2	19.3	19.4
3	10.13	9.55	9.28	9.12	9.01	8.89
4	7.71	6.94	6.59	6.39	6.26	6.09
5	6.61	5.79	5.41	5.19	5.05	4.88
6	5.99	5.14	4.76	4.53	4.39	4.21
7	5.59	4.74	4.35	4.12	3.97	3.79
8	5.32	4.46	4.07	3.84	3.69	3.50
9	5.12	4.26	3.86	3.63	3.48	3.29
10	4.96	4.10	3.71	3.48	3.33	3.14
11	4.84	3.98	3.59	3.36	3.20	3.01
12	4.75	3.89	3.49	3.26	3.11	2.91
13	4.67	3.81	3.41	3.18	3.03	2.83
14	4.60	3.74	3.34	3.11	2.96	2.76
15	4.54	3.68	3.29	3.06	2.90	2.71
16	4.49	3.63	3.24	3.01	2.85	2.66
17	4.45	3.59	3.20	2.96	2.81	2.61
18	4.41	3.55	3.16	2.93	2.77	2.58
19	4.38	3.52	3.13	2.90	2.74	2.54
20	4.35	3.49	3.10	2.87	2.71	2.51
22	4.30	3.44	3.05	2.82	2.66	2.46
24	4.26	3.40	3.01	2.78	2.62	2.42
26	4.23	3.37	2.98	2.74	2.59	2.39
28	4.20	3.34	2.95	2.71	2.56	2.36
30	4.17	3.32	2.92	2.69	2.53	2.33
40	4.08	3.23	2.84	2.61	2.45	2.25
60	4.00	3.15	2.76	2.53	2.37	2.17
90	3.95	3.10	2.71	2.47	2.32	2.11
120	3.92	3.07	2.68	2.45	2.29	2.09
∞	3.84	3.00	2.60	2.37	2.21	2.01

закінчення табл. 4а

k	10	15	20	30	60	120
1	241.9	245.9	248.0	250.1	252.2	253.3
2	19.4	19.4	19.4	19.5	19.5	19.5
3	8.79	8.70	8.66	8.62	8.57	8.55
4	5.96	5.86	5.80	5.75	5.69	5.66
5	4.74	4.62	4.56	4.50	4.43	4.40
6	4.06	3.94	3.87	3.81	3.74	3.70
7	3.64	3.51	3.44	3.38	3.30	3.27
8	3.35	3.22	3.15	3.08	3.01	2.97
9	3.14	3.01	2.94	2.86	2.79	2.75
10	2.98	2.85	2.77	2.70	2.62	2.58
11	2.85	2.72	2.65	2.57	2.49	2.45
12	2.75	2.62	2.54	2.47	2.38	2.34
13	2.67	2.53	2.46	2.38	2.30	2.25
14	2.60	2.46	2.39	2.31	2.22	2.18
15	2.54	2.40	2.33	2.25	2.16	2.11
16	2.49	2.35	2.28	2.19	2.11	2.06
17	2.45	2.31	2.23	2.15	2.06	2.01
18	2.41	2.27	2.19	2.11	2.02	1.97
19	2.38	2.23	2.16	2.07	1.98	1.93
20	2.35	2.20	2.12	2.04	1.95	1.90
22	2.30	2.15	2.07	1.98	1.89	1.84
24	2.25	2.11	2.03	1.94	1.84	1.79
26	2.22	2.07	1.99	1.90	1.80	1.75
28	2.19	2.04	1.96	1.87	1.77	1.71
30	2.16	2.01	1.93	1.84	1.74	1.68
40	2.08	1.92	1.84	1.74	1.64	1.58
60	1.99	1.84	1.75	1.65	1.53	1.47
90	1.94	1.78	1.69	1.59	1.46	1.39
120	1.91	1.75	1.66	1.55	1.43	1.35
∞	1.83	1.67	1.57	1.46	1.32	1.22

12.

Таблиця 4б. Критичні точки розподілу Фішера (F-розподілу). Рівень значущості 0.01

k	1	2	3	4	5	7
1	4052	4999	5403	5625	5764	5928
2	98.5	99.0	99.2	99.2	99.3	99.4
3	34.1	30.8	29.5	28.7	28.2	27.7
4	21.2	18.0	16.7	16.0	15.5	15.0
5	16.26	13.27	12.06	11.39	10.97	10.46
6	13.75	10.92	9.78	9.15	8.75	8.26
7	12.25	9.55	8.45	7.85	7.46	6.99
8	11.26	8.65	7.59	7.01	6.63	6.18
9	10.56	8.02	6.99	6.42	6.06	5.61
10	10.04	7.56	6.55	5.99	5.64	5.20
11	9.65	7.21	6.22	5.67	5.32	4.89
12	9.33	6.93	5.95	5.41	5.06	4.64
13	9.07	6.70	5.74	5.21	4.86	4.44
14	8.86	6.51	5.56	5.04	4.69	4.28
15	8.68	6.36	5.42	4.89	4.56	4.14
16	8.53	6.23	5.29	4.77	4.44	4.03
17	8.40	6.11	5.18	4.67	4.34	3.93
18	8.29	6.01	5.09	4.58	4.25	3.84
19	8.18	5.93	5.01	4.50	4.17	3.77
20	8.10	5.85	4.94	4.43	4.10	3.70
22	7.95	5.72	4.82	4.31	3.99	3.59
24	7.82	5.61	4.72	4.22	3.90	3.50
26	7.72	5.53	4.64	4.14	3.82	3.42
28	7.64	5.45	4.57	4.07	3.75	3.36
30	7.56	5.39	4.51	4.02	3.70	3.30
40	7.31	5.18	4.31	3.83	3.51	3.12
60	7.08	4.98	4.13	3.65	3.34	2.95
90	6.93	4.85	4.01	3.53	3.23	2.84
120	6.85	4.79	3.95	3.48	3.17	2.79
∞	6.64	4.61	3.78	3.32	3.02	2.64

Закінчення табл. 4б

k	10	15	20	30	60	120
1	6056	6157	6209	6261	6313	6339
2	99.4	99.4	99.4	99.5	99.5	99.5
3	27.2	26.9	26.7	26.5	26.3	26.2
4	14.5	14.2	14.0	13.8	13.7	13.6
5	10.05	9.72	9.55	9.38	9.20	9.11
6	7.87	7.56	7.40	7.23	7.06	6.97
7	6.62	6.31	6.16	5.99	5.82	5.74
8	5.81	5.52	5.36	5.20	5.03	4.95
9	5.26	4.96	4.81	4.65	4.48	4.40
10	4.85	4.56	4.41	4.25	4.08	4.00
11	4.54	4.25	4.10	3.94	3.78	3.69
12	4.30	4.01	3.86	3.70	3.54	3.45
13	4.10	3.82	3.66	3.51	3.34	3.25
14	3.94	3.66	3.51	3.35	3.18	3.09
15	3.80	3.52	3.37	3.21	3.05	2.96
16	3.69	3.41	3.26	3.10	2.93	2.84
17	3.59	3.31	3.16	3.00	2.83	2.75
18	3.51	3.23	3.08	2.92	2.75	2.66
19	3.43	3.15	3.00	2.84	2.67	2.58
20	3.37	3.09	2.94	2.78	2.61	2.52
22	3.26	2.98	2.83	2.67	2.50	2.40
24	3.17	2.89	2.74	2.58	2.40	2.31
26	3.09	2.81	2.66	2.50	2.33	2.23
28	3.03	2.75	2.60	2.44	2.26	2.17
30	2.98	2.70	2.55	2.39	2.21	2.11
40	2.80	2.52	2.37	2.20	2.02	1.92
60	2.63	2.35	2.20	2.03	1.84	1.73
90	2.52	2.24	2.09	1.92	1.72	1.60
120	2.47	2.19	2.03	1.86	1.66	1.53
∞	2.32	2.04	1.88	1.70	1.47	1.32

13.

ЛІТЕРАТУРА

1. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. − М.: ЮНИТИ, 1998. − 1022 с.
2. Андерсен Т. Введение в многомерный статистический анализ. − М.: Физматгиз, 1963. − 500 с.
3. Аптон Г. Анализ таблиц сопряженности. − М.: Финансы и статистика, 1982. − 144 с.
4. Аренс Х., Лейтер Ю. Многомерный дисперсионный анализ. – М.: Финансы и статистика, 1985. – 231 с.
5. Бард Й. Нелинейное оценивание параметров. − М.: Финансы и статистика, 1979. − 349 с.
6. Бахрушин В.Є. Аналіз даних. − Запоріжжя: ГУ "ЗІДМУ", 2006. − 170 с.
7. Бендат Дж., Пирсол А. Применение корреляционного и спектрального анализа. – М.: Мир, 1979. – 311 с.
8. Бендат Дж., Пирсол А. Прикладной анализ случайных данных. − М.: Мир, 1989. − 540 с.
9. Боровиков В.П. Популярное введение в программу STATISTICA. − М.: Компьютер−Пресс, 1998. − 267 с.
10. Брандт З. Анализ данных: Статистические и вычислительные методы для научных работников и инженеров. − М.: Мир, ООО "Издательствово АСТ", 2003. − 686 с.
11. Вучков И., Бояджиева Л., Солаков Е. Прикладной линейный регрессионный анализ. – М.: Финансы и статистика, 1987. – 239 с.
12. Гаек Я., Шидак З. Теория ранговых критериев. − М.: Наука, 1971. − 376 с.
13. Гайдышев И. Анализ и обработка данных: Специальный справочник. − С.Пб.: Питер, 2001. − 752 с.
14. Гирко В.Л. Многомерный статистический анализ. – К.: Высшая школа, 1988. – 320 с.
15. Гутер Р.С., Овчинский Б. В. Элемента численного анализа и математической обработки данных эксперимента. - М. 1971.
16. Дрейпер Н., Смит Г. Прикладной регрессионный анализ: В 2 т. − М.: Финансы и статистика, 1986. − Т. 1. – 366 с.; 1987. − Т. 2. − 351 с.
17. Дюран Б., Оделл П. Кластерный анализ. – М.: Статистика, 1977.
18. Енюков И.С. Методы, алгоритмы, программы многомерного статистического анализа. – М.: Финансы и статистика, 1986.

19. ЖлуктенкоВ.І., Наконечний С.І., Савіна С.С. Теорія ймовірностей і математична статистика. – К.:КНЕУ, 2001.-336 с.
20. Іващенко П.О., Семеняк І.В., Іванов В.В. Багатовимірний статистичний аналіз. – Харків: Основа, 1992. – 144 с.
21. Иберла К. Факторный анализ. – М.: Статистика, 1980.
22. Кендалл М. Многомерный статистический анализ и временные ряды. – М.: Наука, 1976. – 511 с.
23. Кендалл М.Дж., Стюарт А. Статистические выводы и связи. – М.: Наука, 1973. – 899 с.
24. Королюк В.С., Боровских Ю.В. Асимптотический анализ распределений статистик. – К.: Наукова думка, 1984. – 301 с.
25. Лапач С.Н., Чубенко А.В., Бабич П.Н. Статистика в науке и бизнесе. – К.: МОРИОН, 2002. – 640с.
26. Литтл Р.Дж., Рубин Д.Б. Статистический анализ данных с пропусками. − М.: Финансы и статистика, 1991. − 336 с.
27. Макарова Н.В., Трофимец В.Я. Статистика в Excel. М.: Финансы и статистика, 2002. – 368с.
28. Многомерный статистический анализ в экономике / Под ред. В.Н. Тамашевича – М.: ЮНИТИ, 1999. – 598 с.
29. Новицкий П.В., Зограф И.А. Оценка погрешностей результатов измерений. – Л.: Энергоатомиздат, 1991. − 304 с.
30. Паніотто В.І., Максименко В.С., Харченко Н.М. Статистичний аналіз соціологічних даних. – К.: Вид. дім ”КМ Академія”, 2004. – 270с.
31. Плюта В. Сравнительный многомерный анализ в эконометрическом моделировании. – М.: Финансы и статистика, 1989. – 175 с.
32. Тьюки Дж. Анализ результатов наблюдений. Разведочный анализ. – М.: Мир, 1981. – 693 с.
33. Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере. − М.: ИНФРА−М, 2003. − 544 с.
34. Факторный, дискриминантный и кластерный анализ. – М.: Финансы и статистика, 1989. – 215 с.
35. Ферстер Э., Ренц Б. Методы кореляционного и регресионного анализа. М.: Финансы и статистика, 1983.
36. Химмельбау Дж. Анализ процессов статистическими методами. – М.: Мир, 1973. – 957 с.
37. Холлендер М., Вульф Д. Непараметрические методы статистики. − М.: Финансы и статистика, 1983. − 518 с.

38. Хьюбер П. Робастность в статистике. − М.: Мир, 1984. − 304 с.
39. Худсон Д. Статистика для физиков. - М.: Мир, 1967. - 242 с.
40. Эфрон Б. Нетрадиционные методы многомерного статистического анализа. – М.: Финансы и статистика, 1988. – 262 с.

Навчальне видання

Грицюк Петро Михайлович
Остапчук Оксана Петрівна

АНАЛІЗ ДАНИХ

Навчальний посібник

Друкується в авторській редакції

Підписано до друку 28.03.2008 р. Формат $60 \times 84\ 1/16$.
Папір друкарський №1. Гарнітура Times. Друк різографічний.
Ум.-друк. арк. 12,7. Обл.-вид. арк.13,3.
Тираж 100 прим. Зам № 836.

Редакційно-видавничий центр
Національного університету
водного господарства та природокористування
33028, Рівне, вул. Соборна, 11.

Свідоцтво про внесення суб'єкта видавничої справи до державного реєстру видавців, виготівників і розповсюджувачів видавничої продукції РВ №31 від 26.04.2005 р.