КЛАСИЧНИЙ ПРИВАТНИЙ УНІВЕРСИТЕТ

В.Є. Бахрушин

МЕТОДИ АНАЛІЗУ ДАНИХ

Навчальний посібник

Запоріжжя
Класичний приватний університет
2011

ББК 22.172.517.8:32.973:65.05
УДК 519.2:681.3
Б 30

Рецензенти:
С.І. Гоменюк, д. т. н., професор
Запорізького національного університету;
Г.В. Корніч, д. ф.-м. н., професор
Запорізького національного технічного університету;
В.В. Слесарєв, д. т. н., професор
Національного гірничого університету.

Бахрушин В.Є.

Б 30 Методи аналізу даних : навчальний посібник для студентів / В.Є. Бахрушин. – Запоріжжя : КПУ, 2011. – 268 с.

ISBN 978-966-414-103-8

Подано відомості про основні поняття, теоретичні підґрунтя та математичні методи аналізу даних. Розглянуто основні параметри описової статистики, методи побудови емпіричних функцій розподілу, принципи побудови й критерії перевірки гіпотез про однорідність вибірок та їх відповідність певним законам розподілу, теоретичні основи та базові алгоритми дисперсійного, кореляційного, регресійного та факторного аналізу, а також методи класифікації даних.

Призначено для студентів, аспірантів та науковців в галузі системних наук та інформаційних технологій.

УДК 519.2:681.3
ББК 22.172.517.8:32.973:65.05

ISBN 978-966-414-103-8

ЗМІСТ

ВСТУП

Математичні методи аналізу даних широко використовують при дослідженні різноманітних систем і процесів – природних, технічних, екологічних, економічних, соціальних тощо. З огляду на це формування відповідних знань та навичок є необхідною складовою підготовки фахівців у галузі системних наук і кібернетики, інформатики та багатьох інших галузей знань.

Про застосування статистичних методів аналізу даних вперше згадується у Книзі чисел. Основи сучасних методів аналізу даних були закладені Томасом Байєсом (байєсівський підхід, байєсівські оцінки), Даніїлом Бернуллі (застосування нормального розподілу в теорії похибок, перші таблиці нормального розподілу, поділ похибок спостережень на випадкові й систематичні тощо), Карлом Гаусом (метод найменших квадратів); Андрієм Миколайовичем Колмогоровим (статистичні методи контролю за якістю, статистика Колмогорова – Смирнова, узагальнена відстань Колмогорова тощо), Адрієном Марі Лежандром (метод найменших квадратів), Вільфредо Парето (розподіл Парето, діаграма Парето), Френсисом Гальтоном (теорія кореляції), Карлом Пірсоном (теорія кореляції, критерії згоди, метод головних компонент), Чарльзом Спірменом (техніка факторного аналізу, рангова кореляція), Рональдом Фішером (метод максимальної правдоподібності, критерії згоди тощо). Помітний внесок у розвиток цих методів зробив видатний український математик Михайло Васильович Остроградський, який у середині XIX ст. сформулював основні ідеї статистичного контролю за якістю виробництва.

Сучасні методи аналізу даних були розвинені у працях Ю.П. Адлера, С.А. Айвазяна, Т. Андерсона, Й. Барда, Л.М. Большева, Б.В. Гнеденко, Н. Дрейпера, А.М. Дуброва, К. Іберли, І.А. Ібрагімова, А.Г. Івахненка, Дж. Кіфера, К.Х. Крамера, М. Кендалла, Г. Куллдорфа, Б.Ю. Лемешка, Ю.В. Лінника, Г.В. Мартинова, В.В. Налімова, М.С. Нікуліна, О.І. Орлова, І.М. Парасюка, Е. Пітмена, Ю.В. Прохорова, Е. Пятецького-Шапіро, С.Р. Рао, Г. Смита, А. Стьюарта, Дж. Тьюкі, Г. Хоттелінга, П. Хьюбера, А. Хьютсона, О.О. Чупрова, Д.У. Юла та багатьох інших дослідників.

Останнім часом значного поширення набувають нові технології й методи аналізу даних, зокрема методи інтелектуального аналізу даних (data mining), які використовують для виявлення прихованих закономірностей у великих масивах даних, та нейроінформатики, а також методики й засоби статистичного контролю за якістю на виробництві та в управлінні організаціями.

Основні процедури аналізу даних найчастіше реалізують за допомогою сучасних комп'ютерних технологій. При цьому дослідники або самі будують розрахункові алгоритми й пишуть відповідні комп'ютерні програми, або ви-

користовують наявне програмне забезпечення – електронні таблиці MS Excel, спеціалізовані пакети SPSS, STATISTICA, математичні пакети загального призначення MatLab, MathCad тощо. Але навіть при застосуванні спеціалізованих пакетів досліднику необхідно володіти теоретичними основами математичних методів аналізу даних, оскільки зазвичай це передбачає необхідність вибору оптимальних алгоритмів та певних параметрів їх реалізації, іноді з декількох сотень можливих варіантів. Це зумовлює необхідність вивчення майбутніми фахівцями основних понять та алгоритмів аналізу даних.

Курс "Аналіз даних" є нормативною складовою стандартів підготовки бакалаврів, що навчаються за напрямами "Системний аналіз", "Інформатика", "Прикладна математика" тощо. Окремі питання аналізу даних входять також до програм підготовки фізиків, інженерів, економістів, соціологів, психологів і фахівців інших галузей знань.

Передбачається, що студенти володіють основними поняттями й методами математичного аналізу, лінійної алгебри, аналітичної геометрії, теорії ймовірності й математичної статистики, чисельних методів, програмування.

У посібнику розглянуто основні типи даних, що підлягають аналізу; методи побудови описової статистики й емпіричних функцій розподілу; критерії перевірки статистичних гіпотез щодо однорідності вибірок та порівняння емпіричних функцій розподілу з теоретичними моделями; критерії та методи перевірки наявності статистичного зв'язку між ознаками; теоретичні основи та основні методи регресійного й факторного аналізу; методи класифікації даних. Наведено приклади вирішення типових завдань із застосуванням сучасних програмних засобів.

Для отримання додаткової інформації з питань теоретичних основ, програмного забезпечення й практики застосування сучасних методів аналізу даних можна використовувати інформацію, що розміщена на таких сайтах мережі Інтернет:

1) http://datan.ucoz.ru (сайт циклу дисциплін "Аналіз даних", "Математичні методи аналізу даних", "Комп'ютеризовані технології аналізу даних");

2) http://www.basegroup.ru (сайт BaseGroup Labs – провідного російського розробника програмного забезпечення з аналізу даних);

3) http://www.statsoft.ru/home/textbook/default.htm (електронний підручник статистики фірми StatSoft – провідного розробника статистичного програмного забезпечення);

4) http://orlovs.pp.ru (сайт "Високі статистичні технології" професора О.І. Орлова);

5) http://www.aup.ru/books/m163/ (підручник О.І. Орлова "Прикладна статистика");

6) http://www.ami.nstu.ru/~headrd (сторінка професора Б.Ю. Лемешка на сайті Новосибірського державного технічного університету);

7) http://uk.wikipedia.org, http://ru.wikipedia.org, http://www.wikipedia.org (Вікіпедія, необхідно перейти на потрібні сторінки за ключовими словами, наприклад "математична статистика", "критерій Колмогорова", тощо; є багато посилань на інші електронні ресурси);

8) http://www.biometrica.tomsk.ru (Біометрика – сайт доказової біології та медицини; багато матеріалів із застосування методів аналізу даних в біології та медицині);

9) http://dvo.sut.ru/libr/opds/i130hodo_part1/index.htm;

10) http://www.dvo.sut.ru/libr/opds/i130hod2/index.htm (навчальний посібник Г.Б. Ходасевича "Обробка експериментальних даних на ЕОМ");

11) http://lib.socio.msu.ru/l/library?e=d-000-00---001ucheb--00-0-0-0prompt-10---4------0-1l--1-ru-50---20-help---00031-001-1-0windowsZz-1251-10&a=d&c=01 ucheb &cl=CL1&d=HASHe10c3b36c7d751dd18704b (навчальний посібник з роботи у пакеті SPSS);

12) http://www.gmdh.net/gmdh.htm (сайт з методу групового врахування аргументів, розробленого відомим українським математиком А.Г. Іваненко);

13) http://www.machinelearning.ru (сайт, присвячений методам машинного навчання, розпізнавання образів та інтелектуального аналізу даних);

14) http://riskcontrol.ru (сайт Центру статистичних досліджень – розробника статистичного пакету "Евріста");

15) http://attestatsoft.narod.ru/index.htm (сайт розробника статистичного програмного забезпечення AtteStatSoft);

16) http://www.medstatistica.com (сайт "Статистика в медико-біологічних дослідженнях).

Автор вдячний за обговорення матеріалу посібника й окремих питань аналізу даних чл.-кор. НАН України В.Г. Литовченку та В.В. Сльозову, професорам і докторам наук М.С. Блантеру, О.М. Горбаню, Г.В. Корнічу, О.С. Куценко, Д.М. Левіну, Д.І. Левінзону, Л.М. Любчику, О.І. Михальову, Л.Н. Сергєєвій, В.В. Слесарєву, Л.Д. Чумакову, доцентам В.М. Буйницькій, М.О. Ігнахіній, І.М. Нацюку.

1. ОСНОВНІ ПОНЯТТЯ Й ЗАВДАННЯ АНАЛІЗУ ДАНИХ. ЗАГАЛЬНА МЕТОДОЛОГІЯ ДОСЛІДЖЕННЯ

Успішність застосування будь-якого методу аналізу даних залежить від відповідності аналізованих даних його вихідним припущенням. Методи, придатні для одного типу даних, можуть призводити до серйозних помилок при їх використанні для даних інших типів.

1.1. Класифікація ознак за шкалами вимірювання

Першим етапом аналізу будь-яких даних зазвичай є визначення їх типу. Основною є класифікація даних за шкалами їх вимірювання. Згідно з нею розрізняють такі типи ознак.

Номінальні ознаки (**ознаки з невпорядкованими станами**, **класифікаційні ознаки**) – це дані, що вимірюють 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. Прикладами номінальних ознак є назви біологічних видів, назви навчальних дисциплін, кольори тощо. З погляду автоматизації аналізу даних і застосування стандартних алгоритмів доцільно обирати такі позначення класів: 0, 1, 2, … Але з цими числами не можливо виконувати будь-які дії, крім перевірки їх рівності або нерівності.

Порядкові ознаки (**ознаки з упорядкованими станами**, **ординальні ознаки**) – це дані, що вимірюють в порядкових шкалах. Ці дані можуть порівнюватися між собою у певному відношенні: "більше – менше", "легше – важче", "правіше – лівіше" тощо. Прикладами порядкових ознак є сила землетрусу, військові звання, оцінки студентів тощо. Якщо значення порядкової ознаки є числами, то вони можуть застосовуватися і для порівняння ступеня вияву 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. Дії, що можуть виконуватися з числовими характеристиками даних, залежать від шкали вимірювань.

В узагальненому вигляді характеристики основних типів даних, згідно з [44], наведено в табл. 1.1.

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 вимі-

ряні у шкалах різного типу. Зворотну операцію – перетворення даних, що виміряні у нижчих шкалах, до вищих – вважають некоректною. Зниження шкали призводить до втрати частини наявної інформації про досліджувані ознаки.

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, збереження відношення інтервалів, періодичність	Зсув $y = x + nb$, b = const, n = 0, 1, 2, …	Обчислення δ_{ij}, рангів R_i та інтервалів (різниць між даними)	Арифметичні дії над інтервалами
Відношень	Еквівалентність, перевага, збереження відношення інтервалів, збереження відношення двох значень	Розтягання $y = ax$, $a > 0$	Усі арифметичні операції	Будь-яка придатна 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 аргументу логарифма	Будь-яка потрібна обробка

Важливими типами класифікації є поділ ознак за дискретністю або неперервністю теоретичної функції розподілу, законом розподілу тощо.

Як характеристики вибірки можна використовувати точкові та інтервальні оцінки. **Точковими оцінками** параметрів вибірки називають такі оцінки, що визначаються одним числом. Прикладами таких оцінок є середні арифметичні й медіани вибірок. При малих обсягах вибірок, а також при їх значному відхиленні від нормального закону розподілу точкові оцінки можуть істотно відхилятися від істинних значень оцінюваних параметрів. Тому поряд з ними використовують інтервальні оцінки параметрів. **Інтервальні оцінки** визначаються двома числами – межами інтервалу, до якого із заданою ймовірністю потрапляє оцінюваний параметр.

1.2. Описова статистика

Описова статистика – це набір основних статистичних показників емпіричної вибірки значень кількісної ознаки. Стандартні методи їх розрахунку, як правило, розроблені, виходячи із припущення, що розподіл є нормальним. Причиною цього є наявність зручного математичного апарату для обробки відповідних даних. Не меншу роль у надмірно широкому застосуванні методів, призначених для аналізу нормально розподілених даних, відіграє необґрунтоване припущення, що майже всі випадкові дані підпорядковуються нормальному закону розподілу. Щодо так званого закону похибок, згідно з яким їх розподіл завжди є нормальним, відомий французький фізик Г. Ліппман ще більше ста років тому так прокоментував цю ситуацію: “Всі вірять у закон похибок, бо експериментатори думають, що він є математичною теоремою, а теоретики вважають, що його встановлено експериментальним шляхом” [44]. Але припущення про нормальний розподіл часто виявляється помилковим. Якщо розподіл даних істотно відрізняється від нормального, необхідно використовувати інші методи та формули. У зв’язку із цим процедуру аналізу емпіричної вибірки завжди слід починати з перевірки закону розподілу на нормальність. Відповідні методи будуть розглянуті нижче.

На практиці найчастіше мають справу з **вибірковими характеристиками**, які розраховують за обмеженою кількістю значень досліджуваного показника, що становлять певну вибірку з генеральної сукупності. Вони є оцінками відповідних **генеральних статистичних характеристик** (параметрів розподілу). На відміну від останніх, вибіркові характеристики є випадковими величинами і змінюються від вибірки до вибірки. Зазвичай для позначення оцінки параметра θ використовують позначення $\tilde{\theta}$. З метою спрощення формул надалі таке позначення буде використовуватися лише у випадках, коли з тексту незрозуміло, що мається на увазі – оцінка параметра чи його істинне значення.

Оцінку називають **конзистентною** (**спроможною**), якщо із збільшенням обсягу вибірки вона наближається (за ймовірністю) до оцінюваного параметра. Для перевірки спроможності часто використовують її до-

статню умову: оцінка є спроможною, якщо при прямуванні обсягу вибірки до нескінченності її математичне сподівання наближається до істинного значення досліджуваного параметра, а її дисперсія – до нуля.

Оцінку називають **незміщеною**, якщо для будь-якого обсягу вибірки її математичне сподівання дорівнює оцінюваному параметру. В деяких випадках, зокрема при побудові багатофакторних регресійних моделей, використовують зміщені оцінки параметрів через нестійкість алгоритмів отримання незміщених оцінок.

Якщо для деякого параметра існує декілька незміщених оцінок, то **більш ефективною** з них вважають ту, що має найменшу дисперсію.

Основними завданнями описової статистики є визначення центру, ширини, симетрії й протяжності розподілу.

Центр статистичного розподілу характеризують його математичне сподівання, середнє значення, медіана та мода. Іноді його визначають як середнє арифметичне мінімального й максимального значень вибірки (**центр розмаху**), але похибка такого методу зазвичай є неприпустимо високою.

Математичне сподівання неперервної кількісної ознаки визначають як центр тяжіння розподілу й визначають за формулою:

$$\overline{X} = \int_{-\infty}^{+\infty} xf\left(x\right)dx, \tag{1.1}$$

де $f(x)$ – щільність розподілу.

Деякі розподіли, наприклад розподіл Коші з функцією щільності $f\left(x\right) = \dfrac{a}{\pi\left(a^2 + x^2\right)}$ (a = const), не мають математичного сподівання, оскільки інтеграл, що стоїть у правій частині (1.1) розбігається. У загальному випадку умовою існування математичного сподівання є те, що крива щільності розподілу $f(x)$ має спадати швидше, ніж $1/x^2$.

Для дискретних випадкових величин, що виміряні у кількісних шкалах, замість математичного сподівання використовують **середнє значення** (**середнє арифметичне**), яке обчислюють за формулою:

$$\mu = \sum_{i=0}^{\infty} x_i p_i, \tag{1.2}$$

де x_i – значення, яких випадкова величина, розподілена на інтервалі $\left(-\infty; +\infty\right)$, набуває з імовірностями p_i.

Вибіркове середнє арифметичне та його середнє квадратичне відхилення обчислюють за формулами:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,\ s_{\bar{x}} = \frac{\sigma}{\sqrt{n}}, \tag{1.3}$$

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, то вибіркове середнє обчислюють за формулою:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} b_i \nu_i \,, \tag{1.4}$$

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

Вибіркове середнє є спроможною незміщеною оцінкою математичного сподівання, у випадку, якщо останнє існує. Широке застосування вибіркового середнього як оцінки центру розподілу також пов'язано з тим, що воно є єдиною оцінкою, для якої існує аналітичний вираз, що може бути використаний в інших співвідношеннях і формулах. Ще однією причиною є те, що середнє арифметичне та інші види середніх значень дають змогу усунути випадкові коливання показника й отримати величини, які точніше характеризують об'єкт дослідження. Водночас, якщо зміни досліджуваного показника з часом чи при змінюванні інших параметрів є істотними для досягнення цілей дослідження, застосування середніх значень може виявитися необґрунтованим.

Закон розподілу середнього арифметичного при $n \geq 30$ є близьким до нормального незалежно від виду розподілу вихідних даних, якщо значення контрексцесу вибірки відмінно від нуля [44]. Недоліком середнього арифметичного як оцінки центру розподілу є те, що на його значення істотно впливають екстремальні значення, які можуть виявитися помилковими. Застосування середнього арифметичного може виявитися незручним, якщо різні емпіричні точки мають різну важливість, або діапазон змінювання значень ознаки є занадто широким.

Тому поряд із середнім арифметичним використовують інші види середніх величин:

– **середнє гармонічне**:

$$h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}; \tag{1.5}$$

– **середнє геометричне** (**середнє пропорційне**):

$$g = \sqrt[n]{\prod_{i=1}^{n} x_i} = \prod_{i=1}^{n} \sqrt[n]{x_i}\,; \tag{1.6}$$

– **степеневе середнє**:

$$\omega_\alpha = \sqrt[\alpha]{\frac{1}{n}\sum_{i=1}^{n} x_i^\alpha},\ \alpha > 0; \tag{1.7}$$

– **зважені степеневі середні**:

$$\Omega_\alpha = \sqrt[\alpha]{\frac{\sum_{i=1}^{n} w_i x_i^\alpha}{\sum_{i=1}^{n} w_i}},\ \alpha > 0, \tag{1.8}$$

де w_i $(i = 1,2,...,n)$ – вагові коефіцієнти; степеневі середні є окремим випадком зважених степеневих середніх.

Для розглянутих величин виконується нерівність $h \le g \le \bar{x} \le \omega_\alpha$ (остання нерівність виконується, якщо $\alpha > 1$).

Розрахуємо різні характеристики центру розподілу для даних, наведених у табл. 1.2.

Таблиця 1.2

Значення елементів вибірки

9	12	7	5	6	8	11	9	1	12
7	20	2	6	4	1	8	13	15	3

Згідно з формулами (1.3, 1.5–1.7), одержимо: вибіркове середнє $\bar{x}$ = 7,95; середнє гармонічне h = 4,144417; середнє геометричне g = 6,188113; степеневі середні w_2 = 9,270922; w_3 = 10,35166.

Середнє геометричне часто використовують для визначення середнього темпу зростання досліджуваного показника, а також у випадках, коли досліджувані значення утворюють геометричну прогресію або змінюються за експоненціальним законом.

Степеневі середні використовують з метою збільшення або зменшення внеску певних значень у результат. Відомим у статистиці прикладом степеневого середнього є стандартне відхилення генеральної сукупності $\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n}}$, що є середньоквадратичним відхиленням від середнього арифметичного. Іншим відомим прикладом є довжина вектора $a = \sqrt{x_1^2 + x_2^2 + ... + x_n^2} = \sqrt{n}\omega_2$, яка є нормованим середнім квадратичним значенням його проекцій.

Слід зазначити, що середнє геометричне використовують лише у випадках, коли ознака x_i може набувати тільки додатних значень, оскільки для від'ємних значень функції, що стоять у правій частині наведених вище

виразів, визначені не для всіх n та α. Якщо елементи можуть набувати нульових значень, то відповідні середні дорівнюватимуть нулю незалежно від значень усіх інших елементів, що неприйнятно з погляду їх змістової інтерпретації як центрів розподілу. При побудові розрахункових алгоритмів необхідно враховувати, що для вибірок великого обсягу добуток, який використовується у першій формулі для середнього геометричного, може перевищити граничне значення, допустиме для певного типу даних. У цьому випадку використовувати другу формулу.

Середнє гармонічне часто використовують, коли шуканим показником є величина, обернена значенню усередненої ознаки. Як і у попередньому випадку, його визначають лише для вибірок, утворених з додатних значень. Прикладами практичного застосування середнього гармонічного є визначення електричного опору n паралельно з'єднаних резисторів: $R = \dfrac{1}{\sum\limits_{i=1}^{n} \dfrac{1}{R_i}} = \dfrac{h_R}{n}$ та електричної ємності n послідовно з'єднаних конденсаторів: $C = \dfrac{1}{\sum\limits_{i=1}^{n} \dfrac{1}{C_i}} = \dfrac{h_C}{n}$.

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

$$\mu \in \left[\bar{x} - t_{n-1,\alpha} \frac{s_{\bar{x}}}{\sqrt{n-1}}; \bar{x} + t_{n-1,\alpha} \frac{s_{\bar{x}}}{\sqrt{n-1}} \right], \tag{1.9}$$

де α – рівень значущості, $t_{n-1,\alpha}$ – значення оберненої функції t-розподілу.

Якщо стандартне відхилення оцінюють за самою вибіркою і $n < 30$, то одержувані значення $s_{\bar{X}}$ мають занадто великий розкид і застосування формули (1.9) стає неправомірним. При $n \geq 30$, як було зазначено вище, розподіл середнього арифметичного стає близьким до нормального, тому формулу (1.9) без великої похибки можна застосовувати й для інших типів розподілу вихідних даних.

Якщо М елементам сукупності загальним обсягом N властива певна дихотомічна якісна ознака (значення 1), а іншим вона не властива (значення 0), то середнє значення сукупності $\mu = M / N$.

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:

$$\int_{-\infty}^{m} f(x)\,dx = \int_{m}^{+\infty} f(x)\,dx = \frac{1}{2}. \tag{1.10}$$

Медіана є найбільш загальною й фундаментальною характеристикою центра розподілу, оскільки вона базується на принципі симетрії.

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

$$\sum_{i=0}^{m-1} p_i \le \frac{1}{2};\ \sum_{i=0}^{m} p_i \ge \frac{1}{2}. \tag{1.11}$$

Вона може бути визначена як розв'язок рівняння:

$$F_n(x) = \frac{1}{2}, \tag{1.12}$$

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

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

$$me_x = \begin{cases} X_{(m)},\ n = 2k+1; \\ \frac{1}{2}\left[X_{(m-1)} + X_{(m)}\right],\ n = 2k \end{cases}, \tag{1.13}$$

де $X_{(m)}$ – елементи варіаційного ряду, n – його чисельність.

Іноді вважають, що для парних n існує дві медіани, і для визначеності за медіану беруть меншу з них.

Модою m_x називають точку максимуму емпіричної функції щільності розподілу. Як характеристику центру моду можна використовувати лише для розподілів із симетричною кривою щільності розподілу.

Для даних табл. 1.2 медіана дорівнює 7,5, а моди – 1, 7, 9 та 12 (тобто маємо чотири моди).

В окремих випадках можуть спостерігатися багатомодальні розподіли. В реальних ситуаціях для достатньо великої кількості даних це зазвичай свідчить про неоднорідність вибірки, тобто досліджувана вибірка може розглядатися як суміш декількох однорідних вибірок. Наприклад, якщо ми побудуємо функцію розподілу за зростанням для вибірки, що складається з приблизно рівних кількостей дорослих осіб та дітей певного віку, то одержимо функцію щільності розподілу з двома максимумами (модами). Існують також розподіли, зокрема рівномірний, які не мають моди.

При застосуванні стандартних програмних пакетів для визначення моди розподілу слід мати на увазі, що для багатомодальних розподілів вони часто виявляють лише один з них. При цьому його значення залежить від алгоритму пошуку. Зокрема, функція "МОДА" електронних таблиць MS Excel для даних, що розглядаються, отримає значення моди 9, оскільки воно буде знайдено першим.

Співвідношення між середнім арифметичним, медіаною та модою розподілу залежить від знака коефіцієнта асиметрії розподілу. Якщо він додатний, то $m_x < me_x < \bar{x}$, у протилежному випадку – $\bar{x} < me_x < m_x$. Якщо ж коефіцієнт асиметрії дорівнює нулю, то ці три показники центру розподілу є рівними.

Ефективність різних методів оцінювання центру розподілу залежить від його виду [44]. Застосування середнього арифметичного з погляду ефективності, оцінюваної як мінімум дисперсії відповідних оцінок, обґрунтовано для одномодальних розподілів, близьких до нормального з контрексцесом від 0,515 до 0,645. Використання медіани є ефективнішим за інші оцінки для гостровершинних одномодальних розподілів з величиною контрексцесу $\chi < 0{,}515$, а для плосковершинних й двомодальних розподілів ефективність медіанного оцінювання наближається до нуля. У цьому разі доцільно використовувати центр згинів. Для обмежених розподілів найбільш ефективною оцінкою їх центру є центр розмаху. Методи визначення двох останніх характеристик будуть наведені нижче.

При виборі методу оцінювання центру розподілу, крім ефективності оцінки, слід також враховувати її стійкість до **промахів** (даних, що помилково потрапили до досліджуваної вибірки). Найбільш чутливою до них величиною є центр розмаху, оскільки промахи найчастіше виявляються найбільш віддаленими від центру розподілу точками. Найбільш стійкими є квантильні оцінки – медіана та центр згинів.

Для усунення промахів часто використовують **цензурування** вибірки, під яким розуміють відкидання найбільш віддалених від центру розподілу елементів вибірки. У найпростішому випадку цензурування здійснюють за правилом 3σ, згідно з яким промахами вважають усі елементи, що знаходяться на відстані понад 3σ від центру. При цьому стандартне відхилення визначають після видалення сумнівних даних. Це правило є обґрунтованим для близьких до нормального розподілів, але може виявитися занадто жорстким в інших випадках. У праці [1] пропонується визначати стандартне відхилення за всією вибіркою, а межі цензурування встановлювати залежно від її обсягу: 4σ, якщо кількість елементів знаходиться в межах від 7 до 100; 4,5σ – при $100 < n \le 1000$; 5σ – при $1000 < n \le 10000$. У загальному випадку межі цензурування необхідно визначати з урахуванням закону розподілу даних. Зокрема для рівномірного розподілу

$\sigma_x^2 = \dfrac{(b-a)^2}{12}$, звідки $b-a = 2\sqrt{3}\sigma$. Тому помилками можна вважати будь-які точки, розташовані на відстані понад $\sim 1{,}8\sigma$ від центру. З іншого боку, для розподілу Лапласа, імовірність отримання даних, віддалених від центру на відстань 3σ, є занадто високою, щоб вважати їх помилковими.

Як показники ширини розподілу найчастіше використовують дисперсію і стандартне відхилення вибірки. Поряд з ними застосовують також середні відхилення, середню різницю Джині, квантильні та інші оцінки.

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 – $\sigma^2 = \int\limits_{-\infty}^{+\infty} x^2 f(x)\,dx - \mu^2$; (1.14)

– для дискретного випадку – $\sigma^2 = \sum\limits_{i=0}^{\infty} x_i^2 p_i - \mu^2$. (1.15)

Якщо середнє значення сукупності відоме, то:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2. \qquad (1.16)$$

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

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{\sum\limits_{i=1}^{n} x_i^2 - n\left(\sum\limits_{i=1}^{n} x_i\right)^2}{n(n-1)}. \qquad (1.17)$$

Похибку (стандартне відхилення) дисперсії визначають за формулою: $s_{\sigma^2} = \sigma^2\sqrt{\dfrac{2}{n}}$.

Основною перевагою дисперсії як характеристики вибірки є те, що дисперсія суми статистично незалежних вибірок є сумою їх дисперсій:

$$\sigma_\Sigma^2 = \sum_{i=1}^{n}\sigma_i^2$$

незалежно від законів розподілу складових вибірок.

У більш загальному випадку дисперсія суми двох вибірок, як відомо з курсу математичної статистики, дорівнює:

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

де ρ – коефіцієнт кореляції.

Деякі розподіли не мають скінченної дисперсії. Для її існування необхідно, щоб при $x \to \infty$ крива щільності розподілу спадала швидше за $1/x^3$. В іншому випадку дисперсія дорівнює нескінченності. Порівнюючи цю умову з умовою існування математичного сподівання можна зробити висновок, що для окремих розподілів математичне сподівання існує, а скінченна дисперсія – ні. У цьому разі неможливо правильно визначити значення математичного сподівання за вибіркою скінченного обсягу.

Середнє квадратичне (стандартне) відхилення (похибка):

$$s = \sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}, \tag{1.18}$$

або, якщо середнє значення відомо з незалежних оцінок:

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu\right)^2}. \tag{1.19}$$

Середнє квадратичне відхилення 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:

$$\sigma^2 = \frac{1}{n-1}\left[\sum_{i=1}^{k}\nu_i b_i^2 - \frac{1}{n}\left(\sum_{i=1}^{k}\nu_i b_i\right)^2\right], \tag{1.20}$$

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 вводять **поправку Шеппарда** і визначають уточнене значення за формулою:

$$s'^2 = \sigma^2 - \frac{d^2}{12}, \tag{1.21}$$

де d – інтервал між групами, який за рівних відстаней між групами збігається з величиною класового інтервалу. Аналогічні поправки вводять і для деяких інших вибіркових характеристик при їх оцінюванні за згрупованими даними.

Отримані нами результати [12] дають підстави стверджувати, що формула (1.20) не завжди дає завищену оцінку дисперсії. Середнє значення відношення дисперсій, що розраховують за формулами (1.17) та (1.20), збільшується із зростанням кількості елементів вибірки і практично не залежить від їх середнього значення і стандартного відхилення. Імовірність того, що це значення буде менше за одиницю, збільшується із зменшен-

ням обсягу вибірки і є достатньо високою для вибірок обсягом менше ніж 1000 елементів. Така поведінка пов'язана зі зменшенням середнього значення цього відношення, а також зростанням його дисперсії при зменшенні обсягу досліджуваної вибірки.

Довірчий інтервал для дисперсії у випадку двобічної гіпотези:

$$\sigma^2 \in \left[\frac{ns^2}{\chi^2_{n-1,\alpha/2}}; \frac{ns^2}{\chi^2_{n-1,1-\alpha/2}} \right], \tag{1.22}$$

де χ^2_{n-1} – значення оберненої функції χ^2-розподілу; α – рівень значущості.

Для якісних ознак стандартне відхилення можна обчислити за формулою:

$$\sigma = \sqrt{p(1-p)}, \tag{1.23}$$

де p – частка відповідної ознаки.

Дисперсія і стандартне відхилення є розмірними величинами, що не зручно для порівняння варіабельності величин, які мають різну розмірність. Тому додатково використовують **коефіцієнт варіації** вибірки $C_v = s/\bar{x}$. Коефіцієнт варіації часто розглядають як міру однорідності вибірок. Для близьких до нормального розподілів вважають, що при $C_v \le 1/3$ вибірка є однорідною, а при $C_v \ge 1/2$ – неоднорідною.

Вихідні дані іноді доцільно подавати у **стандартизованому вигляді**: $z = \frac{x - \bar{x}}{s}$. Це дає змогу привести їх до безрозмірного вигляду і одного масштабу, що часто дає можливість покращити роботу алгоритмів їх подальшої обробки.

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

$$d = \frac{1}{n} \sum_{i=1}^{n} \left| x_i - \bar{x} \right|. \tag{1.24}$$

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

$$g = \frac{1}{n(n-1)} \sum_{\substack{i,j=1 \\ i \ne j}}^{n} \left| x_i - x_j \right|. \tag{1.25}$$

Останні дві характеристики використовують досить рідко.

У деяких випадках застосовують граничні, або максимальні оцінки ширини розподілу. Для обмежених розподілів (рівномірного, трикутного, трапецієподібного, арксинусоїдального тощо) вони є теоретично обґрунтованими, але на практиці такі розподіли можна використовувати лише як ідеалізовані моделі реальних. Зазвичай граничні оцінки монотонно зростають із збільшенням обсягу вибірки.

Асиметрією (**вибірковим коефіцієнтом скісності**) називають міру відхилення симетричного розподілу стосовно максимальної ординати. Для будь-якого симетричного розподілу вона дорівнює нулю. Від'ємні значення відповідають розширенню лівої гілки щільності розподілу, а додатні – її правої гілки. Асиметрію розраховують як основний момент третього порядку, а її стандартне відхилення – за формулою $s_{As} = \sqrt{\frac{6}{n+3}}$. Її часто застосовують як критерій відхилення розподілу від нормальності.

Як кількісну оцінку ступеня відхилення емпіричної кривої розподілу від теоретичної також застосовують **показник** (**коефіцієнт**) **ексцесу**. Його також називають **вибірковим коефіцієнтом гостроверхості**, але насправді він характеризує не гостороверхість, а протяжність розподілу [44]. Нормальному розподілу відповідає нульове значення показника ексцесу. Від'ємні значення свідчать про більш пологу, а додатні – більш гостру вершину максимуму розподілу. Показник ексцесу та його стандартне відхилення визначають за формулами:

$$E = \varepsilon - 3;\ s_E = 2s_{As} = 2\sqrt{\frac{6}{n+3}}, \qquad (1.26)$$

де $\varepsilon = r4$ – основний момент четвертого порядку (ексцес). Поряд з ексцесом використовують контрексцес $\chi = 1/\sqrt{\varepsilon}$.

Відмінність асиметрії й показника ексцесу від нуля вважають істотними, якщо вони перевищують за абсолютною величиною свої стандартні відхилення більше, ніж у 1,5–2 рази.

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

$$P = \frac{s_{\bar{x}}}{\bar{x}} \times 100\%;\ s_P = P\sqrt{\frac{1}{2n} + \left(\frac{P}{100}\right)^2}\ . \qquad (1.27)$$

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

Моментами розподілу називають середні значення відхилень даних:

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

Порядок моменту дорівнює степеню k, до якого підносять відповідні відхилення. На практиці, як правило, обмежуються моментами перших чотирьох порядків, оскільки із збільшенням порядку моменту істотно зростає похибка його визначення за емпіричними даними.

За вибірковими даними моменти розраховують, використовуючи такі формули:

$$\mu_k = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^k}{n};\ m_k = \frac{\sum_{i=1}^{n}\left(x_i - C\right)^k}{n};\ b_k = \frac{\sum_{i=1}^{n} x_i^k}{n}. \qquad (1.28)$$

Початковий момент першого порядку b_1 є середнім арифметичним, а центральний момент другого порядку μ_2 – зміщеною оцінкою дисперсії.

Знання початкових моментів дає змогу побудувати **твірну функцію моментів** $M'(x)$, яка дає повну інформацію про розподіл досліджуваної ознаки. За визначенням:

$$M_x'(t) = \overline{e^{xt}}. \qquad (1.29)$$

Розкладаючи експоненту в ряд:

$$M_x' = \overline{1 + xt + \frac{(xt)^2}{2!} + \ldots + \frac{(xt)^n}{n!} + \ldots} = 1 + b_1 t + b_2 \frac{t^2}{2!} + \ldots + b_n \frac{t^n}{n!} + \ldots, \qquad (1.30)$$

бачимо, що b_n – це коефіцієнт при $\frac{t^n}{n!}$ у такому розкладі. Звідси також випливає, що

$$b_n = \frac{d^n M_x'(0)}{dt^n}. \qquad (1.31)$$

Аналогічно можна визначити й **твірну функцію для центральних моментів**:

$$M_x(t) = \overline{e^{(x-\bar{x})t}}. \qquad (1.32)$$

Між цими двома функціями існує очевидний зв’язок:

$$M_x'(t) = e^{\bar{x}t} M_x(t). \qquad (1.33)$$

Величини $r_k = \frac{\mu_k}{s^k}$ називають **основними моментами порядку *k***. Основний момент порядку 3 є коефіцієнтом асиметрії, а основний момент порядку 4 використовується для розрахунку показника ексцесу. 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.

1.3. Варіаційна статистика

Варіаційною статистикою називають обчислення числових та функціональних характеристик емпіричного розподілу.

Варіаційний ряд (**порядкова статистика**) – це ряд даних, впорядкований за незгасанням. Він може бути побудований з кількісних або порядкових вибірок. **Дискретний** та **інтервальний** варіаційні ряди є таблицями розподілу кількості даних за класами. У першому випадку ці кількості належать до певних значень ознак, які можуть бути нечисловими, а в другому – до інтервалів зміни ознаки (класових інтервалів).

Для практичної побудови інтервального варіаційного ряду необхідно здійснити попереднє **групування даних**. Для цього визначають найбільше ($x_{\max}$) та найменше ($x_{\min}$) значення ознаки у вибірці та її **розкид** $R = x_{\max} - x_{\min}$. Після цього задають кількість класів (груп, інтервалів) *k*.

Задачу вибору кількості інтервалів групування емпіричних даних для їх статистичної обробки можна сформулювати як задачу оптимальної фільтрації випадкових відхилень гістограми розподілу емпіричної вибірки від гладкої кривої щільності розподілу генеральної сукупності. Як правило, використовують інтервали рівної ширини або рівної ймовірності. Можна довести, що існує оптимальна для даної вибірки кількість інтервалів певного типу. Але її значення залежить від того, якою є мета групування (побудова гістограми, порівняння вибірок за певним критерієм тощо). Розроблено багато емпіричних та напівемпіричних формул, які у різних випадках застосовують для визначення оптимальної кількості інтервалів групування [44].

Суб'єктивним критерієм правильності обрання кількості класів є точність відображення характеру розподілу емпіричних частот досліджуваної сукупності. Для інтервалів рівної ширини часто застосовують емпіричні правила, згідно з якими *k* обирають у межах 10−20 або 9−15. При цьому для симетричних розподілів слід брати непарні значення *k*. Рекомендують, щоб найменший за кількістю елементів інтервал, містив принаймні 10 точок. Але допускається, щоб крайні інтервали містили не менше, ніж п'ять точок. Якщо кількість спостережень є меншою, їх доцільно об'єднувати. Це ускладнює подальшу обробку, але є необхідним при застосуванні деяких методів, зокрема критерію χ^2.

Часто використовують **правило Стержесса**:

$$k = 1,44 \ln n + 1, \tag{1.34}$$

де n – кількість елементів сукупності. Але це правило є евристичним і не має ні теоретичного обґрунтування, ні експериментального підтвердження.

І.У. Алексєєвою на основі мінімізації ентропійного коефіцієнта, що використовувався як критерій близькості гістограми до теоретичного розподілу, для оптимальної кількості інтервалів рівної ширини отримано формулу:

$$k = \frac{\varepsilon + 1,5}{6} n^{0,4} = An^{0,4}, \tag{1.35}$$

особливістю якої є те, що вона враховує не тільки обсяг вибірки, а й значення ексцесу розподілу. При цьому вплив ексцесу є суттєвішим за вплив обсягу вибірки. Істотною проблемою для використання формули (1.35) є те, що кількість інтервалів групування зазвичай необхідно вибирати до того, як будуть отримані оцінки моментів розподілу, у т. ч. величина ексцесу. Для найбільш поширених типів розподілу значення коефіцієнта A в (1.35) змінюється у межах від 0,55 (рівномірний) до 1,25 (розподіл Лапласа), що дає можливість істотно зменшити кількість варіантів вибору, враховуючи додаткову вимогу щодо непарності кількості інтервалів.

Для двомодальних розподілів рекомендується збільшити кількість інтервалів у 1,5–2 рази.

Згідно з Г. Манном, А. Вальдом і К. Уїльямсом оптимальна для застосування критерію χ^2 кількість інтервалів рівної імовірності при $n \to \infty$ може бути визначена за формулою:

$$k = \xi \sqrt[5]{2} \left(n / t \right)^{0,4}, \tag{1.36}$$

де ξ внаслідок пологості оптимуму можна варіювати у межах від 2 до 4;

t – безрозмірний квантиль нормального розподілу, що відповідає заданій імовірності $p = 1 - \alpha$; α – прийнятий рівень значущості.

Вплив типу розподілу на оптимальну кількість для інтервалів рівної імовірності є значно меншим завдяки тому, що ширина інтервалу в цьому випадку приблизно обернено пропорційна щільності ймовірності.

Після визначення кількості класів обирають величини інтервалів. Зазвичай їх беруть рівними. Тоді $d = R / k$, де d – величина інтервалу групування (ширина класового інтервалу, класовий інтервал, довжина інтервалу групування). Для нерівних інтервалів i-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 й меж класового інтервалу не мають сенсу.

Теоретичною функцією розподілу випадкової величини x називають функцію дійсного аргументу, що задається як $F(x)=P(X\le x)$. На рис. 1.1 подано приклади графіків теоретичних функцій розподілу.

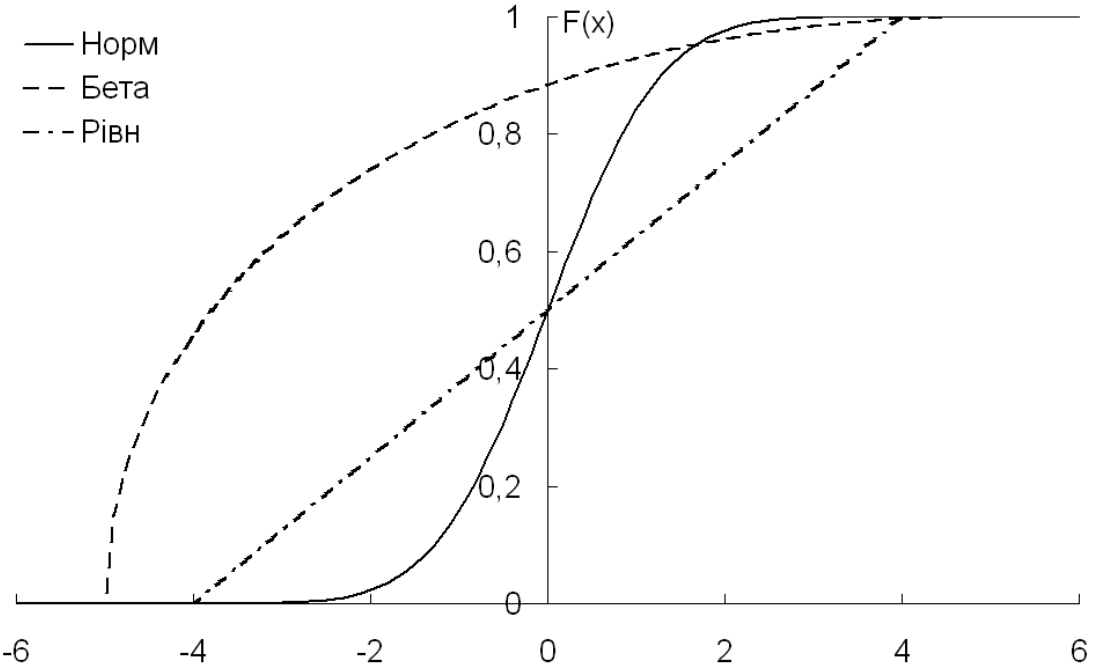


Рис. 1.1. Графіки теоретичних функцій розподілу для нормального, бета- та рівномірного розподілів

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

$$F_n(x)=\begin{cases}0, & x<x_0;\\ n_{x_k}/n, & x_k<x\le x_{k+1},\ 0\le k\le n;\\ 1, & x>x_n,\end{cases} \qquad (1.37)$$

де n_x – кількість елементів вибірки, що є меншими за x (**нагромаджені, або кумулятивні абсолютні частоти**); n – загальна кількість елементів вибірки. Вона є східчастою функцією, яка має стрибки у точках $x_0, x_1, ..., x_n$. Із збільшенням обсягу вибірки емпірична функція розподілу наближається до теоретичної.

Величини n_{x_k} називають **нагромадженими (кумулятивними) відносними частотами**, або **інтенсивностями**. Їх можна подавати у частках або у відсотках.

Теоретична й емпірична функції розподілу виявляють такі властивості:

– якщо $x_1<x_2$, то $F(x_1)\le F(x_2)$;

– $\lim\limits_{x\to-\infty}F(x)=0,\ \lim\limits_{x\to+\infty}F(x)=1$.

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

Функцією виживання називають таку функцію $S(x)$, значення якої є ймовірностями того, що випадкова величина набуде значень, більших, ніж x. Вона пов'язана з функцією розподілу співвідношенням:

$$S(x) = 1 - F(x). \tag{1.38}$$

Похідну від теоретичної функції розподілу $f(x) = \dfrac{dF(x)}{dx}$ називають **функцією щільності розподілу,** або просто **щільністю розподілу**. Іноді її також називають **диференціальною функцією розподілу**. На рис. 1.2 наведено приклади графіків функцій щільності розподілу.

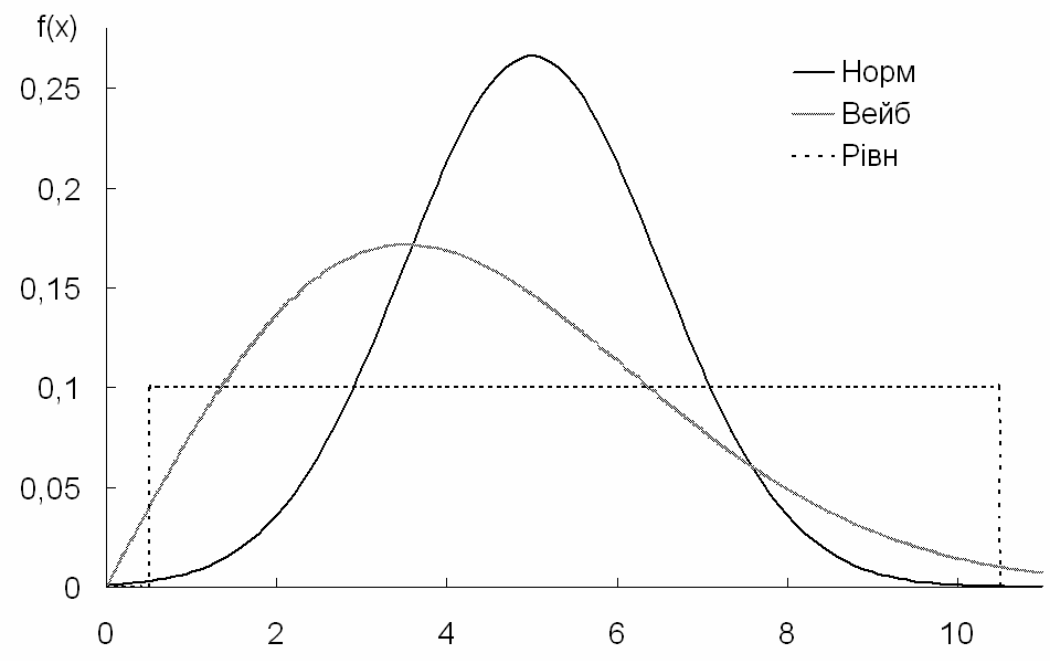


Рис. 1.2. Графіки функцій щільності розподілу для нормального, рівномірного розподілів та розподілу Вейбулла

Ця функція має такі властивості:

$$f(x) \geq 0;$$

$$\int_{-\infty}^{\infty} f(x)\,dx = 1;$$

$$\int_{-\infty}^{x^*} f(x)\,dx = F(x^*).$$

Функцію $x = G(\alpha) = G(F(x))$, значеннями якої є такі числа x, що випадкова величина не перевищує їх з імовірністю α, називають **оберненою функцією розподілу**, або **функцією квантилів**. Обернені функції розподілу широко використовують у статистичному аналізі даних для визначення критичних значень статистик, що відповідають заданому рівню значущості або заданому довірчому рівню.

Оберненою функцією виживання $Z(\alpha)$ називають функцією, значенням якої при аргументі α є число, що може бути перевищене випадковою величиною з імовірністю α. Вона пов'язана з оберненою функцією розподілу співвідношенням:

$$Z(\alpha) = G(1-\alpha). \tag{1.39}$$

Аналогічні функції можна ввести й для багатовимірних розподілів. Наприклад, **двовимірна функція розподілу**:

$$F(x^*, y^*) = P\{x \le x^*, y \le y^*\}. \tag{1.40}$$

Двовимірна функція щільності ймовірності:

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y}. \tag{1.41}$$

Вона має такі властивості:

$$f(x,y) \ge 0;$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dxdy = 1;$$

$$F(x^*, y^*) = \int_{-\infty}^{x^*}\int_{-\infty}^{y^*} f(x,y)\,dxdy.$$

Якщо двовимірна функція щільності розподілу є відомою, то **частинні щільності розподілу** за кожним з параметрів можна визначити за формулами:

$$f(x) = \int_{-\infty}^{\infty} f(x,y)\,dy; \;\; f(y) = \int_{-\infty}^{\infty} f(x,y)\,dx. \tag{1.42}$$

Якщо компоненти x та y є незалежними, то:

$$f(x,y) = f(x)f(y) \tag{1.43}$$

На рис. 1.3 наведено приклад графіка двовимірної функції щільності розподілу.

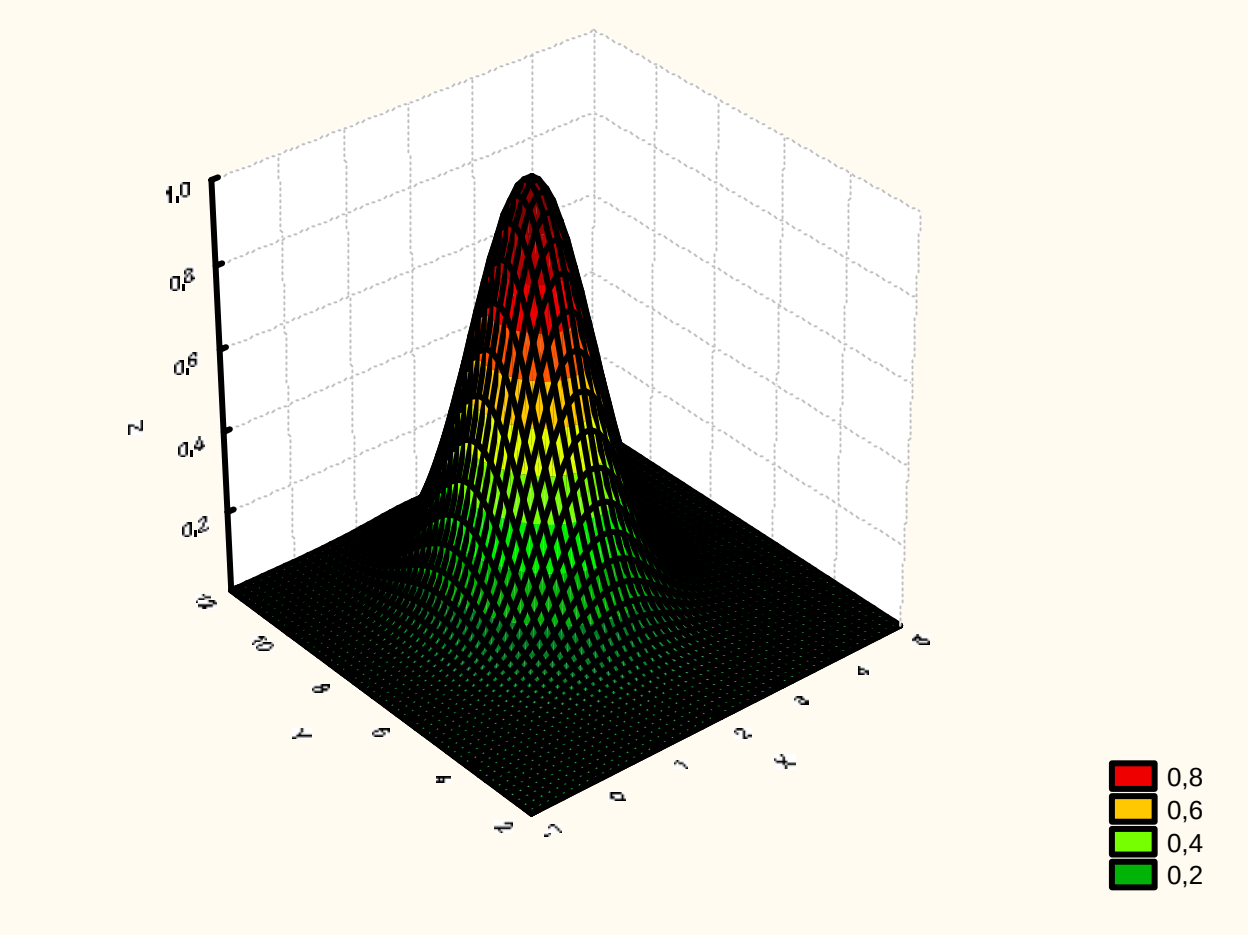


Рис. 1.3. Приклад графіка двовимірної функції щільності розподілу

У загальному багатовимірному випадку **функцією розподілу випадкового вектора *X*** називають величину:

$$F\left(\boldsymbol{X}\right)=P\left(x_1<x_1^*,...,x_k<x_k^*\right)=P\left(\boldsymbol{X}<\boldsymbol{X}^*\right). \qquad (1.44)$$

Вона має такі властивості:

– $0\le F\left(\boldsymbol{X}\right)\le 1$;

– $F\left(\boldsymbol{X}\right)=0$, якщо існує $x_j=-\infty$;

– $\mathrm{F}\left(\boldsymbol{X}\right)=1$, якщо $\forall x_j=+\infty \quad \left(j=1,...,k\right)$.

Корінь X_q рівняння

$$P\left(X\le X_q\right)=F\left(X_q\right)=q \qquad (1.45)$$

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**; порядку 10%, 20%, …, 90% – **децилями**, а 1%, 2%, …, 99% – **процентилями** (**персентилями**).

Величину $X_{0,75} - X_{0,25}$ називають **розмахом розподілу стосовно центрального значення**. Часто її подають у нормованому вигляді: $\frac{X_{0,75} - X_{0,25}}{X_{0,5}}$. Величину $\frac{X_{0,25} + X_{0,75}}{2}$ називають **центром згинів** й іноді використовують як оцінку центру для розподілів, що не мають математичного сподівання і дисперсії.

Інтервал d_α значень x між $X_{q/2}$ та $X_{1-q/2}$ $(0 < q < 1)$ називають **інтерквантильним проміжком з довірчою ймовірністю** $\alpha = 1 - q$. Наприклад, $d_{0,95} = X_{0,975} - X_{0,025}$ є інтерквантильним проміжком з довірчою ймовірністю 0,95. Величина $d_{0,9}$ має унікальну властивість, яка полягає в тому, що для широкого класу найчастіше використовуваних законів розподілу вона однозначно пов'язана з дисперсією співвідношенням $d_{0,9} \approx 3,2\sigma$. Тому за відсутності даних про вид розподілу рекомендують застосовувати саме цей показник [44]. Квантильні оцінки є основою для створення основних критеріїв перевірки статистичних гіпотез. Слід зазначити, що для негаусових вибірок вони є більш робастними мірами ширини розподілу, ніж середнє та стандартне відхилення.

Практичне визначення величини d_α полягає в тому, що із вибірки відкидають q найбільш віддалених від центра й, відповідно, найменш надійних значень. Достовірність квантильних оцінок різко підвищується із зменшенням довірчої ймовірності, а також при збільшенні обсягу вибірки. Можна показати [44], що значення довірчої імовірності не перевищує величини:

$$\alpha_{\max} \le \frac{n - 1 - 2n_{\text{âiäê}}}{n + 1}, \tag{1.46}$$

де n – обсяг вибірки; $n_{\text{відк}}$ – кількість відкинутих з кожного боку крайніх елементів вибірки. При цьому реальне значення довірчої ймовірності може бути значно нижчим, ніж гранична величина, що розраховується за формулою (1.46).

За допомогою квартилів розподілу можна отримати швидку оцінку коефіцієнта асиметрії:

$$As^* = \frac{X_{0,75} + X_{0,25} - 2X_{0,5}}{X_{0,75} - X_{0,25}}. \tag{1.47}$$

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

$$P\left(X \le r\right) = \sum_{i=0}^{r} p_i ;$$

$$P\left(r < X \le s\right) = \sum_{i=r+1}^{s} p_i ;$$

$$\sum_{i=0}^{\infty} p_i = 1.$$

Для неперервних випадкових величин аналогічні властивості можна записати таким чином:

$$P\left(X \le a\right) = F\left(x\right) = \int_{-\infty}^{a} f\left(x\right) dx ;$$

$$P\left(a < X \le b\right) = \int_{a}^{b} f\left(x\right) dx = F\left(b\right) - F\left(a\right).$$

Для двовимірного розподілу ймовірність потрапляння випадкових величин до певної області Ω:

$$P\left\{x, y \in \Omega\right\} = \iint_{\Omega} f\left(x, y\right) dxdy . \qquad (1.48)$$

При аналізі багатовимірних ознак слід враховувати, що їх окремі компоненти можуть бути пов’язані одна з одною. Більш докладно ці питання будуть розглядатися у розділі, присвяченому кореляційному аналізу.

Частотами (**абсолютними**, або **груповими**) розподілу ν_i називають кількості елементів вибірки, що потрапили до *i*-го класу. **Відносними частотами**, або **частостями,** називають величини $f_i = \nu_i / n$, де n – загальна кількість елементів вибірки. При великих n вони наближаються до ймовірностей реалізації відповідних значень параметрів (подій).

Відносна частота є незміщеною оцінкою частки p елементів генеральної сукупності, що мають певну ознаку. Дисперсія такої оцінки дорівнює:

$$D_{f_i} = \frac{p\left(1-p\right)}{n-1} \frac{N-n}{n}, \qquad (1.49)$$

де n – обсяг вибірки; N – обсяг генеральної сукупності.

Для неперервних розподілів графік абсолютних частот доцільно зображувати як стовпчикову діаграму, а для дискретних – як гістограму, точковий або лінійчатий графік.

На практиці для побудови емпіричних функцій розподілу застосовують два способи.

1. Задають класові інтервали і розподіляють дані вихідної вибірки за класами. Потім будують масив відносних частот, послідовне підсумовування елементів якого дає масив функції розподілу. Графічно він зображується як неспадний східчастий графік, по вісі абсцис якого відкладені середини класових інтервалів, а по вісі ординат – значення частот. Для використання цього способу необхідно, щоб вихідний обсяг даних був достатньо великим (не менше ніж 100 елементів). Інакше побудована функція може неправильно відображати характер розподілу. Аналогічний графік можна побудувати й для абсолютних частот. Його називають діаграмою накопичених частот. Приклади відповідних графіків для вибірки з 200 елементів, що підпорядковується стандартному нормальному розподілу, показано на рис. 1.4.

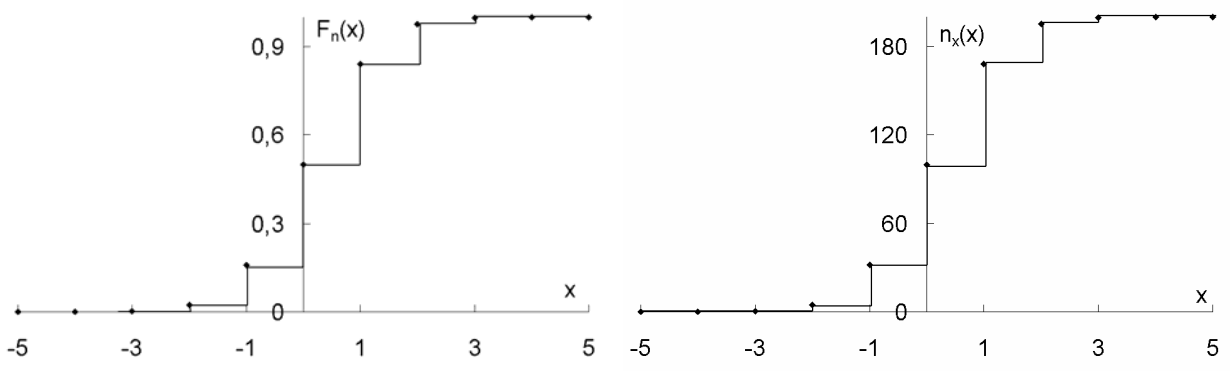


Рис. 1.4. Приклади емпіричної функції розподілу й діаграми накопичених частот

2. Вихідну вибірку впорядковують за зростанням. Потім будують графік, по вісі абсцис якого відкладають значення елементів вибірки, а по вісі ординат – відношення їх номерів до загальної кількості елементів вибірки. Для малих вибірок такий спосіб є єдиним, придатним для отримання функції розподілу, яку можна використовувати у подальших розрахунках. Але і для вибірок великого обсягу його застосування зазвичай є доцільнішим, оскільки із збільшенням обсягу вибірки побудована таким чином емпірична функція розподілу наближається до теоретичної.

На рис. 1.5 показано приклади побудованих цим способом емпіричних функцій розподілу для вибірок, що підпорядковуються стандартному нормальному розподілу й рівномірному розподілу на відрізку [−3, 3]. Якість наближення емпіричної функції розподілу до теоретичної ілюструє рис. 1.6, з якого видно, що вона значно покращується із збільшенням обсягу досліджуваної вибірки.

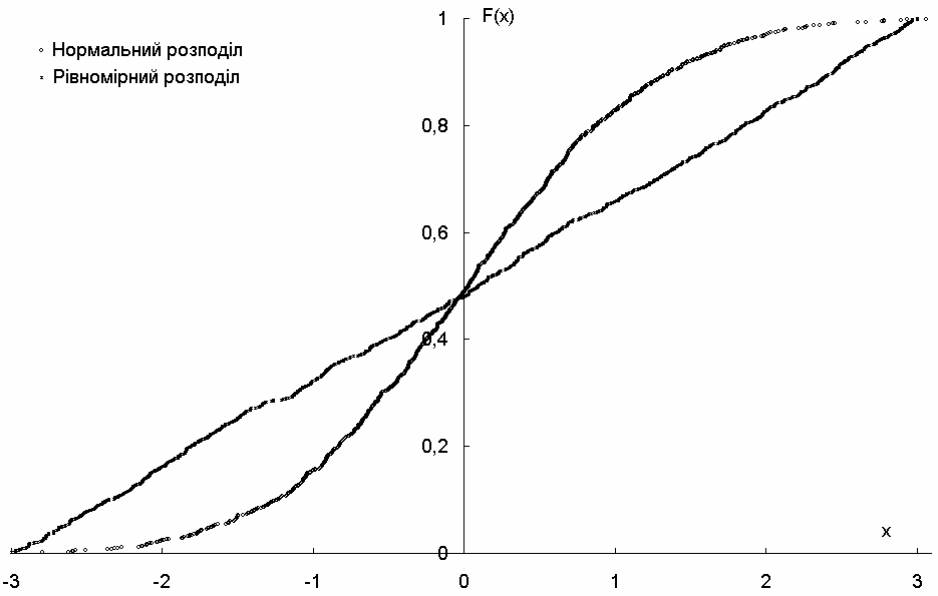


Рис. 1.5. Емпіричні функції розподілу для нормально та рівномірно розподілених вибірок

Аналогом функції щільності розподілу є **гістограма вибірки** (**гістограма відносних частот**). Але, на відміну від емпіричної функції розподілу, значення відносних частот не наближаються до значень щільності розподілу, оскільки при $n \to \infty,\ d \to 0$ відносні частоти $f_i \to f\left(x_i\right) \times d$.

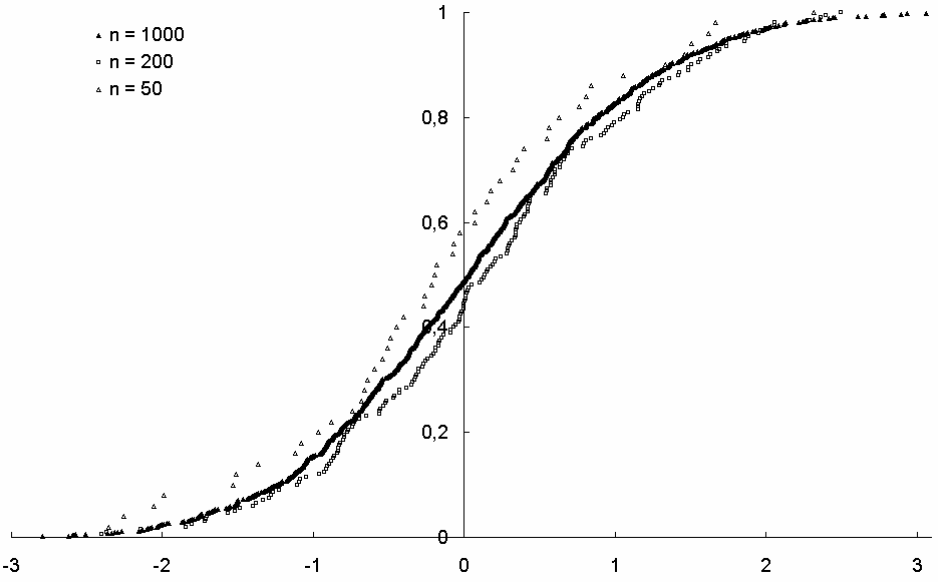


Рис. 1.6. Емпіричні функції розподілу нормально розподілених вибірок, що містять різну кількість елементів

При побудові гістограми наявний діапазон даних поділяють на інтервали рівної ширини. Кількість інтервалів k вибирають за одним із правил, наведених вище. Центральний стовпчик розташовують симетрично стосовно центру розподілу. Ширину стовпчика беруть рівною: $d = 2\Delta X_m / k$, де

ΔX_m – відстань від центру до найбільш віддаленої від нього точки. Центр розподілу часто визначають як центр розмаху $\frac{x_{\min} + x_{\max}}{2}$.

Для розрахунку ширини стовпчика часто використовують формулу $d = \frac{x_{\max} - x_{\min}}{k}$. У випадку необхідності отримане значення округляють у більший бік (якщо округлення здійснювати у менший бік, то крайні точки не потраплять до гістограми).

Площа під гістограмою абсолютних частот має дорівнювати обсягу досліджуваної вибірки, а під гістограмою відносних частот – одиниці.

Гістограму будують у вигляді стовпчикової діаграми, де ліва й права межа стовпчика відповідають межам відповідного інтервалу, а висота стовпчика дорівнює кількості емпіричних даних, що потрапили до цього інтервалу (гістограма абсолютних частот) або ймовірності потрапляння даних до нього (гістограма відносних частот). Приклади гістограм абсолютних й відносних частот для вибірки з 200 елементів, що підпорядковується стандартному нормальному розподілу, наведені на рис. 1.7.

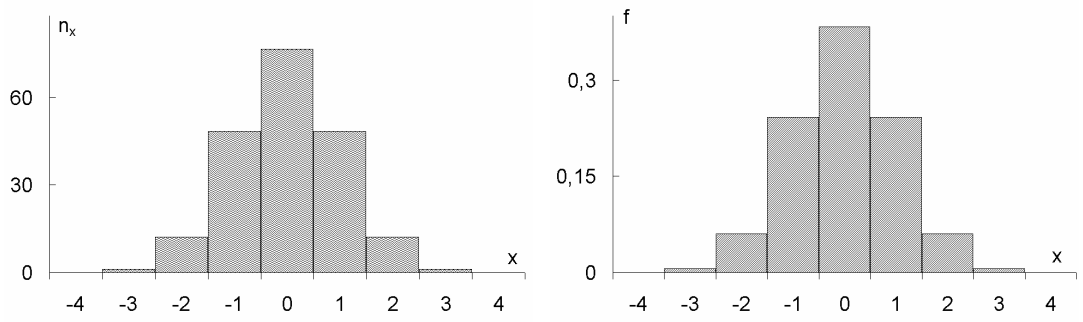


Рис. 1.7. Приклади гістограм абсолютних й відносних частот

Для графічного зображення даних використовують також полігони частот і частостей, а також накопичених частот і частостей. **Полігоном частот** називають ламану, що з'єднує точки (x_i, ν_i), а **полігоном частостей** – ламану, яка з'єднує точки (x_i, f_i). Ці графіки є аналогами графіку функції щільності розподілу досліджуваної ознаки. При побудові полігонів частот і частостей за згрупованими даними як значення x_i зазвичай беруть середини відповідних класових інтервалів. Накопиченими частотами й частостями називають величини $\eta_i = \sum_{j=1}^{i} \nu_j$ та $\varphi_i = \sum_{j=1}^{i} f_j$, відповідно. **Полігонами накопичених частот** і **накопичених частостей** називають ламані, що з'єднують, відповідно, точки (x_i, η_i) та (x_i, φ_i). Полігон накопичених час-

тостей називають також **кумулятивною кривою**. Ці графіки є аналогами графіку функції розподілу досліджуваної вибірки. Приклади полігонів частостей та накопичених частостей для вибірки з 200 елементів, що підпорядковуються стандартному нормальному розподілу, показано на рис. 1.8.

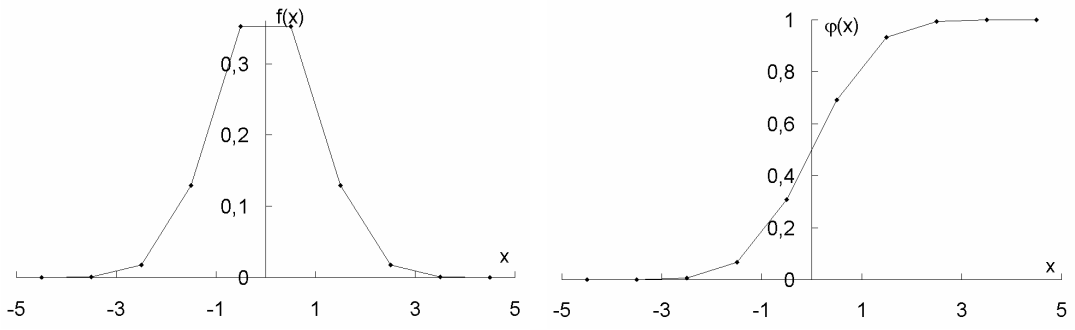


Рис. 1.8. Приклади полігонів частостей та накопичених частостей

Огівою (**багатокутником накопичених частот**) називають ламану, вершини якої мають абсциси, що збігаються з правими межами інтервалів групування, а ординати – зі значеннями накопичених частот для відповідних інтервалів.

Аналіз двовимірних даних зазвичай починають з побудови **поля розсіювання**. Для цього на площині *xOy* наносять емпіричні точки. Для зручності візуального аналізу іноді здійснюють попереднє центрування даних, переходячи до нових змінних $x' = x - \overline{x}$, $y' = y - \overline{y}$. На рис. 1.9 показано приклади полів розсіювання для незалежних і пов'язаних ознак.

Після цього складають таблиці двовимірного розподілу. Для цього здійснюють розбиття осей *Ox* й *Oy* на окремі інтервали довжиною Δx і Δy. Потім підраховують кількість точок n_{m_1,m_2}, що потрапили до кожного з прямокутників утвореної сітки, – абсолютні частоти, а також відповідні відносні частоти $\frac{n_{m_1,m_2}}{n}$. Ці таблиці використовують для побудови тривимірних гістограм та діаграм накопичених частот, що є емпіричними аналогами двовимірних щільності й функції розподілу.

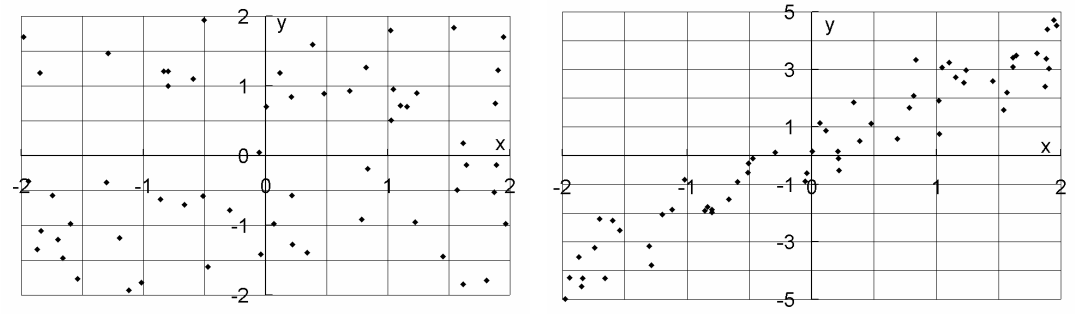


Рис. 1.9. Приклади полів розсіювання для незалежних і пов'язаних ознак

За допомогою таблиці двовимірного розподілу можна отримати вихідні дані для побудови гістограм кожної з її компонент. Для цього достатньо підсумувати значення таблиці за кожним стовпчиком або рядком.

При роботі з порядковими даними часто використовують рангові методи. **Рангом спостереження** називають номер, який відповідне спостереження отримає після впорядкування даних за певним правилом. Якщо кілька спостережень мають рівні значення ознаки, за якою здійснюють ранжирування, 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 підсумовування та ділення не визначені. Для кількісних даних ранжирування знижує вихідну шкалу до порядкової, що призводить до часткової втрати інформації.

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

$$T = \sum_{i=1}^{g} t_i \left(t_i^2 - 1 \right), \tag{1.50}$$

де g – кількість зв'язок (груп збігів); t_i – кількість збігів у i-й зв'язці.

Вибірку називають **репрезентативною**, якщо її параметри збігаються з параметрами генеральної сукупності в межах заданої допустимої похибки. При невідомому обсязі генеральної сукупності достатню кількість елементів вибірки з близьким до нормального законом розподілу можна оцінити за формулою: $N = \left(\frac{t_\infty \sigma}{\Delta} \right)^2$, де t_∞ – значення t-розподілу для нескінченної кількості степенів вільності; σ – вибіркове середнє квадратичне відхилення; Δ – задана допустима абсолютна похибка визначення середнього арифметичного значення. Значення σ і Δ вводять в іменованих одиницях, тобто вказують одиниці їх вимірювання, які для цих двох величин мають бути однаковими.

1.4. Приклад побудови описової статистики

Розглянемо такий приклад. Необхідно побудувати вибірку обсягом 300 елементів як суму двох вибірок, перша з яких підпорядковується нормальному розподілу з параметрами: m = 20, σ = 3; а друга – рівномірному розподілу на відрізку $[0; 5]$. Для отриманої вибірки визначити основні показники описової статистики, побудувати гістограму відносних частот, а також емпіричну функцію розподілу.

Для побудови вихідних вибірок використаємо генератор випадкових чисел пакета аналізу електронних таблиць MS Excel, обираючи у його меню (Сервіс/Аналіз даних/Генерація випадкових чисел) необхідні закони

розподілу й параметри вибірок. На рис. 1.10 показано вікно для задання параметрів розподілів, що генерують.

Для знаходження середнього арифметичного значення використовуємо формулу:

$$\bar{x} = \frac{1}{300}\sum_{i=1}^{300} x_i = 22,77671.$$

Середнє гармонічне розраховуємо за формулою:

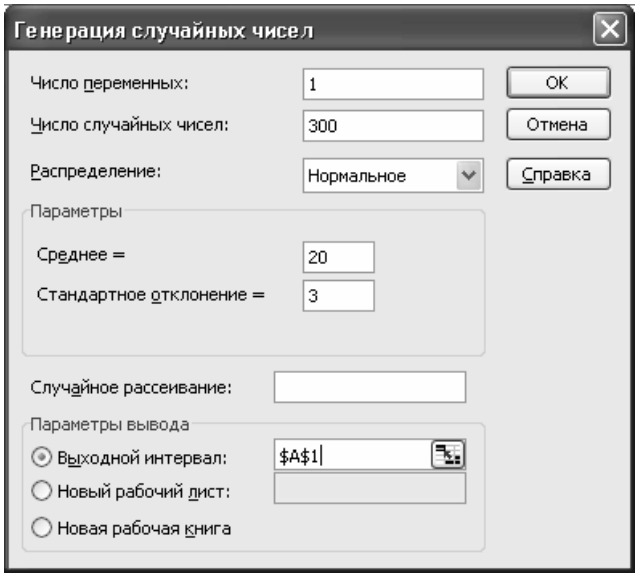


Рис. 1.10. Вікно задання параметрів для побудови випадкових послідовностей

$$h = \frac{300}{\sum_{i=1}^{300} \frac{1}{x_i}} = 21,79387.$$

Для визначення середнього геометричного використовуємо формулу:

$$g = \prod_{i=1}^{300} \sqrt[300]{x_i} = 22,29728.$$

Перевіряємо виконання співвідношення між середніми значеннями: $h \le g \le \bar{x}$. Бачимо, що у нашому випадку воно виконується.

Для визначення медіани впорядковуємо досліджувану вибірку за зростанням, використовуючи пункти меню електронних таблиць MS Excel “Дані/Сортування”, і обчислюємо її, як:

$$m = \frac{x_{150} + x_{151}}{2} = 23,0388.$$

В електронних таблицях MS Excel середні значення й медіану можна обчислити, використовуючи вбудовані функції СРЗНАЧ (), СРЗНАЧА (),

СРГАРМ (), СРГЕОМ (), МЕДИАНА (). Літера А наприкінці функції СРЗНАЧА () та в інших подібних випадках означає, що логічні й текстові значення не ігноруються, а беруться як рівні нулю.

Вибіркову дисперсію розраховуємо за формулою:

$$\sigma^2 = \frac{1}{300-1}\sum_{i=1}^{300}\left(x_i - \bar{x}\right)^2 = 20,82036\,.$$

Вибіркове стандартне відхилення:

$$s = \sigma = \sqrt{\frac{1}{300-1}\sum_{i=1}^{300}\left(x_i - \bar{x}\right)^2} = 4,5629\,.$$

Вибіркове середнє відхилення:

$$d = \frac{1}{300}\sum_{i=1}^{300}\left|x_i - \bar{x}\right| = 3,8021$$

В електронних таблицях MS Excel вибіркові дисперсію, стандартне відхилення й середнє відхилення можна знайти, використовуючи вбудовані функції ДИСП (), ДИСПА (), СТАНДОТКЛОН (), СТАНДОТКЛОН (А), СРОТКЛ ().

Коефіцієнти асиметрії та ексцесу розраховуємо, відповідно, за формулами:

$$As = \frac{\sum_{i=1}^{300}\left(x_i - \bar{x}\right)^3}{300s^3} = -0,09246\,;\; E = \frac{\sum_{i=1}^{300}\left(x_i - \bar{x}\right)^4}{300s^4} - 3 = 0,76648\,.$$

Слід зазначити, що ці оцінки є дещо зміщеними, оскільки отримані на основі формул, призначених для розрахунку моментів 3-го й 4-го порядків генеральної сукупності.

В електронних таблицях MS Excel вибіркові коефіцієнти асиметрії та ексцесу можна знайти, використовуючи вбудовані функції СКОС () та ЭКСЦЕСС ().

Крім того, в електронних таблицях MS Excel для побудови описової статистики можна використовувати засіб “Описова статистика” пакета аналізу. Для вибірки, що аналізується, результат його застосування наведено у табл. 1.3.

Таблиця 1.3

Описова статистика досліджуваної вибірки, отримана за допомогою пакета аналізу електронних таблиць MS Excel

Середнє	22,77671
Стандартна помилка	0,263441
Медіана	23,0388
Мода	19,92967

Продовження табл. 1.3

Стандартне відхилення	4,562934
Дисперсія вибірки	20,82036
Ексцес	– 0,74392
Асиметричність	– 0,09339
Інтервал	21,78896
Мінімум	11,7057
Максимум	33,49466
Сума	6833,012
Рахунок	300
Найбільший (1)	33,49466
Найменший (1)	11,7057
Рівень надійності (95,0%)	0,518433

У пакеті SPSS для визначення параметрів описової статистики можна використовувати діалогове вікно Analyze/Descriptive Statistics/Frequencies. У цьому вікні необхідно задати змінну, для якої будується описова статистика, а також параметри, які треба знайти. Приклад результату роботи цієї процедури для вибірки, що аналізується, наведено у табл. 1.4.

Таблиця 1.4

Приклад результатів, одержуваних за допомогою пакета SPSS

N	Valid	300
	Missing	0
Mean		22,7767
Std. Error of Mean		,26344
Median		23,0388
Mode		19,93
Std. Deviation		4,56293
Variance		20,82036
Skewness		–,093
Std. Error of Skewness		,141
Kurtosis		–,744
Std. Error of Kurtosis		,281
Range		21,79
Minimum		11,71
Maximum		33,49
Sum		6833,01
Percentiles	25	19,3176
	50	23,0388
	75	26,4383

Для побудови гістограми відносних частот поділяємо діапазон між мінімальним та максимальним значеннями елементів вибірки на інтервали рівної ширини. Кількість інтервалів обираємо згідно з правилом Стержеса:

$$k = 1,44 \ln 300 + 1 \approx 9 .$$

Далі, за формулою $d = \frac{x_{\max} - x_{\min}}{k}$ визначаємо ширину інтервалів й розраховуємо їх межі. Обчислюємо відносні частоти потрапляння даних до отриманих інтервалів і за допомогою майстра діаграм електронних таблиць MS Excel будуємо гістограму (рис. 1.11).

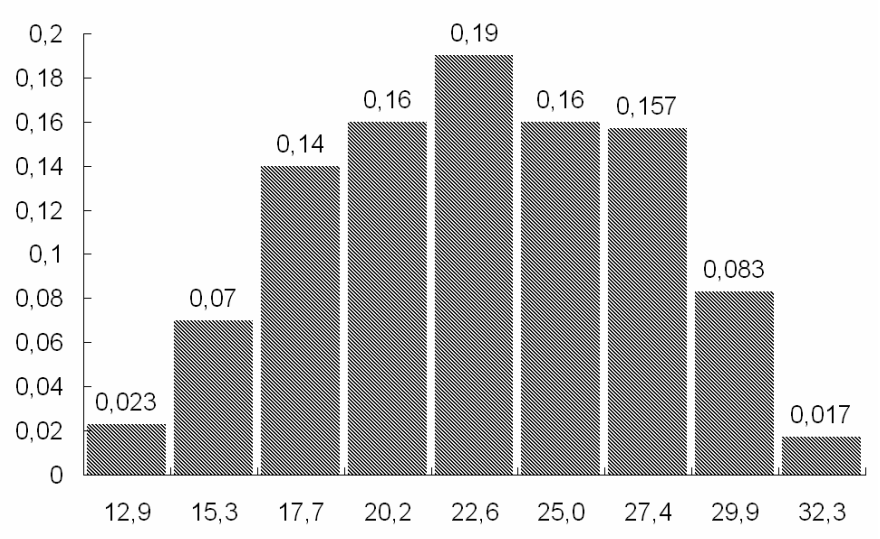


Рис. 1.11. Гістограма відносних частот досліджуваної вибірки

У пакеті SPSS для побудови діаграми відносних частот можна використовувати діалогове вікно Graphs/Interactive/Histogram (рис. 1.12). У ньому слід вказати змінну, для якої будують гістограму, (вісь *X*); які результати відкладати по вісі *Y* – абсолютні (count) чи відносні (percent) частоти. У вікні “Histogram” (рис. 1.13) можна додатково задати побудову нормальної кривої та кількість інтервалів. Приклад гістограми, побудованої за допомогою пакета SPSS, наведено на рис. 1.14.

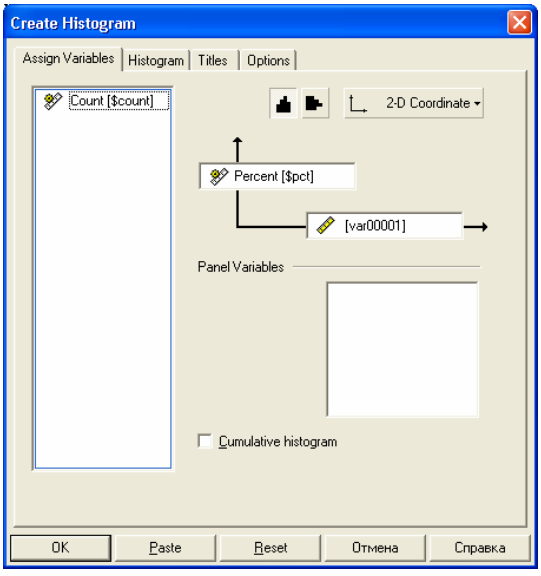


Рис. 1.12. Вікно створення гістограми пакету SPSS

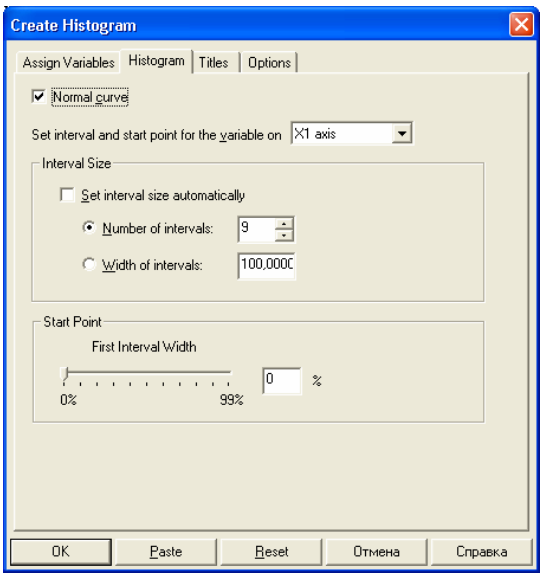


Рис. 1.13. Вікно задання параметрів гістограми

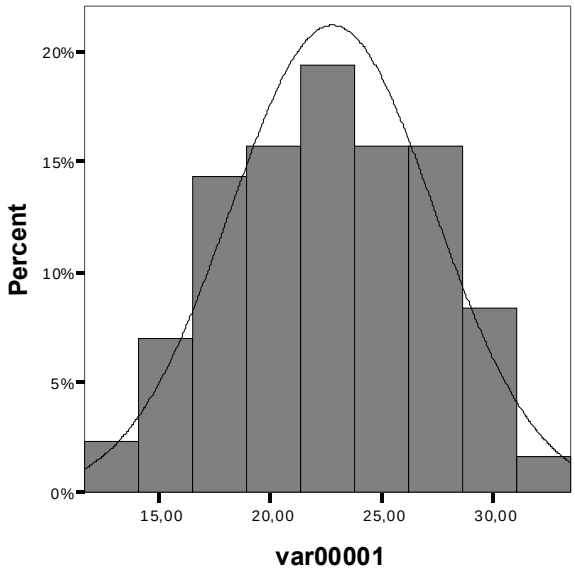


Рис. 1.14. Гістограма відносних частот досліджуваної вибірки, побудована у пакеті SPSS

Для побудови емпіричної функції розподілу за допомогою електронних таблиць MS Excel кожному елементу впорядкованої за зростанням вихідної вибірки ставимо у відповідність число, що дорівнює відношенню його порядкового номера до загальної кількості елементів вибірки. Потім будуємо точкову діаграму, де за віссю *Ox* відкладаємо значення елементів вибірки, а за віссю *Oy* – значення отриманого відношення, які їм відповідають. Емпірична функція розподілу для аналізованої вибірки показана на рис. 1.15.

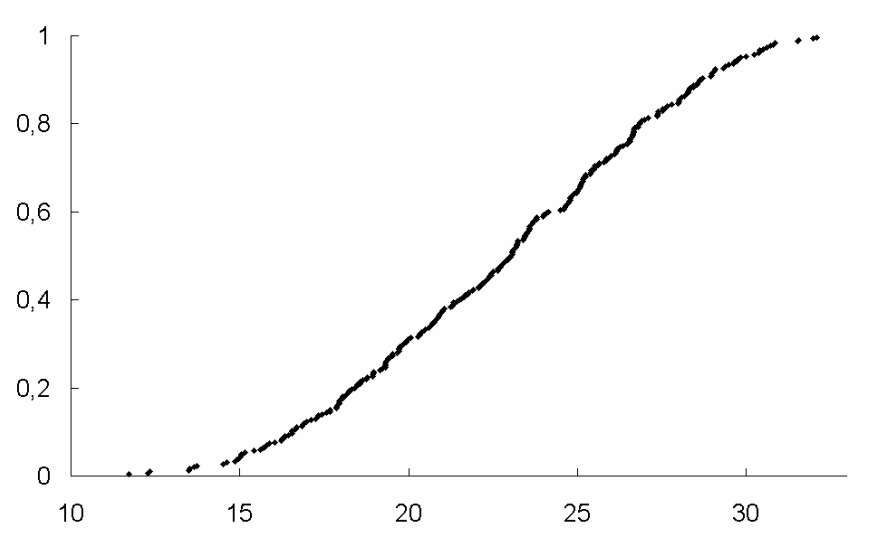


Рис. 1.15. Емпірична функція розподілу досліджуваної вибірки

Контрольні питання

1. Які операції допустимі для номінальних, порядкових та кількісних ознак? Навести приклади.
2. Які оцінки називають точковими? Наведіть приклади.
3. У яких випадках можна використовувати точкові оцінки параметрів?
4. Які оцінки називають інтервальними? Наведіть приклади.
5. У яких випадках застосовують інтервальні оцінки параметрів?
6. Що називають описовою статистикою?
7. Чому процедуру аналізу емпіричної вибірки необхідно починати з перевірки закону її розподілу на нормальність?
8. У чому полягає різниця між вибірковими характеристиками та параметрами генеральних сукупностей?
9. Які оцінки називають конзистентними (спроможними)?
10. Як перевірити конзистентність оцінки?
11. Які оцінки називають незміщеними?
12. За яким критерієм здійснюють порівняння ефективності оцінок?
13. Якими є основні завдання описової статистики?
14. Які параметри характеризують центр розподілу?
15. Що називають математичним сподіванням випадкової величини?
16. Чи завжди розподіл має математичне сподівання?
17. Що називають вибірковим середнім?
18. Від чого залежить стандартне відхилення вибіркового середнього?
19. У чому полягають переваги й недоліки вибіркового середнього як показника центра розподілу?
20. Які види середніх значень, крім вибіркового середнього, використовують як показники центра розподілу? Наведіть приклади їх застосування.

21. Що називають медіаною вибірки? У яких випадках її доцільно використовувати як показник центру розподілу?

22. У яких випадках існує медіана вибірки?

23. Що називають модою вибірки? У яких випадках її доцільно використовувати як показник центру розподілу?

24. Чи можливі випадки розподілів, що не мають моди або мають декілька мод? Наведіть приклади.

25. Які фактори визначають ефективність різних методів оцінювання центра розподілу?

26. Які методи оцінювання центра розподілу є найбільш стійкими до викидів?

27. Які параметри використовують для характеристики ширини розподілу? Наведіть приклади їх застосування.

28. Що називають дисперсією розподілу? Від яких параметрів залежить стандартне відхилення дисперсії?

29. Чи завжди існує дисперсія розподілу?

30. Що називають коефіцієнтом варіації вибірки? Наведіть приклади застосування цього показника.

31. Як перетворити вихідні дані до стандартизованого вигляду? З якою метою здійснюють таке перетворення?

32. Яку величину називають середнім відхиленням вибірки? Наведіть приклади її застосування.

33. Яку величину називають середньою різницею Джині? Наведіть приклади її застосування.

34. Що називають асиметрією вибірки? Які властивості розподілу характеризує цей показник? Від яких факторів залежить його стандартне відхилення?

35. Що називають коефіцієнтом ексцесу вибірки? Які властивості розподілу характеризує цей показник? Від яких факторів залежить його стандартне відхилення?

36. Що називають показником точності експерименту? Які властивості розподілу характеризує цей показник? Від яких факторів залежить його стандартне відхилення?

37. Якими є основні види моментів розподілів? Як моменти розподілу пов’язані з параметрами описової статистики?

38. Що називають варіаційним рядом? Для чого застосовують варіаційні ряди?

39. Якими є основні види варіаційних рядів? Для яких типів ознак їх застосовують?

40. З якою метою здійснюють попереднє групування даних?

41. Якою є загальна методика формування класів для групування даних?

42. Які основні типи інтервалів використовують при аналізі даних?

43. Як вибрати кількість інтервалів групування даних?

44. Що називають теоретичною функцією розподілу даних?

45. Що називають емпіричною функцією розподілу даних?

46. Які основні властивості мають емпірична й теоретична функції розподілу?

47. Як можна побудувати емпіричну функцію розподілу?

48. Що називають нагромадженими, або кумулятивними відносними частотами? Для чого їх використовують? Наведіть приклади.

49. Що називають функцією виживання? Для чого її використовують? Наведіть приклади.

50. Що називають функцією щільності розподілу? Які властивості мають одновимірна та двовимірна функції щільності розподілу?

51. Що називають функцією квантилів? Для чого її використовують? Наведіть приклади.

52. Що називають функцією оберненою функцією виживання? Для чого її використовують? Наведіть приклади.

53. Що називають багатовимірною функцією розподілу? Які властивості має ця функція?

54. Що називають вибірковим квантилем розподілу? Які основні типи квантилів використовують у практиці?

55. Що називають центром згинів? Для чого використовують цей показник? Наведіть приклади.

56. Що називають розмахом розподілу? Для чого використовують цей показник? Наведіть приклади.

57. Що називають інтерквантильними проміжками? Для чого використовують ці показники? Наведіть приклади.

58. Що називають частотами та відносними частотами розподілу? Для чого використовують ці показники? Наведіть приклади.

59. Як будують графіки абсолютних частот для дискретних та неперервних розподілів?

60. Що називають гістограмою вибірки? Для чого використовують гістограми? Наведіть приклади.

61. Як будують гістограму вибірки?

62. Що називають полігонами частот і частостей? Для чого використовують полігони? Наведіть приклади.

63. Що називають огівою? Для чого використовують огіви? Наведіть приклади.

64. Що називають полем розсіювання? Для чого використовують поля розсіювання? Наведіть приклади.

65. Що називають рангом спостереження? Для чого використовують ранги спостереження? Наведіть приклади.

66. Які вибірки називають репрезентативними?

2. ПЕРЕВІРКА СТАТИСТИЧНИХ ГІПОТЕЗ

2.1. Основні поняття

Існує велика кількість різноманітних методів перевірки статистичних гіпотез. При виборі методу для вирішення певного конкретного завдання необхідно виходити з відповідей на такі питання:

- якою є мета перевірки гіпотези;
- у яких шкалах виміряні аналізовані дані;
- чи є аналізовані вибірки незалежними або спряженими;
- скільки вибірок необхідно порівняти.

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

Гіпотезу, що перевіряють, називають **нульовою гіпотезою** (H_0). Прикладами нульових гіпотез можуть бути такі твердження: "Середні значення двох вибірок суттєво не відрізняються одне від одного"; "Дисперсія першої вибірки суттєво перевищує дисперсію другої"; "Розподіл вибірки відповідає нормальному закону з певними параметрами" тощо. Гіпотезу, що суперечить нульовій, називають **конкуруючою**, або **альтернативною гіпотезою** (H_1). Для вказаних вище нульових гіпотез конкуруючими можуть бути такі твердження: "Середні значення двох вибірок суттєво розрізняються одне від одного"; "Дисперсія першої вибірки не перевищує істотно дисперсію другої"; "Розподіл вибірки не відповідає нормальному закону із вказаними параметрами". Для однієї нульової гіпотези у загальному випадку можна сформулювати багато різних альтернативних гіпотез.

Розрізняють прості та складні гіпотези. **Простою** називають гіпотезу, що містить тільки одне твердження. **Складні гіпотези** складаються з декількох простих (при цьому кількість простих гіпотез може бути нескінченно великою).

Зазвичай при перевірці нульової гіпотези використовують певні модельні розподіли, що приблизно відповідають розподілу досліджуваного параметра. Їх називають **статистичними критеріями**. На практиці як критерії найчастіше використовують нормальний розподіл, χ^2-розподіл, розподіли Стьюдента і Фішера. **Значенням критерію, що спостерігається**, називають його величину, яку розраховують за досліджуваними вибірками.

Для перевірки гіпотези весь вибірковий простір поділяють на дві області, що не перетинаються: критичну (w) та область прийняття ($W - w$). **Критичною областю** називають сукупність значень критерію, за яких нульову гіпотезу слід відхилити. **Областю прийняття гіпотези** (**областю допустимих значень**) називають сукупність значень критерію, за яких нульову гіпо-

тезу приймають. Перевірка гіпотези передбачає розрахунок значення критерію і перевірку його потрапляння до області прийняття гіпотези.

Вирізняють **двобічні** й **однобічні** (**лівобічні**, **правобічні**) критичні області (рис. 2.1, 2.2). Їх використання залежить від вибору конкуруючої гіпотези.

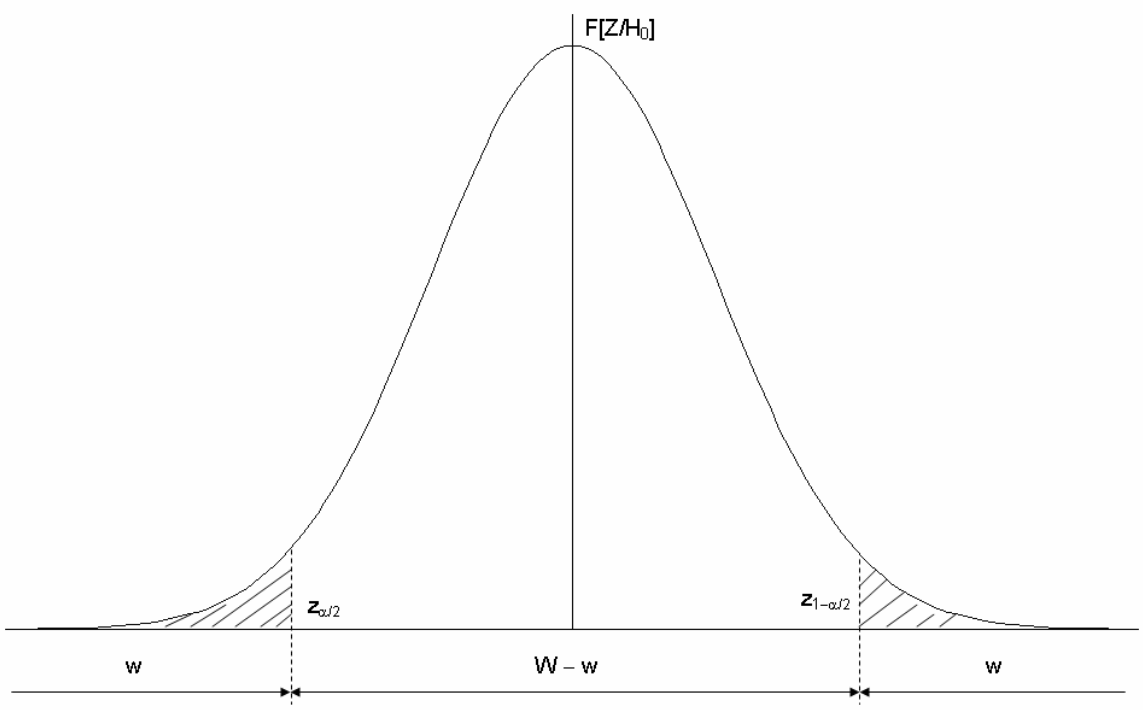


Рис. 2.1. Приклад двобічної критичної області

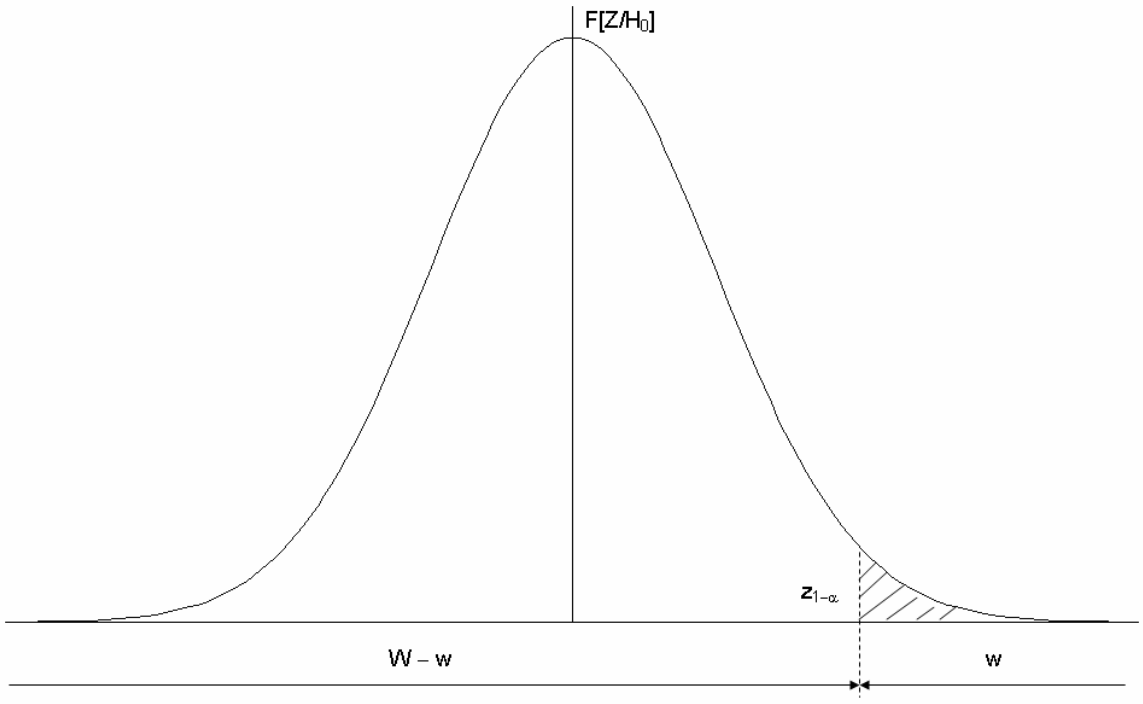


Рис. 2.2. Приклад правобічної критичної області

Якщо розподіл імовірності спостережень, що відповідає нульовій гіпотезі H_0, є відомим, то критичну область визначають так, щоб при виконанні H_0 імовірність її відхилення була рівною заздалегідь заданій малій величині (**рівню значущості**) α.

$$P\left(x \in w \,|\, \mathrm{H}_0\right) = \alpha, \tag{2.1}$$

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

Критерії, що базуються на використанні заздалегідь заданого рівня значущості, називають **критеріями значущості**. Рівень значущості визначає розмір критичної області: що більшим є рівень значущості, то ширшою буде критична область.

Розглядають два типи помилок, що можуть виникати при перевірці статистичних гіпотез:

– **помилкою першого роду** є відхилення правильної нульової гіпотези, рівень значущості α є ймовірністю такої помилки;

– **помилкою другого роду** є прийняття помилкової нульової гіпотези.

У деяких застосуваннях помилки першого та другого роду називають, відповідно, **ризиком виробника** та **ризиком споживача**.

Зменшення ймовірності помилки першого роду водночас призводить до підвищення ймовірності помилки другого роду β. З огляду на це додатково вводять поняття **потужності критерію** $1 - \beta$, яка є імовірністю відхилення помилкової нульової гіпотези, тобто ймовірністю потрапляння критерію до критичної області за умови, що правильною є конкуруюча гіпотеза:

$$P\left\{x \in w \,|\, \mathrm{H}_1\right\} = 1 - \beta. \tag{2.2}$$

Потужність критерію можна підвищити, збільшуючи обсяг вибірки. При визначенні критичної області її зазвичай будують так, щоб максимізувати потужність обраного критерію. За наявності декількох критеріїв, що можуть використовуватися для перевірки досліджуваної гіпотези, рекомендується обирати більш потужні з них, якщо їх застосування є обґрунтованим.

Важливим завданням є визначення обсягу вибірки, який дає змогу гарантувати певне значення похибки першого роду за умови, що похибка другого роду не перевищує заданого значення. Для цього необхідно розв’язати таку систему:

$$\begin{cases} P\left[Z \in w / \mathrm{H}_0\right] = \alpha; \\ P\left[Z \in W / \mathrm{H}_1\right] = \beta. \end{cases} \tag{2.3}$$

Її аналітичне розв’язання можливо лише у найпростіших випадках. Але, в багатьох випадках істотне спрощення можна отримати, якщо замінити ймовірності α й β значеннями меж відповідних критичних інтервалів.

Загальна методика отримання висновків при перевірці гіпотез передбачає, що на першому етапі необхідно задати рівень значущості. Найчастіше його беруть рівним 0,01; 0,05 або 0,1. Обираючи рівень значущості, слід пам'ятати, що його зменшення знижує ймовірність помилки першого роду, але збільшує ймовірність помилки другого роду. Тому, виходячи з конкретних умов, потрібно знайти певний компроміс між ймовірностями припустити помилки різного типу.

На другому етапі за даними вибірки розраховують значення критерію та порівнюють його з обчисленими для заданого рівня значущості межами критичної області. Якщо розраховане значення критерію потрапляє до них, то нульову гіпотезу відхиляють. В іншому випадку вважають, що немає підстав для відхилення нульової гіпотези і або приймають її на заданому рівні значущості, або здійснюють додаткову перевірку. Для визначення меж критичної області застосовують спеціальні таблиці або розраховують їх на основі відомих законів розподілу використовуваних критеріїв.

Можливості сучасної комп'ютерної техніки та наявного програмного забезпечення дають змогу отримувати висновки іншим шляхом. Якщо за наявними емпіричними даними розрахувати значення критерію, то на наступному етапі можна визначити, для якого рівня значущості це значення буде критичним. Ураховуючи, що рівень значущості є ймовірністю відхилення правильної нульової гіпотези, ми можемо за його значенням зробити висновок про ймовірність правильності або помилковості нульової гіпотези. Залежно від того, задовольняє нас отримана ймовірність помилки чи ні, нульову гіпотезу приймають або відхиляють.

При перевірці гіпотез доцільно застосовувати різні методи, призначені для вирішення одних й тих самих завдань та однакових типів даних. Причинами розбіжності отримуваних при цьому результатів зазвичай є:

- помилки при введенні даних;
- непридатність окремих методик для типу даних, що розглядають;
- алгоритмічні помилки у програмах, що використовують 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 в 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 вдвічі більшим, ніж для відповідних однобічних. При використанні однобічних критеріїв рекомендується спочатку 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);

– випускники різних шкіл (при аналізі результатів зовнішнього незалежного оцінювання).

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

– дані опитування громадської думки до й після певної суспільно значущої події;

– дві партії однотипної продукції, виготовлені одними й тими самими працівниками на одному й тому самому обладнанні до й після внесення певних змін до технології;

– одна й та сама партія виробів до і після певної технологічної обробки.

2.2. Параметричні тести

Критерії й тести, що застосовують для порівняння вибірок, поділяють на дві групи: параметричні й непараметричні. Особливістю параметричних критеріїв є припущення, що розподіл ознаки в генеральній сукупності підпорядковується певному відомому закону. Ця відповідність має бути доведена до застосування будь-якого з параметричних тестів. Переважна більшість параметричних тестів розроблена для нормально розподілених даних. Але для деяких типів гіпотез існують параметричні тести, призначені для вибірок, що підпорядковуються іншим законам розподілу.

Як правило, параметричні критерії є потужнішими за непараметричні. Застосування непараметричних критеріїв у випадках, коли можна використовувати параметричні, призводить до збільшення ймовірності прийняття помилкової нульової гіпотези, тобто помилки другого роду.

Якщо досліджувані вибірки підпорядковуються нормальному закону розподілу з відомими дисперсіями, то як критерій рівності їх середніх значень можна використовувати величину:

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \qquad (2.4)$$

де n_1, n_2 – кількість елементів у вибірках. Нульова гіпотеза полягає у рівності середніх.

Z-критерій є випадковою величиною, що підпорядковується стандартному нормальному розподілу. Якщо конкуруючою гіпотезою є $\bar{x}_1 \neq \bar{x}_2$, то праву критичну точку можна визначити з умови:

$$P\left(0 < Z < z_{\text{ïðàâ.êð.}}\right) = \Phi\left(z\right) = \left(1-\alpha\right)/2, \qquad (2.5)$$

де $\Phi(z)$ є функцією Лапласа (інтегралом ймовірностей), що пов'язана з функцією стандартного нормального розподілу $F(z)$ співвідношенням $\hat{O}\left(z\right) = F\left(z\right) - 1/2$. Ліва й права критичні точки для одного й того самого рівня значущості пов'язані умовою $z_{\text{ë³â.êð}} = -z_{\text{ïðàâ.êð}}$.

Нехай, наприклад, при вимірюванні певного параметра у двох серіях однакових виробів, виготовлених на різних установках, отримали такі значення: $\bar{x}_1 = 184$, $\bar{x}_2 = 181$, $\sigma_1^2 = 121$, $\sigma_2^2 = 107$, $n_1 = n_2 = 40$. Як нульову гіпотезу візьмемо твердження, що немає істотної різниці між параметрами виробів, виготовлених на різних установках. Тоді:

$$Z = \frac{184 - 181}{\sqrt{\frac{121}{40} + \frac{107}{40}}} \approx 1{,}26.$$

Беручи рівень значущості рівним 0,05 (5%), отримаємо для визначення критичної точки умову:

$$\Phi\left(z_{\text{ïðàâ.êð}}\right) = \left(1-\alpha\right)/2 = 0{,}475.$$

Для визначення величини $z_{\text{ïðàâ.êð}}$ можна використати функцію НОРМСТОБР електронних таблиць MS Excel, беручи як значення аргументу величину (0,5 + 0,475). Після цього одержимо: $z_{\text{ïðàâ.êð}}$ = 1,96. Оскільки $Z < z_{\text{ïðàâ.êð}}$, ми можемо прийняти нульову гіпотезу на рівні значущості 0,05.

Використовуючи вбудовані функції електронних таблиць MS Excel, ми можемо отримати більш точну оцінку ймовірності відхилення правильності нульової гіпотези. Величина $z_{\text{ïðàâ.êð}}$ = 1,26 для двобічної гіпотези відповідає рівню значущості $\alpha \approx 0{,}209$. Тобто, якщо ми не приймаємо нульову гіпотезу, то ймовірність помилки першого роду становить приблизно 21%. Таку імовірність у більшості випадків вважають занадто високою. Це пов'язано з тим, що зазвичай висновок формулюють як наявність чи відсутність достатніх підстав для відхилення (а не для прийняття) нульової гіпотези. При необхідності більш чіткого обґрунтування її прийняття виконують додаткові перевірки.

Якщо конкуруючою гіпотезою є: $\bar{x}_1 > \bar{x}_2$, то:

$$\Phi\left(z_{\text{êð}}\right) = \left(1-2\alpha\right)/2, \qquad (2.6)$$

і нульову гіпотезу приймають, якщо $Z < z_{\text{êð}}$.

Якщо конкуруючою гіпотезою є: $\bar{x}_1 < \bar{x}_2$, то критичну точку z' визначають з умови (2.6), враховуючи, що $z'_{кр} = -z_{кр}$. Нульову гіпотезу приймають, якщо $Z > -z_{кр}$.

Z-критерій можна застосовувати також для порівняння середніх значень довільно розподілених незалежних вибірок великого обсягу ($n_{1,2} \geq 30$), враховуючи, що в цьому разі вибіркові середні мають приблизно нормальний розподіл, а вибіркові дисперсії є достатньо точними оцінками генеральних дисперсій.

Для порівняння середніх значень вибірок застосовують ***t*-критерій Стьюдента**. Його запропоновано американським статистиком Уїльямом Госсетом в 1908 р. за результатами дослідження проблеми скорочення кількості проб, які потрібно взяти при контролі за якістю продукції пивоварного заводу за умови забезпечення виконання вимог стандартів.

Розглядають дві незалежні нормальні вибірки з генеральних сукупностей, що мають рівні або нерівні, але відомі чи рівні невідомі дисперсії.

Значення критерію Стьюдента розраховують за формулою:

$$t = \frac{\left|\bar{x}_A - \bar{x}_B\right|}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}, \qquad (2.7)$$

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 $m(n-1)$.

У випадку, коли обсяги вибірок є малими або істотно розрізняються, а їх дисперсії є рівними, останні замінюють вибірковим середнім квадратичним відхиленням, яке розраховують за формулою:

$$s^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}; \qquad (2.8)$$

якщо стандартні відхилення вибірок оцінюють за самими вибірками, або:

$$s^2 = \frac{s_1^2 n_1 + s_2^2 n_2}{n_1 + n_2 - 2}, \qquad (2.9)$$

якщо їх оцінюють незалежно. Формула для визначення розрахункового значення критерію у цьому разі набуває вигляду:

$$t = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}. \qquad (2.10)$$

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

При застосуванні критерію Стьюдента у вигляді (2.10) необхідно спочатку перевірити гіпотезу про рівність дисперсій.

Критичні точки симетричні стосовно нуля. Нульову гіпотезу відхиляють: якщо $|t| < t_{êð}(\alpha/2, k)$ при конкуруючій гіпотезі $\bar{x}_1 \neq \bar{x}_2$; якщо $t > t_{ïðàâ.êð}(\alpha, k)$ при конкуруючій гіпотезі $\bar{x}_1 > \bar{x}_2$; якщо $t < -t_{ïðàâ.êð}(\alpha, k)$ при конкуруючій гіпотезі $\bar{x}_1 < \bar{x}_2$.

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 щодо рівності дисперсій при цьому не висувається. Нульова гіпотеза полягає у відсутності різниці між середніми. Значення критерію розраховують за формулою:

$$t = \frac{\sum_{i=1}^{n} \delta_i}{\sqrt{\frac{n\sum_{i=1}^{n} \delta_i^2 - \left(\sum_{i=1}^{n} \delta_i\right)^2}{n-1}}}, \qquad (2.11)$$

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 дисперсії або їх відношення невідомі й припущення про рівність дисперсій є необґрунтованим, то виникає так звана **проблема Беренса – Фішера**, що полягає у перевірці нульової гіпотези про рівність вибіркових середніх за таких умов. Одним з підходів до її вирішення є застосування **критерію Уелча** (**Крамера – Уелча**), запропонованого Б. Уелчем в 1947 р. Його значення розраховують за формулою:

$$d = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \qquad (2.12)$$

де s_1^2, s_2^2 – розраховані за вибірками оцінки дисперсії.

Статистика цього критерію є приблизно такою самою, як для розподілу Стьюдента з кількістю степенів вільності:

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}. \qquad (2.13)$$

Порівняння з (2.7) вказує, що основною відмінністю критерію Уелча з погляду прикладного аналізу є зміна кількості степенів вільності.

***F*-критерій Фішера** запропоновано британським біологом і статистиком Рональдом Фішером в 1920 р. Його використовують для порівняння дисперсій двох вибірок. Його значення розраховують за формулою:

$$F = s_1^2 / s_2^2, \qquad (2.14)$$

де s_1^2, s_2^2 – значення оцінок більшої та меншої дисперсій, відповідно. Кількості степенів вільності для пошуку критичного значення обирають рівними $n_1 - 1$ та $n_2 - 1$. Гіпотезу про рівність дисперсій порівнюваних сукупностей відхиляють, якщо обчислене значення перевищує табличне при заданому довірчому рівні. При цьому, якщо конкуруючою є однобічна гіпотеза $s_1^2 > s_2^2$, то як критичну точку беруть значення оберненого розподілу Фішера, що відповідає рівню значущості α при заданій кількості степенів вільності. Якщо ж конкуруючою є двобічна гіпотеза $s_1^2 \neq s_2^2$, то критичною точкою буде значення оберненого розподілу Фішера, що відповідає рівню значущості α / 2.

Якщо перевіряють гіпотезу про рівність виправленої дисперсії вибірки з гіпотетичною генеральною дисперсією генеральної сукупності, то значення критерію розраховують як:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}. \qquad (2.15)$$

При цьому, якщо конкуруючою є однобічна гіпотеза $\sigma^2 > \sigma_0^2$, то як критичну точку беруть значення оберненого χ^2 розподілу, що відповідає рівню значущості α при $k = n - 1$ степенях вільності. Нульову гіпотезу приймають, якщо $\chi^2 < \chi^2_{êð}(\alpha, n-1)$.

Якщо конкуруючою є однобічна гіпотеза $\sigma^2 < \sigma_0^2$, то як критичну точку беруть значення оберненого χ^2 розподілу, що відповідає рівню значущості $1 - \alpha$ при $k = n - 1$ степенях вільності. Нульову гіпотезу приймають, якщо $\chi^2 > \chi^2_{êð}(1-\alpha, n-1)$.

Якщо ж конкуруючою є двобічна гіпотеза $\sigma^2 > \sigma_0^2$, то критичні точки розраховують за рівняннями:

$$P\left(\chi^2 < \chi^2_{\text{ë³â.ê}\text{ð}}\left(\alpha/2, n-1\right)\right) = \alpha/2;$$
$$P\left(\chi^2 > \chi^2_{\text{ï ðàâ.ê}\text{ð}}\left(\alpha/2, n-1\right)\right) = \alpha/2. \qquad (2.16)$$

Нульову гіпотезу приймають, якщо:

$$\chi^2_{\text{ë³â.ê}\text{ð}}\left(\alpha/2, n-1\right) < \chi^2 < \chi^2_{\text{ï ðàâ.ê}\text{ð}}\left(\alpha/2, n-1\right).$$

При цьому можна враховувати, що внаслідок симетрії розподілу:

$$P\left(\chi^2 < \chi^2_{\text{ë³â.ê}\text{ð}}\right) = 1 - P\left(\chi^2 > \chi^2_{\text{ï ðàâ.ê}\text{ð}}\right) = 1 - \alpha/2. \qquad (2.17)$$

Для оцінювання значущості отриманого значення χ^2 використовують **критерій В.І. Романовського** (запропонований радянським математиком Всеволодом Івановичем Романовським в 1928 р.):

$$R = \frac{\chi^2 - k}{\sqrt{2k}}. \qquad (2.18)$$

При $R \geq 3$ значення χ^2 вважають значущим, а порівнювані вибірки – істотно різними.

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.

2.3. Непараметричні тести

У багатьох випадках емпіричні дані не задовольняють нормальний розподіл. 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.

При перевірці нульової гіпотези про однорідність вибірок числових даних рекомендується [45] використовувати омега-квадрат критерій або (за відсутності необхідних таблиць та програмного забезпечення) критерій Смірнова.

Критерій омега-квадрат (**критерій Крамера – фон Мізеса**) базується на розгляді відхилення між двома емпіричними функціями розподілу (або між емпіричною й теоретичною функціями розподілу при ідентифікації закону розподілу). Вперше його запропонували у 1928–1930 р.

шведський математик Карл Харальд Крамер та американський математик і механік Ріхард фон Мізес, який народився у Львові.

У загальному вигляді критерій має вигляд:

$$\omega^2 = \int_{-\infty}^{\infty} \left[F_n(x) - F^*(x) \right]^2 dF^*(x), \qquad (2.19)$$

де $F_n(x)$, $F^*(x)$ – відповідно, значення емпіричної та теоретичної функцій розподілу в точці x.

Двохвибірковий варіант критерію запропоновано в 1951 р. американським статистиком Еріхом Лео Леманом та досліджено у 1952 р. американським математиком Мюреєм Розенблаттом. Тому його іноді називають також **критерієм Лемана – Розенблатта**. На сьогодні його вважають [45] найпотужнішим з критеріїв, призначених для перевірки гіпотези про однорідність незалежних вибірок.

Спочатку визначають ранги елементів i, j для кожної вибірки. Потім елементи обох вибірок об'єднують і визначають їх ранги r_i, s_j у загальній вибірці. Розрахункове значення визначають за формулою:

$$A = m\omega^2 = \frac{1}{mn(m+n)} \left[m \sum_{i=1}^{m} (r_i - i)^2 + n \sum_{j=1}^{n} (s_j - j)^2 \right] - \frac{4mn-1}{6(m+n)}, \qquad (2.20)$$

де m, n – обсяги порівнюваних вибірок. Нульову гіпотезу відхиляють, якщо розрахункове значення критерію перевищить критичне для заданого рівня значущості.

Критерій Смірнова запропонований в 1939 р. радянським математиком Миколою Васильовичем Смірновим. Він призначений для перевірки гіпотези про однорідність двох вибірок з неперервним законом розподілу. Значення статистики Смірнова можна розрахувати за формулою:

$$D_{m,n} = \sup_x \left| F_m(x) - G_n(x) \right|, \qquad (2.21)$$

де $Fm(x)$, $Gn(x)$ – значення функцій розподілу вибірок у точці x.

На практиці використовують емпіричні функції розподілу вибірок і розраховують значення критерію за формулами:

$$D^+_{m,n} = \max_{1 \le r \le n} \left[\frac{r}{n} - F_m(y'_r) \right] = \max_{1 \le s \le m} \left[G_n(x'_s) - \frac{s-1}{m} \right];$$

$$D^-_{m,n} = \max_{1 \le r \le n} \left[F_m(y'_r) - \frac{r-1}{n} \right] = \max_{1 \le s \le m} \left[\frac{s}{m} - G_n(x'_s) \right]; \qquad (2.22)$$

$$D_{m,n} = \max \left\{ D^+_{m,n}; D^-_{m,n} \right\},$$

де x'_s, y'_r – впорядковані за зростанням елементи досліджуваних вибірок.

Нульову гіпотезу відхиляють, якщо розрахункове значення критерію перевищить критичне для відповідного рівня значущості.

Для незалежних вибірок можна застосовувати **критерій рандомізації компонент**. Він розроблений Р. Фішером у 1920 р. для аналізу вибірок малого обсягу.

При порівнянні спряжених вибірок основою методу є перебирання можливих результатів, побудованих з різницевих оцінок. Нульова гіпотеза полягає у рівності вибіркових середніх. 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:

$$s_j = \sum_{i=1}^{n} a_{ji} \left| x_j - y_j \right| \ \left(j = 1, 2, ..., 2^n\right), \tag{2.23}$$

де $a_{ji} \ \left(j = 1, ..., 2^n;\ \mathrm{i} = 1, ..., n\right)$ – елементи матриці можливих результатів, розраховані згідно з методикою побудови повного ортогонального плану експерименту. Вона є матрицею з 2^n рядків та n стовпців. При цьому i-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 $S = \sum_{j=1}^{2^n} s_j$. Кількість сприятливих результатів:

$$N = \sum_{j=1}^{2^n} n_j, \quad n_j = \begin{cases} 0 \ \left(\mathrm{s_j} < S\right); \\ 1, \ \left(\mathrm{s_j} \geq S\right). \end{cases} \tag{2.24}$$

Однобічне P-значення розраховують за формулою $p = N / 2^n$ і порівнюють із заданим значенням довірчого рівня p^*. Якщо $\mathrm{p} > \mathrm{p}^*$, то нульову гіпотезу приймають на рівні значущості $1 - \mathrm{p}^*$. Для застосування однобіч-

ного критерію обсяг вибірки має бути не нижчим ніж 5 або 6 при рівнях значущості 0,01 та 0,05, відповідно. Для двобічних критеріїв і тих самих рівнів значущості мінімальні обсяги вибірки дорівнюють, відповідно, 7 та 8. За великих обсягів вибірок час обчислень швидко збільшується і доцільно використовувати інші критерії.

При порівнянні незалежних вибірок нульова гіпотеза полягає в належності двох досліджуваних вибірок до генеральних сукупностей з однаковими середніми. Нехай є дві вибірки: $x_i \ (i = 1, 2, ..., n_x)$ та $y_j \ (j = 1, 2, ..., n_y)$, де n_x, n_y – кількість елементів у них. Методика тесту основана на перебиранні всіх комбінацій даних. Знаходимо величину:

$$S = \min\left\{\sum_{i=1}^{n_x} x_i; \sum_{j=1}^{n_y} y_j\right\}. \tag{2.25}$$

Кількість сприятливих результатів визначаємо за формулою:

$$N = 2\sum_{j=1}^{C_n^m} n_j, \quad n_j = \begin{cases} 0 \ (s_j < S); \\ 1 \ (\mathrm{s_j} \geq S), \end{cases} \tag{2.26}$$

де n_j – оцінка j-го результату;

C_n^m – загальна кількість результатів;

$n = n_x + n_y$ – чисельність об'єднаної вибірки;

m – чисельність вибірки, що відповідає мінімальному значенню $s_j = \sum_{i=1}^{m} z_{ji} \ (j = 1, 2, ..., C_n^m)$;

z_{ji} – масив сполучень з об'єднаної вибірки, який будують подібно до розглянутої раніше процедури побудови матриці можливих результатів для спряжених вибірок. Однобічне значення p розраховують як $p = N / C_n^m$. Його порівнюють із заданим рівнем значущості α. Нульову гіпотезу відхиляють, якщо $p < \alpha$ або $p > 1 - \alpha$. Як і у попередньому випадку, критерій застосовують для відносно малих вибірок. Їх мінімальний допустимий обсяг є таким самим, як і для спряжених вибірок.

***И*-критерій Уїлкоксона (критерій рангових сум)** запропонований в 1945 р. американським хіміком і статистиком Френком Уїлкоксоном. Його 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 числовими значеннями даних, а їх рангами – місцями у впорядкованих за згасанням або зростанням рядах даних. При його застосуванні

передбачається, що розподіли вибірок є неперервними, а нульова гіпотеза полягає в тому, що функції розподілу вибірок збігаються одна з одною.

Процедура обчислення значення критерію є близькою до обчислення критерію рандомізації компонент. Різниця полягає в тому, що замість вихідних даних використовують їх ранги. Ранжирування порівнюваних вибірок здійснюють сумісно. Вихідні дані об'єднують до однієї вибірки, впорядковують, визначають ранги елементів об'єднаної вибірки. Потім формують дві нові вибірки, елементами яких є ранги відповідних елементів вихідних вибірок. Якщо деякі значення збігаються, то відповідним спостереженням призначають середній ранг. Обчислення статистики критерію здійснюють за формулою:

$$W = \min\left\{\sum_{i=1}^{n_1} R_i, \sum_{i=1}^{n_2} S_i\right\}, \qquad (2.27)$$

де R_i – ранги вибірки, що має найменшу, а S_i – вибірки, яка має найбільшу суму рангів. Для вибірок малого обсягу (до 25 елементів) суму рангів W' вибірки, що має меншу кількість елементів, порівнюють з критичним значенням, яке визначають за спеціальними таблицями [25]. Нульову гіпотезу відхиляють, якщо: $W' < w_{íèæí.êð}(\alpha/2, n_1, n_2)$ або $W' > w_{âåðõ.êð}(\alpha/2, n_1, n_2)$ при нульовій гіпотезі $F_1(x) \neq F_2(x)$; $W' < w_{íèæí.êð}(\alpha, n_1, n_2)$ при нульовій гіпотезі $F_1(x) > F_2(x)$; $W' > w_{âåðõ.êð}(\alpha, n_1, n_2)$ при нульовій гіпотезі $F_1(x) < F_2(x)$.

Статистика $W^* = \left|\dfrac{W - \mu_W}{\sigma_W}\right|$, де $\mu_W = \dfrac{n_1(N+1)}{2}$ – математичне сподівання, $\sigma_W^2 = \dfrac{n_1 n_2 (N+1)}{12}$ – дисперсія, $N = n_1 + n_2$, при збільшенні N наближається до стандартного нормального розподілу. Для вибірок великого обсягу (понад 25 елементів) 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 $W + 0{,}5$.

При застосуванні W-критерію Уїлкоксона слід мати на увазі, що згідно з [45] він належить до так званих критеріїв зсуву. Тобто найбільш потужним він є при виявленні різниці, пов'язаної з тим, що одну з вибірок отримано додаванням одного й того самого числа до всіх елементів іншої вибірки.

Він є нечутливим до різниці дисперсій порівнюваних вибірок, а також коефіцієнтів їх асиметрії та ексцесу. Зокрема, якщо дві вибірки мають симетричні функції розподілу з однаковими середніми значеннями, але різними стандартними відхиленнями, то в об'єднаної послідовності елементи однієї вибірки матимуть підвищену кількість елементів з високими та

низькими рангами. Елементи іншої вибірки будуть мати підвищену кількість елементів із середніми значеннями рангів. Але суми рангів усіх елементів для цих двох вибірок можуть бути приблизно однаковими.

***U*-критерій Манна – Уїтні** запропонований в 1947 р. американськими математиками Г.Б. Манном та Д.Р. Уїтні. Він 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. Цей критерій є найпотужнішим непараметричним аналогом *t*-критерію Стьюдента для незалежних вибірок. У деяких випадках його потужність може бути навіть більшою, ніж у *t*-критерію.

Обчислення здійснюють за формулами:

$$\begin{aligned} U_1 &= n_1 n_2 + n_1\left(n_1+1\right)/2 - R_1; \\ U_2 &= n_1 n_2 + n_2\left(n_2+1\right)/2 - R_2; \\ U &= \max\left\{U_1, U_2\right\}, \end{aligned} \qquad (2.28)$$

де R_1, R_2 – суми рангів вибірок;

n_1, n_2 – кількість елементів у них. Якщо $n_1, n_2 > 20$, то розподіл вибірки для *U*-статистики наближається до нормального. Правильність обчислення величин U_1, U_2 можна перевірити за формулою:

$$U_1 + U_2 = n_1 n_2 . \qquad (2.29)$$

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

***T*-критерій Уїлкоксона** (**одновибірковий**, **знаковий ранговий критерій Уїлкоксона**) запропонований Ф. Уїлкоксоном в 1945 р. Його застосовують для порівняння вибірок з попарно спряженими значеннями. Він є непараметричним аналогом *t*-критерію Стьюдента для спряжених вибірок. Перевіряють нульову гіпотезу про симетричність розподілу різниць спряжених значень стосовно нуля. Методика розрахунку є близькою до розрахунку значення *W*-критерію Уїлкоксона, але в цьому випадку оперують модулями різниць відповідних значень. Масив модулів різниць, з якого вилучені нульові значення, ранжирують. Потім рангам надають знаки різниць, обчислюють суми додатних (W^+) і від'ємних (W^-) рангів, й беруть $W^* = \min(W^-, W^+)$. Цю величину порівнюють з критичним значенням.

Для перевірки правильності розрахунків можна використовувати тотожність:

$$W^+ + W^- = \frac{N(N+1)}{2}. \tag{2.30}$$

Статистика $\frac{W^* - \mu_{W^*}}{\sigma_{W^*}}$, де $\mu_{W^*} = \frac{N(N+1)}{4}$ – математичне сподівання, $\sigma^2_{W^*} = \frac{N(N+1)(2N+1)}{24}$ – дисперсія, N – чисельності вибірок, має стандартний нормальний розподіл.

Розглянемо докладніше дві проблеми, що виникають при застосуванні критерію Уїлкоксона та подібних йому рангових критеріїв [23]. Першою з них є проблема рівних рангів, яка існує лише при застосуванні нормальної або іншої апроксимації критерію і не виникає при точному обчисленні відповідних статистик. Для *W*-критерію Уїлкоксона у цьому випадку необхідно врахувати поправку до дисперсії та розраховувати її за формулою:

$$\sigma^2_W = \frac{n_1 n_2}{12}\left[N + 1 - \frac{T}{N(N-1)}\right], \tag{2.31}$$

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

$$\sigma^2_{W^*} = \frac{2N(N+1)(2N+1) - T}{48}, \tag{2.32}$$

де N – чисельність кожного ряду.

Інша проблема виникає тільки для *T*-критерію Уїлкоксона і полягає в тому, що наявність збігів одночасно в обох аналізованих спряжених вибірках призводить до нульових різниць. На сьогодні ця проблема залишається недостатньо 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 (**хі-квадрат**) запропонований в 1900 р. видатним британським математиком, біологом та філософом Карлом Пірсоном. Його використовують для перевірки нульової гіпотези про однаковість розподілу досліджуваних випадкових величин. Його широко застосовують у дисперсійному аналізі та інших методах аналізу даних. Цей критерій оперує не первинними даними, а їх розподілом за класами. З огляду на це 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:

$$\chi^2 = \sum_{i=1}^{N} \frac{\left(f_i - g_i\right)^2}{f_i + g_i}, \tag{2.33}$$

де f_i, g_i $\left(i = 1, 2, ..., N\right)$ – частоти розподілів порівнюваних вибірок;

N – кількість класів.

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

$$\chi^2 = \frac{1}{n_1 n_2} \sum_{i=1}^{N} \frac{\left(n_2 f_i - n_1 g_i\right)^2}{f_i + g_i}, \tag{2.34}$$

де n_1, n_2 – кількості спостережень у порівнюваних масивах.

Критерій χ^2 можна застосовувати також і для порівняння вибірок значень номінальних ознак. У цьому випадку аналізують дані, подані у вигляді таблиці спряженості ознак. Елементами таблиці є числа, рівні кількостям елементів досліджуваних вибірок, для яких досліджувана ознака набуває значень, котрі відповідають певному 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:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(a_{ij} - e_{ij}\right)^2}{e_{ij}}, \tag{2.35}$$

де a_{ij} $\left(i = 1, 2, ..., r;\ j = 1, 2, ..., c\right)$ – елементи таблиці спряженості ознак;

r – кількість вибірок (стовпців); c – кількість класів (рядків);

e_{ij} – очікувані величини, що відповідають значенням a_{ij}.

Їх обчислюють як доданок i-го вектора-стовпця на j-й вектор-рядок, поділений на суму елементів усієї таблиці $\sum_{i=1}^{r} \sum_{j=1}^{c} a_{ij}$. Кількість степенів вільності при обчисленні P-значення статистики χ^2 беруть рівною $\left(r-1\right)\left(c-1\right)$.

Критерій серій Вальда – Волфовиця розроблений в 1940 р. американськими математиками Абрахамом Вальдом і Джекобом Волфовицем. Його використовують для перевірки нульової гіпотези про те, що дві неза-

лежні випадкові вибірки обсягами n_1 та n_2 не відрізняються одна від одної за досліджуваною ознакою.

Результати спостережень записують як варіаційний ряд об'єднаної вибірки, а їх належність до вихідних вибірок помічають за допомогою додаткової змінної, яка може набувати два значення, наприклад "0" та "1". Послідовність її значень називають **послідовністю кодів**.

Серією послідовності кодів називають будь-яку послідовність її однакових значень. Наприклад, у послідовності 00101111011 є такі серії: 00, 1, 0, 1111, 0, 11. Очевидно, що за умови справедливості нульової гіпотези кількість серій N має бути великою, а за умови її помилковості – відносно малою. Якщо обсяги вибірок є достатньо великими $(n_1, n_2 > 20)$ для перевірки нульової гіпотези можна використовувати статистику:

$$Z = \frac{\left| N - \left(\frac{2n_1 n_2}{n_1 + n_2} \right) + 1 \right| - \frac{1}{2}}{\sqrt{\frac{2n_1 n_2 \left(2n_1 n_2 - n_1 - n_2 \right)}{\left(n_1 + n_2 \right)^2 \left(n_1 + n_2 - 1 \right)}}}, \tag{2.36}$$

яка має стандартний нормальний розподіл.

Критерій знаків запропонований Ф. Уїлкоксоном. Його використовують для перевірки нульової гіпотези про однорідність двох спряжених вибірок. Нехай x_i і y_i – значення відповідних елементів цих вибірок. Якщо вибірки є однорідними, то ймовірності появи додатних та від'ємних різниць $x_i - y_i$ є рівними. Імовірність появи нульових значень цих різниць вважається нульовою, оскільки передбачається, що розподіл досліджуваної ознаки є неперервним. Якщо внаслідок випадкових похибок або округлення результатів такі різниці з'являються, то відповідні спостереження виключають з подальшого аналізу. За умови справедливості нульової гіпотези ймовірність p появи знаків певного типу (наприклад знаків "+") різниць $x_i - y_i$ підпорядковується біноміальному розподілу з параметрами: $p = 1/2;\ m$, де m – кількість різниць, що аналізують.

Нульовою гіпотезою є $\mathrm{H}_0 : p = 1/2$. Як конкуруючі можна розглядати такі гіпотези: $\mathrm{H}_1^{(1)} : p > 1/2;\ \mathrm{H}_1^{(2)} : p < 1/2;\ \mathrm{H}_1^{(3)} : p \neq 1/2$. Нульову гіпотезу відхиляють на рівні значущості α, якщо:

– при конкуруючій гіпотезі $\mathrm{H}_1^{(1)}$ виконується нерівність:

$$\sum_{i=r}^{m} C_m^i \left(\frac{1}{2} \right)^m \leq \alpha, \tag{2.37}$$

де r – кількість додатних різниць;

C_m^i – кількість сполучень;

– при конкуруючій гіпотезі $H_1^{(2)}$ виконується нерівність:

$$\sum_{i=0}^{r} C_m^i \left(\frac{1}{2}\right)^m \le \alpha, \tag{2.38}$$

– при конкуруючій гіпотезі $H_1^{(3)}$ виконується одна з нерівностей:

$$\sum_{i=0}^{r} C_m^i \left(\frac{1}{2}\right)^m \le \frac{\alpha}{2}; \tag{2.39}$$

$$\sum_{i=r}^{m} C_m^i \left(\frac{1}{2}\right)^m \le \frac{\alpha}{2}. \tag{2.40}$$

Для обчислення кількості сполучень можна використовувати такі апроксимації біноміального розподілу нормальним:

$$\sum_{i=0}^{r} C_m^i \left(\frac{1}{2}\right)^m \approx \Phi\left(\frac{r - m/2}{\sqrt{m/4}}\right), \text{ якщо } m > 50; \tag{2.41}$$

$$\sum_{i=0}^{r} C_m^i \left(\frac{1}{2}\right)^m \approx \Phi\left(\frac{r - m/2 + 0{,}5}{\sqrt{m/4}}\right), \text{ якщо } m \le 50, \tag{2.42}$$

де $\Phi(x)$ – функція розподілу для стандартного нормального закону.

2.4. Визначення моделей розподілу емпіричних даних

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

Особливу роль відіграє перевірка розподілу на нормальність, оскільки її прийняття дає змогу застосовувати більш досліджені параметричні критерії перевірки наступних гіпотез.

Для перевірки відповідності емпіричного розподілу теоретичному застосовують так звані **критерії згоди**: ω^2, Смірнова, χ^2, Ястремського, Бернштейна та інші.

Критерій ω^2 (Крамера – фон Мізеса) запропонований в 1928–1930 р. К. Крамером та Р. фон Мізесом. Його використовують у випадках, коли необхідно перевірити нульову гіпотезу про відповідність ви-

бірки певному відомому закону розподілу. Розрахункове значення обчислюють за формулою:

$$n\omega^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left[F\left(x_i\right) - \frac{2i-1}{2n} \right]^2, \qquad (2.43)$$

де $F(x)$ = теоретична функція розподілу, n – обсяг вибірки.

При $n > 40$ критичні значення визначають згідно з табл. 2.1 [37]. Якщо значення параметрів розподілу визначають за вибіркою, то критичні значення суттєво зменшуються.

Таблиця 2.1

Критичні значення статистики ω^2

α	0,900	0,950	0,990	0,995	0,999
$n\omega^2(\alpha)$	0,3473	0,4614	0,7435	0,8694	1,1679

Критерій Смірнова у вигляді (2.21) застосовують, якщо емпіричну функцію розподілу будують за масивом частот. У випадку, коли її побудову здійснюють безпосередньо за вихідною вибіркою (при цьому чисельність вихідної вибірки та масиву функції розподілу збігаються), для розрахунку критерію при двобічній гіпотезі застосовують формули:

$$\begin{gathered} D_n = \max_{1\le m\le n}\left\{ D_n^{(1)}, D_n^{(2)} \right\}; \\ D_n^{(1)} = \max_{1\le m\le n}\left\{ \frac{m}{n} - F\left(x_m\right) \right\};\ D_n^{(2)} = \max_{1\le m\le n}\left\{ F\left(x_m\right) - \frac{m-1}{n} \right\}, \end{gathered} \qquad (2.44)$$

а при однобічній – $D_n = D_n^{(1)}$.

Функція розподілу D_n є однією й тією самою для всіх неперервних розподілів, а функція розподілу величини $K = D_n\sqrt{n}$ за великих n збігається до **статистики Колмогорова – Смірнова**, тому в літературі його часто називають **критерієм Колмогорова – Смірнова**. Критерій Смірнова за певних умов можна використовувати також для порівняння емпіричних функцій розподілу двох вибірок (перевірка гіпотези про їх однорідність).

Обмеженнями для застосування цього критерію є:

– вимога щодо неперервності теоретичної функції розподілу;

– необхідність мати достатньо представницькі вибірки $\left(n > 200\right)$;

– необхідність незалежних, тобто отриманих не за самою вибіркою, оцінок параметрів розподілу (проста гіпотеза) для порівняння емпіричної й теоретичної функцій розподілу.

Якщо обсяги вибірок $n > 35$, то критичне значення статистики Колмогорова – Смірнова, що відповідає рівню значущості α, можна розраховувати за формулою:

$$K_\alpha \approx \sqrt{-\frac{\ln\frac{\alpha}{2}}{2}}. \qquad (2.45)$$

При застосуванні критерію Смірнова для перевірки нульової гіпотези про однорідність двох вибірок достатньо великого обсягу $(n_1, n_2 > 40)$ можна використовувати [19] такі критичні значення:

$$\text{при } \alpha = 0,05 \;\; D_c = 1,36\sqrt{\frac{n_1 n_2}{n_1 + n_2}}; \tag{2.46}$$

$$\text{при } \alpha = 0,10 \;\; D_c = 1,22\sqrt{\frac{n_1 n_2}{n_1 + n_2}}. \tag{2.47}$$

Для складних гіпотез уводять модифіковані статистики Колмогорова, але їх функції розподілу є різними для різних типів неперервних розподілів. На відміну від більшості інших критеріїв, критерій Смірнова за достатнього обсягу досліджуваної вибірки дає змогу встановити різницю емпіричної й теоретичної функцій розподілу незалежно від параметрів, що її зумовлюють. Зокрема він є чутливим до різниці як вибіркових середніх, так і стандартних відхилень вибірок, коефіцієнтів їх асиметрії та ексцесу. Але така універсальність досягається за рахунок зменшення потужності критерію.

Критерій χ^2 як критерій згоди застосовують для порівняння емпіричної та теоретичної функцій розподілу. Він оперує не первинними даними, а їх розподілом за класами рівної ширини. Тому необхідно враховувати вимогу щодо мінімальних обсягів ряду спостережень і кількості класів. За різними оцінками, мінімальна допустима кількість класів знаходиться у межах 4–7, а кількість елементів у ряді спостережень – в межах 20–200.

Значення критерію розраховують за формулою:

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(\nu_i - np'_i\right)^2}{np'_i}, \tag{2.48}$$

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 $k - r - 1$, 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 на основі її аналізу визначають r параметрів, які необхідні для побудови теоретичної функції розподілу. Далі визначають межі класових інтервалів, абсолютні частоти ν_i й теоре-

тичні ймовірності p'_i. Після цього розраховують значення критерію χ^2 і порівнюють його з критичним.

Розглянуті вище критерії можна застосовувати лише для вибірок достатньо великого обсягу. Для перевірки нормальності вибірок обсягом 3–50 значень використовують ***W*-критерій Шапіро – Уїлка**. Він запропонований в 1965 р. американським статистиком Самуелом Шапіро й канадським статистиком Мартіном Уїлком і є одним з найефективніших методів вирішення цього завдання. Цей критерій базується на регресії порядкових статистик. Його обчислення здійснюють за формулами:

$$W = b^2 / S^2;\ S^2 = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2;\ b = \sum_{i=1}^{k} a_{n-i+1}\left(x_{n-i+1} - x_i\right), \qquad (2.49)$$

де x_i $\left(i=1, 2,..., n\right)$ – ранжируваний ряд; n – обсяг вибірки, параметр k беруть рівним $n/2$ для парних i $\left(n-1\right)/2$ для непарних n, a_{n-i+1} $\left(i=1, 2,..., k;\ n=3, 4,..., 50\right)$ – 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 $\approx \mu + 2\sigma$.

Найпростіші способи перевірки нормальності вибірки базуються на розрахунку значень коефіцієнтів його асиметрії та ексцесу. Вважають, що розподіл є близьким до нормального, якщо $|A|, |E| < 0{,}1$, і сильно відрізняється від такого, коли $|A|, |E| > 0{,}5$.

Ще одним способом перевірки типу розподілу є побудова емпіричної функції розподілу в певних координатах, що лінеаризують її графік. У статистичних пакетах SPSS, Statistica та інших реалізовано спеціальні засоби такої перевірки. Як приклад на рис. 2.3, 2.4 наведені лінеаризовані графіки емпіричних функцій розподілу вибірок з генеральних сукупностей, що підпорядковуються рівномірному розподілу на відрізку [−3, 3] і стандартному нормальному розподілу.

На рис. 2.3 графіки побудовано в координатах, що мають лінеаризувати функцію нормального розподілу, а на рис. 2.4 – функцію рівномірного розподілу.

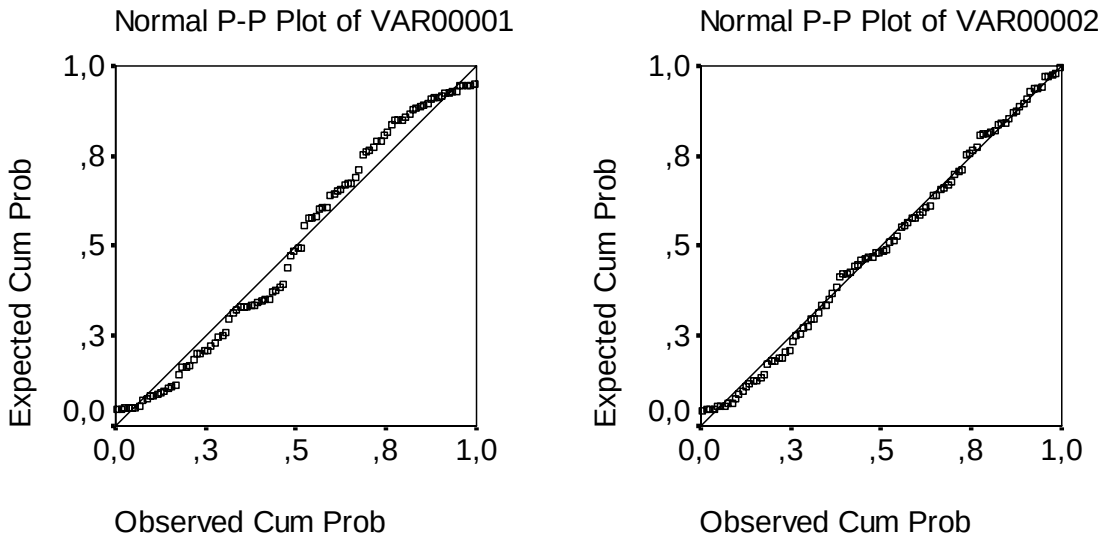


Рис. 2.3. P-P діаграми рівномірно й нормально розподілених вибірок у координатах, що лінеаризують функцію нормального розподілу

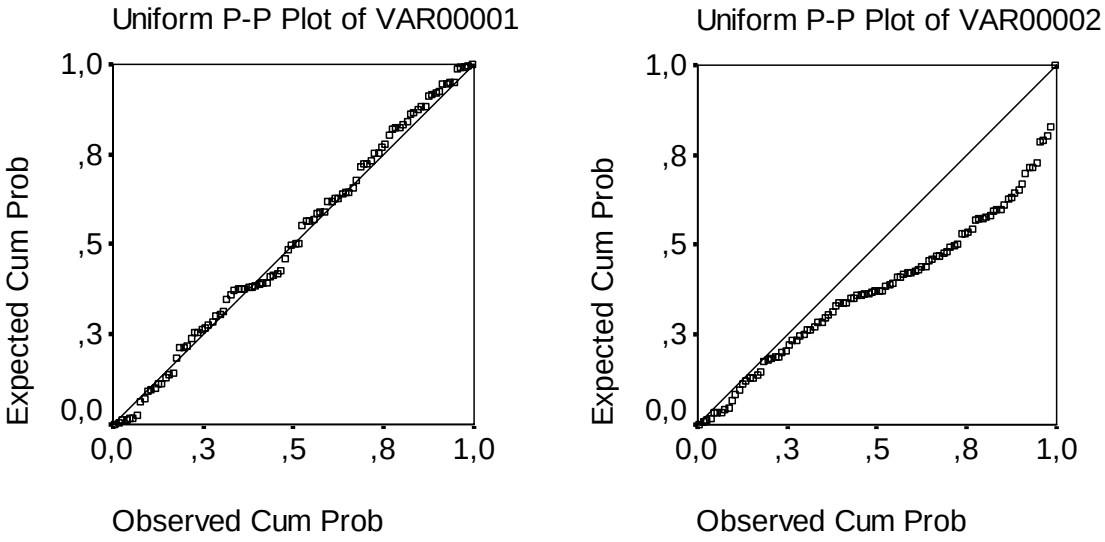


Рис. 2.4. P-P діаграми рівномірно й нормально розподілених вибірок у координатах, що лінеаризують функцію рівномірного розподілу

2.5. Приклад ідентифікації функції розподілу однорідної вибірки

Ідентифікація емпіричних функцій розподілу є відносно простою для однорідних вибірок. У цьому випадку її алгоритм може бути таким.

1. За допомогою P-P діаграм статистичних пакетів (SPSS, Statistica тощо) підбираємо найбільш придатний тип розподілу.

2. Використовуючи статистичні пакети або мінімізуючи суму квадратів залишків моделі за допомогою процедури “Пошук розв’язку” електронних таблиць MS Excel, уточнюємо параметри розподілу.

3. Перевіряємо адекватність підібраної моделі розподілу, використовуючи критерії Крамера – Уелча та Фішера (для нормального розподілу), ω^2 або Смирнова (для інших типів розподілу).

4. Розглянемо як приклад завдання ідентифікації моделі розподілу питомого електричного опору епітаксійних шарів кремнієвих композицій [13].

До робочого вікна пакету SPSS (рис. 2.5) уводимо значення елементів досліджуваної вибірки, що є значеннями питомого електричного опору епітаксійного шару досліджуваної серії виробів. У головному меню обираємо пункти Graphs / P-P Plots. При цьому відчиняється діалогове вікно (рис. 2.6).

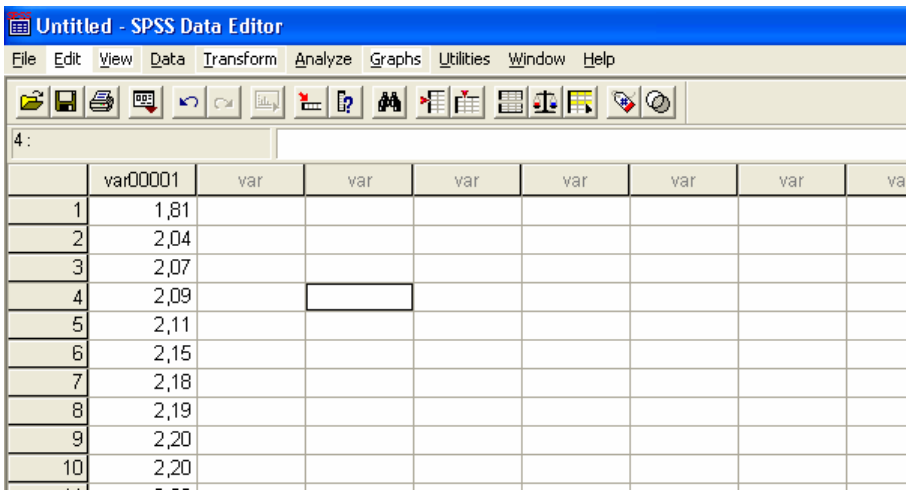


Рис. 2.5. Головне вікно для уведення даних пакету SPSS

У цьому вікні зазначаємо, для якої вибірки треба побудувати P-P діаграму, а також, який саме тип розподілу перевірятимемо. При цьому можливо вибрати такі типи розподілу: бета, χ^2, експоненціальний, гама, напівнормальний, Лапласа, логістичний, логнормальний, нормальний, Парето, Стьюдента, Вейбула, рівномірний.

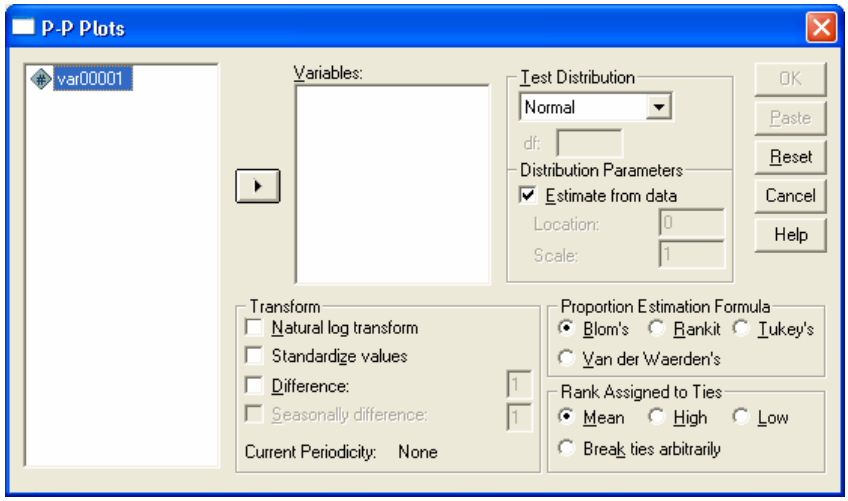


Рис. 2.6. Діалогове вікно побудови P-P діаграм

Для досліджуваної вибірки одержуємо такі результати. Бета розподіл підібрати не вдається (параметри, що визначаються при мінімізації цільового функціонала, не відповідають обмеженням цього розподілу). Розподіли χ^2, експоненціальний, напівнормальний, Парето, Стьюдента й рівномірний (рис. 2.7–2.12) суттєво відрізняються від розподілу досліджуваної вибірки.

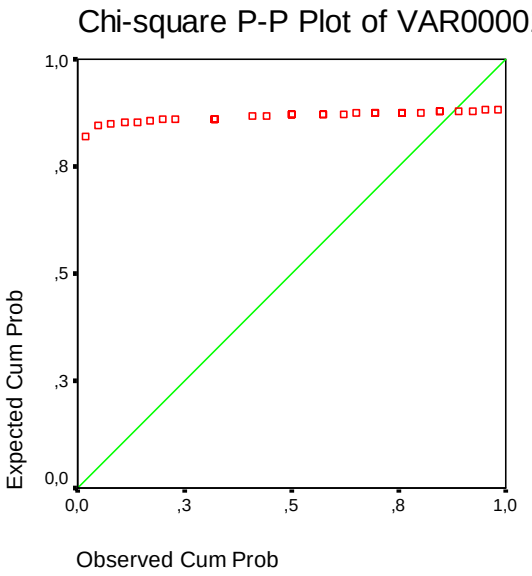


Рис. 2.7. P-P діаграма досліджуваної вибірки для χ^2 розподілу

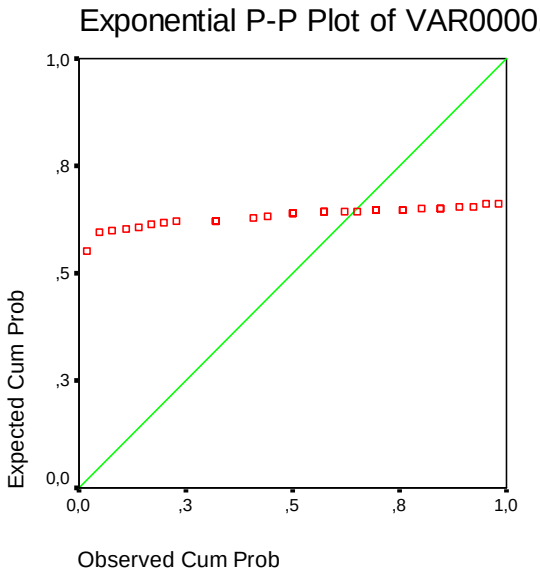


Рис. 2.8. P-P діаграма досліджуваної вибірки для експоненціального розподілу

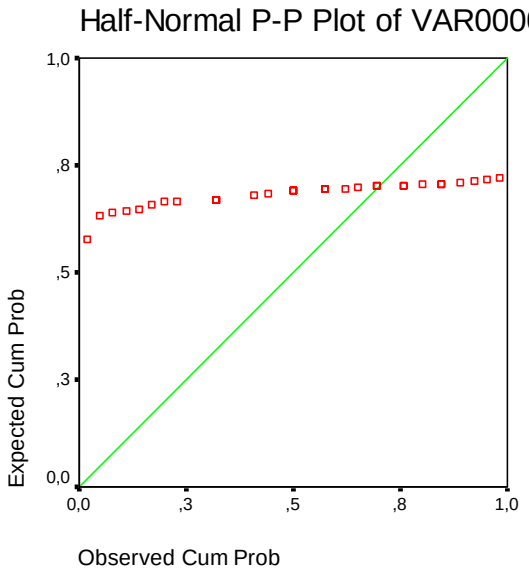


Рис. 2.9. P-P діаграма досліджуваної вибірки для напівнормального розподілу

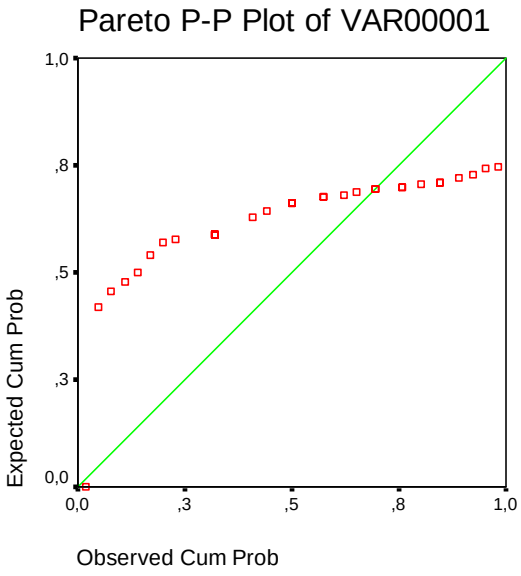


Рис. 2.10. P-P діаграма досліджуваної вибірки для розподілу Парето

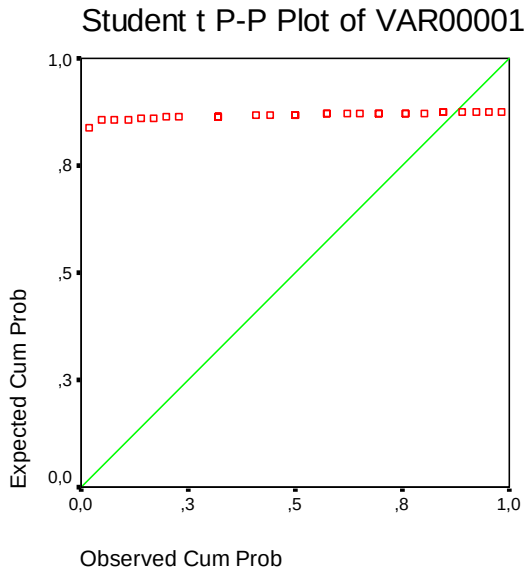


Рис. 2.11. P-P діаграма досліджуваної вибірки для розподілу Стьюдента

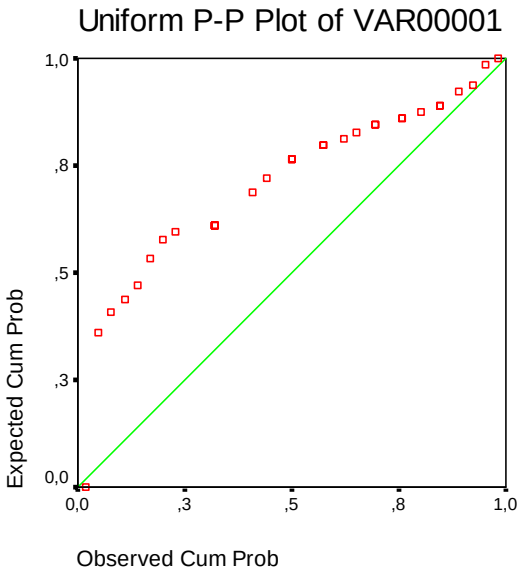


Рис. 2.12. P-P діаграма досліджуваної вибірки для рівномірного розподілу

Більш придатними є розподіли гама з параметром форми 276,9 й параметром масштабу 122,4 (рис. 2.13), Лапласа (рис. 2.14), логістичний з параметром розташування 2,26 й параметром масштабу 0,0749 (рис. 2.15), логнормальний з параметром масштабу 2,257 й параметром форми 0,0627 (рис. 2.16), нормальний (рис. 2.17) і Вейбулла з параметром масштабу 2,324 й параметром форми 19,08 (рис. 2.18). При цьому для нормального розподілу і розподілу Лапласа параметри моделі не визначаються.

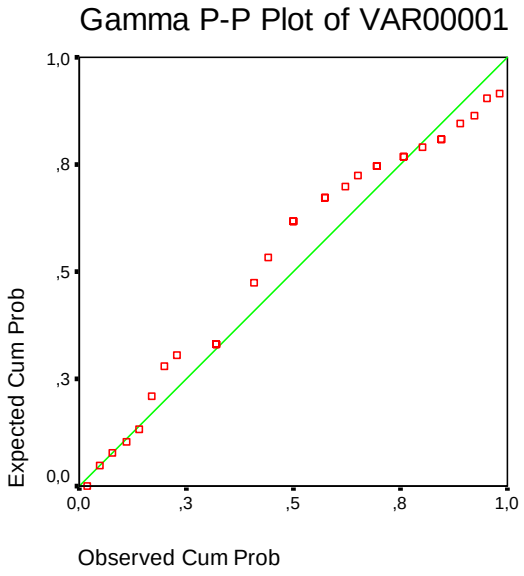


Рис. 2.13. P-P діаграма досліджуваної вибірки для гама розподілу

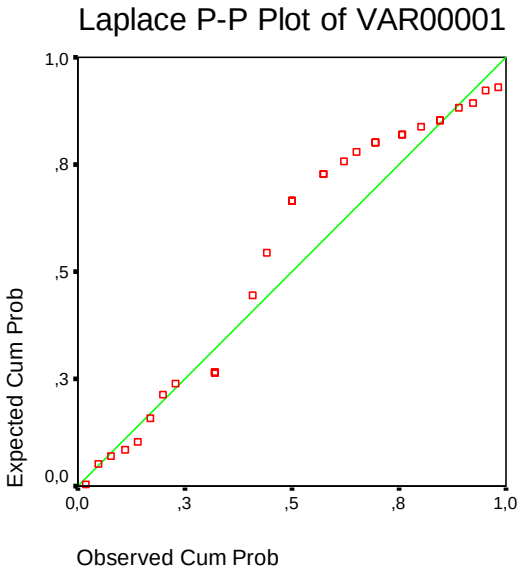


Рис. 2.14. P-P діаграма досліджуваної вибірки для розподілу Лапласа

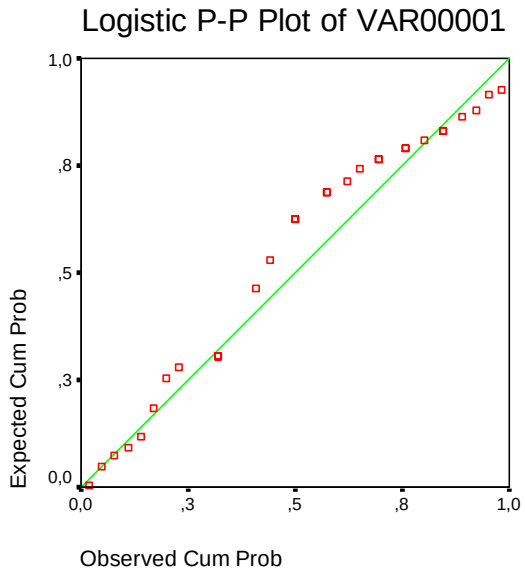


Рис. 2.15. P-P діаграма досліджуваної вибірки для логістичного розподілу

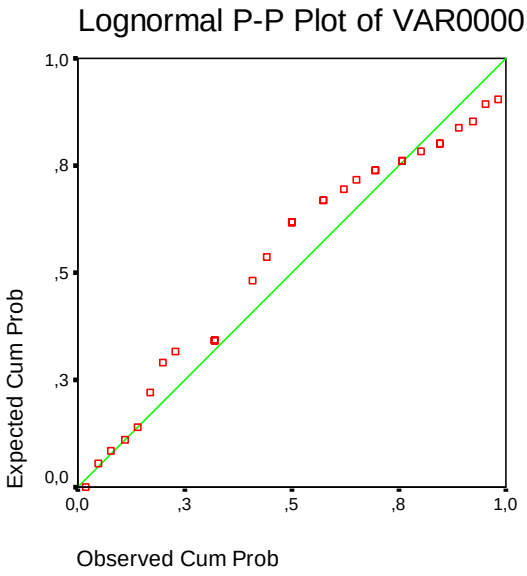


Рис. 2.16. P-P діаграма досліджуваної вибірки для логнормального розподілу

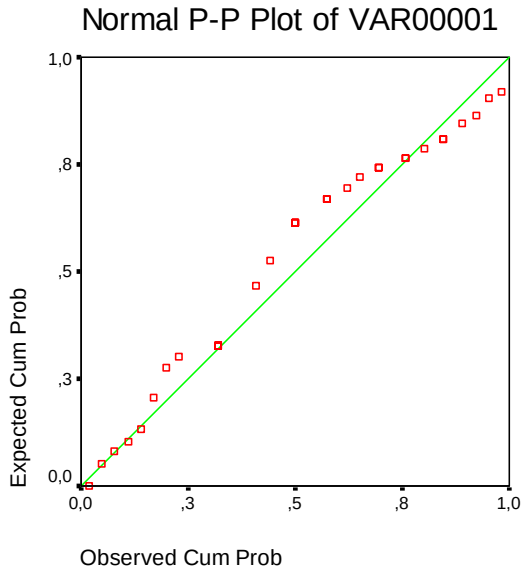


Рис. 2.17. P-P діаграма досліджуваної вибірки для нормального розподілу

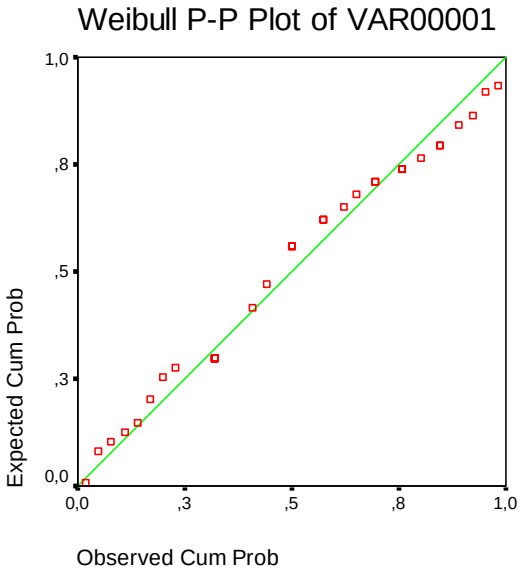


Рис. 2.18. P-P діаграма досліджуваної вибірки для розподілу Вейбулла

З наведених P-P діаграм бачимо, що найбільш придатним для опису досліджуваної вибірки є розподіл Вейбулла.

Для перевірки гіпотези, що вибірка підпорядковується розподілу Вейбулла, використовуємо електронні таблиці MS Excel. Для цього у робочому вікні сформуємо стовпчик даних, що містить елементи досліджуваної вибірки. Потім впорядкуємо дані за зростанням. Для цього у головному меню обираємо пункт “Дані” і підпункт “Сортування”. При цьому з’являється діалогове вікно (рис. 2.19).

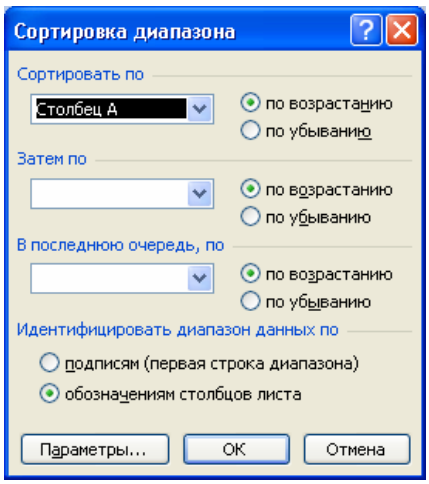


Рис. 2.19. Діалогове вікно сортування даних

Після цього у сусідньому стовпчику поруч з кожним значенням вихідної вибірки наведемо величину n / N, де n – ранг спостереження, N – обсяг вибірки. Отримані величини є значеннями емпіричної функції розподілу f (x_i), а відповідні елементи досліджуваної вибірки – значеннями аргументу x_i.

Для перевірки гіпотези про тип розподілу розрахуємо значення функції розподілу Вейбулла. Для цього у третьому стовпчику кожному значенню x_i приведемо у відповідність величину, що визначається формулою:

$$=(1–EXP (–СТЕПЕНЬ (A4/\$N\$20;\$O\$20))), \qquad (2.50)$$

де \$N\$20, \$O\$20 – посилання на комірки, куди уведено значення параметрів розподілу. Як початкові, будемо використовувати значення, розраховані у пакеті SPSS. Далі формуємо стовпчик модулів різниць відповідних значень емпіричної й теоретичної функцій розподілу. В окремій комірці розраховуємо суму квадратів цих різниць. Далі використовуємо цю комірку як цільову у процедурі “Пошук розв’язку”. Для її застосування у головному меню обираємо пункт “Сервіс” і підпункт “пошук розв’язку”. При цьому з’являється діалогове вікно (рис. 2.20).

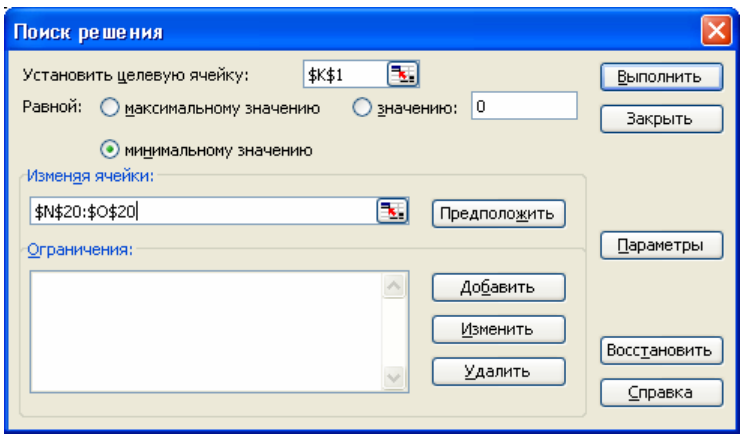


Рис. 2.20. Діалогове вікно процедури “Пошук розв’язку”

У цьому вікні зазначаємо, що шукатимемо мінімальне значення, і й робимо посилання на цільову комірку й комірки, де розташовані значення параметрів розподілу, які треба змінювати для отримання мінімального значення цільової комірки. Крім того встановлюємо значення параметрів процедури. Для цього натискуємо кнопку параметри, і з'являється діалогове вікно (рис. 2.21).

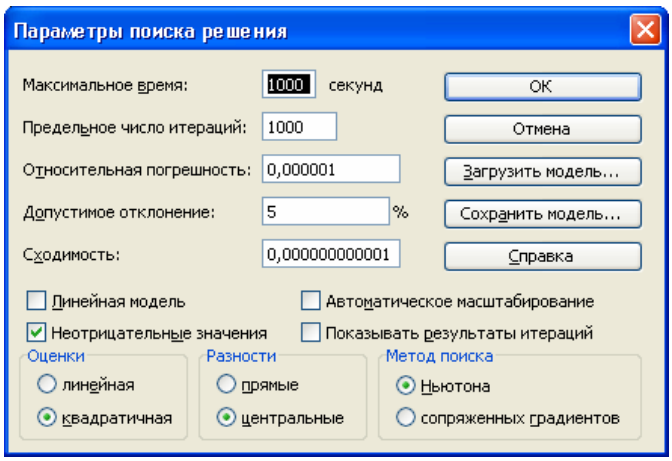


Рис. 2.21. Діалогове вікно встановлення параметрів процедури "Пошук розв'язку"

Для задачі, що розглядається, необхідно задати такі параметри:

– максимальний час;
– гранична кількість ітерацій;
– збіжність;
– невід'ємні значення;
– квадратичні оцінки;
– центральні різниці;
– метод Ньютона.

Інші параметри не задаємо, оскільки для нашої задачі вони є несуттєвими.

Максимальний час та граничну кількість ітерацій можна залишити рівними 100 (значення, що встановлюються за умовчанням). При підборі значень параметрів однорідних вибірок час та кількість ітерацій зазвичай не перевищують цих значень. У більш складних випадках ці параметри доцільно брати більшими. Значення параметра "Збіжність" дає максимальну величину різниці між двома послідовними значеннями, одержуваними у цільовій комірці, після досягнення якої ітерації зупиняються. Чим меншою буде ця величина, тим більшими будуть час та кількість ітерацій, необхідних для отримання розв'язку. Обмеження на невід'ємні значення пов'язано з тим, що у нашому випадку від'ємні значення не відповідають задачі, що розв'язується.

У досліджуваному випадку після оптимізації отримуємо значення параметра масштабу 2,319 й параметра форми – 20,67. Вони дещо відрізняються від значень, одержаних у пакеті SPSS.

Якщо мінімізувати не суму квадратів різниць значень емпіричної й теоретичної функцій розподілу, а максимальний модуль цих різниць, знов одержимо дещо інші значення параметрів розподілу: параметра масштабу 2,321 й параметра форми – 18,51. Бачимо, що результат визначення параметра масштабу є більш стійким і менше залежить від способу його визначення.

Для перевірки адекватності моделі розрахуємо значення критерію Смірнова, яке дорівнює в останньому випадку 0,49. Воно є значно нижчим, ніж критичне значення, яке при розрахунку параметрів розподілу безпосередньо за вибіркою дорівнює [21] 0,895 при рівні значущості 0,05.

Аналогічно ми можемо перевірити гіпотезу про інший тип розподілу. Зокрема для перевірки гіпотези про відповідність вибірки нормальному закону розподілу у комірки для розрахунку значень теоретичної функції розподілу уведемо формули:

$$=\text{НОРМРАСП (A6;\$M\$3;\$M\$7; ИСТИНА)}, \qquad (2.51)$$

де \$M\$3, \$M\$7 – посилання на комірки, зі значеннями математичного сподівання й стандартного відхилення. Ці значення можна розрахувати за допомогою функцій =СРЗНАЧ (A1: A33) та =СТАНДОТКЛОН (A1: A33), де A1: A33 – посилання на діапазон комірок, в яких розташовано елементи досліджуваної вибірки.

У цьому випадку розрахункове значення критерію Смірнова дорівнює 0,73, що також менше, ніж критичне значення, але значно вище, ніж розрахункове значення для розподілу Вейбулла.

На рис. 2.22 наведено графіки емпіричної функції розподілу, а також підібраних теоретичних моделей. Не зважаючи на помітну різницю між ними, ми не маємо підстав для відхилення нульових гіпотез про відповідність розглянутих моделей однорідних розподілів наявним емпіричним даним, оскільки в обох випадках розрахункове значення критерію Смірнова не перевищує критичного.

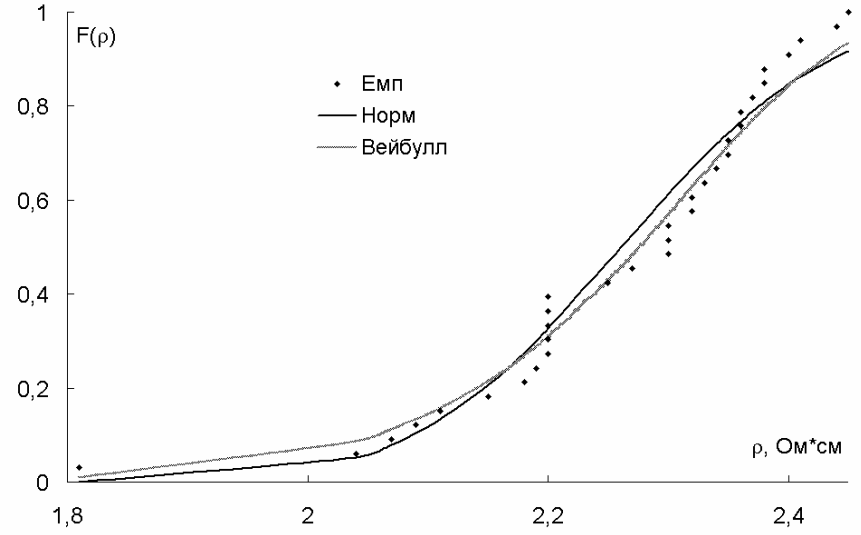


Рис. 2.22. Емпірична й теоретичні функції розподілу досліджуваної вибірки

2.6. Приклад ідентифікації функції розподілу неоднорідної вибірки

Для неоднорідних вибірок P-P діаграми зазвичай не дають змоги виявити більш-менш придатний для подальшої перевірки тип розподілу. Аналіз даних у цьому випадку доцільно розпочинати з побудови емпіричної функції розподілу й гістограми вибірки.

Як приклад, розглянемо результати єдиного державного екзамену з математики у Російській Федерації (ЄДЕ) у 2008 р. [14]. На рис. 2.23, 2.24 наведено функцію їх розподілу й відповідну гістограму. Гістограма свідчить про істотну неоднорідність досліджуваної вибірки. Тому далі розглядатимемо модель суміші розподілів.

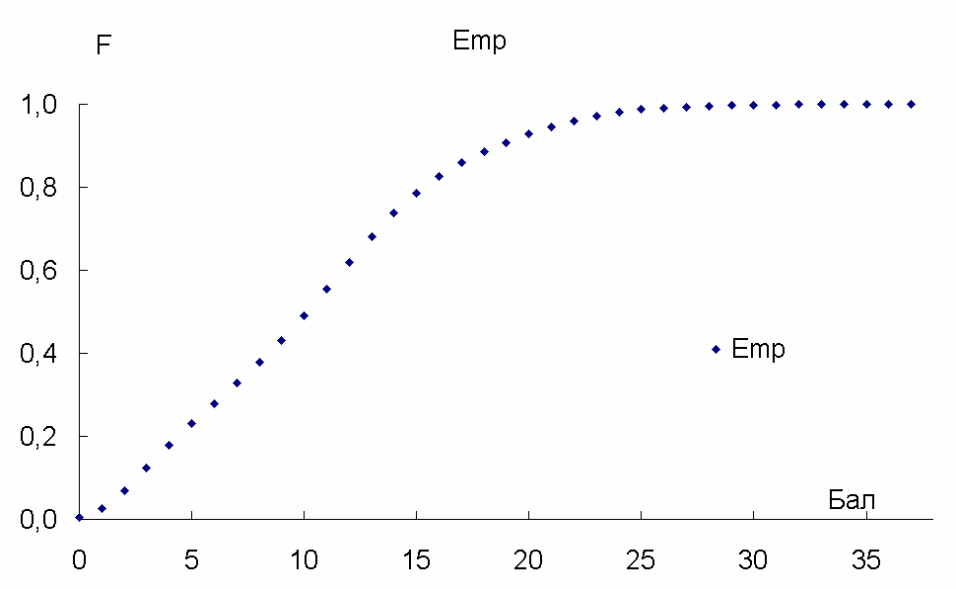


Рис. 2.23. Функція розподілу результатів ЄДЕ-2008 з математики

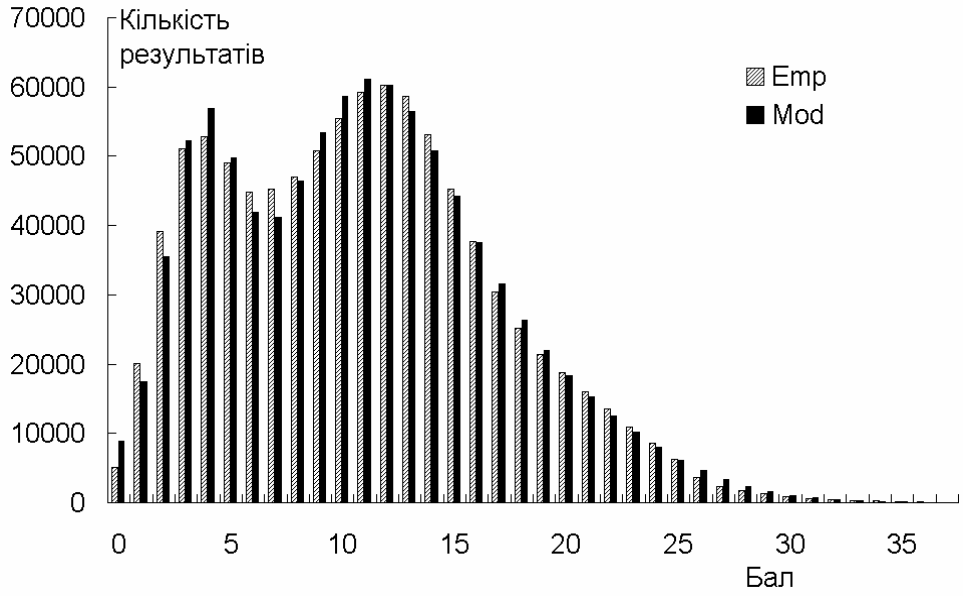


Рис. 2.24. Гістограма розподілу результатів ЄДЕ-2008 з математики

Шукатимемо модель у вигляді суміші нормально розподілених компонент:

$$F_M = \sum_i \alpha_i N(\mu_i, s_i), \tag{2.52}$$

де α_i – вагові коефіцієнти, що мають задовольняти вимогу $\sum_i \alpha_i = 1$, μ_i – математичні сподівання компонент, s_i – їх стандартні відхилення.

Для підбору параметрів моделі використовуємо процедуру "Пошук розв'язку" електронних таблиць MS Excel. Для моделей неоднорідного розподілу результат її застосування зазвичай є нестійким і залежить від вибору початкового наближення. У випадку, що розглядається, аналіз гістограми дає змогу достатньо точно визначити початкові наближення математичного сподівання і стандартного відхилення для двох компонент. Але перевірка адекватності отриманої після оптимізації моделі за критеріями Смірнова та χ^2, а також порівняння гістограми з моделлю показує, що модель не є адекватною. Тому було перевірено модель, яка містила три компоненти. Початкові значення параметрів третьої компоненти були визначені на основі аналізу різниці між емпіричною функцією розподілу і двохкомпонентною моделлю. Після оптимізації було отримано таку модель розподілу:

$$F_M = 0,204N(3,49;1,63) + 0,507N(10,60;3,82) + \\ + 0,288N(16,54;5,57) \tag{2.53}$$

Розрахункове значення критерію Смірнова дорівнює 0,03 і є значно меншим, ніж критичне. Емпірична й модельні гістограми розподілу достатньо добре відповідають одна одній (рис. 2.24). Також спостерігається добра відповідність між емпіричною й теоретичною функцією розподілу (рис. 2.25).

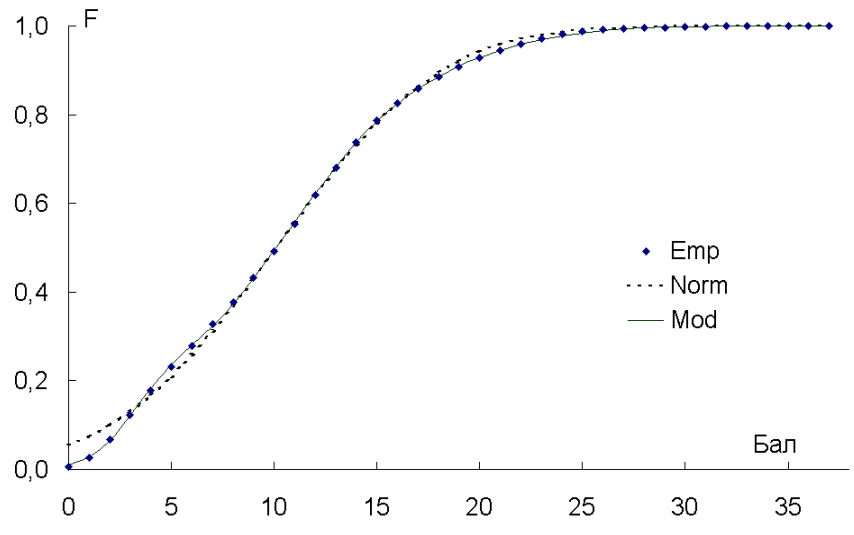


Рис. 2.25. Порівняння емпіричної та теоретичної функцій розподілу результатів єдиного державного екзамену з математики

У пакеті SPSS не передбачено спеціальних засобів для ідентифікації неоднорідних функцій розподілу. Для вирішення цієї проблеми можна використати засоби кластерного аналізу, які будуть докладно розглядатися у розділі 6.

Для ілюстрації розглянемо вибірку, побудовану, як суміш двох нормально розподілених компонент з математичними сподіваннями $a = 10$ й $b = 15$ або 20 та стандартними відхиленнями $s_{1,2} = 3$, обсягом по 100 елементів.

Уведемо досліджувані вибірки одна за одною до вікна даних пакету SPSS, як значення var0001 (рис. 2.26). Позначимо у сусідньому стовпчику (var0002), до якої з вихідних вибірок належить відповідне значення var0001. Це полегшує подальший аналіз результатів кластеризації.

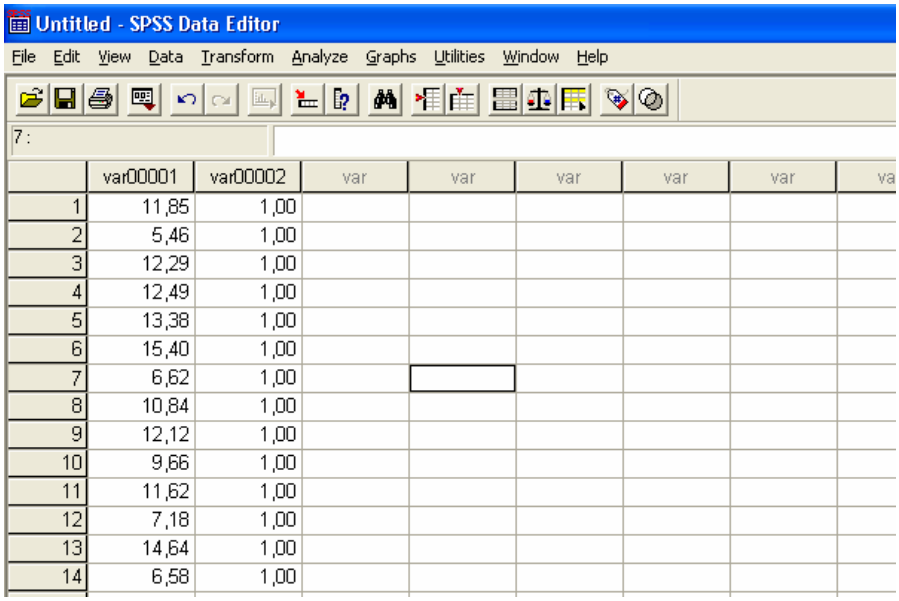


Рис. 2.26. Фрагмент вікна з вихідними даними

Далі у головному меню обираємо Analyze/Classify/K-means Cluster Analysis. При цьому з'являється діалогове вікно (рис. 2.27).

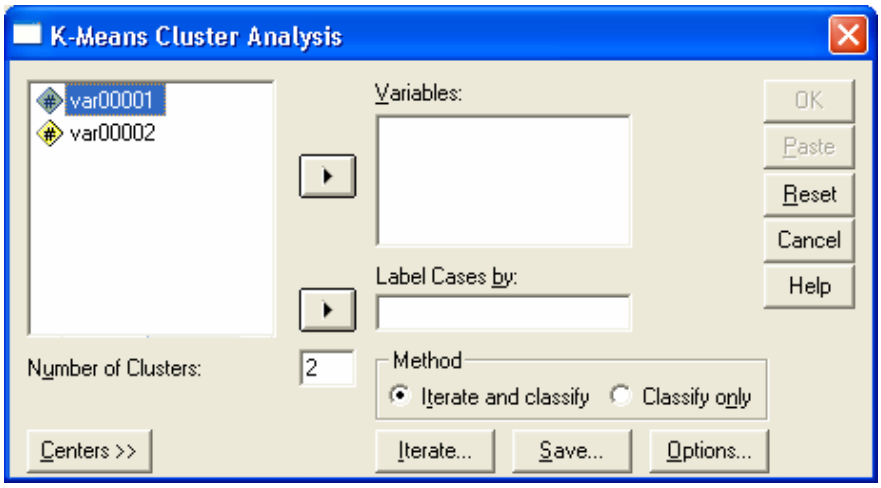


Рис. 2.27. Діалогове вікно кластерного аналізу

У цьому вікні зазначаємо змінну, значення якої необхідно поділити на кластери, метод класифікації, а також необхідність виведення окремих даних у файл або вікно результатів аналізу.

У випадку $b = 15$ отримано такі результати:

– кількість елементів першого кластера – 84, а другого – 116;

– центри кластерів 9,16 і 15,15, тобто похибки визначення центрів дорівнюють, відповідно, 8,4% й 1%;

– кількість помилково класифікованих значень першого кластера 27, другого кластера – 11;

– загальна імовірність помилкової класифікації – 19%.

У випадку $b = 20$ отримано такі результати:

– кількість елементів першого кластера – 102, а другого – 98;

– центри кластерів 10,20 і 20,45, тобто похибки визначення центрів дорівнюють, відповідно, 2% й 2,25%;

– кількість помилково класифікованих значень першого кластера – 2, другого кластера – 4;

– загальна імовірність помилкової класифікації – 3%.

Таким чином ми бачимо, що застосування цієї процедури дає прийнятні результати лише у випадку, коли кластери є достатньо добре відмежованими один від одного.

Після виокремлення кластерів для отримання більш повної інформації про їх параметри використаємо засоби побудови описової статистики. Для цього виберемо у головному меню Analyze/Descriptive Statistics/Frequencies. При цьому з'являється діалогове вікно побудови описової статистики (рис. 2.28).

У підменю Statistics (рис. 2.29) зазначимо, які саме параметри необхідно розрахувати. У підменю Charts (рис. 2.30) зазначимо, що необхідно побудувати гістограми розподілів й нанести на них графіки нормального розподілу.

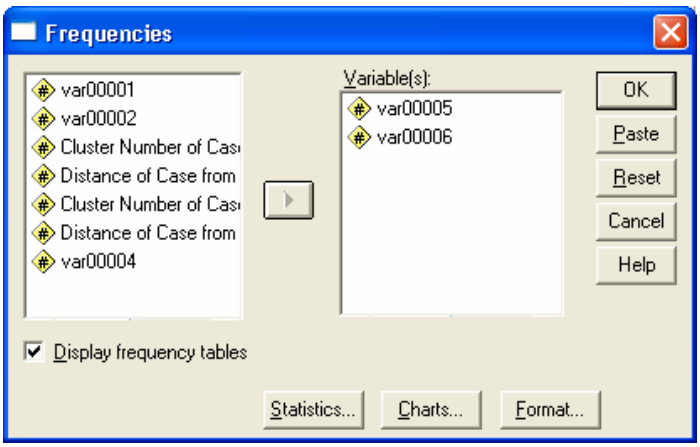


Рис. 2.28. Діалогове вікно побудови описової статистики

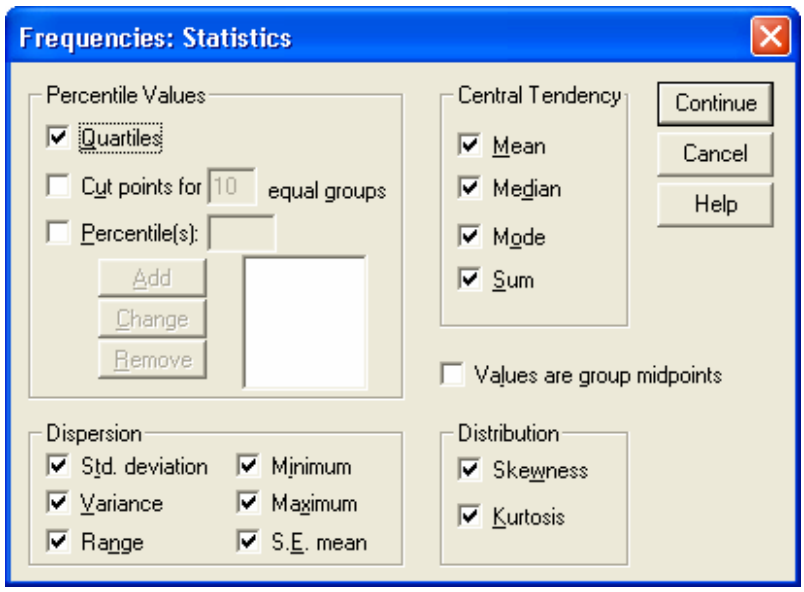


Рис. 2.29. Підменю для розрахунку параметрів описової статистики

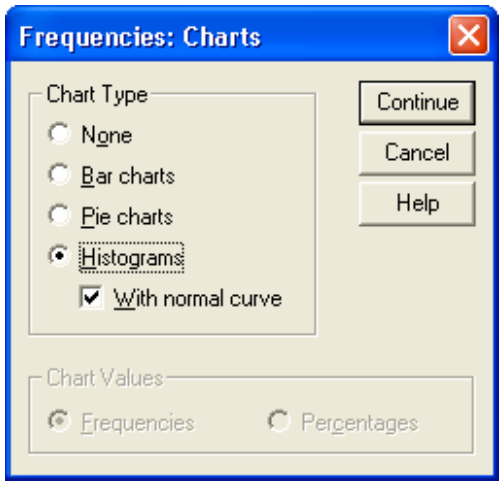


Рис. 2.30. Підменю Charts

На рис. 2.31 наведено вікно з результатами розрахунку описової статистики, а на рис. 2.32, 2.33 – гістограми розподілу частот для отриманих класів. Наведені дані свідчать, що математичні сподівання й стандартні відхилення є близькими до заданих параметрів, а розподіл значень у класах – до нормального.

На підставі отриманих результатів можна побудувати модель суміші розподілів у вигляді:

$$F(x) = 0,51N(10,2;2,86) + 0,49N(20,45;2,93), \qquad (2.54)$$

де $N(\mu;s)$ – функція нормального розподілу з математичним сподіванням μ та стандартним відхиленням s.

Значення вагових коефіцієнтів отриманої формули взято пропорційними чисельності отриманих класів з урахуванням умови $\alpha_1 + \alpha_2 = 1$.

Statistics

		VAR00005	VAR00006
N	Valid	98	102
	Missing	102	98
Mean		20,4493	10,1982
Std. Error of Mean		,29559	,28354
Median		20,7644	10,8251
Mode		16,49	2,27[a]
Std. Deviation		2,92618	2,86365
Variance		8,56251	8,20052
Skewness		,378	-,778
Std. Error of Skewness		,244	,239
Kurtosis		-,147	,381
Std. Error of Kurtosis		,483	,474
Range		13,74	12,70
Minimum		15,40	2,27
Maximum		29,14	14,97
Sum		2004,03	1040,21
Percentiles	25	18,0383	8,7034
	50	20,7644	10,8251
	75	22,1956	12,2346

a. Multiple modes exist. The smallest value is shown

Рис. 2.31. Результати розрахунку описової статистики

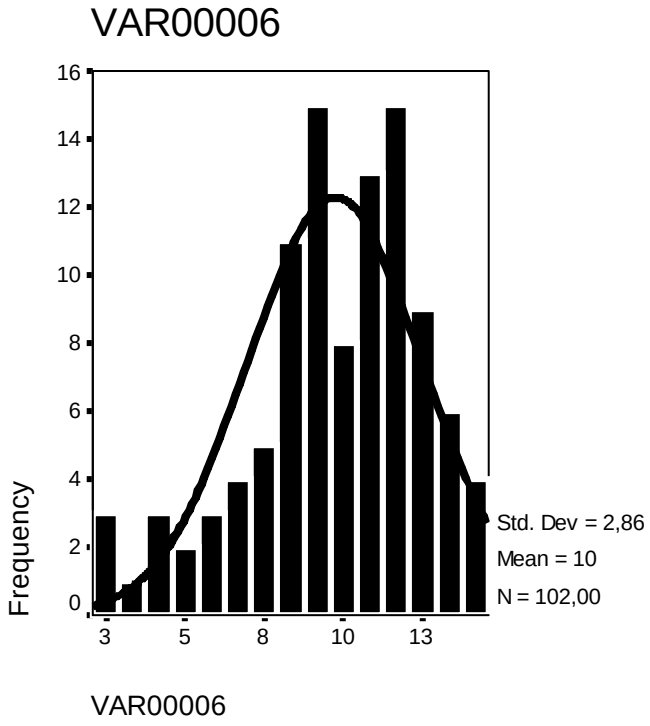


Рис. 2.32. Гістограма абсолютних частот для першого класу

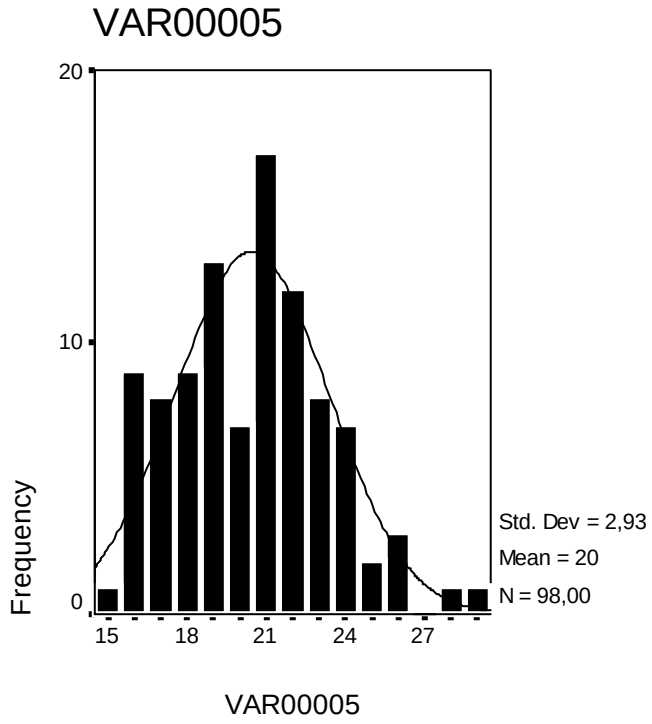


Рис. 2.33. Гістограма абсолютних частот для другого класу

Контрольні питання

1. З чого слід виходити при виборі методу перевірки статистичної гіпотези?

2. Які гіпотези називають нульовою та конкуруючою? Наведіть приклади. Скільки конкуруючих гіпотез можна сформулювати для однієї й тієї самої нульової гіпотези?

3. Які гіпотези називають простими та складними? Наведіть приклади.

4. Що називають статистичним критерієм? Які основні типи критеріїв використовують для перевірки гіпотез?

5. У чому полягає різниця між однобічними та двобічними критеріями?

6. Які величини називають рівнем значущості й довірчим рівнем? Для чого використовують ці параметри?

7. Які типи помилок розглядають при перевірці статистичних гіпотез? Як вони пов’язані між собою?

8. Що називають потужністю критерію? Для чого використовують цей параметр?

9. Як можна підвищити потужність критерію?

10. Якою є загальна методика перевірки статистичних гіпотез?

11. Які вибірки називають незалежними та спряженими? Наведіть приклади.

12. Які критерії називають параметричними? У яких випадках доцільно використовувати параметричні критерії?

13. Які гіпотези можна перевіряти за допомогою *Z*-критерію? Якими є умови правомірності його застосування?

14. Які гіпотези можна перевіряти за допомогою *t*-критерію Стьюдента? Якими є умови правомірності його застосування?

15. Які гіпотези можна перевіряти за допомогою критерію Уелча? Якими є умови правомірності його застосування?

16. Які гіпотези можна перевіряти за допомогою *F*-критерію Фішера? Якими є умови правомірності його застосування?

17. Які гіпотези можна перевіряти за допомогою критерію рандомізації компонент Фішера? Якими є умови правомірності його застосування?

18. Які гіпотези можна перевіряти за допомогою *W*-критерію Уїлкоксона? Якими є умови правомірності його застосування?

19. Які гіпотези можна перевіряти за допомогою *U*-критерію Манна – Уїтні? Якими є умови правомірності його застосування?

20. Які гіпотези можна перевіряти за допомогою *T*-критерію Уїлкоксона? Якими є умови правомірності його застосування?

21. Які гіпотези можна перевіряти за допомогою критерію χ^2? Якими є умови правомірності його застосування?

22. Які гіпотези можна перевіряти за допомогою критерію серій Вальда – Волфовиця? Якими є умови правомірності його застосування?

23. Які гіпотези можна перевіряти за допомогою критерію знаків? Якими є умови правомірності його застосування?

24. Які критерії можна використовувати для перевірки нормальності розподілу даних? Охарактеризуйте їх переваги й недоліки.

25. Яку гіпотезу можна перевіряти за допомогою ω^2 критерію Мізеса? Якими є умови правомірності його застосування?

26. Яку гіпотезу можна перевіряти за допомогою критерію Смірнова? Якими є умови правомірності його застосування?

27. Які гіпотези можна перевіряти за допомогою критерію χ^2? Якими є умови правомірності його застосування?

28. Яку гіпотезу можна перевіряти за допомогою *W*-критерію Шапіро – Уїлка? Якими є умови правомірності його застосування?

3. ДИСПЕРСІЙНИЙ АНАЛІЗ

Дисперсійний аналіз є сукупністю статистичних методів, призначених для перевірки гіпотез про зв'язок між певною ознакою та досліджуваними факторами, які не мають кількісного опису, а також для встановлення ступеня впливу факторів та їх взаємодії. У спеціальній літературі дисперсійний аналіз часто називають ANOVA (від англомовної назви Analysis of Variations). Вперше цей метод було розроблено Р. Фішером в 1925 р.

Факторами називають контрольовані чинники, що впливають на кінцевий результат. **Рівнем фактора**, або **способом обробки**, називають значення, що характеризують конкретний прояв цього фактора. Ці значення зазвичай подають у номінальній або порядковій шкалі вимірювань. Значення вимірюваної ознаки називають **відгуком**.

Часто вихідні значення факторів вимірюють у кількісних або порядкових шкалах. Тоді постає проблема групування вихідних даних у ряди спостережень, що відповідають приблизно однаковим значенням фактора. Якщо кількість груп взяти надмірно великою, то кількість спостережень у них може виявитися недостатньою для отримання надійних результатів. Якщо її взяти надмірно малою, це може призвести до втрати суттєвих особливостей впливу досліджуваного фактора на систему. Загальну методологію групування описано в розділі 1. Вибір конкретного способу групування даних залежить від їх обсягу і характеру варіювання значень фактора.

Кількість і розміри інтервалів при однофакторному аналізі найчастіше визначають за принципом рівних інтервалів або за принципом рівних частот. При багатофакторному аналізі застосовують три типи групування:

- групи з рівною кількістю спостережень;
- групи з різною кількістю спостережень;
- групи, кількості спостережень у яких відповідають певній пропорції.

При цьому існують певні особливості обробки даних, залежно від типу групування, які не розглядаються у цьому посібнику.

3.1. Однофакторний аналіз

Основною метою однофакторного аналізу зазвичай є оцінка величини впливу конкретного фактора на досліджуваний відгук. Іншою метою може бути порівняння двох або декількох факторів один з одним з метою визначення різниці їх впливу на відгук, яку часто називають **контрастом факторів**. Попереднім етапом є перевірка нульової гіпотези про відсутність будь-якого впливу досліджуваного фактора (факторів), тобто гіпотези про те, що зміни значень ознаки в порівнюваних вибірках є випадковими, і всі дані належать до однієї генеральної сукупності.

Якщо нульову гіпотезу відкидають, то наступним етапом є кількісне оцінювання впливу досліджуваного фактора і побудова довірчих інтервалів для отриманих характеристик. У випадку, 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 його маскувати.

При **однофакторному дисперсійному аналізі** вихідні дані подають у вигляді таблиць, у 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. При цьому виходять з припущення, що результати спостережень для різних рівнів є вибірками з нормально розподілених сукупностей, середні значення та дисперсії яких є однаковими і не залежать від рівнів. Завданням аналізу є перевірка нульової гіпотези про рівність середніх значень сукупностей, що розглядаються.

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	$x_{n_1 1}$	$x_{n_1 2}$	…	$x_{n_1 k}$

Метод базується на основній тотожності дисперсійного аналізу, згідно з якою сума квадратів відхилень спостережень від загального середнього **(загальна варіація**) дорівнює:

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij}-\bar{x}\right)^2=\sum_{j=1}^{k}n_j\left(\left\langle x_j\right\rangle-\bar{x}\right)^2+\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij}-\left\langle x_j\right\rangle\right)^2, \qquad (3.1)$$

де $\bar{x}=\frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{n_j}x_{ij}$ – загальне середнє; $\left\langle x_j\right\rangle=\frac{1}{n_j}\sum_{i=1}^{n_j}x_{ij}\quad j=1,\ ...,\ k$;

$N=\sum_{j=1}^{k}n_j$ – загальна чисельність; k – кількість вибірок; n_j (j = 1, 2, …, k) – кількість елементів у j-й вибірці; $\left\langle x_j\right\rangle$ – середнє значення j-ї вибірки.

У правій частині (3.1) перший доданок (**факторна**, або **міжгрупова варіація**) є зваженою сумою квадратів відхилень групових середніх від

загального середнього. Він характеризує коливання значень, зумовлені фактором, на основі якого здійснено групування даних. Другий доданок (**залишкова**, або **внутрішньогрупова варіація**) є сумою квадратів відхилень спостережень від відповідних групових середніх. Він характеризує коливання значень досліджуваної ознаки, зумовлені неврахованими факторами або випадковими чинниками.

Сутність методу полягає в тому, що за умови правильності нульової гіпотези величини

$$\sigma_1^2 = \frac{1}{N-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij} - \left\langle x_j \right\rangle\right)^2 \tag{3.2}$$

та

$$\sigma_2^2 = \frac{1}{k-1}\sum_{j=1}^{k} n_j \left(\left\langle x_j \right\rangle - \bar{x}\right)^2 \tag{3.3}$$

є незміщеними оцінками дисперсії похибок спостережень σ^2 і мають бути приблизно рівними одна одній. Перша з них 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. Друга оцінка характеризує варіацію між вибірками. При справедливості нульової гіпотези $\sigma_2^2 \approx \sigma^2$, а при її порушенні неї величина σ_2^2 є тим більшою, чим більше відхилення від неї.

Значення критерію розраховують за формулою:

$$F = \frac{(N-k)\sum_{j=1}^{k} n_j \left(\left\langle x_j \right\rangle - \bar{x}\right)^2}{(k-1)\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij} - \left\langle x_j \right\rangle\right)^2}. \tag{3.4}$$

Ця величина має *F*-розподіл Фішера з параметрами $k-1$ та $N-k$. Нульову гіпотезу відхиляють, якщо ймовірність $P\left(F \geq F^*\right)$, де F^* – значення, розраховане за емпіричними даними за формулою (3.3) є достатньо малою.

Непараметричним аналогом однофакторного дисперсійного аналізу є **ранговий однофакторний аналіз Краскела – Уолліса**. Він розроблений американськими математиком Вільямом Краскелом та економістом Вільсоном Уоллісом в 1952 р. Цей критерій призначено 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. Нульова гіпотеза полягає в тому, що *k* вибірок обсягами $n_1, n_2, ..., n_k$ отримані з однієї і тієї самої генеральної сукупності. Критерій

Краскела – Уолліса є узагальненням *U*-критерію Манна – Уїтні на випадок, коли кількість вибірок $k > 2$.

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}. При цьому одержують табл. 3.2.

Таблиця 3.2

Загальний вигляд вихідної таблиці рангового однофакторного аналізу

№ результату	№ вибірки			
	1	2	…	k
1	r_{11}	r_{12}	…	r_{1k}
2	r_{21}	r_{22}	…	r_{2k}
…	…	…	…	…
n_i	$r_{n_1 1}$	$r_{n_1 2}$	…	$r_{n_1 k}$

Для кожного рівня фактора, тобто для кожного стовпця, розраховують суму рангів $R_j = \sum_{i=1}^{n_j} r_{ij}$ або відповідні середні ранги $\langle R_j \rangle = \frac{1}{n_j} \sum_{i=1}^{n_j} r_{ij}$.

Для контролю можна використовувати тотожність:

$$\sum_{i=1}^{k} R_i = \frac{N(N+1)}{2}, \tag{3.5}$$

де $N = \sum_{i=1}^{k} n_i$ – загальна чисельність.

Якщо між стовпцями немає систематичної різниці, то останні будуть близькими до середнього рангу, розрахованого за усією сукупністю, який дорівнює $(N+1)/2$. Тому величини $\langle R_j \rangle - (N+1)/2$ мають бути відносно малими, якщо нульова гіпотеза є правильною.

Обчислення критерію здійснюють за формулами:

$$\mathrm{H} = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1), \tag{3.6}$$

або

$$\mathrm{H} = \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j \left(\langle R_j \rangle - \frac{N+1}{2} \right)^2. \tag{3.7}$$

За $n_i \geq 5$ й $k \geq 4$ статистика критерію асимптотично наближається до χ^2-розпоїлу з кількістю степенів вільності $k-1$. У цьому випадку 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 $k-1$ степенем вільності. При $k=2$ статистика Краскела – Уолліса стає еквівалентною статистиці W Уїлкоксона.

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 у випадку, коли вони не є цілими числами, їх не округляють. Якщо кількість збігів велика, використовують модифіковану форму статистики Краскела – Уолліса:

$$H' = \frac{H}{1 - \left(\sum_{j=1}^{g} \frac{T_j}{N^3 - N} \right)}, \tag{3.8}$$

де g – кількість груп спостережень, що збігаються;

$T_j = \left(t_j^3 - t_j \right)$; t_j – кількість спостережень, що збігаються в j-ї групі.

Критерій Джонкхієра (**Джонкхієра – Терпстра**) запропонований незалежно один від одного нідерландським математиком Т.Дж. Терпстрою в 1952 р. й британським психологом Е.Р. Джонкхієром в 1954 р. Його застосовують тоді, коли заздалегідь відомо, що наявні групи результатів упорядковані за зростанням впливу досліджуваного фактора, який вимірюють у 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 тощо, останній стовпчик відповідає найбільшому рівню. При виконанні таких припущень критерій Джонкхієра є більш потужним, ніж критерій Краскела – Уолліса, стосовно гіпотези про монотонний вплив фактора.

Спочатку для кожної пари вибірок з номерами u, v $\left(1 \leq u < v \leq k\right)$, де k – кількість рівнів фактора, розраховують статистику Манна – Уїтні:

$$U_{u,v} = \sum_{\substack{i=1,\ldots,n_u \\ j=1,\ldots,n_v}} \varphi\left(x_{iu}, y_{jv} \right), \tag{3.9}$$

де

$$\varphi\left(x_i, y_j \right) = \begin{cases} 1 & \left(x_i < y_j \right); \\ 1/2 & \left(x_i = y_j \right); \\ 0 & \left(x_i > y_j \right). \end{cases} \tag{3.10}$$

Потім розраховують статистику Джонкхієра:

$$J = \sum_{1 \le u \le v \le k} U_{u,v} \,. \tag{3.11}$$

Великі значення J свідчать проти гіпотези про однорідність вибірок.

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

$$MJ = \frac{1}{4}\left(N^2 - \sum_{j=1}^{k} n_j^2 \right); \quad DJ = \frac{1}{72}\left[N^2 \left(2N+3\right) - \sum_{j=1}^{k} n_j^2 \left(2n_j + 3\right) \right]. \tag{3.12}$$

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

***M*-критерій Бартлетта** запропонований британським статистиком Маурісом Стівенсоном Бартлеттом в 1937 р. Його застосовують для перевірки нульової гіпотези про рівність дисперсій кількох нормальних генеральних сукупностей, з яких взяті досліджувані вибірки, що у загальному випадку мають різні обсяги (обсяг кожної вибірки має бути не менше чотирьох). Обчислення критерію здійснюють за формулою:

$$B = V / C, \tag{3.13}$$

де

$$V = k \ln \overline{s^2} - \sum_{i=1}^{\ell} k_i \ln s_i^2;$$

$$C = 1 + \frac{1}{3\left(\ell - 1\right)}\left[\sum_{i=1}^{\ell} \frac{1}{k_i} - \frac{1}{k} \right],$$

$\overline{s^2} = \dfrac{\sum_{i=1}^{\ell} k_i s_i^2}{k}$ – зважене за кількістю степенів вільності середнє арифметичне стандартних відхилень вибірок; $k = \sum_{i=1}^{\ell} k_i$ – загальна чисельність; k_i – чисельність i-ї вибірки; ℓ – кількість вибірок; s_i^2 – стандартне відхилення i-ї вибірки. При великих n_j статистика критерію асимптотично наближається до χ^2-розпоілу з кількістю ступенів вільності $\ell - 1$.

***G*-критерій Кокрена (Кочрена)** запропонований американським статистиком Вільмом Геммелом Кочреном в 1941 р. Його 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:

$$G = \frac{\max\limits_{1 \leq j \leq k} \sigma_j^2}{\sum\limits_{j=1}^{k} \sigma_j^2}, \tag{3.14}$$

де σ_j^2 – дисперсія j-ї вибірки. Для вибірок рівного обсягу він є потужнішим за критерій Бартлетта. Критичні точки критерію Кокрена визначають за спеціальними таблицями.

Непараметричний **критерій Левене**, запропонований американським математиком Ховардом Левене в 1960 р. є альтернативою критерію Бартлетта в умовах, коли немає впевненості у тому, що досліджувані вибірки підпорядковуються нормальному розподілу. Розрахункове значення критерію обчислюють за формулою:

$$W = \frac{(N-k)\sum\limits_{j=1}^{k} n_j \left(Z_j^* - \overline{Z}\right)^2}{(k-1)\sum\limits_{j=1}^{k}\sum\limits_{i=1}^{n_j} n_j \left(Z_{ij} - \overline{Z}_j\right)^2}, \tag{3.15}$$

де $Z_{ij} = \left| x_{ij} - \overline{x}_j \right|$; x_{ij} – значення i-го спостереження в j-ої вибірці; $\overline{x}_j$ – середнє арифметичне спостережень, що потрапили до j-ої вибірки; $\overline{Z}$ – загальне середнє арифметичне значень Z за усіма спостереженнями; $\overline{Z}_j$ – середнє арифметичне Z за спостереженнями, що потрапили до j-ої вибірки. Розрахункове значення критерію порівнюють з відповідним квантилем F-розподілу з кількостями степенів вільності $(k - 1)$ та $(N - k)$.

В 1974 р. американські статистики Мортон Б. Браун та Алан Б. Форсайт запропонували більш робастний тест (**критерій Брауна – Форсайта**), який відрізняється від критерію Левене тим, що значення $Z_{ij} = \left| x_{ij} - \tilde{x}_j \right|$, де $\tilde{x}_j$ – медіана спостережень, що потрапили до j-ої 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:

$$x_{ij} = a_j + \varepsilon_{ij} \quad (j = 1, ..., k;\ i = 1, ..., n), \tag{3.16}$$

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 (3.16) можна записати у вигляді:

$$x_{ij} = \mu + \tau_j + \varepsilon_{ij} \quad (j = 1, ..., k;\ i = 1, ..., n), \tag{3.17}$$

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 залишилася рівною k, оскільки відхилення пов'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**:

$$z_{ij} = med\left\{x_{ui} - x_{vj}\right\} \quad \left(u = 1, ..., n_i;\ v = 1, ..., n_j\right). \tag{3.18}$$

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**):

$$W_{ij} = \overline{\Delta_i} - \overline{\Delta_j}, \tag{3.19}$$

де величини

$$\overline{\Delta_i} = \frac{\sum_{u=1}^{k} n_u z_{iu}}{N} \quad \left(i = 1, ..., k\right) \tag{3.20}$$

відображають зсув i-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.

Якщо гіпотезу про рівність середніх відхиляють, то наступним кроком може бути визначення вибірок, для яких ця різниця є суттєвою. Для цього використовують метод лінійних контрастів.

Лінійним контрастом у моделі адитивного впливу фактора на відгук називають лінійну функцію середніх значень k незалежних нормальних вибірок з невідомими рівними дисперсіями:

$$L = \sum_{j=1}^{k} c_j m_j \,, \tag{3.21}$$

де c_i – відомі сталі, які задовольняють вимогу $\sum_{i=1}^{k} c_i = 0$; m_j – математичне сподівання для j-ої вибірки, яке для оцінювання лінійного контрасту замінюють відповідним середнім арифметичним.

Оцінку дисперсії лінійного контрасту розраховують за формулою:

$$s_L^2 = \sigma_1^2 \sum_{j=1}^{k} \frac{c_j^2}{n_j}, \tag{3.22}$$

де σ_1^2 визначається формулою (3.2).

Довірчим інтервалом для лінійного контрасту є:

$$L \pm s_L \sqrt{(k-1) F_{1-\alpha}(k-1;\, N-k)}\,. \tag{3.23}$$

Найпростішим прикладом лінійних контрастів є різниці $m_i - m_j$, яким відповідають значення: $c_i = 1$, $c_j = -1$, $c_u = 0$ при всіх $u \neq i, j$. Нульові гіпотези, що перевіряються, полягають у тому, що $m_i = m_j$ для всіх можливих пар вибірок. Їх приймають, якщо нульове значення потрапляє до відповідного довірчого інтервалу.

Для встановлення вибірок, що належать певній множині даних, дисперсії яких є різними, найчастіше застосовують **метод множинних порівнянь** (**Шеффе**), запропонований американським статистиком Генрі Шеффе. Обчислення критерію при перевірці нульової гіпотези $L = L_0$ здійснюють за формулою:

$$t = \frac{\sum_{i=1}^{k} c_i \overline{x_i} - L_0}{\sqrt{M \sum_{i=1}^{k} \frac{c_i^2}{n_i}}}, \tag{3.24}$$

де N – загальна кількість; n_i – кількість елементів в i-ї вибірці;

$\overline{x_i}$ – середнє значення для i-ї вибірки;

$M = \frac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij} - \overline{x_i}\right)^2$ – середній квадратичний залишок. Розрахо-

ване значення критерію для порівняння з критичним необхідно брати за модулем.

3.2. Двофакторний аналіз

Двофакторний дисперсійний аналіз застосовують для пов'язаних нормально розподілених вибірок. Дані подають у вигляді табл. 3.3, у стовпчиках якої наводять дані, що відповідають певному рівню першого фактора, а в 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 3.3

Таблиця даних двофакторного дисперсійного аналізу

Рівні фактора В	Рівні фактора А			
	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 другого фактора є суттєвим. На практиці часто використовують і складніші таблиці двофакторного дисперсійного аналізу, зокрема такі, у яких кожна комірка містить набір даних (повторні вимірювання), що відповідають фіксованим значенням рівнів обох факторів.

Для опису даних табл. 3.3 часто можна застосовувати адитивну модель, яка передбачає, що значення відгуку є сумою внесків окремо кожного із факторів b_i і t_j, а також незалежної від факторів випадкової компоненти ε_{ij}:

$$x_{ij} = b_i + t_j + \varepsilon_{ij}. \qquad (3.25)$$

На практиці модель (3.24) часто подають в еквівалентному вигляді:

$$x_{ij} = \mu + \beta_i + \tau_j + \varepsilon_{ij}, \qquad (3.26)$$

де $\mu = \frac{1}{kn}\sum_{j=1}^{k} x_{ij}$ – загальне середнє за всіма спостереженнями;

β_i і τ_j – відхилення від середнього, зумовлені факторами А і В, відповідно.

У випадку, коли випадкова компонента ε_{ij} підпорядковується нормальному розподілу з нульовим середнім і рівними для всіх i, j дисперсіями σ^2 застосовують **двофакторний дисперсійний аналіз** (**дисперсійний аналіз за двома ознаками**).

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 $H_{02}: \beta_1 = \beta_2 = ... = \beta_n = 0$, тобто в першому випадку припускають відсутність впливу фактора А, а у другому – фактора В.

Як і у випадку однофакторного дисперсійного аналізу в цьому разі розраховують дві оцінки дисперсії. При перевірці гіпотези H_{01} першу з них розраховують за формулою:

$$\sigma_1^2 = \frac{1}{(n-1)(k-1)} \sum_{i=1}^{n} \sum_{j=1}^{k} \left(x_{ij} - \langle x_i \rangle - \langle x_j \rangle + \langle x \rangle \right)^2, \qquad (3.27)$$

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 *i*-м рядком;

$\langle x_j \rangle = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$ – середнє за *j*-м стовпчиком. Оцінка дисперсії σ_1^2 є незміщеною і не залежить від справедливості нульової гіпотези.

Другу оцінку розраховують за формулою:

$$\sigma_2^2 = \frac{n}{k-1} \sum_{j=1}^{k} \left(\langle x_j \rangle - \langle x \rangle \right)^2. \qquad (3.28)$$

Вона є незміщеною лише за умови справедливості нульової гіпотези. Що більшою є різниця між результатами дії фактора А, то більшим є значення, розраховане за формулою (3.28).

Для перевірки справедливості гіпотези H_{01} необхідно розрахувати відношення дисперсій:

$$F = \frac{\sigma_2^2}{\sigma_1^2} = \frac{n(n-1)(k-1) \sum_{j=1}^{k} \left(\langle x_j \rangle - \langle x \rangle \right)^2}{(k-1) \sum_{i=1}^{n} \sum_{j=1}^{k} \left(x_{ij} - \langle x_i \rangle - \langle x_j \rangle + \langle x \rangle \right)^2}. \qquad (3.29)$$

Воно має *F*-розподіл Фішера з кількостями степенів вільності $(k-1)$ і $(n-1)(k-1)$. Нульову гіпотезу приймають на рівні значущості α, якщо $F < F_{1-\alpha}$, де $F_{1-\alpha}$ – α-квантиль *F*-розподілу з відповідними кількостями степенів вільності.

Для перевірки нульової гіпотези H_{02} можна використовувати формулу (3.28), в якій необхідно попарно поміняти місцями величини n і k, а також i та j.

Якщо припущення, необхідні для застосування двофакторного дисперсійного аналізу, не виконуються, то використовують непараметричний **ранговий критерій Фрідмана** (**Фрідмана, Кендалла та Сміта**), розроблений американським економістом Мілтоном Фридманом наприкінці 1930 р. Цей критерій не залежить від типу розподілу. Передбачається 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 якій n рядків відповідають 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 випадку 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 i, отримуючи при цьому k значень відповідних рангів r_{ij}. Це дає можливість усунути вплив фактора В, значення якого для кожного рядка є однаковим.

Обчислення критерію здійснюють за формулою:

$$S = \left[\frac{12}{nk(k+1)} \sum_{j=1}^{k} \left(\sum_{i=1}^{n} r_{ij} \right)^2 \right] - 3n(k+1). \qquad (3.30)$$

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

При справедливості нульової гіпотези і $n \to \infty$ S-статистика Фрідмана 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), запропонований американським статистиком Є.Б. Пейджем в 1963 р., призначений для перевірки нульової гіпотези $H_{01}: \tau_1 = \tau_2 = ... = \tau_k = 0$ або $H_{02}: \beta_1 = \beta_2 = ... = \beta_n = 0$, проти альтернатив $\tau_1 \le \tau_2 \le ... \le \tau_k$ 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 Фрідмана. Значення критерію обчислюють за формулами:

$$L_1 = \sum_{j=1}^{k} jr_j \; ; \; L_2 = \sum_{i=1}^{n} ir_i \, , \tag{3.31}$$

де $r_j = \sum_{i=1}^{n} r_{ij}$, $r_i = \sum_{j=1}^{k} r_{ij}$.

Для великих вибірок застосовують апроксимацію статистики Пейджа:

$$L_1^* = \frac{L - nk\left(k+1\right)^2/4}{\sqrt{n\left(k^3-k\right)^2/144\left(k-1\right)}}; \quad L_2^* = \frac{L - nk\left(n+1\right)^2/4}{\sqrt{k\left(n^3-n\right)^2/144\left(n-1\right)}}, \tag{3.32}$$

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** запропонований В. Кочреном в 1937 р. Його 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:

$$Q = \frac{\left(c-1\right)\left(c\sum_{j=1}^{c} T_j^2 - \left(\sum_{j=1}^{c} T_j\right)^2\right)}{c\sum_{i=1}^{r} T_i - \sum_{i=1}^{r} T_i^2}, \tag{3.33}$$

де $T_j = \sum_{i=1}^{r} x_{ij}$ $\left(j = 1, 2, ..., c\right)$ – суми стовпців; $T_i = \sum_{j=1}^{c} x_{ij}$ $\left(i = 1, 2, ..., r\right)$ – суми рядків; c – кількість стовпців (вибірок); r – кількість рядків (параметрів). Довірчий рівень визначається функцією розподілу χ^2 з кількістю степенів вільності, яка дорівнює $c - 1$.

Двофакторний дисперсійний аналіз дає можливість визначити існування ефектів обробки, проте не дає змоги встановити, для яких саме стовпців існує цей ефект.

При вирішенні цієї проблеми застосовують метод множинних порівнянь Шеффе для пов'язаних вибірок. Значення критерію розраховують за формулою:

$$t = \frac{\left(\sum_{i=1}^{r} c_i \overline{x_i}\right)^2}{\frac{(r-1)S}{c}\sum_{i=1}^{r} c_i^2}, \qquad (3.34)$$

де c_i $(i = 1, 2, ..., r)$ – константи;

$S = \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}^2 - \frac{T^2}{rc}$ – 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. Приклад виконання дисперсійного аналізу

За допомогою вбудованого генератора випадкових чисел електронних таблиць MS Excel сформуємо чотири нормально розподілених вибірки обсягом по 200 елементів з параметрами: $\overline{x_1} = 50;\ s_1 = 5;\ \overline{x_2} = 49;\ s_2 = 5;\ \overline{x_3} = 51;\ s_3 = 6;\ \overline{x_4} = 49{,}5;\ s_4 = 5{,}5$.

На першому етапі перевіряємо відповідність вибірок нормальному закону розподілу. Оскільки приклади такої перевірки було розглянуто в попередньому розділі й ураховуючи, що ми будували саме нормально розподілені послідовності, будемо вважати, що вони задовольняють цю вимогу. Тому можемо використовувати параметричні методи аналізу.

Спочатку перевіримо нульову гіпотезу про рівність середніх значень досліджуваних вибірок При розрахунку безпосередньо за формулою (3.4) маємо: $n_1 = n_2 = n_3 = n_4 = 200$; $N = 800$; $k = 4$; $F = 1{,}544$. Кількості степенів вільності $k - 1 = 3$; $N - k = 796$. За допомогою функції =FРАСПОБР (0,05;3;796), що знаходиться у бібліотеці статистичних функцій електронних таблиць MS Excel, визначаємо критичне значення для рівня значущості 0,05 $F_{кр} = 2{,}62$. Бачимо, що розрахункове значення критерію є меншим, ніж критичне. За допомогою статистичної функції = FРАСП (F7;3;796) ми також можемо визначити рівень значущості, якому відповідає розрахункове значення критерію $F^* = 0{,}202$. Бачимо, що імовірність

припущення помилки першого роду при відхиленні нульової гіпотези є досить високою. Тому приймаємо нульову гіпотезу про рівність середніх значень.

В пакеті аналізу електронних таблиць MS Excel є вбудована процедура “Однофакторний дисперсійний аналіз”. На рис. 3.1 показано результати її застосування, для згенерованих вибірок.

Однофакторный дисперсионный анализ

ИТОГИ

Группы	*Счет*	*Сумма*	*Среднее*	*Дисперсия*
Столбец 1	200	10068,6	50,343	26,802066
Столбец 2	200	9874,115	49,37058	23,848429
Столбец 3	200	9974,982	49,87491	32,535911
Столбец 4	200	9877,329	49,38664	27,908546

Дисперсионный анализ

Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Между группами	128,6168	3	42,87225	1,5436256	0,20180279	2,616088979
Внутри групп	22107,9	796	27,77374			
Итого	22236,51	799				

Рис. 3.1. Вікно виведення результатів однофакторного дисперсійного аналізу MS Excel

Крім визначених раніше параметрів, таблиця результатів застосування процедури однофакторного дисперсійного аналізу містить також середні арифметичні й дисперсії вибірок, значення міжгрупової та внутрішньогрупової дисперсії (MS) й відповідних варіацій (SS), які збігаються з величинами, обчислюваними за формулами (3.1–3.3), та деякі інші дані.

Для визначення однорідності вибірок нам необхідно перевірити також нульову гіпотезу про рівність їх дисперсій. Оскільки у нашому випадку обсяги вибірок є рівними, цю гіпотезу перевірятимемо за критерієм Кокрена. В електронних таблицях MS Excel немає вбудованих засобів для розрахунку критеріїв Бартлетта й Кокрена. Тому розрахункове значення обчислюємо за формулою (3.14). Воно дорівнює $G = 0,293$.

Критичне значення критерію Кокрена для рівня значущості 0,05 можна визначити з відповідних таблиць. Для випадку, що розглядається $G_c < 0,05$ (при тих самих умовах $G_c \approx 0,0495$, якщо обсяги вибірок дорівнюють 120), тобто є значно меншим, ніж розрахункове значення. Тому ми маємо відхилити нульову гіпотезу про рівність дисперсій і зробити висновок, що досліджувані вибірки не є однорідними.

У пакеті SPSS також передбачено можливість здійснення однофакторного дисперсійного аналізу. Для цього необхідно використовувати вікно Analyze/Compare means/One way ANOVA. Вихідні вибірки розміщуємо по-

слідовно одна за одною в одному й тому самому стовпчику, як значення змінної VAR00001. Як значення змінної VAR00002 беремо номери вибірок, до яких належить відповідне значення VAR00001. Додатково ми можемо отримати основні показники описової статистики для кожної з вибірок та усієї сукупності в цілому; перевірити однорідність дисперсій (тест Левена); визначити, які саме вибірки істотно відрізняються від інших (тест Дункана), й отримати іншу корисну інформацію. Якщо для вибірок, що розглядаються у прикладі, замовити у пункті меню "Options" розрахунок параметрів описової статистики й перевірку однорідності дисперсій, а пункті "Post Hoc" – виконання тесту Дункана, одержимо результати, показані на рис. 3.2.

Descriptives

VAR00001

		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum	Between-Component Variance
						Lower Bound	Upper Bound			
1,00		200	50,3430	5,17707	,36607	49,6211	51,0649	36,15	64,18	
2,00		200	49,3706	4,88349	,34531	48,6896	50,0515	36,24	60,87	
3,00		200	49,8749	5,70403	,40334	49,0796	50,6703	34,38	62,45	
4,00		200	49,3866	5,28285	,37355	48,6500	50,1233	35,89	64,92	
Total		800	49,7438	5,27546	,18652	49,3777	50,1099	34,38	64,92	
Model	Fixed Effects			5,27008	,18633	49,3780	50,1095			
	Random Effects				,23150	49,0071	50,4805			,07549

Test of Homogeneity of Variances

VAR00001

Levene Statistic	df1	df2	Sig.
1,132	3	796	,335

ANOVA

VAR00001

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	128,617	3	42,872	1,544	,202
Within Groups	22107,895	796	27,774		
Total	22236,512	799			

VAR00001

	VAR00002	N	Subset for alpha = .05
			1
Duncan[a]	2,00	200	49,3706
	4,00	200	49,3866
	3,00	200	49,8749
	1,00	200	50,3430
	Sig.		,093

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 200,000.

Рис. 3.2. Результати, отримані у пакеті SPSS

Бачимо, що результати однофакторного дисперсійного аналізу збігаються з даними, отриманими за допомогою MS Excel, а результати перевірки однорідності дисперсій є точнішими й не потребують додаткового застосування таблиць. Крім того, додатково визначено, немає поділу вибірок на групи, що суттєво відрізняються одна від іншої за середнім значенням.

3.4. Приклад виконання рангового однофакторного аналізу

Нехай ми маємо чотири вибірки, сформовані за допомогою пакета аналізу MS Excel як суміш 100 елементів нормально розподіленої вибірки з параметрами $\bar{x} = 20$, $s = 3$ та рівномірно розподілених вибірок обсягом по 100 елементів кожна заданих, відповідно, на відрізках: [17; 22], [18; 22], [18; 22] та [17; 23]. В електронних таблицях MS Excel немає вбудованих засобів для реалізації рангового однофакторного аналізу Краскела – Уолліса. Але його неважко здійснити за допомогою наявних функцій. Спочатку необхідно перетворити таблицю вихідних даних у таблицю значень рангів. Для цього можна використати функцію РАНГ (). Вікно задання її параметрів показано на рис. 3.3.

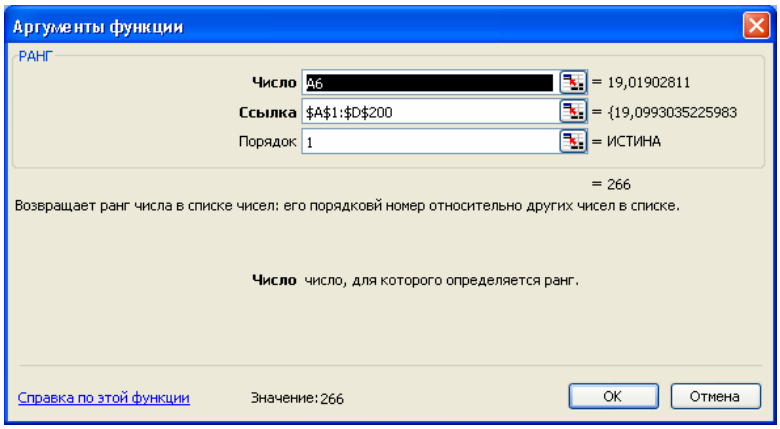


Рис. 3.3. Вікно задання параметрів функції РАНГ ()

У комірці “Число” вказуємо для якого саме значення таблиці вихідних даних необхідно обчислити ранг. У комірці “Ссылка” даємо посилання на весь діапазон, що містить вихідні значення (це посилання має бути абсолютним). У комірці “Порядок” зазначаємо порядок ранжирування: 0 – за убуванням, інше число – за зростанням. Після цього за формулами (3.6) або (3.7) розраховуємо значення критерію. У нашому випадку воно дорівнює 7,22. Як критичне візьмемо значення оберненої функції розподілу χ^2, яке можна визначити за допомогою функції = ХИ2ОБР (0,05;3) = 7,82. Її аргументами є рівень значущості (0,05) та кількість степенів вільності (у нашому випадку 3). Бачимо, що розрахункове значення критерію дещо менше, ніж критичне. Тому немає підстав для відхилення нульової гіпотези про однорідність досліджуваних вибірок.

У пакеті SPSS реалізовано можливість перевірки однорідності вибірок числових даних, розподіл яких відрізняється від нормального закону, а також порядкових даних за допомогою рангового однофакторного аналізу Краскела – Уолліса.

Для реалізації цієї процедури необхідно ввійти до діалогового вікна Analyze/Nonparametric Tests/K Independent Samples й обрати в ньому тест Краскела – Уолліса. Додатково ми можемо здійснити перевірку за медіанним тестом і критерієм Джонкхієра – Терпстри. Результати наведено на рис. 3.4–3.6. Бачимо, що за результатами всіх тестів рівень значущості, що відповідає розрахунковому значенню критерію, є вищим, ніж 0,05. Тому ми можемо прийняти нульову гіпотезу про однорідність досліджуваних вибірок на рівні значущості 0,05.

Ranks

	VAR00002	N	Mean Rank
VAR00001	1,00	200	370,98
	2,00	200	394,42
	3,00	200	432,20
	4,00	200	404,40
	Total	800	

Test Statistics[a,b]

	VAR00001
Chi-Square	7,221
df	3
Asymp. Sig.	,065

a. Kruskal Wallis Test

b. Grouping Variable: VAR00002

Рис. 3.4. Результати перевірки однорідності вибірок
за критерієм Краскела – Уолліса

Frequencies

		VAR00002			
		1,00	2,00	3,00	4,00
VAR00001	> Median	89	95	110	106
	<= Median	111	105	90	94

Test Statistics[b]

	VAR00001
N	800
Median	20,0603
Chi-Square	5,640[a]
df	3
Asymp. Sig.	,131

a. 0 cells (,0%) have expected frequencies less than 5. The minimum expected cell frequency is 100,0.

b. Grouping Variable: VAR00002

Рис. 3.5. Результати перевірки однорідності вибірок за медіанним тестом

Jonckheere-Terpstra Test[a]

	VAR00001
Number of Levels in VAR00002	4
N	800
Observed J-T Statistic	127124,00
Mean J-T Statistic	120000,00
Std. Deviation of J-T Statistic	3654,221
Std. J-T Statistic	1,950
Asymp. Sig. (2-tailed)	,051

a. Grouping Variable: VAR00002

Рис. 3.6. Результати перевірки однорідності вибірок за критерієм Джонкхієра – Терпстрі

Контрольні питання

1. Які завдання вирішують за допомогою дисперсійного аналізу?
2. Що називають факторами й відгуками у дисперсійному аналізі? Наведіть приклади.
3. Що називають рівнем фактора? Наведіть приклади.
4. Як визначають кількість і розміри інтервалів в однофакторному дисперсійному аналізі?
5. Які типи групування використовують у багатофакторному дисперсійному аналізі?
6. Якими є основні умови застосування однофакторного дисперсійного аналізу?
7. Доведіть основну тотожність дисперсійного аналізу.
8. Що являє собою факторна варіація та яку властивість даних вона характеризує?
9. Що являє собою залишкова варіація та яку властивість даних вона характеризує?
10. Які властивості даних характеризують оцінки дисперсії похибок, що використовуються у дисперсійному аналізі?
11. Які завдання вирішують за допомогою рангового однофакторного аналізу Краскела – Уолліса?
12. За яких умов можна використовувати ранговий однофакторний аналіз Краскела – Уолліса?
13. Який критерій є аналогом рангового однофакторного аналізу Краскела – Уолліса при порівнянні двох вибірок?
14. Які завдання вирішують за допомогою критерію Джонкхієра? За яких умов його доцільно використовувати?

15. Які завдання вирішують за допомогою критерію Бартлетта? За яких умов його можна використовувати?

16. Які завдання вирішують за допомогою *G*-критерію Кокрена? За яких умов його можна використовувати?

17. Яким є загальний вигляд адитивної моделі кількісного впливу досліджуваного фактора на відгук?

18. Що називають зсувом у дисперсійному аналізі? Як можна оцінити зсув кількісно?

19. Що називають лінійним контрастом в адитивній моделі? Наведіть приклади лінійних контрастів.

20. Які фактори визначають дисперсію лінійного контрасту?

21. Які завдання вирішують за допомогою методу множинних порівнянь Шеффе?

22. Які завдання вирішують за допомогою двофакторного дисперсійного аналізу? За яких умов можна використовувати цей метод?

23. Які властивості даних перевіряють за допомогою рангового критерію Фрідмана? За яких умов можна використовувати цей критерій?

24. Які властивості даних перевіряють за допомогою критерію Пейджа? За яких умов можна використовувати цей критерій?

25. Які властивості даних перевіряють за допомогою *Q*-критерію Кокрена? За яких умов можна використовувати цей критерій?

4. КОРЕЛЯЦІЙНИЙ АНАЛІЗ

Кореляцією (**кореляційним зв'язком**) між випадковими величинами (ознаками) називають наявність статистичного або ймовірнісного 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. Тому його можна застосовувати для формалізованого подання моделей зв'язків між окремими компонентами системи або між окремими процесами, що відбуваються в ній. Наявність кореляційного зв'язку не означає існування причинно-наслідкового зв'язку між досліджуваними ознаками. Вона може бути зумовлена тим, що обидві ознаки мають причинно-наслідковий зв'язок з певним іншим фактором. Наприклад, існує кореляція між цінами на нафту й на золото. Проте вона пояснюється тим, що обидві ціни виражаються у доларах США й залежать від динаміки його індексу. Кореляція також може бути випадковою.

Сучасну класифікацію мір подібності запропонували австрійський та американський біостатистик та антрополог Роберт Сокал та британський таксономіст Пітер Сніс у 1963 р. Згідно з нею виокремлюють такі типи мір подібності [58]:

– міри асоціації, що відбивають різні співвідношення кількості ознак, що збігаються до загальної кількості ознак, а також близькі до них коефіцієнти спряженості (квантифіковані коефіцієнти зв'язку);

– вибіркові коефіцієнти зв'язку типу кореляції (нормовані косинусні міри);

– показники відстані у метричному просторі.

Перевірку зв'язку можна здійснювати лише для пов'язаних вибірок. Це означає, що між елементами обох досліджуваних вибірок існує взаємно однозначна відповідність, а кількість елементів у вибірках є однаковою.

Замість гіпотези про наявність кореляційного зв'язку часто розглядають протилежну гіпотезу про відсутність зв'язку між досліджуваними величинами. Нехай ознака A має r рівнів $A_1, A_2, \ldots A_r$, а ознака B – s рівнів $B_1, B_2, \ldots B_s$. Їх вважають **незалежними**, якщо події "ознака A набуває значення A_i" та "ознака B набуває значення B_j" є незалежними для всіх можливих пар i, j, тобто:

$$P\left(A_i, B_j\right) = P\left(A_i\right) P\left(B_j\right). \tag{4.1}$$

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

$$P\left(B_j / A_i\right) = P\left(B_j\right), \quad \forall\left(A_i, B_j\right). \tag{4.2}$$

Кореляційний аналіз здійснюють на початковому етапі вирішення всіх основних проблем статистичного аналізу даних [4]. У проблемі статистичного аналізу залежностей і побудови регресійних моделей він дає змогу встановити сам факт існування зв'язку між змінними та оцінити ступінь його прояву. У проблемі класифікації даних за допомогою кореляційного аналізу отримують вихідну інформацію у вигляді коваріаційних і кореляційних матриць та інших характеристик парних порівнянь. Це дає змогу визначити подібні один до одного або до певних еталонів об'єкти, сформувати класи подібних об'єктів і здійснити класифікацію. У проблемі зменшення розмірності досліджуваного простору ознак також за допомогою коваріаційних і кореляційних матриць визначають ознаки, що можуть бути без втрати суттєвої інформації подані через інші наявні дані.

Загальна методика перевірки гіпотези про існування зв'язку між ознаками передбачає три основних етапи: визначення типу даних; перевірку гіпотези про відсутність зв'язку і, в разі її відхилення, оцінювання сили зв'язку. Тип вихідних даних суттєво впливає на вибір методів і критеріїв, які можна застосовувати на наступних етапах аналізу.

Для визначення сили зв'язку використовують різноманітні показники. Зазвичай їх прагнуть вибрати такими, щоб вони змінювалися від –1 до +1 або від 0 до 1. Значення, що є близькими за модулем до одиниці, свідчать про наявність сильного зв'язку. Близькі до нуля значення вказують або на відсутність будь-якого зв'язку, або на відсутність зв'язку того типу (найчастіше лінійного), для якого розроблено відповідний коефіцієнт. Знак коефіцієнта вказує на напрям зв'язку: прямий (для додатних значень) або зворотний (для від'ємних).

4.1. Кореляційний аналіз кількісних ознак

Методику кількісного оцінювання кореляції між ознаками вперше було запропоновано британським географом, антропологом та психологом Френсисом Гальтоном в 1888 р.

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

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

де

$$s_y^2 = \frac{1}{n}\sum_{k=1}^{n}\left(y_k - \bar{y}\right)^2, \tag{4.4}$$

n – кількість спостережень, а вибіркове значення дисперсії нев'язок ε обчислюють за однією з таких формул:

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

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

$$s_\varepsilon^2 = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{\nu_j}\sum_{i=1}^{\nu_j}\left(y_{ij} - \bar{y}_{j*}\right)^2, \tag{4.6}$$

де ν_j – кількість даних, що потрапили до j-го інтервалу групування;

y_{ji} – значення i-го спостереження досліджуваної ознаки, що потрапило до j-го інтервалу;

$\bar{y}_{j*} = \dfrac{\sum_{i=1}^{\nu_j} y_{ji}}{\nu_j}$ – її середнє значення за спостереженнями, які потрапили до j-го інтервалу;

m – кількість інтервалів.

Формулу (4.5) застосовують у випадку, коли за результатами попереднього аналізу встановлено, що умовна дисперсія $D\left(\varepsilon \mid \mathbf{X}\right) = \sigma_\varepsilon^2 = const$, 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. У цьому випадку необхідно попередньо здійснити групування даних. Для цього їх впорядковують за зростанням значень однієї з ознак (ознаки X). Потім задають кількість та межі інтервалів для цієї ознаки. Підраховують кількості точок, що потрапили до кожного інтервалу $\left(\nu_j\right)$, для змінної Y обчислюють загальне середнє $\bar{y}$ та середні за інтервалами $\bar{y}_{j*}$ й розраховують значення коефіцієнта детермінації за формулами (4.3, 4.4, 4.6).

Величина коефіцієнта детермінації може змінюватися в межах від нуля до одиниці й відображає частку загальної дисперсії досліджуваної ознаки, яка зумовлена зміною функції регресії $f(\mathbf{X})$. При цьому нульове значення коефіцієнта детермінації відповідає відсутності будь-якого зв'язку, а його рівність одиниці – наявності строго функціонального зв'язку. Оскільки цей коефіцієнт є універсальним показником зв'язку, він має відбивати й такі зв'язки, що є немонотонними функціями. Тому питання напряму зв'язку у цьому випадку не має сенсу.

Слід зазначити, що для обмеженого набору даних часто можна побудувати декілька різних адекватних регресійних моделей. Групування

даних також можна здійснювати різними способами. Тому існує певна невизначеність коефіцієнтів детермінації: при застосуванні різних регресійних моделей або різних способів групування ми будемо отримувати дещо різні значення коефіцієнта детермінації.

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

Розрізняють парні та частинні кореляційні характеристики. Парні характеристики розраховують за результатами вимірювань тільки досліджуваної пари ознак. Тому вони не враховують опосередкованого або спільного впливу інших ознак. Частинні характеристики є очищеними від впливу інших факторів, але для їх розрахунку необхідно мати вихідну інформацію не тільки про досліджувані ознаки, а й про всі інші, вплив яких необхідно усунути.

Для кількісних ознак найчастіше застосовують коефіцієнти кореляції Пірсона і Фехнера. **Коефіцієнт кореляції Пірсона** (**коефіцієнт кореляційного відношення Пірсона, парний коефіцієнт кореляції**, **вибірковий коефіцієнт кореляції**, **коефіцієнт Бравайса – Пірсона**) вимірює ступінь лінійного кореляційного зв'язку між кількісними скалярними ознаками. Він був запропонований К. Пірсоном у 1896 р. Часто, посилаючись на згадування К. Пірсона про ідеї математичного подання зв'язку, висловлені в 1846 р. відомим французьким фізиком та кристалографом Огюстом Браве, цей показник називають коефіцієнтом Бравайса – Пірсона (Бравайс – це викривлена транскрипція від французького Bravais, що закріпилася в літературі з кореляційного аналізу). Цей коефіцієнт розраховують за формулою:

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

Коефіцієнт Пірсона можна виразити також через дисперсії σ_y і $\sigma_{\Delta y}$, друга з яких характеризує розкид емпіричних точок стосовно рівняння лінійної регресії $y = ax + b$, де a та b – коефіцієнти, визначені за методом найменших квадратів:

$$r = \frac{1}{\sqrt{1 + \left(\sigma_{\Delta y} / \sigma_y\right)^2}}. \qquad (4.8)$$

За умови достатньо великого обсягу спостережень $\left(N \geq 30\right)$ стандартне відхилення коефіцієнта кореляції Пірсона можна визначити за формулою:

$$\sigma_r = \frac{1 - r^2}{\sqrt{N}}. \qquad (4.9)$$

На рівні значущості 0,01 гіпотезу про наявність кореляційного зв'язку приймають, якщо $|r|/\sigma_r \geq 2,6$.

Застосування коефіцієнта Пірсона як міри зв'язку є обґрунтованим лише за умови, що спільний розподіл пари ознак є нормальним. Тому перед його розрахунком слід перевірити виконання цієї гіпотези. Якщо вона справедлива, то квадрат коефіцієнта кореляції Пірсона дорівнює коефіцієнту детермінації.

Значення коефіцієнта кореляції може змінюватися від −1 до +1. Значення −1 та +1 відповідають чіткій 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 випадку знаменник дорівнює нулю. Що ближчим є значення коефіцієнта кореляції до −1 або +1, то 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 серії точок, координати яких відповідають двом парам спряжених вибірок.

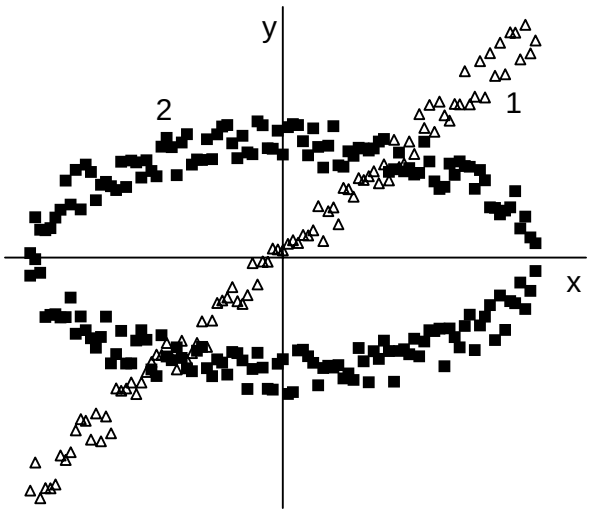


Рис. 4.1. Графічне зображення двох наборів тестових даних

Для обох пар вибірок 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 пари вибірок пов'язана не з відсутністю зв'язку, а з його нелінійністю. Для порівняння, коефіцієнти детермінації для тих самих пар вибірок дорівнюють 0,98 та 1,00.

Показаний приклад свідчить, що в багатьох випадках для попереднього аналізу припущення про наявність і тип зв'язку між певними ознаками доцільно нанести наявні дані на графік.

Як видно, близькість коефіцієнта кореляції Пірсона до нуля в загальному випадку не є доказом незалежності ознак. Але можна довести, що у випадку, коли сумісний розподіл випадкових величин (x, y) є нормальним, рівність $r = 0$ свідчить про статистичну незалежність x і y.

Коефіцієнт кореляції Пірсона часто розглядають як універсальну міру кореляційного зв'язку. У багатьох пакетах загального призначення, зокрема в електронних таблицях MS Excel, не передбачено інших засобів його вимірювання. Але, як випливає з наведених вище даних, насправді сфера його обґрунтованого застосування є досить вузькою, оскільки лінійність залежності й нормальний розподіл даних навколо неї є скоріше винятком, ніж правилом.

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:

$$r \in \left[\tanh\left(z(r) - \frac{N_{\frac{1+p}{2}}}{\sqrt{n-3}} \right); \tanh\left(z(r) + \frac{N_{\frac{1+p}{2}}}{\sqrt{n-3}} \right) \right], \qquad (4.10)$$

де n – обсяг вибірки;

$N_{\frac{1+p}{2}}$ – квантіль нормального розподілу;

p – значення довірчого рівня;

z (r) – z-перетворення (перетворення Фішера) вибіркового коефіцієнта кореляції r.

Коефіцієнт кореляції Пірсона можна застосовувати для перевірки гіпотези про значущість зв'язку. Для нормально розподілених вихідних даних величину вибіркового коефіцієнта кореляції вважають значимо відмінною від нуля, якщо виконується нерівність:

$$r^2 > \left[1 + (n-2)/t_\alpha^2 \right]^{-1}, \qquad (4.11)$$

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

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

$$\varphi_{r(n)}(r) = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-2}{2}\right)} \left(1 - r^2\right)^{\frac{n-4}{2}} \quad (-1 < r < 1). \qquad (4.12)$$

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

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

$$\rho_{yx}^2 = s_{\bar{y}(x)}^2 / s_y^2, \tag{4.13}$$

де

$$s_{\bar{y}(x)}^2 = \frac{1}{n}\sum_{j=1}^{s} \nu_j \left(\bar{y}_{j*} - \bar{y}\right)^2;$$

$$s_y^2 = \frac{1}{n}\sum_{j=1}^{s}\sum_{i=1}^{\nu_j}\left(y_{ji} - \bar{y}\right)^2;$$

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{s} \nu_j \bar{y}_{j*};$$

$$\bar{y}_{j*} = \left(\sum_{i=1}^{\nu_j} y_{ij}\right) / \nu_j,$$

n – обсяг вибірки; *s* – кількість інтервалів групування по вісі абсцис; v_j – кількість точок, що потрапили до *j*-го інтервалу. З погляду термінології, уведеної у попередньому розділі, кореляційне відношення є квадратним коренем з відношення факторної варіації ознаки до її загальної варіації.

Кореляційне відношення може змінюватися в інтервалі від нуля до одиниці. Із $\rho_{yx} = 1$ 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 групові 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, другий – одиниці. Зок-

рема, це може спостерігатися для парних функцій за умови, що функція розподілу значень незалежної змінної є симетричною стосовно нуля. Для даних, що наведені на рис. 4.1, кореляційне відношення першої серії дорівнює приблизно 0,98 і в межах похибки обчислень 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 з 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:

$$v^2 = \frac{(n-k)\left(\rho_{yx}^2 - r^2\right)}{(k-2)\left(1-\rho_{yx}^2\right)}, \tag{4.14}$$

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 (4.13) значення перевищує точку v_α^2 розподілу $F(s-2, n-s)$, то гіпотезу про лінійний зв'язок відхиляють на рівні значущості α. Слід зазначити, що у зв'язку з можливістю різних способів групування даних значення кореляційного відношення, як і значення коефіцієнта детермінації, у загальному випадку є дещо невизначеним.

Коефіцієнт кореляції Фехнера розраховують за формулою:

$$r_F = \frac{C-H}{C+H} = \frac{2C-n}{n} = \frac{2C}{n} - 1, \tag{4.15}$$

де C – кількість збігів знаків відхилень варіант від відповідних середніх;

H – кількість знаків, що не збігаються. Цей показник було запропоновано німецьким психологом Густавом Фехнером у 1860 р.

Значення коефіцієнта Фехнера можуть змінюватися в межах від -1 до $+1$. Як і коефіцієнт Пірсона, він показує наявність лінійного зв'язку: що ближчим до одиниці за модулем є значення коефіцієнта, то сильніший 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 наведених на рис. 4.1 наборів даних 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

застосовують досить рідко, але у певних випадках, коли інформація про збіги й незбіги знаків відхилень потрібна й для інших цілей, він може виявитися зручнішим за критерій Пірсона.

Коваріацією називають змішаний момент другого порядку. Її розраховують за формулою:

$$Cov(X,Y)=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right). \qquad (4.16)$$

На відміну від інших показників, що характеризують наявність статистичного зв'язку, вона не є безрозмірною величиною. Також немає будь-яких обмежень на її значення. У загальному випадку за інших рівних умов вона збільшується (за модулем) із зростанням середніх значень досліджуваних показників. Це робить коваріацію незручною для застосування як показника сили зв'язку. Але у багатьох алгоритмах її використовують як проміжний показник, що застосовують у подальших розрахунках. У таких випадках важливою перевагою коваріації є необхідність виконання значно меншої кількості елементарних обчислень, ніж для аналогічних показників кореляції, таких як коефіцієнт Пірсона. Крім того вона має важливе теоретичне значення, що також у певних випадках приводить до доцільності її використання в аналізі даних.

Коваріація вибірки із самою собою є дисперсією. З наведеної формули можна отримати, що коефіцієнт кореляції Пірсона $r=\frac{Cov(X,Y)}{\sigma_X\sigma_Y}$, де σ_X^2, σ_Y^2 – дисперсії вибірок.

При аналізі багатовимірних вибірок часто застосовують **коваріаційні матриці**:

$$C=\begin{pmatrix} c_{11} & c_{12} & \dots & c_{1n} \\ c_{21} & c_{22} & \dots & c_{2n} \\ \dots & \dots & \dots & \dots \\ c_{n1} & c_{n2} & \dots & c_{nn} \end{pmatrix}, \qquad (4.17)$$

де $c_{ij}=\mathrm{cov}\left(x_i,x_j\right)$. Діагональні елементи матриці (4.17) є дисперсіями $c_{ii}=\sigma^2\left(x_i\right)$ відповідних рядів спостережень. Коваріаційна матриця є симетричною, тобто $c_{ij}=c_{ji}$.

4.2. Кореляційний аналіз порядкових ознак

Під **ранговою кореляцією** розуміють статистичний зв'язок між порядковими ознаками. Вихідні дані зазвичай подають у вигляді табл. 4.1, де елемент x_{ik} є рангом *i*-го об'єкта за *k*-ю властивістю.

Таблиця 4.1

Таблиця вихідних даних для рангового кореляційного аналізу

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}

Завданнями аналізу в цьому випадку можуть бути: вивчення структури досліджуваних об'єктів; перевірка сукупної узгодженості ознак та умовне ранжирування 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 $\mathbf{X}^{(k)} = \left(x_1^{(k)},\ x_2^{(k)},\ ...,\ x_n^{(k)} \right),\ k = 0,\ 1,\ ...,\ p$ у n-вимірному просторі ознак. Найхарактернішими типами структури 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 від одного. Це відповідає наявності декількох таких підмножин ознак, 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.

Вирішення 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 можна розглядати як задачу найкращого у певному розумінні відновлення невідомого ранжирування за наявними емпіричними даними, що зумовлює можливість її розгляду як задачі регресії.

Коефіцієнт рангової кореляції Спірмена (**показник кореляції рангів Спірмена, коефіцієнт кореляції рангів**) запропонований британським

психологом Чарльзом Едвардом Спірменом у 1904 р. Його використовують, якщо досліджується зв'язок між рядами даних, виміряними за порядковою шкалою. Його можна застосовувати також і для кількісних даних, але, як правило, це буває недоцільним. У найпростішому випадку досліджувані об'єкти класифікують за двома ознаками. Наприклад, ми можемо спочатку впорядкувати групу учнів за їх здібностями до математики, а потім – до іноземних мов. Місця, які *i*-й учень займе в обох списках, будуть його рангами r_i та s_i. Якщо досліджувані ознаки взаємопов'язані, то послідовність рангів $r_1, r_2, ..., r_n$ певною мірою корелює з послідовністю рангів $s_1, s_2, ..., s_n$.

Ступінь близькості двох послідовностей відображує величина:

$$S_\rho = \sum_{i=1}^{n} \left(r_i - s_i\right)^2 . \tag{4.18}$$

Якщо для нумерації об'єктів попередньо впорядкувати їх за однією з ознак, наприклад за зростанням рангів r_i, то формула (4.18) може бути записана так:

$$S_\rho = \sum_{k=1}^{n} \left(k - s_k\right)^2 . \tag{4.19}$$

Величина S_ρ набуде 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 послідовність рангів другої $s_i = \{n,\ n-1,\ ...,\ 1\}$. Величину S_ρ незручно застосовувати як міру зв'язку, оскільки на її значення впливає кількість пар варіант досліджуваних рядів *n*.

З огляду на це, як міру зв'язку використовують коефіцієнт рангової кореляції Спірмена, значення якого розраховують за формулою:

$$\rho_s = 1 - \frac{6\left(S_\rho + B_x + B_y\right)}{n^3 - n}, \tag{4.20}$$

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

$$B_i = \frac{1}{12}\sum_{i=1}^{m} n_i \left(n_i^2 - 1\right), \tag{4.21}$$

де m – кількість груп об'єднаних рангів у вибірці;

n_i – кількість рангів у *i*-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.

Коефіцієнт рангової кореляції Спірмена можна застосовувати як показник некорельованості вибірок. У цьому випадку розраховують величину:

$$t_p = \sqrt{n-2}\,\frac{\rho_s}{\sqrt{1-\rho_s^2}}. \qquad (4.22)$$

Для великих за обсягом вибірок ($n > 50$) статистика цього критерію наближається до розподілу Стьюдента з $(n-2)$ степенями вільності. Статистика $\sqrt{n-1}\rho_s$ для великих вибірок наближається до стандартного нормального розподілу.

Інший підхід використовує як міру подібності двох вибірок мінімальну кількість перестановок сусідніх об'єктів, потрібну для переведення послідовності рангів однієї вибірки до послідовності рангів іншої. Можна показати, що вона дорівнює кількості інверсій в однієї з цих послідовностей у випадку, коли інша послідовність впорядкована за зростанням. Нехай, наприклад, $n = 4$, послідовність r_i впорядкована за зростанням, а $s_i = \{4,\ 3,\ 1,\ 2\}$. Інверсіями є: $4 > 3;\ 4 > 1;\ 4 > 2;\ 3 > 1;\ 3 > 2$. Їх кількість $K = 5$. Найменше можливе значення кількості інверсій $K = 0$ відповідає повному збігу рангових послідовностей, а найбільше $K = \dfrac{n(n-1)}{2}$ – їх повній протилежності.

Як і в попередньому випадку, кількість інверсій залежить від обсягу вибірки і є незручною для застосування як показника кореляції. Для цього використовують **коефіцієнт рангової кореляції Кендалла** (**коефіцієнт кореляції рангів**, **ранговий коефіцієнт кореляції**). Він був запропонований британським статистиком Маурисом Кендаллом у 1938 р. Його розраховують за формулою:

$$\tau = 1 - \frac{2K}{\sqrt{\left(\dfrac{n(n-1)}{2} - B_x\right)\left(\dfrac{n(n-1)}{2} - B_y\right)}}, \qquad (4.23)$$

де r_j, s_i – масиви рангів аналізованих рядів;

n – кількість пар варіант у них. B_x, B_y – поправки на об'єднання рангів у відповідних рядах, які обчислюють за формулою:

$$B_i = \frac{1}{2}\sum_{i=1}^{m} n_i\left(n_i - 1\right), \qquad (4.24)$$

де m – кількість груп об'єднаних рангів у вибірці;

n_i – кількість рангів у i-й групі.

Для коефіцієнта рангової кореляції Кендалла у випадку великих вибірок статистика:

$$\tau\sqrt{\frac{9n(n-1)}{2(2n+5)}} \tag{4.25}$$

має розподіл, близький до стандартного нормального закону.

Коефіцієнт рангової кореляції Кендалла призначений для визначення сили кореляційного зв'язку між двома рядами даних за тих самих умов, що і коефіцієнт рангової кореляції Спірмена. Як і для коефіцієнта Спірмена, його значення можуть змінюватися в межах від −1 до +1, при цьому −1 відповідає повній протилежності послідовностей рангів, а +1 – їх повному збігу. Слід зазначити, що обчислення коефіцієнта Кендалла є більш трудомістким, але з іншого боку, він має ряд переваг порівняно із коефіцієнтом Спірмена. Основними з них є такі [23]:

– кращий рівень вивченості його статистичних властивостей, зокрема його вибіркового розподілу;

– можливість його застосування для визначення частинної кореляції;

– більша зручність перерахунку при додаванні нових даних.

4.3. Кореляційний аналіз номінальних ознак

Типовою ситуацією, коли необхідна перевірка 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 проживання тощо.

При дослідженні зв'язків між **категоризованими** ознаками вихідні дані подають у вигляді таблиці спряженості (табл. 4.2). До категоризованих зараховують номінальні ознаки, а також порядкові ознаки, для яких є відомим скінченний набір можливих градацій.

Величини f_{ij} показують, скільки разів зустрічалася комбінація ознак, за якої рівень першої має значення i, а рівень другої – j; m_j є сумами стовпців, а n_i – сумами рядків. За даними табл. 4.2 можна оцінити значення ймовірностей, що входять до формули (4.1):

Таблиця 4.2

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

$$p_{ij} = P\left(A_i B_j\right) = \frac{f_{ij}}{S};\ p_i = P\left(A_i\right) = \sum_{j=1}^{r} p_{ij} = \frac{n_i}{S};$$
$$p_j = P\left(B_j\right) = \sum_{i=1}^{c} p_{ij} = \frac{m_j}{S}. \tag{4.26}$$

Звідси для незалежних ознак маємо:

$$f_{ij} \approx n_i m_j / S. \tag{4.27}$$

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

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r} \frac{\left(f_{ij} - \varphi_{ij}\right)^2}{\varphi_{ij}} = S\left[\sum_{i=1}^{c}\sum_{j=1}^{r} \frac{f_{ij}^2}{n_i m_j} - 1\right], \tag{4.28}$$

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

На практиці частіше використовують **φ-коефіцієнт Пірсона**, або **середньоквадратичну спряженість** $\varphi^2 = \chi^2 / S$, яка може змінюватися від нуля до $\min\left\{r-1, c-1\right\}$.

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

Коефіцієнт спряженості Крамера був запропонований К.Х. Крамером у 1946 р. Його розраховують за формулою:

$$C = \left[\frac{\sum_{i=1}^{c}\sum_{j=1}^{r} \frac{f_{ij}^2}{n_i m_j} - 1}{\min\left(c-1, r-1\right)}\right]^{1/2} = \left[\frac{\varphi^2}{\min\left(c-1, r-1\right)}\right]^{1/2}. \tag{4.29}$$

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:

$$\sigma_C^2 \approx \frac{1}{n \min\left(c-1, r-1\right)}. \tag{4.30}$$

Її довірчий інтервал:

$$\left[C - u_{1-\alpha}\sigma_C; C + u_{1-\alpha}\sigma_C\right], \tag{4.31}$$

де u_q – q-квантиль стандартного нормального розподілу.

Поліхоричний коефіцієнт спряженості Чупрова призначений для дослідження кореляції номінальних ознак у таблиці спряженості $r \times c$. Він був уведений російським статистиком О.О. Чупровим у 1926 р. Його значення розраховують за формулою:

$$T = \frac{J-1}{\sqrt{(r-1)(c-1)}};\ J = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{f_{ij}^2}{n_i m_j}. \tag{4.32}$$

Існує велика кількість коефіцієнтів, що характеризують кореляцію між ознаками у випадку, коли кожна з двох ознак може мати лише два рівні, які найчастіше відповідають наявності та відсутності ознаки. У цьому випадку 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, уведений в 1901 р. французьким геоботаніком Полем Жаккаром, обчислюють за формулою:

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

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

Простий коефіцієнт зустрічальності (показник подібності Сокала й Міченера) запропонований Р. Сокалом та американським ентомологом Чарльзом Дунканом Міченером у 1958 р. Його розраховують за формулою:

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

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

Показник подібності Рассела і Рао запропонували в 1940 р. американський епідеміолог Поль Ф. Рассел та індийський і британський ентомолог Т. Рамакришна Рао. Його обчислюють як:

$$J = \frac{a}{n} = \frac{a}{a+b+c+d}. \tag{4.35}$$

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

Коефіцієнт спряженості Бравайса – Пірсона (**показник подібності Чупрова**) був уведений О.О. Чупровим у 1923 р. Його розраховують за формулою:

$$C = \frac{ad - bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}. \qquad (4.36)$$

Значення цього коефіцієнта може змінюватися в межах від −1 до +1. Від'ємні значення коефіцієнта спряженості означають, що із збільшенням імовірності прояву одної ознаки, зменшується імовірність прояву іншою.

Легко показати, що цей показник є окремим випадком φ-коефіцієнта Пірсона для таблиць 2×2.

Коефіцієнт асоціації Юла був уведений відомим британським статистиком Джорджем Удні Юлом у 1900 р. Його визначають із співвідношення:

$$Q = \frac{ad - bc}{ad + bc}. \qquad (4.37)$$

Коефіцієнт колігації Юла, що також був запропонований Дж.У. Юлом в 1912 р., обчислюють як:

$$Y = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}. \qquad (4.38)$$

Він не має переваг порівняно з коефіцієнтом асоціації. Значення обох коефіцієнтів змінюються в межах від −1 до +1.

Хеммінгова відстань (**метрика Хеммінга**) $H = a + d$ також може застосовуватися для визначення кореляції. Проте, як і коваріація, вона не є безрозмірною величиною і може набувати будь-яких невід'ємних значень (верхньою межею є загальна кількість спостережень n). Цей показник був уведений відомим американським математиком Ричардом Веслі Хеммінгом у 1950 р.

4.4. Кореляційний аналіз змішаних ознак

Коефіцієнт Гауера був запропонований британським статистиком Джоном Кліффордом Гауером у 1971 р. Його застосовують у тому випадку, коли досліджувані ознаки виміряні в різних шкалах. Обчислення елементів матриці подібності здійснюють за формулою:

$$s_{ij} = \frac{\sum_{k=1}^{p} S_{ijk}}{\sum_{k=1}^{p} W_{ijk}}, \ \left(i = 1, ..., n;\ j = 1, ..., n\right), \qquad (4.39)$$

де S_{ijk} $\left(i, j = 1, ..., n;\ k = 1, ..., p\right)$ – 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:

$$S_{ijk} = 1 - \frac{\left| x_{ik} - x_{jk} \right|}{R_k}, \tag{4.40}$$

де x_{ik}, x_{jk} – значення k-ї змінної для об'єктів i та j;

R_k – розкид k-ї ознаки, обчислений за всіма об'єктами.

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

При класифікації за порядковою ознакою бісеріальний коефіцієнт:

$$r_b = \frac{\left(\overline{x_1} - \overline{x} \right) n_1}{n s_x z_k}, \tag{4.41}$$

де $\overline{x_1}$ – середнє за першим рядком;

$\overline{x}$ – загальне середнє за всією таблицею;

s_x – вибіркове середнє квадратичне відхилення;

n_1 – чисельність першого рядка; n – загальна чисельність усіх вибірок;

z_k – ордината щільності нормального розподілу в точці k, де k – розв'язок рівняння:

$$1 - F\left(k\right) = n_1 / n\,. \tag{4.42}$$

Значення бісеріального коефіцієнта кореляції можуть змінюватися від −1 до +1. Його похибку можна визначити за формулою:

$$m_{r_b} = \frac{1 - r_b}{\sqrt{n}}\,. \tag{4.43}$$

Вона має t-розподіл з кількістю степенів вільності $\left(n - 2\right)$.

Бісеріальний коефіцієнт кореляції за таблицею Келлі – Вуда запропонований американським психологом Луїсом Л. Терстоуном (Louis L. Thurstone) в 1928 р. Його розраховують за формулою:

$$r_b = \frac{\left|\overline{x_1} - \overline{x_2}\right| pq}{s_x \varsigma}, \tag{4.44}$$

де $p = n_1 / n$ – частка частот у рядку, що визначається умовою $p > q$;

q – частка частот в іншому рядку;

ς – ордината в точці межі класів частот першого та другого рядків, яка визначається за таблицею Келлі – Вуда. Похибку коефіцієнта визначають за формулою:

$$m_r = \frac{\frac{\sqrt{pq}}{\varsigma} - r_b^2}{\sqrt{n}}. \tag{4.45}$$

У випадку класифікації за номінальною ознакою обчислення бісеріального коефіцієнта кореляції можна здійснить за формулою:

$$r_\eta = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{q} n_i \left(\frac{\mu_i}{\sigma_i}\right)^2 - \left(\frac{\mu_y}{\sigma_y}\right)^2}{1 + \frac{1}{n}\sum_{i=1}^{q} n_i \left(\frac{\mu_i}{\sigma_i}\right)^2}}, \tag{4.46}$$

де n – загальний обсяг даних;

n_i – кількість даних в i-му перетині;

m_i/s_i – оцінка в перетині i, одержувана за таблицею нормального інтеграла від відносної частоти першої з двох якісних ознак;

m_y/s_y – оцінка, одержувана за таблицею нормального інтеграла від відносної частоти першої якісної ознаки за всією таблицею.

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

$$r_{pb} = \frac{\left|\overline{x_p} - \overline{x}\right|}{s_x} \sqrt{\frac{n_p}{n_q}}, \tag{4.47}$$

де $\overline{x_p}$ – середнє варіант кількісної вибірки, які відповідають подіям верхнього (першого) рівня дихотомічної вибірки;

$\overline{x}$ – середнє кількісної вибірки;

s_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 коефіцієнта кореляції Пірсона.

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

$$r_{pb}\sqrt{\frac{n-2}{1-r_{pb}^2}} \geq t_{\alpha}, \tag{4.48}$$

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

4.5. Множинна кореляція

Про множинну кореляцію мова йде в тому випадку, коли певна ознака може бути пов'язана не з однією, а із сукупністю декількох інших ознак.

У реальних дослідженнях можлива ситуація, коли на певну ознаку може впливати не одна, а декілька інших. В таких випадках парні показники кореляції будуть давати неправильну інформацію щодо наявності зв'язку між відповідними показниками, оскільки ці їх значення будуть викривлятися невраховуваними ознаками.

Для уникнення помилок використовують частинні показники кореляції, що усувають такий вплив. Ідея введення таких показників вперше була висунута Г.У. Юлом у 1896 р., а пізніше розвинена ним та К. Пірсоном.

Якщо досліджувані ознаки задовольняють багатовимірний нормальний розподіл, **частинний коефіцієнт кореляції** між двома ознаками *i* та *j* при фіксованих значеннях інших ознак розраховують за формулою:

$$r_{ijX^{(i,\,j)}} = -\frac{R_{ij}}{\sqrt{R_{ii}R_{jj}}}, \tag{4.49}$$

де R_{kl} – алгебраїчне доповнення для елемента r_{kl} у кореляційній матриці. Цей показник запропоновано К. Пірсоном у 1897 р.

Для тривимірної ознаки звідси можна отримати:

$$r_{01(2)} = \frac{r_{01} - r_{02}r_{12}}{\sqrt{\left(1-r_{02}^2\right)\left(1-r_{12}^2\right)}}. \tag{4.50}$$

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:

$$r_{01(2,3,\ldots,k+1)} = \frac{r_{01(2,\ldots,k)} - r_{0k+1(2,\ldots,k)} r_{1k+1(2,\ldots,k)}}{\sqrt{\left(1-r^2_{0k+1(2,\ldots,k)}\right)\left(1-r^2_{1k+1(2,\ldots,k)}\right)}}. \qquad (4.51)$$

Частинні коефіцієнти кореляції мають всі властивості парних коефіцієнтів кореляції. Вони є показниками наявності лінійного зв'язку між двома незалежними ознаками, який не залежить від впливу інших ознак.

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

$$R = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}}, \qquad (4.52)$$

де Y_i – значення змінної, взяті з кореляційної таблиці;

y_i – відповідні значення, розраховані за рівнянням регресії.

Крім того, застосовують множинний коефіцієнт кореляції, якій є мірою лінійної кореляції між певною змінною y та сукупністю величин $X_1, X_2, \ldots, X_n$ і визначається як звичайний парний коефіцієнт кореляції між y та множинною лінійною регресією за $X_1, \ldots, X_n$. При цьому припускають, що досліджувана сукупність підпорядковується багатовимірному нормальному закону.

Множинний коефіцієнт кореляції, запропонований у 1935 р. Г. Хотелінгом, є окремим випадком коефіцієнтів канонічної кореляції. Його можна розрахувати за формулою:

$$R^2_{yX} = 1 - \frac{|R|}{R_{00}}, \qquad (4.53)$$

де $|R|$ – визначник кореляційної матриці.

Його також можна визначити за частинними коефіцієнтами кореляції:

$$R^2_{yX} = 1 - \left(1-r^2_{01}\right)\left(1-r^2_{02(1)}\right)\left(1-r^2_{03(12)}\right)\ldots\left(1-r^2_{0n(1,2,\ldots,n-1)}\right). \qquad (4.54)$$

Наприклад множинний коефіцієнт кореляції між певною ознакою z та двома іншими ознаками (x, y) дорівнює:

$$R_z = R_{z/xy} = \sqrt{\frac{r^2_{zx} + r^2_{zy} - 2r_{xy}r_{zx}r_{zy}}{1-r^2_{xy}}}.$$

Множинний коефіцієнт кореляції мажорує будь-який парний або частинний коефіцієнт кореляції, що характеризує статистичні зв'язки досліджуваної ознаки. Як видно з формули (4.52), додавання нових ознак не може зменшувати коефіцієнт множинної кореляції.

Для багатовимірних нормальних сукупностей виконується рівність:

$$K_d\left(y, X\right) = R_{yX}^2 . \tag{4.55}$$

Подання математичних об'єктів називають **канонічним**, якщо кожному об'єкту однієї множини відповідає один і тільки один об'єкт іншої множини й ця відповідність є взаємно однозначною. **Канонічний кореляційний аналіз** здійснюють між двома сукупностями (групами) вибірок. Він призначений для визначення лінійної функції від перших p компонент і лінійної функції від q компонент, що залишилися, таких, щоб коефіцієнт кореляції між цими лінійними функціями набув найбільшого можливого значення. Чисельності груп (кількість вибірок у першій та другій групах, p та q) можуть різнитися, але необхідною умовою є рівна кількість варіант у всіх вибірках, що становлять обидві групи. Матриця взаємної кореляції двох груп вибірок має вигляд:

$$R = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix}, \tag{4.56}$$

де R_{11} – матриця взаємної кореляції p змінних першої групи розміром $p \times p$; R_{22} – матриця взаємної кореляції q змінних другої групи розміром $q \times q$; R_{12} – матриця взаємної кореляції змінних першої та другої груп розміром $p \times q$.

Розв'язання зводиться до узагальненої проблеми власних значень:

$$R_{12}^T R_{11}^{-1} R_{12} v = \lambda R_{22} v, \tag{4.57}$$

де λ – вектор власних значень розміром q.

Квадратні корені з власних значень називають **канонічними кореляціями**.

Для випадкової вибірки обсягу n з $\left(p+1\right)$-вимірного нормального розподілу коефіцієнт множинної кореляції вважають відмінним від нуля на рівні значущості α, якщо виконується нерівність:

$$R \geq \sqrt{\frac{pF}{\nu + pF}}, \tag{4.58}$$

де F – значення оберненої функції F-розподілу для довірчого рівня $\left(1-\alpha\right)$ та кількості степенів вільності p та $\left(n-p-1\right)$.

Коефіцієнт конкордації призначений для дослідження зв'язків між порядковими ознаками, кількість яких є більшою, ніж два. Як міру узго-

дженості беруть суму квадратів відхилень сум рангів спостережень (об'єктів) від їх спільного середнього рангу:

$$S_W = \sum_{i=1}^{n}\left(S_i - \overline{S}\right) = \sum_{i=1}^{n} S_i^2 - \frac{\left(\sum_{i=1}^{n} S_i\right)^2}{n};$$
$$\overline{S} = \frac{\sum_{i=1}^{n} S_i}{n};\ S_i = \sum_{j=1}^{k} R_{ij}, \tag{4.59}$$

де R_{ij} – ранг i-го спостереження за j-ю ознакою.

Коефіцієнт конкордації Кендалла (***W*-коефіцієнт Кендалла**) обчислюють за формулою:

$$W = \frac{12S_W}{k^2\left(n^3 - n\right)}. \tag{4.60}$$

Цей показник було запропоновано М. Кендаллом у 1939 р. Його значення може змінюватися в межах від нуля до одиниці, при цьому він дорівнює одиниці лише за умови, що всі досліджувані ранжирування збігаються. Коефіцієнт конкордації дорівнює нулю, якщо $k \geq 3$ і всі ранжирування є випадковими впорядкуваннями вихідної вибірки.

Середнє за всіма можливими парами ранжирувань значення коефіцієнта Спірмена за відсутності об'єднаних рангів пов'язано з коефіцієнтом конкордації співвідношенням:

$$\overline{\rho}_s = \frac{kW - 1}{k - 1}. \tag{4.61}$$

Величина $\left(k-1\right)\dfrac{W}{1-W}$ має F-розподіл з кількостями степенів вільності $\left(n-1\right)$ та $\left(\left(n-1\right)\left(k-1\right)-2\right)$. Високі значення функції F-розподілу свідчать про високий рівень узгодженості між ранжируваннями.

При $n > 7$ величина $k\left(n-1\right)W$ за відсутності зв'язку між ознаками має розподіл, близький до χ^2 з $(n-1)$ степенем вільності. Якщо:

$$k\left(n-1\right)W > \chi_\alpha^2\left(n-1\right), \tag{4.62}$$

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

4.6. Приклади здійснення кореляційного аналізу

Перевірка гіпотези про наявність зв'язку між кількісними ознаками

Побудуємо дві послідовності. Перша з них є арифметичною прогресією з першим членом –2 й різницею 0,05. Елементи другої розраховані за формулою: $y_i = 2x_i + \varepsilon_i$, де ε_i – елементи згенерованої за допомогою електронних таблиць MS Excel рівномірної випадкової послідовності, заданої на відрізку $[-0,5;\ 0,5]$.

Для перевірки нульової гіпотези про наявність зв'язку скористаємося відповідною процедурою пакета SPSS (Analyze/Correlate/Bivariate). У відповідному вікні задаємо: вибірки, зв'язок між якими необхідно перевірити; значення коефіцієнтів кореляції, які треба розрахувати; вказуємо характер гіпотези – однобічна чи двобічна; а також, за необхідністю, додаткові опції. На рис. 4.2 наведено результати кореляційного аналізу, отримані у пакеті SPSS, а на рис. 4.3 – графік, з якого видно наявність близького до лінійного зв'язку між ознаками.

Correlations

		VAR00001	VAR00006
VAR00001	Pearson Correlation	1	,991**
	Sig. (2-tailed)		,000
	Sum of Squares and Cross-products	110,700	225,011
	Covariance	1,384	2,813
	N	81	81
VAR00006	Pearson Correlation	,991**	1
	Sig. (2-tailed)	,000	
	Sum of Squares and Cross-products	225,011	465,288
	Covariance	2,813	5,816
	N	81	81

**. Correlation is significant at the 0.01 level (2-tailed).

Correlations

			VAR00001	VAR00006
Kendall's tau_b	VAR00001	Correlation Coefficient	1,000	,927**
		Sig. (2-tailed)	.	,000
		N	81	81
	VAR00006	Correlation Coefficient	,927**	1,000
		Sig. (2-tailed)	,000	.
		N	81	81
Spearman's rho	VAR00001	Correlation Coefficient	1,000	,991**
		Sig. (2-tailed)	.	,000
		N	81	81
	VAR00006	Correlation Coefficient	,991**	1,000
		Sig. (2-tailed)	,000	.
		N	81	81

**. Correlation is significant at the 0.01 level (2-tailed).

Рис. 4.2. Результати кореляційного аналізу, отримані за допомогою пакета SPSS, у випадку лінійного зв'язку

Бачимо, що у цьому випадку, як коефіцієнт кореляції Пірсона, так і рангові коефіцієнти кореляції достатньо точно визначають наявність лінійного зв'язку між ознаками.

В електронних таблицях MS Excel є вбудовані засоби для розрахунку коефіцієнта кореляції Пірсона. Це функція КОРРЕЛ () та процедура "Кореляція" пакета аналізу, який викликають з пункту меню "Сервіс/Аналіз даних". Якщо пакет аналізу не встановлено, то це можна зробити, користуючись пунктом меню "Сервіс/Надбудови".

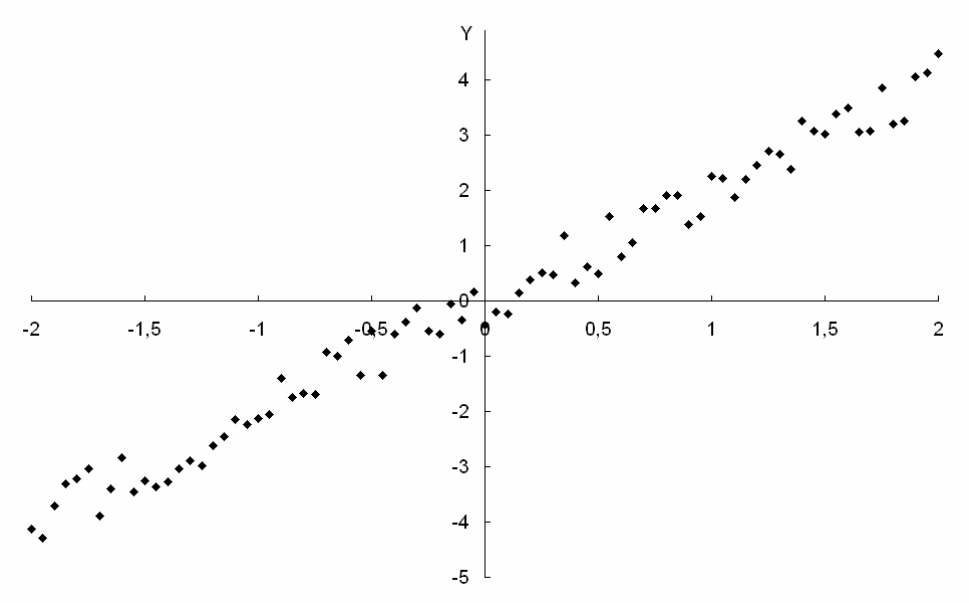


Рис. 4.3. Графік зв'язку між досліджуваними ознаками у випадку лінійного зв'язку

Різниця між цими засобами полягає в тому, що за допомогою пакета аналізу ми отримуємо кореляційну матрицю для даних, що розташовані у декількох сусідніх стовпчиках або рядках робочого аркушу.

При спробі розрахувати коефіцієнти кореляції для даних, між якими є розриви, буде отримано повідомлення про помилку. Якщо ж ми використовуємо функцію КОРРЕЛ (), то вимога щодо відсутності розривів між даними не висувається, але ця функція дає змогу розраховувати лише значення коефіцієнта кореляції між двома множинами даних.

Для прикладу, що розглядається, в обох випадках одержуємо одне й те саме значення коефіцієнта кореляції 0,991.

Розглянемо інший приклад. Елементи другої послідовності у цьому випадку розрахуємо за формулою: $y_i = 2\left(x_i^2 + \varepsilon_i\right)$, де ε_i – елементи згенерованої за допомогою електронних таблиць MS Excel рівномірної випадкової послідовності, заданої на відрізку $\left[-0,5;\, 0,5\right]$.

На рис. 4.4 наведено графік, що демонструє зв'язок між ознаками, а результати, одержані за допомогою пакета SPSS, – на рис. 4.5.

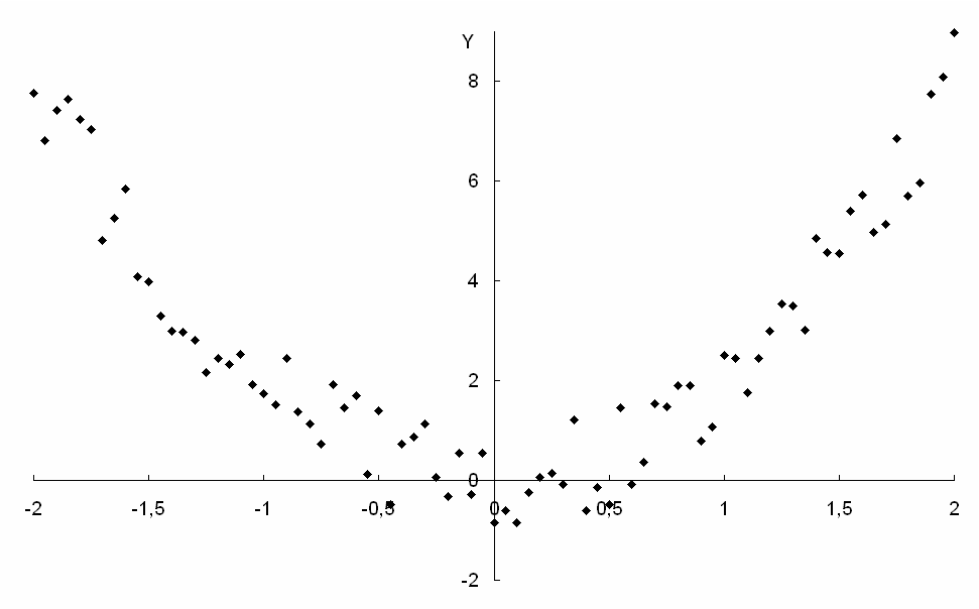


Рис. 4.4. Графік зв’язку між досліджуваними ознаками у випадку нелінійного зв’язку

Correlations

		VAR00001	VAR00003
VAR00001	Pearson Correlation	1	,030
	Sig. (2-tailed)		,791
	Sum of Squares and Cross-products	110,700	7,222
	Covariance	1,384	,090
	N	81	81
VAR00003	Pearson Correlation	,030	1
	Sig. (2-tailed)	,791	
	Sum of Squares and Cross-products	7,222	525,735
	Covariance	,090	6,572
	N	81	81

Correlations

			VAR00001	VAR00003
Kendall's tau_b	VAR00001	Correlation Coefficient	1,000	-,009
		Sig. (2-tailed)	.	,903
		N	81	81
	VAR00003	Correlation Coefficient	-,009	1,000
		Sig. (2-tailed)	,903	.
		N	81	81
Spearman's rho	VAR00001	Correlation Coefficient	1,000	,019
		Sig. (2-tailed)	.	,864
		N	81	81
	VAR00003	Correlation Coefficient	,019	1,000
		Sig. (2-tailed)	,864	.
		N	81	81

Рис. 4.5. Результати кореляційного аналізу, отримані за допомогою пакета SPSS, у випадку нелінійного зв’язку

Бачимо, що у цьому випадку, коефіцієнти кореляції, що розраховуються у пакеті SPSS, не виявляють наявного нелінійного зв'язку. Всі розраховані коефіцієнти є близькими до нуля.

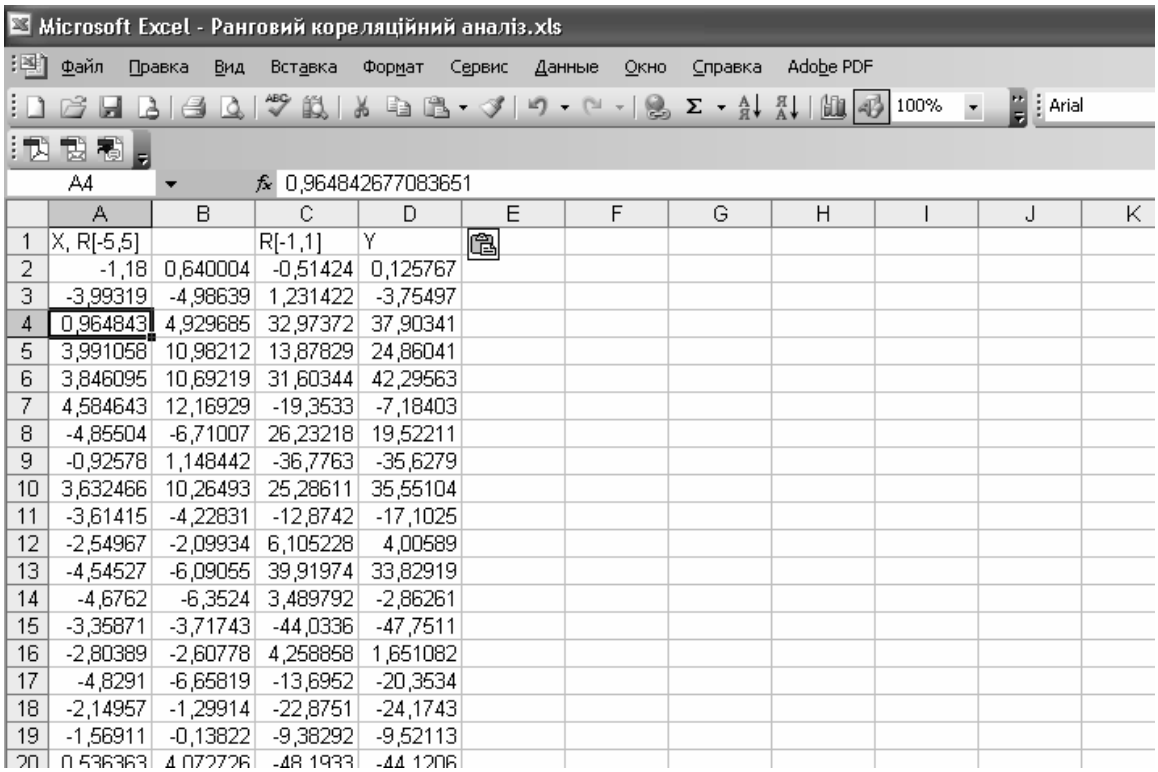
На жаль, у пакеті SPSS та електронних таблицях MS Excel не передбачено можливості розрахунку коефіцієнта детермінації. Тому розрахуємо його за допомогою електронних таблиць MS Excel, використовуючи можливість створення власних розрахункових формул. Вибірковий коефіцієнт детермінації певної ознаки y за вектором незалежних ознак $\mathbf{X} = \left(x^{(1)}, x^{(2)}, ..., x^{(p)} \right)$ можна розрахувати за формулами (4.3; 4.4; 4.6).

У досліджуваному випадку розраховане значення коефіцієнта детермінації дорівнює 0,946, що свідчить про наявність сильного зв'язку між ознаками.

Перевірка гіпотези про наявність зв'язку між порядковими ознаками

Електронні таблиці MS Excel не містять вбудованих доданків, які давали б змогу розраховувати коефіцієнти рангової кореляції. Проте вони дають змогу зробити для цього розрахункову форму.

Проілюструємо це за допомогою такого прикладу. На робочому аркуші електронних таблиць MS Excel сформуємо два стовпчики, що містять елементи досліджуваних вибірок (рис. 4.6).



	A	B	C	D	E	F	G	H	I	J	K
1	X, R[-5,5]		R[-1,1]	Y							
2	-1,18	0,640004	-0,51424	0,125767							
3	-3,99319	-4,98639	1,231422	-3,75497							
4	0,964843	4,929685	32,97372	37,90341							
5	3,991058	10,98212	13,87829	24,86041							
6	3,846095	10,69219	31,60344	42,29563							
7	4,584643	12,16929	-19,3533	-7,18403							
8	-4,85504	-6,71007	26,23218	19,52211							
9	-0,92578	1,148442	-36,7763	-35,6279							
10	3,632466	10,26493	25,28611	35,55104							
11	-3,61415	-4,22831	-12,8742	-17,1025							
12	-2,54967	-2,09934	6,105228	4,00589							
13	-4,54527	-6,09055	39,91974	33,82919							
14	-4,6762	-6,3524	3,489792	-2,86261							
15	-3,35871	-3,71743	-44,0336	-47,7511							
16	-2,80389	-2,60778	4,258858	1,651082							
17	-4,8291	-6,65819	-13,6952	-20,3534							
18	-2,14957	-1,29914	-22,8751	-24,1743							
19	-1,56911	-0,13822	-9,38292	-9,52113							
20	0,536363	4,072726	-48,1933	-44,1206							

Рис. 4.6. Фрагмент робочого аркуша з даними

Перший стовпчик сформуємо як рівномірну випадкову послідовність обсягом 100 елементів, задану на відрізку [–5; 5]. Обираємо в головному меню пункт “Сервіс” й підпункти “Аналіз даних”, “Генерація випадкових чисел”. При цьому з’являється діалогове вікно генератора випадкових чисел (рис. 4.7).

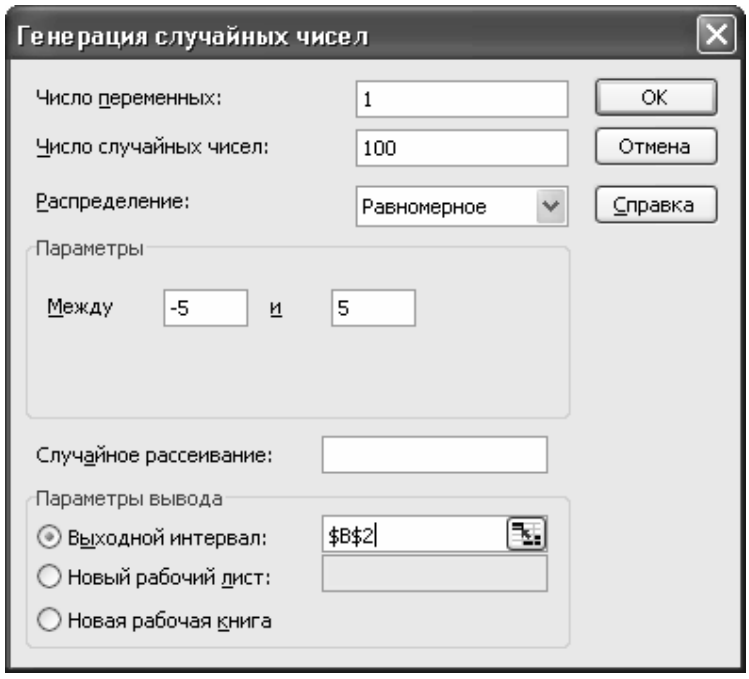


Рис. 4.7. Діалогове вікно генератора випадкових чисел

У цьому вікні задаємо: кількість змінних (кількість вибірок, які необхідно сформувати); кількість випадкових чисел (обсяг кожної вибірки); тип розподілу; параметри розподілу, а також посилання на верхню ліву комірку діапазону, в якому мають бути розташовані згенеровані дані.

Потім згенеруємо елементи другої вибірки, використовуючи формулу:

$$=2*A2+3+C3,$$

де A2 – посилання на комірку, де міститься значення відповідного елемента першої вибірки, а C3 – на комірку з елементом рівномірної випадкової послідовності, заданої на відрізку $[-b; b]$.

Після цього сформуємо стовпчики, що містять значення рангів елементів досліджуваних вибірок. Для цього скористаємося формулою:

$$=РАНГ\ (A2;\$A\$2:\$A\$101;1),$$

де A2 – посилання на комірку, яка містить елемент, ранг якого необхідно визначити; $A$2:$A$101 – посилання на діапазон комірок, де містяться усі елементи досліджуваних вибірок; 1 – вказівка на те, що ранжирування елементів необхідно здійснювати за зростанням.

Далі формуємо стовпчик квадратів різниць рангів, та в окремій комірці записуємо суму квадратів цих різниць. Потім розраховуємо коефіцієнт кореляції Спірмена, використовуючи формулу:

$$=1–6*H102/(K1^3-K1),$$

де H102 – посилання на комірку, у якої записано суму квадратів різниць рангів, K1 – посилання на комірку, де записано кількість елементів у кожній вибірці. На рис. 4.8, 4.9 наведено результати розрахунку рангового коефіцієнта кореляції Спірмена для різних значень параметра *b*, а також відповідні кореляційні поля досліджуваних ознак, а на рис. 4.10 – залежність коефіцієнта кореляції Спірмена від значення параметра *b*.

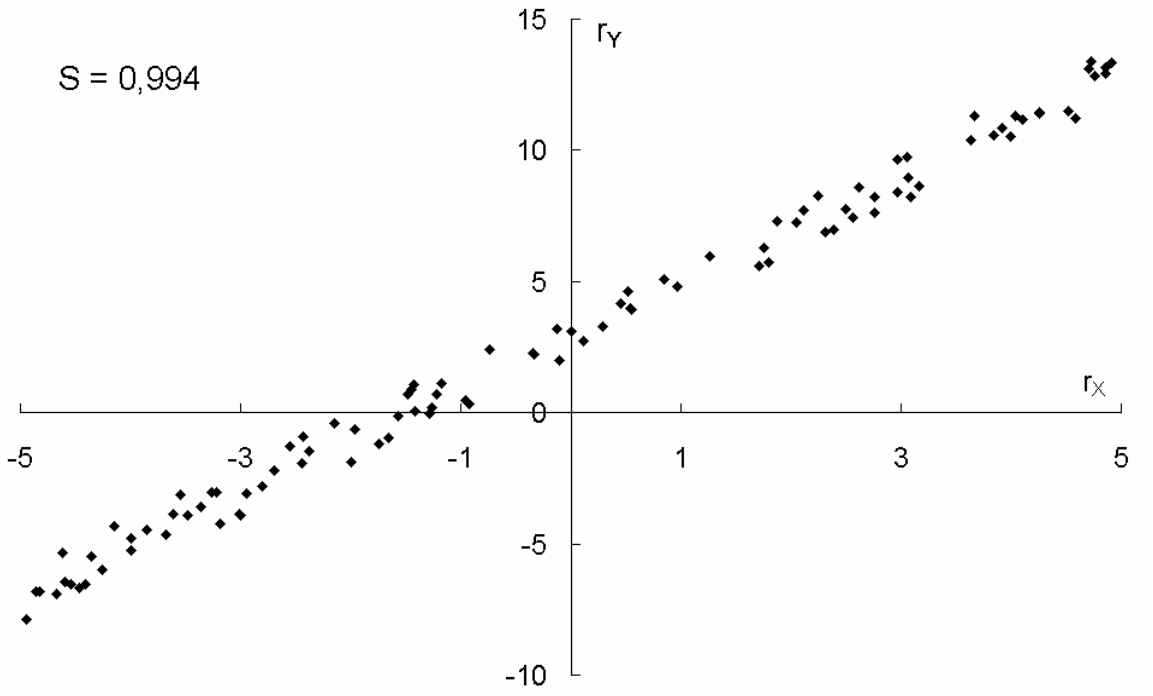


Рис. 4.8. Кореляційне поле досліджуваних ознак при *b* = 1

Більш зручним для дослідження рангової кореляції є застосування спеціалізованих статистичних пакетів, зокрема пакету SPSS. Розглянемо його використання на тому самому прикладі, що і у попередньому випадку.

До вікна даних заносимо стовпчики із значеннями досліджуваних вибірок (рис. 4.11). У пункті меню Analyze обираємо Correlate/ Bivariate Correlations. Після цього з'являється вікно вибору параметрів цієї процедури (рис. 4.12). У ньому треба позначити, між якими змінними шукатимемо кореляцію, які саме коефіцієнти кореляції необхідно розрахувати, а також який тип гіпотези (однобічну чи двобічну) ми розглядаємо.

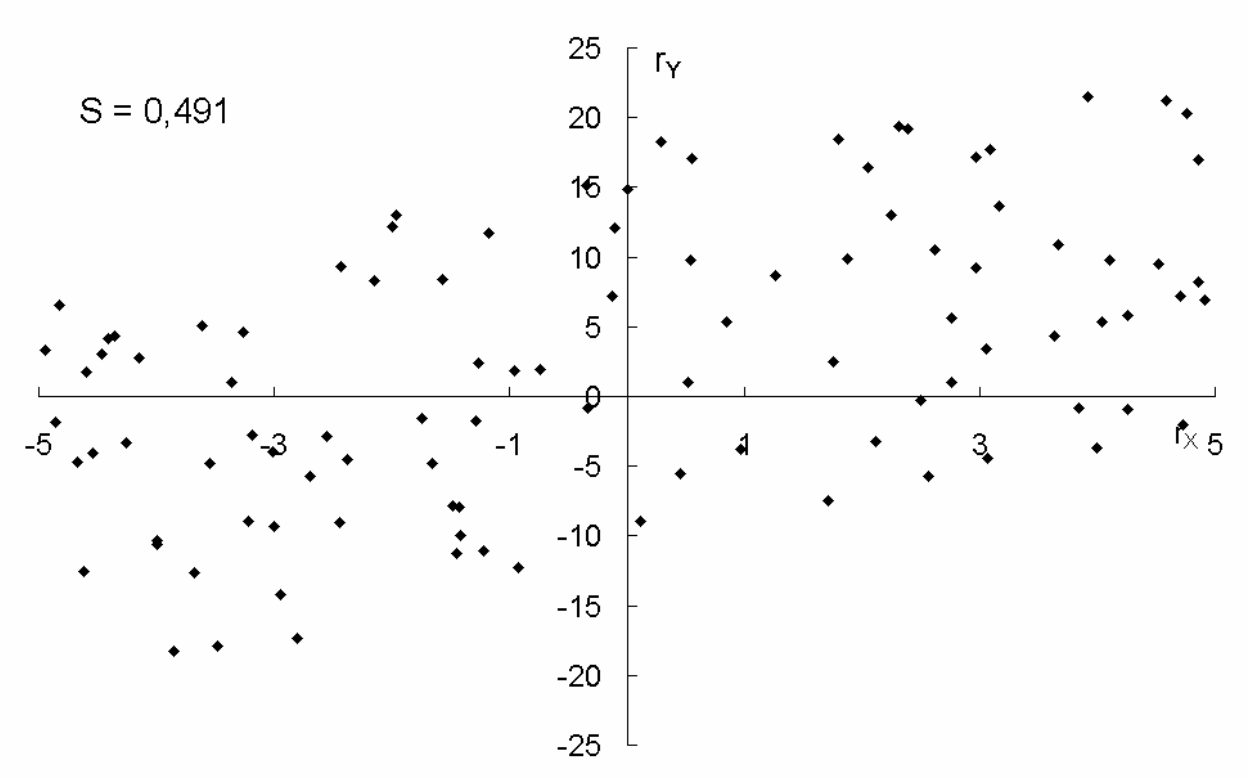


Рис. 4.9. Кореляційне поле досліджуваних ознак при $b = 15$

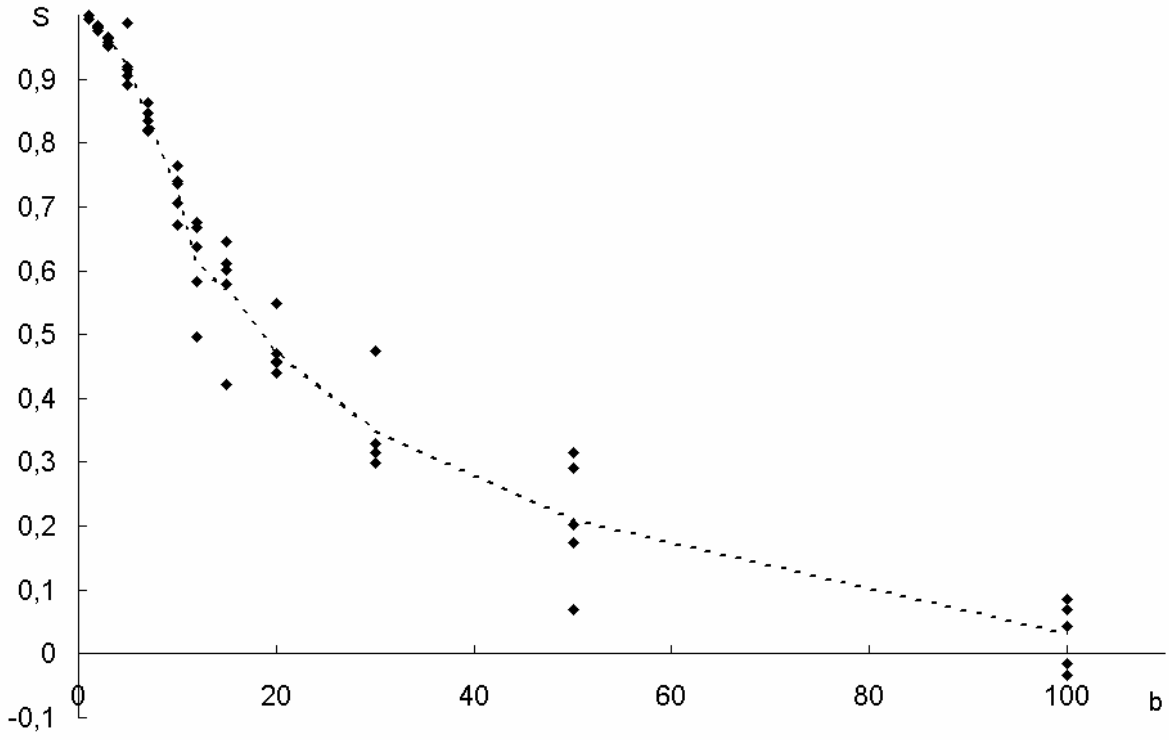


Рис. 4.10. Залежність коефіцієнта Спірмена від параметра b

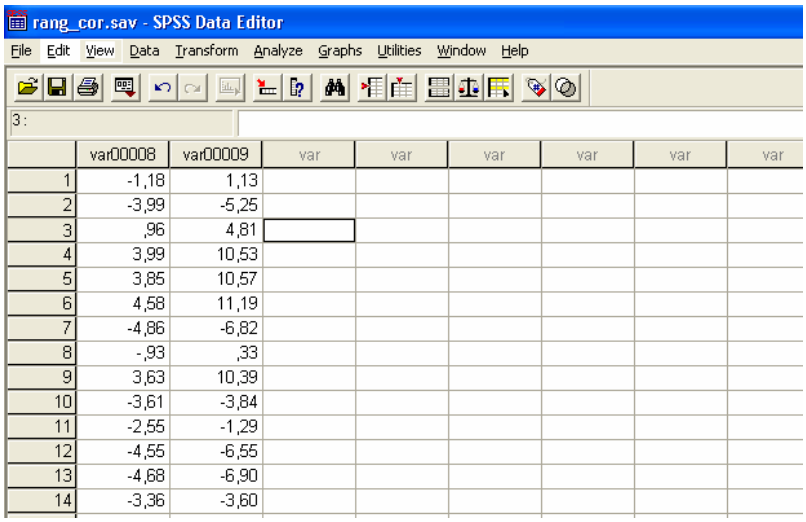


Рис. 4.11. Вікно даних пакету SPSS

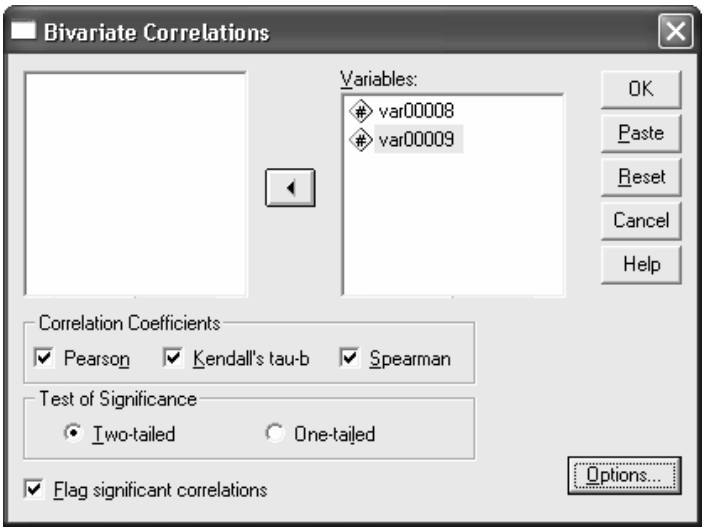


Рис. 4.12. Вікно вибору параметрів кореляційного аналізу

У вікні Options (рис. 4.13) зазначаємо, які додаткові статистичні параметри необхідно розрахувати, а також спосіб обробки пропущених даних.

Результати аналізу наведено на рис. 4.14. Бачимо, що всі коефіцієнти, що розраховувалися є достатньо близькими між собою. Це пов'язано насамперед з тим, що зв'язок між досліджуваними змінними є близьким до лінійного.

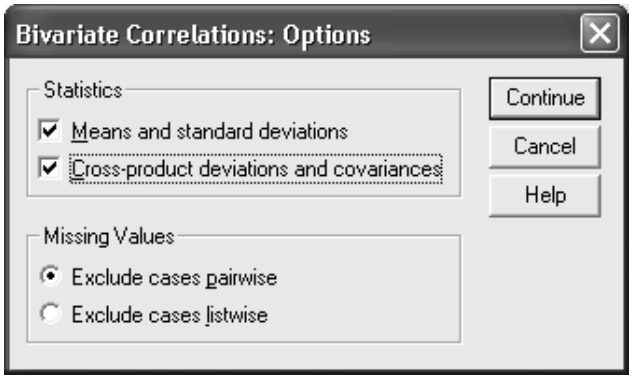


Рис. 4.13. Вікно задання додаткових параметрів кореляційного аналізу

Descriptive Statistics

	Mean	Std. Deviation	N
VAR00008	-,1454	3,09236	100
VAR00009	2,6854	6,23077	100

Correlations

		VAR00008	VAR00009
VAR00008	Pearson Correlation	1	,996**
	Sig. (2-tailed)	,	,000
	Sum of Squares and Cross-products	946,704	1899,604
	Covariance	9,563	19,188
	N	100	100
VAR00009	Pearson Correlation	,996**	1
	Sig. (2-tailed)	,000	,
	Sum of Squares and Cross-products	1899,604	3843,427
	Covariance	19,188	38,822
	N	100	100

**. Correlation is significant at the 0.01 level (2-tailed).

Correlations

			VAR00008	VAR00009
Kendall's tau_b	VAR00008	Correlation Coefficient	1,000	,939**
		Sig. (2-tailed)	,	,000
		N	100	100
	VAR00009	Correlation Coefficient	,939**	1,000
		Sig. (2-tailed)	,000	,
		N	100	100
Spearman's rho	VAR00008	Correlation Coefficient	1,000	,994**
		Sig. (2-tailed)	,	,000
		N	100	100
	VAR00009	Correlation Coefficient	,994**	1,000
		Sig. (2-tailed)	,000	,
		N	100	100

**. Correlation is significant at the .01 level (2-tailed).

Рис. 4.14. Результати кореляційного аналізу

Розглянемо далі приклад параболічної моделі. Першу вибірку згенеруємо так само, як і у попередньому випадку. Потім згенеруємо елементи другої вибірки, використовуючи формулу:

$$=2*A2*A2+3+C3,$$

де A2 – посилання на комірку, де міститься значення відповідного елемента першої вибірки, а C3 – на комірку з елементом рівномірної випадкової послідовності, заданої на відрізку $[-b; b]$.

На рис. 4.15, 4.16 наведено результати розрахунку рангового коефіцієнта кореляції Спірмена для різних значень параметра b, а також відповідні кореляційні поля досліджуваних ознак. Але для коефіцієнта Спірме-

на у цьому випадку розраховували три значення: S – відповідає усієї сукупності даних; S_- – значенням $x < 0$, S_+ – значенням $x > 0$.

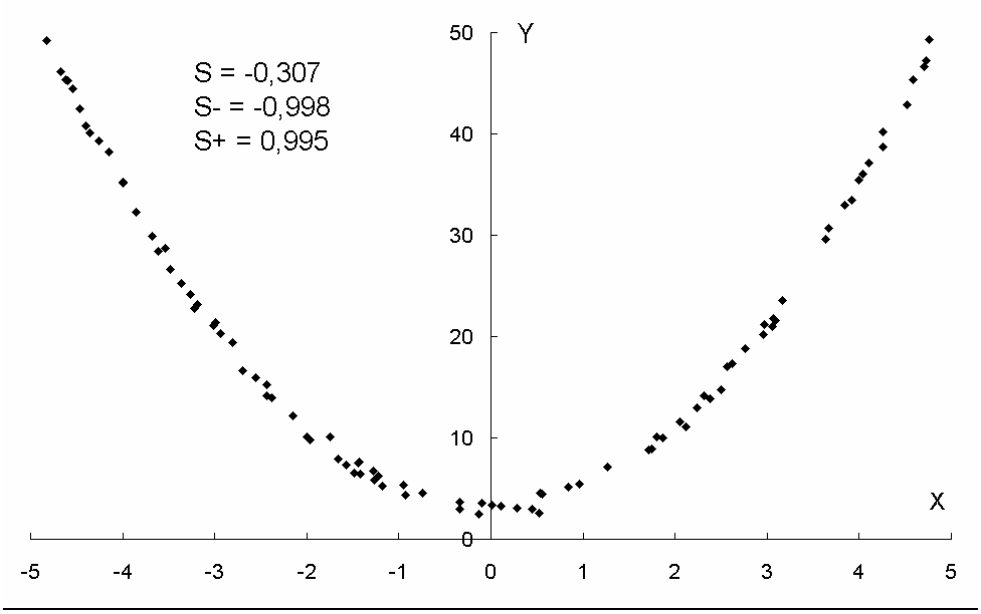


Рис. 4.15. Кореляційне поле досліджуваних ознак при $b = 1$

Безпосереднє застосування процедур кореляційного аналізу пакету SPSS не надає значних переваг при проведенні кореляційного аналізу досліджуваних ознак. Але цей пакет дає змогу визначити коефіцієнт детермінації, що відповідає окремим типам моделей зв'язку між досліджуваними ознаками. Застосування відповідних процедур можливо лише для кількісних даних.

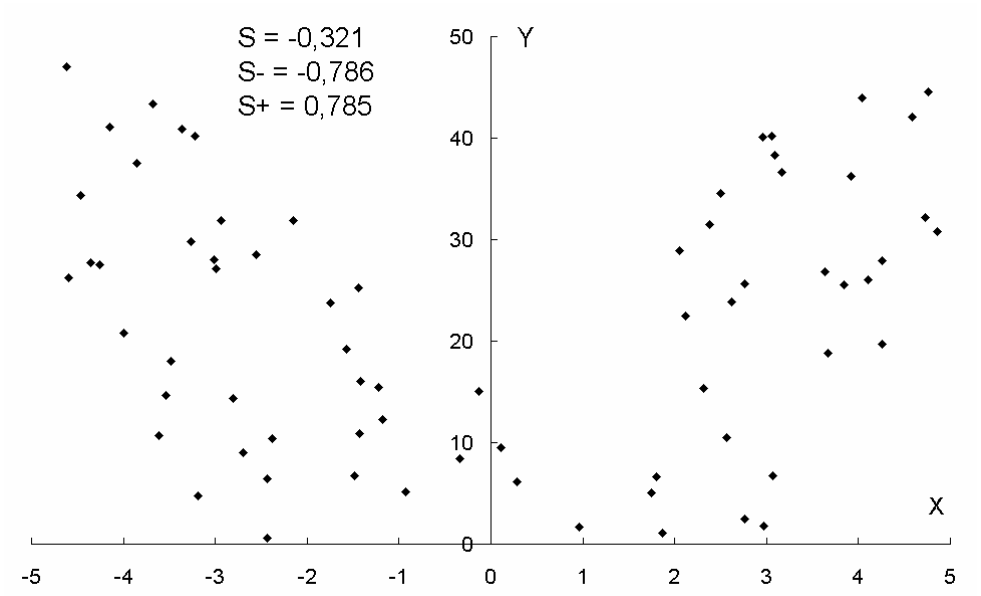


Рис. 4.16. Кореляційне поле досліджуваних ознак при $b = 20$

Розглянемо більш докладно методику визначення коефіцієнта детермінації та інших характеристик моделі зв'язку в пакеті SPSS. У пункті Analyze головного меню обираємо Regression/Curve Estimation. Після цього з'являється діалогове вікно підбору моделі зв'язку (рис. 4.17).

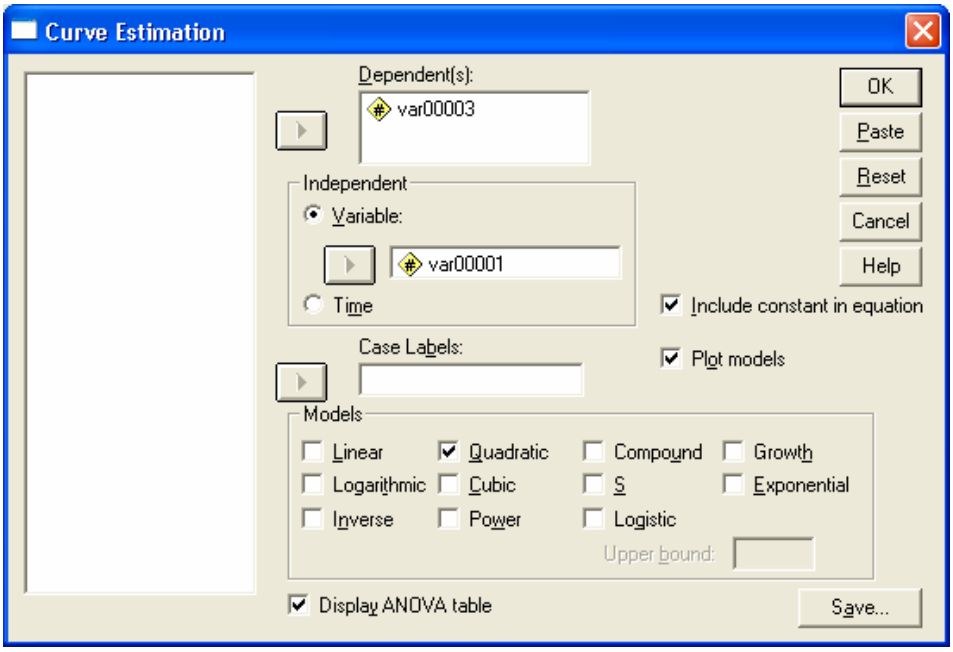


Рис. 4.17. Діалогове вікно підбору моделі зв'язку

У цьому вікні зазначаємо залежну й незалежну змінні, тип моделі, необхідність виведення таблиці ANOVA, графіка, а також збереження результатів на сторінці даних. Результати для розглянутих вище значень параметра *b* наведено на рис. 4.18–4.21.

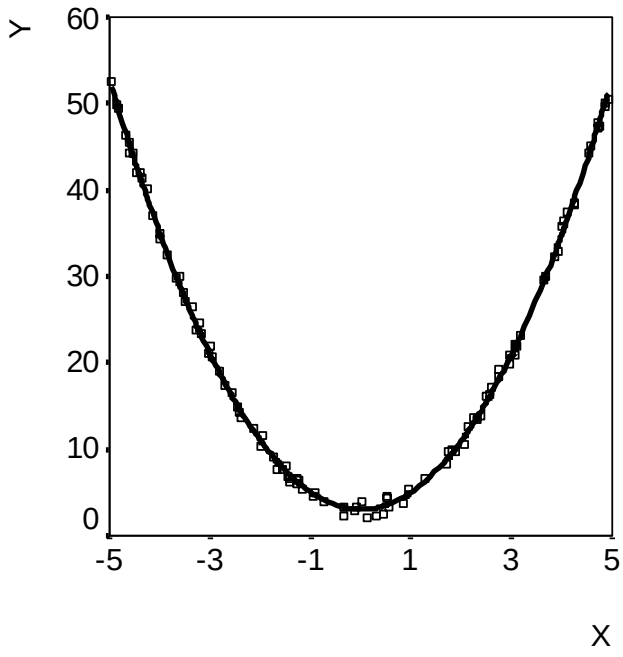


Рис. 4.18. Параболічна модель для $b = 1$

```
MODEL:  MOD_1.

Dependent variable.. VAR00003          Method.. QUADRATI

Listwise Deletion of Missing Data

Multiple R           ,99933
R Square             ,99867
Adjusted R Square    ,99864
Standard Error       ,57182

            Analysis of Variance:

               DF   Sum of Squares      Mean Square

Regression      2        23761,560      11880,77995
Residuals      97           31,717           ,32698

F =   36334,75101       Signif F =  ,0000

-------------------- Variables in the Equation -------------------

Variable                   B        SE B       Beta         T  Sig T

VAR00001     4,24885409E-05     ,018585  8,475E-06       ,002 ,9982
VAR00001**2       1,995714     ,007403    ,999333   269,570 ,0000
(Constant)        3,016919     ,090626                33,290 ,0000
```

Рис. 4.19. Характеристики параболічної моделі для $b = 1$

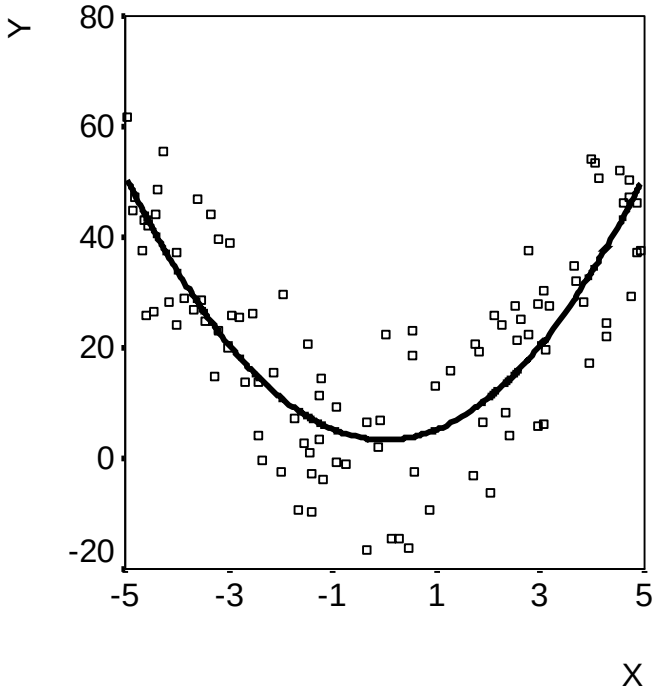


Рис. 4.20. Параболічна модель для $b = 20$

```
MODEL:  MOD_2.

Dependent variable.. VAR00009          Method.. QUADRATI

Listwise Deletion of Missing Data

Multiple R           ,79548
R Square             ,63278
Adjusted R Square    ,62521
Standard Error     11,43645

            Analysis of Variance:

               DF   Sum of Squares      Mean Square

Regression      2        21861,932        10930,966
Residuals      97        12686,870          130,792

F =      83,57488       Signif F =  ,0000

-------------------- Variables in the Equation -------------------

Variable                   B        SE B       Beta        T  Sig T

VAR00008            ,000850     ,371697    ,000141     ,002  ,9982
VAR00008**2        1,914277     ,148066    ,795477    12,928 ,0000
(Constant)         3,338370    1,812522                1,842  ,0686
```

Рис. 4.21. Характеристики параболічної моделі для $b = 20$

Наведені дані свідчать про те, що значення коефіцієнту рангової кореляції, розраховані окремо за спадною та висхідною гілками параболи, є близькими до квадратного кореня з коефіцієнта детермінації. Це підтверджує те, що у даному випадку коефіцієнт рангової кореляції Спірмена є адекватною мірою нелінійного зв'язку між ознаками, якщо він розраховується за інтервалами, що відповідають ділянкам монотонного спадання або згасання функції регресії.

Контрольні питання

1. Що називають кореляцією двох випадкових величин?
2. Які ознаки вважають статистично незалежними?
3. Які проблеми аналізу даних потребують попередньої перевірки наявності статистичного зв'язку між досліджуваними ознаками?
4. Якою є загальна методика перевірки гіпотези про наявність статистичного зв'язку?
5. З якою метою на початковому етапі кореляційного аналізу перевіряють тип даних?
6. Що є універсальною характеристикою статистичного зв'язку між кількісними ознаками?

7. Якими є переваги й недоліки застосування коефіцієнта детермінації?

8. Для заданого набору даних визначити коефіцієнт детермінації і зробити висновок про наявність кореляційного зв'язку.

9. У чому полягає різниця між парними та частинними кореляційними характеристиками?

10. Що вимірює парний коефіцієнт кореляції Пірсона?

11. Для заданого набору даних розрахувати значення парного коефіцієнта кореляції Пірсона і зробити висновок про наявність кореляційного зв'язку.

12. Що називають кореляційним відношенням двох випадкових величин? Які властивості зв'язку характеризує цей показник?

13. Що вимірює коефіцієнт кореляції Фехнера?

14. Для заданого набору даних розрахувати значення коефіцієнта кореляції Фехнера і зробити висновок про наявність кореляційного зв'язку.

15. Що називають коваріацією випадкових величин? Як цей показник пов'язаний з парним коефіцієнтом кореляції Пірсона?

16. Яких значень можуть набувати показники корельованості кількісних ознак? Які висновки можна зробити на основі значень цих показників?

17. Що називають ранговою кореляцією?

18. На якій властивості корельованих ознак ґрунтується коефіцієнт рангової кореляції Спірмена?

19. На якій властивості корельованих ознак ґрунтується коефіцієнт рангової кореляції Кендалла?

20. Яких значень можуть набувати показники рангової кореляції і про що свідчать їх значення?

21. Для заданого набору даних розрахувати значення коефіцієнтів рангової кореляції Спірмена й Кендалла і зробити висновок про наявність кореляційного зв'язку.

22. Які показники використовують для перевірки корельованості номінальних ознак? У чому полягають особливості застосування окремих показників?

23. Для заданого набору даних перевірити корельованість даних за допомогою критерію χ^2, а також коефіцієнта Крамера й поліхоричного коефіцієнта спряженості Чупрова.

24. Які показники використовують для перевірки корельованості ознак, що виміряні у шкалах різного типу? Якими є особливості їх застосування?

25. Які показники використовують для дослідження корельованості декількох ознак? Якими є особливості їх застосування?

5. ФАКТОРНИЙ АНАЛІЗ

При дослідженні складних систем часто немає можливості безпосередньо вимірювати величини, що визначають їх властивості (**фактори**). Більше того, нерідко є невідомими кількість та зміст цих факторів. Але можуть вимірюватися інші величини, що залежать від них. Якщо невідомий фактор впливає на декілька вимірюваних ознак, останні виявляють певний зв'язок, наприклад корельованість, між собою. Тому загальна кількість факторів може бути значно меншою, ніж кількість вимірюваних ознак. Для виявлення таких факторів використовують факторний аналіз. Зменшення кількості факторів може бути необхідним також для забезпечення збіжності алгоритмів подальшого аналізу даних, скорочення ресурсів пам'яті ЕОМ та часу, потрібних для їх обробки, бажанням візуалізувати отримані результати тощо.

Основні ідеї факторного аналізу було сформульовано Ф. Гальтоном наприкінці XIX ст. Пізніше значний внесок у розвиток його методології зробили Р. Кеттелл, К. Пірсон, Ч. Спірмен, Л. Терстоун, Г. Хотеллінг та інші фахівці.

Формально задачу факторного аналізу можна записати у такій спосіб. Є масив p-вимірних спостережень:

$$X_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \dots \\ x_{ip} \end{pmatrix}, \quad i = 1, 2, ..., n, \tag{5.1}$$

де n – кількість спостережень (об'єктів, станів);

p – кількість параметрів, що характеризують кожне спостереження.

Необхідно подати результати у вигляді нового масиву:

$$Z_i = \begin{pmatrix} z_{i1} \\ z_{i2} \\ \dots \\ z_{ip'} \end{pmatrix}, \quad i = 1, ..., n, \tag{5.2}$$

такого, що p' є значно нижчим, ніж p. Компоненти вектора Z' називають **факторами**. На практиці зазвичай прагнуть, щоб виконувалася одна з умов: $p' = (0,1...0,25)p$ або $p' = 1...3$.

Першим етапом факторного аналізу зазвичай є вибір нових ознак (факторів), які є лінійними комбінаціями старих і відображають переваж-

ну частку загальної мінливості вихідних даних. Тому вони зберігають основну частину інформації, що містили ці дані. Другим етапом є обертання факторів з метою спрощення їх інтерпретації.

Об'єктом дослідження методами факторного аналізу, як правило, 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 подання 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.

Розрізняють ***R*-техніку** та ***Q*-техніку** факторного аналізу. Перша з них розроблена британським психологом Реймондом Б. Кеттеллом і передбачає розрахунок коефіцієнтів кореляції між параметрами (ознаками), що утворюють матрицю вихідних даних. Її використовують для зменшення кількості параметрів. *Q*-техніку запропонував британський психолог В. Стефенсон в 1935–1936 р. й докладно описав Р.Б. Кеттелл у 1946 р. За її допомогою вивчають кореляцію між об'єктами або станами об'єктів. Її застосовують для зменшення кількості об'єктів. З формального погляду в першому випадку шукають кореляцію між стовпчиками таблиці спостережень (табл. 5.1), а у другому – між її рядками.

Таблиця 5.1

Загальний вигляд таблиці спостережень для факторного аналізу

Номери об'єктів (станів)	Параметри об'єктів (станів)			
	1	2	...	p
1	x_{11}	x_{12}	…	x_{1p}
2	x_{21}	x_{22}	…	x_{2p}
...	…	…	…	…
n	x_{n1}	x_{n2}	…	x_{np}

Крім того, розроблено ***P*-техніку**, яку використовують при дослідженні одного об'єкта, якщо значення його ознак вимірюють у різні моменти часу, а також ***O*-техніку**, ***S*-техніку** та ***T*-техніку**, які також докладно були описані Р. Кеттеллом в 1940 р. Понад 50% усіх завдань, що передбачають застосування факторного аналізу, вирішують за допомогою *R*-техніки.

Основними методами факторного аналізу є методи головних компонент, головних факторів, 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.

Метою факторного аналізу є зменшення кількості змінних та визначення структури взаємозв'язків між змінними (класифікація даних). З формального погляду його метою є одержання матриці факторного відображення. Її рядки є координатами кінців векторів, що відповідають *n* змінним у *p*'-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 її 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.

Обов'язковими умовами факторного аналізу є такі:

– всі досліджувані ознаки мають бути кількісними;

– кількість ознак має бути принаймні вдвічі більшою, ніж кількість змінних;

– вибірка має бути однорідною;

– вихідні змінні повинні мати симетричний розподіл.

5.1. Метод головних компонент

Метод головних компонент, або **компонентний аналіз** вперше був запропонований К. Пірсоном у 1901 р., який розглядав задачу найкращої (з погляду мінімізації суми квадратів відхилень) апроксимації сукупності точок прямими та площинами. Потім він був докладно розроблений американським статистиком й економістом Гарольдом Хотеллінгом у 1933 р. Його важливою перевагою є те, що він є єдиним математично обґрунтованим методом факторного аналізу.

За своєю сутністю метод полягає у виборі нової ортогональної системи координат у просторі спостережень. Як першу головну компоненту обирають напрям, вздовж якого масив спостережень має найбільшу дисперсію. Кожну наступну компоненту обирають також з умови максимізації частки дисперсії, що залишилася, вздовж неї, доповненої умовою ортогональності всім раніше обраним компонентам. При цьому 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:

$$\widetilde{Z} = LX, \tag{5.3}$$

де X – p-вимірна випадкова величина з вектором середніх значень $\vec{a} = \left(a_1, \ ..., \ a_p\right)$ і коваріаційною матрицею $\Sigma = \left(\sigma_{ij}\right)$, $i, j = 1, \ \ldots, p$;

$$L = \begin{pmatrix} \ell_{11} & \ldots & \ell_{1p} \\ \ldots & \ldots & \ldots \\ \ell_{p'1} & \ldots & \ell_{p'p} \end{pmatrix} \tag{5.4}$$

– **матриця факторного відображення**, рядки якої задовольняють умову ортогональності.

Мірою інформативності факторів є величина:

$$I_{p'}\left(Z\right) = \frac{D_{z_1} + \ldots + D_{z_{p'}}}{D_{x_1} + \ldots + D_{x_p}}, \tag{5.5}$$

де D – дисперсія.

Матрицю факторного відображення вибирають з умови:

$$I_{p'}\left(\widetilde{Z}\right) = \max_{Z(X) \in F} I_{p'}\left(Z\right), \tag{5.6}$$

де $F(X)$ – клас допустимих перетворень досліджуваних ознак.

Першою головною компонентою $\widetilde{Z}_1\left(X\right)$ досліджуваної системи показників $X = \left(x_1, ..., x_p\right)^T$ називають таку нормовано-центровану лінійну комбінацію цих показників, яка серед усіх інших нормовано-центрованих лінійних комбінацій змінних $x_1, \ \ldots, \ x_p$ має найбільшу дисперсію.

***K*-ю головною компонентою** $\widetilde{Z}_k\left(X\right)$ досліджуваної системи показників $X = \left(x_1, ..., x_p\right)^T$ називають таку нормовано-центровану лінійну ком-

бінацію цих показників, яка не корельована з $k-1$ попередніми головними компонентами і серед усіх інших нормовано-центрованих і некорельованих з $k-1$ попередніми головними компонентами лінійних комбінацій змінних $x_1, \ldots, x_p$ має найбільшу дисперсію.

Центрування змінних здійснюють шляхом перетворення:

$$\tilde{x}_i^j = x_i^j - \overline{x^j} = x_i^j - a^j,\ i, j = 1, ..., p. \tag{5.7}$$

Нормування змінних здійснюють перетворенням:

$$x^{*i}_{\ v} = \frac{\tilde{x}_v^i}{\sqrt{\hat{\sigma}_{ii}}},\ \hat{\sigma}_{kj} = \frac{\sum_{v=1}^{n}\left(x_v^k - \overline{x^k}\right)\left(x_v^j - \overline{x^j}\right)}{n}; j, k = 1, ..., p;\ v = 1, ..., n. \tag{5.8}$$

Для визначення головних компонент розраховують власні числа й власні вектори кореляційної матриці, розв'язуючи рівняння:

$$\left|\Sigma - \lambda I\right| = 0, \tag{5.9}$$

де I – одинична матриця.

Для дійсної симетричної матриці розміром $p \times p$ це рівняння має p дійсних коренів $\lambda_1, ..., \lambda_p$, які є власними числами. Можна довести, що $D_{\tilde{z}_1} = \lambda_1$, де $\lambda_1 = \max_j \lambda_j$.

Кореляційну матрицю розміром $p \times p$ обчислюють за формулою:

$$R = \frac{1}{p-1} X^* X^{*T}. \tag{5.10}$$

На головній діагоналі кореляційної матриці R стоять значення, які дорівнюють одиниці.

Розв'язуючи систему:

$$\left(\Sigma - \lambda_1 I\right)\ell_1^T = 0, \tag{5.11}$$

отримуємо значення компонент власного вектора ℓ_1. Потім аналогічно розраховують інші власні вектори.

Враховуючи вказану вище властивість власних чисел, критерій інформативності можна записати у вигляді:

$$I_{p'} = \frac{\lambda_1 + ... + \lambda_{p'}}{\lambda_1 + ... + \lambda_p}, \tag{5.12}$$

де $\lambda_1, ..., \lambda_p$ – власні числа матриці Σ, впорядковані у порядку згасання.

Вибір критерію інформативності в методі головних компонент передбачає, що найбільш важливу інформацію про аналізовану систему можна

відобразити лінійною моделлю, яка відповідає такому вибору системи координат у тому самому просторі, що забезпечує максимальні дисперсії для проекцій досліджуваних об'єктів. Такий підхід є доцільним, якщо більшість вихідних ознак узгоджено впливає на властивість, що вивчається, і пригнічує вплив іррелевантних чинників на розподіл об'єктів. Адекватну модель можна отримати також у випадку, коли кількість пов'язаних інформативних ознак невелика, але вплив інших чинників є неузгодженим. У цьому разі не порушується однорідність еліпсоїда розсіювання, а лише зменшується його довгастість уздовж напрямку досліджуваної властивості.

У факторному аналізі використовують також інші міри інформативності, що дають змогу визначити кількість істотних факторів.

Критерій Кайзера, або критерій власних чисел, запропонований американським психологом Генрі Феліксом Кайзером, передбачає, що до моделі включають тільки фактори, для яких власні числа є не меншими, ніж одиниця. За змістом це означає, що таким факторам відповідає дисперсія, еквівалента принаймні дисперсії одної змінної. У противному випадку виокремлення фактора не має сенсу. Цей критерій іноді залишає в моделі занадто багато факторів.

Критерій кам'янистого осипу (**критерій відсіювання**) передбачає побудову графіка, де по осі абсцис відкладають порядковий номер власного числа, а по осі ординат – його значення. Згідно з Р. Кеттелом необхідно знайти точку найбільшого уповільнення спадання власних значень і враховувати лише фактори, яким відповідають власні числа, розташовані лівіше цієї точки. На відміну від попереднього цей критерій статистично необґрунтовані і часто залишає в моделі не всі істотні фактори. Втім у випадках, коли істотних факторів небагато, а кількість змінних є великою, обидва критерії є придатними для практичного застосування.

На практиці часто здійснюють розрахунки, використовуючи різні критерії, а потім обирають модель, що містить найбільшу кількість факторів, яким можна надати змістову інтерпретацію.

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, пояснюють 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 двохетапну процедуру аналізу. На першому етапі максимальну кількість факторів не задають. Після його проведення аналізують дисперсії, оцінюють приблизну кількість факторів і проводять повторний аналіз.

5.2. Метод головних факторів

Цей метод використовують для зменшення кількості змінних. У його основі лежить припущення, що не всі змінні, які вимірювали при дослідженні системи, є незалежними. Тому можливо формування нових змінних, що достатньо повно відображають наявну інформацію.

На відміну від методу головних компонент, метод головних факторів ґрунтується не на дисперсійному критерії інформативності множини ознак, а на поясненні кореляцій, що існують між цими ознаками. Він враховує, що вихідні дані можуть містити грубі помилки, які у багатовимірному аналізі призводять до помилок інтерпретації.

Тому метод головних факторів застосовують у більш складних випадках, зокрема за наявності сумісного прояву аналізованих й іррелевантних властивостей об'єктів, що є порівнянними за ступенем внутрішньої узгодженості, а також для виділення групи діагностичних показників із вихідної множини ознак.

Основну модель методу головних факторів записують у вигляді:

$$X^* = MF\,. \tag{5.13}$$

Матриця $\boldsymbol{X}^*$ є матрицею нормовано-центрованих значень вихідних ознак, що має розмірність $n \times p$.

У методі головних факторів припускають, що кожний елемент матриці X^* є результатом впливу m гіпотетичних **загальних факторів** та одного **характерного фактора**.

Характерні фактори вважають некорельованими один з одним, а також із загальними факторами. Загальні фактори пов'язані з істотними ваговими коефіцієнтами більше, ніж з одною ознакою. Ті з них, для яких істотними є всі вагові коефіцієнти, називають **генеральними факторами**.

Перший варіант методу, запропонований Ч. Спірменом на початку 1900 р. передбачав існування одного загального та одного характерного факторів. Пізніше у 1920 р. британський та американський психолог Раймонд Бернард Кеттел та американський психолог Карл Джон Хользінгер запропонували біфакторну модель, яка передбачала існування декількох, зазвичай двох, загальних факторів. Сучасний варіант методу головних факторів було запропоновано Г. Томсоном.

Повну факторну матрицю M можна подати як:

$$M = A + D\,, \tag{5.14}$$

де A та D – відповідно, матриці навантажень загальних і характеристичних факторів, які мають розмірність F.

Матриця A є матрицею (значущих або незначущих) вагових коефіцієнтів загальних факторів і записується у вигляді:

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} & 0 & 0 & \dots & 0 \\ a_{21} & a_{22} & \dots & a_{2m} & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_{p1} & a_{p2} & \dots & a_{pm} & 0 & 0 & \dots & 0 \end{pmatrix}. \qquad (5.15)$$

Її ненульова частина розміром $n \times m$ є матрицею факторного відображення. Всі вагові коефіцієнти цієї матриці при характерних факторах дорівнюють нулю. Рівняння (5.13), а також матрицю A називають **факторним відображенням**.

Матриця

$$D = \begin{pmatrix} 0 & 0 & \dots & 0 & d_1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & d_2 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & d_p \end{pmatrix} \qquad (5.16)$$

є матрицею вагових коефіцієнтів характерних факторів. Усі її вагові коефіцієнти для загальних факторів дорівнюють нулю.

Матриця F є матрицею значень факторів (загальних і характерних) для всіх об'єктів. Її можна записати у вигляді:

$$F = \begin{pmatrix} f_{11} & f_{12} & \dots & f_{1n} \\ f_{21} & f_{22} & \dots & f_{2n} \\ \dots & \dots & \dots & \dots \\ f_{m1} & f_{m2} & \dots & f_{mn} \\ \nu_{11} & \nu_{12} & \dots & \nu_{1n} \\ \nu_{21} & \nu_{22} & \dots & \nu_{2n} \\ \dots & \dots & \dots & \dots \\ \nu_{p1} & \nu_{p2} & \dots & \nu_{pn} \end{pmatrix}. \qquad (5.17)$$

Оскільки значення ознак є нормованими, їх загальні дисперсії дорівнюють одиниці. Можна показати, що дисперсія j-ї ознаки:

$$\hat{s}_j^2 = 1 = d_j^2 + h_j^2, \qquad (5.18)$$

де $h_j^2 = \sum_{r=1}^{m} a_{jr}^2$ – **загальність ознаки**, що є сумою відносних внесків m загальних факторів до дисперсії;

d_j^2 – **характерність ознаки**, яка відображає внесок характерного фактора ν_j до дисперсії ознаки x_j^*. У свою чергу, характерність можна подати у вигляді:

$$d_j^2 = b_j^2 + c_j^2, \tag{5.19}$$

де b_j^2 – **специфічність ознаки**, що безпосередньо відображає специфіку відповідного фактора;

c_j^2 – дисперсія похибки. Використовують також поняття **надійності**:

$$r_j^2 = h_j^2 + b_j^2 = 1 - c_j^2, \tag{5.20}$$

яка є часткою загальної дисперсії, непов'язаною із похибкою.

Повнотою факторизації називають величину:

$$k = \frac{V_0}{p} = \frac{\sum_{r=1}^{m} \nu_r}{p}. \tag{5.21}$$

Вихідна матриця $\boldsymbol{X}^*$ дає змогу отримати кореляційну матрицю R:

$$R = \frac{1}{n} X^* X^{*\mathrm{T}}. \tag{5.22}$$

Ураховуючи (5.9), її можна також записати у вигляді:

$$R = MM^T = AA^T + DD^T = R_h + D^2. \tag{5.23}$$

Матриця R є симетрическою й дійсною. Елементами її головної діагоналі є дисперсії відповідних ознак, які дорівнюють одиниці. Тому сумарна дисперсія всіх ознак дорівнює сумі діагональних елементів матриці R.

Матрицю R_h називають **редукованою (кореляційною) матрицею**. У випадку, коли допускається кореляція між загальними факторами,

$$R_h = R - D^2 = ACA^T, \tag{5.24}$$

де $C = \frac{1}{n} FF^T$ – матриця коефіцієнтів кореляції між факторами. Цей вираз називають **фундаментальною теоремою факторного аналізу**. Якщо загальні фактори некорельовані один з одним, то матриця C є одиничною й $R_h = AA^T$.

Діагональними елементами редукованої матриці є загальності. З (5.24) та властивостей кореляційної матриці випливає, що всі $h_j^2 < 1$. Для вирішення завдань факторного аналізу необхідно визначити їх оцінки за вихідними даними. Існує два підходи до їх отримання. У методах голо-

вних факторів, центроїдному та трикутної декомпозиції спочатку визначають загальності, а потім кількість факторів. У методах максимальної правдоподібності й максимальних залишків використовують іншу послідовність: спочатку задають кількість факторів m, а потім підбирають значення загальностей так, щоб ранг матриці R_h наблизився до m.

Нижньою межею загальності є квадрат множинного коефіцієнта кореляції j-ї ознаки з іншими ознаками, а верхньою – квадрат коефіцієнта надійності. Звідси:

$$R_j^2 \le h_j^2 \le r_j^2 . \tag{5.25}$$

При розробці методів оцінювання загальностей виходять із визначення Д. Лоулі й К. Рао, згідно з яким загальності – це величини, які при статистично значущих факторах дають змогу найкращим чином відтворити кореляційну матрицю.

Найчастіше використовують такі методи оцінювання загальностей.

Перший метод ґрунтується на тому, що із зростанням кількості ознак при сталій кількості факторів нижня межа оцінки загальності (5.25) збігається до її істинного значення. Завдяки цьому можна використовувати як оцінку формулу:

$$h_j^2 \approx 1 - \frac{1}{r^{jj}}, \tag{5.26}$$

де r^{jj} – j-й діагональний елемент матриці R^1.

Другий метод базується на припущенні, що за великої кількості ознак як значення загальності можна використовувати найбільший за модулем коефіцієнт кореляції даної ознаки (відповідних рядка або стовпчика) з іншими змінними, взятий зі знаком плюс. Цей метод немає теоретичного підґрунтя, але практика свідчить, що при кількості ознак понад 20 одержувані за його допомогою результати мало відрізняються від тих, що отримують за допомогою більш точних методів.

Згідно з третім методом як значення загальності беруть не максимальне, а середнє за відповідним рядком (стовпчиком) значення коефіцієнта кореляції:

$$h_j^2 = \frac{1}{p-1} \sum_{\substack{k=1 \\ k \ne j}}^{p} r_{jk} . \tag{5.27}$$

Й нарешті, в останньому способі (метод тріад) в j-му рядку (стовпчику) беруть два найбільші (за модулем) значення коефіцієнтів кореляції r_{jk} й r_{jl} $\left(k, l \ne j\right)$. Після цього значення загальності обчислюють за допомогою тріади:

$$h_j^2 = \frac{r_{jk} r_{jl}}{r_{kl}} . \tag{5.28}$$

Першим кроком алгоритму методу головних факторів є отримання матриці парних коефіцієнтів кореляції R, на головній діагоналі якої стоять одиниці. Наступним кроком є одержання редукованої матриці R_h із загальностями на головній діагоналі.

Далі визначають перший загальний фактор, виходячи з умови, що його внесок ν_1 до дисперсії процесу має бути максимальним, тобто:

$$\nu_1 = \sum_{j=1}^{p} a_{j1}^2 \to \max. \tag{5.29}$$

При цьому мають бути виконані умови:

$$r_{jk} = \sum_{r=1}^{m} a_{jr} a_{kr} \quad (j, k = 1, ..., p). \tag{5.30}$$

Для розв'язування цієї задачі можна скористатися методом невизначених множників Лагранжа, який призводить до системи:

$$\sum_{k=1}^{p} r_{jk} a_{k1} - \lambda_1 a_{j1} = 0, \tag{5.31}$$

де $\lambda_1 = \sum_{j=1}^{p} a_{j1}^2$.

Ураховуючи, що $r_{jj} = h_j^2$, її можна записати у вигляді:

$$\begin{cases} \left(h_1^2 - \lambda\right) a_{11} + r_{12} a_{21} + ... + r_{1p} a_{p1} = 0; \\ r_{21} a_{11} + \left(h_2^2 - \lambda\right) a_{21} + r_{2p} a_{p1} = 0; \\ ..... \\ r_{p1} a_{11} + r_{p2} a_{21} + ... + \left(h_p^2 - \lambda\right) a_{p1}. \end{cases} \tag{5.32}$$

Необхідною й достатньою умовою існування нетривіального розв'язку є рівність нулю визначника матриці коефіцієнтів системи (5.32):

$$\begin{vmatrix} h_1^2 - \lambda & r_{12} & ... & r_{1p} \\ r_{21} & h_2^2 - \lambda & ... & r_{2p} \\ ... & ... & ... & ... \\ r_{p1} & r_{p2} & ... & h_n^2 - \lambda \end{vmatrix} = 0. \tag{5.33}$$

Ліва частина (5.33) є характеристичним поліномом системи (5.32), який є поліномом p-ї степені стосовно λ. Оскільки матриця коефіцієнтів є дійсною та симетричною, всі корені (5.33) є дійсними.

Як λ_1 беруть найбільший корінь (5.33) й підставляють його до (5.32). Її розв'язком є вектор $\left(a_{11}, a_{21}, ..., a_{p1}\right)$. Для отримання компонентів першого загального фактора необхідно помножити відповідні елементи цього вектора на вагові коефіцієнти:

$$a_{j1} = a_{j1} \frac{\sqrt{\lambda_1}}{\sqrt{a_{11}^2 + a_{21}^2 + ... + a_{p1}^2}}. \qquad (5.34)$$

Як значення *k*-го фактора беруть значення елементів власного вектора, що відповідає *k*-му за величиною власному значенню характеристичного полінома матриці R_h, нормовані таким чином, як і в попередньому випадку. Процедуру закінчують, якщо сума власних чисел стає рівною сліду матриці R_h.

Значущість отриманої факторної моделі можна перевірити за критерієм Бартлетта. У цьому випадку перевіряють нульову гіпотезу, що *m* загальних факторів достатньо для пояснення вибіркових коефіцієнтів кореляції, що спостерігаються. Розрахункове значення критерію визначають за формулою:

$$\chi^2 = \left[n - \frac{1}{6}(2p+5) - \frac{2}{3}m \right] \ln \frac{\left|AA^T\right|}{|R|}. \qquad (5.35)$$

Нульову гіпотезу відхиляють, якщо отримана величина перевищує критичне значення для кількості степенів вільності $\nu = 0,5\left[\left(p-m\right)^2 - p - m \right]$.

5.3. Інші методи факторного аналізу

У **методі максимуму правдоподібності**, який запропоновано Д. Лоулі, оцінювання загальностей до безпосереднього застосування алгоритму факторного аналізу не здійснюють. Їх визначають за результатами обчислень з 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 однакові або лінійно залежні вектори. Це призводить до виродження матриці характерностей. У такому випадку можна спробувати зняти проблему шляхом виключення з розгляду лінійно залежних параметрів або застосування методу головних факторів.

У попередніх методах максимізується квадратичний критерій. На відміну від них, у **центроїдному методі**, розробленому американським

психологом Луїсом Леоном Терстоуном у 1930 р., максимізують модульний критерій. З погляду змістової інтерпретації ці критерії є еквівалентними. Цікаво зазначити, що Л. Терстоун спочатку вивчав електроніку у Корнельському університеті. Потім працював асистентом видатного американського інженера й винахідника Томаса Едісона. Він винайшов метод озвучування кінофільмів, удосконалив конструкції кінокамери й кінопректора.

Першу координатну вісь у центроїдному методі обирають так, щоб вона проходила через центр ваги сукупності точок (нульову точку вважають відомою). Другу вісь проводять перпендикулярно першій. На головній діагоналі матриці R ставлять найбільші коефіцієнти кореляції кожного рядка (кожного стовпчика). За новою матрицею значення загальностей (точніше їх оцінки знизу) обчислюють за формулою:

$$h_j^2 = \frac{\left(\sum_{k=1}^{p} r_{jk} \right)^2}{\sum_{k=1}^{p} \sum_{l=1}^{p} r_{kl}}. \qquad (5.36)$$

Перевагою центроїдного 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 (одна з таких змінних; спочатку змінна, що має найбільшу кількість від'ємних значень, потім наступна за кількістю тощо; змінна з номером стовпця, що має максимальну за модулем від'ємну суму значень, а потім інші змінні з від'ємними сумами за стовпцем).

За Л. Терстоуном, максимальна кількість факторів, які можуть бути однозначно визначені за наявності n змінних:

$$n = \frac{2p + 1 - \sqrt{8p + 1}}{2}. \qquad (5.37)$$

Її можна оцінити також із співвідношення:

$$p + n < \left(p - n \right)^2. \qquad (5.38)$$

У **методі контрастних груп** необхідно задати початкове наближення матриці факторного перетворення. При цьому вважають, що вона може містити зайві ознаки, які потрібно виключити при подальших розрахунках. Передбачається, що у просторі ознак, що включені до моделі, розподіл об'єктів вписується до певного еліпсоїда розсіювання, витягнутого вздовж напряму тенденції, що діагностується. Припускають також, що вплив зайвих факторів є мінімальним для об'єктів, що знаходяться поблизу полюсів головної діагоналі еліпсоїда розсіювання. Це дає змогу виокремити "контрастні групи" об-

сягом від 1/4 до 1/3 загального обсягу вибірки й перевірити на них вплив кожної ознаки на досліджувану характеристику. Наприклад, при конструюванні тестів використовують [29] **φ-коефіцієнт Пірсона**

$$\varphi_{\tilde{n}} = \sqrt{\chi^2_{\tilde{n}} / N}\,, \qquad (5.39)$$

де χ^2_c – критичне значення розподілу хі-квадрат,

N – обсяг вибірки. Якщо для певної ознаки $\varphi < \varphi_c$, її виключають з моделі. Після перевірки всіх ознак здійснюють корегування вагових коефіцієнтів і знов перевіряють значущість ознак, що залишилися. Збіжність процедури залежить від співвідношення істотних ознак і "шуму" у вихідній моделі.

Сучасні апроксимуючи методи виходять з припущення, що є певне початкове наближення, яке необхідно покращити. Крім розглянутих вище методів головних факторів, найбільшої правдоподібності й контрастних груп до них належать:

– **груповий метод** Л. Гуттмана й П. Хорста, що базується на попередньому відборі груп елементарних ознак;

– **метод мінімальних залишків** Г. Хартмана;

– **метод α-факторного аналізу**, запропонований Г. Кайзером й І. Кеффрі в 1965 р.;

– **метод канонічного факторного аналізу** К. Рао;

– **методи, що оптимізують**.

Всі ці методи дають змогу послідовно покращувати знайдені розв'язки на основі використання статистичних прийомів оцінювання випадкової величини або статистичних критеріїв й передбачають великий обсяг трудомістких обчислень.

5.4. Приклади проведення факторного аналізу

Розглянемо такий приклад. Сформуємо масив, що містить по 100 значень шести змінних $x_1, \ldots, x_6$. Перші три змінних сформуємо за допомогою генератора випадкових чисел електронних таблиць MS Excel як рівномірно розподілені на відрізку [0; 10] випадкові послідовності. Інші три змінних сформуємо, використовуючи формули:

$$x_4 = x_1 + 2x_2 + \varepsilon_4;\ x_5 = x_1 - 3x_2 + x_3 + \varepsilon_5;\ x_6 = 2x_2 - x_3 + \varepsilon_6,$$

де ε_i – елементи нормально розподіленої випадкової послідовності з математичним сподіванням 0 і стандартним відхиленням 5, сформовані за допомогою генератора випадкових чисел MS Excel.

Для здійснення факторного аналізу в пакеті SPSS уводимо змінні $x_1, \ldots, x_6$ до робочого аркуша. Потім, обираючи у меню Analyze/Data Reduction/ Factor, відчиняємо діалогове вікно факторного аналізу. У ньому вказуємо змінні $x_1, \ldots, x_6$, за якими необхідно здійснити аналіз. У додатково-

му вікні “Descriptives” вказуємо, які параметри описової статистики даних необхідно показати в результатах. Зокрема, тут доцільно навести параметри описової статистики та кореляційну матрицю вихідних ознак. У додатковому вікні “Exstruction” необхідно вказати метод факторного аналізу (Principle Components), а також його параметри. Виберемо аналіз кореляційної матриці, добування шести нових факторів, а також виведення матриці факторного перетворення. Можна не вказувати кількість нових факторів безпосередньо, а задати мінімальне значення відповідних власних чисел кореляційної матриці. У цьому випадку будуть визначатися лише найбільш суттєві фактори, що відповідають критерію Кайзера. У додатковому вікні “Rotation” вказуємо необхідність обертання факторної матриці, метод обертання (Varimax) та робимо позначку про необхідність виведення отриманої після обертання матриці факторного перетворення. У додатковому вікні “Scores” робимо позначки, якщо необхідно розрахувати значення отриманих факторів і зберегти їх як нові змінні. У додатковому вікні “Options” вказуємо спосіб обробки пропущених значень та спосіб відображення коефіцієнтів. Деякі результати факторного аналізу для прикладу, що розглядається, наведено на рис. 17–19.

Descriptive Statistics

	Mean	Std. Deviation	Analysis N
VAR00001	5,3433	2,96073	100
VAR00002	4,8615	2,96218	100
VAR0003	5,3713	3,03396	100
VAR0004	14,8861	7,70350	100
VAR0005	-3,3812	9,47603	100
VAR0006	3,6285	6,44891	100

Рис. 17. Описова статистика вихідних даних

Correlation Matrix[a]

		X1	X2	X3	X4	X5	X6
Correlation	X1	1,000	,175	,013	,503	,115	,109
	X2	,175	1,000	,117	,854	-,824	,787
	X3	,013	,117	1,000	,127	,290	-,362
	X4	,503	,854	,127	1,000	-,591	,638
	X5	,115	-,824	,290	-,591	1,000	-,821
	X6	,109	,787	-,362	,638	-,821	1,000
Sig. (1-tailed)	X1		,040	,451	,000	,128	,140
	X2	,040		,124	,000	,000	,000
	X3	,451	,124		,105	,002	,000
	X4	,000	,000	,105		,000	,000
	X5	,128	,000	,002	,000		,000
	X6	,140	,000	,000	,000	,000	

a. Determinant = 2,201E-03

Рис. 18. Кореляційна матриця вихідних даних

Total Variance Explained

Component	Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3,328	55,467	55,467	3,063	51,053	51,053
2	1,388	23,125	78,592	1,249	20,819	71,872
3	,995	16,584	95,176	1,187	19,783	91,654
4	,158	2,629	97,806	,246	4,105	95,759
5	,102	1,698	99,503	,212	3,541	99,300
6	2,981E-02	,497	100,000	4,200E-02	,700	100,000

Extraction Method: Principal Component Analysis.

Рис. 19. Частка дисперсії, що пояснюється окремими факторами

На рис. 17 наведено описову статистику вихідних ознак, а на рис. 18 – їх кореляційну матрицю. Порівнюючи значення коефіцієнтів кореляції з їх стандартними відхиленнями, можна зробити попередній висновок, які з цих коефіцієнтів є значущими. Зокрема коефіцієнт кореляції між x_1 та x_3, який дорівнює 0,013 при стандартному відхиленні 0,451 слід вважати незначущим, а коефіцієнт 0,854 між змінними x_2 та x_4, який має стандартне відхилення 0,000 – значущим. З кореляційної матриці видно наявність істотної кореляції змінної x_2 зі змінними x_4, x_5 та x_6, між змінними x_5 та x_6, а також помітну кореляцію в деяких інших випадках. Це є підставою для спроби зменшити кількість ознак.

На рис. 19 наведено дані про частку загальної дисперсії, що пояснюється визначеними факторами до та після обертання матриці факторного перетворення. Зокрема, видно, що перші два фактора пояснюють більше, ніж 70% загальної дисперсії, а перші три – понад 90%. В останньому випадку вибір кількості факторів буде відповідати критерію Кайзера.

На рис. 20 показано залежність значень власних чисел від їх порядкового номеру.

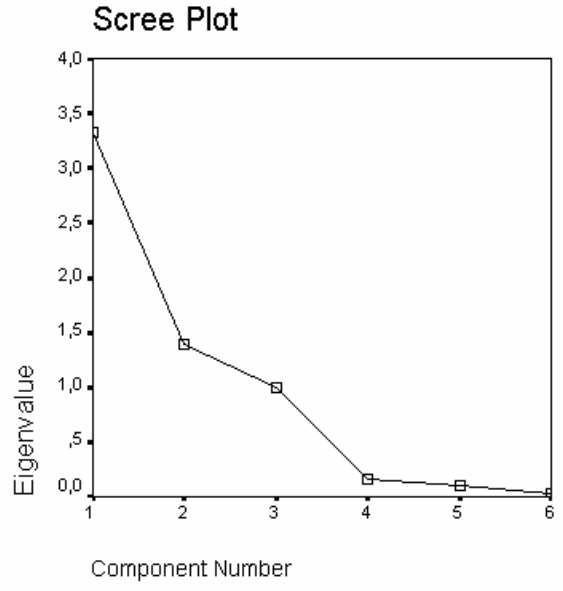


Рис. 20. Залежність значень власних чисел від їх порядкового номеру

Цей графік дає змогу встановити кількість істотних факторів, згідно з критерієм кам'янистого осипу. В нашому випадку вона дорівнює чотирьом, тобто враховує на один фактор більше, ніж при використанні критерію Кайзера.

На рис. 21 показано матрицю навантажень факторів після обертання. Значення у комірках є коефіцієнтами кореляції між відповідними ознаками й факторами.

Rotated Component Matrix

	Component					
	1	2	3	4	5	6
X2	,958	,139	,168	,122	4,174E-02	,137
X5	-,948	,139	,233	9,883E-03	6,665E-02	,151
X6	,838	7,496E-02	-,297	3,686E-02	,451	9,344E-03
X4	,736	,457	,150	,475	3,273E-02	8,332E-03
X1	2,930E-02	,998	4,402E-04	5,281E-02	1,732E-02	1,056E-02
X3	-5,94E-02	1,264E-02	,997	3,375E-02	-4,09E-02	1,210E-02

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 5 iterations.

Рис. 21. Матриця навантажень факторів після обертання

Зокрема бачимо, що шостій фактор практично не корелює з вихідними ознаками, тобто він зумовлений переважно випадковими чинниками. П'ятий має помітну кореляцію лише із шостою ознакою. Це пояснює їх слабкий вплив на загальну дисперсію.

Здійснення факторного аналізу в електронних таблицях MS Excel є незручною, але відповідні завдання можна розв'язувати за допомогою спеціалізованих математичних пакетів.

Розглянемо, як приклад, реалізацію методу головних компонент в пакеті MathCad.

Спочатку задаємо граничне значення критерію інформативності:

$$I := 0,98.$$

Потім уводимо матрицю вихідних даних. Це можна зробити безпосередньо, або використовуючи команди зчитування даних з файлу.

$$X := \begin{pmatrix} 0 & 0.2 & 12 & 35 & 78 & 14 & 21 & 17 & 38 & 40 \\ 1 & 2.1 & 38 & 101 & 26 & 45 & 11 & 93 & 27 & 26 \\ 2 & 3.9 & 8 & 23 & 18 & 62 & 77 & 41 & 89 & 91 \\ 3 & 6.3 & 51 & 99 & 78 & 47 & 32 & 59 & 56 & 55 \\ 4 & 7.5 & 34 & 102 & 84 & 67 & 38 & 92 & 31 & 29 \\ 5 & 10.4 & 62 & 120 & 48 & 90 & 32 & 46 & 58 & 61 \\ 6 & 12.1 & 24 & 71 & 37 & 85 & 64 & 36 & 74 & 76 \\ 7 & 13.7 & 45 & 132 & 37 & 89 & 43 & 57 & 42 & 44 \\ 8 & 15.8 & 13 & 40 & 68 & 99 & 68 & 56 & 21 & 19 \\ 9 & 18.3 & 28 & 83 & 41 & 67 & 39 & 52 & 68 & 71 \end{pmatrix}$$

Далі розраховуємо кореляційну матрицю вихідних ознак:

$i := 0..cols(X) - 1 \qquad j := i$

$CO_{i,j} := corr\left(X^{(i)}, X^{(j)}\right)$

CO =

	0	1	2	3	4	5
0	1	0.999	0.104	0.238	-0.08	0.796
1	0.999	1	0.118	0.24	-0.086	0.785
2	0.104	0.118	1	0.914	0.063	0.212
3	0.238	0.24	0.914	1	$-6.91 \cdot 10^{-3}$	0.281
4	-0.08	-0.086	0.063	$-6.91 \cdot 10^{-3}$	1	-0.242
5	0.796	0.785	0.212	0.281	-0.242	1
6	0.418	0.402	-0.513	-0.504	-0.26	0.576
7	0.016	$-4.626 \cdot 10^{-3}$	0.365	0.517	0.026	0.127
8	0.061	0.079	-0.093	-0.21	-0.511	0.066
9	0.07	0.089	-0.084	-0.197	-0.523	...

Потім розраховуємо власні числа цієї матриці й виводимо їх у порядку спадання:

$V := reverse(sort(eigenvals(CO)))$

V =

	0
0	3.479
1	3.045
2	1.745
3	0.892
4	0.44
5	0.326
6	0.054
7	0.019
8	$2.3 \cdot 10^{-4}$
9	0

На наступному етапі визначаємо кількість компонент, що забезпечують досягнення заданої величини критерію інформативності й виводимо транспоновану матрицю відповідних власних векторів (матрицю факторного перетворення):

```
end(V,I) := | S ← Σ V
            | s ← 0
            | i ← 0
            | while s/S < I
            |     s ← s + V_i
            |     i ← i + 1
            | i
```

$z := 0..end(V,I)$

$L^{(z)} := eigenvec\left(CO, V_z\right)$

$L := L^T$

L =

	0	1	2	3	4	5	6	7	8	9
0	0.28	0.283	-0.196	-0.202	-0.27	0.278	0.452	-0.277	0.407	0.407
1	0.439	0.435	0.347	0.421	$8.691 \cdot 10^{-3}$	0.436	$-7.563 \cdot 10^{-3}$	0.278	-0.163	-0.161
2	0.19	0.173	-0.472	-0.399	0.365	0.109	0.262	-0.031	-0.406	-0.415
3	0.159	0.184	0.244	0.078	0.639	-0.148	-0.196	-0.613	0.11	0.134
4	-0.25	-0.277	0.217	0.017	0.548	0.24	0.512	0.359	0.212	0.141
5	0.285	0.287	-0.256	-0.055	0.223	-0.489	-0.178	0.56	0.277	0.239
6	0.011	-0.11	-0.479	0.776	0.034	-0.194	0.305	-0.148	-0.052	$-8.605 \cdot 10^{-3}$

Після цього виводимо матрицю значень сформованих факторів:

$Z := L \cdot X$

Z =

	0	1	2	3	4	5	6	7	8	9
0	7.283	15.026	25.38	65.288	51.412	83.242	46.63	43.502	55.353	57.574
1	3.757	7.807	79.11	178.535	98.635	104.036	62.578	105.967	105.212	107.603
2	-5.59	-11.343	-8.117	-12.088	-20.482	-51.795	-54.215	-21.626	-53.473	-54.183
3	-0.655	-1.457	0.289	5.653	57.339	7.52	20.075	57.598	19.578	16.965
4	12.148	23.967	57.724	158.529	88.444	162.474	105.537	99.403	101.153	102.711
5	5.273	10.164	17.929	83.826	58.754	47.473	21.29	69.204	3.603	2.442
6	0.725	1.71	20.562	34.581	45.368	-6.177	-8.654	9.498	2.769	0.909

Контрольні питання

1. Які основні завдання вирішують методами факторного аналізу?
2. Як формально записують задачу факторного аналізу?
3. Якими є основні етапи факторного аналізу?
4. Що є об’єктом дослідження в більшості варіантів факторного аналізу?
5. Які техніки застосовують у факторному аналізі? У чому полягають їх основні особливості?
6. Якими є основні припущення основних методів факторного аналізу?
7. Які властивості має задовольняти кореляційна матриця?
8. Якою є основна схема методу головних компонент?
9. Які завдання вирішують за допомогою методу головних компонент?
10. Як визначити кількість головних компонент?
11. Як записують основну модель методу головних компонент?
12. Що використовують як міру інформативності факторів у методі головних компонент?
13. Що називають матрицею факторного відображення?
14. Яку умову використовують для визначення матриці факторного відображення?
15. Що називають першою головною компонентою системи показників?
16. Що називають *k*-ю головною компонентою системи показників?
17. За допомогою якого перетворення здійснюють центрування даних?

18. За допомогою якого перетворення здійснюють нормування даних?

19. Як розраховують власні числа кореляційної матриці?

20. Як розраховують власні вектори кореляційної матриці?

21. Які завдання вирішують за допомогою методу головних факторів?

22. Якими є основні відмінності базових припущень методу головних факторів від методу головних компонент?

23. Як записують основну модель методу головних факторів?

24. З чого складається повна факторна матриця у методі головних факторів? Який зміст мають її складові?

25. Що називають загальністю ознаки?

26. Що називають характерністю ознаки?

27. Що називають специфічністю ознаки?

28. Що називають повнотою факторизації?

29. У чому полягають основні особливості методу максимуму правдоподібності? В яких випадках доцільно використовувати цей метод?

30. У чому полягають основні особливості центроїдного методу? В яких випадках доцільно використовувати цей метод?

6. ЗАВДАННЯ ТА МЕТОДИ КЛАСИФІКАЦІЇ ДАНИХ

У загальному випадку **класифікацією** (**розпізнаванням образів**) називають поділ досліджуваної сукупності об'єктів на однорідні в певному розумінні групи (класи) або зарахування кожного із заданої множини об'єктів до деякого із заздалегідь відомих класів. При цьому вирізняють три групи завдань: дискримінацію, кластеризацію й групування. Останні дві групи є близькими за метою (поділ даних на класи або групи близьких у певному розумінні об'єктів), а також за алгоритмами. Але принципова різниця між ними полягає у тому, що у першому випадку межі класів є природними, а у другому – умовними й їх можна встановлювати суб'єктивно.

При побудові методів розв'язування цієї задачі зазвичай прагнуть мінімізувати ймовірність неправильної класифікації. Для цього можна побудувати функцію втрат $c(j|i)$, що характеризує втрати від помилкового зарахування об'єкта i-го класу до j-го класу [4]. 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:

$$C_n = \sum_{i=1}^{k} \sum_{j=1}^{k} c(j|i) m(j|i), \tag{6.1}$$

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:

$$C = \lim_{n\to\infty} (C_n / n) = \sum_{i=1}^{k} \pi_i \sum_{j=1}^{k} c(j|i) P(j|i) = \sum_{i=1}^{k} \pi_i C^{(i)}, \tag{6.2}$$

де π_i – апріорна ймовірність (питома вага) i-го класу, $P(j|i)$ – імовірність помилкового зарахування об'єкта i-го класу до j-го класу, величина $C(i)$ визначає середні втрати від неправильної класифікації об'єктів i-го класу.

У багатьох випадках втрати є однаковими для будь-якої пари i та j, тобто:

$$c(j|i) = c_0 = const \ \ (\mathrm{i} \neq \mathrm{j}). \tag{6.3}$$

Тоді мінімізація функції втрат еквівалентна максимізації ймовірності правильної класифікації, яка дорівнює $\sum_{i=1}^{k} \pi_i P(i|i)$. З огляду на це, при по-

будові процедур класифікації часто розв'язують задачу мінімізації ймовірності неправильної класифікації:

$$\frac{C}{c_0} = 1 - \sum_{i=1}^{k} \pi_i P(i \mid i). \qquad (6.4)$$

На результат класифікації істотно впливають типи класів. Найчастіше виокремлюють такі типи [39; 45].

1. **Клас типу ядра**, або **згущення**. У цьому випадку всі відстані між об'єктами всередині класу є меншими, ніж їх відстані до будь-якого об'єкта, що не входить до цього класу.

2. **Кластер**, або **згущення у середньому**. Середня відстань між об'єктами всередині класу є меншою, ніж їх середня відстань до всіх інших об'єктів.

3. Для **класу типу стрічки** існує таке $\varepsilon > 0$, що в цьому класі є хоча б один об'єкт x_i, для якого відстань до будь-якого іншого об'єкта цього класу x_j $d_{ij} < \varepsilon$, а відстань до будь-якого об'єкта x_k, що не належить до цього класу, $d_{ij} < \varepsilon$.

4. Характеристичною властивістю **класу із центром** є існування певних граничного значення R і точки x^* у просторі ознак таких, що у багатовимірному шарі радіуса R з центром в точці x^* містяться всі елементи цього класу й немає елементів, які не належать до нього.

Слід зазначити, що один й той самий клас може задовольняти визначення декількох типів. Тому вказана класифікація важлива не стільки з погляду зарахування конкретного класу до певного типу, а з погляду вибору методів розділення декількох класів. Класи, що перетинаються, можуть не задовольняти будь-які з наведених вимог. Але в окремих випадках їх також можна розділяти за допомогою формальних алгоритмів класифікації.

6.1. Параметричні методи класифікації без навчання

У методах класифікації 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 класифікації 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 ймовірностей $f_j\left(X, \Theta_j\right)$ $j = 1, ..., k$ з невідомими значеннями векторних параметрів Θ_i.

Функцію $f(X)$ називають дискретною або неперервною **сумішшю ймовірнісних розподілів**, якщо її можна записати у вигляді, відповідно:

$$f\left(X\right)=\sum_{j=1}^{k}\pi_j f_j\left(X,\Theta\left(j\right)\right) \tag{6.5}$$

або

$$f\left(X\right)=\int f_\omega\left(X,\Theta\left(\omega\right)\right)\pi\left(\omega\right)d\omega\,. \tag{6.6}$$

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 $f_j\left(X,\Theta_j\right)$ для кожного із компонентів за наявною вибіркою спостережень $X_1, X_2, ..., X_n$.

Основною ідеєю більшості методів розв'язування 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 відомі кількість класів, їх апріорні ймовірності та параметричний вигляд функцій щільності ймовірності $f_j\left(X,\Theta_j\right)$, але невідомі значення параметрів Θ_j. Якщо при цьому є навчальні вибірки, 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.

Для розв'язання задачі розщеплення суміші розподілів часто використовують **EM (Expectation – Maximization) алгоритм**, вперше запропонований в 1977 р. американськими статистиками Артуром Демпстером, Неном Лайрдом і Дональдом Рубіном. Цей алгоритм дає змогу визначати методом найбільшої правдоподібності параметри статистичних моделей, що містять певні приховані змінні. Він передбачає здійснення двох кроків на кожній ітерації. Перший крок (Expectation) полягає в обчисленні значення функції правдоподібності за умови, що задані деякі значення прихованих змінних. На другому кроці (Maximization) обчислюють значення параметрів, що максимізують функцію правдоподібності. Обчислення виконують до виконання заданих умов збіжності. Недоліками алгоритму є залежність результату від вибору початкового наближення (якщо функція

правдоподібності не є унімодальною). Крім того, цей алгоритм не дає змоги визначити кількість компонент суміші. Для усунення цих недоліків пізніше було запропоновано різноманітні модифікації EM алгоритму: медіанні, стохастичний (SEM), класифікаційний (CEM), узагальнений (GEM), з додаванням компонент тощо.

На рис. 6.1 наведено приклад гістограми розподілу результатів єдиного державного екзамену для випускників шкіл Російської Федерації з російської мови у 2008 р. Вихідні дані взяті із сайта http://www.ege.ru. Наведений розподіл за результатами нашого аналізу може бути подано як суміш, що відповідає трьом класам з нормально розподіленими балами:

$$f(n) = 0,061N\left(11,9;3,6\right) + 0,278N\left(24,2;6,9\right) + 0,661N\left(39,3;10,0\right),$$

де $N\left(m,\sigma\right)$ – функція щільності нормального розподілу з математичним сподіванням m і стандартним відхиленням σ.

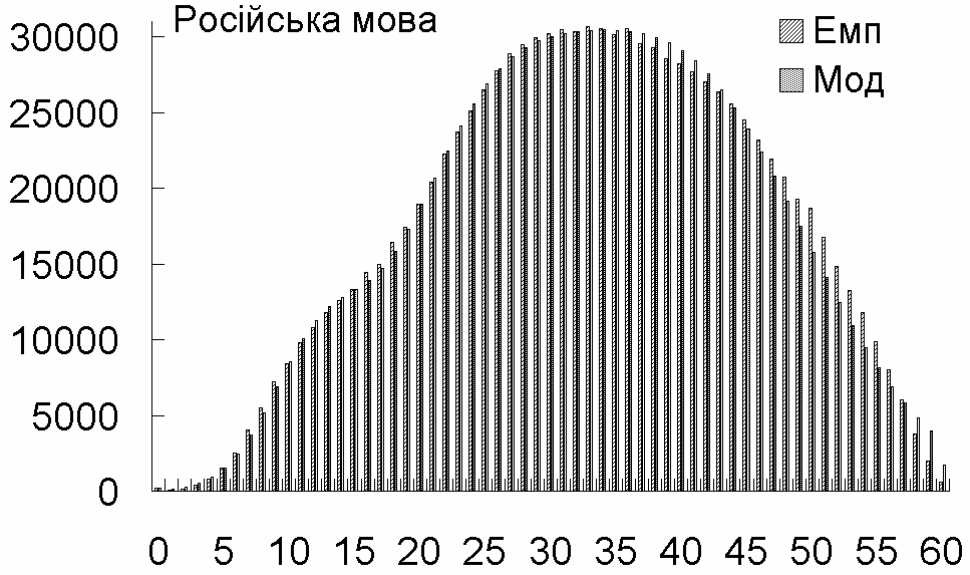


Рис. 6.1. Гістограми розподілу результатів ЄДЕ з російської мови у 2008 р.

6.2. Кластерний аналіз

У непараметричному випадку ми не маємо інформації про загальний вигляд функцій $f_j\left(X,\Theta_j\right)$. Ми можемо мати 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.

Бажано, щоб компоненти вектора X відповідали одному й тому самому типу даних. Для цього зазвичай використовують перехід від кількісних ознак до порядкових та від порядкових до номінальних. Але слід ураховувати, що при цьому втрачається частина корисної інформації.

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 (рис. 6.2). Наочне уявлення про зміст класифікації дає **діаграма Герцшпрунга – Ресселла** (рис. 6.3), яка є основою для однієї з найпоширеніших класифікацій зірок за поєднанням їх світності й кольору (температури або спектрального класу).

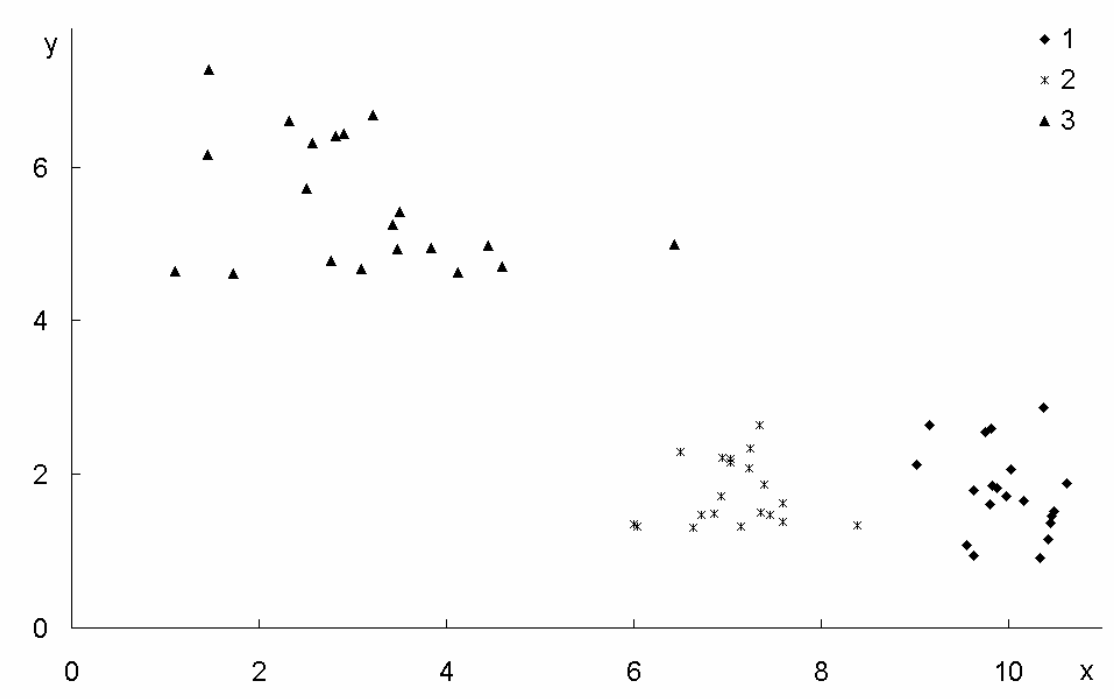


Рис. 6.2. Геометричне зображення сукупності об'єктів,
що характеризуються двома ознаками й утворюють три кластери

Залежно від мети дослідження задачу класифікації можна сформулювати як розбиття аналізованих об'єктів на певну кількість груп, усередині 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, що відповідають наявним спостереженням.

Перша задача завжди має розв'язок, а друга може не мати розв'язку. Це відповідає відсутності природного розшарування досліджуваних об'єктів (наприклад, вони утворюють один кластер або відповідні точки рівномірно заповнюють весь простір ознак).

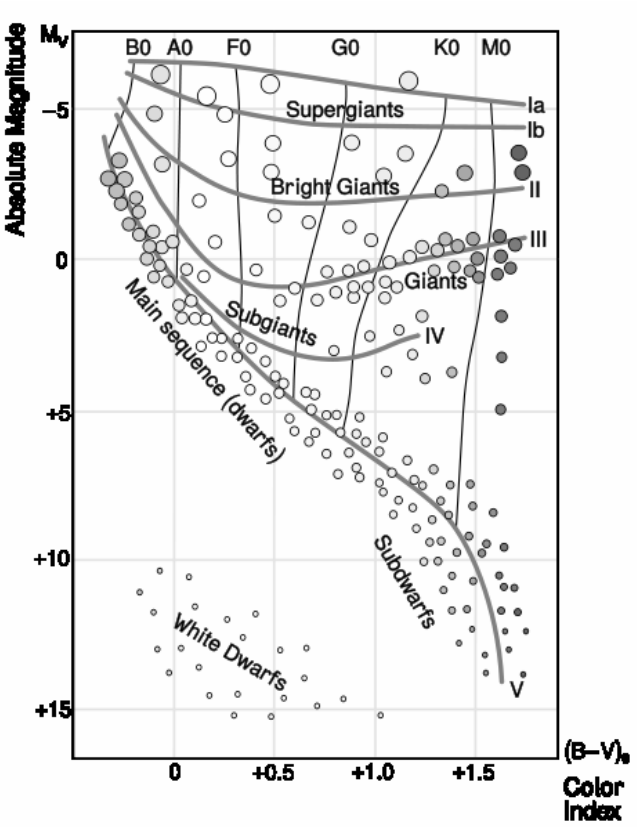


Рис. 6.3. Діаграма Е. Герцшпрунга – Г. Ресселла
(http://uk.wikipedia.org/wiki/ Файл: H-R_diagram.svg)

Класичними непараметричними методами класифікації без навчання є методи **кластерного аналізу** (**таксономії**). За їх допомогою вирішують проблему такого розбиття (класифікації, кластеризації) множини об’єктів, за якого всі об’єкти, що належать до одного класу, були б більш подібними один до одного, ніж до об’єктів інших класів. З формальної точки зору, основне завдання методів кластерного аналізу можна сформулювати, як визначення класів еквівалентності й рознесення за ними досліджуваних об’єктів. Під **класом**, як правило, розуміють генеральну сукупність, що описується одномодальною функцією щільності ймовірності $f(\boldsymbol{X})$ або, у випадку дискретних ознак, – одномодальним полігоном імовірностей. Номери класів не мають змістового навантаження 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 більш подібними один до одного, чим меншою є відстань між ними);

– **міри подібності типу “зв’язок”** (у цьому випадку об’єкти вважають тим більш подібними, чим сильнішим є зв’язок між ними);

– **інформаційна статистика**.

Як міру відстані (метрику) можна використовувати будь-яку функцію $\rho\left(X_i, X_j\right)$, що визначена на множині $\left\{X_1, X_2, ..., X_n\right\}$ і задовольняє такі вимоги:

1) $\rho\left(X_i, X_j\right) \geq 0$ для всіх i, j;

2) $\rho\left(X_i, X_j\right) = 0$ тоді й тільки тоді, коли $X_i = X_j$;

3) $\rho\left(X_i, X_j\right) = \rho\left(X_j, X_i\right)$;

4) $\rho\left(X_i, X_j\right) \leq \rho\left(X_i, X_k\right) + \rho\left(X_k, X_j\right)$.

Вибір міри відстані істотно впливає на результат класифікації. Тому для отримання надійних результатів необхідно враховувати мету дослідження, змістову й статистичну природу вектора спостережень і наявні відомості про характер розподілу досліджуваних ознак. Крім того, після закінчення розрахунків слід перевіряти адекватність отриманої класифікаційної моделі.

Найчастіше використовують евклідову та манхеттенську відстані, супремум-норму, а також відстань Махаланобиса. Вони відображають усе різноманіття підходів до цієї проблеми. Евклідову метрику традиційно застосовують як міру відстані. Манхеттенська відстань є найбільш відомою з класу метрик Мінковського. Відстань Махаланобиса, що не є метрикою, 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, Уорда, *k*-середніх Мак-Куїна.

Як міри зв'язку для кількісних ознак можна обирати коефіцієнт кореляції Пірсона, кореляційне відношення і дисперсію-коваріацію. Застосування коефіцієнта кореляції Пірсона є обґрунтованим лише за умови, що зв'язок між ознаками є лінійним, але в окремих випадках його можна використовувати для нелінійного зв'язку після придатного перетворення вихідних ознак.

Для порядкових ознак призначені коефіцієнти рангової кореляції Спірмена й Кендалла. Їх можна перетворити до мір подібності типу "відстань" за допомогою формул:

$$d_{ij} = 1 - \rho s;\ d_{ij} = 1 - \tau. \qquad (6.7)$$

У цьому випадку їх називають, відповідно, **відстанями Спірмена** й **Кендалла**.

Для дихотомічних ознак та 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 протилежний. Результати класифікації різними методами, як правило, принципово не відрізняються.

Вибір метрики, навпаки, може істотно впливати на результати аналізу. Тому для кожної конкретної задачі його необхідно здійснювати окремо. При цьому треба враховувати головні цілі дослідження, фізичну й 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 Махаланобиса, окремими випадками якої є евклідова, зважена евклідова та хеммінгова відстані.

Дивергенція між двома сукупностями *i* та *j* (**повна середня інформаційна міра різниці двох класів**, **дивергенція Кульбака – Лейблера**) запропонована американськими криптоаналітиками Соломоном Кульбаком та Річардом Лейблером у 1951 р. Вона може бути розрахована за формулою:

$$J_{ij} = \frac{1}{2} tr\left[\left(C_i - C_j\right)\left(C_j^{-1} - C_i^{-1}\right)\right] + \\ + \frac{1}{2} tr\left[\left(C_i^{-1} + C_j^{-1}\right)\left(m_i - m_j\right)\left(m_i - m_j\right)'\right], \tag{6.8}$$

де C_i, C_j – дисперсійно-коваріаційні матриці сукупностей *i* та j;

m_i, m_j – вектори середніх сукупностей *i* та *j*.

Міра Махаланобиса (**відстань Махаланобиса**, **узагальнена евклідова відстань**) – це відстань від точки спостереження до центра ваги в багатовимірному просторі ознак:

$$d_{ij} = \sqrt{\left(X_i - X_j\right)^T \Delta^T \Sigma^{-1} \Delta \left(X_i - X_j\right)}, \tag{6.9}$$

де Δ – векторна симетрична невід'ємно визначена матриця вагових коефіцієнтів, яку найчастіше обирають діагональною;

Σ – коваріаційна матриця генеральної сукупності, до якої належать спостереження. Введена відомим індійським статистиком й антропометристом П.Ч. Махаланобисом у 1938 р. Іноді замість міри Махаланобиса використовують її квадрат.

Розглянемо деякі з інших мір відстані детальніше.

Евклідова відстань (**евклідова метрика**) є відомою із загальноматематичних дисциплін і визначається за формулою:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^2} . \qquad (6.10)$$

Вона збігається з відстанню Махаланобиса у випадку, коли незалежні змінні є некорельованими. Поряд з евклідовою відстанню як міру близькості часто використовують її квадрат.

Евклідову відстань 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.

Недоліком евклідової метрики є те, що у випадках, коли ознаки виміряні у різних одиницях, зміна масштабу одиниць вимірювання може призвести до істотної зміни результатів класифікації. Для запобігання цьому використовують різні методи нормування даних. Найпоширенішими з них є такі [20; 39]:

$$z_1 = \frac{x - \bar{x}}{\sigma};\ z_2 = \frac{x - x_{\min}}{x_{\max} - x_{\min}};\ z_3 = \frac{x - MED}{MAD};\ z_4 = \frac{x}{\bar{x}};\ z_5 = \frac{x}{x_{\max}}. \qquad (6.11)$$

Але слід зазначити, що нормування також впливає на результати класифікації. Зокрема, у випадках, коли кластери істотно розділяються за деякими ознаками й слабко за іншими, нормалізація може призвести до зменшення дискримінуючих можливостей першої групи ознак через збільшення шумового ефекту інших [28].

Якщо ознаки вимірюють у якісно різних одиницях, то застосування евкідової відстані взагалі може виявитися безглуздим.

Зважену евклідову відстань розраховують за формулою:

$$d_{ij}^{*} = \sqrt{\sum_{k=1}^{p} \omega_k \left(x_{ik} - x_{jk} \right)^2} , \qquad (6.12)$$

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 призвести до істотних помилок. Зокрема, залежно від певних незначних варіацій змістової та статистичної природи вихідних даних може бути обґрунтованим надання їм значень, пропорційних середньоквадратичній похибці відповідної ознаки або оберненій до цієї похибки ве-

личині. Тому рекомендують обирати вагові коефіцієнти за результатами експертних опитувань або інших незалежних попередніх досліджень.

Метрика Мінковського, запропонована видатним німецьким математиком і фізиком Германом Мінковським у 1908 р., є узагальненням звичайної евклідової відстані:

$$d_{ij} = \sqrt[r]{\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^r} . \qquad (6.13)$$

У випадку $r = 2$ вона збігається з евклідовою метрикою.

У випадку $r = 1$ метрика Мінковського дає **манхеттенську відстань**:

$$d_{ij} = \sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right| . \qquad (6.14)$$

При $r \to \infty$ метрика Мінковського збігається із **супремум-нормою** (**відстанню Чебишева**), яку було введено видатним російським математиком Пафнутієм Львовичем Чебишевим:

$$d_{ij} = \sup \left\{ \left| x_{ik} - x_{jk} \right| \right\}, \quad k = 1, 2, ..., p. \qquad (6.15)$$

Хеммінгову відстань, яку ввів відомий американський математик Річард Хеммінг у 1950 р., використовують як міру відстані об'єктів, що характеризуються дихотомічними ознаками. Її розраховують за формулою:

$$d_{ij} = \sum_{s=1}^{p} \left| x_{is} - x_{js} \right| , \qquad (6.16)$$

тобто вона збігається із кількістю значень відповідних ознак, що не збігаються, у i-го та j-го об'єктів (у разі, коли ознаки можуть набувати значення 0 або 1).

При конструюванні різноманітних процедур класифікації доцільно використовувати міри близькості кластерів один до одного. Найбільш поширеними з них є відстані, що вимірюють за принципами найближчого і далекого сусідів, середнього зв'язку та за центрами ваги. Вибір міри близькості кластерів є найбільш суттєвим для агломеративних ієрархічних методів кластерного аналізу.

Нехай S_i – i-й кластер, n_i – кількість об'єктів у ньому, $\overline{X}(i)$ – центр ваги i-го кластера, тобто середнє арифметичне векторних спостережень, що його утворюють, $\rho_{\ell m}$ – відстань між класами ℓ і m. Тоді **відстань, що вимірюють за принципом найближчого сусіда** (**nearest neibour**):

$$\rho_{\ell m}^{\min} = \min_{X_i \in S_\ell ; X_j \in S_m} d_{ij} ; \qquad (6.17)$$

відстань, що вимірюють за принципом далекого сусіда (furthest neibour):

$$\rho_{\ell m}^{\max} = \max_{X_i \in S_\ell; X_j \in S_m} d_{ij}; \tag{6.18}$$

відстань, що вимірюють за принципом середнього зв'язку (середнє арифметичне всіх можливих попарних відстаней між представниками класів, які розглядають):

$$\rho_{\ell m}^{m} = \frac{1}{n_\ell n_m} \sum_{X_i \in S_\ell} \sum_{X_j \in S_m} d_{ij}; \tag{6.19}$$

відстань, що вимірюють за центрами ваги:

$$\rho_{\ell m} = d\left(\overline{X}_\ell, \overline{X}_m\right). \tag{6.20}$$

Існує також узагальнена (за О.М. Колмогоровим) формула розрахунку відстаней між класами (**відстань Колмогорова**, **узагальнена *K*-відстань**):

$$\rho_{\ell m}^{K} = \left[\frac{1}{n_\ell n_m} \sum_{X_i \in S_\ell} \sum_{X_j \in S_m} d_{ij}^{\tau}\right]^{1/\tau}. \tag{6.21}$$

При $\tau \to -\infty$ вона переходить до формули (6.17), при $\tau \to +\infty$ – до формули (6.18), а при $\tau = 1$ – до формули (6.19).

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:

$$\rho_{\ell w}^{K} = \left[\frac{n_m \left(\rho_{\ell m}^{K}\right)^{\tau} + n_q \left(\rho_{\ell q}^{K}\right)^{\tau}}{n_m + n_q}\right]^{1/\tau}. \tag{6.22}$$

Для перерахунку відстані між класами використовують також загальну **формулу Ланса та Уїльяма**:

$$\rho_{\ell w} = a_m \rho_{\ell m} + a_q \rho_{\ell q} + b\rho_{mq} + c\left|\rho_{\ell m} - \rho_{\ell q}\right|, \tag{6.23}$$

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:

$$a_m = \frac{n_m}{n_m + n_q};\ a_m = \frac{n_q}{n_m + n_q};\ b = c = 0. \tag{6.24}$$

Для розрахунку ступеня близькості класів використовують також розглянуті вище інформаційну відстань Каллбека (у випадку, коли класи можна розглядати як багатовимірні нормальні сукупності) та відстань Махаланобиса (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 чітких методів обрання функціоналів якості.

У випадку, коли кількість класів є заданою, як функціонали якості найчастіше обирають такі.

Сума внутрішньокласових дисперсій:

$$Q_1(S) = \sum_{\ell=1}^{k} \sum_{X_i \in S_\ell} d^2\left(X_i, \overline{X}(\ell)\right). \tag{6.25}$$

Сума попарних внутрішньокласових відстаней між елементами:

$$Q_2(S) = \sum_{\ell=1}^{k} \sum_{X_i, X_j \in S_\ell} d_{ij}^2 \tag{6.26}$$

або

$$Q_2'(S) = \sum_{\ell=1}^{k} \frac{1}{n_\ell} \sum_{X_i, X_j \in S_\ell} d_{ij}^2 . \tag{6.27}$$

У більшості випадків вона приводить до тих самих результатів, що й попередній критерій.

Узагальнену внутрішньокласову дисперсію можна розраховувати як показник середньоарифметичної або середньогеометричної дисперсії, відповідно, за формулами:

$$\begin{aligned} Q_3(S) &= \det\left(\sum_{\ell=1}^{k} n_\ell \Sigma_\ell\right); \\ Q_4(S) &= \prod_{\ell=1}^{k} \left(\det \Sigma_\ell\right)^{n_\ell}, \end{aligned} \tag{6.28}$$

де елементи вибіркової коваріаційної матриці Σ класу S_ℓ розраховують як:

$$\sigma_{qt}(\ell) = \frac{1}{n_\ell} \sum_{X_i \in S_\ell} \left(x_i^{(q)} - \overline{x}^{(q)}(\ell)\right)\left(x_i^{(t)} - \overline{x}^{(t)}(\ell)\right);\ q, t = 1, 2, ..., p\,. \tag{6.29}$$

Такі функціонали доцільно застосовувати, якщо допускають можливість зосередженості розбитих на класи спостережень у просторі меншої розмірності, 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 – незгасаючою 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**:

$$Z_\tau(S) = \left[\frac{1}{n} \sum_{i=1}^{n} \left(\frac{\nu(X_i)}{n} \right)^\tau \right]^{1/\tau}, \qquad (6.30)$$

де $\nu(X_i)$ – 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 ураховувати, що:

$$Z_{-\infty}(S) = \min_{1 \le i \le k} \left(\frac{n_i}{n} \right);$$

$$Z_{-1}(S) = 1/k;$$

$$\log Z_0(S) = \sum_{i=1}^{k} \frac{n_i}{n} \log \frac{n_i}{n};$$

$$Z_1(S) = \frac{1}{n} \sum_{j=1}^{n} \frac{\nu(X_i)}{n} = \frac{1}{n^2} \sum_{i=1}^{k} n_i^2;$$

$$Z_\infty(S) = \max_{1 \le i \le k} \left(\frac{n_i}{n} \right),$$

де k – кількість різних кластерів у розбитті S;

n_i – кількість елементів у i-му кластері.

Величина Z_0 є природною інформаційною мірою концентрації.

За будь-якого τ міра (6.30) має мінімальне значення $1/n$ при розбитті досліджуваної множини на n одноточкових кластерів і максимальне значення 1 при об'єднанні всіх вихідних даних в один кластер.

Як середню міру внутрішьокласового розсіювання можна використовувати величину:

$$I_{\tau}^{(K)}(S)=\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{v(X_i)}\sum_{X_\ell\in S(X_i)}d_{i\ell}^{\tau}\right]^{1/\tau}. \qquad (6.31)$$

Тоді сумарне розсіювання характеризує величина $n\left(I_{\tau}^{(K)}(S)\right)^{\tau}$ й оптимальною слід вважати таку кількість кластерів, за якої величина набуває мінімального можливого значення:

$$\frac{\Delta\left[n\left(I_{\tau}^{(K)}(S)\right)^{\tau}\right]}{\Delta Z_{\tau}(S)}. \qquad (6.32)$$

Залежно від кількості вихідних спостережень виділяють задачі класифікації невеликих за обсягом (до декількох десятків точок) масивів 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 вирізняють 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.

Найпоширенішими методами кластерного аналізу є:

– ієрархічні методи (ближнього зв'язку, середнього зв'язку Кінга, Уорда, далекого зв'язку);

– ітеративні методи групування (метод *k*-середніх Мак-Куїна);

– 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 випадку реалізацію ієрархічного алгоритму продовжують до досягнення кількості класів, яка дорівнює 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 основними недоліками є громіздкість обчислювальної процедури, яка пов'язана з пе-

ерахунком 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 ієрархічних агломеративних методів кластерного аналізу. Процес класифікації в цьому випадку 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 так само до повного вичерпання матриці подібності. Як правило, робота алгоритму закінчується, коли всі спостереження об'єднані в один клас. Для виокремлення 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 розбиттям досліджуваних об'єктів на кластери.

У методі ближнього зв'язку 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. Іноді це призводить до необґрунтованого зарахування об'єктів до одного й того самого кластера (**ланцюжковий ефект**). У процесі кластеризації можна явно простежити утворення таких ланцюжків. Для запобігання цьому ефекту можна задавати обмеження на максимальну відстань між елементами одного кластера.

Сучасний варіант цього методу, зважений алгоритм найближчих сусідів, запропоновали в 2005 р. іспанські математики Роберто Паредес та Енрікє Відал.

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.

Після проведення кластеризації рекомендується візуалізувати результати шляхом побудови дендрограми, яка дає можливість отримати уявлення про загальну конфігурацію об'єктів.

Результати ієрархічних методів кластерного аналізу стають більш наочними, якщо їх подати у вигляді **дендрограми** (**дендограми**). Типовий вигляд дендрограми наведено на рис. 6.3.

Пари об'єктів при побудові дендрограми з'єднують згідно з рівнем зв'язку, який відкладають по вісі 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.

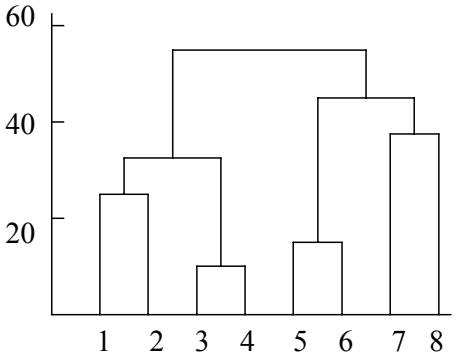


Рис. 6.3. Приклад дендрограми

Метод середнього зв'язку Кінга подібний до методу ближнього зв'язку. Його 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. При використанні методу середнього зв'язку в 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.

***K*-узагальнена ієрархічна процедура** ґрунтується на тому, що перелічені відстані, а також відстань, що вимірюється за принципом далекого сусіда, яка застосовується в методі далекого зв'язку, є окремими випадками узагальненої відстані Колмогорова (6.20), яку й використовують у цьому випадку як міру близькості. Описані вище методи можна розглядати як окремі випадки *K*-узагальненої ієрархічної процедури.

Порогові ієрархічні процедури передбачають задання монотонної послідовності порогів $c_1, c_2, \ldots, c_t$. В агломеративних методах на першому кроці попарно об'єднують елементи, відстані між якими не перевищують c_1. На другому кроці об'єднують елементи або групи елементів, відстані між якими не перевищують c_2 тощо. 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 поділяється на достатньо віддалені одне від одного скупчення точок у досліджуваному просторі ознак.

Метод Уорда, запропонований Дж. Уордом у 1963 р., 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. Як правило, їх застосовують для розв'язання перших двох типів задач кластеризації при великих обсягах досліджуваних сукупностей.

Прикладом послідовної ітеративної процедури є **метод *k*-середніх Мак-Куїна**. Ідея цього методу запропонована в 1956 р. відомим польським математиком Гуго Штейнгаузом, який в 1920–1941 р. працював професором Львівського університету і є одним із засновником львівської математичної школи. Стандартний алгоритм методу розроблено в 1957 р. Стюартом Ллойдом, а назву введено в 1967 р. американським математиком Дж.Б. Мак-Куїном. Ще один поширений алгоритм цього методу запропонований у 1965 р. Г. Боллом та Д. Холлом.

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, розташованим у центрі ваги вихідного кластера й 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 тощо. Другий підхід передбачає багаторазовий повторний вибір вихідних кластерів. При цьому на кожному етапі як вихідні обирають точки, що є найближчими до фінальних кластерів, що найчастіше отримували на попередніх етапах.

Особливістю методу є алгоритмічне гарантування того, що кожний із класифікованих об'єктів буде зарахований лише до одного з кластерів. При застосуванні цього методу 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 трьох), або двовимірних зрізів простору. У багатьох випадках метод *k*-середніх дає змогу отримати розбиття, близьке до найкращого з погляду функціонала якості.

Якщо 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, що дорівнює сумі вагових коефіцієнтів вихідних кластерів. Після закінчення цієї 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. Можна довести, що алгоритм методу Мак-Куїна збігається до локального мінімуму суми внутрішньокласових дисперсій. Глобальний мінімум цього функціоналу може бути досягнутий за допомогою алгоритму Р. Дженсена, який базується на застосуванні динамічного програмування.

Сутність методу **кореляційних плеяд** є такою. Візуально результати класифікації можна подати у вигляді кореляційного циліндра, розсіченого площинами, перпендикулярними його осі. Площини відповідають рівням від нуля до одиниці з кроком 0,1. На цих рівнях об'єднують класифіковані параметри або об'єкти. Метод наближається до методу ближнього зв'язку з фіксованими рівнями об'єднання. Графічно результати зображують у вигляді кіл-зрізів (плеяд) кореляційного циліндра. 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 певного оптимального дерева (дендрита). Цей метод є подібним до методу ближнього зв'язку, але його алгоритм належить до алгоритмів розрізання графів. Якщо мірою подібності обрати коефіцієнт кореляції, ми отримаємо метод найбільшого кореляційного шляху.

У методі вроцлавської таксономії результати розрахунків відображають (рис. 6.4) 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).

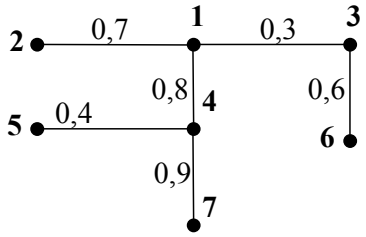


Рис. 6.4. Приклад графа, що відображає зв'язок між параметрами

Аналізуючи отриманий граф, можна зробити висновки про взаємозв'язки тих чи інших параметрів або груп параметрів, зокрема про те, що для графа, наведеного на рис. 6.3, параметри умовно поділяються на три групи: (1, 3, 6), (2) та (4, 5, 7).

Перед застосуванням процедур класифікації рекомендується дослідити наявні ознаки з метою вибору найбільш інформативних з 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.

6.3. Класифікація з навчанням

Методи розпізнавання образів з навчанням (із вчителем) призначені для зарахування некласифікованих об'єктів до заздалегідь описаних класів (кластерів; навчальні 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 класифікації у цьому випадку можна сформулювати так. 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):

$$f(X) = \sum_{j=1}^{k} \pi_j f_j(X). \tag{6.33}$$

Розв'язувальним 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:

$$S_j = \{X : \delta(X) = j\}. \tag{6.34}$$

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'язувальні 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'язувальне правило називають оптимальним (байєсівським), якщо воно забезпечує мінімальні втрати серед усіх можливих процедур класифікації. Оптимальне розв'язувальне правило можна задати так:

$$S_j^{(opt)} = \left\{ X : \sum_{\substack{i=1 \\ i \neq j}}^{k} \pi_i f_i(X) c(j \mid i) = \min_{1 \le \ell \le k} \sum_{\substack{i=1 \\ i \neq j}}^{k} \pi_i f_i(X) c(\ell \mid i) \right\}. \tag{6.35}$$

Це означає, що спостереження X_ν $(\nu = 1, 2, ..., n)$ зараховують до j-го класу у випадку, 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_ν зараховують до j-го класу за умови:

$$\pi_i f_i(X_\nu) = \max_{1 \le \ell \le k} \pi_\ell f_\ell(X_\nu). \tag{6.36}$$

Це розв'язувальне правило можна сформулювати так: спостереження X_ν зараховують до класу j_0, якщо:

$$\frac{f_{j_0}(X_\nu)}{f_j(X_\nu)} \ge \frac{\pi_j}{\pi_{j_0}} \tag{6.37}$$

для всіх $j = 1, 2, \ldots, k$.

Одним з основним методів розпізнавання образів з навчанням є **дискримінантний аналіз**. Він належить до класу лінійних методів, оскільки його модель є лінійною стосовно 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 у випадках, коли навчання здійснюють за допомогою емпіричної класифікації. Якщо встановлено, що окремі об'єкти не належать 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 дають метод k-середніх, який гарантовано будує кластери, що не перетинаються, а також метод ближнього зв'язку.

Під **інформативністю параметрів**, як правило, 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 виокремлення груп, що навчають.

Можливі ситуації, коли кількість параметрів є недостатньою для правильної з погляду дослідника класифікації, або, навпаки, є зайві параметри, що не є обов'язковими для класифікації й призводять до отримання громіздких результатів, які важко інтерпретувати. В окремих випадках об'єкти з вибірок, що навчають, після класифікації можуть бути зараховані не до тих кластерів, куди вони були вміщені на попередньому етапі. Особливо часто

таке відбувається при застосуванні емпіричних класифікацій. У таких ситуаціях необхідно виконати додаткове дослідження стосовно необхідності й достатності тих параметрів, за якими здійснюють класифікацію.

Для практичної реалізації оптимальних розв'язувальних правил (6.35, 6.36) необхідно знати апріорні ймовірності π_j і функції щільності ймовірності $f_j(X)$. 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:

$$\pi_j = n_j / n_{sum}, \tag{6.38}$$

де n_j – обсяг j-ї вибірки;

$n_{sum} = n_1 + n_2 + ... + n_k$ – сумарний обсяг вибірок, що навчають.

При оцінюванні функцій щільності ймовірності застосовують два підходи. У першому (**параметричний дискримінаційний аналіз**) припускають, що всі класи характеризуються функціями щільності ймовірності, які належать до однієї параметричної сім'ї $\{f(X,\Theta)\}$ і розрізняються лише значеннями векторного параметра Θ. У цьому випадку відповідні значення параметра Θ_j оцінюють за спостереженнями, що належать до j-ї вибірки. У другому підході (**непараметричний дискримінантний аналіз**) загальний вигляд функцій $f_j(X)$ є невідомим. Тому 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 з 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:

$$\varphi(X, M, \Sigma) = \frac{1}{2\pi|\Sigma|^{1/2}} e^{-\frac{1}{2}(X-A)^T \Sigma^{-1}(X-A)}, \tag{6.39}$$

де A – матриця, утворена векторами середніх значень;

X – матриця значень ознак. Обидві матриці мають розмір $p \times k$.

Оцінки для векторів середніх значень $\mathbf{a}_j = \left(a_j^{(1)}, ..., a_j^{(p)}\right)^T$ та елементів коваріаційної матриці, отримані методом найбільшої правдоподібності за вибірками, що навчають, мають вигляд:

$$a_j^{(\ell)} = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ji}^{(\ell)}; \tag{6.40}$$

$$\sigma_{\ell q} = \frac{1}{n_{sum} - k} \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(x_{ji}^{(\ell)} - a_j^{(\ell)} \right) \left(x_{ji}^{(q)} - a_j^{(q)} \right); \qquad (6.41)$$

$$\left(\ell = 1, ..., p; \, j = 1, ..., k \right).$$

Якщо функції $f_j(X)$ визначаються формулою (6.39), то розв'язувальне правило (6.37) набуває вигляду:

$$\left[X_\nu - \frac{1}{2} \left(\mathbf{a}_{j_0} + \mathbf{a}_j \right) \right]^T \Sigma^{-1} \left(\mathbf{a}_{j_0} + \mathbf{a}_j \right) \geq \ln \frac{\pi_j}{\pi_{j_0}} \qquad (6.42)$$

для всіх $j = 1, 2, ..., k$.

Для $k = 2$ апріорні ймовірності $\pi_1 = \pi_2 = 0,5$. У цьому випадку спостереження X_v зараховують до першого класу, якщо виконується умова:

$$\left[X_\nu - \frac{1}{2} \left(\mathbf{a}_1 + \mathbf{a}_2 \right) \right]^T \Sigma^{-1} \left(\mathbf{a}_1 + \mathbf{a}_2 \right) \geq 0, \qquad (6.43)$$

і до другого – якщо вона не виконується.

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 (імовірностей) 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 таблиці 6.1 наведено приклад такої матриці для випадку двох класів і трьох ознак. При цьому перша і третя ознаки мають по три розряди, а друга – два розряди.

Таблиця 6.1

Приклад діагностичної матриці для багаторозрядних ознак

D	K_1			K_2		K_3			$P(D)$
	$P(K_{11})$	$P(K_{12})$	$P(K_{13})$	$P(K_{21})$	$P(K_{22})$	$P(K_{31})$	$P(K_{32})$	$P(K_{33})$	
D_1									
D_2									

Розрахунок імовірності зарахування об'єкта до класу D_i здійснюють за формулою:

$$P\left(D_i \mid K^* \right) = \frac{P\left(D_i \right) P\left(K^* \mid D_i \right)}{\sum_{s=1}^{n} P\left(D_s \right) P\left(K^* \mid D_s \right)}, \qquad (6.44)$$

де $K\left(K_1, K_2, ..., K_\nu \right)$ – ряд v багаторозрядних ознак;

K^* – його реалізація;

$P\left(D_i \mid K^*\right)$ – імовірність зарахування об'єкта до класу D_i за умови, що комплекс ознак K набув реалізації K^*; $P\left(K^* \mid D_i\right)$ – імовірність появи комплексу ознак K^* в об'єкта, що належить до класу D_i;

$P(D_i)$ – апріорна ймовірність потрапляння до класу D_i, яка визначається за емпіричними даними; i – номер кластера.

Якщо комплекс ознак містить v ознак, то:

$$P\left(K^* \mid D_i\right) = P\left(K_1^* \mid D_i\right) P\left(K_2^* \mid K_1^* D_i\right) ... P\left(K_\nu^* \mid K_1^* K_2^* ... K_{\nu-1}^* D_i\right). \quad (6.45)$$

У багатьох випадках, навіть за наявності істотних кореляційних зв'язків, можна використовувати формулу Байєса для незалежних ознак. У цьому випадку:

$$P\left(K^* \mid D_i\right) = \prod_{r=1}^{\nu} P\left(K_r^* \mid D_i\right). \quad (6.46)$$

В основі методу **лінійного дискримінантного аналізу Фішера**, запропонованого Р. Фішером у 1936 р., лежить припущення, що класифікацію можна здійснити за допомогою лінійної комбінації **дискримінантних (розрізнюючих) змінних**. Підґрунтям 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:

$$h_k = b_{k0} + \sum_{i=1}^{p} b_{ki} X_i, \quad (6.47)$$

де p – кількість дискримінантних змінних;

b_{ki} – коефіцієнт для i-ї змінної k-го класу:

$$b_{ki} = (n-g) \sum_{j=1}^{p} a_{ij} X_{jk}, \quad (6.48)$$

де n – загальна кількість спостережень за усіма класами;

a_{ij} – елементи матриці, яка є оберненою до матриці розкидів усередині класів і розраховується за формулою:

$$w_{ij} = \sum_{k=1}^{j} \sum_{n=1}^{n_k} \left(X_{ikm} - X_{ik}\right)\left(X_{jkm} - X_{jk}\right), \quad (6.49)$$

g – кількість класів;

nk – кількість спостережень у k-му класі;

X_{jkm} – значення m-го спостереження i-ї змінної у k-му класі;

X_{ik} – середнє значення i-ї змінної у k-му класі.

Для використання методу необхідно виконання таких умов:

– обсяг вибірки має бути більшим, ніж кількість змінних;

– кластери, серед яких здійснюють дискримінацію, підпорядковані багатовимірному нормальному розподілу;

– класи можуть перетинатися, але їх центри мають бути достатньо віддаленими один від одного;

– різниця між коваріаційними матрицями цих кластерів є статистично незначущою.

Останнє припущення спрощує обчислювальну процедуру. Але його необґрунтоване застосування може призвести до втрати найсуттєвіших 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 збігаються з результатами більш складного методу канонічного дискримінантного аналізу.

Канонічний дискримінантний аналіз ґрунтується на знаходженні канонічних дискримінантних функцій:

$$f_{km} = u_0 + \sum_{i=1}^{p} u_i X_{ikm}\,, \tag{6.50}$$

де u_i – коефіцієнти, що визначають за формулою:

$$u_i = v_i \sqrt{n-g}\,,\ u_0 = -\sum_{i=1}^{p} u_i \overline{X}_i\,, \tag{6.51}$$

$\overline{X}_i$ – середнє значення i-ї змінної за всіма класами;

v_i – коефіцієнти, що розраховують як компоненти власних векторів розв'язку узагальненої проблеми власних значень:

$$B\nu = \lambda W \nu\,, \tag{6.52}$$

B – міжгрупова сума квадратів відхилень;

v – власний вектор.

Інші позначення збігаються з тими, що були використані для лінійного дискримінантного аналізу. Кількість дискримінантних функцій може бути меншою або рівною кількості параметрів об'єкта.

Матрицю B визначають як:

$$B = T - W\,, \tag{6.53}$$

де T – матриця сум квадратів і попарних добутків:

$$t_{ij} = \sum_{k=1}^{g} \sum_{m=1}^{n_k} \left(X_{imk} - \overline{X}_i \right) \left(X_{jmk} - \overline{X}_j \right). \qquad (6.54)$$

Зарахування нових некласифікованих об'єктів до заданих кластерів здійснюють 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, передбачають, що коваріаційні матриці різних вибірок є різними. Це істотно ускладнює процедуру розрахунків. Відмова від припущення про статистичну нерозрізненість коваріаційних матриць для кластерів, що навчають, зумовлює необхідність 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 з 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 змістову інтерпретацію одержуваних параметрів.

6.4. Приклади здійснення класифікації даних

Приклад параметричної класифікації даних без навчання розглянуто у п. 2.6.

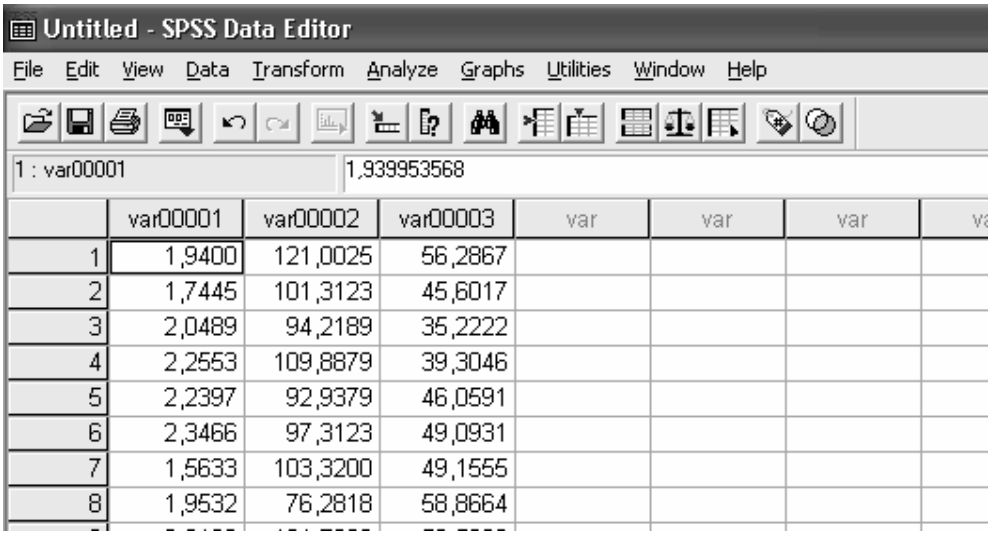
Як приклад здійснення кластерного аналізу розглянемо таку задачу. Сформуємо вибірку, що містить чотири класи об'єктів, кожний з яких характеризується трьома кількісними ознаками. Значення ознак сформуємо як нормально розподілені випадкові величини з параметрами, що наведено у табл. 6.2.

Таблиця 6.2

Параметри розподілу ознак для прикладу кластерного аналізу

Номер кластера	$x_{1\text{ср}}$	s_1	$x_{2\text{ср}}$	s_2	$x_{3\text{ср}}$	s_3	n
1	2	0,2	100	10	49	5	20
2	3	0,4	83	8	78	7	15
3	1	0,3	88	9	32	4	25
4	5	0,5	53	8	37	5	30

Занесемо отримані значення до таблиці вихідних даних пакету SPSS (рис. 6.5)



	var00001	var00002	var00003	var	var	var	va
1	1,9400	121,0025	56,2867				
2	1,7445	101,3123	45,6017				
3	2,0489	94,2189	35,2222				
4	2,2553	109,8879	39,3046				
5	2,2397	92,9379	46,0591				
6	2,3466	97,3123	49,0931				
7	1,5633	103,3200	49,1555				
8	1,9532	76,2818	58,8664				

Рис. 6.5. Таблиця вихідних даних кластерного аналізу

Далі обираємо в меню: Analyze/Classify/K-means Cluster … При цьому з’являється діалогове вікно, показане на рис. 6.6.

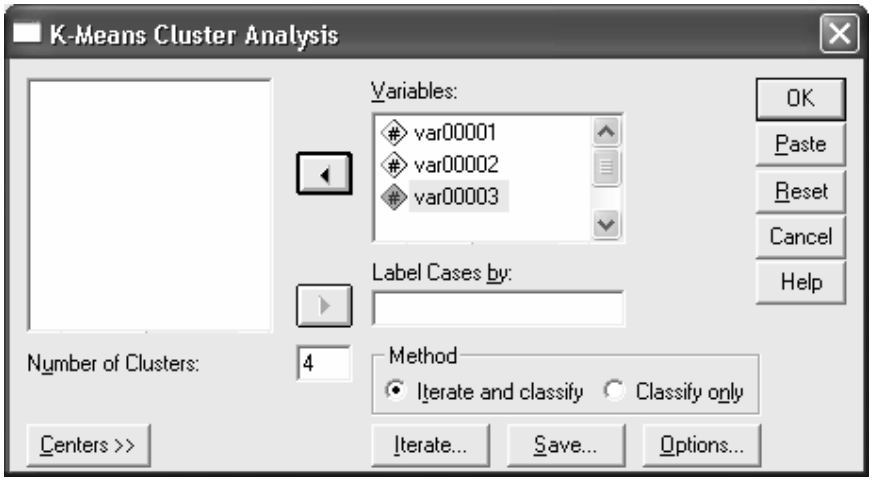


Рис. 6.6. Діалогове вікно кластерного аналізу методом *k*-середніх

У цьому вікні необхідно вказати змінні, що характеризують досліджувані об’єкти, кількість кластерів, що потрібно виокремити та метод кластеризації. Крім того можна встановити додаткові параметри процедури.

За допомогою кнопки “Centers” вказуємо, чи потрібно зчитати з окремого файлу початкові значення центрів кластерів, або записати до файлу кінцеві значення центрів.

За допомогою кнопки “Iterate” відкриваємо вікно установки параметрів ітераційної процедури (рис. 6.7).

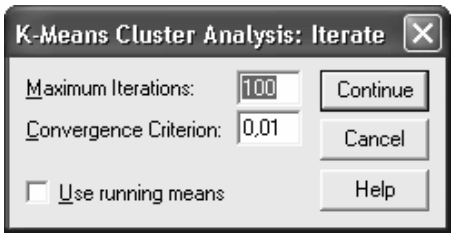


Рис. 6.7. Вікно визначення параметрів ітераційної процедури

У цьому вікні необхідно задати:

– максимальну кількість ітерацій (від 1 до 999);

– критерій збіжності (від 0 до 1); у прикладі, наведеному на рис. 6*, ітерації закінчують, якщо на останній ітерації максимальний зсув центрів кластерів не перевищує 1% від мінімальної різниці між центрами початкових кластерів;

– вказати необхідність застосування ковзних середніх при розрахунку центрів кластерів під час кожної ітерації.

За допомогою кнопки “Save” відкриваємо діалогове вікно (рис. 6.8), у якому вказуємо, чи потрібно зберегти як нові змінні значення, що характеризують приналежність об’єктів кластерам, та центри отриманих кластерів.

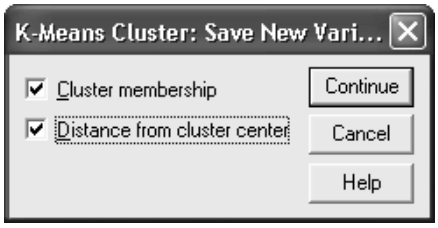


Рис. 6.8. Діалогове вікно запису нових змінних

За допомогою кнопки “Options” відкриваємо діалогове вікно (рис. 6.9), в якому вказуємо додаткові параметри процедури кластеризації: необхідність виводу початкових значень центрів кластерів, таблиці ANOVA й інформації про кількість об’єктів у кожному кластері, а також спосіб обробки пропущених значень.

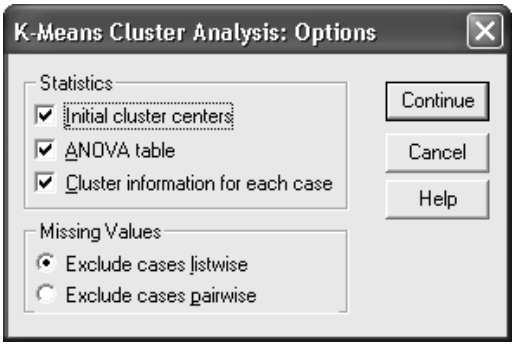


Рис. 6.9. Вікно задання додаткових параметрів процедури кластеризації

Деякі результати наведено на рис. 6.10.

Final Cluster Centers

	Cluster			
	1	2	3	4
VAR00001	1,9177	2,8655	1,0529	5,0436
VAR00002	100,2400	82,0074	88,0278	53,6599
VAR00003	49,2148	74,1121	33,8212	38,5363

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
VAR00001	82,102	3	,165	86	499,014	,000
VAR00002	9716,806	3	63,743	86	152,438	,000
VAR00003	6156,787	3	28,537	86	215,747	,000

Distances between Final Cluster Centers

Cluster	1	2	3	4
1		30,874	19,668	47,891
2	30,874		40,778	45,541
3	19,668	40,778		34,919
4	47,891	45,541	34,919	

Number of Cases in each Cluster

Cluster	1	17,000
	2	16,000
	3	27,000
	4	30,000
Valid		90,000
Missing		,000

Рис. 6.10. Деякі результати кластеризації

З наведених даних бачимо, що отримані результати дещо відрізняються від заданих (при цьому слід мати на увазі, що номера вихідних кластерів можуть не збігатися з номерами, наданими кластерам при використанні процедури кластеризації методом *k*-середніх). Визначити помилково класифіковані дані можна, порівнюючи вихідні дані з отриманою таблицею приналежності. Її фрагмент показано на рис. 6.11. Для випадку, що розглядається: два об’єкти першого кластера помилково зараховані до третього, а один – до другого; всі об’єкти інших кластерів класифіковано правильно. Таким чином, загальна частка помилково класифікованих даних дорівнює приблизно 3,3%, що вважають задовільним результатом. Слід зазначити, що в розглянутому прикладі вихідні дані було взято так, що кластери були достатньо добре відокремлені один від одного. У реальності ситуація часто буває більш складною.

Cluster Membership

Case Number	Cluster	Distance
1	1	21,934
2	1	3,773
3	3	6,425
4	1	13,835
5	1	7,961
6	1	2,961
7	1	3,101
8	2	16,311
9	1	10,433
10	1	4,112
11	1	9,242
12	1	4,009
13	3	7,111

Рис. 6.11. Фрагмент таблиці приналежності

У нашому випадку кількість класів була відомою заздалегідь. Але зазвичай це не так. Для встановлення оптимальної кількості виокремлюваних класів доцільно виконати розрахунки для декількох варіантів, а потім порівняти результати аналізу. Корисну інформацію для цього дає таблиця ANOVA. Виходячи з того, що кластери формуються з умови максимізації дисперсії між класами, як оптимальний варіант кількості кластерів доцільно взяти той, якому відповідають максимальні значення *F*-критерію у таблиці ANOVA. Іншим підґрунтям для визначення оптимальної кількості кластерів є порівняння відстаней аналізованих об'єктів від центрів кластерів, які наводяться в таблиці "Cluster membership" вихідних результатів аналізу, та відстаней між центрами кластерів, що наводяться у таблиці "Distances between Final Cluster Centers". При правильної кластеризації відстані від центрами кластерів мають перевищувати відстані від об'єктів до центрів кластерів. За малої кількості виокремлюваних кластерів зростають максимальні значення відстаней від центрів кластерів до об'єктів, а за великої – зменшуються відстані між центрами кластерів. В обох випадках це призводить до порушення правильного співвідношення між цими параметрами.

Розглянемо розв'язання тієї самої задачі ієрархічними методами в пакеті SPSS. Для цього обираємо у меню: Analyze/Classify/Hierarchial Claster … У результаті з'являється діалогове вікно, показане на рис. 6.12.

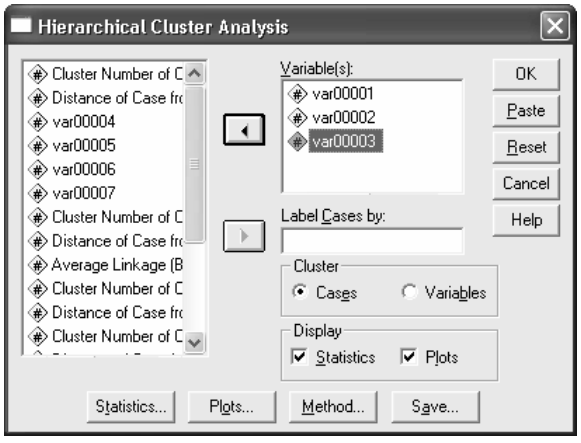


Рис. 6.12. Діалогове вікно ієрархічного кластерного аналізу

У цьому вікні зазначаємо, за якими змінними здійснюватиме кластеризацію; що саме об’єднуватиме в кластери – об’єкти чи змінні; необхідність виведення статистики й графіків. Крім того, за допомогою кнопок “Statistics”, “Plots”, “Method”, “Save” задаємо додаткові параметри кластеризації.

За допомогою кнопки “Statistics” виводимо діалогове вікно (рис. 6.13), в якому задаємо такі параметри:

– Agglomeration schedule – необхідність складу кластерів на кожному кроці ітерацій;

– Proximity matrix – необхідність виводу відстаней чи показників подібності між об’єктами;

– Cluster Membership – необхідність виводу приналежності об’єктів кластерам на однієї чи декількох стадіях агломерації.

За допомогою кнопки “Plots” виводимо діалогове вікно (рис. 6.14), в якому задаємо такі параметри:

– Dendrogram – необхідність виводу дендрограми;

– Icicle – необхідність виводу діаграми, що показує, як об’єкти об’єднуються в кластери на кожному кроці ітерацій. Для цього графіка можливо задати горизонтальну чи вертикальну орієнтацію.

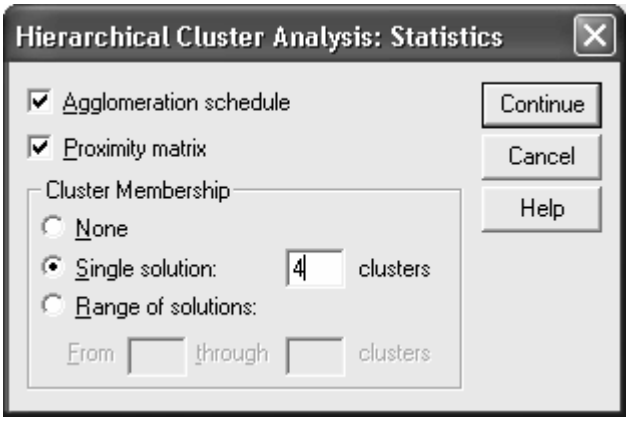


Рис. 6.13. Діалогове вікно статистики ієрархічного кластерного аналізу

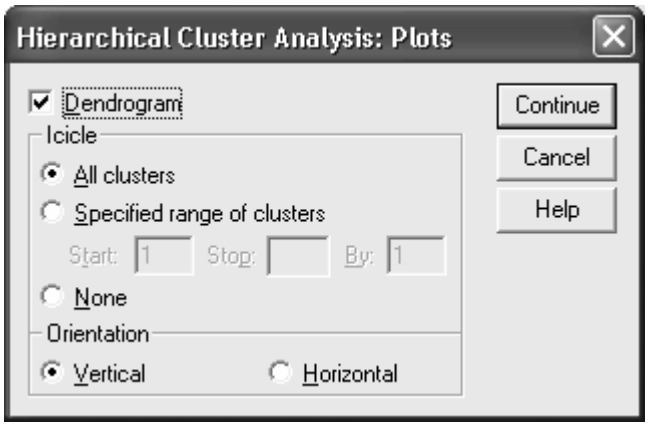


Рис. 6.14. Діалогове вікно задання параметрів графіків ієрархічного кластерного аналізу

За допомогою кнопки “Method” виводимо діалогове вікно (рис. 6.15), в якому задаємо такі параметри:

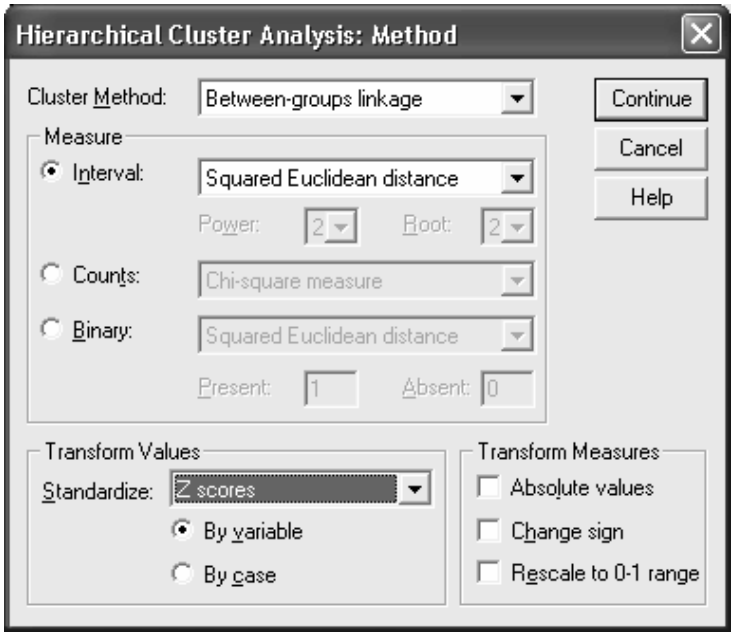


Рис. 6.15. Діалогове вікно методів ієрархічного кластерного аналізу

Cluster Method – дає можливість вибрати метод міжгрупових відстаней; всередінігрупових відстаней; ближчих сусідів; далеких сусідів; центроїдний, медіанний або Уорда.

Measure – дає можливість задати міри відстані чи зв’язку для різних типів даних – інтервальних, порядкових або бінарних.

Transform Values – дає змогу вибрати метод стандартизації вихідних даних (*Z*-перетворення відповідає z_1 у формулі 6.11).

Transform Measures – дає змогу перетворювати отримані відстані.

За допомогою кнопки “Save” виводимо діалогове вікно (рис. 6.16), в якому вказуємо необхідність збереження у вікні даних нових змінних, що характеризують належність об’єктів до кластерів для єдиного розв’язку чи для певного набора розв’язків.

Деякі результати кластеризації показано на рис. 6.17–6.19.

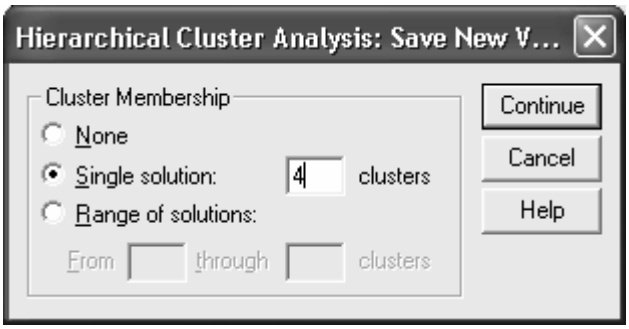


Рис. 6.16. Діалогове вікно збереження нових змінних

Case	1:Case 1	2:Case 2	3:Case 3	4:Case 4	5:Case 5	6:Case 6	7:(
1:Case 1	,000	1,495	3,735	1,578	2,500	1,719	
2:Case 2	1,495	,000	,619	,447	,265	,217	
3:Case 3	3,735	,619	,000	,716	,516	,873	
4:Case 4	1,578	,447	,716	,000	,932	,817	
5:Case 5	2,500	,265	,516	,932	,000	,092	
6:Case 6	1,719	,217	,873	,817	,092	,000	
7:Case 7	1,068	,075	1,119	,687	,474	,302	
8:Case 8	5,168	2,373	3,208	4,561	1,439	1,596	
9:Case 9	1,026	,903	2,669	1,913	,971	,520	
10:Case 10	1,388	,246	1,384	1,200	,339	,170	
11:Case 11	,630	,201	1,314	,384	,765	,461	
12:Case 12	1,876	,082	,737	,887	,159	,169	
13:Case 13	4,503	,829	,389	1,750	,444	,916	
14:Case 14	3,065	,984	1,928	2,594	,553	,577	
15:Case 15	,618	,599	1,678	,297	1,486	1,111	
16:Case 16	3,136	,576	1,079	1,911	,283	,457	

Рис. 6.17. Матриця відстаней між об’єктами

Cluster Membership

Case	4 Clusters
1:Case 1	1
2:Case 2	1
3:Case 3	2
4:Case 4	1
5:Case 5	1
6:Case 6	1
7:Case 7	1
8:Case 8	1
9:Case 9	1
10:Case 10	1
11:Case 11	1

Рис. 6.18. Таблиця приналежності об’єктів кластерам

У цілому результати ієрархічної кластеризації для прикладу, що розглядається, збігаються з результатами, отриманими у попередньому випадку. Як і у попередньому випадку маємо три помилково класифіковані об’єкти, що належать першому кластеру. Аналіз таблиць приналежності об’єктів показує, що помилково класифікованими в обох випадках виявляються об’єкти, що мають великі відхилення від середнього значення за однією або декількома ознаками.

```
* * * * * * H I E R A R C H I C A L  C L U S T E R  A N A L Y S I S *

 Dendrogram using Average Linkage (Between Groups)

                          Rescaled Distance Cluster Combine

   C A S E          0         5        10        15        20        25
 Label     Num  +---------+---------+---------+---------+---------+

 Case 26    26   **
 Case 30    30   **
 Case 25    25   **
 Case 34    34   **
 Case 42    42   **
 Case 41    41   **
 Case 39    39   **
 Case 21    21   **
 Case 35    35   **
 Case 29    29   ****
 Case 37    37   ** *
 Case 23    23   ** *
 Case 27    27   ** *
 Case 33    33   ** *
 Case 13    13   ** ***
 Case 32    32   ** * *
 Case 22    22   ** * *
 Case 24    24   ** * *
 Case 44    44   ** * *
 Case 45    45   ** * *
 Case 28    28   **** *
 Case 31    31   **   *
 Case 36    36   **   *
 Case 38    38   **   *********
 Case 43    43   **   *        *
 Case 3      3   ******       *
 Case 20    20   **   *        *
 Case 40    40   ******       *
 Case 14    14   **           *
 Case 19    19   ****         *
 Case 16    16   ** *****     ****************************
 Case 8      8   ****   *     *                           *
 Case 11    11   **     *     *                           *
 Case 15    15   ****** ***** *                           *
 Case 4      4   **   * *   * *                           *
 Case 5      5   **   * *   * *                           *
 Case 6      6   **** ***   * *                           *
 Case 2      2   ** * *     * *                           *
 Case 7      7   ** * *     ***                           *
 Case 12    12   ** ***     *                             ***********
 Case 17    17   ** *       *                             *          *
 Case 10    10   ** *       *                             *          *
 Case 18    18   ** *       *                             *          *
 Case 9      9   ****       *                             *          *
 Case 1      1   ************                             *          *
 Case 76    76   ******                                   *          *
 Case 86    86   **   ***                                 *          *
```

Рис. 6.19. Фрагмент дендрограми

Контрольні питання

1. Що називають класифікацією даних?
2. Який критерій найчастіше застосовують для розробки процедур класифікації даних?
3. Що називають функцією втрат? Для чого її використовують?
4. У чому полягає основна різниця між методами класифікації з навчанням та без навчання?
5. Якою є основна передумова застосування параметричних методів класифікації без навчання?
6. Що є результатом параметричної класифікації без навчання?
7. У якому вигляді подають вихідну інформацію для використання непараметричних методів класифікації без навчання?
8. Як формулюють задачі класифікації в непараметричних методах класифікації без навчання? Чи завжди ці задачі мають розв'язок?
9. Що називають класом в кластерному аналізі?
10. Які міри відстані найчастіше використовують у кластерному аналізі?
11. Які міри зв'язку найчастіше використовують у кластерному аналізі?
12. Які вимоги має задовольняти функція, яку використовують як міру відстані?
13. Які фактори слід враховувати при виборі метрики? Наведіть приклади.
14. Що називають мірою Махаланобиса? Як вона пов'язана з іншими мірами відстані?
15. Як визначається евклідова відстань? У яких випадках її доцільно застосовувати у кластерному аналізі? Наведіть приклади.
16. Які перетворення використовують для нормування даних у кластерному аналізі? Наведіть приклади.
17. Які міри близькості кластерів один до одного використовують у кластерному аналізі? Наведіть приклади.
18. Які функціонали якості розбиття найчастіше використовують у кластерному аналізі?
19. Які типи задач виділяють у кластерному аналізі з погляду інформації про кількість класів?
20. У чому полягає сутність ієрархічних методів кластерного аналізу?
21. Якими є основні переваги й недоліки ієрархічних методів кластерного аналізу?
22. Яким є базовий алгоритм методу ближнього зв'язку в кластерному аналізі? Наведіть приклади застосування цього методу.
23. Що називають дендрограмою? Як можна побудувати дендрограму?
24. Яким є базовий алгоритм методу середнього зв'язку Кінга у кластерному аналізі? Наведіть приклади застосування цього методу.

25. Яким є базовий алгоритм методу *k*-середніх Мак-Куїна у кластерному аналізі? Наведіть приклади застосування цього методу.

26. У чому полягає сутність методу кореляційних плеяд у кластерному аналізі? Наведіть приклади застосування цього методу.

27. У чому полягає сутність методу вроцлавської таксономії у кластерному аналізі? Наведіть приклади застосування цього методу.

28. Як формулюють задачу побудови оптимальної класифікації у методах класифікації з навчанням?

29. Що називають розв'язувальним правилом, або дискримінантною функцією?

30. У якій формі можна записати розв'язувальні правила?

31. У якому випадку розв'язувальне правило називають оптимальним?

32. У чому полягає сутність дискримінантного аналізу?

33. Якою є базова процедура параметричного дискримінантного аналізу у випадку нормальних класів?

34. Якою є загальна схема байєсівських методів класифікації з навчанням?

35. Якою є базова процедура лінійного дискримінантного аналізу Фішера? У чому полягають передумови його застосування?

36. Якою є базова процедура канонічного дискримінантного аналізу?

7. МЕТОДИ ПОБУДОВИ Й ДОСЛІДЖЕННЯ РЕГРЕСІЙНИХ МОДЕЛЕЙ

Завданням дослідження складних систем і процесів часто є перевірка наявності й 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**) z. Розв'язання таких завдань є предметом регресійного аналізу. Термін "Регресія" вперше був уведений Ф. Гальтоном наприкінці XIX ст. На практиці завдання регресійного аналізу зазвичай формулюють так: необхідно підібрати достатньо просту функцію, що в певному розумінні найкращим чином описує наявну сукупність емпіричних даних.

7.1. Загальна характеристика методів і 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 й параметри є детермінованими, а відгуки – рівноточними (тобто мають однакові дисперсії) некорельованими випадковими величинами. Передбачається також, що всі змінні вимірюють 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 застосування процедур класичного регресійного аналізу і потребує застосування більш складних методів аналізу емпіричних даних.

Постулат про рівноточність і некорельованість відгуків не є обов'язковим. У випадку його невиконання процедура побудови регресійної моделі певною мірою змінюється, але суттєво не ускладнюється.

Більш складною проблемою є вибір моделі та її незалежних змінних. У класичному регресійному аналізі припускають, що набір факторів задається однозначно, всі суттєві змінні наявні в моделі й немає ніяких альтернативних способів обрання факторів. На практиці це припущення не ви-

конується. Тому виникає необхідність розробки формальних та неформальних процедур перетворення й порівняння моделей. Для пошуку оптимальних формальних перетворень використовують методи факторного та дискримінантного аналізу. На сьогодні розроблено комп'ютеризовані технології послідовної побудови регресійних моделей.

Фактори в класичному регресійному аналізі вважають детермінованими, тобто вважається, що дослідник має про них всю необхідну інформацію з абсолютною точністю. На практиці це припущення часто не виконується. Відмова від детермінованості незалежних змінних зумовлює необхідність застосування моделей кореляційного аналізу. В окремих випадках можна використовувати компромісні методи **конфлюентного аналізу**, які передбачають 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. Але на практиці це припущення виконується не завжди. У деяких випадках можна вважати параметри випадковими величинами із заданими законами розподілу. Тоді як оцінки параметрів можна брати їх умовні математичні сподівання для відгуків, що спостерігалися. Умовні розподіли та математичні 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 зумовлює необхідність переходу від випадкових значень відгуків до випадкових послідовностей, випадкових процесів або випадкових полів. Однією з поширених і найпростіших моделей такого типу є **модель авторегресії**, згідно з якою 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 найчастішими є випадки, коли цей закон невідомий чи відомо, що він не є нормальним. Їх дослідження зумовило виникнення непараметричного регресійного 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).

Одним з основних постулатів класичного регресійного аналізу є припущення, що найкращі оцінки параметрів можна одержати, використовуючи метод найменших квадратів. На практиці оцінки, одержані за допомогою цього методу, часто бувають недостатньо точними і містять великі похибки. Причиною цього може бути структура регресійної моделі. Якщо вона є лінійною комбінацією експонент або поліномом високого степеня, то це призводить до поганої зумовленості матриці системи нормальних рівнянь і нестійкості оцінок параметрів. Підвищення стійкості оцінок можна досягти шляхом відмови від вимоги щодо їх незміщеності. Розвиток цього напряму досліджень призвів до виникнення гребеневого, або рідж-регресійного аналізу.

Найчастіше задачу побудови регресійної моделі формулюють так. Необхідно знайти функцію заданого класу, для якої функціонал:

$$F(\alpha) = \sum_{i=1}^{n} \left(z_i(\alpha, X) - y_i \right)^2 \rightarrow \min . \qquad (7.1а)$$

У цьому виразі $z_i(\alpha, X)$ – значення функції, що апроксимує залежність, в i-ї точці, y_i – відповідне значення емпіричної залежності, α – вектор параметрів, які треба знайти, X – вектор незалежних змінних. Одержану функцію $z(\alpha, X)$ називають (**середньоквадратичною**) **регресійною моделлю**. Метод її пошуку, який базується на застосуванні критерію (7.1а), називають методом найменших квадратів.

Іноді замість функціонала (7.1а) для визначення параметрів регресійних моделей розв'язують задачі мінімізації інших функціоналів, зокрема:

$$F(\alpha) = \sum_{i=1}^{n} \left| z_i(\alpha, X) - y_i \right| \rightarrow \min ; \qquad (7.1б)$$

$$F(\alpha) = \max \left| z_i(\alpha, X) - y_i \right| \rightarrow \min . \qquad (7.1в)$$

Одержувані при цьому регресійні моделі називають, відповідно, **середньоабсолютними** (**медіанними**) та **мінімаксними**. Ці моделі найчастіше використовують при побудові робастних алгоритмів регресійного аналізу, але їх практичне застосування обмежується поганою збіжністю таких алгоритмів.

Апроксимуючу функцію у випадку однієї незалежної змінної (моделі простої регресії) часто шукають у вигляді полінома $z(x) = \sum_{j=0}^{M} \alpha_j x^j$, обер-

неного полінома $z(x) = \dfrac{1}{\sum_{j=0}^{M} \alpha_j x^j}$, експоненціальних або показникових функцій $z = \alpha e^x$ чи $z = \alpha b^x$, степеневої функції $z = \alpha x^b$, лінійно-логарифмічної функції $z = \alpha_1 + \alpha_2 x + \alpha_3 1nx$, тригонометричного ряду Фур'є тощо. За наявності декількох незалежних змінних (моделі множинної регресії) найчастіше використовують функції, лінійні як за параметрами, так і за незалежними змінними $z = \alpha_0 + \sum_{i=1}^{p} \alpha_i x_i$, а також поліноміальні моделі, що 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 розкладу функції відгуку в ряд Тейлора. Проте можливе й використання для апроксимації інших видів 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:

$$z = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \varepsilon, \qquad (7.2)$$

де ε – похибка моделі, є лінійною моделлю третього порядку.

Вибір типу регресійної моделі є нетривіальним завданням. Для моделей, що містять одну незалежну змінну, рекомендують спочатку нанести наявні емпіричні дані на графік. Це дає можливість визначити наявність чи відсутність залежності між досліджуваними величинами, а також зробити певні припущення про тип залежності.

На рис. 7.1 як приклади наведено певні набори емпіричних точок, для яких потрібно побудувати регресійні моделі. З наведених графіків видно, що ці моделі доцільно будувати у вигляді лінійної, квадратичної та експоненціальної функцій, відповідно. Але, як правило, визначення типу моделі за графіком емпіричних даних є не настільки очевидним, тому зазвичай доводиться перевіряти декілька варіантів моделі і вибирати кращий з них за певними критеріями.

Часто як попередній етап регресійного аналізу рекомендують за допомогою методів кореляційного аналізу перевіряти наявність значущого зв'язку між досліджуваними змінними. Але при цьому слід ураховувати, що звичайні методи кореляційного аналізу дають змогу перевіряти лише гіпотезу про наявність лінійного зв'язку. Якщо зв'язок є, але він нелінійний, висновки, отримані за допомогою кореляційного аналізу, можуть бути помилковими.

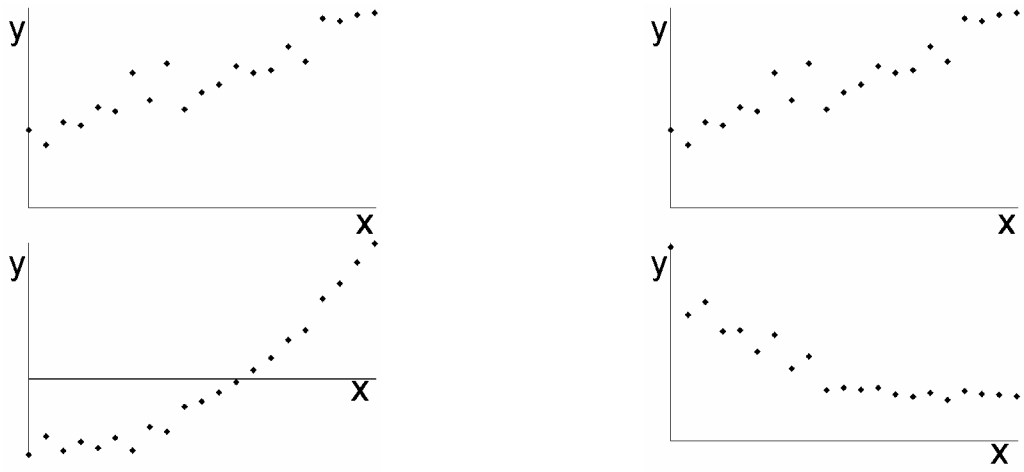


Рис. 7.1. Приклади наборів даних, для яких треба побудувати регресійні моделі

Важливою особливістю регресійних моделей є те, що їх не можна застосовувати поза межами тієї області значень вихідних параметрів, для якої вони були побудовані. При використанні регресійних моделей типу полінома, оберненого полінома, тригонометричного ряду та деяких інших слід враховувати, що, збільшуючи кількість членів ряду, можна одержати скільки завгодно близькі до нуля значення функціоналів (7.1). Проте це не завжди свідчить про якість апроксимації, оскільки ці функціонали не дають інформації про ступінь наближення моделі до емпіричної залежності у проміжках між наявними точками.

Іншою проблемою може бути наявність декількох локальних екстремумів функціоналів (7.1). У таких випадках необхідно враховувати, що більшість стандартних алгоритмів дає можливість знаходити лише локальні, а не глобальні екстремуми функціоналів, і результат мінімізації залежать від вибору початкових умов пошуку. Це часто зумовлює необхідність встановлення додаткових критеріїв вибору моделі, серед яких можуть бути як формальні критерії їх адекватності, так і неформальні критерії, що ґрунтуються на сукупності відомих даних про об’єкт дослідження.

Поліноміальні регресійні моделі, як правило, є формальними. Їх використовують для опису систем і процесів, теорію яких розроблено недостатньо. При цьому спираються на відомі властивості ряду Тейлора для аналітичних функцій. Більш цікавими для дослідників зазвичай є математичні моделі, які відображають структуру та зв’язки у системах, сутність і механізми процесів, що відбуваються у них. Якщо теоретичні основи досліджуваних систем і процесів достатньо розроблені, часто постає проблема визначення окремих параметрів моделі за наявними емпіричними даними. Для її вирішення у багатьох випадках можна використовувати формальні процедури регресійного аналізу.

На практиці часто доводиться користуватися нелінійними за параметрами та багатовимірними моделями. Під багатовимірними тут розуміють моделі, що розглядають декілька відгуків. Задачам, що розв’язуються у межах

відповідних напрямів регресійного аналізу, властиві й інші ускладнення. Так у багатовимірних моделей окремі відгуки можуть бути пов'язані один з одним. Сама регресійна модель часто задається у неявному вигляді та є неаналітичним розв'язком певної системи алгебраїчних або диференціальних рівнянь. Нестійкість оцінок параметрів для нелінійних моделей різко зростає. Як правило, такі задачі мають декілька розв'язків або не мають розв'язків взагалі.

7.2. Лінійні однофакторні моделі

Найпростішим для аналізу і найбільш дослідженим є випадок лінійної кореляційної залежності між двома змінними X та Y. Наявність лінійного зв'язку можна перевірити, розрахувавши коефіцієнт парної кореляції Пірсона (4.7).

Розглянемо детальніше задачу підбору параметрів лінійної моделі:

$$z(x) = \alpha_0 + \alpha_1 x + \varepsilon \tag{7.3}$$

за набором наявних емпіричних точок (x_i, y_i).

У **методі найменших квадратів** (МНК) виходять з припущення, що найкращими значеннями параметрів α_0 і α_1 будуть ті, для яких сума квадратів відхилень емпіричних значень y_i від розрахункових значень $z(x_i)$ набуває мінімального можливого значення. Можна довести, що МНК оцінки мають такі властивості:

– вони є лінійними функціями результатів спостережень і незміщеними оцінками параметрів моделі;

– згідно з теоремою Гауса – Маркова, МНК оцінки мають найменші дисперсії серед усіх інших оцінок, що є лінійними функціями результатів спостережень;

– МНК оцінки збігаються з оцінками, які обчислюють методом найбільшої правдоподібності.

Для знаходження таких значень параметрів необхідно розв'язати систему:

$$\begin{cases} \dfrac{\partial}{\partial \alpha_0} \sum\limits_{i=1}^{n} \left[z(x_i) - y_i \right]^2 = \dfrac{\partial}{\partial \alpha_0} \sum\limits_{i=1}^{n} \left[\alpha_1 x_i + \alpha_0 - y_i \right]^2 = 0 , \\ \dfrac{\partial}{\partial \alpha_1} \sum\limits_{i=1}^{n} \left[z(x_i) - y_i \right]^2 = \dfrac{\partial}{\partial \alpha_1} \sum\limits_{i=1}^{n} \left[\alpha_1 x_i + \alpha_0 - y_i \right]^2 = 0 . \end{cases} \tag{7.4}$$

З (7.4) можна одержати такі вирази для оцінок α^*_0 і α^*_1 коефіцієнтів лінійної залежності:

$$\alpha_1^* = \frac{n \sum\limits_{i=1}^{n} x_i y_i - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} y_i}{n \sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i \right)^2};$$

$$\alpha_0^* = \overline{Y} - \alpha_1 \overline{X} = \frac{\sum_{i=1}^{n} y_i - \alpha_1 \sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} . \qquad (7.5)$$

Перше рівняння в (7.5) є відношенням коваріації $\operatorname{cov}(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)$ до дисперсії $\sigma_X^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2$, тобто:

$$\alpha_1^* = \frac{\operatorname{cov}(x,y)}{\sigma_X^2} = \frac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}. \qquad (7.6)$$

У багатьох випадках, завдяки особливостям округлення чисел у ЕОМ, останній вираз дає змогу отримати точніші оцінки параметрів, ніж (7.5).

У випадку однофакторної лінійної моделі існує зв'язок між коефіцієнтом a_1 моделі, коефіцієнтом кореляції предиктора і відгуку, а також їх дисперсіями:

$$r_{xy} = a_1 \frac{\sigma_x}{\sigma_y}. \qquad (7.7)$$

Як приклад розглянемо таку задачу. У табл. 7.1 подано емпіричні дані, для яких треба побудувати регресійну модель, а також дані, необхідні для розрахунку її параметрів. На рис. 7.2 наведено емпіричні точки та графік залежності, побудований за одержаною методом найменших квадратів моделлю. Як видно з рис. 7.2, одержана модель задовільно описує наявні емпіричні дані.

Таблиця 7.1

Приклад розрахунку регресійної моделі

№ випробування	x_i	y_i	x_i^2	$x_i \cdot y_i$	z	Залишок
1	0	0,310	0	0	0,516	0,206
2	0,3	1,037	0,09	0,311	1,506	0,470
3	0,6	2,513	0,36	1,508	2,497	−0,017
4	0,9	3,843	0,81	3,459	3,487	−0,356
5	1,2	4,840	1,44	5,807	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	15,686	7,447	−0,022
9	2,4	8,889	5,76	21,332	8,438	−0,451
10	2,7	9,25399	7,29	24,98577	9,427681	0,174

Продовження табл. 7.1

11	3	10,39294	9	31,17882	10,41785	0,025
12	3,3	11,11287	10,89	36,67247	11,40801	0,295
Σ	19,8	71,54526	45,54	160,5277		0
α_1^*	3,300548	α_0^*	0,516201			

Для розглянутої регресійної моделі сума залишків $\sum_{i=1}^{n}\left(y_i - z(x_i)\right)$ дорівнює нулю, якщо модель містить вільний член α_0. Виключення вільного члена з моделі зазвичай є невиправданим. Використання моделі з $\alpha_0 = 0$ доцільно лише у випадках, коли з теорії відомо, що для нульових значень предикторів відгук має дорівнювати нулю. Якщо це невідомо, але бажано одержати модель, що не містить вільного члена, більш доцільним є застосування центрування даних.

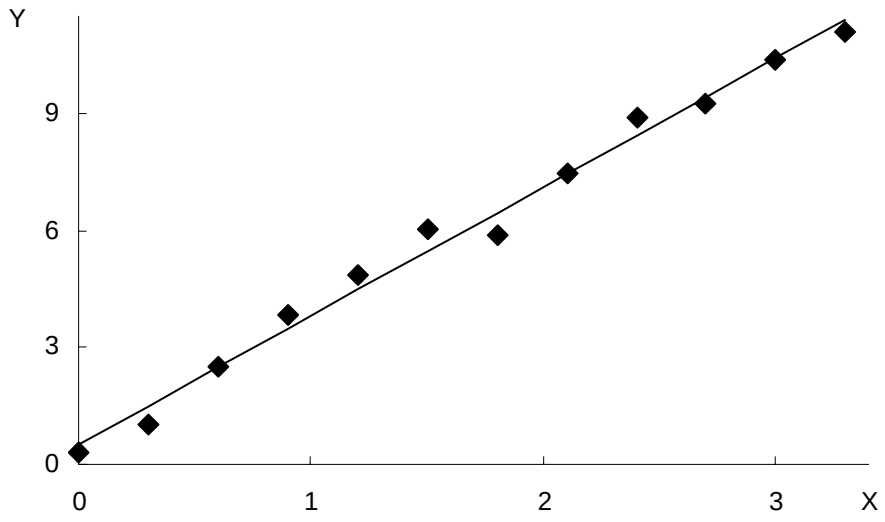


Рис. 7.2. Графіки досліджуваної залежності й лінійної моделі

Підставляючи до моделі $z(x) = \alpha_0 + \alpha_1 x + \varepsilon$ оцінку коефіцієнта α^*_0 з (7.5), можна одержати:

$$Y^* = \alpha_0^* + \alpha_1^* x + \varepsilon = \overline{Y} + \alpha_1^*\left(x - \overline{X}\right) + \varepsilon. \qquad (7.8)$$

Звідси отримуємо **центровану модель**:

$$Y - \overline{Y} = \alpha_1^*\left(x - \overline{X}\right) + \varepsilon, \qquad (7.9)$$

яка не містить вільного члена.

Незважаючи на те, що, як правило, 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 простим і дослідженим. Зокрема,

для нього достатньо повно розроблені процедури визначення статистичних характеристик одержуваних оцінок параметрів (дисперсії, довірчих інтервалів тощо) та перевірки адекватності моделей. По-друге, у багатьох випадках складні залежності можна подати як набір лінійних (на малих відрізках змінювання факторів) залежностей. По-третє, нелінійні залежності у деяких випадках можна перетворити до лінійного вигляду шляхом заміни змінних. Деякі приклади такого перетворення наведено у табл. 7.2.

Таблиця 7.2

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, $\ln \frac{\alpha_0}{\alpha_0 - z}$
$z = \alpha_0 exp(-\alpha_1/x)$	$\ln z = \ln \alpha_0 - \alpha_1/x$	$1/x$, $\ln z$
$z = z = \alpha_0 x^{\alpha_1}$	$\ln z = \ln \alpha_0 + \alpha_1 \ln x$	$\ln x$, $\ln z$
$z = \alpha_0 x + \alpha_1 x^2$	$z/x = \alpha_0 + \alpha_1 x$	x, z/x
$z = \alpha_0 \sin(\alpha_1 x)$	$\arcsin (z/\alpha_0) = \alpha_1 x$	x, $\arcsin (z/\alpha_0)$

Перетворення нелінійних залежностей до лінійних є строго обґрунтованим, якщо вихідні дані є точними. На практиці вони завжди вимірюються з деякою похибкою. Розглянемо модель:

$$z = \alpha_0 x^{\alpha_1} + \varepsilon, \tag{7.10}$$

де ε – похибка вимірювань.

Її лінеаризована форма матиме вигляд:

$$\ln z = \ln \alpha_0 + \alpha_1 \ln x + \varepsilon', \tag{7.11}$$

де ε' є невідомою випадковою величиною. Використання як лінеаризованої форми виразу:

$$\ln z = \ln \alpha_0 + \alpha_1 \ln x \tag{7.12}$$

є коректним лише у тому випадку, коли величина ε' є малою порівняно з іншими доданками правої частини (7.12).

Розглянемо питання про точність оцінок. Для цього запишемо таку тотожність:

$$\left(y_i - \overline{Y}\right) = \left(y_i^* - \overline{Y}\right) + \left(y_i - y_i^*\right). \tag{7.13}$$

Тут y^*_i є оцінкою значення величини у при $x = x_i$. Якщо піднести обидві частини цієї тотожності до квадрата та взяти суму від $i = 1$ до n, то одержимо:

$$\sum_{i=1}^{n}\left(y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n}\left(y_i^* - \overline{Y}\right)^2 + \sum_{i=1}^{n}\left(y_i - y_i^*\right)^2. \tag{7.14}$$

У цієї рівності немає члена $2\sum_{i=1}^{n}\left(y_i^* - \overline{Y}\right)\left(y_i - y_i^*\right)$, оскільки:

$$y_i - \overline{Y} = \alpha_1\left(x_i - \overline{X}\right);$$

$$y_i - y_i^* = y_i - \overline{Y} - \alpha_1\left(x_i - \overline{X}\right);$$

$$\sum_{i=1}^{n}\left(y_i - \overline{Y}\right)\left(y_i - y_i^*\right) = \alpha_1\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right) - \alpha_1^2\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2 = 0,$$

враховуючи (7.6).

Розглянемо склад виразу (7.14). Його ліва частина є сумою квадратів відхилень значень, що спостерігалися, стосовно загального середнього. Перший доданок правої частини є сумою квадратів відхилень оцінок цих значень, зроблених на основі обраної моделі регресії, від загального середнього. Її часто називають сумою квадратів стосовно регресії. Другий доданок правої частини є сумою квадратів відхилень значень, що спостерігалися, від їх оцінок, одержаних з використанням обраної моделі. Цей доданок називають сумою квадратів, що зумовлена регресією. Для того, щоб модель була придатною для прогнозування значень досліджуваної величини, необхідно, щоб він був малим порівняно із сумою квадратів стосовно регресії. У граничному випадку він має дорівнювати нулю.

Будемо вважати **дисперсію залишків** $\sigma^2_{\varepsilon_i}$ і, відповідно, **дисперсію відгуків** $\sigma^2_{Y_i}$ сталими. **Дисперсія** емпіричних точок **стосовно середнього** $\sigma^2_{Y_i}$ буде дорівнювати їх **дисперсії** σ^2_{YX} **стосовно лінії регресії** у випадку, коли постульована модель є істинною. У протилежному випадку $\sigma^2_{Y_i} > \sigma^2_{YX}$. Оцінкою величини σ^2_{YX} є відношення суми квадратів відхилень спостережень від середнього до кількості степенів вільності. Останню беруть рівною різниці між кількістю випробувань і кількістю констант, які визначаються незалежно одна від одної за їх результатами. У випадку, що розглядається, воно дорівнює $n - 2$, оскільки з емпіричних даних визначають два параметри прямої регресії. Тобто:

$$\sigma^2_{XY} = \frac{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}{n-2}. \qquad (7.15)$$

Для прикладу, що наведений у таблиці 7.1, можна одержати $\sigma^2_{YX} = 14{,}02$, $\sigma^2_{Y_i} = 14{,}17$, $\sigma^2_{\varepsilon_i} = 0{,}15$. Таким чином рівність $\sigma^2_{Y_i} = \sigma^2_{YX} + \sigma^2_{\varepsilon_i}$, яка випливає з (7.14), є виконаною.

Нехай p_i є кількістю повторних вимірювань величини $\overline{Y_i}$ при заданому значенні x_i. Тоді квадратична форма, яку мінімізують в методі найменших квадратів:

$$Q = \sum_{i=1}^{n} \left(\overline{Y_i} - z(x_i) \right)^2 p_i = \sum_{i=1}^{n} p_i \left[\overline{Y_i} - \alpha_0 - \alpha_1 \left(x_i - \overline{x} \right) \right]^2 . \qquad (7.16)$$

Розглядаючи її як функцію параметрів α_0, α_1, як і у попередньому випадку, одержуємо оцінки параметрів:

$$\begin{cases} \alpha_0^* = \sum_{i=1}^{n} p_i \overline{Y_i} / \sum_{i=1}^{n} p_i = \overline{Y} ; \\ \alpha_1^* = \sum_{i=1}^{n} p_i \overline{Y_i} \left(x_i - \overline{x} \right) / \sum_{i=1}^{n} p_i \left(x_i - \overline{x} \right)^2 . \end{cases} \qquad (7.17)$$

У припущенні, що умовний розподіл величини $\overline{Y_i}$ при заданому x_i є нормальним, оцінкою дисперсії буде величина:

$$\sigma_{Y_i}^{2\,*} = \frac{1}{n} \sum_{i=1}^{n} p_i \left(\overline{Y_i} - Y_i^* \right)^2 , \qquad (7.18)$$

де $Y_i^* = \alpha_0^* + \alpha_1^* (x_i - \overline{x_i})$.

Слід зазначити, що висновки, одержувані на основі мінімізації дисперсії похибки, є правильними тільки тоді, коли постульована модель коректна. В інших випадках вони можуть виявитися помилковими. Перед прийняттям рішення стосовно моделі треба перевірити гіпотезу, що лінійна модель $z = \alpha_0 + \alpha_1 \left(x - \overline{x} \right)$ задовільно описує емпіричні дані із заданою точністю при заданому рівні значущості η. Для цього визначають міру похибки емпіричних даних:

$$S_a^2 = \frac{1}{n-2} \sum_{i=1}^{n} p_i \left(\overline{Y_i} - Y_i^* \right)^2 . \qquad (7.19)$$

Ця величина є зміщеною оцінкою дисперсії $\sigma_{Y_i}^2$. Іноді її називають **дисперсією неадекватності**.

Незміщеною оцінкою цієї дисперсії є величина:

$$S_e^2 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{p_i} \left(Y_{ij} - \overline{Y_i} \right)^2}{\sum_{i=1}^{n} p_i - n}, \qquad (7.20)$$

де Y_{ij} – j-те одиничне вимірювання при $x = x_i$.

Критерієм адекватності моделі при заданій надійності $1-\eta$ є виконання нерівності:

$$S_a^2 / S_e^2 \le F_{1-\eta}, \tag{7.21}$$

де $F_{1-\eta}$ – відповідне значення функції розподілу Фішера, для кількостей степенів вільності $n_1 = n_2 = n - 1$.

Іноді вважають, що малі значення відношення (7.21) свідчать про адекватність обраної моделі. Але такий висновок може виявитися помилковим. Більш докладно це питання буде розглянуто нижче.

Довірчі інтервали для параметрів α_0, α_1 можна знайти за допомогою коефіцієнтів *t*-розподілу Стьюдента з кількістю степенів вільності $\sum_{i=1}^{n} p_i - 2$:

$$\begin{cases} \alpha^* - t_{1-\eta/2} S_{\alpha^*} \le \alpha \le \alpha^* - t_{1-\eta/2} S_{\alpha^*}\ ; \\ \mathrm{S}_{\alpha^*} = \sqrt{\dfrac{1}{n-2} \sum\limits_{i=1}^{n} p_i \left(\overline{\alpha} - \alpha^*\right)^2}\ . \end{cases} \tag{7.22}$$

Важливим практичним завданням є перевірка гіпотези про збіг двох рівнянь регресії:

$$z_1(x) = \alpha_{01} + \alpha_{11} x;$$

та

$$z_2(x) = \alpha_{02} + \alpha_{12} x.$$

Воно передбачає перевірку трьох простих гіпотез. Спочатку перевіряють гіпотезу про рівність дисперсій неадекватності моделей:

$$\mathrm{H}_0^{(1)} : \sigma_{a1}^2 = \sigma_{a2}^2. \tag{7.23}.$$

Для цього використовують дисперсійний критерій Фішера.

Якщо різниця дисперсій неадекватності є незначущою, то переходять до перевірки гіпотези про рівність кутових коефіцієнтів моделей:

$$\mathrm{H}_0^{(2)} : \alpha_{11} = \alpha_{12}. \tag{7.24}$$

$$t_a = \frac{a_{11} - a_{12}}{\sigma \sqrt{\dfrac{1}{\sum\limits_{i=1}^{k_1} n_{1i} \left(x_i^{(1)} - \overline{x}^{(1)}\right)^2} + \dfrac{1}{\sum\limits_{i=1}^{k_2} n_{2i} \left(x_i^{(2)} - \overline{x}^{(2)}\right)^2}}}, \tag{7.25}$$

При цьому виходять з того, що величина підпорядковується *t*-розподілу Стьюдента з $N_1 + N_2 - 4$ степенями вільності. У (7.25) n_{1i}, n_{2i} – кількість вимірювань для першої та другою моделі в *i*-ї точці; k_1, k_2 – кількість точок для кожної з моделей;

$$\sigma = \frac{(N_1 - 2)\sigma_{a1}^2 + (N_2 - 2)\sigma_{a2}^2}{N_1 + N_2 - 4}; \tag{7.26}$$

$$N_1 = \sum_{i=1}^{k_1} n_{1i}\ ;\ N_2 = \sum_{i=1}^{k_2} n_{2i}\ .$$

Справедливість гіпотези $H_0^{(2)}$ означає, що порівнювані лінії регресії паралельні одна одній. У цьому випадку можна отримати уточнену оцінку коефіцієнта нахилу прямою регресії:

$$\overline{a_1} = \frac{\sum_{i=1}^{k_1} n_{1i}\left(x_i^{(1)} - \overline{x}^{(1)}\right)\left(y_i^{(1)} - \overline{y}^{(1)}\right) + \sum_{i=1}^{k_2} n_{2i}\left(x_i^{(2)} - \overline{x}^{(2)}\right)\left(y_i^{(2)} - \overline{y}^{(2)}\right)}{\sum_{i=1}^{k_1} n_{1i}\left(x_i^{(1)} - \overline{x}^{(1)}\right)^2 + \sum_{i=1}^{k_2} n_{2i}\left(x_i^{(2)} - \overline{x}^{(2)}\right)^2}. \tag{7.27}$$

Після цього необхідно перевірити останню гіпотезу про рівність вільних членів моделей:

$$H_0^{(3)} : a_{01} = a_{02}\,. \tag{7.28}$$

З цією метою використовують те, що величина

$$u = \frac{\widehat{a_1} - \overline{a_1}}{\sigma\left\{\widehat{a_1} - \overline{a_1}\right\}}, \tag{7.29}$$

де

$$\widehat{a_1} = \frac{\overline{y}^{(1)} - \overline{y}^{(2)}}{\overline{x}^{(1)} - \overline{x}^{(2)}},$$

$$\sigma\left\{\widehat{a_1} - \overline{a_1}\right\} =$$

$$= \sigma^2\left[\frac{1}{\left(\overline{x}^{(2)} - \overline{x}^{(1)}\right)^2}\left(\frac{1}{N_1} + \frac{1}{N_2}\right) + \frac{1}{\sum_{i=1}^{k_1} n_{1i}\left(x_i^{(1)} - \overline{x}^{(1)}\right)^2 + \sum_{i=1}^{k_2} n_{2i}\left(x_i^{(2)} - \overline{x}^{(2)}\right)^2}\right]$$

підпорядковується *t*-розподілу Стьюдента з $N_1 + N_2 - 4$ степенями вільності.

7.3. Поліноміальні моделі

У багатьох випадках емпіричні залежності можна описати поліноміальними моделями вигляду:

$$z = \sum_{i=1}^{q} \alpha_i x^i\ . \tag{7.30}$$

Оцінки параметрів таких моделей отримують шляхом розв'язування нормальних рівнянь виду:

$$\begin{pmatrix} n & \sum x_i & \sum x_i^2 & \dots & \sum x_i^q \\ \sum x_i & \sum x_i^2 & \sum x_i^3 & \dots & \sum x_i^{q+1} \\ \dots & \dots & \dots & \dots & \dots \\ \sum x_i^q & \sum x_i^{q+1} & \sum x_i^{q+2} & \dots & x_i^{2q} \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \dots \\ \alpha_q \end{pmatrix} = \begin{pmatrix} \sum Y_i \\ \sum Y_i x_i \\ \dots \\ \sum Y_i x_i^q \end{pmatrix}. \quad (7.31)$$

Зазвичай стовпці, що утворюють матрицю *X*, не є ортогональними. У зв'язку з цим у разі необхідності збільшення степеня полінома необхідно перераховувати оцінки всіх його коефіцієнтів. Тому для поліномів високих степенів більш раціональним методом побудови регресійної моделі є заміна вихідного рівняння (7.30) іншим:

$$z = \sum_{i=1}^{q} \alpha_i' \zeta_i , \quad (7.32)$$

де $\zeta_i = \zeta_i(x)$ є поліномами *i*-го степеня за *x*, які задовольняють умови ортогональності:

$$\begin{cases} \sum_{j=1}^{n} \zeta_{ij} = 0, \;\; i = 1, 2, ..., q ; \\ \sum_{j=1}^{n} \zeta_{ij} \zeta_{i'j} = 0, \;\; i \neq i' \end{cases}, \quad (7.33)$$

ζ_{ij} є *i*-м поліномом для точки x_j.

Квадратична форма, що мінімізується у методі найменших квадратів має вигляд:

$$Q = \sum_{j=1}^{n} \left(Y_j - \alpha'_0 - \alpha'_1 \zeta_{1j} - ... - \alpha'_q \zeta_{qj} \right)^2 . \quad (7.34)$$

Значення, що відповідають мінімуму (7.34), можна знайти, розв'язавши систему:

$$\begin{cases} \dfrac{\partial Q}{\partial \alpha'_0} = -2 \sum_{j=1}^{n} \left(Y_j - \alpha'^*_0 - \alpha'^*_1 \zeta_{1j} - ... - \alpha'^*_q \zeta_{qj} \right) = 0 ; \\ \dfrac{\partial Q}{\partial \alpha'_i} = -2 \left(\sum_{j=1}^{n} Y_j \zeta_{ij} - \alpha'^*_0 \sum_{j=1}^{n} \zeta_{ij} - \alpha'^*_1 \sum_{j=1}^{n} \zeta_{1j} \zeta_{ij} - ... - \alpha'^*_i \sum_{j=1}^{n} \zeta_{ij}^2 - ... - \right. \\ \qquad \left. - \alpha'^*_q \sum_{j=1}^{n} \zeta_{qj} \zeta_{ij} \right) = 0, \;\; i = 1, 2, ..., q \end{cases} \quad (7.35)$$

Звідси, використовуючи умови ортогональності (7.33), одержуємо:

$$\alpha'^{*}_{0} = \overline{Y};\ \ \alpha'^{*}_{i} = \frac{\sum_{j=1}^{n} Y_j \zeta_{ij}}{\sum_{j=1}^{n} \zeta_{ij}^2}\ . \tag{7.36}$$

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

$$\begin{cases} \zeta_{0j} = 1; \\ \zeta_{1j} = v_j - \overline{v_j}; \\ \zeta_{2j} = \zeta_{1j}^2 - \left(n^2 - 1\right)/12\ , \end{cases} \tag{7.37}$$

де $v_j = \left(x_{j+1} - x_1\right)/\omega$.

Поліноми вищих степенів одержують за рекурентною формулою:

$$\zeta_{r+1,j} = \zeta_{1j}\zeta_{rj} - \frac{r^2\left(n^2 - r^2\right)}{4\left(4r^2 - 1\right)}\zeta_{r-1,j}\ . \tag{7.38}$$

Розглянемо такий приклад. У табл. 7.3 наведено результати вимірювання деякої величини.

Таблиця 7.3

Емпіричні дані для побудови поліноміальної регресійної моделі

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

Побудуємо модель досліджуваної залежності у вигляді полінома 5-го степеня:

$$z = \alpha'_0 + \alpha'_1\zeta_1 + \alpha'_2\zeta_2 + \alpha'_3\zeta_3 + \alpha'_4\zeta_4 + \alpha'_5\zeta_5,$$

де $\zeta_i = \sum_{t=1}^{i} \beta_{it} x^t$ – поліноми i-го степеня, які задовольняють умови ортогональності.

Дані, необхідні для розрахунку значень ζ_{ij} і коефіцієнтів α_i, наведено в табл. 7.4.

Легко перевірити, що для одержаних даних виконуються умови ортогональності, тобто суми значень ζ_{ij} $(i \neq 0)$ у кожному рядку й суми за і добутків вигляду $\zeta_{ij}\zeta_{kj}$ $(i \neq k)$ дорівнюють нулю.

За даними таблиці розраховуємо коефіцієнти α'_i (табл. 7.5).

Таблиця 7.4

Результати розрахунку допоміжних параметрів для поліноміальної регресійної моделі та оцінок значень досліджуваної величини

v_j	1	2	3	4	5	6	7	8	9	10
ζ_{0j}	1	1	1	1	1	1	1	1	1	1
ζ_{1j}	−4,5	−3,5	−2,5	−1,5	−0,5	0,5	1,5	2,5	3,5	4,5
ζ_{2j}	12	4	−2	−6	−8	−8	−6	−2	4	12
ζ_{3j}	−25,2	8,4	21	18,6	7,2	−7,2	−18,6	−21	−8,4	25,2
ζ_{4j}	43,2	−52,8	−40,8	7,2	43,2	43,2	7,2	−40,8	−52,8	43,2
ζ_{5j}	−60	140	−10	−110	−60	60	110	10	−140	60
$y_j\zeta_{0j}$	23	29	41	60	79	88	83	61	33	27
$y_j\zeta_{1j}$	−103,5	−101,5	−102,5	−90	−39,5	44	124,5	152,5	115,5	121,5
$y_j\zeta_{2j}$	276	116	−82	−360	−632	−704	−498	−122	132	324
$y_j\zeta_{3j}$	−579,6	243,6	861	1116	568,8	−633,6	−1544	−1281	−277,2	680,4
$y_j\zeta_{4j}$	993,6	−1531	−1673	432	3413	3802	597,6	−2489	−1742	1166
$y_j\zeta_{5j}$	−1380	4060	−410	−6600	−4740	5280	9130	610	−4620	1620
Y_i^*	24,09	30,1	42,23	60,96	79,8	89,74	83,86	61,81	34,38	28,03

Таблиця 7.5

Результати розрахунку параметрів побудованої регресійної моделі

i	0	1	2	3	4	5
$\sum_{j=1}^{n}\zeta_{ij}^2$	10	82,5	528	3089	16474	78000
$\sum_{j=1}^{n}y_j\zeta_{ij}$	524	121	−1550	−845,4	2969	2950
α'_i	52,4	1,467	−2,936	−0,2737	0,1802	0,03782
$\sigma^2_{a'_j}$	0,038	0,0036	0,00072	0,00012	0,000023	0,000005
$\sigma_{a'_j}$	0,20	0,068	0,027	0,011	0,0048	0,0022

Оцінками дисперсії цих коефіцієнтів є величини $\sigma^2_{\alpha'_i} = \dfrac{\sum_{j=1}^{n}\left(y_i - y_i^*\right)^2}{\left(n-q-1\right)\sum_{j=1}^{n}\zeta_{ij}^2}$.

Звідси за формулами (7.32, 7.38) одержуємо оцінки Y^*_i значень досліджуваної величини Y у точках $x = x_j$, які наведені у табл. 7.3 і є досить близькими до її емпіричних значень.

На рис. 7.3 наведено графіки вихідних даних та побудованої регресійної моделі.

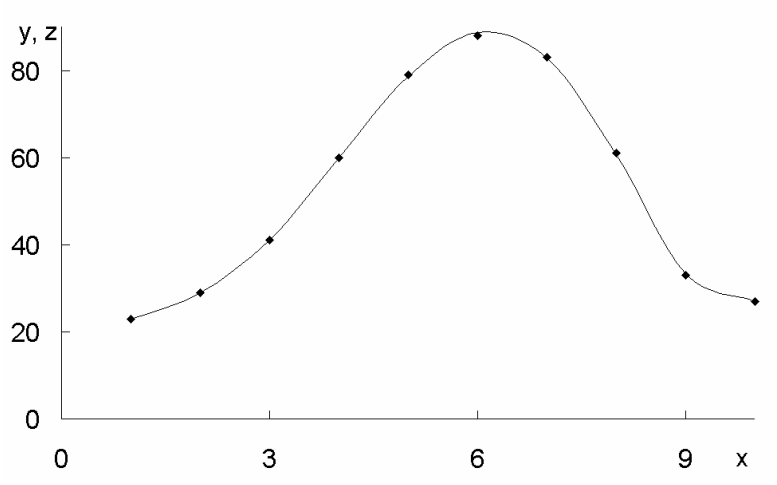


Рис. 7.3. Графіки вихідних даних (крапки) й побудованої поліноміальної регресійної моделі (лінія)

7.4. Однофакторні моделі інших типів

Апроксимацію емпіричних залежностей тригонометричними багаточленами називають **гармонічним аналізом**. У цьому випадку модель має вигляд:

$$z(x) = \alpha_0 + \sum_{k=1}^{r} \alpha_k \cos\frac{2\pi kx}{T} + \sum_{k=1}^{r} \beta_k \sin\frac{2\pi kx}{T}, \qquad (7.39)$$

де T – період спостереження апроксимованої залежності;
r – кількість гармонік ($r < n/2$);
n – кількість частин, на які розділений період T.

Її параметри визначають за формулами:

$$\alpha_0 = \frac{1}{n}\sum_{k=0}^{n} y_k;$$

$$\alpha_m = \frac{2}{n}\sum_{k=0}^{n} y_k \cos\frac{2\pi km}{n}, \quad m = 1, 2, ..., r; \qquad (7.40)$$

$$\beta_m = \frac{2}{n}\sum_{k=0}^{n} y_k \sin\frac{2\pi km}{n}, \quad m = 1, 2, ..., r,$$

де y_k – значення апроксимованої функції у точках $x_k = \frac{kT}{n}$.

На рис. 7.4 наведено графіки емпіричних даних і відповідних регресійних моделей, побудованих у вигляді тригонометричних рядів, для m рівних 2, 3, 4 й 5. Видно, що зі збільшенням кількості членів тригонометричного ряду різниця між моделлю та емпіричними точками зменшується. Добре видно також, що найбільшу похибку модель дає поблизу меж відрізка, на якому визначені емпіричні дані.

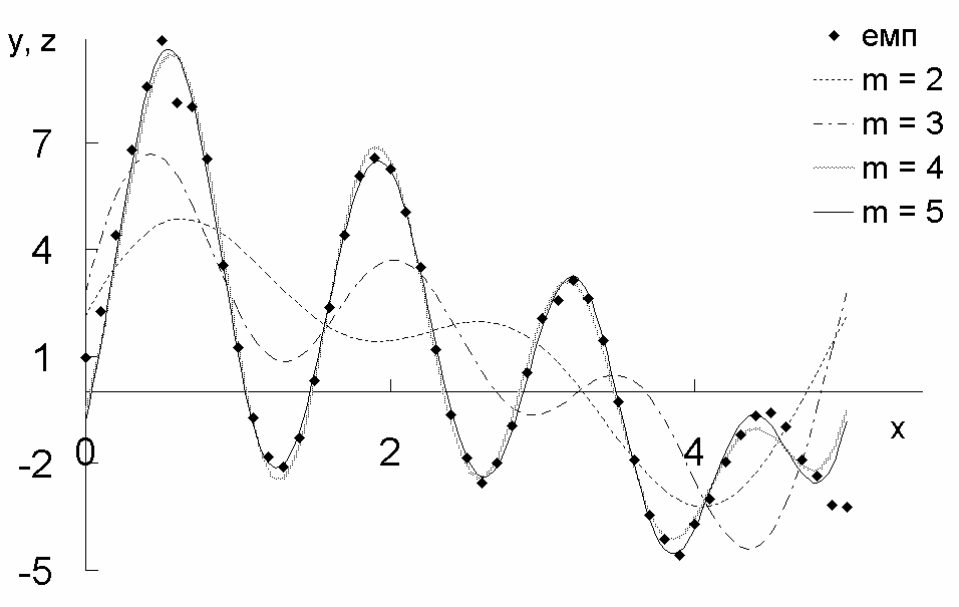


Рис. 7.4. Графіки вихідних даних і побудованої тригонометричної регресійної моделі

Моделі, що мають вигляд **модифікованої показникової функції**, записують як $z(x) = a + bc^x$ або $z(x) = a + bc^{-x}$. Якщо емпіричні точки є рівновіддаленими одна від одної, тобто крок емпіричної залежності за x є сталим, то можна перейти від незалежної змінної x до порядкового номера відповідної точки i та розрахувати параметри моделі першого типу за формулами:

$$c = \frac{(N-1)\sum_{i=1}^{N-1} y_i y_{i+1} - \sum_{i=1}^{N-1} y_i \sum_{i=1}^{N-1} y_{i+1}}{(N-1)\sum_{i=1}^{N-1} y_i^2 - \left(\sum_{i=1}^{N-1} y_i\right)^2};$$

$$b = \frac{N\sum_{i=1}^{N} c^i y_i - \sum_{i=1}^{N} c^i \sum_{i=1}^{N} y_i}{N\sum_{i=1}^{N} c^{2i} - \left(\sum_{i=1}^{N} c^i\right)^2};\ a = \frac{\sum_{i=1}^{N} y_i - b\sum_{i=1}^{N} c^i}{N}. \qquad (7.41)$$

У моделі другого типу параметри необхідно визначати за однією з ітераційних процедур мінімізації функціонала, що характеризує суму квадратів відхилень рівнів емпіричних точок від моделі. Зокрема у цьому випадку можна використовувати метод деформівного багатогранника.

Крива Гомперця описується рівняннями $\hat{y}_t = ab^{c^t}$ або $\hat{y}_t = ab^{c^{-t}}$, які логарифмуванням зводяться до узагальненої показникової функції першого або другого типу, відповідно.

Логістична функція $\hat{y}_t = \dfrac{1}{a + bc^t}$ або $\hat{y}_t = \dfrac{1}{a + bc^{-t}}$ зводиться до модифікованої показникової перетворенням $y^* = 1/\hat{y}_t = a + bc^{\pm t}$.

Модифіковану показникову функцію використовують як модель у випадках, коли досліджуваній залежності властиве насичення, тобто при збільшенні значень незалежної змінної відгук поступово наближається до певного граничного значення, а його прирости наближуються до нуля. У таких випадках існує певний обмежувальний фактор, вплив якого збільшується із зростанням досягнутого рівня. Значення рівня насичення, як правило, можна задати, виходячи з наявних даних про об'єкт дослідження. У такому разі інші параметри моделі можна визначити методом найменших квадратів після її лінеаризації.

Якщо вплив обмежувального фактора виявляється лише після досягнення певного рівня розвитку процесу, слід використовувати **моделі *S*-подібного зростання**, до яких належать крива Гомперця і логістична функція. Вони описують процеси, в яких темп зростання поступово збільшується на початкових стадіях і поступово зменшується в кінці. При цьому слід ураховувати, що крива Гомперця є асиметричною, а логістична крива симетрична стосовно точки перегину. Процес побудови й дослідження логістичної моделі називають логістичним аналізом. Логістичну криву часто називають законом зростання, оскільки вона описує залежність кількості популяції або її біомаси від часу.

При побудові регресійних моделей загального вигляду часто використовують методи, які базуються на мінімізації функціоналів виду (7.1). Для функціонала (7.1а) це зумовлює необхідність розв'язання системи:

$$\left\{ \frac{\partial}{\partial \alpha_k} \sum_{i=1}^{n} \left[z(x_i) - y_i \right]^2 = 0. \right. \tag{7.42}$$

Якщо вона є системою лінійних рівнянь, застосовують звичайні алгоритми розв'язання таких систем – Гауса, простих ітерацій, Зейделя тощо. В окремих випадках, зокрема при логістичному аналізі, розробляють спеціальні алгоритми. Якщо ж система (7.42) є нелінійною, використовують алгоритми нелінійної оптимізації: Ньютона – Рафсона, квазіньютонівські, спряжених градієнтів тощо.

7.5. Лінійні багатофакторні моделі

Як було зазначено вище, лінійну як за параметрами, так і за незалежними змінними регресійну модель можна записати у вигляді:

$$Y = \alpha_0 + \sum_{i=1}^{p} \alpha_j x_j + \varepsilon = X\alpha + \varepsilon, \tag{7.43}$$

де Y – вектор-стовпчик відгуків, який має розмірність n ($n > p$);

$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$ – матриця значень p незалежних змінних при n вимірюваннях;

α – вектор-стовпчик невідомих параметрів моделі, що має розмірність $p + 1$;

ε – вектор-стовпчик похибок моделі, який має розмірність n.

Оцінки параметрів моделі у методі найменших квадратів отримують мінімізацією скалярного добутку:

$$Q = (\boldsymbol{Y} - \boldsymbol{X}\alpha)^{T} (\boldsymbol{Y} - \boldsymbol{X}\alpha), \tag{7.44}$$

де символ "Т" позначає транспонування.

Система (7.42) у цьому випадку набуває вигляду:

$$-2\boldsymbol{X}^{T} (\boldsymbol{Y} - \boldsymbol{X}\alpha) = 0 .$$

Звідси маємо:

$$\boldsymbol{X}^{T}\boldsymbol{Y} = \boldsymbol{X}^{T}\boldsymbol{X}\alpha.$$

Помножуючи обидві частини цієї рівності ліворуч на матрицю $(\boldsymbol{X}^{T}\boldsymbol{X})^{-1}$, одержимо:

$$(\boldsymbol{X}^{T}\boldsymbol{X})^{-1} (\boldsymbol{X}^{T}\boldsymbol{Y}) = (\boldsymbol{X}^{T}\boldsymbol{X})^{-1} (\boldsymbol{X}^{T}\boldsymbol{X})\,\alpha.$$

Добуток $(\boldsymbol{X}^{T}\boldsymbol{X})^{-1} (\boldsymbol{X}^{T}\boldsymbol{X}) = \boldsymbol{E}$, де $\boldsymbol{E}$ – одинична матриця розмірності $p + 1$. Тому остаточно маємо:

$$\alpha = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} . \tag{7.45}$$

Для лінійної моделі (7.45) є незміщеною оцінкою з найменшою дисперсією вектора α.

Коваріаційною матрицею вектора α є:

$$\Sigma = \sigma^{2} (X^{T}X)^{-1}, \tag{7.46}$$

де σ^{2} – дисперсія похибки.

Елементами головної діагоналі коваріаційної матриці є дисперсії компонентів вектора α, а позадіагональні компоненти є значеннями відповідних коефіцієнтів коваріації.

Для перевірки значущості регресії використовують *F*-критерій Фішера, розрахункове значення якого обчислюють за формулою:

$$F = \frac{Q_R / (p+1)}{Q / (n-p-1)}, \tag{7.47}$$

де $Q_R = (X\alpha)^T (X\alpha)$ – сума квадратів відхилень, зумовлена регресією;

Q – сума квадратів відхилень спостережень від регресії, що визначається за формулою (7.44).

За умови виконання нульової гіпотези $H_0: \alpha = 0$ $F < F_{кр}$, де $F_{кр}$ – критичне значення статистики Фішера для заданого рівня значущості й кількостей степенів вільності $(p+1)$ та $(n-p-1)$.

Значущість окремих коефіцієнтів регресії перевіряють за допомогою критерію:

$$t_j = \frac{\alpha_j}{\hat{s}\sqrt{(X^T X)^{-1}_{jj}}}, \tag{7.48}$$

де $\hat{s} = \sqrt{\frac{1}{n-p-1}Q}$ – незміщена оцінка стандартного відхилення залишків моделі. За умови виконання нульової гіпотези $H_0: \alpha_j = 0$ статистика критерію підпорядковується *t*-розподілу Стьюдента з кількістю степенів вільності $n - p - 1$.

Інтервальною оцінкою для коефіцієнта α_j є:

$$\alpha_j \in \left[\widehat{\alpha_j} - t\hat{s}\sqrt{(X^T X)^{-1}_{jj}}; \widehat{\alpha_j} + t\hat{s}\sqrt{(X^T X)^{-1}_{jj}} \right]. \tag{7.49}$$

Одним з ускладнень, що можуть виникати при побудові регресійних моделей за наявності декількох предикторів, є можлива нерівноточність спостережень. У найпростішому випадку це можна врахувати за допомогою коваріаційної матриці такого вигляду:

$$\Omega = \begin{pmatrix} \sigma_1^2 & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & \sigma_n^2 \end{pmatrix}, \tag{7.50}$$

де σ_i^2 – дисперсія *i*-го спостереження. У цьому випадку формула (7.45) набуває вигляду:

$$\alpha = (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} Y. \tag{7.51}$$

Такий метод побудови регресійних моделей називають **зваженим методом найменших квадратів**. Вперше його було запропоновано К. Гаусом в 1809 р.

У цьому випадку коваріаційна матриця визначається за формулою:

$$\Sigma = \sigma^2 \left(X^T \Omega^{-1} X \right)^{-1}. \tag{7.52}$$

Для перевірки значущості регресії використовують *F*-критерій Фішера, розрахункове значення якого обчислюють за формулою (7.47), де $Q_R = \left(X\alpha \right)^{\mathrm{T}} \Omega^{-1} \left(X\alpha \right)$, $Q = \left(Y - X\alpha \right)^{\mathrm{T}} \Omega^{-1} \left(Y - X\alpha \right)$.

Значущість окремих коефіцієнтів регресії перевіряють за допомогою критерію:

$$t_j = \frac{\alpha_j}{\hat{s}^2_{\alpha_j}} = \frac{\alpha_j}{\hat{s}\sqrt{\left(X^T \Omega^{-1} X \right)^{-1}_{jj}}}, \tag{7.53}$$

де $\hat{s}^2_{\alpha_j}$ – *j*-й діагональний елемент коваріаційної матриці.

Інтервальною оцінкою для коефіцієнта α_j є:

$$\alpha_j \in \left[\widehat{\alpha_j} - t\hat{s}_{\alpha_j}; \widehat{\alpha_j} + t\hat{s}_{\alpha_j} \right]. \tag{7.54}$$

Іншим можливим ускладненням є наявність взаємозв'язку між предикторами. Припустимо, що існує лінійна залежність між компонентами вектора ***X***:

$$\nu_1 x_1 + \nu_2 x_2 + ... + \nu_m x_m = 0. \tag{7.55}$$

Що ближчою є ліва частина (7.55) до нульового вектора, то сильнішою є мультиколінеарність. Граничний випадок точного виконання рівності (7.55) називають **строгою мультиколінеарністю**. У цьому випадку визначник $|X^{\mathrm{T}}X| = 0$ й використовувати формулу (7.45) неможливо. Випадок $|X^{\mathrm{T}}X| \approx 0$ називають **мультиколінеарністю**.

Мультиколінеарність зумовлює нестійкість обчислювальної процедури через високу похибку обчислення оберненої матриці, тобто додавання нових даних призводить до істотної зміни оцінок параметрів. Коефіцієнти регресійної моделі у цьому випадку виявляються сильно корельованими один з одним, а довірчий рівень і дисперсія їх оцінок – підвищеними. Внаслідок цього інтерпретація результатів стає неможливою, а значення окремих коефіцієнтів – статистично незначущими.

Наявність мультиколінеарності можна перевірити шляхом дослідження кореляційної матриці *R* нормованих і центрованих вихідних даних [16]. У цьому випадку: $|R| << 1$ і для окремих елементів матриці $|r_{ij}| \geq 0{,}9$ при $i \neq j$.

Свідченням мультиколінеарності є також **погана зумовленість** матриці ($X^{\mathrm{T}}X$), яку визначають як відношення максимального власного числа матриці до мінімального. Якщо $\frac{\lambda_{\max}}{\lambda_{\min}} \geq 10^5$, то це є свідченням сильної мультиколінеарності вихідних даних.

Існує декілька способів корегування мультиколінеарності. Найпростішим є стандартизація й центрування даних.

Інший підхід передбачає залучення додаткової інформації, зокрема, збільшення обсягу вибірки, за якою оцінюють значення параметрів моделі. Проте, в більшості випадків це неможливо, особливо при дослідженні соціально-економічних систем.

Ще один підхід ґрунтується на зменшенні розмірності простору факторів. Найпростіше це можна зробити шляхом виявлення найсильніше корельованих змінних і об'єднання їх в один фактор. Але це дає позитивні результати лише за умови, що таке об'єднання є теоретично обґрунтованим.

В окремих випадках для усунення мультиколінеарності відкидають одну чи декілька сильно пов'язаних змінних, що призводить до появи нових похибок. У такому випадку необхідно визначити, яка з похибок є більш істотною. Для цього можна почергово виключати пов'язані змінні й порівнювати одержувані результати. Інший варіант цього підходу передбачає послідовне додавання нових факторів і перевірку того, покращує він модель чи ні.

Одним з методів відбору найбільш істотних факторів є процедура **покрокової регресії**. Вона може бути організована як у напрямі зменшення кількості факторів, що враховують у моделі, так і в зворотному. У першому випадку спочатку будують модель, що враховує всі фактори і перевіряють їх значущість. Для цього перевіряють нульові гіпотези $\alpha_j = 0$, використовуючи статистику:

$$F_j = \frac{\widetilde{\alpha}_j^{\,2}}{D\left[\widetilde{\alpha}_j\right]}, \tag{7.56}$$

де $D\left[\widetilde{\alpha}_i\right] = \dfrac{\sum_{j=1}^{n}\left(y_j - \widetilde{y}_j\right)^2}{n-p}\left(X^{-1}X\right)_{ii}^{-1}$.

За умови справедливості нульової гіпотези вона має розподіл Фішера з кількістю степенів вільності 1 та $(n - p)$. Найменше значення F_j порівнюють з граничною величиною F_0. Якщо $F_j < F_0$, то j-й фактор виключають із моделі й будують нову модель з меншою кількістю факторів. В іншому випадку модель залишають без змін.

У випадку, коли процедуру організовують у напрямі збільшення кількості факторів, на першому етапі визначають коефіцієнти кореляції між кожним фактором і відгуком, а потім будують однофакторну модель регресії, яка враховує лише фактор x_ℓ, що має найбільший (за модулем) коефіцієнтом кореляції.

Якщо перевірка моделі встановлює значущість обраного фактора, то наступним кроком є визначення F-статистики (7.56) для факторів, що залишилися. До нової моделі включають фактор x_ℓ, а також фактор x_m, для якого значення цієї статистики є найбільшим. Потім визначають значущість отриманої моделі й розраховують F-статистики F_ℓ та F_m. Меншу з них порівнюють з граничним значенням F_0 і за виконання умови $F_j < F_0$ відповідний фактор виключають із моделі.

Таку процедуру здійснюють на кожному наступному кроці. Побудову моделі закінчують, коли найбільше значення F-статистики для факторів, що не включені до неї, не перевищує граничного значення F_0. Процес також закінчують, якщо до моделі включено всі досліджувані фактори або якщо перевищено граничну кількість кроків.

Можна також перетворити множину вихідних факторів до меншої кількості нових взаємноортогональних факторів. У цьому випадку використовують метод головних компонент та інші процедури факторного аналізу.

Розглянуті вище методи усунення мультиколінеарності розраховані на отримання незміщених оцінок параметрів регресійних моделей. Альтернативою їм є **методи зміщеного оцінювання**, зокрема гребеневі оцінки, редуковані оцінки, оцінки Марквардта, оцінки Хоккінса тощо. Якщо коваріаційні матриці є погано зумовленими, при обчисленні оцінок коефіцієнтів регресії застосовують також метод регуляризації О.М. Тихонова. Завданням оцінювання при їх застосуванні є одержання значень параметрів регресії, які були б стійкими навіть за умови сильної спряженості незалежних змінних.

У граничному випадку, коли матриця XX є одиничною, оцінки, одержувані за формулою (7.55), є незміщеними і мають мінімальну дисперсію. Із збільшенням власних чисел матриці XX відстань між оцінками $\hat{\beta}$ та істинними β збільшується. Крім того, стає можливою зміна напряму впливу вхідних змінних на вихідні.

На сьогодні розроблено понад 80 алгоритмів зміщеного оцінювання параметрів лінійних регресійних моделей. Їх поділяють на такі групи: методи звичайних гребеневих оцінок, методи узагальнених гребеневих оцінок, методи оцінок дробового рангу, методи стиснутих оцінок. Всі вони є лінійними перетвореннями оцінок, одержуваних МНК.

У **методах гребеневого аналізу** ставиться завдання отримання оцінок з мінімальною дисперсією. Оцінка гребеневої регресії α_k є лінійним перетворенням оцінки $\boldsymbol{\alpha}$, отриманої МНК, і залежить від параметра k і матриці вихідних даних $\boldsymbol{X}$. Її можна записати у вигляді:

$$\alpha_k = \left(X^T X - kE \right)^{-1} X^T Y . \qquad (7.57)$$

Ефективність методів гребеневих оцінок залежить від статистичних характеристик вихідної інформації й оптимального вибору параметра k. Метод гребеневого аналізу, як і інші методи зміщеного оцінювання параметрів регресійних моделей, за певних умов можуть бути обґрунтовані теоретично, але в більшості практичних ситуацій перевірити виконання цих умов неможливо. Тому застосування цих методів потребує певної обережності.

Параметри нелінійних регресійних моделей за наявності декількох незалежних змінних зазвичай оцінюють з використанням чисельних методів нелінійної мінімізації функціонала (7.1а). При цьому істотне значення має вибір початкового наближення. У багатьох випадках це завдання може бути істотно спрощено з використанням методів планування експерименту, які дають змогу визначити оптимальні для подальшого аналізу плани, тобто точки простору незалежних ознак, у яких потрібно здійснити вимірювання значень відгуків.

7.6. Інші типи багатофакторних моделей

Для статичних систем багатофакторні регресійні моделі зазвичай задають у вигляді полінома:

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{\substack{i=1 \\ i \neq j}}^{k} \beta_{ij} x_i x_j + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \ldots . \qquad (7.58)$$

При цьому найчастіше обмежуються поліномами другого степеня.

У деяких випадках, зокрема при проектуванні теплотехнічних і гідродинамічних систем, а також при дослідженні економічних систем модель задають у вигляді степеневої функції:

$$y = \gamma x_1^{\beta_1} x_2^{\beta_2} \ldots . \qquad (7.59)$$

У такому випадку модель можна лінеаризувати логарифмуванням:

$$\ln y = \ln \gamma + \beta_1 \ln x_1 + \beta_2 \ln x_2 + \ldots$$

та переходом до нових змінних: $y' = \ln y$, $x'_1 = \ln x_1$, $x'_2 = \ln x_2$ … Після цього аналіз моделі здійснюють за допомогою методів, описаних у попередньому підрозділі.

Відомим прикладом моделі такого типу є відома в економічній теорії **функція Кобба – Дугласа**:

$$Q = Q_0 \left(\frac{L}{L_0} \right)^a \left(\frac{K}{K_0} \right)^b , \tag{7.60}$$

де Q – обсяг виробництва;

L – трудові ресурси;

K – капітал.

Індекс "0" відповідає певним фіксованим значенням цих параметрів. Логарифмуванням ця модель приводиться до лінійної:

$$\ln Q = \ln Q_0 + a \ln \left(\frac{L}{L_0} \right) + b \ln \left(\frac{K}{K_0} \right). \tag{7.61}$$

7.7. Перевірка адекватності регресійних моделей

Основні методи перевірки адекватності регресійних моделей ґрунтуються на таких трьох властивостях їх залишків.

По-перше, для адекватної моделі дисперсія залишків має бути близькою до дисперсії емпіричних точок. При цьому припускають, що дисперсії всіх емпіричних точок є однаковими. У випадку, коли для кожної точки здійснюють декілька вимірювань значення відгуку, останнє припущення можна перевірити за допомогою критеріїв Кокрена або Бартлетта.

Причиною неадекватності при невиконанні цієї властивості є використання надмірно спрощених або ускладнених регресійних моделей. Відомо, наприклад, що за наявності n емпіричних точок можна побудувати поліноміальну модель $n - 1$ порядку, яка пройде строго через всі ці точки. Але використовувати такий поліном як регресійну модель за наявності похибок емпіричних даних, очевидно, недоцільно. З іншого боку, якщо степінь полінома буде надто малим, то він не відтворюватиме істотних рис досліджуваної залежності, тому існує певне оптимальне значення степеня такого полінома. Як критерій виконання цієї властивості часто застосовують такий критерій:

$$\frac{S}{\Delta^2} \le F , \; \frac{\Delta^2}{S} \le F , \tag{7.62}$$

де S – значення цільового функціонала (7.1а);

Δ^2 – сума квадратів похибок емпіричних даних по всіх точках;

F – критичне значення критерію Фішера для вибраного рівня значущості й кількостей степенів вільності, що дорівнюють $n - 1$.

Невиконання першої умови свідчить про надмірну спрощеність моделі, зокрема про необхідність збільшення порядку поліноміальної моделі. Невиконання другої умови є свідченням того, що модель треба спростити,

наприклад зменшити порядок полінома. У деяких випадках друга умова може не виконуватися навіть для однофакторних лінійних моделей. Найчастіше це може бути наслідком свідомого підганяння емпіричних даних під заздалегідь задану модель. Це часто роблять у навчальних задачах, але на практиці такий результат свідчить про навмисне викривлення первинних даних. Іншою причиною може бути неправильна (завищена) оцінка похибки емпіричних даних. Це може бути пов'язано, зокрема, з нехтуванням зміною дисперсії емпіричних даних при їх попередній обробці.

Інша властивість залишків, яку перевіряють при визначенні адекватності моделі, полягає в тому, що вони мають підпорядковуватися нормальному закону розподілу з нульовим математичним сподіванням і однаковими дисперсіями. Перевірку цих властивостей можна здійснити за допомогою критеріїв, що описані у розділі 2.

На рис. 7.5 показано деякі типові випадки порушення вказаних властивостей, які можуть бути виявлені при візуальному аналізі ряду залишків.

У випадку а) на графіку є так звані викиди – точки, що аномально сильно відхиляються від середнього значення. У випадку б) дисперсія залишків помітно зменшується при зміщенні вправо. У випадку в) ряд залишків не є випадковим, що свідчить про наявність неврахованих істотних закономірностей у моделі.

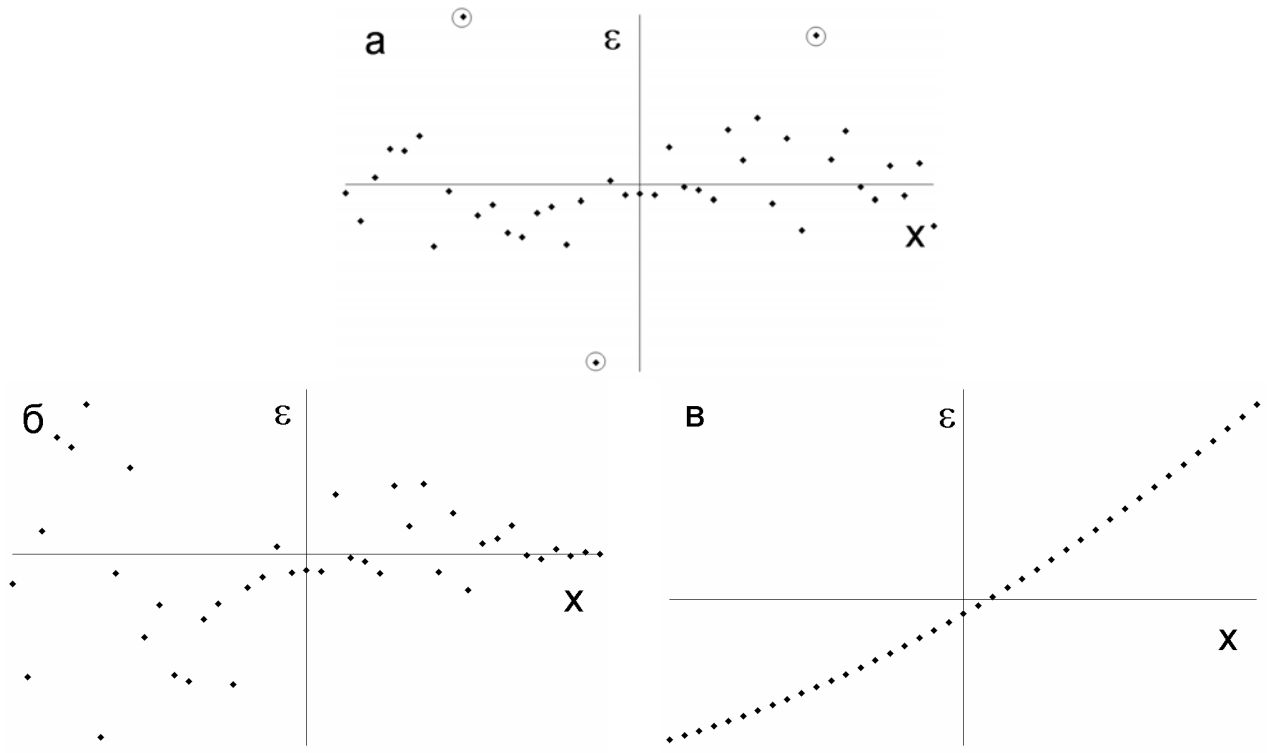


Рис. 7.5. Приклади порушення властивостей залишків неадекватних моделей

Третьою властивістю є те, що залишки адекватної регресійної моделі мають бути некорельованими випадковими величинами. Наявність автокореляції першого порядку перевіряють за допомогою **критерію Дарбіна − Уотсона**:

$$d = \frac{\sum_{i=2}^{n} \left(e_i - e_{i-1}\right)^2}{\sum_{i=1}^{n} e_i^2}, \qquad (7.63)$$

де n – кількість емпіричних точок. Для адекватної моделі має виконуватися умова $d \approx 2$. Близькі до нуля значення d свідчать про наявність додатної автокореляції, а значення, що наближаються до 4, – про наявність від'ємної автокореляції.

Цей критерій було розроблено в 1950 р. британським статистиком Джеймсом Дарбіном та австралійським статистиком Джеффрі Стюартом Уотсоном.

Наявність автокореляції вищих порядків перевіряють шляхом дослідження автокореляційної функції. Про наявність автокореляції в цьому випадку свідчить збільшення абсолютних значень коефіцієнта автокореляції при певних значеннях параметра зсуву. На рис 7.6 показано приклади автокореляційних функцій для ряду, що є білим шумом, (ліворуч) та рядом, який змінюється за синусоїдальним законом, (праворуч).

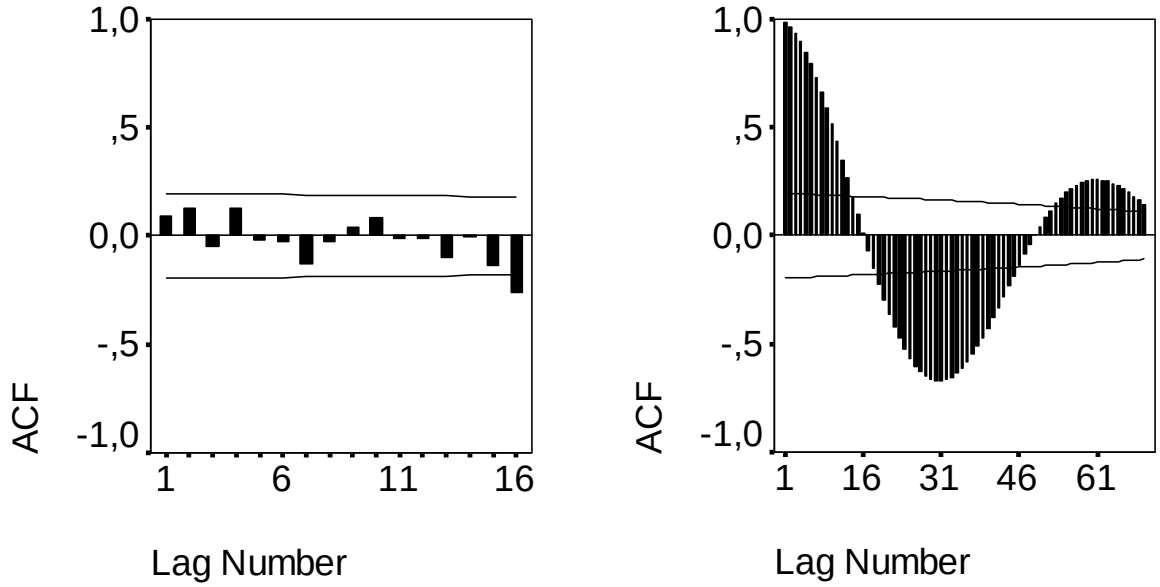


Рис. 7.6. Приклади автокореляційних функцій деяких рядів

Горизонтальними лініями на цих графіках показано довірчі інтервали для нульових значень коефіцієнта автокореляції. З наведених графіків добре видно, що у першому випадку автокореляція є практично відсутньою, а в другому – для певних значень параметра зсуву спостерігається істотна додатна або від'ємна автокореляція.

7.8. Побудова однофакторних регресійних моделей в електронних таблицях MS Excel

Найпростішим випадком є побудова одновимірної лінійної регресійної моделі. Її параметри можна визначити безпосередньо за формулами 7.5. Але в електронних таблицях MS Excel це можна зробити за допомогою вбудованих формул та пакета аналізу.

Розглянемо такий приклад. Нехай ми маємо дві пов'язані вибірки обсягом по 41 елементу. Елементами першої вибірки x_i є числа від –2 до 2, взяті у порядку зростання з кроком 0,1. Елементи другої вибірки розраховуємо за формулою $y_i = 2x_i + 1 + \varepsilon_i$, де ε_i – нормально розподілені випадкові числа з математичним сподіванням 0 та стандартним відхиленням 0,2.

Для побудови лінійної моделі можна скористатися функцією "ЛИНЕЙ()". На рис. 7.7 показано діалогове вікно задання її параметрів.

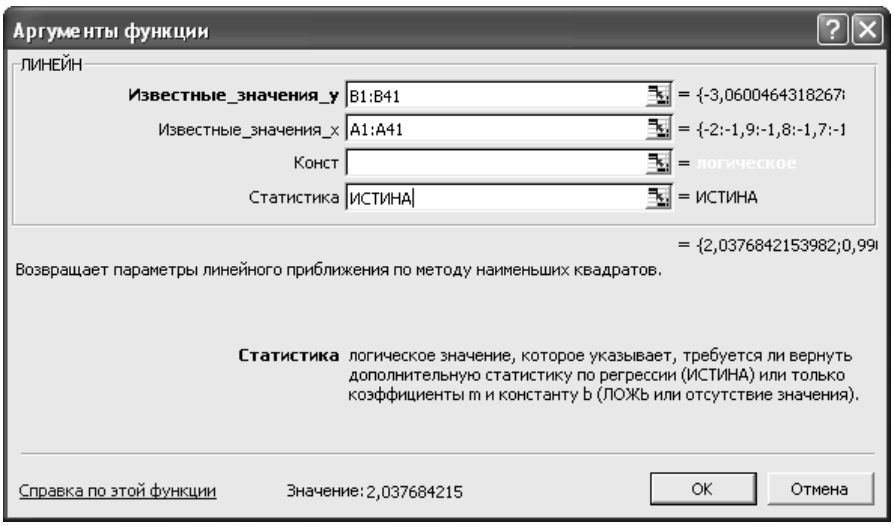
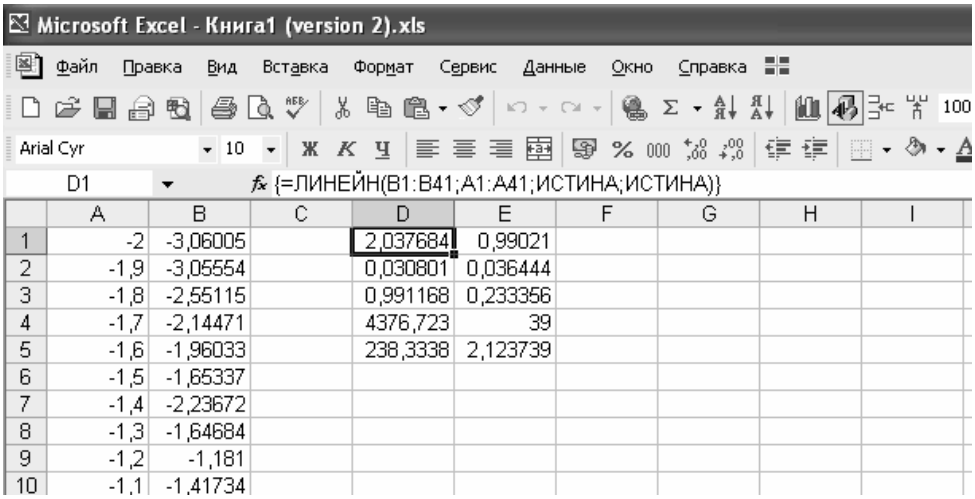


Рис. 7.7. Діалогове вікно задання параметрів функції "ЛИНЕЙ()"

У перших двох комірках задаємо посилання на комірки, що містять значення змінних *Y* та *X*, відповідно. У третій комірці вказуємо необхідність підбору константи (якщо значення дорівнює "ИСТИНА" або відсутнє, константу необхідно обчислити; якщо значення дорівнює "ЛОЖЬ", то константа береться рівною нулю). В останній комірці вказуємо необхідність виведення підсумкової статистики ("ИСТИНА") чи тільки коефіцієнтів моделі ("ЛОЖЬ").

Уведення формули необхідно здійснювати як формулу масиву. Для цього потрібно виділити на робочому аркуші масив суміжних комірок об-

сягом 2×5, записати формулу й після цього натиснути клавішу F2, а потім одночасно клавіші Ctrl+Shift+Enter. При цьому отримуємо масив результатів, наведений на рис. 7.8.



	A	B	C	D	E	F	G	H	I
1	-2	-3,06005		2,037684	0,99021				
2	-1,9	-3,05554		0,030801	0,036444				
3	-1,8	-2,55115		0,991168	0,233356				
4	-1,7	-2,14471		4376,723	39				
5	-1,6	-1,96033		238,3338	2,123739				
6	-1,5	-1,65337							
7	-1,4	-2,23672							
8	-1,3	-1,64684							
9	-1,2	-1,181							
10	-1,1	-1,41734							

Рис. 7.8. Результати обчислення параметрів лінійної регресії

Результати наведено у комірках D1:E5. При цьому: D1, E1 – це значення коефіцієнтів рівняння регресії $y = ax + b$; D2, E2 – стандартні відхилення коефіцієнтів моделі a й b; D3 – коефіцієнт детермінації моделі; E3 – стандартне відхилення для значень y; D4 – F-статистика; E4 – кількість степенів вільності для F-статистики; D5 – регресійна сума квадратів; E5 – сума квадратів залишків.

Крім лінійної, в електронних таблицях MS Excel можна побудувати степеневу регресійну модель вигляду $y = ab^x$, використовуючи вбудовану формулу “=ЛГРФПРИБЛ()”. Застосування цієї формули є таким самим, як і формули “ЛИНЕЙ()”.

Іншим варіантом побудови регресійної моделі в електронних таблицях MS Excel є застосування пакету аналізу. Для цього обираємо у головному меню: Сервіс/Аналіз даних/Регресія. Після цього відкривається діалогове вікно (рис. 7.9).

У цьому вікні позначаємо посилання на комірки, де містяться значення змінних x, y. Позначку “Метки” робимо у випадку, коли перший стовпчик або перший рядок вхідних даних містять заголовки. Позначку “Константа – ноль” робимо у випадку, коли вільний член моделі дорівнює нулю. У комірці “Уровень надежності” задаємо довірчий рівень (за умовчанням він дорівнює 0,95).

Далі позначаємо, куди саме слід виводити результати, а також необхідність виводу статистичних характеристик моделі та побудови графіку нормальної імовірності. Результати для тих самих вихідних даних, що і у попередньому випадку, показано на рис. 7.10.

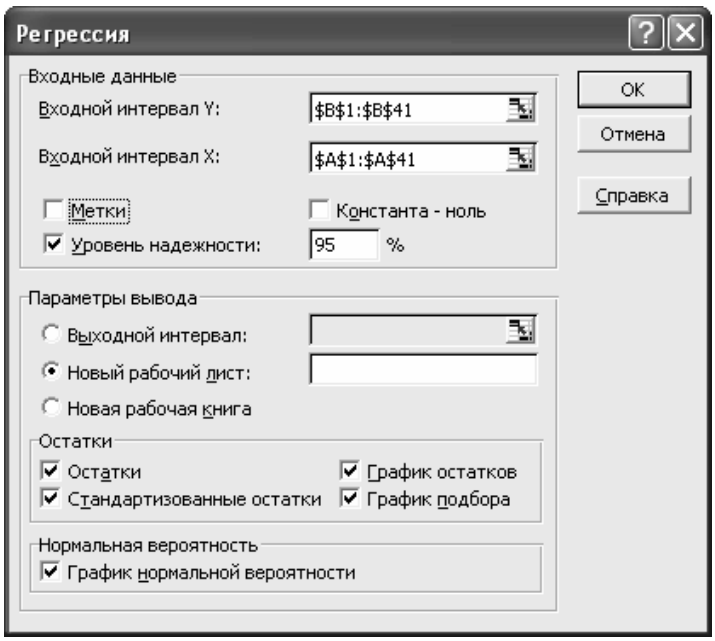


Рис. 7.9. Діалогове вікно побудови регресійної моделі в пакеті аналізу

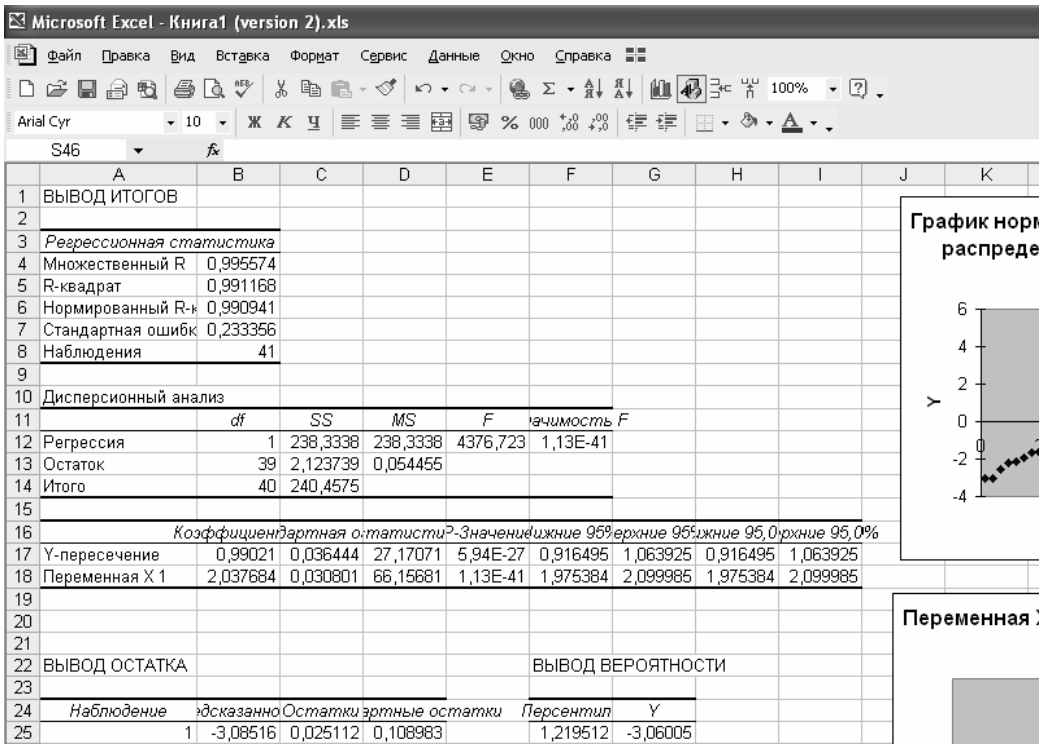


Рис. 7.10. Результати підбору параметрів регресійної моделі

Бачимо, що основні параметри є такими самими, що й у попередньому випадку. Проте слід зазначити, що за допомогою пакету аналізу ми можемо отримати більш докладну інформацію про властивості моделі. Крім того, використання пакету аналізу є простішим. Зокрема, воно не передбачає необхідності заздалегідь розраховувати й виокремлювати на робочому аркуші комірки для виводу результатів.

7.9. Побудова однофакторних регресійних моделей в пакеті SPSS

У пакеті SPSS також є різні засоби побудови регресійних моделей.

Для побудови лінійної моделі заносимо дані аркуш даних (рис. 7.11) й використовуємо пункти меню: Analyze/Regression/Linear.

	VAR00001	VAR00002	var	var	var	var	var
1	-2,00	-3,06					
2	-1,90	-3,06					
3	-1,80	-2,55					
4	-1,70	-2,14					
5	-1,60	-1,96					
6	-1,50	-1,65					
7	-1,40	-2,24					
8	-1,30	-1,65					

Рис. 7.11. Аркуш даних пакету SPSS при побудові регресійної моделі

При цьому відкривається діалогове вікно задання параметрів процедури (рис. 7.12). У цьому вікні необхідно вказати залежну й незалежну (або декілька незалежних) змінну. У випадку декількох незалежних змінних можна згрупувати їх у блоки й обрати методи вводу для різних блоків. Також при застосуванні зваженого методу найменших квадратів можна задати вагові коефіцієнти для незалежних змінних.

У діалоговому вікні “Statistics” (рис. 7.13) задаємо, які статистичні характеристики моделі необхідно показати у вікні результатів. Вивід коефіцієнтів кореляції між змінними й діагностика мультиколінеарності для одно факторної моделі не потрібні.

У вікні “Plots” (рис. 7.14) зазначаємо, які графіки необхідно вивести у вікні результатів.

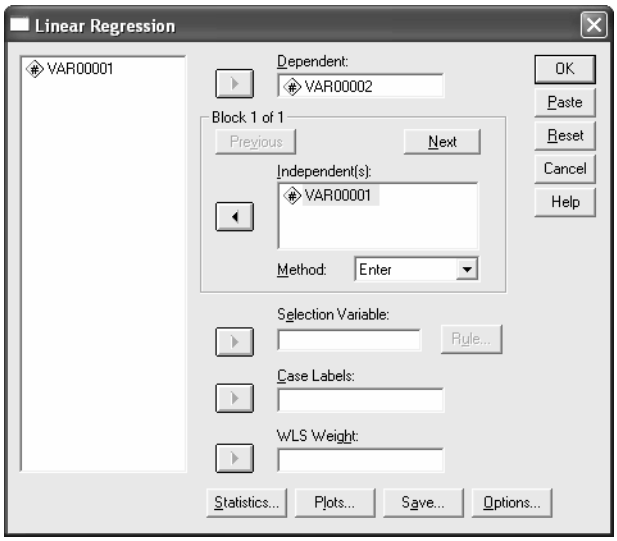


Рис. 7.12. Діалогове вікно задання параметрів побудови лінійної регресійної моделі в пакеті SPSS

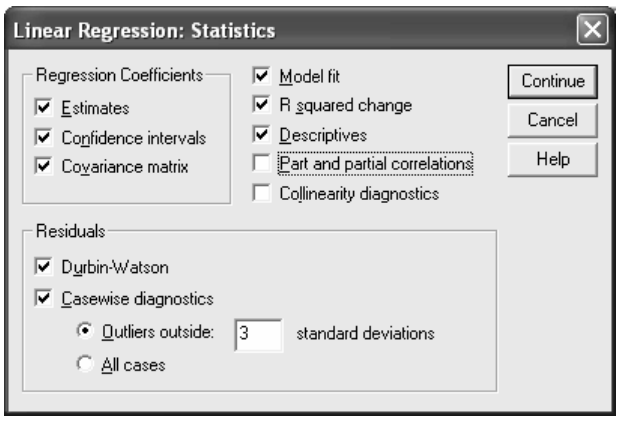


Рис. 7.13. Діалогове вікно задання параметрів статистики лінійної регресійної моделі

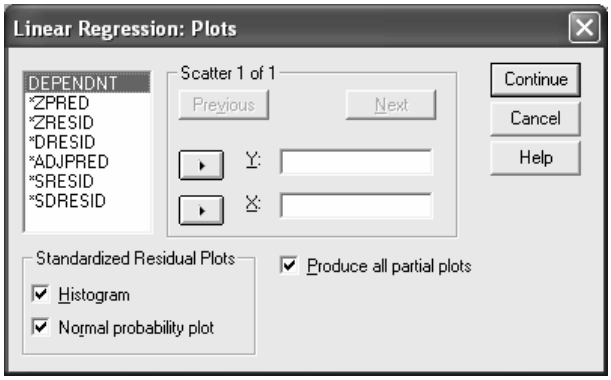


Рис. 7.14. Діалогове вікно задання графіків, які потрібно показати у вікні результатів

У діалоговому вікні “Save” (рис. 7.15) вказуємо, які результати необхідно зберегти у вікні даних як нові змінні. Зокрема можна задати формування таких змінних як передбачувані значення залежної змінної, залишки моделі тощо. Частина змінних потрібна тільки при побудові багатофакторних лінійних моделей.

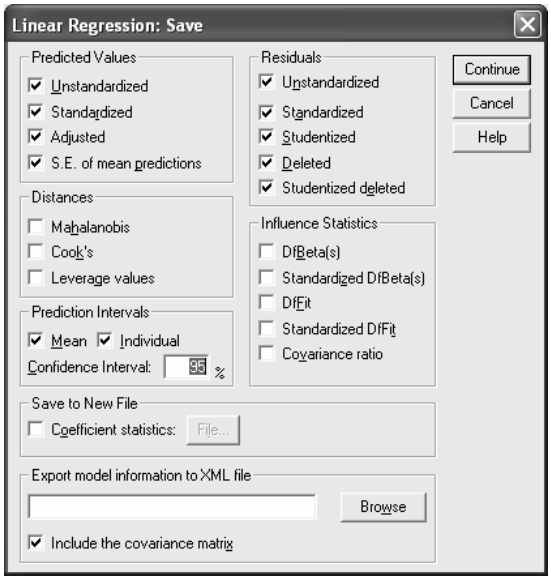


Рис. 7.15. Діалогове вікно задання нових змінних

У діалоговому вікні “Options” (рис. 7.16) задаємо додаткові параметри побудови моделі.

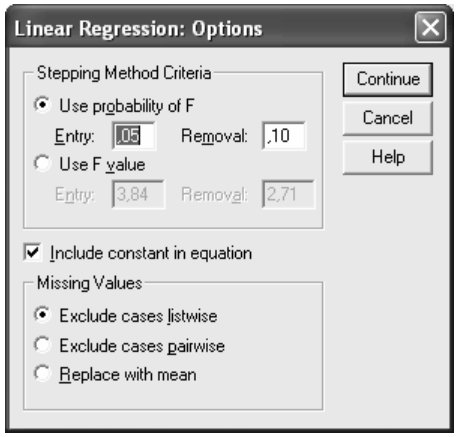


Рис. 7.16. Діалогове вікно задання додаткових параметрів побудови моделі.

Деякі результати наведено на рис. 7.17, 7.18. Бачимо, що вони збігаються з результатами, отриманими в електронних таблицях MS Excel, а також з вихідними даними. Але слід зазначити, що у пакеті SPSS ми маємо можливість отримати значно більше статистичних даних стосовно якості побудованої моделі.

Descriptive Statistics

	Mean	Std. Deviation	N
VAR00002	,9902	2,45182	41
VAR00001	,0000	1,19791	41

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics		Durbin-Watson
					R Square Change	F Change	
1	,996[a]	,991	,991	,23336	,991	4376,723	1,805

a. Predictors: (Constant), VAR00001

b. Dependent Variable: VAR00002

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	238,334	1	238,334	4376,723	,000[a]
	Residual	2,124	39	,054		
	Total	240,458	40			

a. Predictors: (Constant), VAR00001

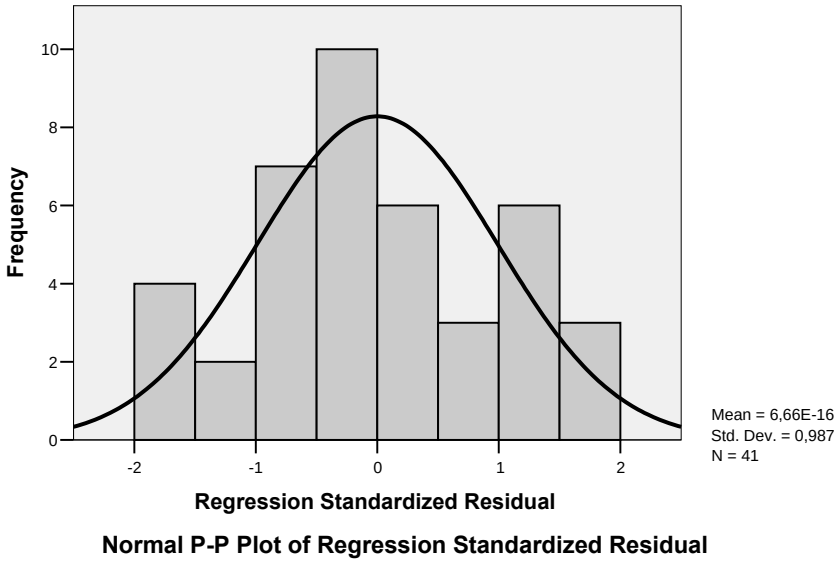
b. Dependent Variable: VAR00002

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-3,0852	5,0656	,9902	2,44097	41
Std. Predicted Value	-1,670	1,670	,000	1,000	41
Standard Error of Predicted Value	,036	,072	,050	,011	41
Adjusted Predicted Value	-3,0878	5,1024	,9899	2,44293	41
Residual	-,39495	,45477	,00000	,23042	41
Std. Residual	-1,692	1,949	,000	,987	41
Stud. Residual	-1,717	1,984	,001	1,013	41
Deleted Residual	-,40664	,47153	,00029	,24275	41
Stud. Deleted Residual	-1,763	2,066	,004	1,031	41
Mahal. Distance	,000	2,787	,976	,883	41
Cook's Distance	,000	,132	,027	,034	41
Centered Leverage Value	,000	,070	,024	,022	41

a. Dependent Variable: VAR00002

Рис. 7.17. Деякі результати побудови лінійної регресійної моделі у пакеті SPSS



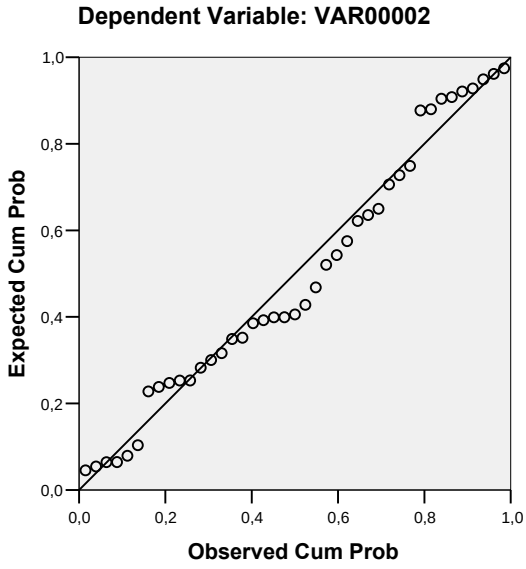


Рис. 7.18. Деякі графіки результатів побудови лінійної регресійної моделі

Розглянемо інший засіб побудови регресійних моделей у пакеті SPSS. Для цього внесемо на аркуш даних такий масив: незалежна змінна X є набором чисел від –2 до 2 з кроком 0,1; значення залежної змінної розраховано за формулою $yi = (2x^3_i - 3x^2_i + 5x_i + 2)\varepsilon_i$, де ε_i – елементим нормально розподіленої випадкової послідовності з математичним сподіванням 1 й стандартним відхиленням 0,2. Оберемо у головному меню: Analyze/Regression/Curve estimation. При цьому відкривається діалогове вікно задання параметрів процедури оцінювання параметрів кривих (рис. 7.19).

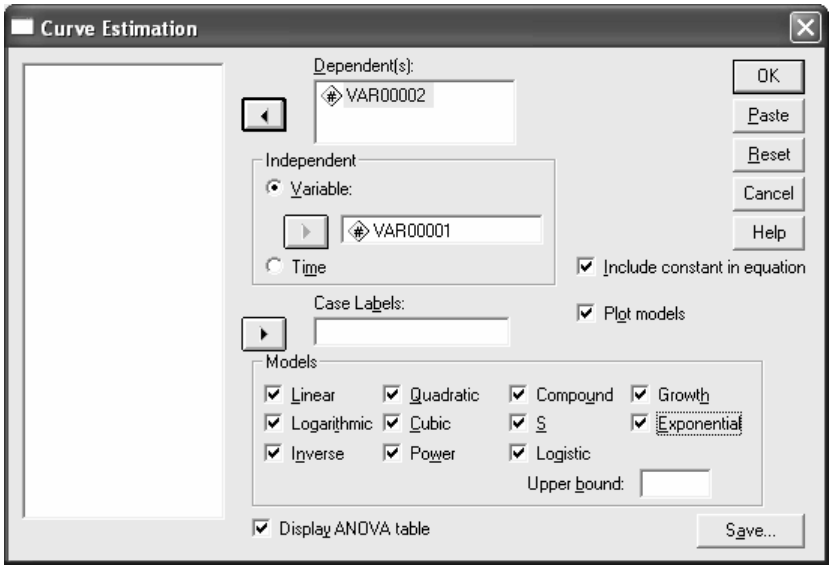


Рис. 7.19. Діалогове вікно задання параметрів процедури оцінювання кривої

У цьому вікні ми вказуємо залежну й незалежну змінні, необхідність включення вільного члена, побудови графіків моделей і таблиці ANOVA, а також типи оцінюваних моделей. Позначку “Case Labels” використовують для вибору типів маркерів при побудові графіків у випадку декількох залежних змінних. Кнопку “Save” використовують для формування нових змінних (передбачувані значення залежної змінної, залишки моделі, довірчі інтервали) на аркуші даних.

У випадку, що розглядається, не всі типи моделей можуть бути побудовані. Про це на аркуші результатів виводиться відповідне повідомлення. Зокрема обернену й *S*-подібну моделі неможливо побудувати через наявність нульового значення незалежної змінної; логарифмічну та степеневу – через наявність від’ємних значень незалежної змінної; складену, степеневу, *S*-подібну, зростання, експоненціальну й логістичну – через наявність від’ємних значень залежної змінної.

Тому залишаються три типи доступних моделей – лінійна, квадратична й кубічна. На рис. 7.20 побудовано їх графіки, з яких видно, що найбільш придатною є кубічна модель. Це відповідає вихідним даним. Основні результати для кубічної моделі наведено на рис. 7.21.

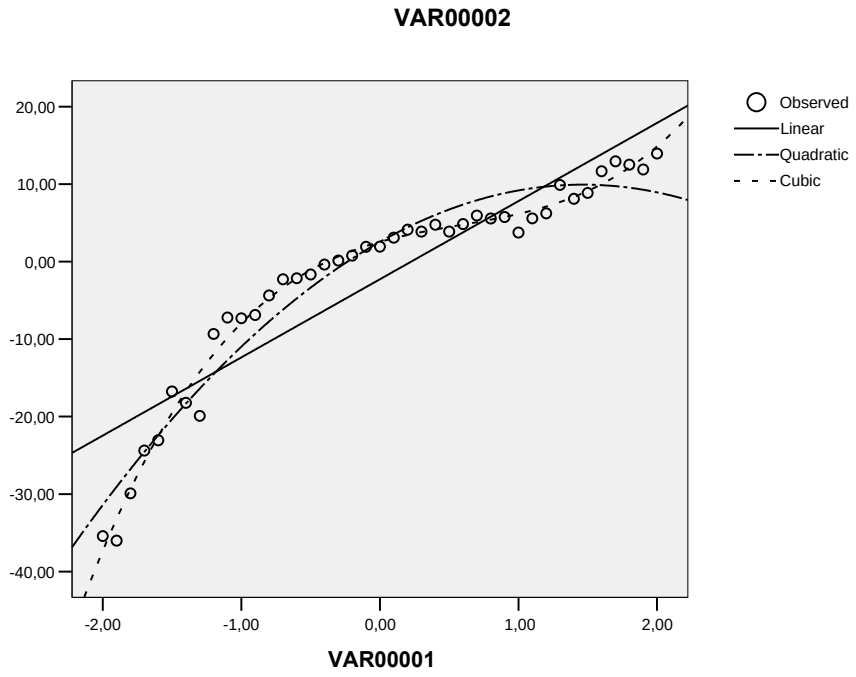


Рис. 7.20. Графіки побудованих моделей

Бачимо, що отримані результати задовільно збігаються з вихідними даними, але є помітні розбіжності у значеннях окремих коефіцієнтів моделі. Це може бути пов’язано з малим обсягом вихідної вибірки й високим рівнем стандартного відхилення при розрахунку вихідних значень залежної змінної.

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
,993	,986	,985	1,625

The independent variable is VAR00001.

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	6874,087	3	2291,362	867,333	,000
Residual	97,748	37	2,642		
Total	6971,835	40			

The independent variable is VAR00001.

Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
VAR00001	5,063	,537	,459	9,420	,000
VAR00001 ** 2	-3,444	,203	-,330	-16,972	,000
VAR00001 ** 3	1,995	,196	,497	10,192	,000
(Constant)	2,540	,381		6,668	,000

Рис. 7.21. Основні результати підбору кубічної моделі

7.10. Побудова однофакторних регресійних моделей в пакеті MathCad

Для побудови лінійної моделі створимо масив даних, що містить 11 точок. Значення змінної x сформуємо у вигляді арифметичної прогресії з першим членом –5 та різницею 1. Значення змінної y сформуємо за формулою:

$$y = 2x + 1 + \varepsilon,$$

де ε – рівномірно розподілена випадкова величина, задана на відрізку [–2; 2].

На рис. 7.22, 7.23 показано робочі вікна з двома варіантами програми побудови моделі, а на рис. 7.24 – результат побудови моделі (однаковий для обох випадків).

$$x := (-5 \ -4 \ -3 \ -2 \ -1 \ 0 \ 1 \ 2 \ 3 \ 4 \ 5)^T$$

$$y := (-9.5 \ -8.6 \ -4.6 \ -3.4 \ 0.5 \ 2.8 \ 1.1 \ 4.6 \ 8.8 \ 7.4 \ 9.9)^T$$

$$line(x, y) = \begin{pmatrix} 0.818 \\ 1.98 \end{pmatrix}$$

$$f(t) := line(x, y)_0 + line(x, y)_1 \cdot x$$

Рис. 7.22. Фрагмент програми побудови лінійної однофакторної моделі за допомогою функції line(x, y)

$$x := (-5 \ -4 \ -3 \ -2 \ -1 \ 0 \ 1 \ 2 \ 3 \ 4 \ 5)^T$$

$$y := (-9.5 \ -8.6 \ -4.6 \ -3.4 \ 0.5 \ 2.8 \ 1.1 \ 4.6 \ 8.8 \ 7.4 \ 9.9)^T$$

$$intercept(x, y) = 0.818$$

$$slope(x, y) = 1.98$$

$$g(x) := intercept(x, y) + slope(x, y)\, x$$

Рис. 7.23. Фрагмент програми побудови лінійної однофакторної моделі за допомогою функцій intercept(x, y) та slope(x, y)

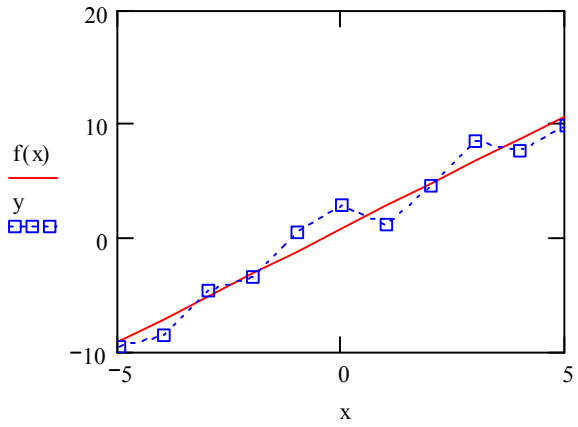


Рис. 7.24. Результат побудови лінійної однофакторної моделі

Функція line(x, y) видає вектор коефіцієнтів лінійної моделі $f(x) = ax+b$ у вигляді вектора $\begin{pmatrix} b \\ a \end{pmatrix}$.

Значенням функції intercept(x, y) є ордината точки перетину моделі з віссю ординат, а значенням функції slope(x, y) – тангенс кута нахилу моделі до осі абсцис.

На рис. 7.25 наведено фрагмент програми для побудови медіанної регресійної моделі за даними, що використовувалися при побудові лінійної моделі методом найменших квадратів.

$$\text{medfit}(x, y) = \begin{pmatrix} 0.867 \\ 2.1 \end{pmatrix}$$

$$q(x) := \text{medfit}(x, y)_0 + \text{medfit}(x, y)_1 \cdot x$$

Рис. 7.25. Фрагмент програми побудови медіанної регресійної моделі

Функція medfit(x, y) видає вектор коефіцієнтів лінійної моделі $f(x) = ax+b$ у вигляді вектора $\begin{pmatrix} b \\ a \end{pmatrix}$. На рис. 7.26 показано результати порівняння звичайної лінійної та медіанної моделей.

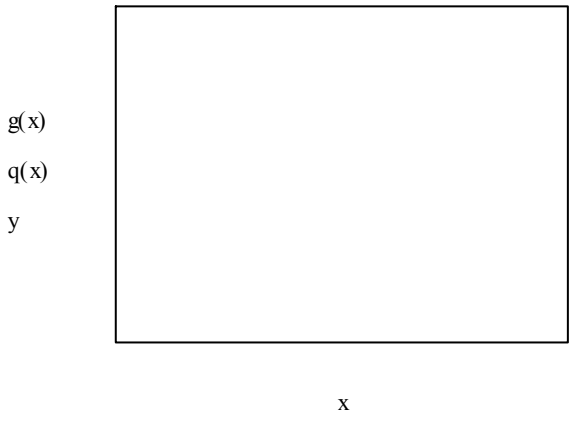


Рис. 7.26. Результат побудови лінійної однофакторної моделі

Для побудови поліноміальної моделі згенеруємо вектор $x1$, який містить 21 елемент, що утворюють арифметичну прогресію з початковим значенням –2,25 й різницею 0,25. Елементи вектора $y1$ розрахуємо за формулою:

$$y1 = -2x1^4 + 3x1^3 + 1,5x1^2 - 7x1 + 2 + \varepsilon,$$

де ε – елементи рівномірної випадкової послідовності, заданої на відрізку [–3; 3].

На рис. 7.27 наведено фрагмент програми, що використовується для побудови поліноміальної моделі.

```
x1 := (−2.5 −2.25 −2 −1.75 −1.5 −1.25 −1 −0.75 −0.5 −0.25 0 0.25 0.5 0.75 1 1.25 1.5 1.75 2 2.25 2.5)^T

y1 := (−96.8 −62.5 −33.8 −13.6 −2.3 5.1 2.6 5.6 7.5 1.6 2.8 −2.3 −3.7 −3.1 −4.2 −6.3 −6.4 −9.3 −13.7 −24.1 −38.1)^T

k := 4

s := regress(x1, y1, 4)

        ⎛  3      ⎞
        ⎜  3      ⎟
        ⎜  4      ⎟
        ⎜ 0.644   ⎟
s =     ⎜−8.548   ⎟
        ⎜ 2.366   ⎟
        ⎜ 3.294   ⎟
        ⎝−2.134   ⎠

A(t) := interp(s, x1, y1, t)
```

Рис. 7.27. Фрагмент програми побудови поліноміальної регресійної моделі

Останні п’ять елементів вектора s є оцінками коефіцієнтів a_0, a_1, a_2, a_3 й a_4 полінома $A(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4$. На рис. 7.28 показано графік отриманої моделі, який достатньо добре узгоджується з вихідними даними.

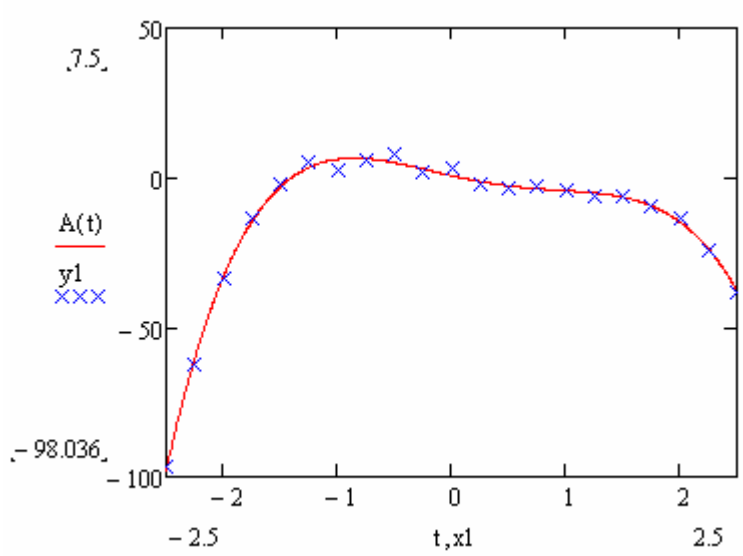


Рис. 7.28. Результат побудови регресійної моделі у вигляді полінома четвертого степеня

Для порівняння на рис. 7.29–7.32 показано графіки моделей у вигляді поліномів першого, другого, третього та п’ятого степенів, побудованих для тих самих вихідних даних.

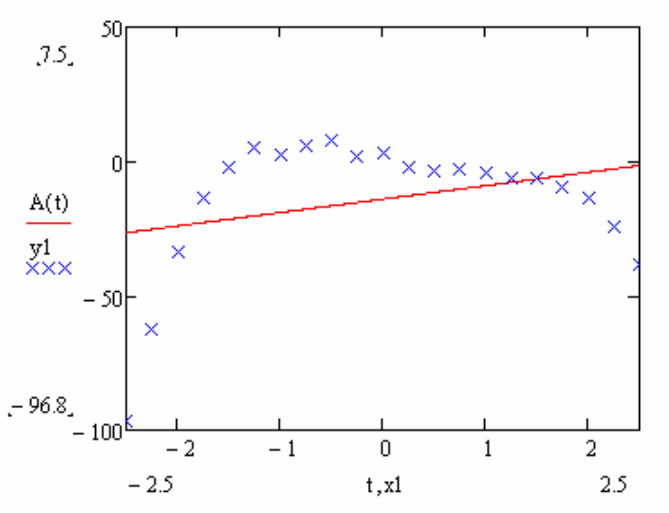


Рис. 7.29. Результат побудови регресійної моделі у вигляді полінома першого степеня (лінійної моделі)

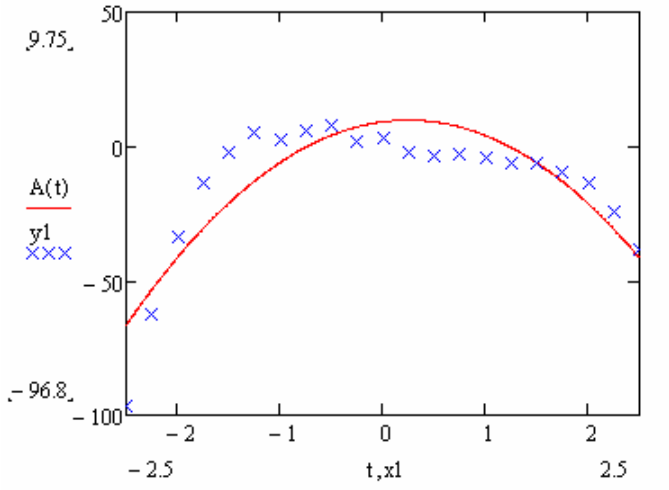


Рис. 7.30. Результат побудови регресійної моделі у вигляді полінома другого степеня (квадратичної моделі)

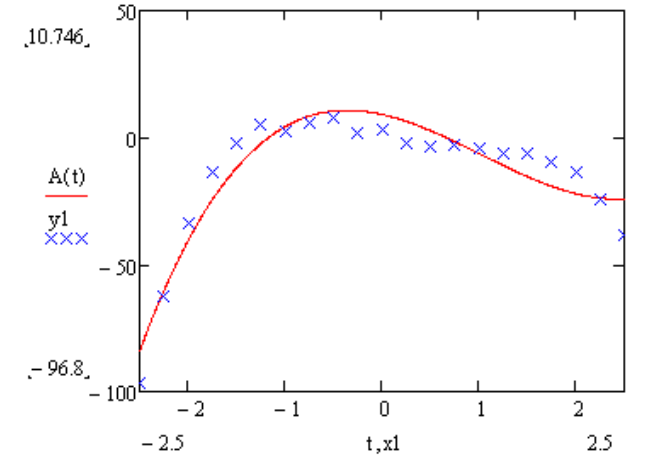


Рис. 7.31. Результат побудови регресійної моделі у вигляді полінома третього степеня (кубічної моделі)

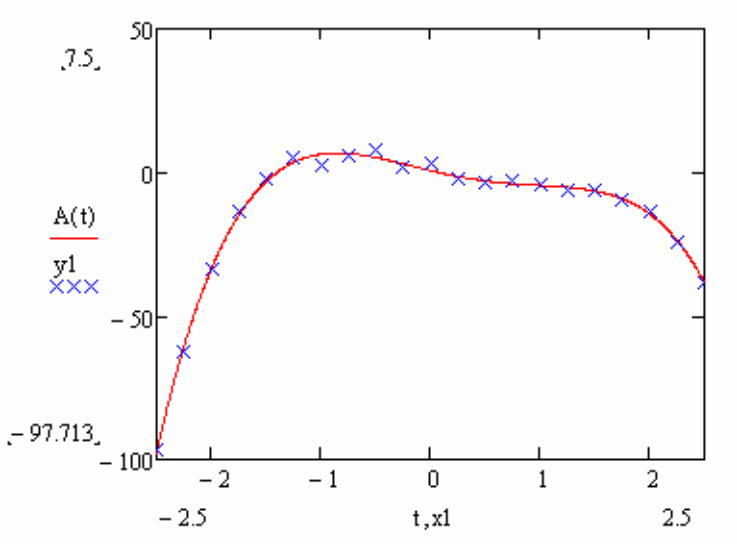


Рис. 7.32. Результат побудови регресійної моделі у вигляді полінома п’ятого степеня

Зіставлення наведених результатів дає підстави зробити висновок, що для наведених даних вже модель третього порядку достатньо добре відображає основні особливості вихідних даних, а модель п’ятого порядку є надлишковою, оскільки немає жодних переваг перед моделлю четвертого порядку, але є більш складною.

Альтернативний варіант передбачає побудову моделі у вигляді відрізків поліномів. Фрагмент програми для побудови такої моделі наведено на рис. 7.33. При цьому використовували ті самі дані, що і для побудови звичайної поліноміальної моделі.

```
s1 := loess(x1, y1, 0.75)

A1(t) := interp(s1, x1, y1, t)
```

Рис. 7.33. Фрагмент програми побудови регресійної моделі
у вигляді відрізків поліномів

На рис. 7.34–7.37 наведено результати побудови моделі для різних значень фактора span, що задає довжину відрізків поліномів. Аналіз наведених даних показує, що при малих значеннях фактора span модель краще відображає наявні дані. Але занадто добра відповідність моделі й вихідних даних може бути непотрібною, оскільки дані містять певну похибку.

Крім того для малих значень фактора span істотно збільшується обсяг розрахунків і при span $\leq$ 0,28 з’являється повідомлення про нестачу пам’яті для завершення операції. Із збільшенням параметра span понад 0,75–0,8 якість моделі погіршується, а при span = 1 відхилення моделі від даних вже є неприпустимо великими.

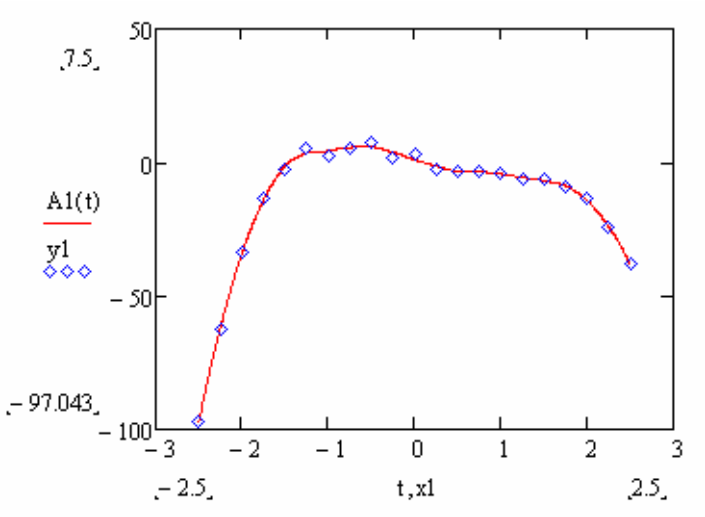


Рис. 7.34. Результат побудови регресійної моделі у відрізків поліномів для span = 0,3

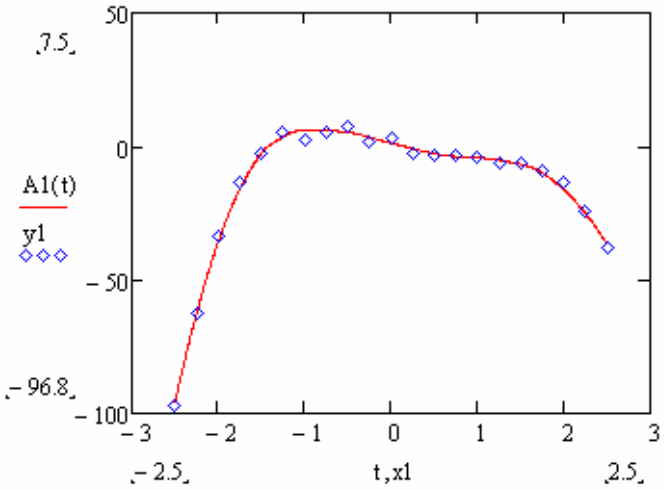


Рис. 7.35. Результат побудови регресійної моделі у відрізків поліномів для span = 0,5

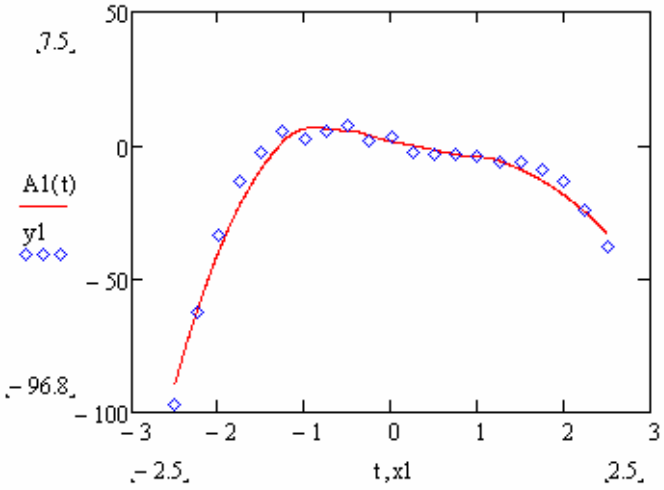


Рис. 7.36. Результат побудови регресійної моделі у відрізків поліномів для span = 0,75

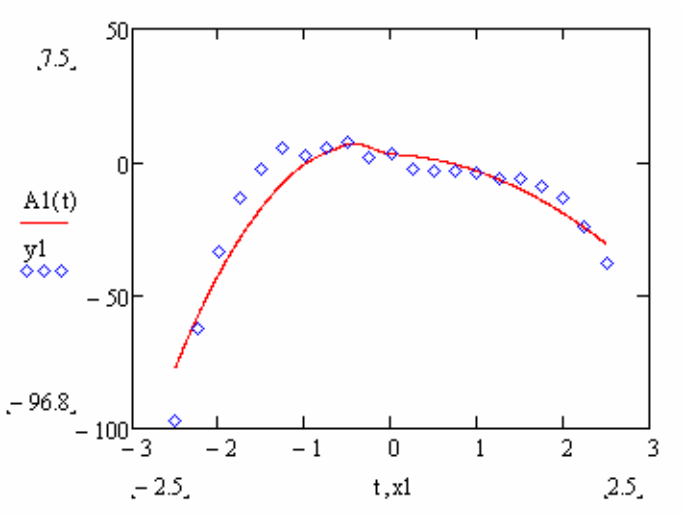


Рис. 7.37. Результат побудови регресійної моделі у відрізків поліномів для span = 1

Для побудови спеціальних моделей можна використовувати спеціальні функції пакету MathCad. Для ілюстрації цього візьмемо попередній вектор $x1$, а вектор $z1$ сформуємо за формулою:

$$z1 = \frac{5\varepsilon}{1 + 2\exp(-3x1)},$$

де ε – випадкова величина, рівномірно розподілена на відрізку [0,9; 1,1].

Для побудови експоненціальної та логістичної регресійних моделей використовували форму, фрагмент якої наведено на рис. 7.38.

$x1 := (-2.5\ \ -2.25\ \ -2\ \ -1.75\ \ -1.5\ \ -1.25\ \ -1\ \ -0.75\ \ -0.5\ \ -0.25\ \ 0\ \ 0.25\ \ 0.5\ \ 0.75\ \ 1\ \ 1.25\ \ 1.5\ \ 1.75\ \ 2\ \ 2.25\ \ 2.5)^T$

$z1 := (0.001\ \ 0.003\ \ 0.006\ \ 0.014\ \ 0.026\ \ 0.061\ \ 0.12\ \ 0.27\ \ 0.53\ \ 0.88\ \ 1.64\ \ 2.34\ \ 3.19\ \ 3.92\ \ 4.83\ \ 5.16\ \ 5.25\ \ 5.37\ \ 4.76\ \ 5.09\ \ 5.33)^T$

$g := \begin{pmatrix} 5 \\ 2 \\ 2 \end{pmatrix}$

$v := expfit(x1, z1, g)$

$v = \begin{pmatrix} 10.233 \\ 0.135 \\ -8.123 \end{pmatrix}$

$f(t) := v_0 \cdot \exp(v_1 \cdot t) + v_2$

$s := lgsfit(x1, z1, g)$

$ff(t) := \dfrac{s_0}{1 + s_1 \cdot \exp(-s_2 \cdot t)}$ $\qquad s = \begin{pmatrix} 5.266 \\ 2.461 \\ 2.865 \end{pmatrix}$

Рис. 7.38. Фрагмент програми побудови експоненціальної та логістичної регресійної моделей

Результати побудови моделей показано на рис. 7.39, 7.40.

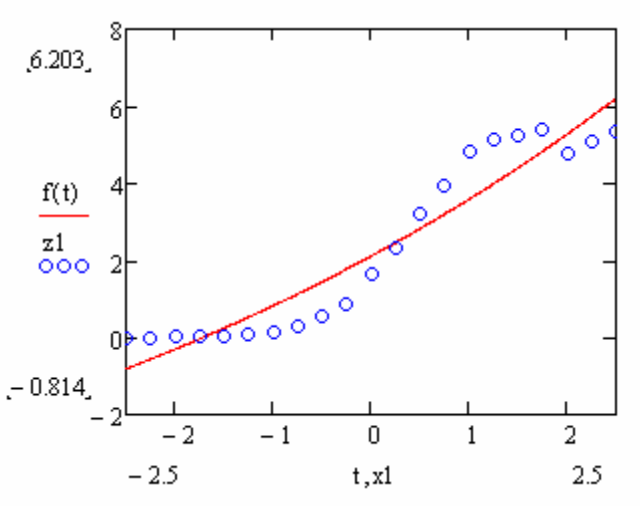


Рис. 7.39. Результат побудови експоненціальної регресійної моделі

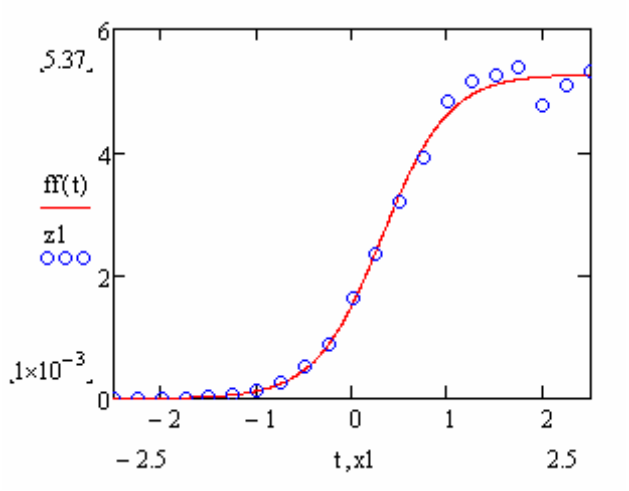


Рис. 7.40. Результат побудови логістичної регресійної моделі

Бачимо, що логістична модель значно краще відображає емпіричні дані, що у даному випадку є цілком природним оскільки вихідні дані побудовано саме на основі логістичної моделі. Дослідження впливу початкових значень параметрів моделей показує, що навіть для досить істотних їх відхилень від правильних значень, результат підбору параметрів моделей зазвичай є одним і тим самим. Але для окремих наборів вихідних значень алгоритм підбору параметрів не збігається.

На рис. 7.41 наведено фрагмент програми, яка дає змогу будувати регресійну модель у вигляді лінійної комбінації двох експонент.

Результат виконання цієї програми наведено на рис. 7.42. Бачимо, що побудована модель достатньо точно описує наведені дані.

Отримані результати дають підстави стверджувати, що пакет MathCad можна застосовувати для побудови однофакторних регресійних моделей різних типів.

$$x := (0\ \ 0.5\ \ 1\ \ 1.5\ \ 2\ \ 2.5\ \ 3\ \ 3.5\ \ 4\ \ 4.5\ \ 5\ \ 5.5\ \ 6\ \ 6.5\ \ 7\ \ 7.5\ \ 8\ \ 8.5\ \ 9\ \ 9.5\ \ 10)^T$$

$$y := (5.4\ \ 4.1\ \ 4.1\ \ 3.6\ \ 2.8\ \ 2.8\ \ 2.3\ \ 2.3\ \ 2\ \ 1.5\ \ 1.46\ \ 1.2\ \ 1.08\ \ 0.99\ \ 0.99\ \ 0.9\ \ 0.8\ \ 0.72\ \ 0.57\ \ 0.54\ \ 0.51)^T$$

$$F(x) := \begin{pmatrix} \exp\left(\frac{-x}{3}\right) \\ \exp\left(\frac{-x}{5}\right) \end{pmatrix}$$

$$C := \text{linfit}(x, y, F)$$

$$C = \begin{pmatrix} 2.343 \\ 2.79 \end{pmatrix}$$

$$f(x) := C_0 \exp\left(\frac{-x}{3}\right) + C_1 \exp\left(\frac{-x}{5}\right)$$

Рис. 7.41. Фрагмент програми побудови моделі у вигляді суми функцій

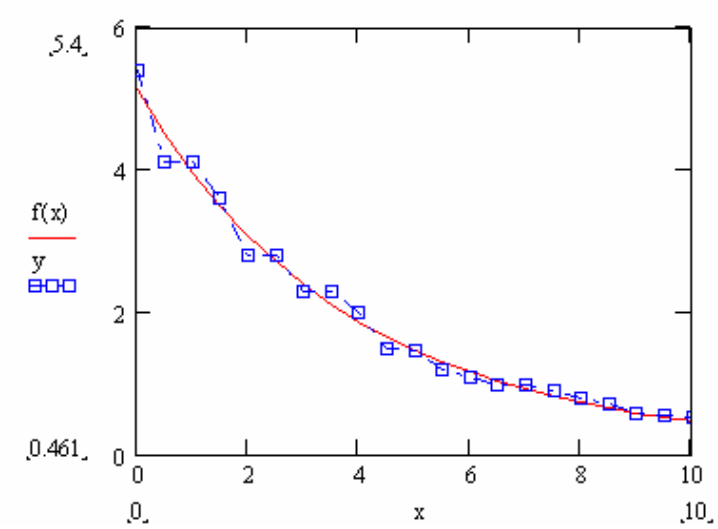


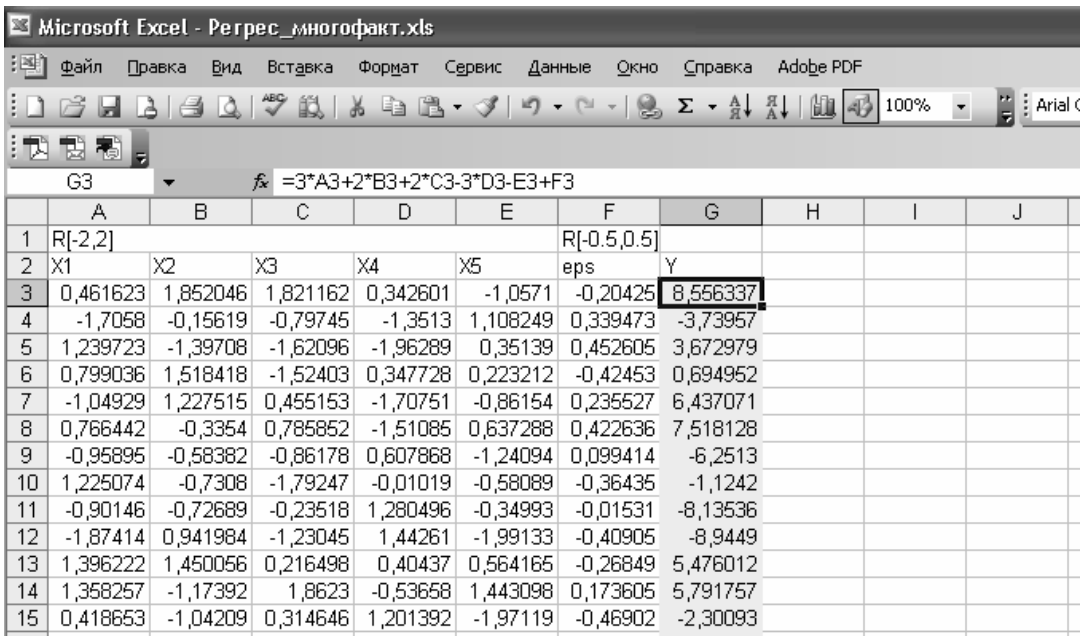
Рис. 7.42. Результат побудови моделі у вигляді суми двох експонент

7.11. Побудова лінійної багатофакторної моделі в електронних таблицях MS Excel

Створимо у на робочому аркуші масив, що містить по 50 значень п’яті змінних, що є елементами рівномірної випадкової послідовності, заданої на відрізку [2; 2]. Значення залежної змінної розрахуємо за формулою:

=3*A3+2*B3+2*C3-3*D3-E3+F3,

де A3–E3 – посилання на комірки зі значеннями сформованих змінних, а F3 – посилання на комірку, де міститься значення елементу рівномірної випадкової величини, яка задана на відрізку [–a; a]. Фрагмент робочого аркуша наведено на рис. 7.43.



	A	B	C	D	E	F	G	H	I	J
1	R[-2,2]					R[-0.5,0.5]				
2	X1	X2	X3	X4	X5	eps	Y			
3	0,461623	1,852046	1,821162	0,342601	-1,0571	-0,20425	8,556337			
4	-1,7058	-0,15619	-0,79745	-1,3513	1,108249	0,339473	-3,73957			
5	1,239723	-1,39708	-1,62096	-1,96289	0,35139	0,452605	3,672979			
6	0,799036	1,518418	-1,52403	0,347728	0,223212	-0,42453	0,694952			
7	-1,04929	1,227515	0,455153	-1,70751	-0,86154	0,235527	6,437071			
8	0,766442	-0,3354	0,785852	-1,51085	0,637288	0,422636	7,518128			
9	-0,95895	-0,58382	-0,86178	0,607868	-1,24094	0,099414	-6,2513			
10	1,225074	-0,7308	-1,79247	-0,01019	-0,58089	-0,36435	-1,1242			
11	-0,90146	-0,72689	-0,23518	1,280496	-0,34993	-0,01531	-8,13536			
12	-1,87414	0,941984	-1,23045	1,44261	-1,99133	-0,40905	-8,9449			
13	1,396222	1,450056	0,216498	0,40437	0,564165	-0,26849	5,476012			
14	1,358257	-1,17392	1,8623	-0,53658	1,443098	0,173605	5,791757			
15	0,418653	-1,04209	0,314646	1,201392	-1,97119	-0,46902	-2,30093			

Рис. 7.43. Фрагмент робочого аркуша з даними

Для побудови лінійної регресійної моделі скористуємося засобом “Регресія” Пакету аналізу електронних таблиць MS Excel. У діалоговому вікні (рис. 7.44) помічаємо комірки, що містять незалежні й залежну змінні, а також які саме результати потрібно вивести.

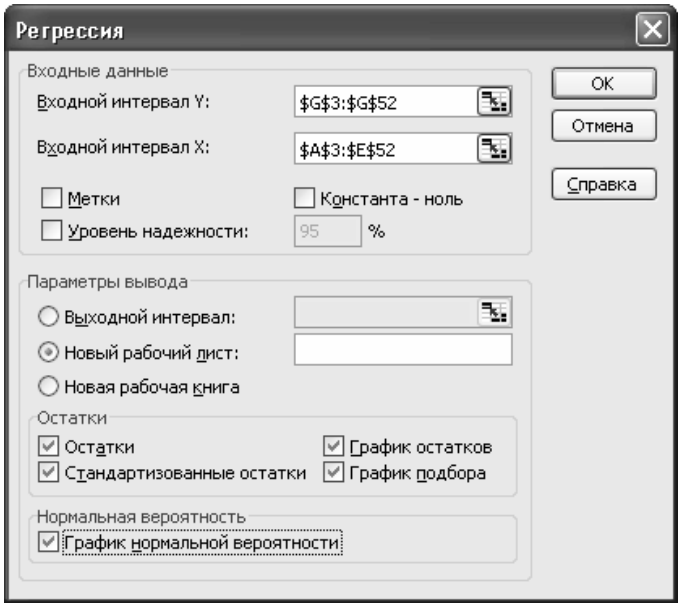


Рис. 7.44. Діалогове вікно підбору параметрів регресійної моделі

Результати роботи програми наведено на рис. 7.45. Звідси бачимо, що модель можна записати у вигляді:

$$Y = 2,98x_1 + 2,02x_2 + 1,99x_3 - 3,05x_4 - 0,96x_5 .$$

Регрессионная статистика								
Множественный R	0,998815							
R-квадрат	0,997631							
Нормированный R-квадрат	0,997362							
Стандартная ошибка	0,327466							
Наблюдения	50							
Дисперсионный анализ								
	df	*SS*	*MS*	*F*	*ачимость F*			
Регрессия	5	1987,279	397,4558	3706,431	1,46E-56			
Остаток	44	4,7183	0,107234					
Итого	49	1991,997						
	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	0,003106	0,047113	0,065927	0,947734	-0,09184	0,098055	-0,09184	0,098055
Переменная X 1	2,98268	0,043683	68,2808	2,73E-46	2,894643	3,070716	2,894643	3,070716
Переменная X 2	2,020953	0,04242	47,64104	1,67E-39	1,935461	2,106446	1,935461	2,106446
Переменная X 3	1,98938	0,044526	44,67861	2,66E-38	1,899642	2,079117	1,899642	2,079117
Переменная X 4	-3,04665	0,046941	-64,9033	2,49E-45	-3,14125	-2,95204	-3,14125	-2,95204
Переменная X 5	-0,9552	0,041872	-22,8127	5,19E-26	-1,03959	-0,87082	-1,03959	-0,87082

Рис. 7.45. Результати підбору лінійної моделі

Стандартне відхилення вільного члена істотно перевищує його значення, тому включати вільний член до моделі недоцільно. Коефіцієнт детермінації побудованої моделі дорівнює 0,997 і є досить близьким до одиниці, що свідчить про адекватність лінійної моделі.

Графіки залишків (рис. 7.46) вказують на їх випадковий характер, що також є свідченням адекватності моделі.

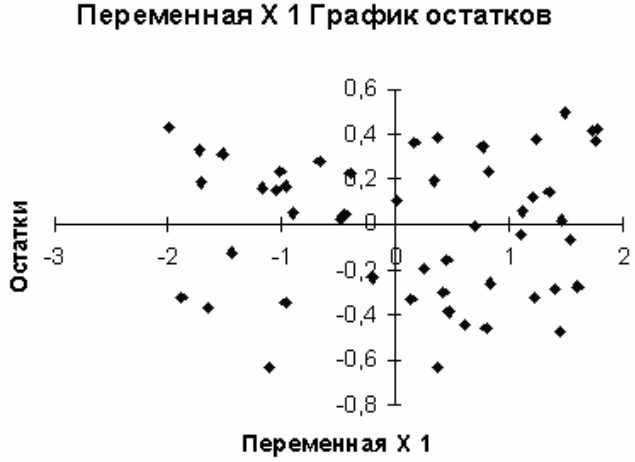


Рис. 7.46. Графік залишків для змінної x_1

Процедура “Регресія” електронних таблиць MS Excel також дає змогу побудувати графіки підбору, що характеризують відповідність моделі вихідним даним. Для досліджуваної моделі вони мають вигляд, наведений на рис. 7.47.

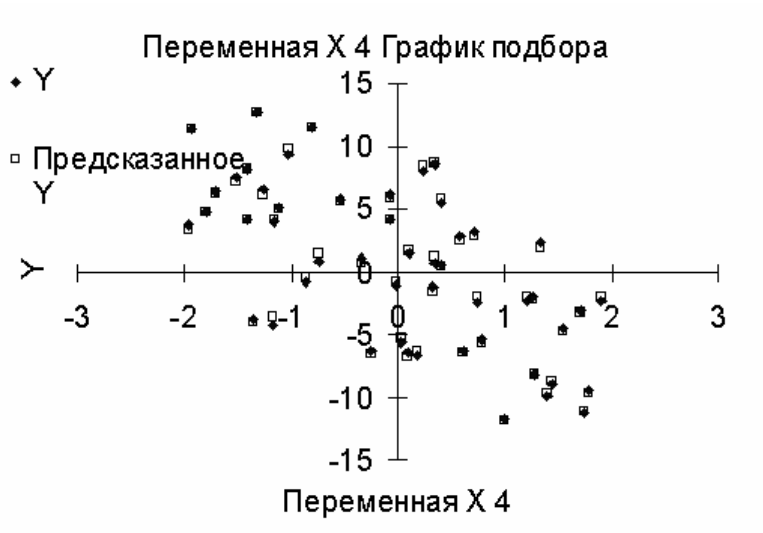


Рис. 7.47. Графік підбору для змінної x_4

На рис. 7.48 показано графік нормального розподілу, який дає змогу перевірити відповідність ряду залишків нормальному розподілу. Але цей графік не зручний для практичного застосування.

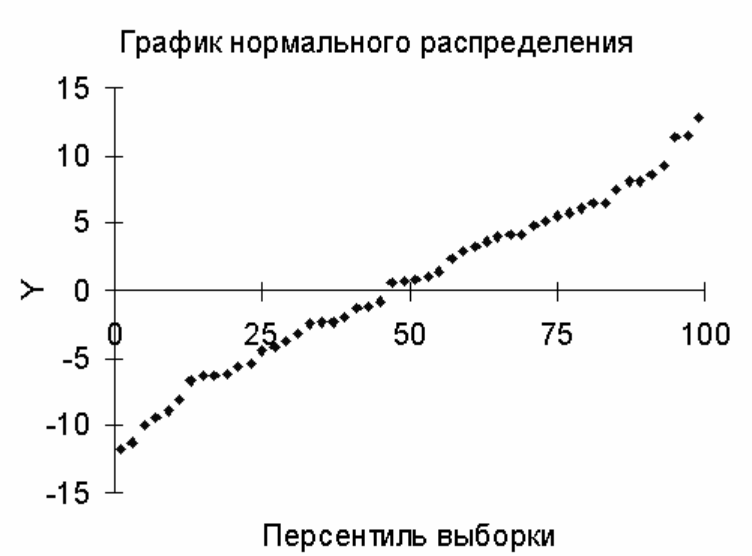
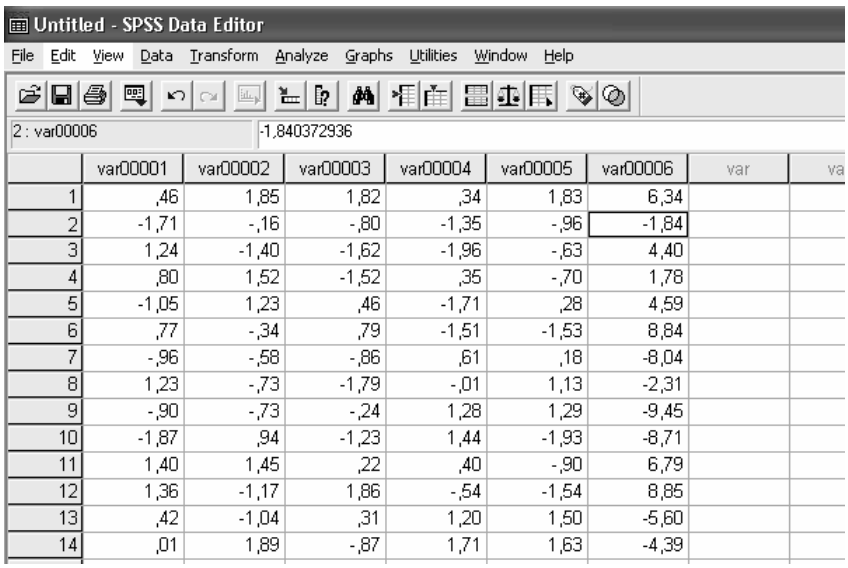


Рис. 7.48. Графік нормального розподілу

7.12. Побудова лінійної багатофакторної моделі в пакеті SPSS

Для побудови й дослідження багатофакторних регресійних моделей також можна також використовувати статистичний пакет SPSS [17]. Проілюструємо це на прикладі розглянутої вище моделі з п’ятма незалежними змінними.

Заносимо на робочий аркуш значення усіх змінних (рис. 7.49)



	var00001	var00002	var00003	var00004	var00005	var00006	var	va
1	,46	1,85	1,82	,34	1,83	6,34		
2	-1,71	-,16	-,80	-1,35	-,96	-1,84		
3	1,24	-1,40	-1,62	-1,96	-,63	4,40		
4	,80	1,52	-1,52	,35	-,70	1,78		
5	-1,05	1,23	,46	-1,71	,28	4,59		
6	,77	-,34	,79	-1,51	-1,53	8,84		
7	-,96	-,58	-,86	,61	,18	-8,04		
8	1,23	-,73	-1,79	-,01	1,13	-2,31		
9	-,90	-,73	-,24	1,28	1,29	-9,45		
10	-1,87	,94	-1,23	1,44	-1,93	-8,71		
11	1,40	1,45	,22	,40	-,90	6,79		
12	1,36	-1,17	1,86	-,54	-1,54	8,85		
13	,42	-1,04	,31	1,20	1,50	-5,60		
14	,01	1,89	-,87	1,71	1,63	-4,39		

Рис. 7.49. Вигляд робочого аркушу SPSS з вихідними даними

Потім у головному меню обираємо Analyze/Regression/Linear regression. Після цього з'являється вікно вибору параметрів моделі (рис. 7.50). У цьому вікні позначаємо залежну й незалежні змінні. У вікні Statistics (рис. 7.51) позначаємо, які параметри моделі необхідно вивести. Зокрема можна передбачити вивід довірчих інтервалів для коефіцієнтів моделі, перевірку мультиколінеарності даних, перевірку автокореляції залишків моделі за критерієм Дарбіна-Уотсона тощо.

У вікні Plots (рис. 7.52) позначаємо які графіки необхідно побудувати й які змінні відкладати за їх осями. Зокрема ми можемо побудувати гістограму залишків і відповідний графік нормального розподілу. У вікні Save (рис. 7.53) зазначаємо, які змінні необхідно додати у вікні даних.

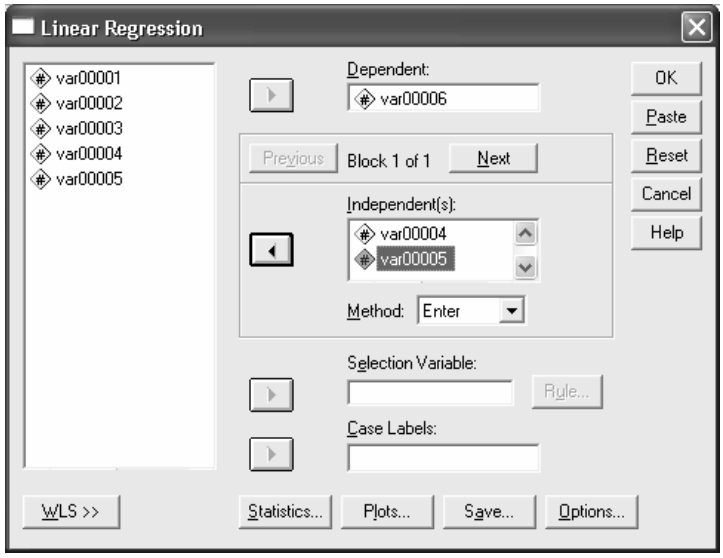


Рис. 7.50. Діалогове вікно вибору параметрів побудови регресійної моделі

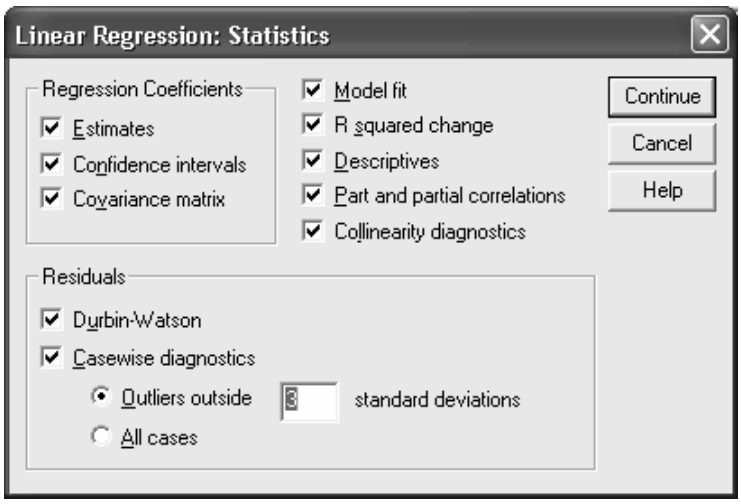


Рис. 7.51. Діалогове вікно задання статистичних параметрів моделі, які необхідно вказати у вікні виводу

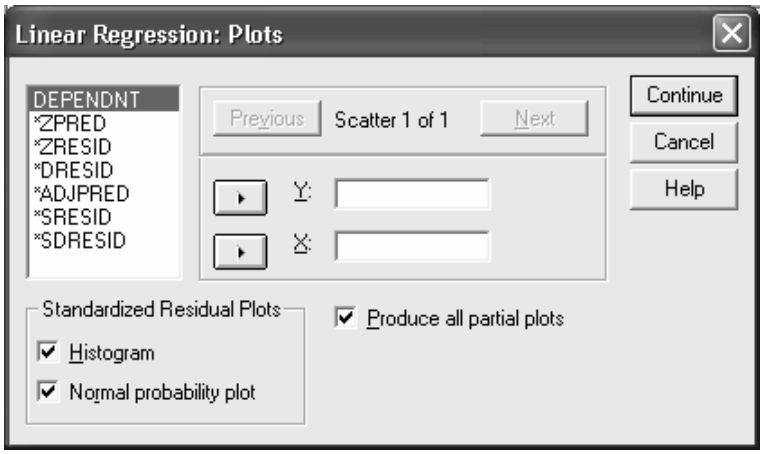


Рис. 7.52. Діалогове вікно задання параметрів графіків

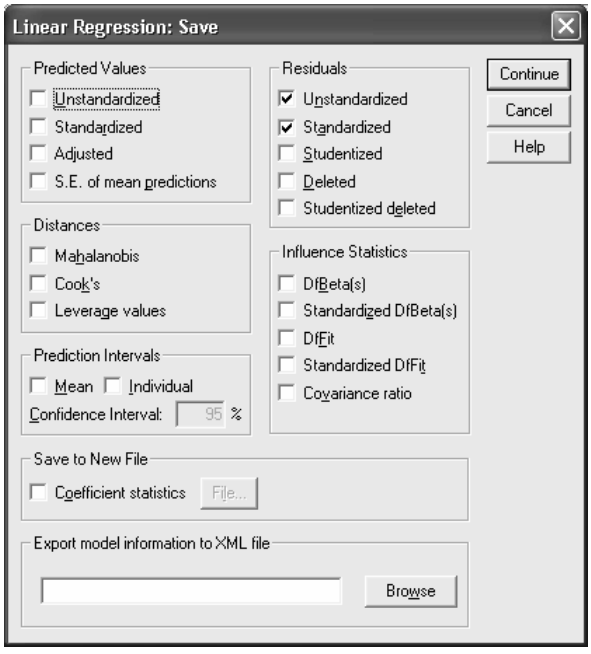


Рис. 7.53. Діалогове вікно задання величин, значення яких треба додати у вікні даних

У вікні Options (рис. 7.54) зазначаємо метод і параметри розрахункової процедури.

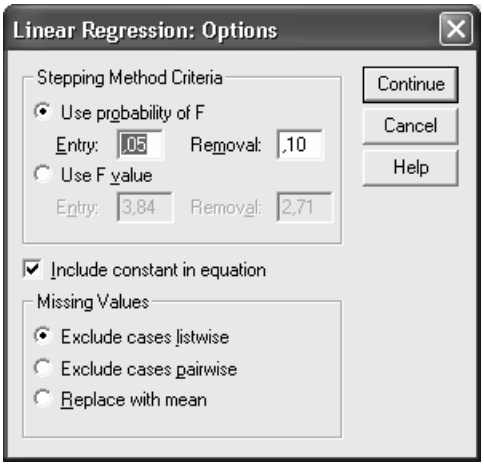


Рис. 7.54. Вікно задання методу і параметрів розрахункової процедури

Деякі результати підбору параметрів та аналізу багатофакторної лінійної моделі наведено на рис. 7.55–7.63.

Correlations

		VAR00006	VAR00001	VAR00002	VAR00003	VAR00004	VAR00005
Pearson Correlation	VAR00006	1,000	,712	,344	,480	-,664	-,312
	VAR00001	,712	1,000	,171	,097	-,192	-,237
	VAR00002	,344	,171	1,000	-,131	,045	,041
	VAR00003	,480	,097	-,131	1,000	-,276	-,057
	VAR00004	-,664	-,192	,045	-,276	1,000	,040
	VAR00005	-,312	-,237	,041	-,057	,040	1,000
Sig. (1-tailed)	VAR00006	.	,000	,007	,000	,000	,014
	VAR00001	,000	.	,118	,251	,090	,049
	VAR00002	,007	,118	.	,182	,377	,388
	VAR00003	,000	,251	,182	.	,026	,347
	VAR00004	,000	,090	,377	,026	.	,391
	VAR00005	,014	,049	,388	,347	,391	.
N	VAR00006	50	50	50	50	50	50
	VAR00001	50	50	50	50	50	50
	VAR00002	50	50	50	50	50	50
	VAR00003	50	50	50	50	50	50
	VAR00004	50	50	50	50	50	50
	VAR00005	50	50	50	50	50	50

Рис. 7.55. Кореляції між змінними

Зокрема з рис. 7.55 бачимо, що немає істотної кореляції між незалежними змінними VAR0001 – VAR0005.

Descriptive Statistics

	Mean	Std. Deviation	N
VAR00006	,3436	6,92320	50
VAR00001	,1572	1,13998	50
VAR00002	,0651	1,14300	50
VAR00003	-,1018	1,10601	50
VAR00004	-,0062	1,12834	50
VAR00005	,0939	1,19823	50

Рис. 7.56. Описова статистика для змінних

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	VAR00005, VAR00004, VAR00002, VAR00003, VAR00001[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: VAR00006

Рис. 7.57. Результати відбору незалежних змінних,
які враховуватимуться при побудові моделі

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	,999[a]	,998	,998	,28419	,998	5807,296	5	44	,000	1,537

a. Predictors: (Constant), VAR00005, VAR00004, VAR00002, VAR00003, VAR00001

b. Dependent Variable: VAR00006

Рис. 7.58. Загальна характеристика моделі

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2345,048	5	469,010	5807,296	,000[a]
	Residual	3,554	44	,081		
	Total	2348,601	49			

a. Predictors: (Constant), VAR00005, VAR00004, VAR00002, VAR00003, VAR00001

b. Dependent Variable: VAR00006

Рис. 7.59. Таблиця ANOVA побудованої моделі

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	3,091E-02	,041		,753	,456	-,052	,114
	VAR00001	2,970	,038	,489	77,836	,000	2,893	3,047
	VAR00002	2,020	,037	,333	55,150	,000	1,946	2,094
	VAR00003	2,077	,039	,332	53,798	,000	1,999	2,154
	VAR00004	-2,984	,038	-,486	-78,425	,000	-3,061	-2,908
	VAR00005	-,988	,035	-,171	-28,225	,000	-1,059	-,918

a. Dependent Variable: VAR00006

Рис. 7.60. Таблиця коефіцієнтів побудованої моделі

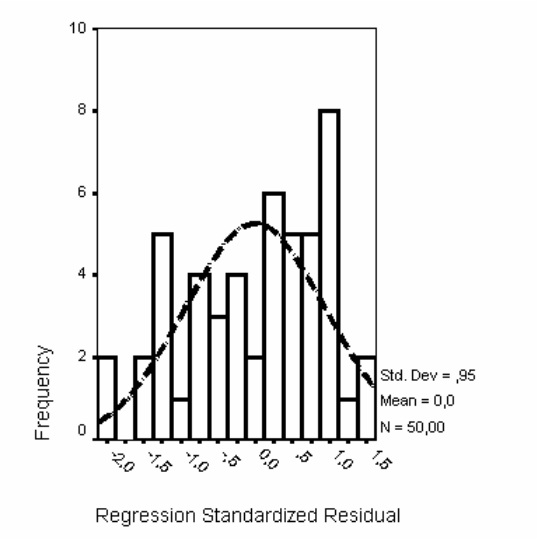


Рис. 7.61. Гістограма залишків

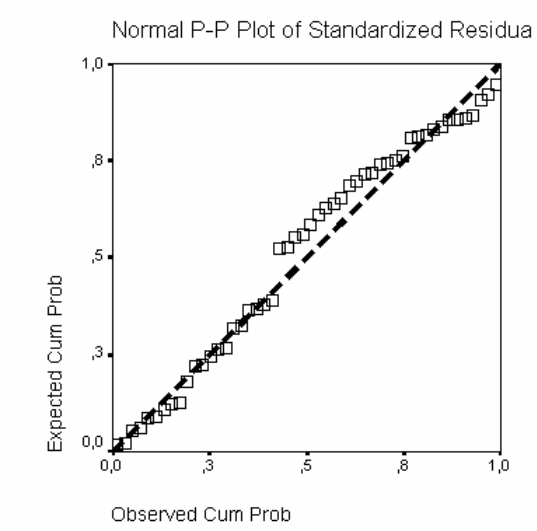


Рис. 7.62. P-P графік стандартизованих залишків моделі для нормального розподілу

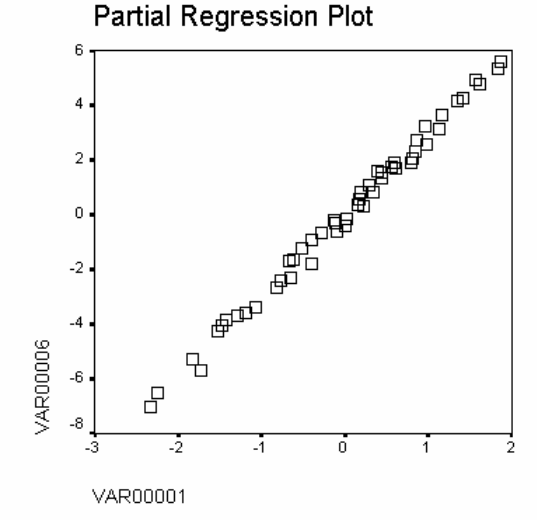


Рис. 7.63. Частинна регресія стосовно першої незалежної змінної

Побудована модель є адекватною, про що свідчить близькість отриманих значень коефіцієнтів, високий коефіцієнт детермінації й практична відсутність автокореляції залишків. Але гістограма залишків дещо відрізняється від нормального закону. Графіки частинної кореляції свідчать про гарне наближення до лінійного зв'язку за усіма незалежними змінними.

З наведених даних можна зробити висновок, що статистичний пакет SPSS надає більше можливостей для аналізу багатофакторних лінійних регресійних моделей, ніж електронні таблиці MS Excel, а одержувані за його допомогою результати є більш зручними для практичного використання.

Контрольні питання

1. Яким є основне завдання регресійного аналізу?
2. У чому полягають основні припущення класичного регресійного аналізу?
3. Якою є звичайна процедура класичного регресійного аналізу?
4. Як формулюється задача побудови регресійної моделі?
5. Які функціонали використовують для визначення параметрів регресійних моделей? У чому полягають переваги й недоліки різних типів таких функціоналів?
6. Якими є основні типи функцій, що використовуються для побудови однофакторних регресійних моделей?
7. Які моделі називають лінійними? Що називають порядком регресійної моделі?
8. Чому регресійні моделі не рекомендують використовувати поза межами тієї області значень вихідних параметрів, для якої вони побудовані?
9. Для заданого набору даних побудувати однофакторну лінійну регресійну модель і перевірити її адекватність.
10. У яких випадках нелінійні однофакторні моделі можна звести до лінійних? Навести приклади відповідних перетворень.
11. Для заданого набору даних побудуйте однофакторну нелінійну регресійну модель і перевірте її адекватність.
12. Як використовують критерій Фішера для перевірки адекватності регресійних моделей?
13. Як визначають довірчі інтервали для коефіцієнтів однофакторних регресійних моделей?
14. Яким є загальний вигляд поліноміальної регресійної моделі?
15. Яким є загальний алгоритм визначення порядку і параметрів поліноміальних регресійних моделей?
16. Для заданого набору даних побудуйте поліноміальну регресійну модель і перевірте її адекватність.

17. У яких випадках використовують регресійні моделі у вигляді тригонометричних поліномів? Яким є загальний алгоритм побудови таких моделей?

18. Для заданого набору даних побудуйте регресійну модель у вигляді тригонометричного поліному і перевірте її адекватність.

19. Якими є загальні алгоритми побудови однофакторних регресійних моделей у вигляді модифікованої показникової функції, кривої Гомперця та логістичної кривої?

20. Яким є загальний алгоритм побудови багатофакторної лінійної регресійної моделі?

21. Для заданого набору даних побудуйте багатофакторну лінійну регресійну модель і перевірте її адекватність.

22. Що називають мультиколінеарністю даних? Наведіть приклади.

23. Для чого застосовують алгоритми зміщеного оцінювання параметрів багатофакторних лінійних регресійних моделей? Наведіть приклади.

24. За якими властивостями перевіряють адекватність регресійних моделей? Якими є основні критерії адекватності?

ДОДАТОК А
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;

– розкладання у ряд Тейлора або Маклорена з наступним інтегруванням;

– кускова апроксимація елементарними функціями з наступним інтегруванням;

– апроксимація за допомогою нейронної сітки.

Пряме інтегрування використовують лише для дискретних розподілів. Відповідні алгоритми зазвичай потребують великого часу розрахунків і в багатьох випадках доцільнішою є апроксимація статистик критеріїв стандартними розподілами.

Пряме інтегрування можна здійснювати за формулами трапецій, Сімпсона, методами Монте-Карло, Гауса тощо.

Нехай потрібно обчислити певну функцію розподілу, яка в загальному вигляду може бути поданою як

$$P(x) = \int_{-\infty}^{x} f(t)\,dt. \qquad \text{(A.1)}$$

При цьому слід ураховувати такі властивості:

– для симетричних відносно $x = 0$ розподілів

$$P(x) = \frac{1}{2} + \int_{0}^{x} f(t)\,dt; \qquad \text{(A.2)}$$

– функція $f(t)$ досить швидко згасає 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 сіток 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.

Нормальний розподіл використовують при поданні випадкових даних, що результатами впливу великої кількості слабких впливів на об'єкт дослідження. Наприклад, якщо вимірювання певної величини здійснюють багато разів при однакових умовах, то результати спостережень будуть мати нормальний розподіл. Тому функцію нормального розподілу називають також **функцією похибок**. Нормальний розподіл є основним теоретичним видом розподілу імовірності, оскільки багато інших типів розподілу наближаються до нього за певних умов. Наприклад при необхідності здійснити апроксимацію дискретного розподілу неперервним використовують заміну біноміального розподілу нормальним з тими самими математичним сподіванням і дисперсією.

Функція щільності нормального розподілу має вигляд:

$$f\left(x\right)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{\left(x-m_x\right)^2}{2\sigma^2}\right],\ \mathrm{x}\in\left(-\infty,+\infty\right),\qquad\text{(A.3)}$$

де m_x – математичне сподівання, σ^2 – дисперсія. Шляхом уведення нормованого відхилення $t=\frac{x-m_x}{\sigma}$ її перетворюють до функції щільності стандартного нормального розподілу

$$f\left(t\right)=\frac{1}{\sqrt{2\pi}}\exp\left[-t^2/2\right].\qquad\text{(A.4)}$$

Функцію стандартного нормального розподілу (функцію Лапласа)

$$\Phi\left(x\right)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left[-y^2/2\right]dy\qquad\text{(A.5)}$$

зазвичай розраховують за допомогою апроксимаційних формул, або розкладанням у ряд. Якщо необхідно отримати підвищену точність результату (5−6 знаків) застосовують пряме інтегрування за методом Сімпсона.

На рис. А.1 наведено графіки функції нормального розподілу й відповідної функції щільності розподілу.

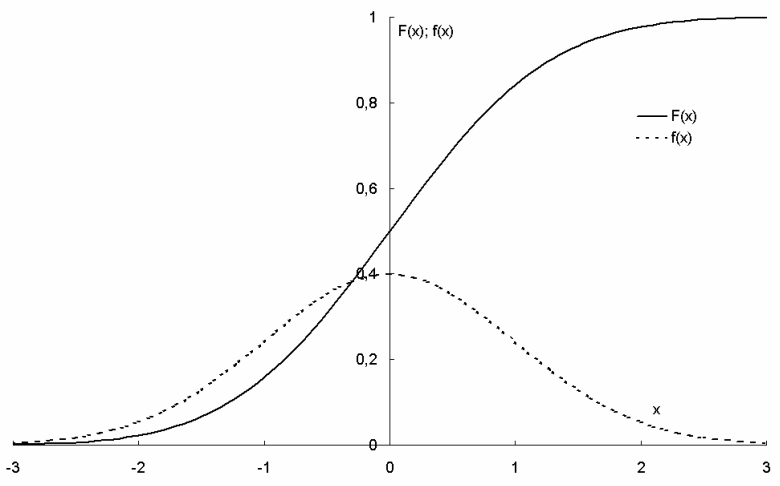


Рис. А.1. графіки функції нормального розподілу
й відповідної функції щільності розподілу

Функція щільності **логнормального розподілу** має вигляд:

$$f\left(x\right)=\frac{1}{\sqrt{2\pi}x\sigma}\exp\left(-\frac{\left[\ln\left(x/me\right)\right]^2}{2\sigma^2}\right), \tag{A.6}$$

де $me>0$ – медіана, $\sigma>0$ – параметр форми. Для логнормального розподілу математичне сподівання $\mu_x=me\sqrt{\omega}$, дисперсія $\sigma^2{}_x=me^2\omega(\omega–1)$, мода $m_x=me/\omega$, коефіцієнт асиметрії $As=\left(\omega+2\right)\sqrt{\omega-1}$, коефіцієнт ексцесу $\varepsilon=\omega^4+2\omega^3+3\omega^2-6$, коефіцієнт варіації $C_v=\sqrt{\omega-1}$, де $\omega=\exp(\sigma^2)$. Основні параметри логнормального розподілу можна оцінити по таким формулам:

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}\ln x_i\ ; \tag{A.7}$$

$$\widehat{me}=\exp\left(\hat{\mu}\right); \tag{A.8}$$

$$\widehat{\sigma^2}=\frac{1}{n-1}\sum_{i=1}^{n}\left(\ln x_i-\hat{\mu}\right)^2\ . \tag{A.9}$$

Щільність **Г-розподілу** (**гамма-розподілу**) визначається формулою:

$$f\left(x\right)=\begin{cases}\dfrac{\beta^{\alpha}}{\Gamma\left(\alpha\right)}x^{\alpha-1}e^{-\beta x},\ x\geq 0\ \left(\alpha,\beta>0\right);\\ 0,\ x<0,\end{cases} \tag{A.10}$$

де $\Gamma(\alpha)$ – гамма-функція, $\alpha>0$ – параметр форми, $\beta>0$ – параметр масштабу (іноді використовують параметр $1/\beta$).

Гамма функцією Ейлера називають функцію виду:

$$\Gamma(\alpha)=\int_0^{\infty} t^{\alpha-1}e^{-t}dt\ \ (x>0). \qquad (A.11)$$

Вона задовольняє співвідношення $\Gamma(x+1)=x\Gamma(x)$. $\Gamma(1)=1$. Звідси для цілих n маємо: $\Gamma(n+1)=n!(n=1,2,\ldots)$. Значення $\Gamma(1/2)=\sqrt{\pi}$ дає змогу отримати значення також для будь-якого напівцілого значення аргументу. Для інших значень аналітичного виразу для обчислення Г-функції не існує. Наближено її можна розрахувати за такими формулами. Для від'ємних значень аргументу $(1-x)$

$$\Gamma(1-x)=\frac{\pi}{\Gamma(x)\sin(\pi x)}=\frac{\pi x}{\Gamma(1+x)\sin(\pi x)}. \qquad (A.12)$$

Для довільного комплексного аргументу Г-функція має полюси у нулі, а також при всіх цілочислових від'ємних значеннях аргументу. Для побудови апроксимації вигляду

$$\Gamma(x+1)=\sqrt{2\pi}\left(x+\gamma+\frac{1}{2}\right)^{x+\frac{1}{2}}\exp\left(-x-\gamma-\frac{1}{2}\right)\left(A_\gamma(x)+\varepsilon\right) \qquad (A.13)$$

враховують кілька перших полюсів. Для цього будують функцію

$$A_\gamma(x)=c_0+\frac{c_1}{x+1}+\frac{c_2}{x+2}+\ldots+\frac{c_{\gamma+1}}{x+\gamma+1}. \qquad (A.14)$$

Для $\gamma=5$ похибка апроксимації для всіх точок, які знаходяться у першій та четвертій чвертях координатної площини, $|\varepsilon|<2\times10^{-10}$.

Для Г-розподілу математичне сподівання $\mu_x=\alpha/\beta$, дисперсія $\sigma^2{}_x=\alpha/\beta^2$, мода $m_x=(\alpha-1)/\beta(\alpha>1)$, коефіцієнт асиметрії $As_x=2\sqrt{\alpha}$, коефіцієнт ексцесу $6/\alpha$, коефіцієнт варіації $C_v=\sqrt{\alpha}$ Для оцінювання параметрів розподілу можна використовувати формули:

$$\hat{\alpha}=\left(\bar{x}/s\right)^2;\ \hat{\beta}=\bar{x}/s^2, \qquad (A.15)$$

де $s^2=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$ – вибіркова дисперсія без поправки.

Розподіл χ^2 є окремим випадком Г-розподілу з параметрами $\alpha=n/2$, $\beta=1/2$ (n – кількість степенів вільності). Іншим окремим випадком є розподіл Ерланга. Форма розподілу визначається кількістю степенів вільності, яке є параметром форми, (рис. А.2).

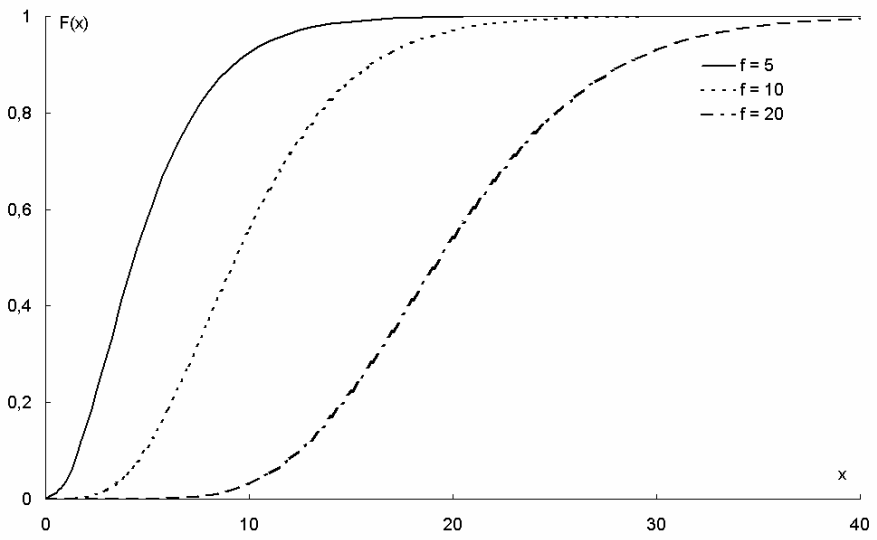


Рис. А.2. Графіки функції розподілу χ^2 для різних значень кількості степенів вільності

Функцію розподілу χ^2 обчислюють за формулою:

$$F_n(x) = \frac{(1/2)^{n/2}}{\Gamma(n/2)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}. \qquad (A.16)$$

Якщо $n = 2$, розподіл χ^2 збігається з експоненціальним розподілом.

При розрахунку значень функції розподілу інтегрування можна здійснювати за формулою трапецій. При цьому для отримання прийнятної точності кількість інтервалів розбиття для малих n має бути достатньо великою. Параметри точності та швидкодії можуть бути підвищені шляхом застосування більш досконалих формул інтегрування або розкладанням підінтегральної функції у ряд.

Математичне сподівання розподілу χ^2 $\mu_x = n$, дисперсія $\sigma^2_x = 2n$, мода $m_x = n - 2$ $(n \geq 2)$, коефіцієнт асиметрії $As_x = \frac{2^{\frac{3}{2}}}{\sqrt{n}}$, коефіцієнт ексцесу $\varepsilon = 12/n$, коефіцієнт варіації $C_v = \sqrt{2/n}$.

Щільність B-розподілу (бета-розподілу) визначається формулою:

$$f(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \qquad (A.17)$$

де

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx \qquad (A.18)$$

– B-фунція Ейлера, $\alpha > 0$, $\beta > 0$ – параметри форми. Для вибірок великого обсягу розрахунки здійснюють за асимптотичними формулами.

Для B-розподілу математичне сподівання $\mu_x = \alpha/(\alpha + \beta)$, дисперсія $\sigma^2_x = \alpha\beta/(\alpha + \beta)^2(\alpha + \beta + 1)$, мода $m_x = \dfrac{\alpha - 1}{\alpha + \beta - 1}$ $(\alpha > 1,\ \beta > 1)$, коефіцієнт варіації $C_v = \sqrt{\dfrac{\beta}{\alpha(\alpha + \beta + 1)}}$, коефіцієнт асиметрії $\dfrac{2(\beta - \alpha)\sqrt{\alpha + \beta + 1}}{(\alpha + \beta + 2)\sqrt{\alpha\beta}}$, коефіцієнт ексцесу $\varepsilon = 6\dfrac{\alpha^3 - \alpha^2(2\beta - 1) + \beta^2(\beta + 1) - 2\alpha\beta(\beta + 2)}{\alpha\beta(\alpha + \beta + 2)(\alpha + \beta + 3)}$.

Для оцінювання параметрів B-розподілу можна використовувати такі формули:

$$\hat{\alpha} = \bar{x}\left[\frac{\bar{x}(1 - \bar{x})}{s^2} - 1\right]; \tag{A.19}$$

$$\hat{\beta} = (1 - \bar{x})\left[\frac{\bar{x}(1 - \bar{x})}{s^2} - 1\right], \tag{A.20}$$

де s^2 – вибіркова дисперсія без поправки.

Функція щільності ***F*-розподілу** (**розподілу Фішера**, **розподілу дисперсійного відношення**) має вигляд:

$$f(x) = \frac{\Gamma\left(\frac{1}{2}(v + w)\right)\left(\frac{v}{w}\right)^{v/2} x^{(v-2)/2}}{\Gamma(v/2)\Gamma(w/2)\left(1 + \frac{vx}{w}\right)^{\frac{v+w}{2}}}, \tag{A.21}$$

де v, w – цілі додатні числа (кількості степенів вільності).

Вираз для функції розподілу також можна записати через *B*-функцію. При чисельних розрахунках зазвичай використовують саме такий запис або розклад в ряд.

Основні параметри *F*-розподілу є такими:

– математичне сподівання $\mu_x = \dfrac{w}{w - 2}$, $w > 2$;

– дисперсія $\sigma^2_x = \dfrac{2w^2(v + w - 2)}{v(w - 2)^2(w - 4)}$, $w > 4$;

– мода $m_x = \dfrac{w(v-2)}{v(w+2)}$, $v>1$;

– коефіцієнт асиметрії $As_x = \dfrac{(2v+w-2)\sqrt{8(w-4)}}{(w-6)\sqrt{v+w-2}}$, $w>6$;

– коефіцієнт варіації $\dfrac{2(v+w-2)}{v\sqrt{w-4}}$, $w>4$.

Щільність ***t*-розподілу Стьюдента** обчислюють за формулою:

$$f(x) = \frac{\Gamma\left(\dfrac{\alpha+1}{2}\right)}{\sqrt{\pi\alpha}\,\Gamma(\alpha/2)}\left(1+\frac{x^2}{\alpha}\right)^{-(\alpha+1)/2}, \qquad \text{(A.22)}$$

де α – параметр форми (ціле додатне число).

При $n\to\infty$ розподіл Стьюдента наближається до нормального. Зазвичай вважають, що його апроксимація нормальним розподілом є прийнятною при $n \geq 30$. При $n = 1$ t-розподіл Стьюдента називають розподілом Коші. Розрахунок функції розподілу можна здійснювати за формулою трапецій або шляхом розкладання у ряд.

Математичне сподівання, мода, коефіцієнти асиметрії та ексцесу розподілу Стьюдента дорівнюють нулю, дисперсія $\sigma_x^2 = \dfrac{\alpha}{\alpha-2}$, $\alpha>2$, середнє відхилення $d_x = \dfrac{\sqrt{\alpha}\,\Gamma\left(\dfrac{\alpha-1}{2}\right)}{\sqrt{\pi}\,\Gamma(\alpha/2)}$.

Багатовимірним узагальненням розподілу Стьюдента є ***T*²-розподіл Хотеллинга**, щільність якого задається формулою:

$$f(x) = \begin{cases} \dfrac{2\Gamma\left(\dfrac{n-q+1}{2}\right)}{(n-q)^{k/2}\,\Gamma\left(\dfrac{k}{2}\right)\Gamma\left(\dfrac{n-q-k+1}{2}\right)} \dfrac{x^{k-1}}{\left(1+\dfrac{x^2}{n-q}\right)^{\frac{n-q+1}{2}}}, & x\in[0,+\infty); \\ 0, & x\in(-\infty,0), \end{cases} \qquad \text{(A.23)}$$

де k – розмірність багатовимірної вибірки; q – кількість лінійних умов. У практиці цю функцію розраховують на основі функції F-розподілу.

Розподіл критичних значень ***G*-критерію Кокрена** визначається за точною формулою

$$g_{k,v,1-\alpha} = \frac{F}{(k-1)+F}, \qquad \text{(A.24)}$$

де k – кількість вибірок; v – обсяг кожної з них; α – рівень значущості; F – значення оберненої функції F-розподілу для v та $(k-1)v$ ступенів вільності й довірчого рівня $\left(1-\frac{\alpha}{k}\right)$.

При $k>2$, $v>10$ можна використовувати також апроксимаційну формулу

$$g_{k,\nu,1-\alpha}=\frac{2x}{\nu(2k-1)-2+x+\frac{(2-\nu)(2+\nu+x)+2x^2}{6(\nu(2k-1)-2)}}, \qquad (A.25)$$

де x – значення оберненої функції χ^2 – розподілу для ν ступенів вільності і довірчого рівня $\left(1-\frac{\alpha}{k}\right)$. У певних випадках апроксимаційна формула дає точніші результати, ніж "точна", оскільки результати останньої залежать від точності розрахунку оберненої функції F-розподілу.

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

$$K(\lambda)=\begin{cases}\sum_{i=-\infty}^{+\infty}(-1)^i e^{-2i^2\lambda^2}, \lambda>0;\\ 0,\ \lambda\le 0.\end{cases} \qquad (A.26)$$

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.

Функція **розподілу Вейбула** дається формулою:

$$F(x)=1-\exp\left(-(x/b)^c\right), \qquad (A.27)$$

де $b>0$ – параметр масштабу (характерний час життя), $c>0$ – параметр форми. Функція щільності розподілу має вигляд:

$$f(x)=\frac{cx^{c-1}}{b^c}\exp\left(-(x/b)^c\right). \qquad (A.28)$$

Математичне сподівання розподілу Вейбула $\mu_x=b\Gamma\left(\frac{c+1}{c}\right)$, дисперсія $\sigma_x^2=b^2\left[\Gamma\left(\frac{c+2}{c}\right)-\left\{\Gamma\left(\frac{c+1}{c}\right)\right\}^2\right]$, мода $m_x=b\left(1-\frac{1}{c}\right)^{1/c}$ при $c\ge 1$ и 0 при $c\le 1$, коефіцієнт варіації $C_v=\sqrt{\frac{\Gamma\left(\frac{c+2}{c}\right)}{\left[\Gamma\left(\frac{c+1}{c}\right)\right]^2}-1}$.

ЛІТЕРАТУРА

1. Агекян Т.А. Основы теории ошибок для астрономов и физиков / Т.А. Агекян. – М. : Наука, 1972. – 172 с.

2. Айвазян С.А. Прикладная статистика. Основы моделирования и первичная обработка данных / С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин. – М. : Финансы и статистика, 1983. – 471 с.

3. Айвазян С.А. Прикладная статистика. Исследование зависимостей / С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин. – М. : Финансы и статистика, 1985. – 487 с.

4. Айвазян С.А. Прикладная статистика и основы эконометрики / С.А. Айвазян, В.С. Мхитарян. – М. : ЮНИТИ, 1998. – 1022 с.

5. Андерсен Т. Введение в многомерный статистический анализ / Т. Андерсен. – М. : Физматгиз, 1963. – 500 с.

6. Аптон Г. Анализ таблиц сопряженности / Г. Аптон. – М. : Финансы и статистика, 1982. – 144 с.

7. Аренс Х. Многомерный дисперсионный анализ / Х. Аренс, Ю. Лейтер. – М. : Финансы и статистика, 1985. – 231 с.

8. Бард Й. Нелинейное оценивание параметров / Й. Бард. – М. : Финансы и статистика, 1979. – 349 с.

9. Бахрушин В.Є. Математичне моделювання / В.Є. Бахрушин. – Запоріжжя : ГУ “ЗІДМУ”, 2003. – 138 с.

10. Бахрушин В.Є. Аналіз даних : навчальний посібник / В.Є. Бахрушин. – Запоріжжя : ГУ “ЗІДМУ”, 2006. – 128 с.

11. Бахрушин В.Є. Математичні основи моделювання систем : навчальний посібник / В.Є. Бахрушин. – Запоріжжя : КПУ, 2009. – 224 с.

12. Бахрушин В.Є. Дослідження властивостей оцінок дисперсії, отриманих за груповим методом / В.Є. Бахрушин // Складні системи і процеси. – 2006. – № 1. – С. 3–7.

13. Бахрушин В.Е. Применение статистических методов при обработке результатов производственного контроля в металлурги полупроводников / В.Е. Бахрушин, М.А. Игнахина // Системні технології. – 2008. – № 3 (56). – Т. 1. – С. 3–7.

14. Бахрушин В.Е. Эмпирические функции распределения результатов тестирования выпускников школ / В.Е. Бахрушин, С.В. Журавель, М.А. Игнахина // Управляющие системы и машины. – 2009. – № 2. – С. 82–84.

15. Бендат Дж. Применение корреляционного и спектрального анализа / Дж. Бендат, А. Пирсол. – М. : Мир, 1979. – 311 с.

16. Бендат Дж. Прикладной анализ случайных данных / Дж. Бендат, А. Пирсол. – М. : Мир, 1989. – 540 с.

17. Большаков А.А. Методы обработки многомерных данных и временных рядов / А.А. Большаков, Р.Н. Каримов. – М. : Горячая линия – Телеком, 2007. – 522 с.

18. Боровиков В.П. Популярное введение в программу STATISTICA / В.П. Боровиков. – М. : Компьютер-Пресс, 1998. – 267 с.

19. Брандт З. Анализ данных: Статистические и вычислительные методы для научных работников и инженеров / З. Брандт. – М. : Мир : АСТ, 2003. – 686 с.

20. Вуколов Э.А. Основы статистического анализа. Практикум по статистическим методам и исследованию операций с использованием пакетов STATISTICA и EXCEL / Э.А. Вуколов. – М. : ФОРУМ : ИНФРА-М, 2004. – 464 с.

21. Вучков И. Прикладной линейный регрессионный анализ / И. Вучков, Л. Бояджиева, Е. Солаков. – М. : Финансы и статистика, 1987. – 239 с.

22. Гаек Я. Теория ранговых критериев / Я. Гаек, З. Шидак. – М. : Наука, 1971. – 376 с.

23. Гайдышев И. Анализ и обработка данных : специальный справочник / И. Гайдышев. – СПб. : Питер, 2001. – 752 с.

24. Гирко В.Л. Многомерный статистический анализ / В.Л. Гирко. – К. : Высшая школа, 1988. – 320 с.

25. Гмурман В.Е. Теория вероятностей и математическая статистика / В.Е. Гмурман. – М. : Высшая школа, 2003. – 479 с.

26. Горянский В.Ф. Математико-статистические методы в анализе эффективности сельскохозяйственного производства / В.Ф. Горянский. – К. : Вища школа, 1980. – 176 с.

27. Дрейпер Н. Прикладной регрессионный анализ : в 2 т. / Н. Дрейпер, Г. Смит. – М. : Финансы и статистика, 1986. – Т. 1. – 366 с.; 1987. – Т. 2. – 351 с.

28. Дубров А.М. Многомерные статистические методы / А.М. Дубров, В.С. Мхитарян, Л.И. Трошин. – М. : Финансы и статистика, 1998. – 352 с.

29. Дюк В.А. Компьютерная психодиагностика / В.А. Дюк. – СПб., 1994. – 364 с.

30. Дюран Б. Кластерный анализ / Б. Дюран, П. Оделл. – М. : Статистика, 1977. – 125 с.

31. Енюков И.С. Методы, алгоритмы, программы многомерного статистического анализа / И.С. Енюков. – М. : Финансы и статистика, 1986. – 232 с.

32. Іващенко П.О. Багатовимірний статистичний аналіз / П.О. Іващенко, І.В. Семеняк, В.В. Іванов. – Х. : Основа, 1992. – 144 с.

33. Иберла К. Факторный анализ / К. Иберла. – М. : Статистика, 1980. – 398 с.

34. Кендалл М. Многомерный статистический анализ и временные ряды / М. Кендалл, А. Стюарт. – М. : Наука, 1976. – 511 с.

35. Кендалл М.Дж. Статистические выводы и связи / М.Дж. Кендалл, А. Стюарт. – М. : Наука, 1973. – 899 с.

36. Кендалл М.Дж. Теория распределений / М.Дж. Кендалл, А. Стюарт. – М. : Наука, 1966. – 566 с.

37. Кобзарь А.И. Прикладная математическая статистика / А.И. Кобзарь. – М. : Физматлит, 2006. – 816 с.

38. Королюк В.С. Асимптотический анализ распределений статистик / В.С. Королюк, Ю.В. Боровских. – К. : Наукова думка, 1984. – 301 с.

39. Лагутин М.Б. Наглядная математическая статистика / М.Б. Лагутин. – М. : БИНОМ, 2007. – 472 с.

40. Лесникович А.И. Корреляции в современной химии / А.И. Лесникович, С.В. Левчик. – Минск : Университетское, 1989. – 118 с.

41. Литтл Р.Дж. Статистический анализ данных с пропусками / Р.Дж. Литтл, Д.Б. Рубин. – М. : Финансы и статистика, 1991. – 336 с.

42. Мандель И.Д. Кластерный анализ / И.Д. Мандель. – М. : Финансы и статистика, 1988.

43. Многомерный статистический анализ в экономике / [под ред. В.Н. Тамашевича]. – М. : ЮНИТИ, 1999. – 600 с.

44. Новицкий П.В. Оценка погрешностей результатов измерений / П.В. Новицкий, И.А. Зограф. – Л. : Энергоатомиздат, 1991. – 304 с.

45. Орлов А.И. Прикладная статистика / А.И. Орлов. – М. : Экзамен, 2006. – 671 с.

46. Перегудов Ф.И. Введение в системный анализ / Ф.И. Перегудов, Ф.П. Тарасенко. – М. : Высшая школа, 1989. – 367 с.

47. Плюта В. Сравнительный многомерный анализ в эконометрическом моделировании / В. Плюта. – М. : Финансы и статистика, 1989. – 175 с.

48. Прикладная статистика. Классификация и снижение размерности / [С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин]. – М. : Финансы и статистика, 1989. – 607 с.

49. Протасов К.В. Статистический анализ экспериментальных данных / К.В. Протасов. – М. : Мир, 2005. – 142 с.

50. Себер Дж. Линейный регрессионный анализ / Дж. Себер. – М. : Мир, 1980. – 456 с.

51. Тьюки Дж. Анализ результатов наблюдений. Разведочный анализ / Дж. Тьюки. – М. : Мир, 1981. – 693 с.

52. Тюрин Ю.Н. Анализ данных на компьютере / Ю.Н. Тюрин, А.А. Макаров. – М. : ИНФРА-М, 2003. – 544 с.

53. Факторный, дискриминантный и кластерный анализ / [Дж.-О. Ким, Ч.У. Мьюллер, У.Р. Клекка]. – М. : Финансы и статистика, 1989. – 216 с.

54. Химмельбау Дж. Анализ процессов статистическими методами / Дж. Химмельблау. – М. : Мир, 1973. – 957 с.

55. Холлендер М. Непараметрические методы статистики / М. Холлендер, Д. Вульф. – М. : Финансы и статистика, 1983. – 518 с.

56. Хьюбер П. Робастность в статистике / П. Хьюбер. – М. : Мир, 1984. – 304 с.

57. Худсон Д. Статистика для физиков / Д. Худсон. – М. : Мир, 1967. – 242 с.

58. Шитиков В.К. Количественная гидроэкология: Методы системной идентификации / В.К. Шитиков, Г.С. Розенберг, Т.Д. Зинченко. – Тольятти : ИЭВБ РАН, 2003. – 463 с.

59. Эфрон Б. Нетрадиционные методы многомерного статистического анализа / Б. Эфрон. – М. : Финансы и статистика, 1988. – 262 с.

ПРЕДМЕТНИЙ ПОКАЖЧИК

Л

М

Н

О

П

Р

С

Т

Ф

Х

Ц

Ч

Ш

ІМЕННИЙ ПОКАЖЧИК

Навчальне видання

Бахрушин Володимир Євгенович

МЕТОДИ АНАЛІЗУ ДАНИХ

Навчальний посібник

Редактор Г.І. Гутман
Коректор Я.О. Рекубрацька
Технічний редактор Т.В. Безденежна

Підписано до друку 20.01.2011.
Формат 60x84/16. Гарнітура Times
Ум.-друк. арк. 15,57. Обл.-вид. арк. 16,01. Тираж 300 прим. Зам. № 67-10Н.

Видавець та виготовлювач
Класичний приватний університет
69002, м. Запоріжжя, вул. Жуковського, 70-б

Свідоцтво суб'єкта видавничої справи
серія ДК, № 3321 від 25.11.2008 р.