

**Міністерство освіти і науки України
Національний авіаційний університет
Кафедра прикладної математики**

П.О. Приставка, О.М.Мацуга

АНАЛІЗ ДАНИХ

Електронний посібник для студентів
спеціальності *«прикладна математика»*

Київ

2010

ЗМІСТ

ВСТУП.....	5
1. ОБРОБКА Й АНАЛІЗ ОДНОВИМІРНИХ ДАНИХ.....	9
1.1. Первинний статистичний аналіз.....	9
1.1.1. Формування варіаційного ряду.....	9
1.1.2. Гістограмна оцінка.....	10
1.1.3. Точкові та інтервальні оцінки.....	18
1.2. Відтворення розподілів.....	26
1.2.1. Методи оцінки параметрів розподілу.....	27
1.2.2. Оцінювання точності оцінок параметрів.....	32
1.2.3. Інтервальне оцінювання теоретичної функції розподілу.....	33
1.2.4. Параметричні розподіли.....	34
Контрольні запитання та завдання.....	46
2. ПЕРЕВІРКА СТАТИСТИЧНИХ ГІПОТЕЗ.....	48
2.1. Головні поняття та визначення.....	48
2.2. Оцінка згоди відтворення розподілів.....	56
2.3. Задача двох вибірок.....	58
2.4. Перевірка збігу середніх.....	59
2.5. Перевірка збігу дисперсій.....	61
2.6. Однофакторний дисперсійний аналіз.....	63
2.7. Критерії порядкових статистик.....	64
Контрольні запитання та завдання.....	67
3. ОБРОБКА Й АНАЛІЗ ДВОВИМІРНИХ ДАНИХ.....	69
3.1. Первинний аналіз.....	69
3.2. Кореляційний аналіз.....	75
3.2.1. Парна кореляція.....	75
3.2.2. Кореляційне відношення.....	78
3.2.3. Парна рангова кореляція.....	80
3.2.4. Коефіцієнти сполучень таблиць.....	83
3.3. Одновимірний регресійний аналіз.....	89
3.3.1. Лінійний регресійний аналіз.....	89
3.3.2. Нелінійний регресійний аналіз.....	105
Контрольні запитання та завдання.....	112

4. ОБРОБКА Й АНАЛІЗ БАГАТОВИМІРНИХ ДАНИХ	114
4.1. Первинний аналіз	114
4.2. Перевірка гіпотез про збіг параметрів багатовимірних даних	119
4.3. Часткові та множинні коефіцієнти кореляції	123
4.4. Основи багатовимірного регресійного аналізу	127
4.5. Компонентний та факторний аналіз	137
4.5.1. Основи компонентного аналізу	137
4.5.2. Розвідницький факторний аналіз	143
Контрольні запитання та завдання	151
5. ОСНОВИ РОЗПІЗНАВАННЯ ОБРАЗІВ	153
5.1. Кластерний аналіз	154
5.1.1. Відстані між об'єктами та кластерами	154
5.1.2. Ієрархічні методи кластеризації	160
5.1.3. Метод К-середніх	163
5.1.4. Оцінка якості кластеризації	165
5.2. Дискримінантний аналіз	167
Контрольні запитання та завдання	174
6. ОСНОВИ АНАЛІЗУ ВИПАДКОВИХ ПРОЦЕСІВ ТА ЧАСОВИХ РЯДІВ	176
6.1. Характеристики випадкового процесу	177
6.2. Визначення та оцінка спектральної щільності	182
6.3. Первинний аналіз часових рядів та випадкових процесів	187
6.3.1. Вилучення аномальних спостережень	187
6.3.2. Ідентифікація тренду процесу	188
6.3.3. Згладжування даних	194
6.3.4. Вилучення поліноміального тренду	200
Контрольні запитання та завдання	203
Додаток А. Процедури знаходження квантилів	204
Додаток Б. Статистичні таблиці	206
Додаток В. Приклади завдань до лабораторних робіт	212
СПИСОК РЕКОМЕНДОВАНОЇ ЛІТЕРАТУРИ	218

ВСТУП

Статистичний аналіз – наука, яка вивчає оточуючий матеріальний світ. Усе, що піддається пізнанню, – предмети, явища, природні чи соціальні процеси – є об'єкт дослідження статистичного аналізу. Дано більш формалізоване визначення поняття **об'єкта**, від якого спробуємо простежити логіку статистичного аналізу як науки.

Об'єкт – це те, що має визначення, дане в результаті спостереження й аналізу.

Один і той же об'єкт може мати різні визначення залежно від характеру спостережень та глибини аналізу. Характер спостережень зумовлюється набором вимірних ознак об'єкта, аналіз – переліком методів інтерпретації реалізацій ознак та висновками.

Статистичний аналіз займається дослідженням об'єктів на основі одержаної в результаті спостереження інформації. Будь-яку зареєстровану інформацію називають **даними**. Залежно від характеру спостережень (кількості вимірних ознак об'єкта) розрізняють одновимірні, двовимірні та багатовимірні дані. Зі збільшенням вимірності даних зростає перелік методів аналізу, спрямованих на опрацювання окремих ознак, їх взаємодії та наслідків такої взаємодії.

Одновимірні набори даних (одна змінна) містять інформацію лише про одну ознаку об'єкта. Такий набір дає можливість знайти типові значення та характеристики варіабельності даних, а також виділити специфічні особливості або аномалії в даних.

Двовимірні дані на додаток до інформації про кожну зі змінних дозволяють вивчити зв'язок між двома ознаками та обчислити значення однієї змінної на основі іншої.

Багатовимірні дані, крім того, дозволяють встановлювати значення однієї змінної на основі значень інших.

Дані, що реєструються як числа, називають **кількісними**. **Дискретна** кількісна змінна може набувати значень тільки з деякого списку конкретних натуральних чисел (0 чи 1 або 1, 2, 3, ...). Кількісну змінну, яка є дійсна, називають **неперервною**.

Прикладом дискретних даних може бути кількість: мікроавтобусів на маршруті; відвідань кінотеатру за місяць, боргів на останній день сесії тощо.

Прикладами неперервних даних є: зріст групи людей; діаметр підшипників (у міліметрах); результати змагань у забігу на 100-метрівці.

Якщо змінна містить інформацію про те, до якої з декількох нечислових категорій належить об'єкт, то вона називається **якісною**. Якщо категорії можна

впорядкувати за змістом, то мова йде про **порядкову** (ординарну) якісну змінну, за відсутності ж такого порядку говорять про **номінальні** дані.

Приклади порядкових даних такі: військові звання (рядовий, сержант, лейтенант, майор, полковник); відповіді на питання анкети («Ставлення до навчання: люблю вчитися; не дуже люблю вчитися; не люблю, але змушую себе; не люблю, тому вчуся, поки не виженуть»).

Як приклади номінальних даних можна розглядати: райони міста; найменування товарів у мережі магазинів предмети, що їх вивчають студенти (математичний аналіз, програмування, філософія).

До кількісних даних можна застосовувати ті самі операції, що й до звичайних чисел: підрахунок частот, ранжування, арифметичні дії, для порядкових – ранжування та підрахунок частот, для номінальних – лише підрахунок частот.

Нарешті, якщо послідовність запису даних має певний сенс, то відповідний набір являє собою **часовий ряд**. Якщо ж послідовність запису даних не важлива, то маємо справу з даними, що містять інформацію про **часовий зріз**.

У термінології статистичного аналізу максимально повна інформація про об'єкт дослідження має назву генеральної сукупності Ω . Звичайно з різних причин доступ до такої інформації обмежений, тому як спостереження використовують деяку підмножину, випадковим чином сформовану з елементів генеральної сукупності. Таку підмножину називають вибіркою Ω_N , зокрема:

$$\Omega_N \subset \Omega,$$

$$\Omega_N = \{x_1, \dots, x_N\} \text{ або } \Omega_{1,N} = \{x_l; l = \overline{1, N}\}.$$

У позначенні $\Omega_{1,N}$ індекс «1» вказує на те, що вибірка одновимірна, тобто випадкова величина $\xi(\omega)$ має відбиття в R_1 .

Формально мають місце такі визначення.

Генеральною сукупністю Ω називають простір усіх елементарних подій.

Вибірка є частина елементарних подій, випадковим чином вибраних із генеральної сукупності. Вибірку називають **репрезентативною**, якщо вона відображає всі властивості генеральної сукупності.

Функція вибірки τ , або **статистика**, – це показник (число), обчислений за даними вибірки:

$$\tau = \varphi(x_1, \dots, x_N).$$

Статистика τ являє собою випадкову величину, оскільки в її основі лежать вибіркові дані й по суті вона є функцією від випадкової величини ξ , тому їй притаманні всі властивості випадкових величин.

За визначенням статистика є результат будь-якого обчислювального перетворення над даними вибірки, проте прагнуть одержати такі статистики, що можуть мати змістову інтерпретацію відносно об'єкта дослідження або аналізу, який проводиться.

Головними етапами статистичного аналізу є :

- 1) планування досліджень, результати яких можуть бути подані у вигляді випадкової вибірки;
- 2) попереднє дослідження даних, що дозволяє в подальшому аналізі адекватно оцінити статистичні характеристики;
- 3) оцінка невідомих числових величин та функцій, яка базується на вихідних даних;
- 4) перевірка статистичних гіпотез, що дозволяє на основі вибірових даних оцінити невизначеність у виборі характеристик простору $\langle \Omega_{1,N}, \mathcal{A}, P_N \rangle$.

Попереднє дослідження даних включає ряд обчислювальних процедур, основні з яких: формування варіаційних рядів та гістограм, редагування даних (як приклад – вилучення аномальних значень), ідентифікація типів розподілів тощо.

Головною задачею дисципліни «Аналіз даних» є перехід від статистичних оцінок та висновків до формальної моделі об'єкта та її інтерпретації. Формальна модель описується математичною залежністю (математичною моделлю).

За певних обставин можна ототожнювати об'єкт та **модель об'єкту**, проте, слід зважати на факт нескінченності варіантів визначень об'єкту, що, у свою чергу, визначає нескінченність процесу пізнання. У будь-якому випадку, при наданні визначення прагнуть надати таке, що однозначно формує тип реакції свідомості, як суб'єкту пізнання, на об'єкт. Якщо модель не забезпечує однозначного реагування на об'єкт, маємо справу з ситуативною невизначеністю, тобто мова йде про різні типи реагування, кожен з яких, так, чи інакше, має відповідну ймовірність.

Математичною моделлю будемо називати визначення, дане в термінах науки математики, при цьому визначення має задовольняти аксіомам математики. Чим більш вдало підібрано математичну модель, тим краще вона відбиває характерні риси об'єкту і тим адекватніші будуть реакції на об'єкт.

Слід зазначити, що вимоги до математичної моделі неоднозначні. З одного боку модель повинна бути достатньо повною – в ній мають бути враховані усі важливі фактори, від яких суттєво залежить повнота пізнання об'єкту. З іншого боку, модель має бути досить простою, щоб можна було б встановлювати залежності (бажано аналітичні) між параметрами, що в неї входять (тут під параметром моделі будемо розуміти визначення тієї, чи іншої характеристики об'єкту).

Автори підручника вбачають основною відмінною рисою дисципліни «Аналіз даних» від традиційного викладення курсу «Математична статистика» саме в прикладному аспекті. Викладення матеріалу спрямоване на формування у студентів сприйняття знань з паралельним їх закріпленням під час створення

інформаційних технологій для автоматизованих систем обробки та аналізу експериментальних даних із використанням сучасних комп'ютерних засобів. Акцентується увага на конкретних задачах (перетворення даних, вилучення аномальних, перехід до незалежних ознак, тощо), вирішення яких дозволяє значно підвищити адекватність кінцевих висновків стосовно об'єкту спостережень. Самі ж обчислювальні схеми методів статистичного оцінювання підбирались з умови одночасної ефективності їх застосування, обчислювальної простоти та можливості реалізації у сучасних програмних середовищах.

Автори вважають, що наведені в підручнику теоретичні та практичні матеріали знайдуть свою реалізацію не тільки в задачах навчального процесу, але й будуть корисні аспірантам, інженерам, економістам, бізнесменам та іншим спеціалістам, які працюють в області обробки статистичної інформації.

Розділ 1. ОБРОБКА Й АНАЛІЗ ОДНОВИМІРНИХ ДАНИХ

Статистичний аналіз одновимірних даних вимагає проведення первинного статистичного аналізу, що є необхідною складовою етапу попереднього дослідження даних, та розв'язання статистичної задачі відтворення функції розподілу.

1.1. Первинний статистичний аналіз

Розглянемо обчислювальні процедури первинного статистичного аналізу, такі як: формування варіаційних рядів та гістограм, вилучення аномальних значень, обчислення статистичних характеристик. Дано визначення понять параметра та оцінки параметра.

1.1.1. Формування варіаційного ряду

Нехай задана вибірка (масив даних) $\Omega_{1,N} = \{x_l; l = \overline{1, N}\}$, де x_l – результати спостережень реалізації випадкової величини ξ .

Побудова варіаційного ряду потребує ранжування результатів спостережень та обчислення відповідних їм частот і відносних частот:

$$\begin{array}{cccc} x_1, & x_2, & \dots & x_r \\ n_1, & n_2, & \dots & n_r \\ p_1, & p_2, & \dots & p_r, \end{array}$$

де

x_l – **варіанта** варіаційного ряду (тобто результат спостереження з вибірки, що не повторюється);

$x_i < x_j$, якщо $i < j$;

r – кількість варіант;

n_l – **частота** x_l , $\sum_{l=1}^r n_l = N$;

$$p_l = \frac{n_l}{N}$$

– **відносна частота** x_l ,

$$\sum_{l=1}^r p_l = 1.$$

Приклад 1.1. Нехай є вибірка $\Omega_{1,10} = \{5, 2, 1, 3, 2, 8, 4, 5, 3, 2\}$. Відповідний варіаційний ряд матиме вигляд

x_l :	1	2	3	4	5	8
n_l :	1	3	2	1	2	1
p_l :	0,1	0,3	0,2	0,1	0,2	0,1

Завжди більшу інформативність несе зображення варіаційного ряду у вигляді **гістограм** відносних частот, коли за віссю абсцис відкладають значення варіант x_l , а за віссю ординат – відповідні значення p_l , що дозволяє швидко візуально оцінити емпіричні ймовірності тих чи інших реалізацій. Із цією метою здійснюється гістограмна оцінка.

1.1.2. Гістограмна оцінка

Для проведення **гістограмної оцінки** на осі реалізацій $x \in R_1$ випадкової величини $\xi(\omega)$ задають рівномірне розбиття

$$\Delta_h : x_i = ih$$

або

$$\tilde{\Delta}_h : x_i = (i + 0,5)h, \quad i \in Z, \quad h > 0,$$

підраховують для кожного i кількість n_i спостережень з $\Omega_{1,N}$, які потрапили до відповідного елемента розбиття:

$$n_i = \sum_{l=1}^N I_i(x_l), \quad \sum_{i \in Z} n_i = N,$$

де

$$I_i(x_l) = \begin{cases} 1, & x_l \in [x_i; x_{i+1}), \\ 0, & x_l \notin [x_i; x_{i+1}), \end{cases}$$

потім визначають на інтервалах $[x_i; x_{i+1})$ відносні частоти p_i :

$$p_i = \frac{n_i}{N}, \quad \sum_{i \in Z} p_i = 1.$$

Тоді величина

$$f_i = \frac{n_i}{Nh} = \frac{p_i}{h}, \quad x \in [x_i; x_{i+1}), \quad i \in Z$$

є оцінкою **усередненого значення** $\bar{f}_i(x)$ функції щільності $f(x)$ на i -му елементі розбиття Δ_h :

$$\begin{aligned} f_i \approx \bar{f}_i(x) &= \frac{1}{h} \int_{x_i}^{x_{i+1}} f(u) du = \\ &= \frac{1}{h} (F(x_{i+1}) - F(x_i)) = \frac{1}{h} P\{x_i \leq \xi(\omega) < x_{i+1}\}, \end{aligned} \quad (1.1)$$

а величина p_i – оцінка ймовірності реалізацій $\xi(\omega)$ в межах інтервалу $[x_i; x_{i+1})$.

Звідси випливає, що на інтервалі

$$[x_{\min}; x_{\max}],$$

де x_{\min} , x_{\max} – відповідно мінімальне та максимальне значення з $\Omega_{1,N}$:

$$x_{\min} \in [x_{i_{\min}} h; x_{i_{\min}+1} h), \quad x_{\max} \in [x_{i_{\max}} h; x_{i_{\max}+1} h),$$

$$i_{\min}, i_{\max} \in Z, \quad i_{\min} < i_{\max},$$

можливе оцінювання функції розподілу $F(x)$ у вигляді **емпіричної функції розподілу** $F_{1,N}(x)$ [8 – 10]:

$$F_{1,N}(x) = \begin{cases} 0, & x < x_{\min}, \\ \sum_{j=i_{\min}}^i p_j, & x_i \leq x < x_{i+1}, \\ 1, & x \geq x_{\max}, \end{cases}$$

причому

$$P \left\{ \lim_{N \rightarrow \infty} \sup_i |F_{1,N}(x_i) - F(x_i)| = 0 \right\} = 1.$$

Відповідно до вищесказаного оцінка визначається за кількості даних результатів спостережень $N \rightarrow \infty$. У реальних задачах обробки статистичної інформації обсяги спостережень скінченні, часто навіть обмежені. У цьому разі

адекватність оцінки функції розподілу ймовірностей випадкової величини $\xi(\omega)$ залежить від того, як проведене розбиття Δ_h осі спостереження, іншими словами від вибору кроку розбиття h для одержання на основі даних вибірки $\Omega_{1,N}$ масиву значень відносних частот (емпіричної функції розподілу), найадекватніших щодо усереднених значень функції щільності (розподілу) на розбитті Δ_h .

Під час обробки вибірки $\Omega_{1,N}$ крок розбиття встановлюють зі співвідношення

$$h = \frac{x_{\max} - x_{\min}}{M},$$

де M – **кількість елементів розбиття Δ_h (класів)**, для яких $p_i \neq 0$.

Величина M досить довільна, проте існує оптимальна кількість класів, яка залежить від обсягу N даних вибірки, типу їх закону розподілу (мається на увазі врахування оцінок асиметрії та ексцесу) або інших будь-яких припущень стосовно $F(x)$.

При $N < 100$ достатньо обмежитися застосуванням формули

$$M = \begin{cases} \left[\sqrt{N} \right], & \text{якщо } \left[\sqrt{N} \right] \text{ непарне,} \\ \left[\sqrt{N} \right] - 1, & \text{якщо } \left[\sqrt{N} \right] \text{ парне,} \end{cases}$$

де $[\cdot]$ – ціла частина.

Більш точно можна визначати M , виходячи з того, що для однорідних даних, вибраних лише з однієї генеральної сукупності Ω (функція щільності розподілу випадкової величини одномодальна), практично завжди

$$M \in (0,55N^{0,4}; 1,25N^{0,4}),$$

отже, зважаючи на те, що M має бути цілочисловою (бажано непарною) величиною, завжди можна оцінити кількість класів вибірки. Якщо ж дані вибірки $\Omega_{1,N}$ неоднорідні (функція щільності багатомодальна), то кількість класів збільшується пропорційно кількості мод.

Доведено, що за існування для функції щільності $f(x)$ обмеженої другої похідної слушне таке співвідношення:

$$M \approx \sqrt[3]{N}.$$

Тому при $N \geq 100$ можна застосовувати формулу

$$M = \begin{cases} \left[\sqrt[3]{N} \right], & \text{якщо } \left[\sqrt[3]{N} \right] \text{ непарне,} \\ \left[\sqrt[3]{N} \right] - 1, & \text{якщо } \left[\sqrt[3]{N} \right] \text{ парне,} \end{cases}$$

Нижче наведений приклад графічного зображення результатів спостережень (рис. 1.1)

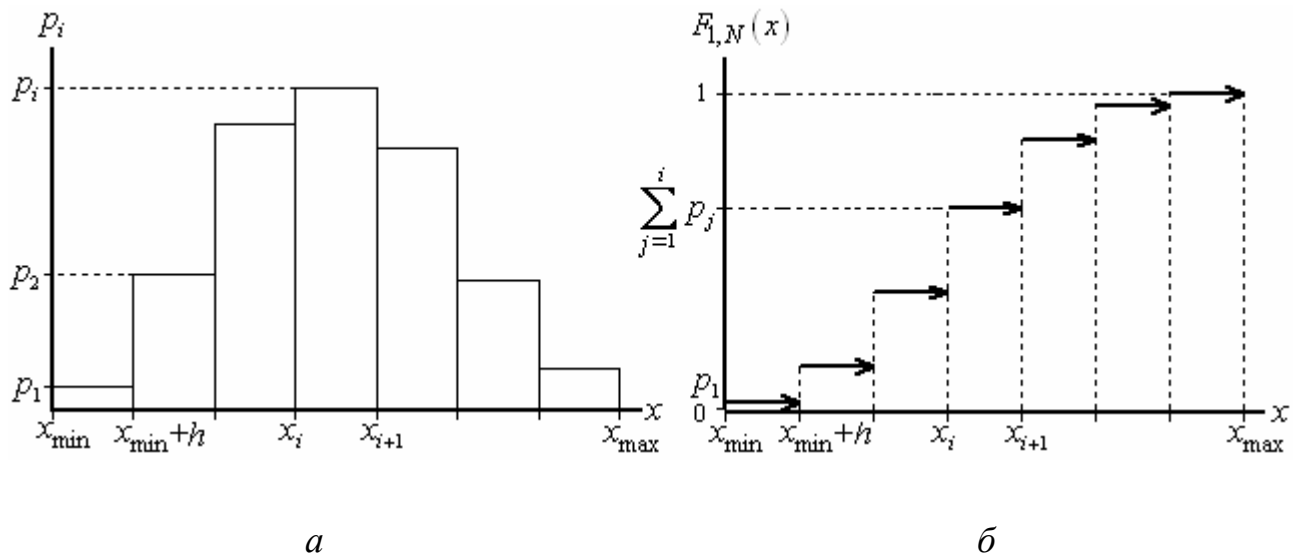


Рис. 1.1. Графічне подання результатів гістограмної оцінки:
а – гістограма відносних частот; б – графік емпіричної функції розподілу

Зауваження 1.1. З огляду на вираз (1.1) відносна частота є з точністю до константи h оцінкою усередненого значення функції щільності $f(x)$ на i -му елементі розбиття Δ_h :

$$p_i \approx \bar{f}_i(x)h.$$

Тому в разі одночасного відображення гістограми та графіка функції щільності слід зводити їх до одного масштабу шляхом нормування або відносних частот, або функції щільності. Щоб не втратити можливість інтерпретації відносних частот як імовірностей реалізації $\xi(\omega)$, рекомендується виконувати нормування функції щільності:

$$\tilde{f}(x) = f(x)h,$$

де $\tilde{f}(x)$ – нормована функція щільності. Надалі, говорячи про нормовану функцію щільності, знак « \sim » будемо опускати.

З аналізу гістограми впливають чотири основні питання:

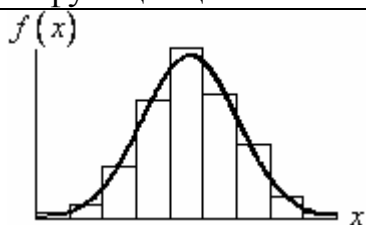
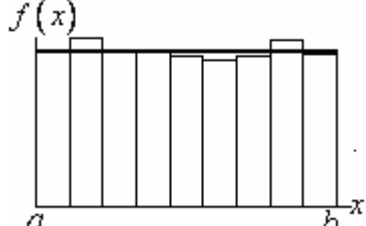
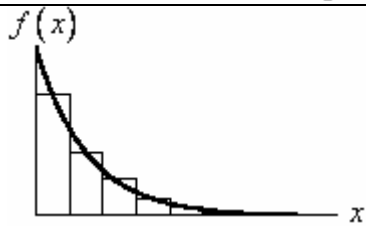

- 1) визначення моделі розподілу випадкової величини;
- 2) визначення однорідності даних;
- 3) перевірка наявності аномальних результатів спостережень;
- 4) необхідність проведення перетворень над даними.

Розглянемо приклади деяких поширених **моделей розподілів** (табл. 1.1). Так, нормальний розподіл має симетричну дзвоноподібну функцію щільності,

тому й відповідна гістограма відзначається схожим виглядом. Функція щільності розподілу для експоненціальної моделі характеризується істотною лівосторонньою асиметрією, так само – і гістограма. Якщо асиметрія гістограми незначна, то це може бути, наприклад, логарифмічно–нормальний розподіл чи розподіл Вейбулла. Якщо ж мод у гістограмі взагалі не спостерігається, мова може йти про рівномірний розподіл.

Таблиця 1.1

**Приклади моделей параметричних розподілів імовірностей
випадкової величини $\xi(\omega)$**

Розподіл	Аналітичне подання	Вигляд гістограми та графіка нормованої функції щільності
Нормальний	$F(x; m, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{(u-m)^2}{2\sigma^2}\right) du,$ $-\infty < x < \infty$	
Рівномірний	$F(x; a, b) = \begin{cases} 0, & -\infty < x < a, \\ \frac{x-a}{b-a}, & a \leq x < b, \\ 1, & b \leq x < \infty \end{cases}$	
Експоненціальний	$F(x; \lambda) = \begin{cases} 0, & -\infty < x < 0, \\ 1 - \exp(-\lambda x), & 0 \leq x < \infty \end{cases}$	
Вейбулла	$F(t; \alpha, \beta) = \begin{cases} 0, & -\infty < x < 0, \\ 1 - \exp\left(-\frac{x^\beta}{\alpha}\right), & 0 \leq x < \infty \end{cases}$	

Якщо кількість мод гістограми більша однієї, це може навести на думку про можливу **неоднорідність даних** (рис. 1.2). Тут слід нагадати, яким чином формуються вибірки. Наприклад, дослідженню підлягає розподіл розміру взуття чоловіків та жінок. У цьому випадку гістограма зазвичай двомодальна. Отже, маємо неправильно сформовану вибірку або вибірку, сформовану з двох різних генеральних сукупностей – чоловіків та жінок. Водночас багатомодальність розподілу не завжди є показником припинення подальшого аналізу саме таких даних (наприклад, дослідження розподілу часу відмов технічного виробу). У

разі виявлення неоднорідності даних подальший аналіз передбачає використання більш складних, ніж наведені (табл. 1.1), моделей розподілу, зокрема суміші нормальних розподілів або сплайн-експоненціального розподілу.

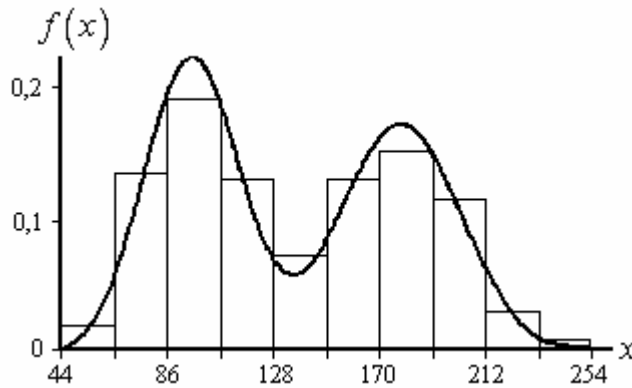


Рис. 1.2. Гістограма та графік нормованої функції щільності розподілу у випадку неоднорідних даних

Вірогідність відтворення функції розподілу величини $\xi(\omega)$ за масивом реалізацій $\Omega_{1,N}$ можна значно підвищити, здійснивши знаходження та вилучення (за наявності) із $\Omega_{1,N}$ **аномальних** результатів спостережень. Варіанта за своїм значенням може різко відхилитися від загальної сукупності варіант, якщо:

1) вона належить до генеральної сукупності, як і основна група, проте є малоймовірною подією (рис. 1.3):

$$x_{ep} \leq x_{\gamma_1} \quad \text{або} \quad x_{ep} \geq x_{\gamma_2},$$

де x_{γ_1} і x_{γ_2} визначаються з інтегральних рівнянь

$$\int_{-\infty}^{x_{\gamma_1}} f(u) du = \gamma_1; \quad \int_{x_{\gamma_2}}^{\infty} f(u) du = \gamma_2;$$

γ_1, γ_2 — помилки в прийнятті рішення про малоймовірність значення x_{ep} ;

2) має місце випадкове порушення умов експерименту.

У будь-якому разі за достатнього обсягу вибірки доцільно вилучати такі значення перед подальшою обробкою. Наприклад, оцінка x_{ep} може бути одержана з зазначених умов на основі апроксимації гістограм відносних частот. Справді, якщо на «хвості» розподілу відносна частота p_i , $i = \overline{1, s_1}$, $i = \overline{s_2, M}$ варіанти розбитого на класи варіаційного ряду менша величини помилки в прийнятті рішення про малоймовірність її значення

$$p_i \approx \int_{x_i - 0.5h}^{x_i + 0.5h} f(u) du < \gamma_1, \quad i = \overline{1, s_1}$$

або

$$p_i \approx \int_{x_i-0,5h}^{x_i+0,5h} f(u) du < \gamma_2, \quad i = \overline{s_2, M},$$

то, очевидно, реалізації вибірки, що потрапили до даного класу, є аномальні.

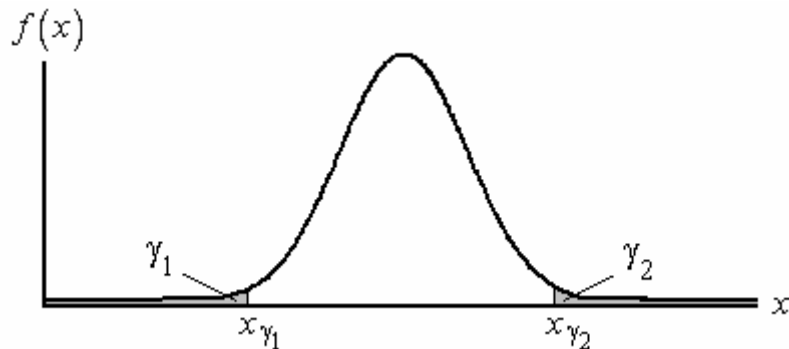


Рис. 1.3. Області малоїмовірних спостережень на графіку функції щільності

Відзначимо, що під варіантами розбитого на класи варіаційного ряду x_i маються на увазі середини класів.

Як видно з рис. 1.4, після вилучення аномальних результатів спостережень можна досягти більш вірогідного відтворення функції щільності.



Рис. 1.4. Гістограма та графік нормованої функції щільності за наявності аномальних результатів спостережень: а – вихідний вигляд; б – після вилучення аномальних значень

У випадку обробки **асиметричних даних** за необхідності зведення даних до вигляду з симетричною функцією щільності рекомендується здійснювати нелінійні перетворення над вихідними масивами, наприклад, шляхом логарифмування за експоненціальною чи десятковою основою:

$$x_l^* = \ln x_l,$$

$$x_l^* = \lg x_l, \quad l = \overline{1, N}$$

або за будь-якою іншою основою $c > 0$:

$$x_l^* = \log_c x_l, \quad l = \overline{1, N}.$$

Зауваження 1.2. Операція логарифмування прийнятна лише для даних, які не містять від'ємних спостережень. За наявності останніх, перед логарифмуванням слід виконати лінійне перетворення (зсув), наприклад:

$$x_l^* = x_l + |x_{\min}| + \varepsilon, \quad \forall \varepsilon > 0, \quad l = \overline{1, N},$$

де x_{\min} – значення найменшого від'ємного значення у $\Omega_{1,N}$.

Операція логарифмування дозволяє «розтягнути» шкалу спостережень поблизу нуля, тим самим перерозподіляючи дані, згруповані в цьому околі. Водночас логарифмування уможливорює перегруповання даних, розташованих на правому «хвості» реалізацій, шляхом «звуження» шкали вимірювання зі зростанням відліку вихідної осі x . Операція логарифмування істотно впливає на вигляд гістограм відносних частот, зводячи (у випадку вдалого вибору основи логарифма) їх до симетричного (рис. 1.5).

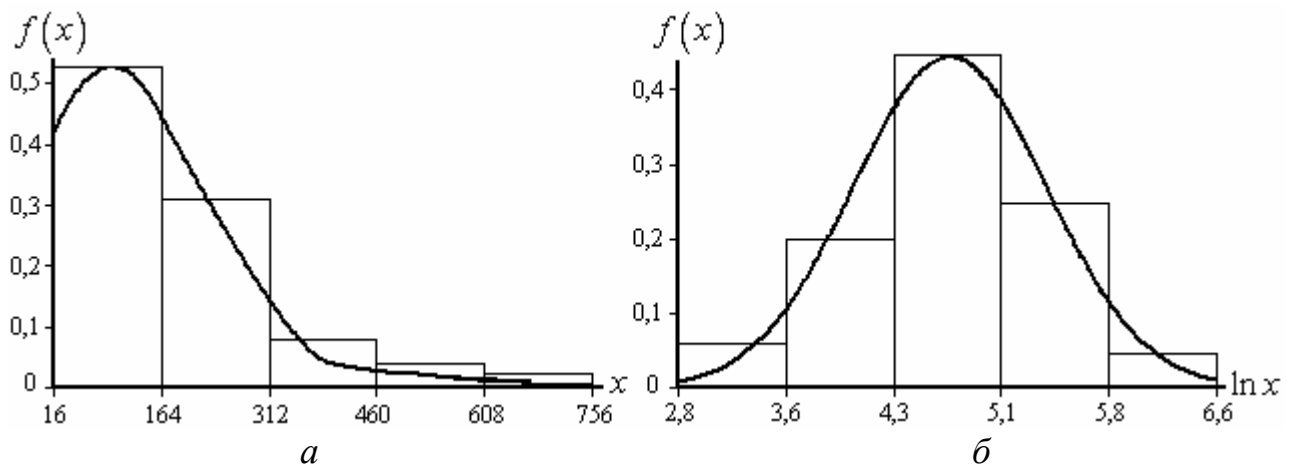


Рис. 1.5. Вигляд гістограми та графіка нормованої функції щільності асиметричного розподілу: *а* – вихідний; *б* – після застосування логарифмічного перетворення

Зазначимо, що крім логарифмування можна застосовувати й інші типи нелінійних перетворень:

$$x_l^* = x_l^c,$$

$$x_l^* = \frac{1}{x_l^c}, \quad x_l^c \neq 0,$$

$$x_l^* = \exp(-x_l) \text{ і т.д.,}$$

які забезпечують симетричність функції щільності розподілу ймовірностей випадкової величини $\xi^*(\omega)$, що є функцією від вихідної $\xi^*(\omega) = \phi(\xi(\omega))$.

1.1.3. Точкові та інтервальні оцінки

Уведемо поняття параметра й оцінки параметра генеральної сукупності та вибірки.

Параметром θ генеральної сукупності називають число, яке визначає характеристики генеральної сукупності. Параметр є невідома та фіксована величина.

Оцінкою параметра $\hat{\theta}$ вибірки називають вибірккову статистику

$$\hat{\theta} = \phi(x_1, \dots, x_N),$$

яка оцінює параметр θ . При цьому вибір функції $\phi(\cdot)$ залежить від методу знаходження оцінок параметра. Реалізація процедур, що визначають $\phi(\cdot)$, дозволяє відшукувати оцінки, які поділяються на точкові та інтервальні.

У випадку **точкового оцінювання**, коли деякому параметру θ ставиться у відповідність оцінка $\hat{\theta}$, виникає питання про адекватність такого зіставлення. Похибкою оцінки називають різницю поміж оцінкою та параметром, звичайно похибка оцінки – невідома величина. Залежно від похибки оцінки параметрів мають такі головні властивості:

1) **незсуненість**, якщо математичне сподівання оцінки $E\{\hat{\theta}\}$ дорівнює параметру генеральної сукупності

$$E\{\hat{\theta}\} = \theta;$$

2) **спроможність** у разі прямування оцінки за ймовірністю до значення параметра для будь-якого $\varepsilon > 0$:

$$P\left\{\left|\hat{\theta}_N - \theta\right| \leq \varepsilon\right\} \xrightarrow[N \rightarrow \infty]{\text{Ймов}} 1,$$

де N – кількість елементів вибірки, на основі яких одержана оцінка;

3) **ефективність**, якщо оцінка має в певному класі серед k інших подібних до неї оцінок мінімальну дисперсію:

$$\hat{\theta} : \min_k D\{\hat{\theta}^{(k)}\}.$$

Наприклад, відносно властивостей гістограмної оцінки слід зауважити,

що вона є спроможною та незсуненою оцінкою усереднених значень функції щільності (розподілу) на розбитті Δ_h .

Як було зазначено, першим кроком в аналізі даних є вивчення варіаційних рядів та гістограм, що дозволяє зробити висновок про повноту даних та припустимі ймовірнісні характеристики. Подальший аналіз даних зводиться до обчислення наведених нижче оцінок характеристик вибірки.

Середнє арифметичне є оцінка математичного сподівання випадкової величини ξ та використовується як показник типового значення в наборі даних:

$$\bar{x} = \frac{1}{N} \sum_{l=1}^N x_l = \frac{1}{N} \sum_{i=1}^c x_i n_i = \sum_{i=1}^c x_i p_i,$$

причому подібна оцінка є незсунена, отже:

$$E\{\bar{x}\} = E\{\xi\}.$$

Як c може використовуватися r (у такому разі x_i – варіанти варіаційного ряду) або M (тоді x_i – варіанти варіаційного ряду, розбитого на класи, тобто середини класу). В останньому випадку має бути врахована поправка Шеппарда на дискретизацію (М. Кендалл, А. Стьюарт, 1966).

Величина середнього арифметичного визначає розташування графіка функції щільності (або гістограми) на осі спостережень (рис. 1.6).

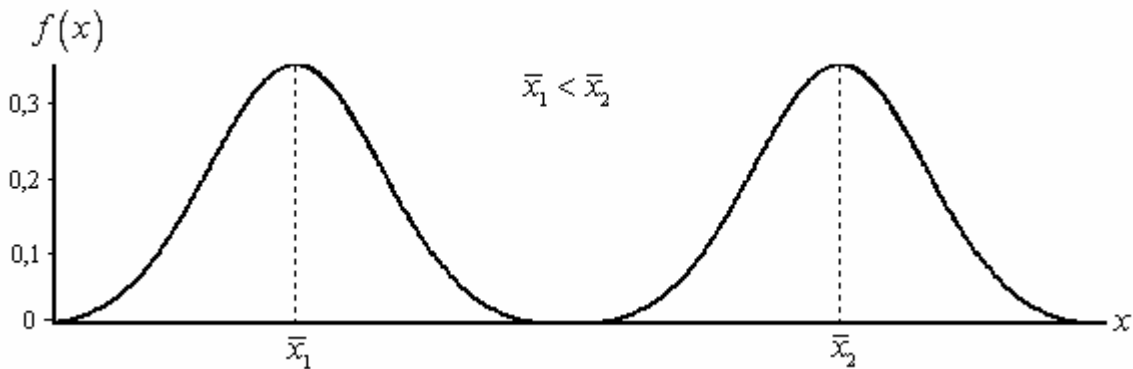


Рис. 1.6. Графік функції щільності залежно від середнього арифметичного

Вибіркова дисперсія та середньоквадратичне відхилення (стандартне відхилення), що характеризують розсіювання вибірових даних відносно середнього (рис. 1.7), можуть бути:

– зсунені:

$$\hat{S}^2 = \frac{1}{N} \sum_{l=1}^N x_l^2 - \bar{x}^2 = \frac{1}{N} \sum_{i=1}^c x_i^2 n_i - \bar{x}^2 = \sum_{i=1}^c x_i^2 p_i - \bar{x}^2, \quad \hat{\sigma} = \hat{S};$$

– незсунені:

$$S^2 = \frac{1}{N-1} \sum_{l=1}^N (x_l - \bar{x})^2 = \frac{1}{N-1} \sum_{i=1}^c (x_i - \bar{x})^2 n_i = \frac{N}{N-1} \sum_{i=1}^c (x_i - \bar{x})^2 p_i, \quad \sigma = S.$$

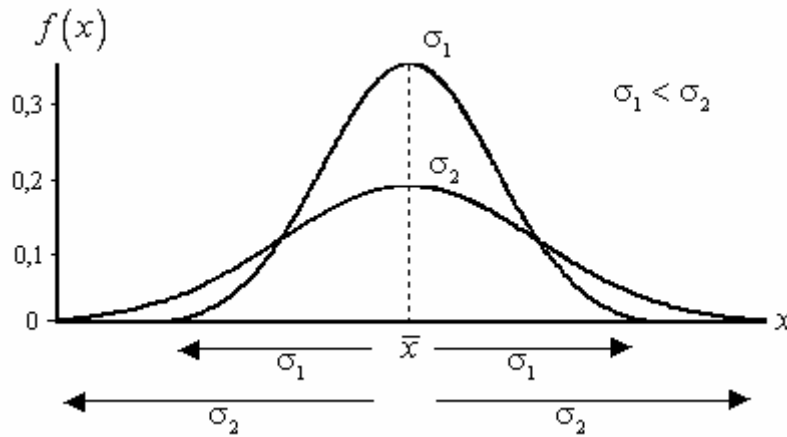


Рис. 1.7. Графік функції щільності залежно від σ

Часто для даних, наведених у різних одиницях виміру, вводять операцію стандартизації:

$$x_l^* = \frac{x_l - \bar{x}}{\sigma},$$

що дозволяє перейти до «безрозмірних» стандартизованих даних, для яких середнє арифметичне дорівнює нулю, а середнє квадратичне відхилення – одиниці. При цьому вигляд гістограми не змінюється.

Коефіцієнт асиметрії, що характеризує асиметричність функції щільності (гістограми) відносно середнього, буває:

– зсунений:

$$\hat{A} = \frac{1}{N\hat{\sigma}^3} \sum_{l=1}^N (x_l - \bar{x})^3 = \frac{1}{N\hat{\sigma}^3} \sum_{i=1}^c (x_i - \bar{x})^3 n_i = \frac{1}{\hat{\sigma}^3} \sum_{i=1}^c (x_i - \bar{x})^3 p_i;$$

– незсунений:

$$\bar{A} = \frac{\sqrt{N(N-1)}}{N-2} \hat{A},$$

причому функція щільності симетрична, якщо $\bar{A} = 0$; у разі $\bar{A} > 0$ функція щільності лівоасиметрична; при $\bar{A} < 0$ – правоасиметрична (рис. 1.8).

Коефіцієнт ексцесу, що характеризує гостровершинність функції щільності вибіркового розподілу (гістограми) відносно теоретичного нормального розподілу (рис. 1.9) є:

– зсунений:

$$\hat{E} = \frac{1}{N\hat{\sigma}^4} \sum_{l=1}^N (x_l - \bar{x})^4 = \frac{1}{N\hat{\sigma}^4} \sum_{i=1}^c (x_i - \bar{x})^4 n_i = \frac{1}{\hat{\sigma}^4} \sum_{i=1}^c (x_i - \bar{x})^4 p_i;$$

– незсунений:

$$\bar{E} = \frac{N^2 - 1}{(N - 2)(N - 3)} \left((\hat{E} - 3) + \frac{6}{N + 1} \right).$$

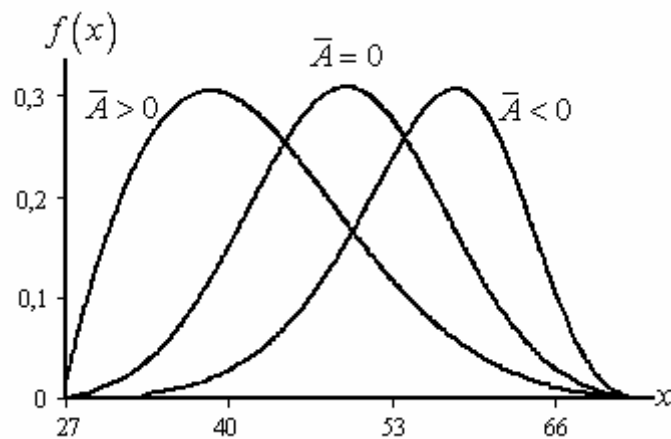


Рис. 1.8. Графік функції щільності залежно від \bar{A}

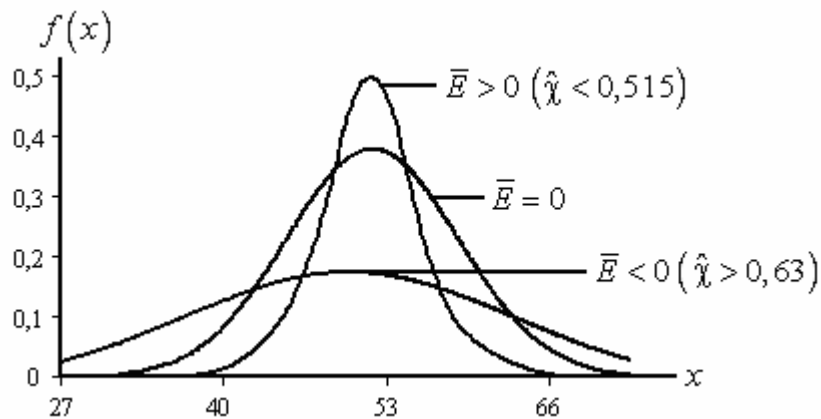


Рис. 1.9. Графік функції щільності залежно від \bar{E} та $\hat{\chi}$

Коефіцієнт контрексесу

$$\hat{\chi} = \frac{1}{\sqrt{|\bar{E}|}}$$

визначає форму розподілу, причому, якщо $\hat{\chi} < 0,515$, розподіл є гостровершинний; при $\hat{\chi} > 0,63$ має місце форма розподілу типу шапїто (приклад – рівномір-

ний розподіл) (рис. 1.9).

Коефіцієнт варіації Пірсона

$$\bar{W} = \frac{\sigma}{\bar{x}}$$

характеризує якість вибірки, відображає відносну варіабельність даних у частках відносно середнього та дозволяє порівнювати варіабельність наборів даних, наведених у різних одиницях виміру. Якщо $\bar{W} < 1$, вибірка вважається якісною, тобто величина розсіювання відповідає середньому арифметичному; поміж двох вибірок кращою вважається та, для якої значення коефіцієнта \bar{W} менше, тобто менша варіабельність (рис. 1.10).

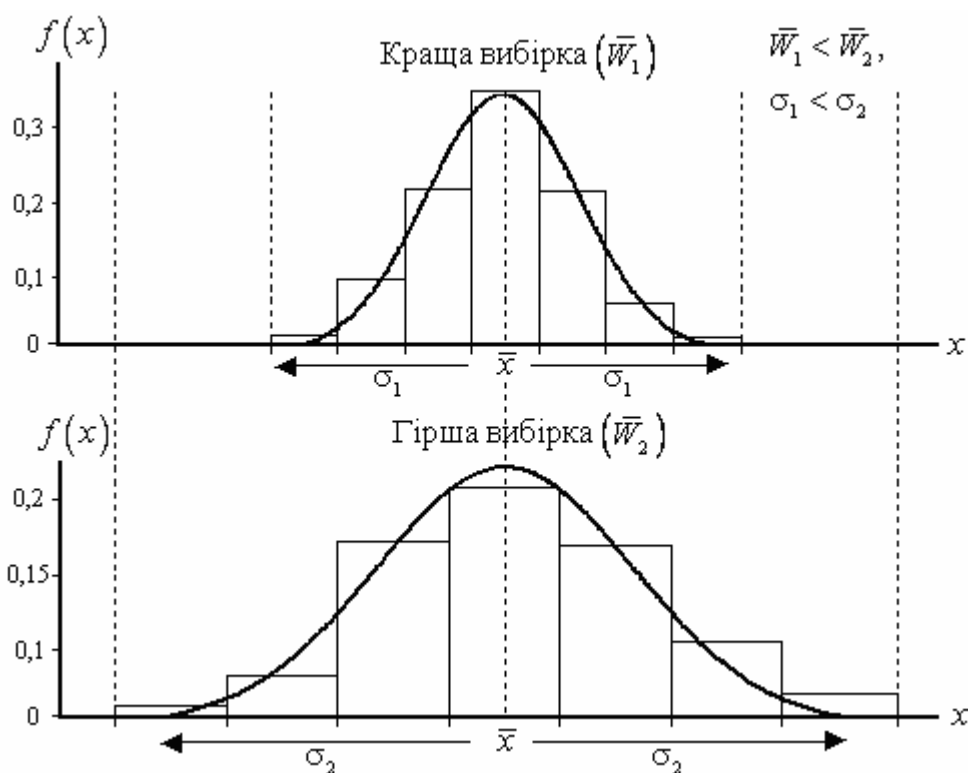


Рис. 1.10. Графік нормованої функції щільності та гістограми залежно від \bar{W}

Зауваження 1.3. Визначення оцінок за формулами з використанням частот і відносних частот застосовується в процесі оцінювання параметрів для дискретних випадкових величин.

За можливості оцінити функцію розподілу $F(x; \bar{\Theta})$ або знайти апроксимацію функції розподілу $F(x)$ шляхом інтерполяції табульованих значень емпіричної функції $F_{1,N}(x_i)$ до характеристик вибірки долучають оцінки **квантилів**

$$\hat{x}_\alpha = F^{-1}(\alpha),$$

де величину ймовірності α зазвичай вибирають такою, що дорівнює 0,05, 0,1, 0,25, 0,5, 0,75, 0,9, 0,95.

При обмежених обсягах інформації ($N \leq 10$) оцінюють лише середнє арифметичне, середнє квадратичне та коефіцієнт варіації Пірсона. Обчислення \bar{x} і \bar{W} проводять за відомими формулами, значення $\hat{\sigma}$ визначають для варіаційного ряду за формулами, наведеними нижче (табл.1.2).

Таблиця 1.2

**Визначення зсуненого середньоквадратичного відхилення
при обмежених обсягах інформації**

Обсяг вибірки	Оцінка $\hat{\sigma}$
2	$0,8862(x_2 - x_1)$
3	$0,5908(x_3 - x_1)$
4	$0,4539(x_4 - x_1) + 0,1102(x_3 - x_2)$
5	$0,3724(x_5 - x_1) + 0,1352(x_4 - x_2)$
6	$0,3175(x_6 - x_1) + 0,1386(x_5 - x_2) + 0,0432(x_4 - x_3)$
7	$0,2778(x_7 - x_1) + 0,1351(x_6 - x_2) + 0,0625(x_5 - x_3)$
8	$0,2476(x_8 - x_1) + 0,1294(x_7 - x_2) + 0,0713(x_6 - x_3) + 0,0230(x_5 - x_4)$
9	$0,2237(x_9 - x_1) + 0,1233(x_8 - x_2) + 0,0750(x_7 - x_3) + 0,0360(x_6 - x_4)$
10	$0,2044(x_{10} - x_1) + 0,1172(x_9 - x_2) + 0,0763(x_8 - x_3) +$ $+0,0436(x_7 - x_4) + 0,0142(x_6 - x_5)$

Будь-яка статистика, обчислена на основі випадкової вибірки, має розподіл імовірностей, який називають **вибірковим розподілом** цієї статистики. Знання вибіркового розподілу дає можливість перейти від інформації про вибірку (одержаної на основі даних) до інформації про генеральну сукупність. У багатьох випадках вибіровий розподіл статистик близький до нормального (середнього, середньоквадратичного та ін.) навіть тоді, коли розподіл окремих об'єктів дослідження є відмінний від нього. Згідно з **центральною граничною теоремою** для випадкової вибірки обсягу N елементів із генеральної сукупнос-

ті слушні твердження:

1) зі збільшенням N розподіл як середнього, так і суми все більше наближається до нормального;

2) середнє та середньоквадратичне відхилення розподілів середнього та суми обчислюють за такими виразами:

	Середнє	Загальна сума
Середнє	$\mu_{\bar{x}} = \mu$	$\mu_{\text{sum}} = \mu \cdot N$
Середньоквадратичне	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$	$\sigma_{\text{sum}} = S\sqrt{N}$

(μ та σ – середнє та середньоквадратичне елементів генеральної сукупності).

Дослідження законів розподілу статистик, які є оцінками параметрів, дозволяє робити висновок відносно ймовірності появи певного значення конкретно обчисленої статистики, а обчислення дисперсії $D\{\hat{\theta}\}$ – стосовно проведення інтервального оцінювання.

Для оцінювання параметрів на основі довірчих інтервалів припускають, що значення параметра генеральної сукупності з деякою ймовірністю γ розташоване поміж оцінками $\hat{\theta}_H$ та $\hat{\theta}_B$:

$$P\{\hat{\theta}_H \leq \theta \leq \hat{\theta}_B\} = \gamma,$$

$$\theta \in [\hat{\theta}_H; \hat{\theta}_B],$$

при цьому інтервал $(\hat{\theta}_H; \hat{\theta}_B)$ називають 100γ -відсотковим **довірчим інтервалом** для θ , а самі $\hat{\theta}_H$, $\hat{\theta}_B$ – відповідно **нижньою** та **верхньою довірчими межами**. Обчислення значень $\hat{\theta}_H$, $\hat{\theta}_B$ проводять, виходячи з закону розподілу оцінки параметра $\hat{\theta}$. Так, якщо закон розподілу оцінки параметра симетричний (нормальний закон або закон розподілу Стюдента), нижню й верхню довірчі межі призначають зі співвідношень

$$\hat{\theta}_H = \hat{\theta} - t_{\alpha/2, v} \sqrt{D\{\hat{\theta}\}} = \hat{\theta} - t_{\alpha/2, v} \sigma\{\hat{\theta}\},$$

$$\hat{\theta}_B = \hat{\theta} + t_{\alpha/2, v} \sqrt{D\{\hat{\theta}\}} = \hat{\theta} + t_{\alpha/2, v} \sigma\{\hat{\theta}\},$$

де

$t_{\alpha/2, v}$ – квантиль t -розподілу Стюдента (додат. А, Б);

$v = N - 1$ (при $N > 60$ замість $t_{\alpha/2, v}$ використовують квантиль $u_{\alpha/2}$ стандар-

тного нормального закону);

$\alpha = 1 - \gamma$ – величина ймовірності «промаху» параметра повз довірчий інтервал.

Таким чином, інтервальну оцінку параметра θ проводять із довірчою ймовірністю γ (найчастіше $\gamma = 0,9$ або $\gamma = 0,95$) на основі нерівності

$$\hat{\theta} - t_{\alpha/2, v} \sigma\{\hat{\theta}\} \leq \theta \leq \hat{\theta} + t_{\alpha/2, v} \sigma\{\hat{\theta}\}$$

або шляхом призначення одnobічних довірчих інтервалів

$$\hat{\theta} - t_{\alpha/2, v} \sigma\{\hat{\theta}\} \leq \theta, \quad \theta \leq \hat{\theta} + t_{\alpha/2, v} \sigma\{\hat{\theta}\}.$$

Поряд із найчастіше використовуваним довірчим рівнем 95% беруться й інші. Вибір рівня – це компроміс між розміром інтервалу (менший інтервал є більш точний, а отже, і більш бажаний) та ймовірністю того, що інтервал включає шуканий параметр генеральної сукупності (вища ймовірність більш бажана).

У процесі інтервального оцінювання вищерозглянутих характеристик вибірки призначають довірчі інтервали з надійною ймовірністю γ . Як величину $\hat{\theta}$ беруть відповідну точкову оцінку, а значення $\sigma\{\hat{\theta}\}$ обчислюють за співвідношеннями

$$\sigma\{\bar{x}\} = \frac{S}{\sqrt{N}},$$

$$\sigma\{S\} = \frac{S}{\sqrt{2N}},$$

$$\sigma\{\bar{A}\} = \sqrt{\frac{6}{N} \left(1 - \frac{12}{2N+7}\right)} \quad \text{або} \quad \sigma\{\bar{A}\} = \sqrt{\frac{6(N-2)}{(N+1)(N+3)}},$$

$$\sigma\{\bar{E}\} = \sqrt{\frac{24}{N} \left(1 - \frac{225}{15N+124}\right)} \quad \text{або} \quad \sigma\{\bar{E}\} = \sqrt{\frac{24N(N-2)(N-3)}{(N+1)^2(N+3)(N+5)}},$$

$$\sigma\{\hat{\chi}\} = \sqrt{\frac{|\hat{E}|}{29N}} \sqrt[4]{|\hat{E}^2 - 1|^3},$$

$$\sigma\{\bar{W}\} = \bar{W} \sqrt{\frac{1 + 2\bar{W}^2}{2N}}.$$

При обмеженому обсязі даних довірче оцінювання коефіцієнта варіації проводять при умові $\bar{W} < 1$:

$$\frac{\bar{W}}{1 + v\sqrt{1 + 2\bar{W}^2}} \leq W \leq \frac{\bar{W}}{1 - v\sqrt{1 + 2\bar{W}^2}},$$

де

$$v = \frac{u_{\alpha/2}}{\sqrt{2(N-1)}};$$

$u_{\alpha/2}$ - квантиль нормального розподілу (додат. А, Б).

Інтервальне оцінювання квантилів здійснюється згідно з нерівністю

$$\hat{x}_\alpha - t_{\gamma/2, v} \frac{\alpha(1-\alpha)}{N(f(x_\alpha))^2} < x_\alpha < \hat{x}_\alpha + t_{\gamma/2, v} \frac{\alpha(1-\alpha)}{N(f(x_\alpha))^2},$$

де

$f(x_\alpha)$ – відповідне значення функції щільності;

γ – імовірність «промаху» значення оцінки повз довірчий інтервал.

Якщо обчислені оцінки квантилів, то можливе **наведення довірчих інтервалів реалізації випадкової величини**. Найчастіше наводять інтервали реалізації з довірчою ймовірністю 0,9 :

$$[\hat{x}_{0,05}; \hat{x}_{0,95}].$$

Інтервал передбачення дозволяє використовувати дані вибірки для прогнозування з відомою ймовірністю значення нового спостереження за умови, що це спостереження одержане тим самим способом, що і попередні. Як міру невізначеності при цьому використовують стандартну похибку передбачення

$$S\sqrt{1 + \frac{1}{N}}$$

– міру варіабельності відстані між середнім значенням вибірки та новим спостереженням. Отже, нове спостереження з імовірністю $1 - \alpha$ буде знаходитись у межах

$$\bar{x} - t_{\alpha/2, v} S\sqrt{1 + \frac{1}{N}} < x_{\text{нове}} < \bar{x} + t_{\alpha/2, v} S\sqrt{1 + \frac{1}{N}}.$$

1.2. Відтворення розподілів

Будемо вважати, що на основі $\Omega_{1, N}$ одержаний масив $\{x_l, F_{1, N}(x_l); l = \overline{1, N}\}$.

Необхідно відтворити функцію розподілу $F(x; \bar{\Theta}) = P\{\omega : -\infty < \xi(\omega) < x\}$ шляхом

знаходження оцінки $F(x; \hat{\Theta})$, де $\hat{\Theta} = \{\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_s\}$, $s \geq 1$.

Обчислювальна схема відтворення розподілу може якісно відрізнятися від поданої нижче схеми або бути її варіацією. Проте кінцева мета кожної з них – одержання **статистичної оцінки функції розподілу** $F(x; \hat{\Theta})$ за вибірковими даними $\Omega_{1,N}$. Розв'язання статистичної задачі відтворення функції розподілу потребує реалізації таких обчислювальних процедур:

- 1) **первинного статистичного аналізу**;
- 2) **знаходження вектора оцінок параметрів** $\hat{\Theta} = \{\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_s\}$ для апріорно заданого або ідентифікованого типу розподілу $F(x; \vec{\Theta})$;
- 3) **оцінювання точності оцінок параметрів** шляхом обчислення дисперсій $D\{\hat{\theta}_i\}$ та довірчих інтервалів для кожного з параметрів θ_i , $i = \overline{1, s}$;
- 4) **обчислення значень статистичної функції розподілу** $F(x; \hat{\Theta})$ у точках варіаційного ряду;
- 5) **інтервального (довірчого) оцінювання теоретичної функції розподілу ймовірностей**;
- 6) **визначення одного або кількох (за необхідності) критеріїв згоди** (критерію χ^2 , уточненого критерію Колмогорова, критерію ω^2 та ін.), що дозволяють оцінити достовірність розподілу $F(x; \hat{\Theta})$.

1.2.1. Методи оцінки параметрів розподілу

Вибираючи метод знаходження оцінок параметрів розподілу, прагнуть, щоб шляхом якомога простіших обчислювальних процедур одержати такі оцінки, для яких властиві були б незсуненість, спроможність та ефективність. Проте така вимога виконується не завжди. Це залежить від методу, на основі якого здійснюється обчислювальна процедура, та від типу відтворюваного розподілу.

Практичне застосування мають методи: максимальної правдоподібності, моментів та найменших квадратів. Як показує досвід, найбільш ефективні є метод максимальної правдоподібності та близький до нього метод найменших квадратів.

Метод максимальної правдоподібності (ММП) полягає в знаходженні оцінок вектора $\bar{\Theta} = \{\theta_1, \theta_2, \dots, \theta_s\}$ з умови

$$\max_{\bar{\Theta}} L_1 = \max_{\bar{\Theta}} \prod_{l=1}^N f(x_l; \theta_1, \dots, \theta_s),$$

еквівалентної

$$\max_{\bar{\Theta}} L = \max_{\bar{\Theta}} \sum_{l=1}^N \ln f(x_l; \theta_1, \dots, \theta_s), \quad (1.2)$$

де $L = \ln L_1$.

Доведено, що для виконання умови (1.2) необхідно, щоб

$$\frac{\partial L}{\partial \theta_1} = 0, \quad \dots, \quad \frac{\partial L}{\partial \theta_s} = 0. \quad (1.3)$$

Розв'язання системи рівнянь (1.3) дає обчислювальну процедуру знаходження оцінок параметрів.

Приклад 1.2. Для експоненціального розподілу, з урахуванням (1.2), функція правдоподібності має вигляд

$$L = N \ln \lambda - \lambda \sum_{l=1}^N x_l,$$

отже,

$$\hat{\lambda} = \frac{N}{\sum_{l=1}^N x_l} = \frac{1}{\bar{x}}.$$

Приклад 1.3. Для нормального розподілу з функцією правдоподібності

$$\begin{aligned} L &= \sum_{l=1}^N \ln \left(\frac{1}{\sigma \sqrt{2\pi}} \exp \left(-\frac{(x_l - m)^2}{2\sigma^2} \right) \right) = \\ &= -\frac{1}{2} N \ln(2\pi) - N \ln \sigma - \frac{1}{2\sigma^2} \sum_{l=1}^N (x_l - m)^2 \end{aligned} \quad (1.4)$$

маємо

$$\begin{cases} \frac{\partial L}{\partial m} = \frac{1}{\sigma^2} \sum_{l=1}^N (x_l - m) = 0, \\ \frac{\partial L}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_{l=1}^N (x_l - m)^2 = 0, \end{cases}$$

звідки, з урахуванням (1.3),

$$\hat{m} = \frac{1}{N} \sum_{l=1}^N x_l = \bar{x},$$

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{l=1}^N (x_l - m)^2 = \frac{1}{N} \sum_{l=1}^N (x_l - \bar{x})^2.$$

Слід зазначити, що оцінка параметра $\hat{\sigma}$, одержана за методом максимальної правдоподібності, є зсунена.

Метод моментів (ММ) базується на властивості рівності теоретичних та статистичних початкових або центральних моментів:

$$v_k = \hat{v}_k, \quad \mu_k = \hat{\mu}_k, \quad k = \overline{1, s},$$

причому можлива їх комбінація. Для одно- та двопараметричних розподілів можна говорити, що оцінки параметрів $\hat{\Theta}$, одержані за методом моментів, мають вигляд

$$\hat{\theta}_i = H_i(\bar{x}, \bar{x}^2), \quad i = \overline{1, s}, \quad (1.5)$$

тобто оцінка параметра є деякою функцією моментів.

Нагадаємо, що

$$\hat{v}_1 = \bar{x}, \quad \hat{\mu}_2 = S.$$

Приклад 1.4. Для експоненціального розподілу є правильне

$$v_1 = \frac{1}{\lambda},$$

отже,

$$\hat{\lambda} = \frac{1}{\bar{x}}.$$

Приклад 1.5. Для нормального закону розподілу є слушне

$$v_1 = m, \quad \mu_2 = \sigma^2,$$

таким чином,

$$\hat{m} = \bar{x}, \quad \hat{\sigma} = S.$$

Метод найменших квадратів (МНК) ефективно реалізується у тому випадку, коли функцію розподілу шляхом деякого перетворення зводять до лінійного вигляду відносно параметрів. Нехай, наприклад, двопараметричний розподіл зведений до вигляду

$$z = \theta_1 + \theta_2 t.$$

Тоді початковий масив варіаційного ряду

$$\{x_l, F_{1,N}(x_l); l = \overline{1, N}\}$$

перетворюється на масив

$$\{t_l, z_l; l = \overline{1, N}\}.$$

За стандартною процедурою методу найменших квадратів з умови мінімізації залишкової дисперсії

$$\min_{\hat{\theta}_1, \hat{\theta}_2} S_{\text{зал}}^2 = \min_{\hat{\theta}_1, \hat{\theta}_2} \frac{1}{N-3} \sum_{l=1}^{N-1} (z_l - \hat{\theta}_1 - \hat{\theta}_2 t_l)^2,$$

тобто з розв'язку системи рівнянь

$$\frac{\partial S_{\text{зал}}^2}{\partial \hat{\theta}_1} = 0; \quad \frac{\partial S_{\text{зал}}^2}{\partial \hat{\theta}_2} = 0,$$

одержують систему лінійних алгебричних рівнянь

$$A\hat{\Theta} = Z,$$

де

$$A = \begin{pmatrix} 1 & \bar{t} \\ \bar{t} & \bar{t}^2 \end{pmatrix}; \quad \hat{\Theta} = \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix}; \quad Z = \begin{pmatrix} \bar{z} \\ \bar{zt} \end{pmatrix};$$

$$\bar{t} = \frac{1}{N-1} \sum_{l=1}^{N-1} t_l; \quad \bar{z} = \frac{1}{N-1} \sum_{l=1}^{N-1} z_l;$$

$$\bar{t}^2 = \frac{1}{N-1} \sum_{l=1}^{N-1} t_l^2; \quad \bar{zt} = \frac{1}{N-1} \sum_{l=1}^{N-1} z_l t_l.$$

Приклад 1.6. Експоненціальний розподіл

$$F(x) = 1 - \exp(-\lambda x), \quad x \geq 0,$$

зводиться до лінійного вигляду

$$\ln \frac{1}{1-F(x)} = \lambda x.$$

Тоді масив $\{x_l, F_{1,N}(x_l); l = \overline{1, N}\}$ перетворюється на масив $\{t_l, z_l; l = \overline{1, N}\}$, де

$t_l = x_l$, $z_l = \ln \frac{1}{1-F_{1,N}(x_l)}$. Оскільки розподіл є однопараметричний, залишкова

дисперсія має вигляд

$$S_{\text{зал}}^2 = \frac{1}{N-2} \sum_{l=1}^{N-1} (z_l - \hat{\lambda} t_l)^2.$$

Реалізуючи умову мінімуму залишкової дисперсії

$$\frac{dS_{\text{зал}}^2}{d\lambda} = \frac{-2}{N-2} \sum_{l=1}^{N-1} (z_l - \hat{\lambda} t_l) t_l = 0,$$

одержують

$$\hat{\lambda} = \frac{\sum_{l=1}^{N-1} z_l t_l}{\sum_{l=1}^{N-1} t_l^2}.$$

Приклад 1.7. Нормальний розподіл у результаті перетворення відносно квантилів набуває такого вигляду:

$$x = m + \sigma u,$$

тобто початковий масив $\{x_l, F_{1,N}(x_l); l = \overline{1, N}\}$ переформовується в масив $\{u_l, x_l; l = \overline{1, N}\}$, де $u_l = F_{1,N}^{-1}(x_l)$.

Тоді

$$S_{\text{зал}}^2 = \frac{1}{N-3} \sum_{l=1}^{N-1} (x_l - \hat{m} - \hat{\sigma} u_l)^2.$$

Із системи рівнянь

$$\begin{cases} \hat{m} + \hat{\sigma} \bar{u} = \bar{x}, \\ \hat{m} \bar{u} + \hat{\sigma} \overline{u^2} = \overline{xu} \end{cases}$$

визначають

$$\hat{m} = \frac{\bar{x} \overline{u^2} - \bar{u} \cdot \overline{xu}}{\overline{u^2} - (\bar{u})^2},$$

$$\hat{\sigma} = \frac{\overline{xu} - \bar{x} \cdot \bar{u}}{\overline{u^2} - (\bar{u})^2}.$$

Подібні обчислювальні процедури застосовуються, крім того, до таких розподілів імовірностей: логарифмічно-нормального, Вейбулла, екстремального та ін.

1.2.2. Оцінювання точності оцінок параметрів

За умови, що реалізуються такі методи визначення оцінок параметрів, як метод максимальної правдоподібності, найменших квадратів, оцінювання дисперсій та коваріацій оцінок параметрів двопараметричного розподілу здійснюється на основі дисперсійно-коваріаційної матриці

$$DC = \begin{pmatrix} D\{\hat{\theta}_1\} & \text{cov}\{\hat{\theta}_1, \hat{\theta}_2\} \\ \text{cov}\{\hat{\theta}_2, \hat{\theta}_1\} & D\{\hat{\theta}_2\} \end{pmatrix},$$

причому спосіб визначення DC може бути різний залежно від методу (ММП, ММ чи МНК).

У **методі максимальної правдоподібності** матрицю DC знаходять на основі матриці, зворотної до інформаційної матриці I :

$$DC = -I^{-1},$$

де

$$I = \begin{pmatrix} \frac{\partial^2 L}{\partial \theta_1^2} & \frac{\partial^2 L}{\partial \theta_1 \partial \theta_2} \\ \frac{\partial^2 L}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 L}{\partial \theta_2^2} \end{pmatrix}.$$

Визначення $\frac{\partial^2 L}{\partial \theta_i^2}$ та $\frac{\partial^2 L}{\partial \theta_i \partial \theta_j}$ не викликає труднощів, тому обчислення матриці DC – процедура здійснення: спочатку слід знайти в числовому вигляді матрицю I при $\theta_i = \hat{\theta}_i$, $i = 1, 2$, а вже потім – DC .

Приклад 1.8. Для нормального розподілу з функцією правдоподібності (1.4) дисперсійно-коваріаційна матриця має вигляд

$$DC = \begin{pmatrix} \frac{\sigma^2}{N} & 0 \\ 0 & \frac{\sigma^2}{2N} \end{pmatrix},$$

отже, узявши за оцінку параметра σ незсунене значення S , одержують

$$D(\hat{m}) = D(\bar{x}) = \frac{S^2}{N}, \quad D(\hat{\sigma}) = \frac{S^2}{2N}.$$

Таким чином, за методом максимальної правдоподібності знаходять оцінки точності середнього та середньоквадратичного.

У разі реалізації **методу найменших квадратів** точність оцінок параметрів $\hat{\theta}_1$, $\hat{\theta}_2$ впливає з дисперсійно-коваріаційної матриці вигляду

$$DC = S_{\text{Зал}}^2 A^{-1}.$$

Якщо ж для знаходження оцінок параметрів розподілу застосовують **метод моментів**, то для одно- та двопараметричних розподілів оцінки параметрів визначають через моменти \bar{x} і $\overline{x^2}$ як функції вигляду (1.5). При цьому

$$\begin{aligned} D\{\hat{\theta}_i\} = & \left(\frac{\partial H_i}{\partial \bar{x}} \right)^2 D\{\bar{x}\} + \left(\frac{\partial H_i}{\partial \overline{x^2}} \right)^2 D\{\overline{x^2}\} + \\ & + 2 \left(\frac{\partial H_i}{\partial \bar{x}} \right) \left(\frac{\partial H_i}{\partial \overline{x^2}} \right) \text{cov}\{\bar{x}, \overline{x^2}\}, \quad i = 1, 2, \end{aligned}$$

де

$$\begin{aligned} D\{\bar{x}\} &= \frac{v_2 - v_1^2}{N}; & D\{\overline{x^2}\} &= \frac{v_4 - v_2^2}{N}; \\ \text{cov}\{\bar{x}, \overline{x^2}\} &= \frac{v_3 - v_1 v_2}{N}. \end{aligned}$$

Коваріація оцінок параметрів у ММ обчислюється на основі виразу

$$\begin{aligned} \text{cov}\{\hat{\theta}_1, \hat{\theta}_2\} = & \frac{\partial H_1}{\partial \bar{x}} \frac{\partial H_2}{\partial \bar{x}} D\{\bar{x}\} + \frac{\partial H_1}{\partial \overline{x^2}} \frac{\partial H_2}{\partial \overline{x^2}} D\{\overline{x^2}\} + \\ & + \left(\frac{\partial H_1}{\partial \bar{x}} \frac{\partial H_2}{\partial \overline{x^2}} + \frac{\partial H_1}{\partial \overline{x^2}} \frac{\partial H_2}{\partial \bar{x}} \right) \text{cov}\{\bar{x}, \overline{x^2}\}. \end{aligned}$$

1.2.3. Інтервальне оцінювання теоретичної функції розподілу

Довірче оцінювання теоретичної функції розподілу за результатами відтворення здійснюється шляхом призначення довірчого інтервалу, нижня та верхня межі якого знаходяться за виразом

$$F_{\text{н,в}}(x; \bar{\Theta}) = F(x; \hat{\Theta}) \mp u_{\alpha/2} \sqrt{D\left\{F(x; \hat{\Theta})\right\}},$$

де

$D\left\{F(x; \hat{\Theta})\right\}$ – оцінка дисперсії відтвореного розподілу;

$u_{\alpha/2}$ – квантиль нормального розподілу.

Процедура обчислення $D\left\{F\left(x; \hat{\Theta}\right)\right\}$ залежить від кількості параметрів розподілу. У випадку **однопараметричного** розподілу дисперсія статистичної функції розподілу ймовірностей визначається згідно зі співвідношенням

$$D\left\{F\left(x; \hat{\theta}\right)\right\} = \left(\frac{dF}{d\theta}\right)_{\theta=\hat{\theta}}^2 D\left\{\hat{\theta}\right\}. \quad (1.6)$$

Для **двопараметричного** розподілу має місце

$$\begin{aligned} D\left\{F\left(x; \hat{\theta}_1, \hat{\theta}_2\right)\right\} = & \left(\frac{\partial F}{\partial \theta_1}\right)_{\theta_1=\hat{\theta}_1}^2 D\left\{\hat{\theta}_1\right\} + \left(\frac{\partial F}{\partial \theta_2}\right)_{\theta_2=\hat{\theta}_2}^2 D\left\{\hat{\theta}_2\right\} + \\ & + 2\left(\frac{\partial F}{\partial \theta_1}\right)\left(\frac{\partial F}{\partial \theta_2}\right)_{\substack{\theta_1=\hat{\theta}_1 \\ \theta_2=\hat{\theta}_2}} \text{cov}\left\{\hat{\theta}_1, \hat{\theta}_2\right\}. \end{aligned}$$

Значення дисперсій оцінок $D\left\{\hat{\theta}\right\}$, $D\left\{\hat{\theta}_1\right\}$, $D\left\{\hat{\theta}_2\right\}$, $\text{cov}\left\{\hat{\theta}_1, \hat{\theta}_2\right\}$ обчислюють відповідно до третього пункту схеми відтворення розподілу.

1.2.4. Параметричні розподіли

Наведемо приклади деяких класичних розподілів, їх характеристики та найпростіші підходи до відтворення в процесі автоматизації розрахунків.

Експоненціальний розподіл

У задачах надійності, масового обслуговування та оцінки рідкісних явищ найчастіше застосовується експоненціальний розподіл.

До характеристик експоненціального розподілу належать функції:

1) щільності розподілу ймовірностей (рис. 1.11)

$$f(x; \lambda) = \begin{cases} 0, & -\infty < x < 0, \\ \lambda \exp(-\lambda x), & 0 \leq x < \infty; \end{cases}$$

2) розподілу ймовірностей

$$F(x; \lambda) = \begin{cases} 0, & -\infty < x < 0, \\ 1 - \exp(-\lambda x), & 0 \leq x < \infty. \end{cases}$$

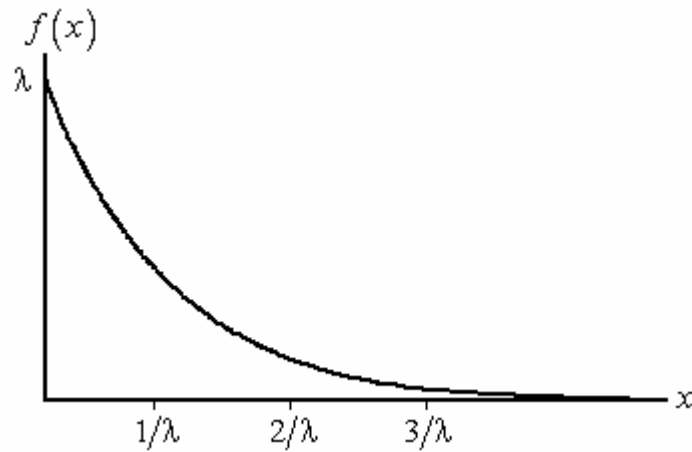


Рис. 1.11. Графік функції щільності експоненціального розподілу

Важливими характеристиками розподілу є:

1) математичне сподівання

$$E\{\xi\} = \frac{1}{\lambda};$$

2) дисперсія

$$D\{\xi\} = \frac{1}{\lambda^2};$$

3) коефіцієнт асиметрії

$$A = \frac{\mu_3}{\mu_2^{3/2}} = 2;$$

4) коефіцієнт ексцесу

$$E = \frac{\mu_4}{\mu_2^2} - 3 = 6.$$

У процесі відтворення функції розподілу за вибіркою $\Omega_{1,N}$ виникає необхідність знаходження значення параметра λ . Його обчислюють за методом моментів:

$$\hat{\lambda} = \frac{1}{\bar{x}} = H(\bar{x}).$$

Точність оцінки $\hat{\lambda}$ визначають у такий спосіб:

$$D\{\hat{\lambda}\} = \left(\frac{dH}{d\bar{x}} \right)^2 D\{\bar{x}\} = (-\hat{\lambda}^2)^2 \frac{1}{\hat{\lambda}^2 N} = \frac{\hat{\lambda}^2}{N},$$

де

$$D\{\bar{x}\} = \frac{1}{\hat{\lambda}^2 N}.$$

Довірче оцінювання $F(x)$ виконується за формулою (1.6) з урахуванням

$$D\{F(x; \hat{\lambda})\} = \left(\frac{dF}{d\lambda} \right)_{\lambda=\hat{\lambda}}^2 D\{\hat{\lambda}\} = x^2 \exp(-2\hat{\lambda}x) \frac{\hat{\lambda}^2}{N}.$$

Розподіл Релея

Розподіл модуля вектора на площині, координати якого є незалежними випадковими величинами, що мають нормальний закон розподілу з нульовим середнім та одиничною дисперсією, описується за законом Релея. Розподіл Релея реалізують, коли похибки вимірювання за координатами x та y незалежні і нормально розподілені з однаковими дисперсіями.

Головні характеристики розподілу Релея:

1) функція щільності розподілу ймовірностей (рис. 1.12)

$$f(x; \sigma) = \frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right), \quad 0 \leq x < \infty;$$

2) функція розподілу ймовірностей

$$F(x; \sigma) = 1 - \exp\left(-\frac{x^2}{2\sigma^2}\right), \quad 0 \leq x < \infty.$$

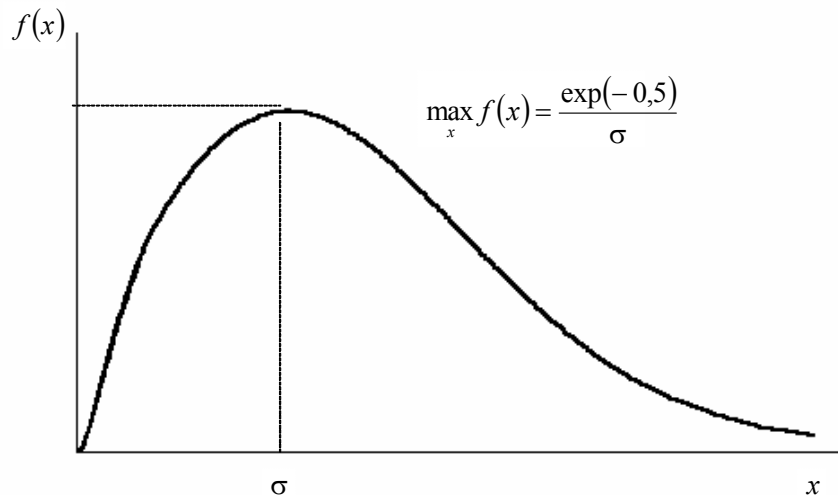


Рис.1.12. Графік функції щільності розподілу Релея

Кількісні характеристики розподілу Релея знаходять так:

1) математичне сподівання $E\{\xi\} = 1,25\sigma;$

2) дисперсія $D\{\xi\} = 0,43\sigma^2;$

- 3) коефіцієнт асиметрії $A = 0,63$;
 4) коефіцієнт ексцесу $E = -0,3$.

Параметр $\hat{\sigma}$ оцінюють за формулами:

$$\hat{\sigma} = \frac{2}{\sqrt{\pi}} \bar{x}, \quad D\{\hat{\sigma}\} = \frac{4}{\pi} D\{\bar{x}\}.$$

При визначенні довірчих інтервалів для функції розподілу використовують вираз

$$D\{F(x; \hat{\sigma})\} = \left(-\frac{x^3}{\hat{\sigma}^3} \exp\left(-\frac{x^2}{2\hat{\sigma}^2}\right) \right)^2 D\{\hat{\sigma}\}.$$

Нормальний розподіл

В основі практично всієї класичної теорії ймовірностей та прикладного статистичного аналізу лежить нормальний розподіл, який є межевою формою численних розподілів.

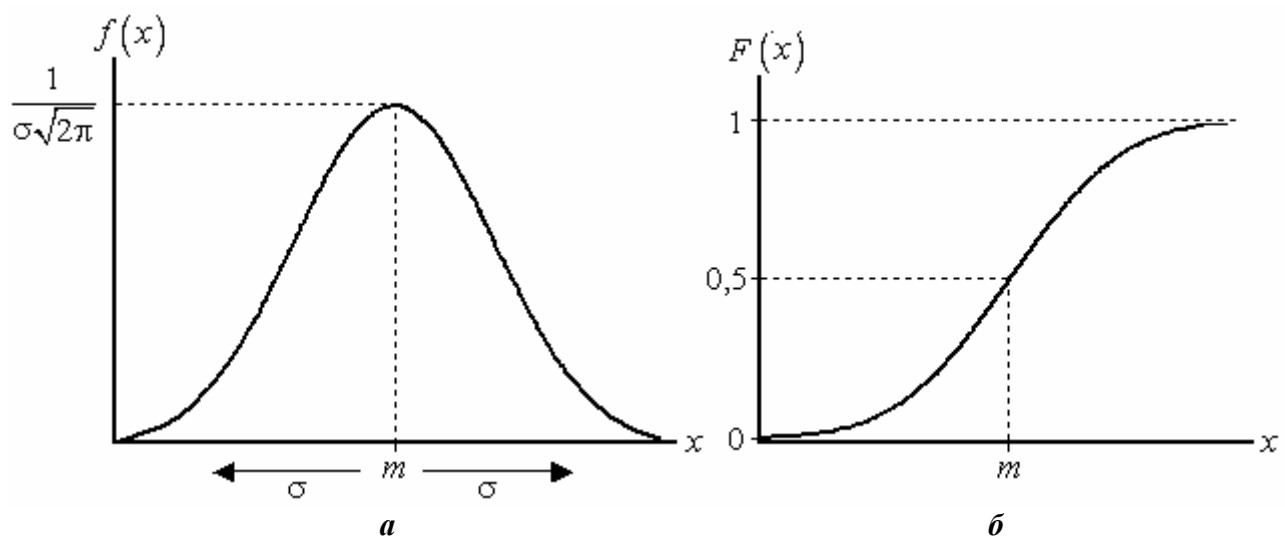


Рис. 1.13. Графіки функцій нормального розподілу:
 а – функції щільності розподілу; б – функції розподілу

Головні характеристики нормального розподілу такі:

- 1) функція щільності розподілу ймовірностей (рис. 1.13, а)

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

- 2) функція розподілу ймовірностей (рис. 1.13, б)

$$F(x; m, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{(u-m)^2}{2\sigma^2}\right) du = \Phi\left(\frac{x-m}{\sigma}\right),$$

де $\Phi(\cdot)$ – функція Лапласа.

У літературі функція нормального розподілу часто позначається таким чином:

$$F(x; m, \sigma) \equiv N(x; m, \sigma).$$

До кількісних характеристик розподілу належать:

1) математичне сподівання

$$E\{\xi\} = m;$$

2) дисперсія

$$D\{\xi\} = \sigma^2;$$

3) коефіцієнт асиметрії

$$A = 0;$$

4) коефіцієнт ексцесу

$$E = 0, \hat{E} = 3.$$

Оцінки параметрів нормального розподілу мають вигляд

$$\hat{m} = \bar{x}, \quad \hat{\sigma} = \frac{N}{N-1} \sqrt{x^2 - \bar{x}^2}.$$

Дисперсії оцінок параметрів обчислюють згідно зі співвідношеннями

$$D\{\hat{m}\} = \frac{\hat{\sigma}^2}{N}, \quad D\{\hat{\sigma}\} = \frac{\hat{\sigma}^2}{2N},$$

$$\text{cov}\{\hat{m}, \hat{\sigma}\} = 0.$$

Довірчі інтервали для функції розподілу знаходять з урахуванням виразів

$$\frac{\partial F}{\partial m} = -\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right),$$

$$\frac{\partial F}{\partial \sigma} = -\frac{x-m}{\sigma^2\sqrt{2\pi}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right).$$

Для визначення функції Лапласа припустима апроксимація:

$$\Phi(u) = 1 - \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) (b_1 t + b_2 t^2 + b_3 t^3 + b_4 t^4 + b_5 t^5) + \varepsilon(u),$$

де

$$u \geq 0; \quad t = \frac{1}{1 + \rho u};$$

$$\rho = 0,231\,641\,9; \quad |\varepsilon(u)| \leq 7,8 \cdot 10^{-8};$$

$$b_1 = 0,319\,381\,53; \quad b_2 = -0,356\,563\,782; \quad b_3 = 1,781\,477\,937;$$

$$b_4 = -1,821\,255\,978; \quad b_5 = 1,330\,274\,429.$$

У разі, якщо $u < 0$, слушне співвідношення

$$\Phi(u) = 1 - \Phi(|u|).$$

Квантилі $u_{\alpha/2}$ нормального розподілу можна визначити так:

$$u_p = t - \frac{c_0 + c_1 t + c_2 t^2}{1 + d_1 t + d_2 t^2 + d_3 t^3} + \varepsilon_\alpha, \quad (1.7)$$

де

$$p = \alpha/2; \quad t = \sqrt{\ln \frac{1}{p^2}}; \quad |\varepsilon_\alpha| \leq 4,5 \cdot 10^{-4};$$

$$c_0 = 2,515\,517; \quad c_1 = 0,802\,853; \quad c_2 = 0,010\,328;$$

$$d_1 = 1,432\,788; \quad d_2 = 0,189\,265\,9; \quad d_3 = 0,001\,308.$$

Під час розв'язання задачі моделювання випадкових величин постає потреба в обчисленні одnobічного квантиля u_α (табл. Б.1). Його визначення здійснюється на основі виразу (1.7) з урахуванням того, що за $\alpha \leq 0,5$

$$u_\alpha = -u_p, \quad \text{де } p = \alpha,$$

при $\alpha > 0,5$

$$u_\alpha = u_p,$$

де $p = 1 - \alpha$.

Розподіл Лапласа

Розподіл Лапласа описує різницю двох незалежних випадкових величин $\xi = \xi_1 - \xi_2$, які мають експоненціальний розподіл.

Розподіл Лапласа характеризують такі функції:

1) щільності розподілу ймовірностей (рис. 1.14)

$$f(x; \lambda, \mu) = \frac{\lambda}{2} \exp(-\lambda|x - \mu|);$$

2) розподілу ймовірностей

$$F(x; \lambda, \mu) = \begin{cases} \frac{1}{2} \exp(\lambda(x - \mu)), & -\infty < x \leq \mu, \\ 1 - \frac{1}{2} \exp(-\lambda(x - \mu)), & \mu < x < \infty. \end{cases}$$

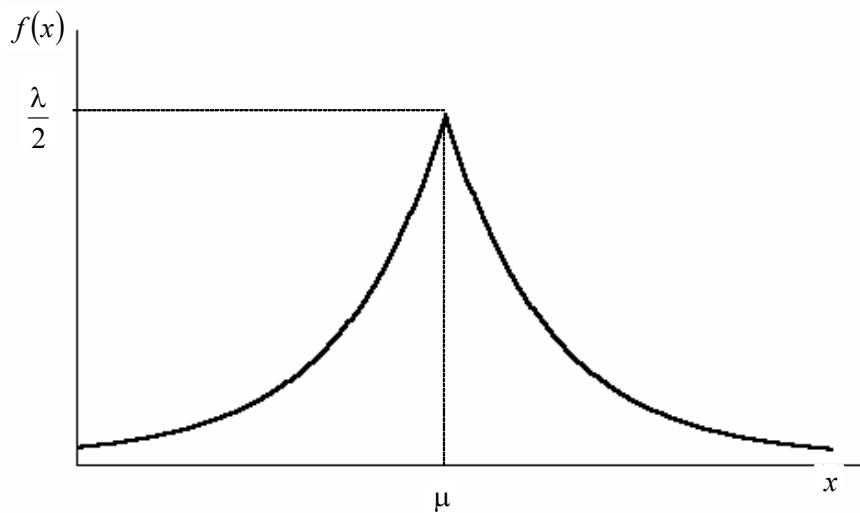


Рис.1.14. Графік функції щільності розподілу Лапласа

До кількісних характеристик розподілу Лапласа належать:

- 1) математичне сподівання $E\{\xi\} = \mu$;
- 2) дисперсія $D\{\xi\} = \frac{2}{\lambda^2}$;
- 3) коефіцієнт асиметрії $A = 0$;
- 4) коефіцієнт ексцесу $E = 3$.

Оцінювання параметрів виконують за методом моментів, згідно з яким

$$\hat{\mu} = \bar{x},$$

$$\hat{\lambda} = \frac{\sqrt{2}}{\sqrt{x^2 - \bar{x}^2}}.$$

Дисперсії оцінок параметрів розподілу Лапласа мають вигляд

$$D\{\hat{\lambda}\} = \frac{5\hat{\lambda}^2}{N},$$

$$D\{\hat{\mu}\} = \frac{2}{N\hat{\lambda}^2},$$

$$\text{cov}\{\hat{\lambda}, \hat{\mu}\} = -\frac{3}{2N}.$$

Довірче оцінювання функції розподілу виконують за виразами

$$\frac{\partial F}{\partial \lambda} = \begin{cases} \frac{1}{2}(x - \mu) \exp(\lambda(x - \mu)), & -\infty < x \leq \mu, \\ \frac{1}{2}(x - \mu) \exp(-\lambda(x - \mu)), & \mu < x < \infty, \end{cases}$$

$$\frac{\partial F}{\partial \mu} = \begin{cases} -\frac{\lambda}{2} \exp(\lambda(x - \mu)), & -\infty < x \leq \mu, \\ -\frac{\lambda}{2} \exp(-\lambda(x - \mu)), & \mu < x < \infty. \end{cases}$$

Розподіл Вейбулла

Розподіл Вейбулла можна назвати найбільш універсальним серед вищезгаданих. Залежно від параметра β його функція щільності може бути унімодальною ($\beta \leq 1$) чи одномодальною ($\beta > 1$), а одномодальна – симетричною, правасиметричною або лівасиметричною. Якщо $\beta = 1$ та $\alpha = 1/\lambda$, то розподіл Вейбулла зводиться до експоненціального.

Розподіл Вейбулла характеризують такі функції:

1) щільності розподілу ймовірностей (рис. 1.15)

$$f(x; \alpha, \beta) = \frac{\beta}{\alpha} x^{\beta-1} \exp\left(-\frac{x^\beta}{\alpha}\right), \quad 0 \leq x < \infty, \quad \alpha, \beta > 0;$$

2) розподілу ймовірностей

$$F(x; \alpha, \beta) = 1 - \exp\left(-\frac{x^\beta}{\alpha}\right), \quad 0 \leq x < \infty, \quad \alpha, \beta > 0.$$

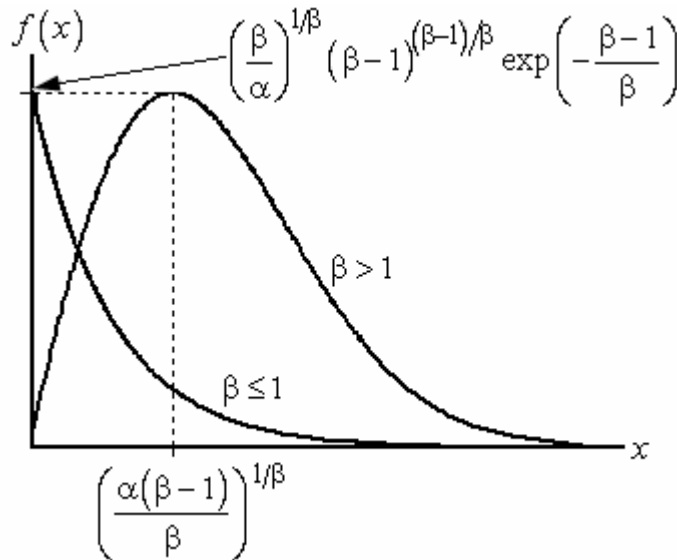


Рис. 1.15. Графік функції щільності розподілу Вейбулла

Кількісними характеристиками розподілу є:

1) математичне сподівання

$$E\{\xi\} = \alpha^{2/\beta} \Gamma\left(1 + \frac{1}{\beta}\right);$$

2) дисперсія

$$D\{\xi\} = \alpha^{2/\beta} \left(\Gamma\left(1 + \frac{2}{\beta}\right) - \Gamma^2\left(1 + \frac{1}{\beta}\right) \right);$$

3) коефіцієнт асиметрії

$$A = \frac{\mu_3}{\mu_2^{3/2}};$$

4) коефіцієнт ексцесу

$$E = \frac{\mu_4}{\mu_2^2} - 3.$$

Оскільки аналітичний вигляд функції розподілу Вейбулла можна звести до лінійної форми, то найпростіше визначати оцінки параметрів цього розподілу за методом найменших квадратів.

Зводячи функцію розподілу до лінійної форми

$$\ln\left(\ln\frac{1}{1-F(x)}\right) = -\ln\alpha + \beta\ln x,$$

одержуємо процедуру знаходження оцінок параметрів $\hat{\alpha}$, $\hat{\beta}$ з умови мінімуму залишкової дисперсії у вигляді

$$S_{3ал}^2 = \frac{1}{N-3} \sum_{l=1}^{N-1} \left(\ln \left(\ln \frac{1}{1-F_{1,N}(x_l)} \right) - \hat{A} - \hat{\beta} \ln x_l \right)^2,$$

де

$$\hat{A} = -\ln \hat{\alpha},$$

звідси

$$\hat{\alpha} = \exp(-\hat{A}).$$

З урахуванням умов мінімуму необхідне розв'язання системи рівнянь

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} \hat{A} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix},$$

де

$$a_{11} = N-1; \quad a_{12} = a_{21} = \sum_{l=1}^{N-1} \ln x_l; \quad a_{22} = \sum_{l=1}^{N-1} \ln^2 x_l;$$

$$b_1 = \sum_{l=1}^{N-1} \ln \left(\ln \frac{1}{1-F_{1,N}(x_l)} \right); \quad b_2 = \sum_{l=1}^{N-1} \ln x_l \ln \left(\ln \frac{1}{1-F_{1,N}(x_l)} \right).$$

Тоді дисперсії оцінок параметрів такі:

$$D\{\hat{A}\} = \frac{a_{22}S_{3ал}^2}{a_{11}a_{22} - a_{12}a_{21}},$$

$$D\{\hat{\beta}\} = \frac{a_{11}S_{3ал}^2}{a_{11}a_{22} - a_{12}a_{21}}.$$

Коваріацію визначаємо за співвідношенням

$$\text{cov}\{\hat{A}, \hat{\beta}\} = -\frac{a_{21}S_{3ал}^2}{a_{11}a_{22} - a_{12}a_{21}} = -\frac{a_{12}S_{3ал}^2}{a_{11}a_{22} - a_{12}a_{21}}.$$

Беручи до уваги зв'язок між $\hat{\alpha}$ та \hat{A} , маємо

$$D\{\hat{\alpha}\} = \exp(-2\hat{A}) \cdot D\{\hat{A}\},$$

$$\text{cov}\{\hat{\alpha}, \hat{\beta}\} = -\exp(\hat{A}) \cdot \text{cov}\{\hat{A}, \hat{\beta}\}.$$

Довірчі інтервали для функції розподілу призначають з огляду на такі формули для частинних похідних:

$$\frac{\partial F}{\partial \alpha} = -\frac{x^\beta}{\alpha^2} \exp\left(-\frac{x^\beta}{\alpha}\right),$$

$$\frac{\partial F}{\partial \beta} = \frac{x^\beta}{\alpha} \ln x \exp\left(-\frac{x^\beta}{\alpha}\right).$$

Рівномірний розподіл

Рівномірний розподіл є статистична модель, що описує події, які з однаковою ймовірністю можуть з'явитись у будь-який момент у заданому інтервалі. Якщо апріорно невідомий тип розподілу випадкової величини, то часто вважають, що має місце або рівномірний, або нормальний розподіл.

Головними характеристиками рівномірного розподілу є функції:

1) щільності розподілу ймовірностей (рис. 1.14, а)

$$f(x; a, b) = \begin{cases} 0, & -\infty < x < a, \\ \frac{1}{b-a}, & a \leq x < b, \\ 0, & b \leq x < \infty, \end{cases}$$

2) розподілу ймовірностей (рис. 1.14, б)

$$F(x; a, b) = \begin{cases} 0, & -\infty < x < a, \\ \frac{x-a}{b-a}, & a \leq x < b, \\ 1, & b \leq x < \infty. \end{cases}$$

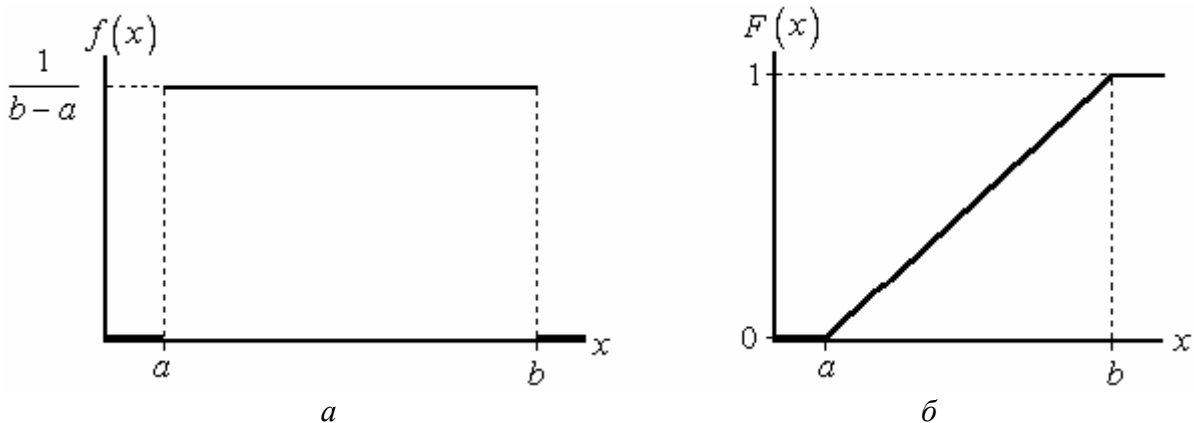


Рис. 1.14. Графік функцій рівномірного розподілу:
а – функції щільності розподілу; б – функції розподілу

Кількісним характеристикам розподілу відповідають такі співвідношення:

1) математичне сподівання

$$E\{\xi\} = \frac{a+b}{2};$$

2) дисперсія

$$D\{\xi\} = \frac{(b-a)^2}{12};$$

3) коефіцієнт асиметрії

$$A = 0;$$

4) коефіцієнт ексцесу

$$E = -1,2.$$

Оцінки параметрів функції рівномірного розподілу визначають за методом моментів, згідно з яким мають таку систему лінійних рівнянь відносно a і b :

$$\begin{cases} \frac{a+b}{2} = \bar{x}, \\ \frac{b-a}{2\sqrt{3}} = \sqrt{\overline{x^2} - \bar{x}^2}. \end{cases}$$

Отже,

$$\hat{a} = \bar{x} - \sqrt{3(\overline{x^2} - \bar{x}^2)} = H_1(\bar{x}, \overline{x^2}),$$

$$\hat{b} = \bar{x} + \sqrt{3(\overline{x^2} - \bar{x}^2)} = H_2(\bar{x}, \overline{x^2}).$$

Значення $D\{\hat{a}\}$, $D\{\hat{b}\}$, $\text{cov}\{\hat{a}, \hat{b}\}$ знаходять з урахуванням таких виразів:

$$\frac{\partial H_1}{\partial \bar{x}} = 1 + 3 \frac{\hat{a} + \hat{b}}{\hat{b} - \hat{a}},$$

$$\frac{\partial H_1}{\partial \overline{x^2}} = -\frac{3}{\hat{b} - \hat{a}},$$

$$\frac{\partial H_2}{\partial \bar{x}} = 1 - 3 \frac{\hat{a} + \hat{b}}{\hat{b} - \hat{a}},$$

$$\frac{\partial H_2}{\partial \overline{x^2}} = \frac{3}{\hat{b} - \hat{a}},$$

$$D\{\bar{x}\} = \frac{(\hat{b} - \hat{a})^2}{12N},$$

$$\text{cov}\{\bar{x}, \overline{x^2}\} = \frac{(\hat{a} + \hat{b})(\hat{b} - \hat{a})^2}{12N},$$

$$D\{\overline{x^2}\} = \frac{1}{180N} \left((\hat{b} - \hat{a})^4 + 15(\hat{a} + \hat{b})^2 (\hat{b} - \hat{a})^2 \right).$$

Призначаючи довірчі інтервали для функції розподілу, беруть до уваги вираз

$$D\{F(x; \hat{a}, \hat{b})\} = \frac{(x - \hat{b})^2}{(\hat{b} - \hat{a})^4} D\{\hat{a}\} + \frac{(x - \hat{a})^2}{(\hat{b} - \hat{a})^4} D\{\hat{b}\} - 2 \frac{(x - \hat{a})(x - \hat{b})}{(\hat{b} - \hat{a})^4} \text{cov}\{\hat{a}, \hat{b}\}.$$

Контрольні запитання та завдання

1. Що таке об'єкт спостережень? Які його ознаки?
2. Що називається даними? Які існують класифікації типів даних?
3. Дати визначення випадкової величини та функції розподілу, навести їх властивості.
4. У чому полягає різниця між вибіркою та генеральною сукупністю? Яка вибірка називається репрезентативною?
5. Для функції розподілу ймовірностей неперервної випадкової величини довести $F(a) \leq F(b)$ за умови $a < b$.
6. У чому полягає різниця між параметром генеральної сукупності та оцінкою параметра?
7. Сформулювати властивості оцінок параметрів.
8. У який спосіб визначається кількість класів за гістограмною оцінкою?
9. Яка сутність візуальної ідентифікації моделі розподілу за гістограмою?
10. У чому полягає зв'язок між функцією щільності та відносною частотою варіаційного ряду, розбитого на класи?
11. Як використовується та в яких одиницях вимірюється коефіцієнт варіації?
12. Що показує середньоквадратичне відхилення вибірки?
13. Чому дорівнює результат ділення зсуненої оцінки коефіцієнта асиметрії на незсунену?
14. Як призначається довірчий інтервал на параметр генеральної сукупності?
15. Визначити довірчий інтервал на середнє та середньоквадратичне вибірки.
16. Чому дорівнює середньоквадратичне відхилення середньоквадратичного відхилення, одержаного на основі одновимірної вибірки?
17. Чому дорівнює зсунена оцінка середньоквадратичного відхилення суми елементів вибірки?
18. Для якої вибірки незсунена оцінка середньоквадратичного відхилення дорівнює 1?

19. Обсяг вибірки 100. Визначити межі 90%-го довірчого інтервалу для незсуненої оцінки середньоквадратичного відхилення.

20. Як зміниться середньоквадратичне відхилення в результаті додавання та множення сталої до кожної варіанти?

21. Дати визначення квантиля.

22. Які результати спостережень називають аномальними? У який спосіб вони вилучаються з процесу обробки?

23. Чим відрізняється емпірична функція розподілу від теоретичної та відтвореної статистичної?

24. Указати, чому дорівнює значення функції Лапласа в точці 0.

25. Чим відрізняється функція Лапласа від функції розподілу Лапласа?

26. Записати функцію вибірки та умови досягнення максимуму функції вибірки для розподілу Вейбулла та нормального розподілу.

27. Записати функцію щільності нормального розподілу з параметрами $(0;1)$.

28. Схарактеризувати модель експоненціального закону розподілу та її властивості, указати приклади застосування.

29. Вказати зворотню функцію до функції розподілу Релея.

30. Звести до лінійної форми функцію розподілу Вейбулла.

31. Навести модель рівномірного розподілу, графіки функцій щільності та розподілу.

32. Визначити оцінку параметра експоненціального закону розподілу за методом найменших квадратів.

33. Реалізувати метод максимальної правдоподібності для оцінки параметрів експоненціального розподілу.

34. За допомогою методу найменших квадратів знайти оцінки параметрів розподілу Вейбулла.

Розділ 2. ПЕРЕВІРКА СТАТИСТИЧНИХ ГІПОТЕЗ

Розглянемо основи теорії статистичних гіпотез. За класичним підходом подамо обчислювальні процедури для розв’язання однієї із задач перевірки статистичних гіпотез – задачі перевірки однорідності статистичних масивів даних. Необхідність наведення такого матеріалу пов’язана із завданням формування об’єднаних масивів даних представницького обсягу для подальшої обробки, наприклад, під час відтворення розподілів. Також охарактеризуємо критерії згоди відтворення розподілів.

2.1. Головні поняття та визначення

Дослідження законів розподілу статистик дозволяє робити висновок відносно ймовірності появи значення конкретно обчисленої статистики. Такий висновок дозволяє говорити, наприклад, про адекватність оцінки параметра, або про вірогідність того чи іншого припущення (гіпотези) відносно об’єкта дослідження.

Статистична гіпотеза – це будь-яке припущення щодо функції частот (функції щільності розподілу ймовірностей) або кількісних характеристик спостережуваних змінних.

У теорії перевірки статистичних гіпотез вихідні є поняття **головної** та **конкуруючої (альтернативної) гіпотез**. Конкуруючих гіпотез може бути більше однієї. Розглянемо формальні визначення таких гіпотез.

Нехай маємо сукупність (множину) $\Omega_N \subset \Omega$ реалізацій випадкової величини ξ (для визначеності – неперервної випадкової величини). Із курсу теорії ймовірностей відомо, що під час роботи з випадковими величинами можемо говорити про існування закону розподілу випадкової величини $F(X)$, де $X \in \mathbb{R}_m$, $m \geq 1$. У більшості випадків вигляд $F(X)$ невідомий, проте є вказівка стосовно належності функції розподілу до деякого класу \mathfrak{F} . Будемо вважати, що розподіли, які входять до класу \mathfrak{F} , відрізняються значеннями деякого вектора параметрів $\bar{\Theta}$. Іншими словами, для генеральної сукупності існує такий вектор параметрів, значення якого визначають функцію розподілу випадкової величини ξ через функціональну залежність величини ймовірності від параметрів та реалізацій

$$P\{\xi < X\} = F(X; \bar{\Theta}),$$

і якщо

$$\vec{\Theta}_1 \neq \vec{\Theta}_2,$$

то

$$F(X; \vec{\Theta}_1) \neq F(X; \vec{\Theta}_2).$$

Нехай $\varpi \in R_s$, $s \geq 1$ – множина всіх можливих значень вектора параметрів $\vec{\Theta} = \{\theta_1, \dots, \theta_s\}$. Розглянемо розбиття ϖ на дві підмножини:

$$\varpi = \varpi_0 \cup \varpi_1,$$

причому до ϖ_0 входить деякий вектор параметрів $\vec{\Theta}$

$$\vec{\Theta} \in \varpi_0,$$

а до ϖ_1 – не входить:

$$\vec{\Theta} \notin \varpi_1.$$

Гіпотезу H_0 називають головною, якщо вона характеризує розподіл $F(X; \vec{\Theta})$, де $\vec{\Theta} \in \varpi_0$, у протилежному разі ($\vec{\Theta} \in \varpi_1$) гіпотезу H_1 називають конкуруючою (альтернативною).

Усі гіпотези поділяються на **прості** та **складні**. Гіпотезу називають простою, якщо вона без будь-яких винятків визначає розподіл $F(X; \vec{\Theta})$, у протилежному випадку маємо складну гіпотезу.

Якщо $\vec{\Theta}$ визначає точку множини ϖ :

$$\varpi : \vec{\Theta} = \{\theta_1, \dots, \theta_s\},$$

то гіпотеза є проста, якщо ж визначає область

$$\vec{\Theta} = \{\underline{\theta}_1 < \theta_1 < \overline{\theta}_1, \dots, \underline{\theta}_s < \theta_s < \overline{\theta}_s\}$$

множини ϖ , то гіпотеза є складна.

Наприклад, гіпотеза $H_0 : \lambda = 0, 2$, де λ – параметр експоненціального розподілу, проста, а гіпотеза $H_0 : \lambda \in [0, 1; 0, 3]$ – складна.

Головна гіпотеза H_0 являє собою твердження відносно вектора параметрів $\vec{\Theta}$, яке приймається тоді, коли немає переконливих аргументів для його відхилення. Альтернативну гіпотезу H_1 приймають тільки за наявності статистичного доведення, яке відхиляє нульову гіпотезу.

У зв'язку з тим, що значення $\vec{\Theta}$ наперед невідоме, головна гіпотеза формулюється в термінах статистик, а саме: робиться припущення щодо рівності

величині $\vec{\Theta}$, одержаній на основі вибірки Ω_N , оцінки вектора параметрів $\hat{\vec{\Theta}}$ (за будь-якої альтернативи):

$$H_0 : \vec{\Theta} = \hat{\vec{\Theta}}.$$

Зазначимо, що в даному випадку мова йде про вибірку Ω_N будь-якої розмірності.

Оскільки оцінка $\hat{\vec{\Theta}}$ – випадкова величина, існує можливість побудови правила перевірки гіпотез, виходячи з аналізу законів розподілу $\hat{\vec{\Theta}}$. **Статистичним критерієм** називають без винятків визначене правило обробки статистичного матеріалу (або правило аналізу закону розподілу оцінок параметрів), на основі якого одна з гіпотез приймається, а всі інші відхиляються.

Формальні деталі процедури перевірки гіпотези визначають у термінах різних допустимих помилок. Оскільки рішення про прийняття чи відхилення гіпотези приймається на основі вибірки (опосередковано через функцію вибірки – статистику), існує ймовірність припуститися помилки, бо гіпотеза являє собою твердження про генеральну сукупність Ω . В основі кожного з типів розглянутих нижче помилок лежать різні припущення відносно того, яка з гіпотез дійсно є правильна.

Розрізняють **помилки першого та другого роду**. За неформальним визначенням помилка першого роду полягає в тому, що гіпотеза H_0 відхиляється, коли насправді вона правильна. За помилки другого роду гіпотезу H_0 приймають тоді, коли вона не є істинна. Кількісно помилку оцінюють за ймовірністю. Із зазначеного випливає, що в процесі перевірки гіпотези може виникнути одна з таких ситуацій (табл. 2.1):

- 1) гіпотеза H_0 правильна й приймається;
- 2) має місце гіпотеза H_1 , проте приймається гіпотеза H_0 , яка не є правильною (помилка другого роду);
- 3) має місце гіпотеза H_0 , проте приймається гіпотеза H_1 , яка не є правильною (помилка першого роду);
- 4) гіпотеза H_1 правильна й приймається.

Таблиця 2.1

Прийняття рішень щодо гіпотез

Рішення	Гіпотеза H_0 правильна	Гіпотеза H_1 правильна
Прийняти H_0	Правильно	Неправильно (помилка другого роду)
Відхилити H_0	Неправильно (помилка першого роду)	Правильно

Відносно гіпотези ніколи не говорять, що вона «ймовірно» правильна чи неправильна. Мова йде про помилки та про ймовірності помилок (α та β відповідно для помилок першого та другого роду). У гіпотезі нема нічого випадкового, проте вибір гіпотези є випадковий, так само як і вибіркова статистика. Посправжньому правильна гіпотеза є невідома, як і вектор параметрів генеральної сукупності.

Нехай маємо головну гіпотезу H_0 , за якою стверджується, що для вибірки Ω_N існує конкретний вигляд розподілу $F(X; \vec{\Theta})$. Отже, виникає задача про перевірку гіпотези

$$H_0 : \vec{\Theta} \in \varpi_0$$

або

$$H_0 : \theta_1 = \hat{\theta}_1, \dots, \theta_s = \hat{\theta}_s$$

за однієї або кількох альтернатив: $H_k, k \geq 1$.

Функцією потужності $W(\vec{\Theta})$ критерію називають імовірність того, що головна гіпотеза H_0 буде відхилена в той час, як буде правильна альтернатива H_1 :

$$W(\vec{\Theta}) = P\{\Omega_N \in \Omega_1 / \vec{\Theta}\},$$

де множина Ω_1 – критична область.

Критичною областю Ω_1 називають множину можливих значень результатів експерименту (або множину можливих значень оцінки вектора параметрів чи статистичної характеристики), при яких головна гіпотеза H_0 відхиляється. **Статистична характеристика гіпотези** – це функція вибірки, на основі якої перевіряється головна гіпотеза H_0 . Як правило, до статистичних характеристик відносять різні перетворення над оцінками параметрів або ж інші статистики.

Із наведених визначень випливає, що вся множина Ω експерименту має розбиття на дві підмножини Ω_0 і Ω_1 , що не перетинаються. Вибірка $\Omega_N \in \Omega$ може належати як до Ω_0 , так і до Ω_1 . Сказане стосується, наприклад, простих гіпотез H_0 та H_1 : якщо вибірка $\Omega_N \in \Omega_0$, то гіпотеза H_0 приймається, а якщо $\Omega_N \in \Omega_1$, вона відхиляється і приймається H_1 . Множину Ω_0 називають **допустимою областю** прийняття гіпотези H_0 .

Поняття функції потужності, статистичного критерію, допустимої та критичної областей дозволяють формально визначити помилки першого та другого роду через імовірності α, β . Не зменшуючи загальності, розглянемо клас однопараметричних функцій розподілу ймовірностей $F(X; \vec{\Theta})$, де $\vec{\Theta} = \{\theta\}$, а також просту гіпотезу $H_0 : \theta_0 = \hat{\theta}$ і в протиставленні їй альтернативну гіпотезу

$H_1 : \theta_1 = \hat{\theta}$. Тоді на основі вибірки необхідно визначити $\vec{\Theta} \in \varpi_0$ або $\vec{\Theta} \in \varpi_1$, тобто $\Omega_N \in \Omega_0$ або $\Omega_N \in \Omega_1$.

За вибіркою Ω_N завжди можна одержати оцінку параметра $\hat{\theta}$ з функцією щільності розподілу ймовірностей $f(\hat{\theta})$. Вважаючи, що одержана оцінка параметра $\hat{\theta}$ має властивість незсуненості, доходять висновку, що функція щільності $f(\hat{\theta})$ є симетрична, а закон розподілу оцінки $\hat{\theta}$ близький до нормального $N(\hat{\theta}; E\{\hat{\theta}\}; \sigma\{\hat{\theta}\})$. Якщо виявиться, що гіпотеза $H_0 : \theta_0 = \hat{\theta}$ правильна, то функція щільності розподілу $f(\hat{\theta})$ буде мати максимум у точці, що визначає параметр θ_0 (рис. 2.1).

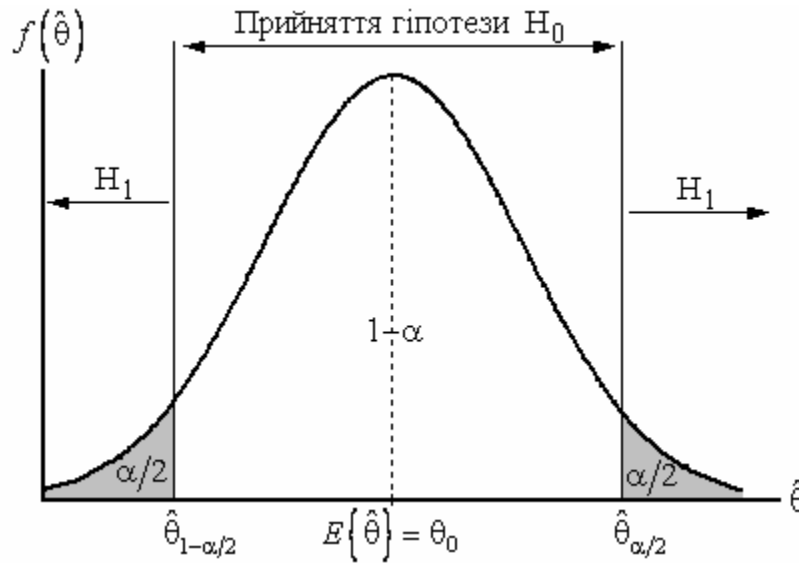


Рис. 2.1. Графік функції щільності розподілу оцінки параметра $\hat{\theta}$ за умови правильності гіпотези H_0

Тоді помилка першого роду визначається згідно з таким виразом:

$$\begin{aligned} \alpha &= P\{\Omega_N \in \Omega_1 / \theta_0\} = P\{\varpi_1 | H_0\} = P\{\hat{\theta} < \hat{\theta}_{1-\alpha/2}\} + P\{\hat{\theta} > \hat{\theta}_{\alpha/2}\} = \\ &= \int_{-\infty}^{\hat{\theta}_{1-\alpha/2}} f(\hat{\theta}) d\hat{\theta} + \int_{\hat{\theta}_{\alpha/2}}^{\infty} f(\hat{\theta}) d\hat{\theta}. \end{aligned}$$

При цьому ймовірність правильного рішення, яке полягає в прийнятті гіпотези H_0 , визначається так:

$$P\{\Omega_N \in \Omega_0 / \theta_0\} = P\{\varpi_0 | H_0\} = P\{\hat{\theta}_{1-\alpha/2} \leq \hat{\theta} \leq \hat{\theta}_{\alpha/2}\} = \int_{\hat{\theta}_{1-\alpha/2}}^{\hat{\theta}_{\alpha/2}} f(\hat{\theta}) d\hat{\theta}.$$

Імовірність помилки першого роду α називають **рівнем значущості**, у практичних задачах її задають у вигляді чисел: 0,1; 0,05; 0,01; 0,001 та ін. У разі прийняття альтернативної гіпотези говорять, що результат перевірки є статистично значущий на рівні α . У літературі та програмному забезпеченні довірчу ймовірність $1 - \alpha$ рішення, відповідно до якої приймають головну гіпотезу, називають **p-значенням**.

Для перевірки гіпотези $H_0 : \theta_0 = \hat{\theta}$ про рівність параметра значенню оцінки параметра (інша назва такої перевірки – **t-тест**) вводять статистичну характеристику гіпотези t , що являє собою стандартизоване значення оцінки $\hat{\theta}$:

$$t = \frac{\theta_0 - \hat{\theta}}{\sigma\{\hat{\theta}\}}.$$

Відомо, що при $N \rightarrow \infty$ для вибірки $\Omega_{1,N}$ статистика t має нормальний розподіл $N(t; 0, 1)$, якщо ж N скінченне, то t розподіляється за законом Стюдента з кількістю степенів вільності $v = N - 1$ (табл. Б.2). В останньому випадку довірна ймовірність того, що оцінка $\hat{\theta}$ збігається з величиною параметра θ_0 , визначається так:

$$1 - \alpha = P\{t_{1-\alpha/2, v} \leq t \leq t_{\alpha/2, v}\} = P\{|t| \leq t_{\alpha/2, v}\}.$$

Уведення t -статистики дозволяє сформулювати загальне правило побудови довірчих інтервалів для незсунених оцінок параметрів (іншими словами – проведення інтервальної оцінки параметрів на основі t -тесту). Із нерівності

$$|t| \leq t_{\alpha/2, v}$$

одержують

$$\hat{\theta} - t_{\alpha/2, v} \cdot \sigma\{\hat{\theta}\} \leq \theta_0 \leq \hat{\theta} + t_{\alpha/2, v} \cdot \sigma\{\hat{\theta}\},$$

зокрема, при $N > 60$

$$\hat{\theta} - u_{\alpha/2} \cdot \sigma\{\hat{\theta}\} \leq \theta_0 \leq \hat{\theta} + u_{\alpha/2} \cdot \sigma\{\hat{\theta}\},$$

де $u_{\alpha/2}$ – квантиль стандартного нормального розподілу (табл. Б.1).

Якщо значення θ_0 розташоване в межах довірчого інтервалу, то приймають рішення про те, що гіпотеза H_0 є правильна.

Приклад 2.1. Нехай на гірничо збагачувальному підприємстві упродовж тривалого періоду середній вміст заліза в руді забезпечувався на рівні 68%. Припустимо, що запропоновано нову технологію збагачення руди, проведено 25 експериментів з відбором проб. Отримано, що за новим підходом середній вміст заліза складає 68,9% і середньоквадратичне відхилення по замірам 3,1%. Необхідно визначити, чи значущим є середній приріст вмісту заліза в руді за но-

вою технологією.

Позначимо середній вміст заліза в руді упродовж тривалого періоду $\bar{\theta}$, середній приріст вмісту заліза в руді за новою технологією $\hat{\theta}$ середньоквадратичне відхилення по замірам S . Головна гіпотеза H_0 така:

$$H_0: \bar{\theta} = \hat{\theta},$$

для перевірки використаємо статистику

$$t = \frac{\bar{\theta} - \hat{\theta}}{\sigma\{\hat{\theta}\}} = \frac{\bar{\theta} - \hat{\theta}}{S} \sqrt{N},$$

де N – кількість експериментів.

Визначивши статистичну характеристику t

$$t = \frac{68 - 68,9}{3,1} \sqrt{25} = \frac{4,5}{3,1} = 1,4516$$

та задаючи $\alpha = 0,1$ (зважаючи на невелике значення N), неважко переконатись, що

$$|t| \leq t_{\alpha/2, v},$$

де $t_{\alpha/2, v} = 1,71$.

Отже головну гіпотезу не спростовано, а отже слід вважати, що нова технологія не забезпечує статистично значущого приросту вмісту заліза у руді після збагачення

У разі потреби провести оцінку справжнього значення параметра θ за вибіркою $\Omega_{1,N}$ має місце помилка другого роду. Як уже зазначалося, оцінка справжнього значення параметра θ на основі вибірки може бути здійснена через довірчий інтервал, тобто шляхом доведення того, що з певною надійністю правильне значення знаходиться в інтервалі $[\theta_0 - \Delta; \theta_0 + \Delta]$. Оскільки $\hat{\theta}$ – випадкова величина відносно величин θ_0 , $\theta_0 - \Delta$, $\theta_0 + \Delta$, то мають місце близькі до нормальних розподіли

$$N(\hat{\theta}; E\{\hat{\theta}\}, \sigma\{\hat{\theta}\}),$$

$$N(\hat{\theta} - \Delta; E\{\hat{\theta} - \Delta\}, \sigma\{\hat{\theta}\}),$$

$$N(\hat{\theta} + \Delta; E\{\hat{\theta} + \Delta\}, \sigma\{\hat{\theta}\})$$

оцінки параметра $\hat{\theta}$ (рис. 2.2). З огляду на це помилка другого роду формально визначається за виразом

$$\beta = P\{\Omega_N \in \Omega_0 / \theta_1\} = P\{\varpi_0 | H_1\}.$$

Імовірність правильного рішення, яке полягає у відхиленні неістинної гі-

потези, визначається так:

$$1 - \beta = P\{\Omega_N \in \Omega_1 / \theta_1\} = P\{\varpi_1 | H_1\}.$$

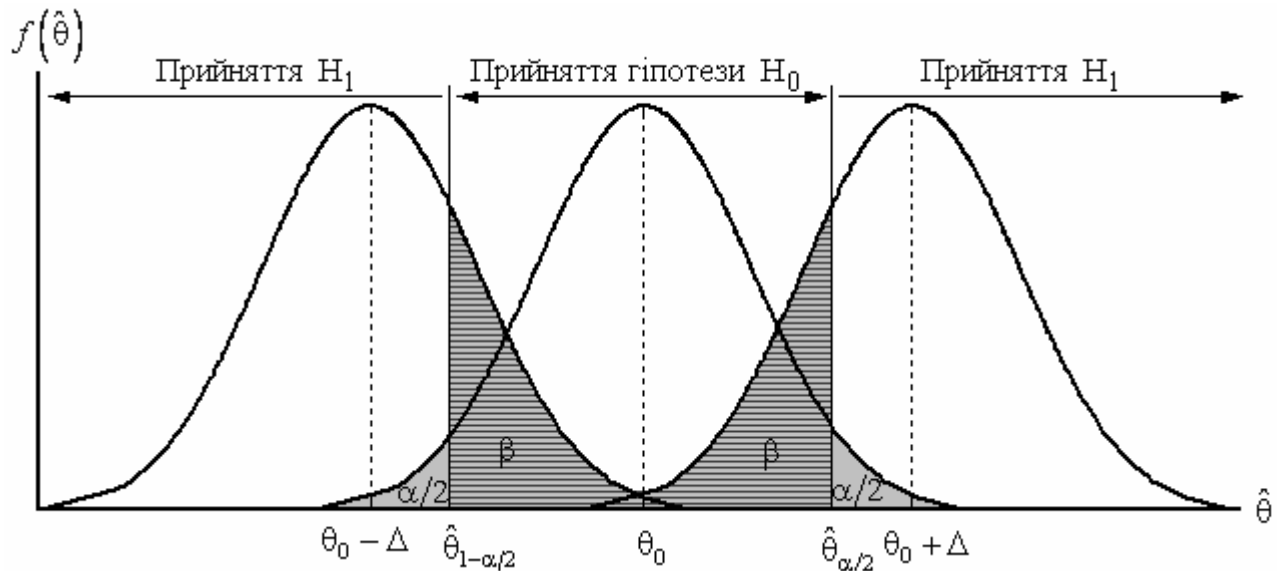


Рис. 2.2. Визначення помилок під час перевірки гіпотез

Якщо α , β – ймовірності помилок першого та другого роду, то $1 - \alpha$, $1 - \beta$ – відповідно потужності статистичного критерію відносно гіпотез H_1 , H_0 .

З аналізу графіка (рис. 2.2) й залежності ймовірностей випливає, що не можна однозначно стверджувати довільність помилки другого роду, якщо задана помилка першого роду. Проте можна навести нескінченну кількість критичних областей Ω_1 із заданою помилкою першого роду α та вибрати з них таку Ω_1^0 , яка дає

$$\max_{\theta} P\{\Omega_N \in \Omega_1 / \theta_1\} = \max_{\theta} P\{\varpi_1 | H_1\} = 1 - \beta_0,$$

у результаті чого буде одержаний критерій, який при заданій потужності $1 - \alpha$ відносно H_0 відзначатиметься найбільшою потужністю $1 - \beta$ відносно H_1 . Такі критерії мають назву **найбільш потужних**.

Якщо

$$\lim_{N \rightarrow \infty} P\{\varpi_1 | H_1\} = \lim_{N \rightarrow \infty} \max_{\theta} P\{\Omega_N \in \Omega_1 / \theta_1\} = 1,$$

то такий критерій називають обґрунтованим.

Вищенаведені вирази стосуються двобічного критерію перевірки головної гіпотези. Відповідним чином розв'язується задача побудови однобічних критеріїв для перевірки гіпотези

$$H_0 : \theta < \hat{\theta}$$

або

$$H_0 : \theta > \hat{\theta}.$$

Уведені поняття дозволяють запропонувати такий алгоритм побудови статистичного критерію перевірки гіпотези:

- 1) визначення статистичної характеристики гіпотези;
- 2) визначення або задання помилки першого роду α (критичний рівень);
- 3) формулювання альтернативної гіпотези;
- 4) визначення критичної області для статистичної характеристики з огляду на необхідність мінімізації помилки другого роду;
- 5) порівняння статистичної характеристики з критичним значенням і прийняття рішення про правильність головної чи альтернативної гіпотези.

2.2. Оцінка згоди відтворення розподілів

Головна процедура під час з'ясування вірогідності одновимірного статистичного розподілу – реалізація **критеріїв згоди**. Критерії згоди дозволяють для вибірки $\Omega_{1,N} = \{x_l; l = \overline{1, N}\}$ розв'язувати задачу порівняння емпіричної функції $F_{1,N}(x_l)$, $l = \overline{1, N}$ і табульованих значень відтвореної функціональної залежності $F(x_l; \hat{\Theta})$, $l = \overline{1, N}$ ($\hat{\Theta}$ – вектор оцінок параметрів теоретичної функції розподілу).

У цьому разі головна гіпотеза формулюється у вигляді

$$H_0 : F(x) = F_{1,N}(x).$$

Умовно критерії згоди можна поділити на дві групи. Для перевірки перших використовуються статистики, що є функціоналами від різниці функцій емпіричного та відтвореного розподілів. Це критерії: уточнений Колмогорова, Реньє, Андерсена–Дарлінга, ω^2 Мізеса. У процесі реалізації другої групи критеріїв враховується різниця між емпіричними та відтвореними (теоретичними) частотами. До них окрім критерію χ^2 належать його різні модифікації: критерії Берштейна, Романовського, Ястремського. Охарактеризуємо два найпоширеніші з названих критеріїв.

Один із найефективніших є **уточнений критерій згоди Колмогорова**, який потребує обчислення уточненої функції розподілу Колмогорова

$$K(z) = 1 + 2 \sum_{k=1}^{\infty} (-1)^k \exp(-2k^2 z^2) \left(1 - \frac{2k^2 z}{3\sqrt{N}} - \frac{1}{18N} ((f_1 - 4(f_1 + 3))k^2 z^2 + 8k^4 z^4) \right) +$$

$$+\frac{k^2 z}{27\sqrt{N^3}}\left(\frac{f_2^2}{5}-\frac{4(f_2+45)k^2 z^2}{15}+8k^4 z^4\right)+O\left(\frac{z^{13}}{N^2}\right),$$

де

$$f_1 = k^2 - 0,5\left(1 - (-1)^k\right);$$

$$f_2 = 5k^2 + 22 - 7,5\left(1 - (-1)^k\right);$$

$$z = \sqrt{N} \max\{D_N^-, D_N^+\};$$

$$D_N^+ = \max_l \left| F_{1,N}(x_l) - F(x_l; \hat{\Theta}) \right|;$$

$$D_N^- = \max_l \left| F_{1,N}(x_l) - F(x_{l-1}; \hat{\Theta}) \right|.$$

На основі статистичної характеристики z та її функції розподілу $K(z)$ складається процедура реалізації критерію згоди Колмогорова для перевірки вірогідності збігу емпіричного розподілу з теоретичним, яка потребує:

- 1) обчислення функцій розподілу $F_{1,N}(x_l)$ та $F(x_l; \hat{\Theta})$ і подальшого знаходження на їх основі значення статистики z ;
- 2) обчислення значення функції $K(z)$ та значення ймовірності узгодження

$$P(z) = 1 - K(z);$$

- 3) перевірки умови $P(z) \geq \alpha$, тобто умови збігу емпіричної функції розподілу з теоретичною, де α – критичний рівень значущості (якщо $N > 100$, то беруть $\alpha = 0,05$, при $N < 30$ рекомендується $\alpha = 0,3$);

- 4) побудови для теоретичного розподілу $F(x_l; \bar{\Theta})$ довірчого інтервалу

$$F_{\text{н.в.}}(x_l; \bar{\Theta}) = F(x_l; \hat{\Theta}) \mp D_{N\alpha},$$

де

$$D_{N\alpha} = \frac{z_\alpha}{\sqrt{N}};$$

z_α – критичне значення статистики Колмогорова, що встановлюється за значенням α (якщо $\alpha = 0,05$, то $z_\alpha = 1,36$, при $\alpha = 0,3$ $z_\alpha = 0,97$).

Критерій згоди χ^2 (Пірсона) реалізується лише для варіаційного ряду, розбитого на класи, та базується на обчисленні статистики

$$\chi^2 = \sum_{i=1}^M \frac{(n_i - n_i^0)^2}{n_i^0},$$

де

n_i – значення частот i -го класу, знайдені під час гістограмної оцінки;

$n_i^0 = Np_i$ – значення теоретичних частот;

$$p_i = F(x_i; \hat{\Theta}) - F(x_{i-1}; \hat{\Theta});$$

x_i та x_{i-1} – відповідно права та ліва межі i -го класу;

M – кількість класів.

Функція розподілу статистики χ^2 має вигляд

$$P(\chi^2 < x) = \frac{1}{2^{N/2} \Gamma(N/2)} \int_0^x u^{N/2-1} \exp\left(-\frac{u}{2}\right) du.$$

Перевірка головної гіпотези H_0 на основі даного критерію згоди полягає в обчисленні статистики χ^2 та порівнянні її з критичним значенням $\chi_{\alpha, \nu}^2$ (табл. Б.3), де $\nu = M - 1$. Виконання нерівності $\chi^2 \leq \chi_{\alpha, \nu}^2$ вказує на збіг емпіричної функції розподілу з теоретичною. Значення $P(\chi^2 < x) = \gamma$ відповідає ймовірності узгодження.

2.3. Задача двох вибірок

Задачу однорідності й незалежності в більшості випадків можна звести до задачі **двох вибірок**.

Нехай маємо дві генеральні сукупності Ω_1, Ω_2 , із яких вибрані вибірки $\Omega_{1, N_1} = \{x_1, \dots, x_{N_1}\}$ та $\Omega_{2, N_2} = \{y_1, \dots, y_{N_2}\}$. Відносно Ω_1 і Ω_2 припускаються розподіли відповідно $F(x)$ і $G(y)$. Необхідно перевірити гіпотезу $H_0 : F(x) \equiv G(y)$ за альтернативи $H_1 : F(x) \neq G(y)$.

Таке подання задачі є загальне, і її розв'язок одержують за допомогою як параметричних, так і непараметричних критеріїв. Розглянемо розв'язання такої задачі за параметричним критерієм. Припустимо, що закони розподілів $F(x)$, $G(y)$ є нормальні, а їх функції щільності такі:

$$f(x) = \frac{1}{\sigma_1 \sqrt{2\pi}} \exp\left(-\frac{(x-m_1)^2}{2\sigma_1^2}\right),$$

$$g(y) = \frac{1}{\sigma_2 \sqrt{2\pi}} \exp\left(-\frac{(y-m_2)^2}{2\sigma_2^2}\right).$$

Для того щоб $F(x)$ і $G(y)$ були однаковими, необхідний збіг їх відповідних параметрів. У цьому випадку гіпотези H_0 , H_1 можемо переписати у вигляді

$$H_0 : m_1 = m_2, \sigma_1 = \sigma_2$$

за альтернативи

$$H_1 : m_1 \neq m_2, \sigma_1 \neq \sigma_2.$$

Для перевірки гіпотез H_0 , H_1 існують критерії, розглянуті нижче.

2.4. Перевірка збігу середніх

Перевірка збігу середніх двох вибірок здійснюється за t -тестом, проведення якого для аналізованих вибірок потребує певних операцій перетворення. Будемо розрізняти випадки залежних і незалежних вибірок Ω_{1,N_1} , Ω_{1,N_2} .

Випадок залежних вибірок. Такий варіант дозволяє оцінювати вибірки, що характеризують однакові фізичні процеси або явища, які вивчаються різними методами. Для цього вимірюють один і той же параметр за різними методами, одержуючи вибірки однакового обсягу відносно x_l та y_l , $l = \overline{1, N}$.

Обчисливши різницю $z_l = x_l - y_l$, одержують нову вибірку $\Omega_{1,N} = \{z_l; l = \overline{1, N}\}$, для якої визначають

$$\bar{z} = \frac{1}{N} \sum_{l=1}^N z_l,$$

$$S_z^2 = \frac{1}{N-1} \sum_{l=1}^N (z_l - \bar{z})^2.$$

Оскільки x_i та y_i – реалізації випадкових величин ξ та η , які мають нормальні розподіли з $N_1(x; m_1, \sigma_1)$, $N_2(y; m_2, \sigma_2)$, маємо, що z_l – реалізація випадкової величини ζ , для якої $E\{\zeta\} = E\{\xi\} - E\{\eta\}$. Тоді гіпотезу $H_0 : m_1 = m_2$ переписують у вигляді $H_0 : m_1 - m_2 = 0$ або $H_0 : E\{\zeta\} = 0$ і для її перевірки використовують таку статистичну характеристику:

$$t = \frac{\bar{z}\sqrt{N}}{S_z}.$$

Результат порівняння $|t| > t_{\alpha/2, v}$ ($v = N - 2$) свідчить про те, що значення статистичної характеристики потрапило до критичної області, отже, головну гіпотезу слід відхилити. Подальший висновок відносно того, яке із середніх більше, робиться за знаком \bar{z} .

Випадок незалежних вибірок. Даний варіант дозволяє оцінювати вибірки, які характеризують різні фізичні процеси або явища. У такому разі обсяги вибірок можуть відрізнятися. Можливі два випадки:

- 1) обсяг вибірок є представницький;
- 2) обсяг вибірок обмежений.

Нехай **вибірки є представницькі**. Враховуючи, що різниця

$$\bar{z} = \bar{x} - \bar{y}$$

розподілена нормально з дисперсією

$$S_z^2 = S_x^2 + S_y^2 = \frac{S_x^2}{N_1} + \frac{S_y^2}{N_2},$$

для перевірки головної гіпотези H_0 на основі t -тесту застосовують статистику

$$t = \frac{\bar{z}}{S_z} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{S_x^2}{N_1} + \frac{S_y^2}{N_2}}},$$

яка має t -розподіл Стюдента з кількістю степенів вільності $v = N_1 + N_2 - 2$.

За **обмеженого обсягу вибірок** ($N_1 + N_2 \leq 25$) оцінюють зважене середнє S^2 оцінок S_x^2 , S_y^2 :

$$S^2 = \frac{(N_1 - 1)S_x^2 + (N_2 - 1)S_y^2}{N_1 + N_2 - 2},$$

де

$$S_x^2 = \frac{S_x^2}{N_1},$$

$$S_y^2 = \frac{S_y^2}{N_2}.$$

Як статистичну характеристику використовують величину

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{(N_1 - 1)S_x^2 + (N_2 - 1)S_y^2}{N_1 + N_2 - 2}}} \sqrt{\frac{N_1 N_2}{N_1 + N_2}},$$

що має t -розподіл Стюдента з кількістю степенів вільності $\nu = N_1 + N_2 - 2$. Подальша процедура перевірки не становить труднощів.

Приклад 2.2. Нехай студентів університету зважили на вагах A , а потім – на вагах B . Тим самим одержали дві залежні вибірки. У випадку, коли на вагах A зважили спочатку хлопців, а потім дівчат, мають місце дві незалежні вибірки.

2.5. Перевірка збігу дисперсій

Поряд з t -тестом у статистичній теорії перевірки гіпотез особливе місце займають параметричні критерії, що базуються на F -статистиках, розподілених за законом розподілу Фішера, – так звані **F -тести**. За наявності S_1^2 , S_2^2 – незалежних оцінок для дисперсій σ_1^2 , σ_2^2 – F -тест дозволяє перевіряти гіпотезу про їх збіг

$$H_0 : \sigma_1^2 = \sigma_2^2.$$

Для перевірки головної гіпотези вводять статистичну характеристику, що являє собою відношення оцінок двох дисперсій. Якщо таке відношення більше табульованого значення реалізацій випадкової величини, розподіленої за законом розподілу Фішера, то головна гіпотеза має бути відкинута.

Під час розв'язання задачі перевірки збігу дисперсій двох вибірок за статистичну характеристику беруть значення

$$f = \begin{cases} \frac{S_x^2}{S_y^2}, & \text{якщо } S_x^2 \geq S_y^2, \\ \frac{S_y^2}{S_x^2}, & \text{якщо } S_x^2 < S_y^2. \end{cases}$$

Статистика f має F -розподіл Фішера з кількістю степенів вільності $\nu_1 = N_1 - 1$ та $\nu_2 = N_2 - 1$. Враховуючи, що $f > 0$, за відомого α обчислюють критичне значення f_{α, ν_1, ν_2} (табл. Б.4) і, якщо

$$f \leq f_{\alpha, \nu_1, \nu_2},$$

приймають головну гіпотезу.

Зауваження 2.1. В процесі побудови обчислювальної процедури для перевірки гіпотези про однорідність двох вибірок потрібно **спочатку реалізувати**

перевірку збігу дисперсій. Якщо головна гіпотеза підтверджується, то проводять перевірку збігу середніх.

За необхідності перевірити гіпотезу **збіг дисперсій k вибірок**

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 = \sigma^2$$

за альтернативи

$$H_1 : \sigma_1^2 \neq \sigma_2^2 \neq \dots \neq \sigma_k^2 \neq \sigma^2$$

використовують **критерій Бартлетта**. Нехай заданий багатовимірний набір даних $\{x_{i,j}; i = \overline{1, k}, j = \overline{1, N_i}\}$, що являє собою k вибірок (можливо, різного обсягу).

Для перевірки головної гіпотези спочатку обчислюють значення

$$\bar{x}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} x_{i,j}, \quad i = \overline{1, k},$$

$$S_i^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (x_{i,j} - \bar{x}_i)^2, \quad i = \overline{1, k},$$

$$S^2 = \frac{\sum_{i=1}^k (N_i - 1) S_i^2}{\sum_{i=1}^k (N_i - 1)}.$$

За статистичну характеристику беруть величину

$$\chi^2 = \frac{B}{C},$$

яка має розподіл χ^2 . Значення B і C одержують за такими формулами:

$$B = - \sum_{i=1}^k (N_i - 1) \ln \frac{S_i^2}{S^2},$$

$$C = 1 + \frac{1}{3(k-1)} \left(\sum_{i=1}^k \frac{1}{N_i - 1} - \frac{1}{\sum_{i=1}^k (N_i - 1)} \right).$$

Для заданого рівня значущості α і кількості степенів вільності $v = k - 1$ знаходять критичне $\chi_{\alpha, v}^2$ (табл. Б.3) і приймають головну гіпотезу, якщо

$$\chi^2 \leq \chi_{\alpha, v}^2.$$

2.6. Однофакторний дисперсійний аналіз

Однофакторний дисперсійний аналіз застосовують для перевірки того, чи різняться поміж себе значення середніх множини k незалежних вибірок, що є реалізаціями відповідних нормально розподілених випадкових величин. Однофакторний дисперсійний аналіз порівнює два джерела варіації даних: міжгрупову варіацію (варіацію поміж вибірками) та варіацію всередині кожної вибірки.

Джерелом міжгрупової варіації є факт відмінності одне від одного генеральних сукупностей, з яких вилючено вибірки, що аналізуються. З іншого боку, чим більшою є варіація всередині кожної з вибірок, тим більш випадковими є дані, важче встановити, чи дійсно різняться генеральні сукупності.

Припускаючи, що дисперсії всіх k вибірок однакові:

$$\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 = \sigma^2,$$

висувають головну гіпотезу

$$H_0 : m_1 = m_2 = \dots = m_k$$

за альтернативи

$$H_1 : m_i \neq m_j, \forall i, j, i \neq j.$$

Міжгрупова варіація S_M^2 дає оцінку відмінностей середніх вибірок, що аналізуються:

$$S_M^2 = \frac{1}{k-1} \sum_{i=1}^k N_i (\bar{x}_i - \bar{x})^2,$$

де

N_i – обсяг i -ї вибірки;

\bar{x}_i – оцінка математичного сподівання i -ї вибірки;

\bar{x} – загальне середнє

$$\bar{x} = \frac{1}{N} \sum_{i=1}^k N_i \bar{x}_i,$$

де

$$N = \sum_{i=1}^k N_i.$$

Вочевидь, міжгрупова варіація S_M^2 дорівнює нулю, якщо середні рівні між собою і є тим більшою, чим сильніше середні різняться. Отже, величина S_M^2 представляє собою основу міри вираїції вибірових середніх.

Варіація всередині кожної вибірки S_B^2 визначається згідно з виразом

$$S_B^2 = \frac{1}{N-k} \sum_{i=1}^k (N_i - 1) S_i^2,$$

де

S_i^2 – оцінка дисперсії i -ї вибірки.

Перевірка головної гіпотези проводиться на основі статистичної характеристики

$$F = \frac{S_M^2}{S_B^2},$$

яка має розподіл Фішера з кількістю степенів вільності $v_1 = k - 1$, $v_2 = N - k$.

Головну гіпотезу H_0 приймають у разі виконання умови

$$F \leq f_{\alpha, v_1, v_2},$$

роблячи висновок, що середні вибірок невеликою мірою різняться поміж собою.

Якщо остання нерівність не виконується, роблять висновок про існування істотної різниці між вибірковими середніми, а отже, про неможливість пояснити розходження в їх значеннях лише випадковістю. Подальший аналіз може полягати у визначенні того, які саме вибірки попарно різняться між собою. Останнє з'ясовується на основі t -статистик, уведених для випадку незалежних вибірок, з урахуванням наявних обсягів аналізованих вибірок.

2.7. Критерії порядкових статистик

Наведені нижче критерії однорідності належать до так званих **рангових**. Вони ґрунтуються на вивченні послідовності реалізацій випадкової величини та можуть застосовуватися навіть у тих випадках, коли закони розподілу аналізованих вибірок відмінні від нормального.

З усього різноманіття процедур відібрані найбільш прості в реалізації, які дають змогу зробити надійні висновки про однорідність вибірок.

Задачу перевірки однорідності двох вибірок реалізують за одним або за всіма разом ранговими критеріями, при цьому головна гіпотеза формулюється так: дві вибірки $\Omega_{1, N_1} = \{x_i; i = \overline{1, N_1}\}$, $\Omega_{1, N_2} = \{y_j; j = \overline{1, N_2}\}$ вибрані з генеральних сукупностей з однаковим законом розподілу

$$H_0 : F(x) \equiv G(y).$$

Критерії Вілкоксона та U -критерій Манна–Уїтні є найчастіше використовувани. Їх реалізують під час перевірки гіпотез:

- 1) про наявність тренда в ряді спостережень;
- 2) однорідність вибірок.

Для перевірки головної гіпотези про значущість різниці двох незалежних вибірок з останніх формують загальний варіаційний ряд (обсягом $N = N_1 + N_2$), приписуючи кожному значенню варіанти ранг $r(x_i)$ або $r(y_j)$, тобто порядковий номер.

Приклад 2.3. Нехай задані дві вибірки $\Omega_{1,5} = \{12, 3, 18, -1, 20\}$, $\Omega_{1,6} = \{15, 7, 0, 10, 25, 9\}$. Сформуємо загальний варіаційний ряд і визначимо ранги:

Загальний варіаційний ряд:	x_1	y_1	x_2	y_2	y_3	y_4	x_3	y_5	x_4	x_5	y_6
	-1	0	3	7	9	10	12	15	18	20	25
Ранги:	1	2	3	4	5	6	7	8	9	10	11

Зауваження 2.2. Якщо в загальному варіаційному ряді виявляється декілька варіант, які збігаються, то кожній присвоюють ранг, що дорівнює середньому арифметичному їх порядкових номерів у сумісній послідовності.

Приклад 2.4. Нехай задані дві вибірки $\Omega_{1,7} = \{10, 3, 18, -1, 20, 10, 3\}$, $\Omega_{1,6} = \{15, 7, 0, 10, 25, 9\}$. Відповідно загальний варіаційний ряд і ранги такі:

Загальний варіаційний ряд:	x_1	y_1	x_2	x_3	y_2	y_3	x_4	x_5	y_4	y_5	x_6	x_7	y_6
	-1	0	3	3	7	9	10	10	10	15	18	20	25
Ранги:	1	2	3,5	3,5	5	6	8	8	8	10	11	12	13

Тоді, порівнюючи ранги вибірки Ω_{1,N_1} з рангами вибірки Ω_{1,N_2} , можна з'ясувати, різняться вибірки систематично чи випадково.

Критерій суми рангів Вілкоксона базується на обчисленні статистичної характеристики W , що визначається як сума рангів вибірки Ω_{1,N_1} (або Ω_{1,N_2}) у загальному варіаційному ряді:

$$W = \sum_{i=1}^{N_1} r(x_i).$$

Для головної гіпотези H_0 статистична характеристика W має симетричний відносно $E\{W\}$ закон розподілу, причому при $N > 25$ закон розподілу W прямує до нормального з параметрами

$$E\{W\} = \frac{N_1(N+1)}{2},$$

$$D\{W\} = \frac{N_1 N_2 (N+1)}{12}.$$

Порівнюючи значення

$$w = \frac{W - E\{W\}}{\sqrt{D\{W\}}}$$

з критичним значенням u_α нормального закону розподілу, головну гіпотезу приймають або відхиляють.

В основі ***U*-критерію Манна–Уїтні** лежить дослідження кількості способів, за допомогою яких в одній вибірці можна знайти значення, що перевищує значення в іншій вибірці. Аналізуючи загальний ряд даних, встановлюють, що має місце перерозподіл значень випадкових величин. Ступінь перерозподілу x та y визначають через інверсію. Якщо у варіаційному ряді деякому x передуює y , то таке явище називають однією інверсією, якщо ж певному x передуює k значень y , говорять, що значення x має k інверсій. Під час реалізації *U*-критерію Манна–Уїтні розраховують статистичну характеристику U , яка визначає кількість інверсій відносно x (або y) у загальному ряду:

$$U = \sum_{j=1}^{N_2} \sum_{i=1}^{N_1} z_{i,j},$$

$$z_{i,j} = \begin{cases} 1, & \text{якщо } x_i > y_j, \\ 0, & \text{якщо } x_i \leq y_j. \end{cases}$$

Слід відзначити, що поміж статистиками U та W існує така залежність:

$$U = N_1 N_2 + \frac{N_1(N_1-1)}{2} - W.$$

Якщо головна гіпотеза є правильна, то при $N > 25$ закон розподілу характеристики U прямує до нормального з параметрами

$$E\{U\} = \frac{N_1 N_2}{2},$$

$$D\{U\} = \frac{N_1 N_2 (N+1)}{12}.$$

Для перевірки гіпотези H_0 обчислюють статистичну характеристику

$$u = \frac{U - E\{U\}}{\sqrt{D\{U\}}},$$

значення якої порівнюють із критичним u_α нормального закону.

Зауваження 2.3. Якщо обсяг загального варіаційного ряду $N < 25$, слід застосовувати точні апроксимації законів розподілу статистик W та U або звертатися до їх табульованих значень.

Поряд із критеріями Вілкоксона та Манна–Уїтні існує й може бути застосований **критерій різниці середніх рангів вибірок** Ω_{1,N_1} та Ω_{1,N_2} . Для перевірки головної гіпотези вводять статистичну характеристику v , яка при $N > 20$ має нормальний закон розподілу. Для значення

$$v = \frac{\bar{r}_x - \bar{r}_y}{N \sqrt{\frac{N+1}{12N_1N_2}}},$$

де

$$\bar{r}_x = \frac{1}{N_1} \sum_{i=1}^{N_1} r(x_i),$$

$$\bar{r}_y = \frac{1}{N_2} \sum_{j=1}^{N_2} r(y_j),$$

перевіряють виконання умови

$$|v| \leq u_\alpha$$

і приймають головну гіпотезу в разі слушності наведеної нерівності.

Контрольні запитання та завдання

1. Дати визначення статистичної гіпотези. Відносно чого – генеральної сукупності чи вибірки – висувається статистична гіпотеза?
2. У чому полягає відмінність нульової гіпотези від альтернативної? Яка з них підлягає доведенню?
3. Дати визначення помилки першого роду. Чи можна нею керувати?
4. Що називають областями допустимих та критичних значень?
5. Що таке функція потужності статистичного критерію?

6. Яким чином обчислюється t -статистика для проведення t -тесту?
7. Для експоненціально розподілених даних обсягу $N = 50$ перевірити гіпотезу $H_0 : \lambda = \hat{\lambda}$, якщо $\lambda = 0,65$, $\hat{\lambda} = 0,9$, де λ – параметр моделі розподілу.
8. Навести статистику та обчислювальну схему реалізації уточненого критерію згоди Колмогорова.
9. Призначити 95%-й довірчий інтервал для одновимірної функції розподілу на основі реалізації критерію згоди Колмогорова.
10. Подати статистику та обчислювальну схему реалізації критерію згоди Пірсона. До яких варіаційних рядів застосовують цей критерій?
11. У чому полягає відмінність між залежними та незалежними вибірками в задачі перевірки однорідності двох вибірок?
12. На основі якої статистики перевіряють збіг двох дисперсій?
13. Яка гіпотеза перевіряється в однофакторному дисперсійному аналізі? Що таке міжгрупова варіація?
14. Навести процедуру реалізації критерію Бартлетта.
15. Чим відрізняються непараметричні критерії від параметричних?
16. Визначити статистики Вілкоксона та Манна–Уїтні.

Розділ 3. ОБРОБКА Й АНАЛІЗ ДВОВИМІРНИХ ДАНИХ

Розглянемо питання обробки та аналізу двовимірних масивів спостережень. Під час опрацювання таких масивів звичайно виникає три типи задач:

1) первинний аналіз, що включає побудову варіаційного ряду, перетворення даних, вилучення аномальних результатів спостережень, гістограмну оцінку та перевірку нормальності розподілу двовимірної випадкової величини;

2) встановлення наявності стохастичного зв'язку між складовими двовимірного випадкового вектора;

3) за наявності стохастичного зв'язку між складовими випадкового вектора – задачі ідентифікації та відтворення регресії.

3.1. Первинний аналіз

Беручи за основу реалізацію ймовірнісної оцінки одновимірної випадкової величини, можна узагальнити подібну оцінку для випадку обробки масивів реалізацій двовимірних випадкових величин. Так, для реалізації

$$\Omega_{2,N} = \{(x_l, y_l); l = \overline{1, N}\}$$

двовимірного випадкового вектора

$$\vec{\zeta} = (\xi(\omega), \eta(\omega))$$

з функцією розподілу

$$F(x, y) = P\{\omega: -\infty < \xi(\omega) < x, -\infty < \eta(\omega) < y\}$$

у припущенні незалежності складових $\xi(\omega)$ та $\eta(\omega)$

$$F(x, y) = P\{\omega: -\infty < \xi(\omega) < x\} P\{\omega: -\infty < \eta(\omega) < y\}$$

можна розглядати одновимірні масиви

$$\xi(\omega): \{x_l; l = \overline{1, N}\}$$

та

$$\eta(\omega) : \{y_l; l = \overline{1, N}\},$$

за кожним із яких можна провести побудову варіаційних рядів, розбитих на класи. Отже, визначаючи рівномірні розбиття Δ_{h_x} , Δ_{h_y} з кроками h_x , h_y відповідно за осями реалізацій величин $\xi(\omega)$ та $\eta(\omega)$, автоматично задаємо рівномірне розбиття Δ_{h_x, h_y} площини реалізацій двовимірної випадкової величини $\bar{\zeta}$.

Двовимірний варіаційний ряд

	x_1	...	x_i	...	x_{m_x}
y_1	$n_{1,1}, p_{1,1}$...	$n_{i,1}, p_{i,1}$...	$n_{m_x,1}, p_{m_x,1}$
...
y_j	$n_{1,j}, p_{1,j}$...	$n_{i,j}, p_{i,j}$...	$n_{m_x,j}, p_{m_x,j}$
...
y_{m_y}	n_{1,m_y}, p_{1,m_y}	...	n_{i,m_y}, p_{i,m_y}	...	n_{m_x,m_y}, p_{m_x,m_y}

визначений за розбиттям Δ_{h_x, h_y} , має такий алгоритм побудови.

1. За **варіанту** ряду $\{(x_i, y_j); i = \overline{1, M_x}, j = \overline{1, M_y}\}$, де M_x , M_y – кількість елементів розбиття (класів) за відповідними осями, беруть центральну точку (i, j) -го елемента розбиття Δ_{h_x, h_y} (рис. 3.1).

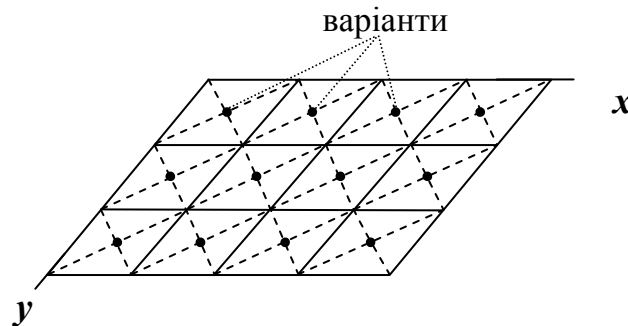


Рис. 3.1. Розбиття Δ_{h_x, h_y} площини реалізації $\bar{\zeta}$

2. Із нижченаведених співвідношень визначають **відносну частоту** $p_{i,j}$:

$$p_{i,j} = \frac{n_{i,j}}{N}, \quad \sum_{i=1}^{M_x} \sum_{j=1}^{M_y} p_{i,j} = 1,$$

де $n_{i,j}$ – кількість точок вихідного масиву спостережень $\Omega_{2,N}$, що потрапили в межі (i,j) -го елемента розбиття Δ_{h_x, h_y} .

Зауваження 3.1. Якщо (x_i, y_j) – центральна точка (i,j) -го елемента розбиття Δ_{h_x, h_y} , тоді

$$\bar{f}_{i,j}(x, y) = \frac{1}{h_x h_y} \int_{x_i - 0,5h_x}^{x_i + 0,5h_x} \int_{y_j - 0,5h_y}^{y_j + 0,5h_y} f(u, w) du dw$$

– усереднене значення функції щільності розподілу ймовірностей $\bar{\xi}$ у зазначеній області й має місце такий зв'язок із відносною частотою варіанти:

$$p_{i,j} \approx \bar{f}_{i,j}(x, y) h_x h_y = P\{\omega : x_i - 0,5h_x \leq \xi(\omega) < x_i + 0,5h_x, y_j - 0,5h_y \leq \eta(\omega) < y_j + 0,5h_y\}.$$

Отже, як і у випадку одновимірних даних, відносні частоти з точністю до константи $h_x \cdot h_y$ є **оцінкою усередненого значення функції щільності $f(x, y)$** для неперервної випадкової величини $\bar{\xi}$.

3. На основі відносних частот одержують **емпіричну оцінку $F_{2,N}(x, y)$ функції розподілу $\bar{\xi}$** :

$$F_{2,N}(x_i, y_j) = \sum_{a=1}^i \sum_{b=1}^j p_{a,b}, \quad i = \overline{1, M_x}, \quad j = \overline{1, M_y}.$$

Побудований таким чином варіаційний ряд можна зобразити у вигляді **двовимірної гістограми відносних частот** (рис. 3.2). У разі практичної реалізації достатньо подавати вид зверху на гістограму (рис. 3.3).

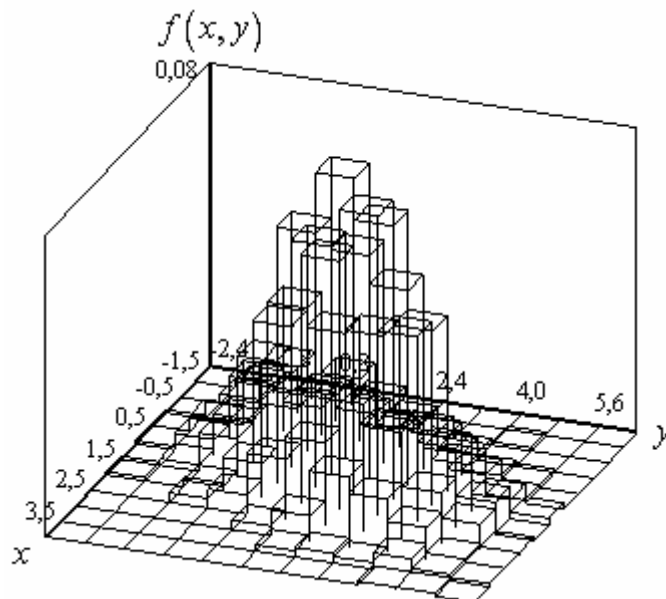


Рис. 3.2. Двовимірна гістограма відносних частот

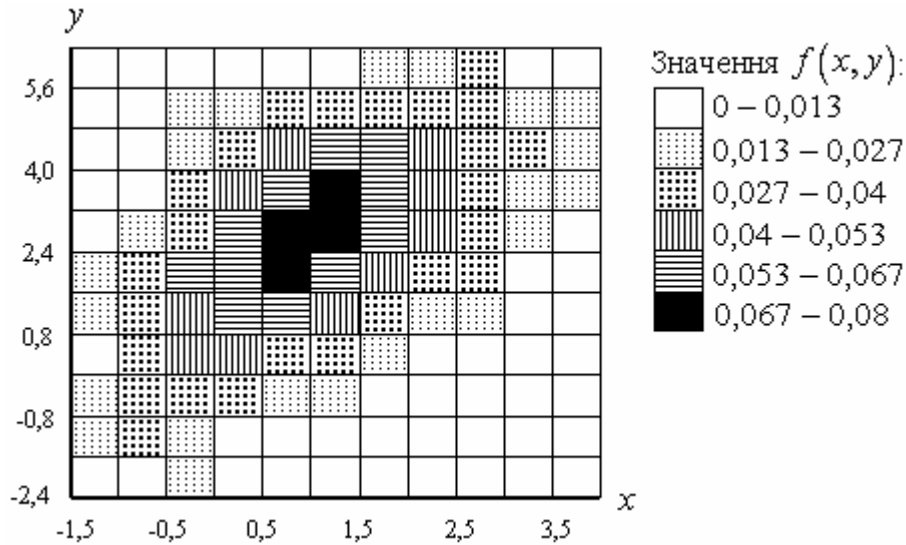


Рис. 3.3. Вид зверху на двовимірну гістограму відносних частот

Щодо кількості класів, то величини M_x , M_y визначаються за виразами, аналогічними одновимірному випадку. Перетворення даних у двовимірному випадку також зводиться до перетворень одновимірних складових вектора спостережень, а задача видалення аномальних значень розв'язується на основі гістограмної оцінки шляхом аналізу величин відносних частот та порівняння їх із заданою ймовірністю появи аномального результату спостереження в ряді. Якщо виконується нерівність

$$p_{i,j} \leq \alpha,$$

де α – ймовірність появи аномального значення, відповідні спостереження, що входять до (i, j) -го класу, видаляються з подальшого процесу обробки.

Найпростішими **точковими оцінками** за масивом $\Omega_{2,N} = \{(x_l, y_l); l = \overline{1, N}\}$ є оцінка вектора математичного сподівання $\hat{E}\{\vec{\xi}\} = (\bar{x}, \bar{y})$, де

$$\bar{x} = \frac{1}{N} \sum_{l=1}^N x_l,$$

$$\bar{y} = \frac{1}{N} \sum_{l=1}^N y_l,$$

який характеризує геометричний центр тяжіння однорідної сукупності спостережень, та оцінка дисперсійно-коваріаційної матриці

$$\hat{DC}\{\vec{\xi}\} = \begin{pmatrix} \hat{D}\{\xi\} & \text{cov}\{\xi, \eta\} \\ \text{cov}\{\xi, \eta\} & \hat{D}\{\eta\} \end{pmatrix} = \begin{pmatrix} \hat{\sigma}_x^2 & \hat{\sigma}_x \hat{\sigma}_y \hat{r}_{x,y} \\ \hat{\sigma}_x \hat{\sigma}_y \hat{r}_{x,y} & \hat{\sigma}_y^2 \end{pmatrix},$$

де

$\hat{\sigma}_x^2, \hat{\sigma}_y^2$ – незсунені оцінки дисперсій

$$\sigma_x^2 = \frac{1}{N-1} \sum_{l=1}^N (x_l - \bar{x})^2,$$

$$\sigma_y^2 = \frac{1}{N-1} \sum_{l=1}^N (y_l - \bar{y})^2;$$

$\hat{r}_{x,y}$ – оцінка парного коефіцієнта кореляції (розглядається далі).

У даному випадку оцінки $\hat{\sigma}_x^2, \hat{\sigma}_y^2$ характеризують розсіювання відповідних реалізацій відносно середнього (\bar{x}, \bar{y}) , а оцінка парного коефіцієнта кореляції $\hat{r}_{x,y}$ визначає міру лінійного зв'язку двох ознак.

Нарешті, якщо випадкова величина $\vec{\zeta} = (\xi(\omega), \eta(\omega))$ має **двовимірний нормальний розподіл**, то функція щільності, одержана за результатами обробки масиву $\Omega_{2,N}$, буде визначатися (рис. 3.4) в такий спосіб:

$$f(x, y) = \frac{1}{2\pi\hat{\sigma}_x\hat{\sigma}_y\sqrt{1-\hat{r}_{x,y}^2}} \exp\left(-\frac{1}{2(1-\hat{r}_{x,y}^2)}\left(\left(\frac{x-\bar{x}}{\hat{\sigma}_x}\right)^2 - 2\hat{r}_{x,y}\frac{x-\bar{x}}{\hat{\sigma}_x}\frac{y-\bar{y}}{\hat{\sigma}_y} + \left(\frac{y-\bar{y}}{\hat{\sigma}_y}\right)^2\right)\right).$$

Окремо звернемо увагу на те, що за незалежності випадкових величин $\xi(\omega)$ та $\eta(\omega)$ їх сумісна функція щільності така:

$$f(x, y) = \frac{1}{2\pi\hat{\sigma}_x\hat{\sigma}_y} \exp\left(-\frac{1}{2}\left(\left(\frac{x-\bar{x}}{\hat{\sigma}_x}\right)^2 + \left(\frac{y-\bar{y}}{\hat{\sigma}_y}\right)^2\right)\right).$$

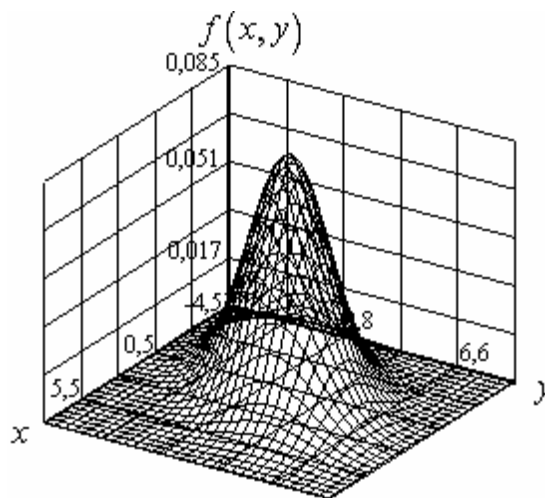


Рис. 3.4. Графік функції щільності двовимірного нормального розподілу

Для оцінки адекватності відтворення двовимірної функції нормального розподілу застосовується критерій χ^2 . Відповідні статистики мають вигляд:

1) за змінною y при фіксованій x :

$$\chi_i^2 = \sum_{j=1}^{M_y} \frac{(p_{i,j} - p_{i,j}^*)^2}{p_{i,j}^*}, \quad p_{ij}^* \neq 0;$$

2) за змінною x у разі фіксованої y :

$$\chi_j^2 = \sum_{i=1}^{M_x} \frac{(p_{i,j} - p_{i,j}^*)^2}{p_{i,j}^*}, \quad p_{ij}^* \neq 0;$$

3) одночасно за змінними x та y :

$$\chi^2 = \sum_{i=1}^{M_x} \sum_{j=1}^{M_y} \frac{(p_{i,j} - p_{i,j}^*)^2}{p_{i,j}^*}, \quad p_{ij}^* \neq 0,$$

де

$p_{i,j}$ – відносна частота варіаційного ряду, розбитого на класи;

$p_{i,j}^* = \bar{f}_{i,j}(x, y) \cdot h_x \cdot h_y$ – відтворена відносна частота;

$\bar{f}_{i,j}(x, y)$ – оцінка на основі відтворення нормального розподілу усередненого значення функції щільності.

Зауваження 3.2. Оцінка $\bar{f}_{i,j}(x, y)$ знаходиться як значення функції щільності нормального розподілу в центральній точці (i, j) -го класу.

Задаючи рівень помилки α та порівнюючи значення статистик із відповідним квантилем $\chi_{\alpha, \nu}^2$, $\nu = N - 2$ розподілу χ^2 , можна говорити про адекватність відтворення двовимірного нормального розподілу як за окремими розрізами, так і за всією областю визначення в цілому.

Масив $\Omega_{2,N} = \{(x_l, y_l); l = 1, N\}$ змодельованих нормально розподілених випадкових чисел із параметрами $m_1, m_2, \sigma_x, \sigma_y, r_{x,y}$ можна одержати за формулами

$$x = m_1 + \sigma_1 z_1,$$

$$y = m_2 + \sigma_2 \left(z_2 \sqrt{1 - r_{x,y}^2} + z_1 r_{x,y} \right),$$

де z_1, z_2 – реалізації одновимірних стандартизованих нормально розподілених випадкових величин ξ та η :

$$z_1 = \frac{\xi(\omega) - m_1}{\sigma_1},$$

$$z_2 = \frac{\frac{\eta(\omega) - m_2}{\sigma_2} - r_{x,y} \frac{\xi(\omega) - m_1}{\sigma_1}}{\sqrt{1 - r_{x,y}^2}} \xi(\omega).$$

3.2. Кореляційний аналіз

Головна задача кореляційного аналізу – оцінка стохастичних зв'язків між змінними за підсумками спостережень. Залежно від закону розподілу спостережуваних змінних вводяться різні типи коефіцієнтів кореляції, розглянуті нижче.

3.2.1. Парна кореляція

Найпростіший стохастичний зв'язок між двома випадковими величинами $\xi(\omega)$ та $\eta(\omega)$ є лінійний зв'язок, який визначається **коефіцієнтом кореляції**

$$r = \frac{E\{(\xi - E\{\xi\})(\eta - E\{\eta\})\}}{\sqrt{D\{\xi\}D\{\eta\}}} = \frac{\text{cov}\{\xi, \eta\}}{\sigma\{\xi\}\sigma\{\eta\}}.$$

Передумовою саме лінійного кореляційного аналізу є те, що випадкові величини $\xi(\omega)$ та $\eta(\omega)$ повинні бути нормально розподіленими.

Коефіцієнт кореляції (парної кореляції) має властивості:

- 1) $|r| \leq 1$;
- 2) $r = 0$, якщо $\xi(\omega)$ та $\eta(\omega)$ – незалежні випадкові величини;
- 3) $r = \pm 1$, якщо між $\xi(\omega)$ та $\eta(\omega)$ існує лінійний функціональний зв'язок,

у протилежному разі – випадковий лінійний регресійний

$$\eta = \alpha + \beta\xi + \varepsilon,$$

де ε – похибка.

Оцінка параметра r за масивом $\Omega_{2,N}$ здійснюється так:

$$\hat{r}_{x,y} = \frac{N}{N-1} \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\hat{\sigma}_x \hat{\sigma}_y},$$

де

$$\overline{xy} = \frac{1}{N} \sum_{l=1}^N x_l \cdot y_l.$$

Оцінка парного коефіцієнта кореляції має **геометричну інтерпретацію** як косинус кута φ_{xy} поміж векторами спостережень

$$\vec{X} = \{x_l; l = \overline{1, N}\}$$

та

$$\vec{Y} = \{y_l; l = \overline{1, N}\}.$$

І справді,

$$\cos \varphi_{x,y} = \frac{\vec{X} \cdot \vec{Y}}{|\vec{X}| \cdot |\vec{Y}|} = \frac{\sum_{l=1}^N x_l \cdot y_l}{\sqrt{\sum_{l=1}^N x_l^2} \cdot \sqrt{\sum_{l=1}^N y_l^2}},$$

тоді, якщо $\bar{x} = 0$ та $\bar{y} = 0$ при $N \rightarrow \infty$, вираз для оцінки $\hat{r}_{x,y}$ є еквівалентний наведеному для $\cos \varphi_{x,y}$.

Ідентифікація наявності зв'язку між $\xi(\omega)$ та $\eta(\omega)$ може бути здійснена візуально після побудови кореляційного поля, що являє собою графічне зображення масиву $\Omega_{2,N}$, коли за віссю абсцис відкладаються значення x_l , а за віссю ординат – відповідні значення y_l . Кореляційне поле у вигляді кола або овалу свідчить про те, що $\xi(\omega)$ та $\eta(\omega)$ нормально розподілені. Якщо поле вписується в коло (рис. 3.5, а), то можна вважати, що зв'язок між $\xi(\omega)$ та $\eta(\omega)$ відсутній, кут $\varphi_{x,y} = 90^\circ$. Поле у вигляді овалу дає можливість говорити про наявність лінійного зв'язку, а нахил овалу – про додатний (рис 3.5, б) чи від'ємний зв'язок (рис. 3.5, в). Поле складної конфігурації (рис 3.5, г, д) свідчить про нелінійний зв'язок між $\xi(\omega)$ та $\eta(\omega)$ і можливу потребу в перетворенні даних. Якщо в межах кола виділяється декілька сукупностей (рис. 3.5, е), це вказує на неоднорідність даних.

Статистичне значення $\hat{r}_{x,y}$ завжди є відмінне від нуля. Тому виникає задача **перевірки значущості коефіцієнта кореляції**, отже, висувається гіпотеза

$$H_0 : r = 0,$$

для перевірки якої реалізують t -тест на основі статистики

$$t = \frac{\hat{r}_{x,y} \sqrt{N-2}}{\sqrt{1-\hat{r}_{x,y}^2}}.$$

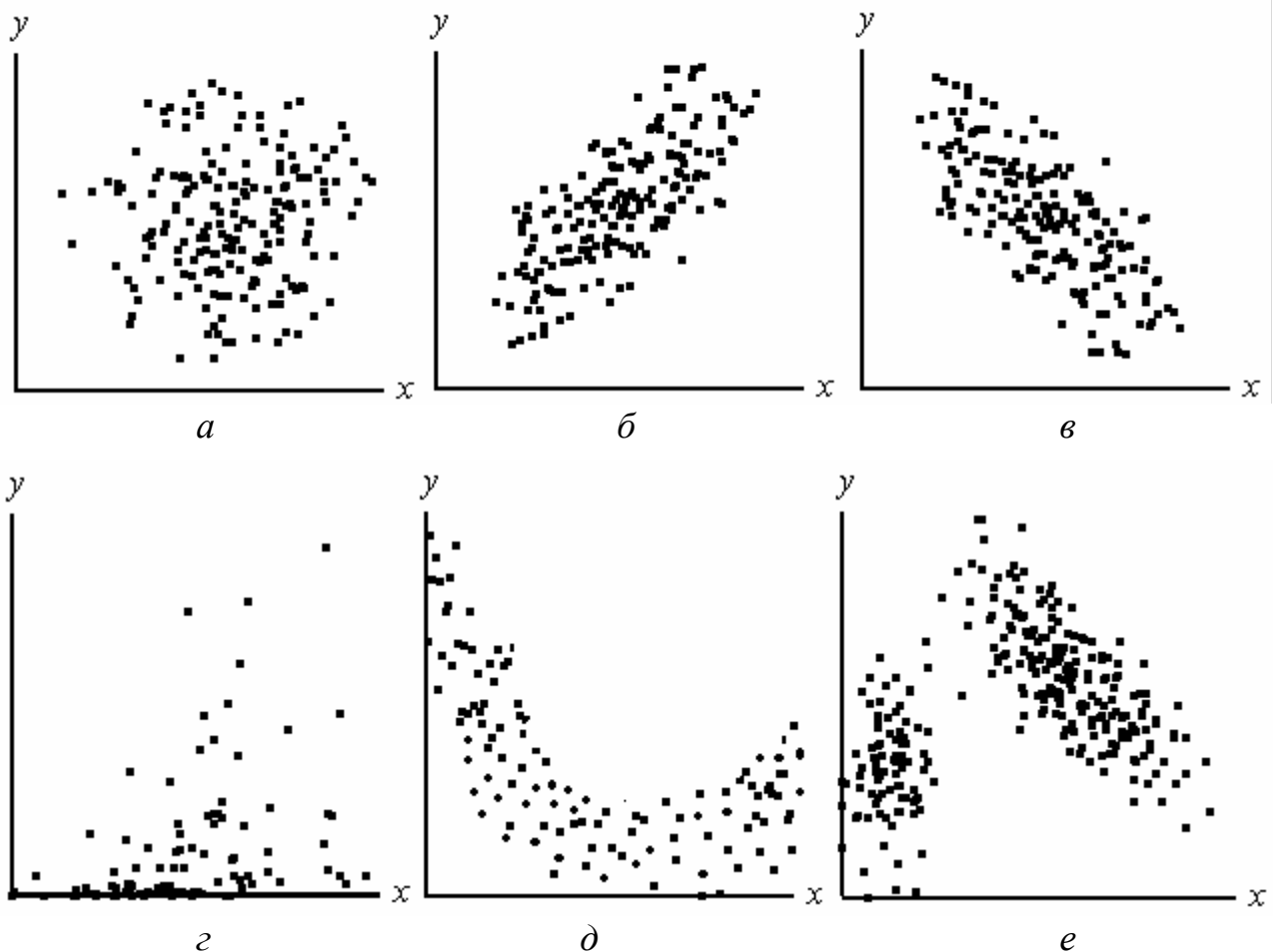


Рис. 3.5. Кореляційні поля: a – зв'язок відсутній; $б$ – додатний лінійний зв'язок; $в$ – від'ємний лінійний зв'язок; $г, д$ – нелінійний зв'язок; $е$ – випадок неоднорідних даних

Інтервальне оцінювання коефіцієнта кореляції здійснюється шляхом призначення довірчого інтервалу з межами

$$r_{н,в} = \hat{r}_{x,y} \pm \frac{\hat{r}_{x,y} (1 - \hat{r}_{x,y}^2)}{2N} \mp u_{\alpha/2} \frac{1 - \hat{r}_{x,y}^2}{\sqrt{N-1}}.$$

На практиці дані можуть формуватись у вигляді k масивів $\Omega_{2,N_j} = \{(x_l, y_l); l = \overline{1, N_j}\}$, $j = \overline{1, k}$, тоді виникає задача про формування єдиного масиву даних (за умови збігу відповідних середніх та середньоквадратичних масивів). Під час розв'язання такої задачі можливі випадок перевірки парами та загальний випадок, за яких на основі Ω_{2,N_j} обчислюють масив $\{\hat{r}_j, j = \overline{1, k}\}$.

Формування парами зумовлює перевірку статистичної гіпотези

$$H_0 : r_j = r_s, \quad j \neq s$$

з огляду на статистичну характеристику

$$u = \frac{z_j - z_s}{\sqrt{\frac{1}{N_j - 3} + \frac{1}{N_s - 3}}},$$

де

$$z_i = \frac{1}{2} \ln \frac{1 + \hat{r}_i}{1 - \hat{r}_i}, \quad i = j, s.$$

Величина u нормально розподілена, отже, для заданої помилки першого роду α перевіряють виконання умови

$$|u| \leq u_{\alpha/2}.$$

Якщо нерівність виконується, приймають рішення, що коефіцієнти r_j, r_s статистично не різняться. У цьому випадку масиви початкових даних об'єднують в один, за яким переобчислюють коефіцієнт кореляції.

Для загального випадку здійснюється перевірка гіпотези

$$H_0 : r_1 = r_2 = \dots = r_k$$

на основі характеристики

$$\chi^2 = \sum_{i=1}^k (N_i - 3) z_i^2 - \frac{\left(\sum_{i=1}^k (N_i - 3) z_i \right)^2}{\sum_{i=1}^k (N_i - 3)},$$

яка має χ^2 -розподіл із кількістю степенів вільності $\nu = k - 1$. Якщо має місце $\chi^2 \leq \chi_{\alpha, \nu}$, то головна гіпотеза є правильна і необхідне формування єдиного масиву, за яким обчислюють \hat{r} із подальшою статистичною оцінкою.

3.2.2. Кореляційне відношення

Якщо залежність поміж випадковими величинами η, ξ не лінійна, то для оцінки такого зв'язку на основі масиву $\{(x_i, y_{i,j}); i = \overline{1, k}, j = \overline{1, m_i}\}$ обчислюють **коефіцієнт кореляційного відношення** ρ :

$$\hat{\rho}_{\eta/\xi}^2 = \frac{\sum_{i=1}^k m_i (\bar{y}_i - \bar{y})^2}{\sum_{i=1}^k \sum_{j=1}^{m_i} (y_{i,j} - \bar{y})^2} = \frac{S_{\bar{y}(x)}^2}{S_y^2},$$

де $S_{\bar{y}(x)}^2$ – оцінка міжгрупової дисперсії.

Кореляційне відношення має такі властивості:

- 1) $0 \leq \rho \leq 1$;
- 2) $\rho_{\eta/\xi} \geq |r_{x,y}|$, $\rho_{\xi/\eta} \geq |r_{x,y}|$;
- 3) якщо кореляційний зв'язок відсутній, то $\rho_{\eta/\xi} = \rho_{\xi/\eta} = 0$;
- 4) якщо поміж η та ξ існує лінійний регресійний зв'язок, то $\rho_{\eta/\xi} = \rho_{\xi/\eta} = |r_{x,y}|$.

Статистичне значення $\hat{\rho}$ є випадкова величина і за досить великого $N = \sum_{i=1}^k m_i$ має нормальний розподіл із параметрами

$$E\{\hat{\rho}\} = \rho, \quad D\{\hat{\rho}\} = \frac{1-\rho^2}{N-2}.$$

Це дозволяє запропонувати статистичну характеристику

$$t = \frac{\hat{\rho}\sqrt{N-2}}{\sqrt{1-\hat{\rho}^2}}$$

для перевірки гіпотези

$$H_0 : \rho = 0.$$

Значення t має t -розподіл з $\nu = N - 2$ степенями вільності. Якщо

$$|t| \leq t_{\alpha/2, \nu},$$

то стверджують, що кореляційний зв'язок поміж η , ξ відсутній.

Правило перевірки наявності стохастичного зв'язку між двома змінними таке:

- 1) обчислюють значення $\hat{r}_{x,y}$ та оцінюють його значущість;
- 2) якщо $\hat{r}_{x,y}$ не є значуще, обчислюють $\hat{\rho}_{\eta/\xi} = \hat{\rho}$ та перевіряють його значущість;
- 3) у разі правильності гіпотези $H_0 : \rho = 0$ роблять висновок про відсутність стохастичного зв'язку поміж η , ξ .

Зауваження 3.3. Для одержання масиву $\{(x_i, y_{i,j}); i = \overline{1, k}, j = \overline{1, m_i}\}$ на основі $\Omega_{2,N} = \{(x_l, y_l); l = \overline{1, N}\}$ можна провести розбиття осі X з деяким кроком Δx . Тоді $x_i = x_{\min} + (i - 0,5)\Delta x$. Відповідні $y_{i,j}$ знайдемо з використанням варіант $\Omega_{2,N}$, для яких $x_l \in [x_i - 0,5\Delta x; x_i + 0,5\Delta x]$.

3.2.3. Парна рангова кореляція

Процедури рангової кореляції реалізуються в тому випадку, коли передумови лінійного кореляційного аналізу **не виконуються**. Так, якщо розподіли випадкових величин η та ξ відмінні від нормального, то обчислюють ранговий коефіцієнт Спірмена, поряд із яким реалізують коефіцієнт Кендалла. Попередньо початковий масив даних $\Omega_{2,N} = \{(x_l, y_l); l = \overline{1, N}\}$ переформовують у масив рангів

$$\{(r_{x,l}, r_{y,l}); l = \overline{1, N}\},$$

де $r_{x,l}$, $r_{y,l}$ – ранги, тобто порядкові номери варіант у варіаційних рядах за x та y .

При цьому кожному $r_{x,l}$ приписується номер $r_{y,l}$, що відповідає значенню y_l , або, навпаки, кожному $r_{y,l}$ приписується відповідний $r_{x,l}$.

На практиці можливий випадок збігу рангів. Такі ранги називаються зв'язаними, а їх група – зв'язкою. Для зв'язаних рангів здійснюють їх усереднення і кожному зв'язаному рангу приписують середнє значення.

Приклад 3.1. Нехай заданий масив $\Omega_{2,7} = \{(10,13), (7,5), (11,10), (3,5), (7,8), (12,15), (5,9)\}$. Підсумком ранжування змінної X будуть такі ранги:

Значення x_l : 3 5 7 7 10 11 12

Ранги r_x : 1 2 3,5 3,5 5 6 7

У результаті ранжування змінної Y одержуємо

Значення y_l : 5 5 8 9 10 13 15

Ранги r_y : 1,5 1,5 3 4 5 6 7

Після зіставлення рангів за змінною X остаточно маємо

r_x : 1 2 3,5 3,5 5 6 7

r_y : 1,5 4 1,5 3 6 5 7

Нижчеподана обчислювальна схема визначає ступінь стохастичного зв'язку поміж r_x , r_y через наведені коефіцієнти рангової кореляції.

1. Значення оцінки **рангового коефіцієнта кореляції Спірмена** $\hat{\tau}_c$ обчислюють за формулою

$$\hat{\tau}_c = 1 - \frac{6}{N(N^2 - 1)} \sum_{l=1}^N d_l^2,$$

де

$$d_l = r_{x,l} - r_{y,l}.$$

За наявності зв'язаних рангів оцінка $\hat{\tau}_c$ визначається таким чином:

$$\hat{\tau}_c = \frac{\frac{1}{6} N(N^2 - 1) - \sum_{l=1}^N (r_{x,l} - r_{y,l})^2 - A - B}{\sqrt{\left(\frac{1}{6} N(N^2 - 1) - 2A \right) \left(\frac{1}{6} N(N^2 - 1) - 2B \right)}},$$

де

$$A = \frac{1}{12} \sum_{j=1}^z (A_j^3 - A_j);$$

$$B = \frac{1}{12} \sum_{k=1}^p (B_k^3 - B_k);$$

z – кількість зв'язок поміж рангами r_x ; j – порядковий номер зв'язки;

A_j – кількість однакових значень x у зв'язці; так, якщо в другій зв'язці за r_x є три однакових x , то $A_2 = 3$; це саме стосується і p , k і B_k за y і r_y .

Коефіцієнт рангової кореляції Спірмена має такі властивості:

1) $-1 \leq \tau_c \leq 1$;

2) якщо $r_{x,l} = r_{y,l}$, $l = \overline{1, N}$, то $\tau_c = 1$, що означає повну узгодженість між X і Y ;

3) у разі $\tau_c = -1$ має місце протилежне впорядкування послідовностей рангів, тобто повна неузгодженість (від'ємна кореляція);

4) при відсутності кореляції $\tau_c = 0$.

Значущість $\hat{\tau}_c$ визначається на основі гіпотези

$$H_0 : \tau_c = 0 ,$$

для перевірки якої вводиться статистична характеристика

$$t = \frac{\hat{\tau}_c \sqrt{N-2}}{\sqrt{1-\hat{\tau}_c^2}} ,$$

яка має t -розподіл з кількістю степенів вільності $\nu = N - 2$.

2. Оцінка **рангового коефіцієнта Кендалла** $\hat{\tau}_k$ визначається за виразом

$$\hat{\tau}_k = \frac{2S}{N(N-1)} ,$$

де

S – алгебрична сума кількості найвищих рангів відносно кожного нижчого рангу;

$$S = \sum_{l=1}^{N-1} v_l = \sum_{l=1}^{N-1} \sum_{j=l+1}^N v_{l,j} ;$$

$$v_{l,j} = \begin{cases} 1 & , \quad r_{y,l} < r_{y,j} , \\ -1 & , \quad r_{y,l} > r_{y,j} . \end{cases}$$

Для встановлення значущості $\hat{\tau}_k$ перевіряють гіпотезу

$$H_0 : \tau_k = 0$$

із використанням статистичної характеристики

$$u = \frac{3\hat{\tau}_k}{\sqrt{2(2N+5)}} \sqrt{N(N-1)} ,$$

яка має стандартний нормальний розподіл $N(u; 0, 1)$.

Отже, якщо $|u| \leq u_{\alpha/2}$, то оцінка $\hat{\tau}_k$ не є значуща.

Коефіцієнт кореляції Кендалла має ті самі властивості, що й коефіцієнт Спірмена. Завжди для одних і тих же масивів $\hat{\tau}_c > \hat{\tau}_k$, а у випадку досить великого N

$$\hat{\tau}_c \approx \frac{3}{2} \hat{\tau}_k .$$

Приклад 3.2. Для наведеного вище прикладу 3.1 правильне таке:

$$z = 1, \quad A_1 = 2 ,$$

$$p = 1, \quad B_1 = 2 ,$$

значення рангового коефіцієнта Спірмена дорівнює

$$\hat{\tau}_c = 0,809.$$

У процесі оцінювання рангового коефіцієнта Кендалла має місце

$$v_1 = \sum_{j=2}^7 v_{1,j} = 5 - 0 = 5, \quad v_4 = \sum_{j=5}^7 v_{4,j} = 3 - 0 = 3,$$

$$v_2 = \sum_{j=3}^7 v_{2,j} = 3 - 2 = 1, \quad v_5 = \sum_{j=6}^7 v_{5,j} = 1 - 1 = 0,$$

$$v_3 = \sum_{j=4}^7 v_{3,j} = 4 - 0 = 4, \quad v_6 = \sum_{j=7}^7 v_{6,j} = 1 - 0 = 1,$$

$$S = 14,$$

значення коефіцієнта становить

$$\hat{\tau}_k = 0,667.$$

Наведені вирази не потребують лінійної кореляції поміж змінними. Обмежуючою вимогою є монотонність функції регресії. Слід відзначити, що процедури рангової кореляції є ефективні під час оцінки стохастичних зв'язків як для кількісних, так і для якісних ознак.

3.2.4. Коефіцієнти сполучень таблиць

Чимало прикладних задач з області економіки соціології, медицини, тощо, потребують обробки та аналізу категорійного типу. Наприклад, об'єктивним показником стану людини є температура тіла t . Визначальним в постановці діагнозу є перевищення показника температури значення деякого порогу (наприклад $t > 36,6^\circ C$). В даному прикладі ніщо не заважає звести масив спостережень реалізації неперервної випадкової величини (заміри температури) до бінарного масиву з елементами

$$s = \begin{cases} 0, & t \leq 36,6 \\ 1, & t > 36,6. \end{cases}$$

Опрацювання реалізацій неперервних випадкових величин, що можуть бути співставлені бінарним масивам або деяким шкалам можливо проводити за використанням так званих **таблиць сполучень**.

Коли двовимірною випадковою величиною подана масивом реалізації двох дихотомізованих змінних X , Y , які задано бінарним масивом $\{0;1 \times 0;1\}$ («false»,

«true»; «так», «ні»; «добре», «погано» та ін.), формується наступна таблиця сполучень 2×2 (табл.3.1):

Таблиця 3.1

Таблиця сполучень 2x2

$Y \setminus X$	0	1	
0	N_{00}	N_{01}	N_0
1	N_{10}	N_{11}	N_1
	M_0	M_1	N

де

$N_{00}, N_{01}, N_{10}, N_{11}$ – кількість відповідних двійок даних;

$$N_0 = N_{00} + N_{01}; \quad N_1 = N_{11} + N_{10}; \quad M_0 = N_{00} + N_{10};$$

$$M_1 = N_{01} + N_{11}; \quad N = N_0 + N_1 = M_0 + M_1.$$

Для оцінки стохастичного зв'язку дихотомізованих змінних X, Y до розгляду вводять різного типу коефіцієнти кореляції або сполучень (асоціації):

1. Індекс Фехнера найпростіший, який реалізується не тільки для бінарних даних, а й для кількісних $\Omega_{2,N} = \{(x_l, y_l); l = \overline{1, N}\}$.

Якщо має місце масив дійсних чисел, то обчислюють середнє (\bar{x}, \bar{y}) та визначають знаки двійок, $(x_l - \bar{x}, y_l - \bar{y}) : \{++; --; +-; -+\}$. Для оцінки зв'язку між складовими двовимірної випадкової величини вводиться такий індекс Фехнера

$$I = \frac{V - W}{V + W},$$

де

V, W – кількість збігу та незбігу знаків.

Індекс I має властивість $|I| \leq 1$. При $I > 0$ - додатна кореляція, при $I < 0$ - від'ємна, $I = 0$ - відсутність зв'язку, отже, X та Y - незалежні. Обчислення індексу рекомендовано для приблизної оцінки зв'язку.

2. Коефіцієнт сполучень Φ («Фі»), який запропоновано К.Пірсоном, обчислюється за формулою

$$\Phi = \frac{N_{00}N_{11} - N_{01}N_{10}}{\sqrt{N_0N_1M_0M_1}}.$$

Коефіцієнт Φ інтерпретується як коефіцієнт кореляції:

- завжди $|\Phi| \leq 1$;
- при $N_{00} = N_{11} = 0$, $\Phi = -1$;
- при $N_{01} = N_{10}$, $\Phi = 1$.

Для перевірки значущості коефіцієнта Φ ($H_0 : \Phi = 0$) обчислюють статистичну характеристику

$$\chi^2 = N\Phi^2,$$

яка має χ^2 -розподіл з $\nu = 1$ степенем вільності. Якщо $\chi^2 \geq \chi_{\alpha,1}^2$, то оцінка коефіцієнта Φ є значущою.

Слід зазначити, що реалізація статистичної характеристики χ ставить вимоги: частота ознаки має бути більшою, ніж 5, а обсяг вибірки – не менше, ніж 40. Якщо вимоги не виконуються, то для перевірки головної гіпотези реалізують таку статистику

$$\chi^2 = N \frac{(N_{00}N_{11} - N_{01}N_{10} - 0,5)^2}{N_0N_1M_0M_1}.$$

3. Коефіцієнти зв'язку Юла Q та Y , які мають властивості коефіцієнта Φ та обчислюються наступним чином:

$$Q = \frac{N_{00}N_{11} - N_{01}N_{10}}{N_{00}N_{11} + N_{01}N_{10}},$$

$$Y = \frac{\sqrt{N_{00}N_{11}} - \sqrt{N_{01}N_{10}}}{\sqrt{N_{00}N_{11}} + \sqrt{N_{01}N_{10}}}.$$

Існує зв'язок

$$Q = \frac{2Y}{(1+Y)^2}.$$

Значущість коефіцієнтів Q , Y (гіпотези $H_0 : Q = 0$, $H_0 : Y = 0$) перевіряється на підставі наступних нормально розподілених статистик:

$$u_Q = \frac{Q}{S_Q}, \quad u_Y = \frac{Y}{S_Y},$$

де

$$S_Q = \frac{1}{2}(1-Q^2) \sqrt{\frac{1}{N_{00}} + \frac{1}{N_{01}} + \frac{1}{N_{10}} + \frac{1}{N_{11}}};$$

$$S_Y = \frac{1}{4}(1 - Y^2) \sqrt{\frac{1}{N_{00}} + \frac{1}{N_{01}} + \frac{1}{N_{10}} + \frac{1}{N_{11}}}.$$

Головні гіпотези мають бути прийняті, якщо $|u_Q| \leq u_{\alpha/2}$, $|u_Y| \leq u_{\alpha/2}$.

На відміну від таблиць сполучень 2×2 , **таблиці перехресного табулювання** $n \times m$ описують експеримент, коли має місце класифікація за двома змінними (властивостями, факторами) X , Y , при цьому змінна X має $m \geq 2$ класів (рівнів), Y має $n \geq 2$ класів (рівнів). Рівні визначаються за шкалами найменувань або порядку.

Нехай кількості спостережень n_{ij} $i = \overline{1, n}$ $j = \overline{1, m}$, що потрапили до певних класів зведено до таблиці розмірності $n \times m$ (табл.3.2):

Таблиця 3.2

Таблиця сполучень $n \times m$

$Y \setminus X$	1	2	...	j	...	m	
1	n_{11}	n_{12}	...	n_{1j}	...	n_{1m}	n_1
2	n_{21}	n_{22}	...	n_{2j}	...	n_{2m}	n_2
...
i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{im}	n_i
...
n	n_{n1}	n_{n2}	...	n_{nj}	...	n_{nm}	n_n
	m_1	m_2	...	m_j	...	m_m	N

Стохастичний зв'язок поміж X , Y визначається мірою зв'язаності, яка базується на критеріях χ^2 Пірсона, Кендалла та ін.

На першому етапі за даними (табл.3.2) перевіряється гіпотеза про незалежність X , Y на підставі статистичної характеристики

$$\chi^2 = \sum_{i=1}^n \sum_{j=1}^m \frac{(n_{ij} - N_{ij})^2}{N_{ij}},$$

де

$$N_{ij} = \frac{n_i \cdot m_j}{N}.$$

Гіпотеза $H_0: \chi^2 = 0$ є вірною, коли для статистичної характеристики χ^2 ,

яка має розподіл Пірсона з $v = (n-1)(m-1)$ степенями вільності, виконується нерівність $\chi^2 \leq \chi_{\alpha, v}^2$, у супротивному випадку приймається рішення про наявність зв'язку поміж ознаками X , Y .

Якщо має місце зв'язок поміж X , Y , то обчислюють коефіцієнти сполучення. Найпростішими є коефіцієнти Пірсона та Кендала.

1. Коефіцієнт сполучень Пірсона обчислюється на підставі значення наведеної статистики χ^2 :

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}}$$

та який має такі властивості:

- завжди $0 \leq C \leq 1$;
- якщо $C = 0$, то зв'язок поміж ознаками X , Y відсутній;
- якщо $C = 1$ – має місце суттєва зв'язаність.

Значущість коефіцієнта C , а отже, перевірка гіпотези $H_0 : C = 0$ визначається на підставі статистичної характеристики χ^2 -Пірсона.

2. Міра зв'язку Кендалла обчислюється для випадку, коли $m = n$, за формулою

$$\tau_b = \frac{P - Q}{\sqrt{\left(\left(\frac{1}{2}n(n-1) - T_1\right)\left(\frac{1}{2}n(n-1) - T_2\right)\right)},$$

де

$$P = \sum_{i=1}^n \sum_{j=1}^m n_{ij} \left(\sum_{k=i+1}^n \sum_{l=j+1}^m n_{kl} \right);$$

$$Q = \sum_{i=1}^n \sum_{j=1}^m n_{ij} \left(\sum_{k=i+1}^n \sum_{l=1}^{j-1} n_{kl} \right);$$

$$T_1 = \frac{1}{2} \sum_{i=1}^n n_i (n_i - 1);$$

$$T_2 = \frac{1}{2} \sum_{j=1}^m m_j (m_j - 1).$$

Середньоквадратичне відхилення дорівнює

$$\sigma(\tau_b) = \sqrt{\frac{4n+10}{9(n^2-n)}}.$$

3. Якщо $m \neq n$, то для оцінки зв'язку реалізується **статистика Стюарда**

$$\tau_{st} = \frac{2(P-Q)\min(m,n)}{n^2(\min(m,n)-1)},$$

для якої

$$\sigma(\tau_{st}) = \frac{2\min(m,n)}{n^3(\min(m,n)-1)} \sqrt{n^2 \sum_{i=1}^n \sum_{j=1}^m n_{ij} (A_{ij} - B_{ij})^2 - 4n(P-Q)},$$

де

$$A_{ij} = \sum_{k=i+1}^n \sum_{l=j+1}^m n_{kl} + \sum_{k=1}^{i-1} \sum_{l=1}^{j-1} n_{kl};$$

$$B_{ij} = \sum_{k=i+1}^n \sum_{l=1}^{j-1} n_{kl} + \sum_{k=1}^{i-1} \sum_{l=j+1}^m n_{kl}.$$

4. Коефіцієнт **рангової кореляції Спірмена** для таблиць сполучень, що оцінює зв'язаність ознак X , Y , визначається за формулою

$$\tau_s = \frac{12 \sum_{i=1}^n \sum_{j=1}^m n_{ij} \left(\sum_{k=1}^{i-1} n_k + \frac{1}{2} n_k - \frac{1}{2} n \right) \left(\sum_{l=1}^{j-1} m_l + \frac{1}{2} m_l - \frac{1}{2} n \right)}{\sqrt{\left(n^3 - n - \sum_{i=1}^n (n_i^3 - n_i) \right) \left(n^3 - n - \sum_{j=1}^m (m_j^3 - m_j) \right)}},$$

при цьому

$$\sigma(\tau_s) = \frac{1 - \tau_s^2}{\sqrt{n-2}}.$$

Значення τ_b , τ_{st} , τ_s змінюються у діапазоні від -1 до $+1$, отже, мають усі властивості лінійного парного коефіцієнта кореляції. Враховуючи, що коефіцієнти τ_b , τ_c , τ_s розподілені асимптотично нормально, їх істинне значення з довірчою ймовірністю $\beta = 1 - \alpha$ оцінюється через довірчий інтервал, межами якого є (з точністю до оцінки)

$$\tau_{n,\beta} = \hat{\theta} \pm u_{\alpha/2} \sigma\{\hat{\theta}\}.$$

3.3. Одновимірний регресійний аналіз

Подальший аналіз змінних, для яких встановлена наявність стохастичного зв'язку, передбачає ідентифікацію та відтворення регресійної залежності за ними.

3.3.1. Лінійний регресійний аналіз

Найпростіша форма оцінки стохастичного зв'язку – одновимірний лінійний регресійний аналіз, за яким формуються обчислювальні процедури відтворення лінії регресії. Припускається, що дві нормально розподілені випадкові величини η та ξ зв'язані **лінійною регресійною залежністю**

$$\eta = \theta_1 + \theta_2 \xi + \varepsilon, \quad (3.1)$$

де ε – похибка, яка має нормальний розподіл, причому

$$E\{\varepsilon\} = 0; \quad D\{\varepsilon\} = \sigma_\varepsilon^2 = \text{const}.$$

Якщо обробці підлягає масив даних $\Omega_{2,N} = \{(x_l, y_l); l = \overline{1, N}\}$, **лінійна регресійна модель має вигляд**

$$\bar{y}(x) = a + bx, \quad (3.2)$$

тоді **оцінкою** наведеної залежності є

$$\hat{\bar{y}}(x) = \hat{a} + \hat{b}x,$$

де \hat{a} , \hat{b} – оцінки вектора параметрів регресії $\vec{\Theta} = \{\theta_1, \theta_2\}$ (параметрів a , b).

Відповідно до визначення **регресія** – це залежність середнього значення однієї випадкової величини від однієї або кількох інших:

$$\bar{y}(x) = E\{\eta/\xi = x\}.$$

Неформальне визначення таке: регресія – це лінія (або крива), відносно якої розсіювання даних мінімальне (рис. 3.6). З огляду на це лінія, позначена пунктиром (рис. 3.6), не може бути лінією регресії.

Проведення регресійного аналізу не обмежується відтворенням лінійної залежності. Можлива оцінка залежностей

$$\eta = \sum_{i=0}^s \theta_i \xi^i + \varepsilon, \quad (3.3)$$

чи будь-яких інших нелінійних залежностей:

$$\eta = \varphi(\xi; \vec{\Theta}), \quad \vec{\Theta} = \{\theta_i; i = \overline{0, s}\}.$$

Слід зазначити, що відтворення саме залежностей типу (3.1), (3.3) має найбільше поширення у відповідному програмному забезпеченні. Пояснюється це тим, що обчислювальні схеми відтворення регресії зазвичай базуються на методі найменших квадратів оцінки параметрів.

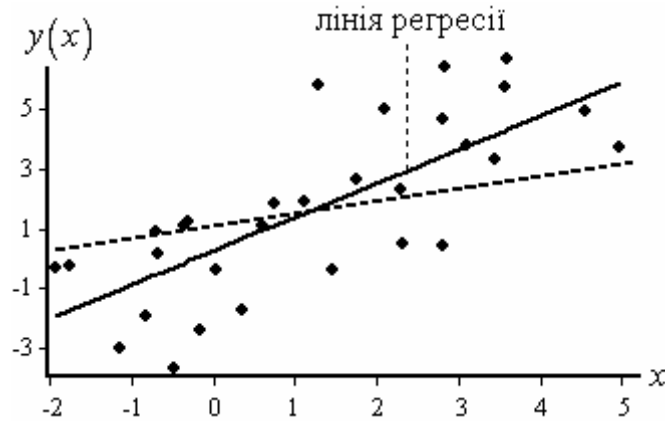


Рис. 3.6. Графік лінійної регресійної залежності

Етапами обчислювальної схеми відтворення функції регресії є:

- 1) перевірка виконання початкових умов регресійного аналізу;
- 2) ідентифікація вигляду регресійної залежності;
- 3) вибір типу функції регресії $\bar{y}(x) = \varphi(x; \bar{\Theta})$ та оцінка вектора параметрів $\hat{\bar{\Theta}}$;
- 4) дослідження якості відтворення регресії.

Для переліку задач обробки даних вводиться процедура порівняння двох або кількох регресійних залежностей. Якщо мають місце нелінійні залежності, то процедури знаходження оцінок параметрів та довірчого оцінювання відрізняються від процедури лінійної оцінки.

Початкові умови регресійного аналізу. Умови, що забезпечують застосування методів параметричного регресійного аналізу (наприклад, методу найменших квадратів), такі:

1. Сумісний розподіл випадкових величин η , ξ має бути **нормальним**.
2. **Дисперсія залежної змінної** y залишається сталою під час зміни значення аргументу x , отже,

$$D\{y/x\} = \sigma_y^2 = \text{const} \quad (3.4)$$

або пропорційною деякій відомій функції від x :

$$D\{y/x\} = \sigma_y^2 h^2(x), \quad (3.5)$$

де $h(x)$ – саме така функція.

3. Підсумки спостережень (x_l, y_l) стохастично незалежні, таким чином, результати, одержані на l -му кроці експерименту, не пов'язані з попереднім $(l-1)$ -м кроком і не містять інформації для $(l+1)$ -го кроку.

Нижче подана ілюстрація зазначених вимог (рис. 3.7).

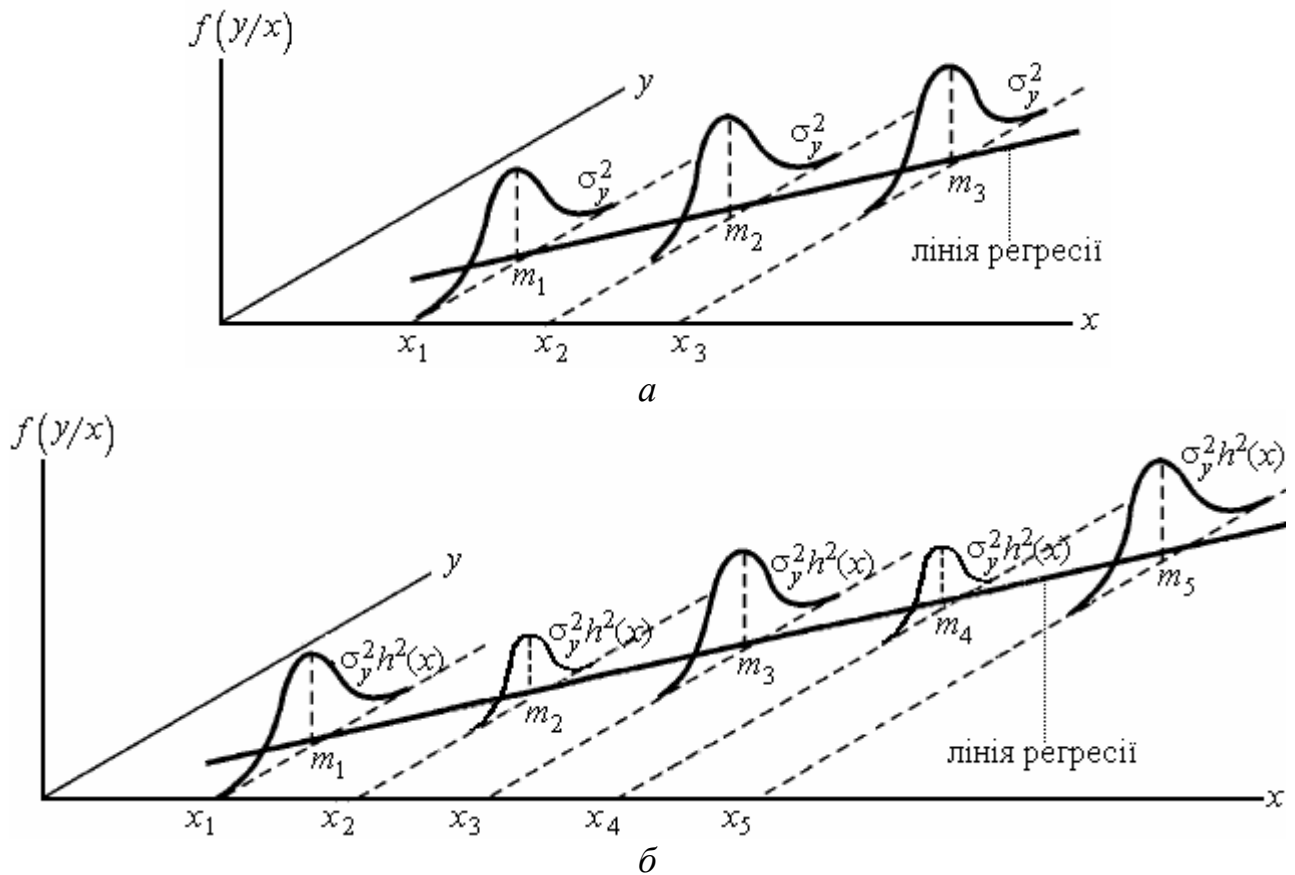


Рис. 3.7. Графічне зображення початкових умов регресійного аналізу:
 a – дисперсія y стала; b – дисперсія y пропорційна $h(x)$

На практиці допускається формальне відхилення від указаних вимог. Наприклад, якщо обсяг вибірок досить великий, можливе порушення першої умови. Перевірка виконання першої та третьої умов не викликає труднощів. Для перевірки другої використовують критерій однорідності для дисперсій (критерій Бартлетта). Розглянемо його використання для даної задачі.

Нехай для кожного з $X = \{x_i; i = \overline{1, k}\}$ зафіксовані $Y = \{y_{i,j}; i = \overline{1, k}, j = \overline{1, m_i}\}$ значень залежної змінної. Загальний обсяг експериментальних даних Y за всіма x_i дорівнює

$$N = \sum_{i=1}^k m_i,$$

отже, використовується масив $\Omega_{2,N} = \{(x_i, y_{i,j}); i = \overline{1, k}, j = \overline{1, m_i}\}$.

Зауваження 3.4. Відносно формування масиву $\{(x_i, y_{i,j}); i = \overline{1, k}, j = \overline{1, m_i}\}$ на основі $\{(x_l, y_l); l = \overline{1, N}\}$ див. заув. 3.3.

Як статистичну характеристику гіпотези

$$H_0 : D\{y/x_1\} = \dots = D\{y/x_k\} = \sigma^2$$

використовують статистику

$$\Lambda = -\frac{1}{C} \sum_{i=1}^k m_i \ln \frac{S_{\bar{y}(x_i)}^2}{S^2},$$

яка при $m_i \geq 3$ приблизно має χ^2 -розподіл із кількістю степенів вільності $\nu = k - 1$. Константа C та відхилення $S_{\bar{y}(x_i)}^2$, S^2 визначаються за формулами

$$C = 1 + \frac{1}{3(k-1)} \left(\sum_{i=1}^k \frac{1}{m_i} - \frac{1}{N} \right),$$

$$S_{\bar{y}(x_i)}^2 = \frac{1}{m_i - 1} \sum_{j=1}^{m_i} (y_{i,j} - \bar{y}_i)^2,$$

де

$$\bar{y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{i,j};$$

$$S^2 = \frac{1}{N - k} \sum_{i=1}^k (m_i - 1) S_{\bar{y}(x_i)}^2.$$

Якщо виявиться, що $\Lambda > \chi_{\alpha, \nu}^2$, де α – помилка першого роду, то гіпотезу H_0 відкидають, отже, порушена умова (3.4). У цьому випадку висувають гіпотезу відносно умови (3.5):

$$H_0 : \frac{D\{y/x_1\}}{h^2(x_1)} = \dots = \frac{D\{y/x_k\}}{h^2(x_k)} = \sigma^2.$$

Як статистичну характеристику використовують статистику

$$\Lambda' = -\frac{1}{C} \sum_{i=1}^k m_i \ln \frac{S_{\bar{y}(x_i)}'^2}{S'^2},$$

де

$$S'^2_{\bar{y}(x_i)} = \frac{S^2_{\bar{y}(x_i)}}{h^2(x_i)};$$

$$S'^2 = \frac{1}{N-k} \sum_{i=1}^k (m_i - 1) S'^2_{\bar{y}(x_i)}.$$

Наступна процедура перевірки гіпотези аналогічна розглянутій вище. Якщо і в даному випадку головна гіпотеза буде відкинута, маємо порушення другої умови. У цьому разі необхідно реалізовувати непараметричні процедури відтворення регресії.

Ідентифікація регресії. Метою процедури ідентифікації вигляду регресії є:

- 1) виявлення зв'язку поміж X та Y ;
- 2) за наявності зв'язку проведення класифікації на лінійність або нелінійність як відносно змінних X та Y , так і щодо вектора параметрів $\vec{\Theta}$.

Процедура ідентифікації зумовлює реалізацію і візуальної схеми, і кількісної оцінки зв'язку. У процесі візуалізації оцінюються початкові масиви, які відображаються у вигляді кореляційного поля (див. рис. 3.5).

Якщо кореляційне поле вписується в коло, то зв'язок між X та Y відсутній. Для поля у вигляді овалу має місце лінійна регресійна залежність. Для кореляційного поля складної конфігурації необхідно здійснити підбір нелінійної функції. Вибираючи вигляд регресії, слід комбінувати дослідження розташування точок кореляційного поля з логіко-професійним аналізом, тобто приймати рішення щодо вигляду кривої згідно з виглядом кореляційного поля. Найпростіші є процедури, що описують лінійний зв'язок відносно оцінюваного вектора параметрів. Практично це алгебричні поліноми порядку, не вищого за четвертий.

Під час проведення ідентифікації за допомогою числових методів реалізується двохетапна процедура. На першому етапі здійснюється статистичний аналіз, підсумком якого є знаходження оцінок $\hat{r}_{x,y}$, \hat{r} та перевірка їх значущості. Наприклад, за умови, що коефіцієнт парної кореляції $\hat{r}_{x,y}$ значущий, висувається твердження про лінійний регресійний зв'язок поміж Y і X . Якщо ж ідентифікується нелінійна регресійна залежність, то її тип уточнюється процедурою візуалізації кореляційного поля та накладенням на нього типових кривих.

Статистичний аналіз, який ґрунтується на процедурах перевірки статистичних гіпотез про загальний вигляд регресійної залежності, проводиться на другому етапі. Найбільш потужні критерії перевірки гіпотези про вигляд функції регресії запропоновані для лінійної залежності (див. далі перевірку адекватності відтвореної регресійної моделі).

Відтворення лінійної регресійної залежності. Загальноприйнятим методом оцінки параметрів регресії є МНК. Нехай на основі процедури іденти-

фікації встановлено, що поміж Y , X існує лінійний зв'язок

$$\bar{y}(x) = a + bx.$$

При цьому оцінки параметрів регресійної моделі знаходять з умови мінімуму функціонала залишкової дисперсії

$$S_{\text{зал}}^2 = \frac{1}{N-2} \sum_{l=1}^N (y_l - \hat{\bar{y}}(x_l))^2 = \frac{1}{N-2} \sum_{l=1}^N (y_l - \hat{a} - \hat{b}x_l)^2,$$

що формується як сума квадратів відхилень результатів спостережень від лінії регресії (рис. 3.8).

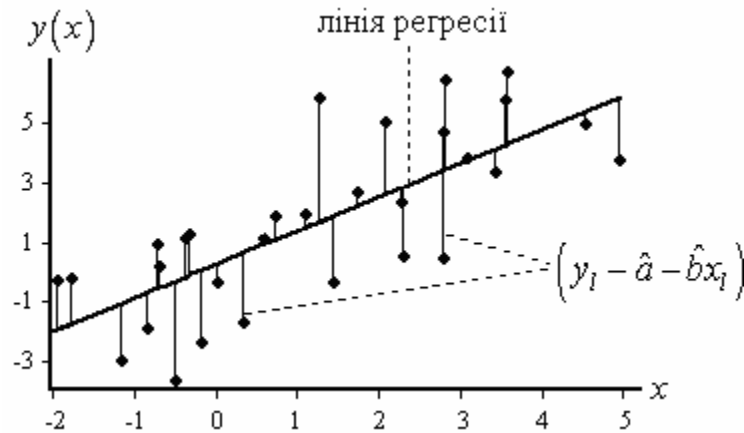


Рис. 3.8. Графічне зображення відхилення результатів спостережень від лінії регресії

Необхідна та достатня умова $\min_{a,b} S_{\text{зал}}^2$ визначається системою лінійних рівнянь

$$\begin{cases} \frac{\partial S_{\text{зал}}^2}{\partial \hat{a}} = 0, \\ \frac{\partial S_{\text{зал}}^2}{\partial \hat{b}} = 0 \end{cases}$$

або

$$\begin{pmatrix} 1 & \bar{x} \\ \bar{x} & \overline{x^2} \end{pmatrix} \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} \bar{y} \\ \overline{xy} \end{pmatrix},$$

звідки

$$\hat{a} = \frac{\overline{yx^2} - \bar{x}\overline{xy}}{\overline{x^2} - \bar{x}^2},$$

$$\hat{b} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2},$$

тобто

$$\hat{a} = \bar{y} - \hat{b}\bar{x},$$

$$\hat{b} = \hat{r}_{x,y} \frac{\hat{\sigma}_y}{\hat{\sigma}_x}.$$

Якщо початкові дані подані у вигляді масиву $\{(x_i, y_{i,j}); j = \overline{1, m_i}, i = \overline{1, k}\}$, то оцінки лінійної регресії обчислюють з умови

$$\min_{\hat{a}, \hat{b}} S_{\text{зал}}^2 = \min_{\hat{a}, \hat{b}} \frac{1}{N-2} \sum_{i=1}^k \sum_{j=1}^{m_i} (y_{i,j} - \hat{a} - \hat{b}x_i)^2,$$

яка визначає

$$\hat{a} = \frac{\sum_{i=1}^k \sum_{j=1}^{m_i} y_{i,j} \sum_{i=1}^k m_i x_i^2 - \sum_{i=1}^k \sum_{j=1}^{m_i} y_{i,j} x_i \sum_{i=1}^k m_i x_i}{N \sum_{i=1}^k m_i x_i^2 - \left(\sum_{i=1}^k m_i x_i \right)^2},$$

$$\hat{b} = \frac{N \sum_{i=1}^k \sum_{j=1}^{m_i} y_{i,j} x_i - \sum_{i=1}^k \sum_{j=1}^{m_i} y_{i,j} \sum_{i=1}^k m_i x_i}{N \sum_{i=1}^k m_i x_i^2 - \left(\sum_{i=1}^k m_i x_i \right)^2}.$$

Можна показати, що

$$\hat{a} = \bar{y} - \hat{b}\bar{x},$$

$$\hat{b} = \hat{r} \frac{\sigma_y}{\sigma_x},$$

де

$$\hat{r} = \frac{\frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{m_i} y_{i,j} x_i - \bar{x}\bar{y}}{\sigma_x \sigma_y}.$$

Якщо має місце $D\{y/x\} = \sigma^2 h^2(x)$, то початковий масив даних $\{(x_i, y_{i,j}); i = \overline{1, k}, j = \overline{1, m_i}\}$ переформовують у $\{(x_i, y_{i,j}), \omega_i; j = \overline{1, m_i}, i = \overline{1, k}\}$, де

$\omega_i = \frac{1}{h^2(x_i)}$. Подальша процедура одержання оцінок параметрів \hat{a} , \hat{b} зводиться до знаходження

$$\min_{\hat{a}, \hat{b}} S_{3\text{ал}}^2 = \min_{\hat{a}, \hat{b}} \frac{1}{N-2} \sum_{i=1}^k \sum_{j=1}^{m_i} \omega_i (y_{i,j} - \hat{a} - \hat{b}x_i)^2.$$

Реалізуючи МНК, розв'язують таку систему лінійних рівнянь:

$$\begin{pmatrix} \sum_{i=1}^k \omega_i m_i & \sum_{i=1}^k \omega_i m_i x_i \\ \sum_{i=1}^k \omega_i m_i x_i & \sum_{i=1}^k \omega_i m_i x_i^2 \end{pmatrix} \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^k \omega_i m_i \bar{y}_i \\ \sum_{i=1}^k \omega_i m_i \bar{y}_i x_i \end{pmatrix},$$

де

$$\bar{y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{i,j}.$$

Із розв'язку наведеної системи одержують

$$\hat{a} = \frac{\sum_{i=1}^k \omega_i m_i \bar{y}_i \sum_{i=1}^k \omega_i m_i x_i^2 - \sum_{i=1}^k \omega_i m_i x_i \sum_{i=1}^k \omega_i m_i \bar{y}_i x_i}{\sum_{i=1}^k \omega_i m_i \sum_{i=1}^k \omega_i m_i x_i^2 - \left(\sum_{i=1}^k \omega_i m_i x_i \right)^2},$$

$$\hat{b} = \frac{\sum_{i=1}^k \omega_i m_i \sum_{i=1}^k \omega_i m_i \bar{y}_i x_i - \sum_{i=1}^k \omega_i m_i x_i \sum_{i=1}^k \omega_i m_i \bar{y}_i}{\sum_{i=1}^k \omega_i m_i \sum_{i=1}^k \omega_i m_i x_i^2 - \left(\sum_{i=1}^k \omega_i m_i x_i \right)^2}.$$

Аналіз наведених процедур знаходження оцінок параметрів лінійної регресії дозволяє керувати обчислювальним процесом відтворення лінії регресії залежно від типу початкового масиву даних і вигляду оцінки $D\{y/x\}$.

Дослідження якості відтворення лінії регресії для випадку $D\{y/x\} = \sigma_y^2 = \text{const}$ зумовлює реалізацію таких процедур:

- 1) обчислення коефіцієнта детермінації R^2 ;
- 2) дослідження значущості й точності оцінок параметрів \hat{a} , \hat{b} ;
- 3) оцінювання відхилень окремих значень y_l , $l = \overline{1, N}$ залежної змінної від емпіричної регресії $\hat{y}(x_l)$;
- 4) побудови довірчого інтервалу для прогнозу нового спостереження;

5) **побудови довірчого інтервалу для лінії регресії** $\bar{y}(x) = a + bx$ з урахуванням її оцінки $\hat{y}(x) = \hat{a} + \hat{b}x$;

6) **перевірки адекватності** даним відтвореної моделі регресії $\hat{y}(x) = \varphi(x, \hat{\Theta})$.

Коефіцієнт детермінації R^2 – показник, що визначає, якою мірою варіабельність ознаки Y пояснюється поведінкою X . Більш точно, R^2 – це та частка дисперсії Y , яка пояснюється впливом X . Значення коефіцієнта детермінації обчислюють шляхом піднесення до квадрата значення оцінки коефіцієнта парної кореляції:

$$R^2 = \hat{r}_{x,y}^2 \cdot 100\%.$$

Зрозуміло, що $\hat{r}_{x,y}^2 \in [0;1]$ і більші значення R^2 свідчать про «якісне» відтворення лінійної регресії.

Дослідження точності оцінок параметрів \hat{a} , \hat{b} становить результат процедури перевірки гіпотез про значущість

$$H_0 : a = 0,$$

$$H_0 : b = 0$$

та гіпотез про рівність оцінок деяким значенням параметрів

$$H_0 : a = \hat{a},$$

$$H_0 : b = \hat{b}.$$

Зазначені гіпотези перевіряються на основі t -тесту з урахуванням середньоквадратичних оцінок параметрів \hat{a} , \hat{b} :

$$S_a = S_{\text{зал}} \sqrt{\frac{1}{N} + \frac{\bar{x}^2}{\sigma_x^2 (N-1)}},$$

$$S_b = \frac{S_{\text{зал}}}{\sigma_x \sqrt{N-1}}.$$

Тоді відповідні t -статистики, як завжди, дорівнюють

$$t_a = \frac{\hat{a} - a}{S_a},$$

$$t_b = \frac{\hat{b} - b}{S_b}.$$

Слід відзначити, що в разі спростування гіпотези $H_0 : b = 0$ (невиконання нерівності

$$|t_b| \leq t_{\alpha/2, v},$$

$v = N - 2$) говорять про значущість регресійного зв'язку.

Інтервальне оцінювання параметрів лінійної регресії здійснюють, виходячи з нерівностей ($v = N - 2$)

$$\hat{a} - t_{\alpha/2, v} S_a \leq a \leq \hat{a} + t_{\alpha/2, v} S_a,$$

$$\hat{b} - t_{\alpha/2, v} S_b \leq b \leq \hat{b} + t_{\alpha/2, v} S_b.$$

Оцінка відхилень окремих значень спостережень y_i від лінії регресії дозволяє вказати стандартну похибку регресійної оцінки. Значення $S_{\text{зал}}$ приблизно вказує величину залишків для наявних даних у тих же одиницях, у яких вимірюється Y . Крім того, оцінка відхилень зумовлює побудову **припустимих (або толерантних) меж** на основі оцінки $S_{\text{зал}}^2$ (по суті, дисперсії σ_ε^2 похибки ε в моделі (3.1)). Значення оцінки стандартного відхилення похибки обчислюють зі співвідношення для знаходження залишкової дисперсії:

$$\hat{\sigma}_\varepsilon = S_{\text{зал}} = \sqrt{\frac{1}{N-2} \sum_{l=1}^N (y_l - \hat{y}(x_l))^2} = \hat{\sigma}_y \sqrt{(1 - \hat{r}_{x,y}^2) \frac{N-1}{N-2}}.$$

У ході **інтерпретації** величина σ_ε дозволяє припускати розташування 95% спостережень у толерантних межах (рис. 3.9) на такій відстані від лінії регресії, яка не перевищує приблизно $2\sigma_\varepsilon$ (відповідно дві третини даних розташовані на відстані, не більшій ніж σ_ε).

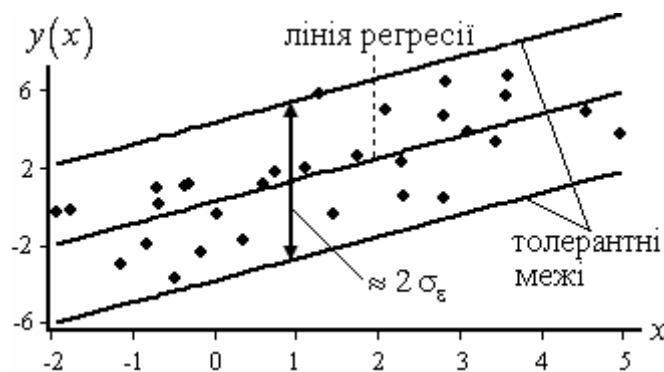


Рис. 3.9. Графічне зображення толерантних меж для лінійної регресії

Толерантні межі $\hat{y}_{\min}(x)$, $\hat{y}_{\max}(x)$ для y_l , $l = \overline{1, N}$ визначаються за виразами:

$$\hat{y}_{\min}(x) = \hat{y}(x) - t_{\alpha/2, v} \hat{\sigma}_{\varepsilon},$$

$$\hat{y}_{\max}(x) = \hat{y}(x) + t_{\alpha/2, v} \hat{\sigma}_{\varepsilon}$$

де $v = N - 2$.

Необхідність побудови **довірчого інтервалу для прогнозу нового спостереження** виникає в разі використання моделі регресії для знаходження y за деякого заданого x_0 . У такій ситуації існує два джерела невизначеності. По-перше, оскільки \hat{a} та \hat{b} являють собою оцінки, то $\hat{a} + \hat{b}x_0$ містить елемент невизначеності. По-друге, присутня похибка ε , яка є частиною лінійної моделі і яку також треба враховувати, аналізуючи окремі спостереження. З огляду на це величина $S_{(y|x_0)}$ стандартної похибки y при заданому x_0 обчислюється так:

$$S_{(y|x_0)} = \sqrt{\hat{\sigma}_{\varepsilon}^2 \left(1 + \frac{1}{N}\right) + S_b^2 (x_0 - \bar{x})^2}.$$

Відповідний довірчий інтервал для нового спостереження y за певного x_0 (рис. 3.10)

$$\hat{y}(x_0) - t_{\alpha/2, v} S_{(y|x_0)} \leq y \leq \hat{y}(x_0) + t_{\alpha/2, v} S_{(y|x_0)},$$

де $v = N - 2$.



Рис. 3.10. Графічне зображення довірчого інтервалу для прогнозу нового спостереження у випадку лінійної регресії

Інтервальне оцінювання лінійної регресії здійснюється шляхом призначення довірчого γ -імовірного ($\gamma = 1 - \alpha$) інтервалу. На відміну від попереднього випадку, оцінюється середнє значення $\bar{y}(x)$ при $\forall x$. У такій ситуації під час оцінки $S_{(\bar{y}|x)}$ стандартної похибки $\bar{y}(x)$ не враховується випадкова похибка ε (згідно з моделлю (3.2)):

$$S_{(\bar{y}|x)} = \sqrt{\hat{\sigma}_\varepsilon^2 \frac{1}{N} + S_b^2 (x - \bar{x})^2}.$$

Тоді довірчий інтервал визначається з нерівності

$$\hat{\bar{y}}(x) - t_{\alpha/2, \nu} S_{(\bar{y}|x)} \leq \bar{y}(x) \leq \hat{\bar{y}}(x) + t_{\alpha/2, \nu} S_{(\bar{y}|x)},$$

де $\nu = N - 2$.

Слід наголосити на існуванні двох закономірностей (рис. 3.11):

- 1) чим більша є для $\forall x$ різниця $|x - \bar{x}|$, тим ширша є величина довірчого інтервалу, отже, довірчий інтервал розходиться відносно віддалення x від \bar{x} ;
- 2) чим більший обсяг вибірки N , тим менша є величина довірчого інтервалу.

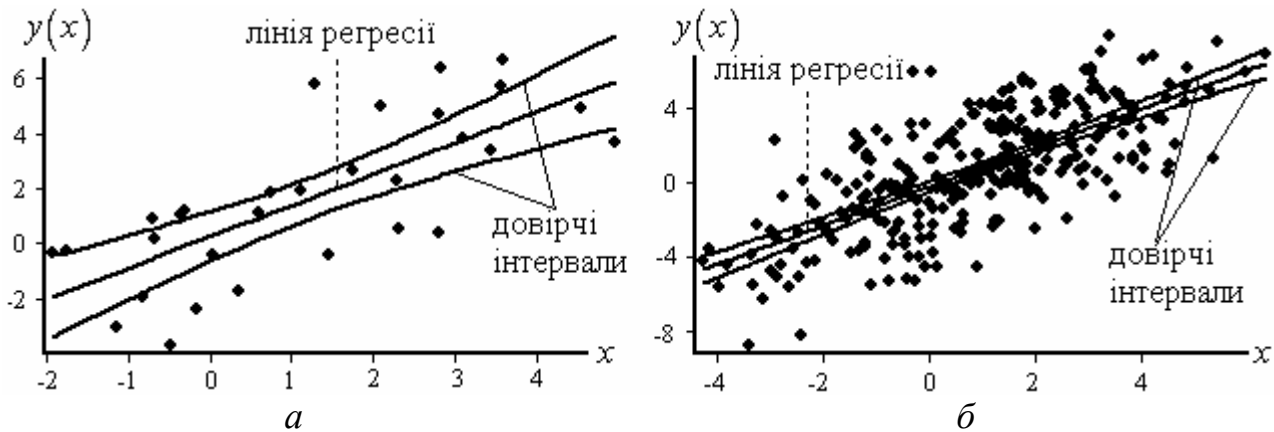


Рис. 3.11. Графічне зображення інтервального оцінювання лінійної регресії:
а – $N = 30$; б – $N = 300$

Для наочності нижче наведені толерантні межі, довірчі інтервали для лінії регресії та прогнозного значення (рис. 3.12).



Рис. 3.12. Графічне зображення довірчого оцінювання лінійної регресії

Із метою перевірки адекватності відтвореної моделі регресії $\hat{y}(x) = \varphi(x, \hat{\Theta})$ висувається статистична гіпотеза $H_0 : \bar{y}(x) = \hat{y}(x)$ про вигляд регресійної залежності. Критерій перевірки гіпотези базується на статистиці f :

$$f = \frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_y^2},$$

яка має F -розподіл Фішера з кількістю степенів вільності $\nu_1 = N - 1$, $\nu_2 = N - 3$.

Значення f порівнюють із критичним f_{α, ν_1, ν_2} і за виконання нерівності

$$f \leq f_{\alpha, \nu_1, \nu_2}$$

роблять висновок про адекватність та значущість відтвореної залежності.

Зауваження 3.5. Аналогічна процедура може бути реалізована під час розв'язання задачі про відповідність даним деякої конкретної регресійної моделі (не обов'язково одержаної в результаті відтворення, а, наприклад, суто евристичної).

Як правило, критерій, що враховує **конкретний вигляд регресійної залежності** $\hat{y}(x) = \varphi(x, \hat{\Theta})$, використовують на етапі попередньої ідентифікації моделі регресійної залежності для перевірки гіпотези

$$H_0 : \bar{y}(x) = \varphi(x; \hat{\Theta}).$$

Не зменшуючи загальності, розглянемо дані у вигляді масиву $\{x_i, y_{i,j}; i = \overline{1, k}, j = \overline{1, m_i}\}$. У випадку $D\{y/x\} = \sigma^2 = \text{const}$ для перевірки головної гіпотези реалізується статистична характеристика

$$f = \frac{(N - k) \sum_{i=1}^k m_i \left(\bar{y}_i - \hat{y}(x_i; \hat{\Theta}) \right)^2}{(k - s - 1) \sum_{i=1}^k \sum_{j=1}^{m_i} (y_{i,j} - \bar{y}_i)^2},$$

яка має F -розподіл із кількістю степенів вільності $\nu_1 = k - s - 1$, $\nu_2 = N - k$. Якщо $f \leq f_{\alpha, \nu_1, \nu_2}$, то запропонована регресійна залежність є значуща.

Якщо $D\{y/x\} = \sigma_y^2 h^2(x)$, для перевірки H_0 реалізують статистику

$$f' = \frac{(N - k) \sum_{i=1}^k \omega_i m_i \left(\bar{y}_i - \hat{y}(x_i; \hat{\Theta}) \right)^2}{(k - s - 1) \sum_{i=1}^k \omega_i \sum_{j=1}^{m_i} (y_{i,j} - \bar{y}_i)^2},$$

де

$$\omega_i = \frac{1}{h^2(x_i)},$$

що має F -розподіл із кількістю степенів вільності $v_1 = k - s - 1$, $v_2 = N - k$. Процедура перевірки гіпотези є еквівалентна вищенаведеній.

Зауваження 3.6. У випадку перевірки гіпотези про лінійний зв'язок $s = 2$.

До задач лінійного регресійного аналізу обробки даних належить **процедура порівняння двох або більше регресійних залежностей**. Слід відзначити, що подібна задача є актуальна, коли з однієї генеральної сукупності одержані різні вибірки. Отже, нехай за вибірковими даними $\Omega_{2,N_1} = \{x_{1,l}, y_{1,l}; l = \overline{1, N_1}\}$, $\Omega_{2,N_2} = \{x_{2,l}, y_{2,l}; l = \overline{1, N_2}\}$ відтворені лінії регресії:

$$\hat{y}_1(x) = \hat{a}_1 + \hat{b}_1(x - \bar{x}_1),$$

$$\hat{y}_2(x) = \hat{a}_2 + \hat{b}_2(x - \bar{x}_2),$$

залишкова дисперсія для яких відповідно визначається так:

$$S_{1,3ал}^2 = \frac{1}{N_1 - 2} \sum_{l=1}^{N_1} \left(y_{1,l} - \hat{a}_1 - \hat{b}_1(x_{1,l} - \bar{x}_1) \right)^2,$$

$$S_{2,3ал}^2 = \frac{1}{N_2 - 2} \sum_{l=1}^{N_2} \left(y_{2,l} - \hat{a}_2 - \hat{b}_2(x_{2,l} - \bar{x}_2) \right)^2.$$

Необхідно оцінити, чи істотна різниця поміж $\hat{y}_1(x)$ і $\hat{y}_2(x)$.

Процедура перевірки гіпотези

$$H_0 : \bar{y}_1(x) = \bar{y}_2(x)$$

має розбиття на декілька етапів:

1. Спочатку перевіряється гіпотеза про збіг залишкових дисперсій, отже, про рівність дисперсій залишків:

$$H_0 : \sigma_{1,\varepsilon}^2 = \sigma_{2,\varepsilon}^2.$$

Перевірка здійснюється з урахуванням статистичної характеристики

$$f = \begin{cases} \frac{S_{1,3ал}^2}{S_{2,3ал}^2}, & \text{якщо } S_{1,3ал}^2 > S_{2,3ал}^2, \\ \frac{S_{2,3ал}^2}{S_{1,3ал}^2}, & \text{якщо } S_{1,3ал}^2 < S_{2,3ал}^2, \end{cases}$$

яка має розподіл Фішера зі степенями вільності $v_1 = N_1 - 2$, $v_2 = N_2 - 2$. У разі $f \leq f_{\alpha, v_1, v_2}$ головна гіпотеза правильна, при цьому обчислюється зведена оцінка дисперсії залишків:

$$S^2 = \frac{(N_1 - 2)S_{1, \text{зал}}^2 + (N_2 - 2)S_{2, \text{зал}}^2}{N_1 + N_2 - 4}.$$

2. У випадку рівності залишкових дисперсій реалізується обчислювальна схема перевірки гіпотези

$$H_0 : b = \hat{b}_1 = \hat{b}_2$$

на основі статистичної характеристики

$$t = \frac{\hat{b}_1 - \hat{b}_2}{S \sqrt{\frac{1}{(N_1 - 1)\hat{\sigma}_{x_1}^2} + \frac{1}{(N_2 - 1)\hat{\sigma}_{x_2}^2}}}, \quad (3.6)$$

де $\hat{\sigma}_{x_1}^2$, $\hat{\sigma}_{x_2}^2$ – незсунені оцінки дисперсій x_1 , x_2 .

Статистична характеристика (3.6) має t -розподіл із $v = N_1 + N_2 - 4$ степенями вільності, тоді:

1) якщо $|t| \leq t_{\alpha/2, v}$, то гіпотеза H_0 правильна, таким чином, регресійні прямі є паралельні, а лінії регресії можуть збігатись або різнитися постійними коефіцієнтами \hat{a}_1 , \hat{a}_2 ;

2) при $|t| > t_{\alpha/2, v}$ гіпотеза H_0 повинна бути відкинута, отже, регресійні прямі мають різні кути нахилу.

У разі прийняття H_0 обчислюється $\hat{b}_1 = \hat{b}_2 = \hat{b}$:

$$\hat{b} = \frac{(N_1 - 1)\hat{\sigma}_{x_1}^2 \hat{b}_1^2 + (N_2 - 1)\hat{\sigma}_{x_2}^2 \hat{b}_2^2}{(N_1 - 1)\hat{\sigma}_{x_1}^2 + (N_2 - 1)\hat{\sigma}_{x_2}^2}.$$

3. На завершальному етапі перевіряється

$$H_0 : a = \hat{a}_1 = \hat{a}_2$$

на основі статистичної характеристики

$$t = \frac{\hat{b} - \hat{b}_0}{S_0}, \quad (3.7)$$

де

$$\hat{b}_0 = \frac{\bar{y}_1 - \bar{y}_2}{\bar{x}_1 - \bar{x}_2};$$

$$S_0^2 = S^2 \left(\frac{1}{(N_1 - 1)\hat{\sigma}_{x_1}^2 + (N_2 - 1)\hat{\sigma}_{x_2}^2} + \frac{1}{(\bar{x}_1 - \bar{x}_2)^2} \left(\frac{1}{N_1} + \frac{1}{N_2} \right) \right).$$

Статистична характеристика (3.7) має t -розподіл з $v = N_1 + N_2 - 4$ степенями вільності, тому якщо $|t| \leq t_{\alpha/2, v}$, то обидві регресійні прямі вважаються ідентичними, у противному разі має місце статистично значущий незбіг.

Якщо дисперсії залишків $S_{1,3ал}^2$, $S_{2,3ал}^2$ різняться істотно, а отже, гіпотеза про рівність дисперсій залишків не підтверджується, то для порівняння регресійних прямих $\hat{y}_1(x)$, $\hat{y}_2(x)$ адекватних статистичних критеріїв не існує. У цьому випадку рекомендується застосовувати процедуру порівняння регресій на основі наближених формул шляхом перевірки двох гіпотез. Аналогічно попередньому алгоритму перевіряється гіпотеза

$$H_0 : b = \hat{b}_1 = \hat{b}_2$$

з урахуванням статистичної характеристики

$$t = \frac{\hat{b}_1 - \hat{b}_2}{S \sqrt{\frac{S_{1,3ал}^2}{N_1 \hat{\sigma}_{x_1}^2} + \frac{S_{2,3ал}^2}{N_2 \hat{\sigma}_{x_2}^2}}},$$

яка має t -розподіл із кількістю степенів вільності

$$v = \left[\left(\frac{C_0^2}{N_1 - 2} + \frac{(1 - C_0)^2}{N_2 - 2} \right)^{-1} \right],$$

де

$$C_0 = \frac{S_{1,3ал}^2}{N_1 \hat{\sigma}_{x_1}^2} \bigg/ \left(\frac{S_{1,3ал}^2}{N_1 \hat{\sigma}_{x_1}^2} + \frac{S_{2,3ал}^2}{N_2 \hat{\sigma}_{x_2}^2} \right);$$

$[\cdot]$ – ціла частина.

Якщо $|t| \leq t_{\alpha/2, v}$, то правильна гіпотеза про збіг кутових коефіцієнтів ліній регресій.

Розглянута нижче процедура полягає в перевірці гіпотези $H_0 : a = \hat{a}_1 = \hat{a}_2$ на основі статистичної характеристики

$$u = \frac{\hat{b} - \hat{b}_0}{S_{10}}, \quad (3.8)$$

де

$$\hat{b} = \left(\hat{b}_1 \frac{N_1 \hat{\sigma}_{x_1}^2}{S_{1,3ал}^2} + \hat{b}_2 \frac{N_2 \hat{\sigma}_{x_2}^2}{S_{2,3ал}^2} \right) / \left(\frac{N_1 \hat{\sigma}_{x_1}^2}{S_{1,3ал}^2} + \frac{N_2 \hat{\sigma}_{x_2}^2}{S_{2,3ал}^2} \right);$$

$$\hat{b}_0 = \frac{\bar{y}_1 - \bar{y}_2}{\bar{x}_1 - \bar{x}_2};$$

$$S_{10}^2 = \frac{N_2 S_{1,3ал}^2 + N_1 S_{2,3ал}^2}{N_1 N_2 (\bar{x}_1 - \bar{x}_2)^2} + \frac{S_{1,3ал}^2 S_{2,3ал}^2}{N_1 \hat{\sigma}_{x_1}^2 S_{2,3ал}^2 + N_2 \hat{\sigma}_{x_2}^2 S_{1,3ал}^2}.$$

Статистична характеристика (3.8) має нормальний розподіл, тому H_0 правильна, коли $|u| \leq u_{\alpha/2}$. Якщо дві наведені гіпотези правильні, робиться висновок про їх випадкову різницю, у противному разі має місце істотна розбіжність поміж $\hat{y}_1(x)$ і $\hat{y}_2(x)$.

3.3.2. Нелінійний регресійний аналіз

У багатьох випадках у процесі ідентифікації кореляційного поля виявляється, що треба відтворювати **нелінійну регресійну залежність**. При цьому підбір кривої може бути здійснений на основі:

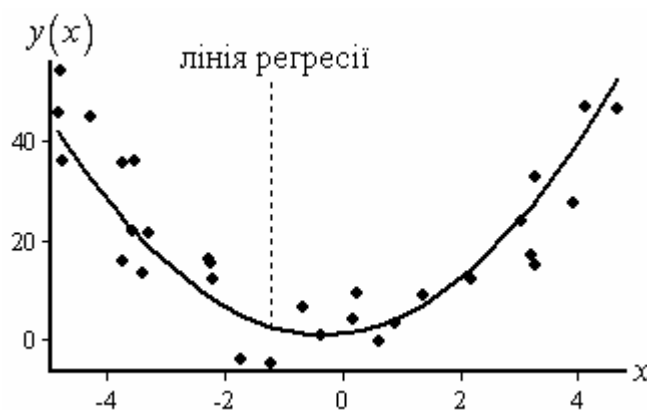


Рис. 3.13. Графік поліноміальної регресійної залежності другого порядку

1) **поліноміальної регресії** другого (рис. 3.13):

$$\bar{y}(x) = a + bx + cx^2 \quad (3.9)$$

або більш високого порядку:

$$\bar{y}(x) = a_0 + a_1 x + a_2 x^2 + \dots + a_k x^k, \quad k \geq 3; \quad (3.10)$$

2) **нелінійних залежностей** як **відносно параметрів**, так і **відносно аргументів лінії регресії**. Цей тип поділяється на регресії:

- ті, що зводяться до лінійної форми відносно параметрів (квазілінійні функції);
- нелінійні функції відносно параметрів, які не зводяться до лінійної форми.

Для нелінійних функцій, що зводяться до лінійної форми відносно оцінок параметрів, реалізуються різні перетворення координат (логарифмування, заміна змінних та ін.). Після переформування масиву даних до них можна застосувати МНК. Регресії, що характеризуються нелінійністю за оцінюваними параметрами зводяться до нелінійних рівнянь, одержаних за МНК, і для їх відтворення застосовуються ітераційні методи або методи апроксимації параметрів. Ортодоксальної теорії нелінійної регресії не існує. Проте зведення до лінійної форми відносно шуканих параметрів дозволяє реалізовувати статистичні критерії лінійної регресії.

Відтворення параболічної регресії

Безпосереднє застосування обчислювальної схеми МНК до регресійної залежності (3.9) не відрізняється від лінійної. Для залежності (3.10) обчислювальний процес відтворення емпіричної лінії регресії ускладнюється.

Розглянемо питання оцінки регресії типу (3.9) на основі масиву даних $\{(x_l, y_l); l = \overline{1, N}\}$. Реалізуючи МНК, з умови

$$\min_{\hat{a}, \hat{b}, \hat{c}} S_{3ал(1)}^2 = \min_{\hat{a}, \hat{b}, \hat{c}} \frac{1}{N-3} \sum_{l=1}^N (y_l - \hat{a} - \hat{b}x_l - \hat{c}x_l^2)^2,$$

еквівалентної

$$\frac{\partial S_{3ал(1)}^2}{\partial \hat{a}} = 0, \quad \frac{\partial S_{3ал(1)}^2}{\partial \hat{b}} = 0,$$

$$\frac{\partial S_{3ал(1)}^2}{\partial \hat{c}} = 0,$$

одержують

$$\hat{a} = \bar{y} - \hat{b}\bar{x} - \hat{c}\bar{x}^2,$$

де \hat{b} , \hat{c} отримують із системи рівнянь

$$\begin{cases} \hat{b} \sum_{l=1}^N (x_l - \bar{x})^2 + \hat{c} \sum_{l=1}^N (x_l^2 - \bar{x}^2)(x_l - \bar{x}) = \sum_{l=1}^N (y_l - \bar{y})(x_l - \bar{x}), \\ \hat{b} \sum_{l=1}^N (x_l^2 - \bar{x}^2)(x_l - \bar{x}) + \hat{c} \sum_{l=1}^N (x_l^2 - \bar{x}^2)^2 = \sum_{l=1}^N (y_l - \bar{y})(x_l^2 - \bar{x}^2). \end{cases}$$

Ця система є еквівалентна такій:

$$\begin{pmatrix} \hat{\sigma}_x^2 & (\overline{x^3} - \overline{x^2}\overline{x}) \\ (\overline{x^3} - \overline{x^2}\overline{x}) & (\overline{x^4} - (\overline{x^2})^2) \end{pmatrix} \begin{pmatrix} \hat{b} \\ \hat{c} \end{pmatrix} = \begin{pmatrix} \hat{r}_{x,y} \hat{\sigma}_x \hat{\sigma}_y \\ \frac{\hat{r}_{x,y} \hat{\sigma}_x \hat{\sigma}_y}{(y - \bar{y})(x^2 - \overline{x^2})} \end{pmatrix}, \quad (3.11)$$

де

$$\overline{x^k} = \frac{1}{N} \sum_{l=1}^N x_l^k, \quad k = 1, 2, 3, 4;$$

$$\overline{(y - \bar{y})(x^2 - \overline{x^2})} = \frac{1}{N} \sum_{l=1}^N (x_l^2 - \overline{x^2})(y_l - \bar{y}).$$

Із розв'язку системи (3.11) знаходять оцінки параметрів регресії \hat{b} , \hat{c} :

$$\hat{b} = \frac{\left(\overline{x^4} - (\overline{x^2})^2 \right) \hat{r}_{x,y} \hat{\sigma}_x \hat{\sigma}_y - (\overline{x^3} - \overline{x^2}\overline{x}) \overline{(y - \bar{y})(x^2 - \overline{x^2})}}{\hat{\sigma}_x^2 \left(\overline{x^4} - (\overline{x^2})^2 \right) - (\overline{x^3} - \overline{x^2}\overline{x})^2},$$

$$\hat{c} = \frac{\hat{\sigma}_x^2 \overline{(y - \bar{y})(x^2 - \overline{x^2})} - (\overline{x^3} - \overline{x^2}\overline{x}) \hat{r}_{x,y} \hat{\sigma}_x \hat{\sigma}_y}{\hat{\sigma}_x^2 \left(\overline{x^4} - (\overline{x^2})^2 \right) - (\overline{x^3} - \overline{x^2}\overline{x})^2}.$$

Наведені вирази і визначають обчислювальну процедуру відтворення параболічної регресії у вигляді (3.9).

Іншим прикладом обчислювальної процедури, що ґрунтується на МНК, може бути наступна. Подамо залежність (3.9) у такому вигляді:

$$\bar{y}(x) = a_1 + b_1 \varphi_1(x) + c_1 \varphi_2(x), \quad (3.12)$$

де

$$\begin{aligned} \varphi_1(x) &= x - \bar{x}; \\ \varphi_2(x) &= x^2 - \frac{\sum_{l=1}^N x_l^3 - \bar{x} \sum_{l=1}^N x_l^2}{\sum_{l=1}^N x_l^2 - N \bar{x}^2} (x - \bar{x}) + \\ &= x^2 - \frac{\overline{x^3} - \overline{x^2}\overline{x}}{\sigma_x^2} (x - \bar{x}) - \overline{x^2}. \end{aligned}$$

З умови

$$\min_{\hat{a}_1, \hat{b}_1, \hat{c}_1} S_{\text{зал}(2)}^2 = \min_{\hat{a}_1, \hat{b}_1, \hat{c}_1} \frac{1}{N-3} \sum_{l=1}^N \left(y_l - \hat{a}_1 - \hat{b}_1 \varphi_1(x_l) - \hat{c}_1 \varphi_2(x_l) \right)^2$$

знаходять оцінки параметрів регресії (3.12):

$$\begin{aligned} \hat{a}_1 &= \frac{1}{N} \sum_{l=1}^N y_l = \bar{y}, \\ \hat{b}_1 &= \frac{\sum_{l=1}^N (x_l - \bar{x}) y_l}{\sum_{l=1}^N (x_l - \bar{x})^2} = \frac{\overline{(x - \bar{x})y}}{\hat{\sigma}_x^2}, \\ \hat{c}_1 &= \frac{\sum_{l=1}^N \varphi_2(x_l) y_l}{\sum_{l=1}^N \varphi_2^2(x_l)} = \frac{\overline{\varphi_2(x)y}}{\overline{\varphi_2^2(x)}}. \end{aligned} \quad (3.13)$$

З аналізу формули (3.13) випливає, що оцінки \hat{a}_1 , \hat{b}_1 повністю збігаються з оцінками для лінійної регресії у вигляді

$$\bar{y}(x) = a + b(x - \bar{x}).$$

Оцінка точності та значущості параметрів \hat{a}_1 , \hat{b}_1 , \hat{c}_1 , як і для лінійної регресії, проводиться шляхом перевірки гіпотез

$$H_0 : a_1 = \hat{a}_1,$$

$$H_0 : b_1 = \hat{b}_1,$$

$$H_0 : c_1 = \hat{c}_1$$

на основі статистик

$$\begin{aligned} t_{a_1} &= \frac{\hat{a}_1 - a_1}{S_{\text{зал}(2)}} \sqrt{N}, \\ t_{b_1} &= \frac{\hat{b}_1 - b_1}{S_{\text{зал}(2)}} \sqrt{\sum_{l=1}^N \varphi_1^2(x_l)} = \frac{(\hat{b}_1 - b_1) \sigma_x}{S_{\text{зал}(2)}} \sqrt{N}, \\ t_{c_1} &= \frac{\hat{c}_1 - c_1}{S_{\text{зал}(2)}} \sqrt{\sum_{l=1}^N \varphi_2^2(x_l)} = \frac{(\hat{c}_1 - c_1)}{S_{\text{зал}(2)}} \sqrt{N \overline{\varphi_2^2(x)}}. \end{aligned} \quad (3.14)$$

Значущість оцінок параметрів перевіряють, вважаючи $a_1 = 0$, $b_1 = 0$, $c_1 = 0$, на основі умови

$$|t_{a_1}| \leq t_{\alpha/2, \nu}, \quad |t_{b_1}| \leq t_{\alpha/2, \nu}, \quad |t_{c_1}| \leq t_{\alpha/2, \nu},$$

де $\nu = N - 3$. Якщо хоча б одна з нерівностей порушується, говорять про «втрату» відповідного члена параболи.

З урахуванням статистичних характеристик (3.14) проводять інтервальне оцінювання відповідних коефіцієнтів регресії:

$$a_{н,в} = \hat{a}_1 \mp t_{\alpha/2, \nu} \frac{S_{3ал(2)}}{\sqrt{N}},$$

$$b_{н,в} = \hat{b}_1 \mp t_{\alpha/2, \nu} \frac{S_{3ал(2)}}{\sigma_x \sqrt{N}},$$

$$c_{н,в} = \hat{c}_1 \mp t_{\alpha/2, \nu} \frac{S_{3ал(2)}}{\sqrt{N \phi_2^2(x)}}.$$

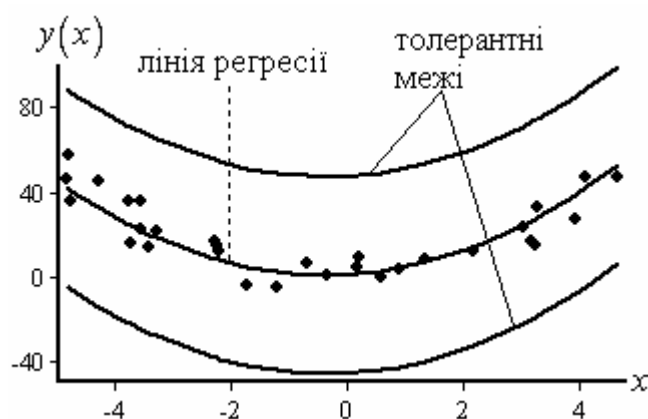


Рис. 3.14. Графічне зображення толерантних меж для параболічної регресії

Відхилення окремих значень від оцінки параболічної регресії (рис. 3.14) оцінюється за аналогією з лінійною регресією шляхом призначення толерантних інтервалів, межі яких визначають зі співвідношень

$$\hat{y}_{\min}(x) = \hat{a}_1 + \hat{b}_1 \phi_1(x) + \hat{c}_1 \phi_2(x) - t_{\alpha/2, \nu} S_{3ал(2)},$$

$$\hat{y}_{\max}(x) = \hat{a}_1 + \hat{b}_1 \phi_1(x) + \hat{c}_1 \phi_2(x) + t_{\alpha/2, \nu} S_{3ал(2)},$$

Відхилення оцінки регресії $\hat{y}(x)$ від теоретичної оцінюють на основі статистичної характеристики

$$t(x) = \frac{\hat{\bar{y}}(x) - \bar{y}(x)}{S_{(\bar{y}|x)}},$$

де (за повною аналогією з лінійною моделлю)

$$\begin{aligned} S_{(\bar{y}|x)} &= \sqrt{\frac{1}{N} \hat{\sigma}_\varepsilon^2 + S_{b_1}^2 \varphi_1^2(x) + S_{c_1}^2 \varphi_2^2(x)} = \\ &= \frac{S_{3ал(2)}}{\sqrt{N}} \sqrt{1 + \frac{\varphi_1^2(x)}{\sigma_x^2} + \frac{\varphi_2^2(x)}{\varphi_2^2(x)}}; \end{aligned}$$

$$\hat{\sigma}_\varepsilon^2 = S_{3ал(2)}^2; \quad S_{b_1}^2 = \frac{S_{3ал(2)}^2}{N\sigma_x^2};$$

$$S_{c_1}^2 = \frac{S_{3ал(2)}^2}{N\varphi_2^2(x)}.$$

Якщо $|t(x)| \leq t_{\alpha/2, \nu}$, де $\nu = N - 3$, то правильна гіпотеза

$$H_0: \bar{y}(x) = \hat{\bar{y}}(x)$$

і проводиться інтервальне оцінювання параболічної регресії (рис. 3.15). Межі довірчого інтервалу визначаються так:

$$\hat{\bar{y}}_{н,в}(x) = \hat{\bar{y}}(x) \mp t_{\alpha/2, \nu} S_{(\bar{y}|x)}.$$

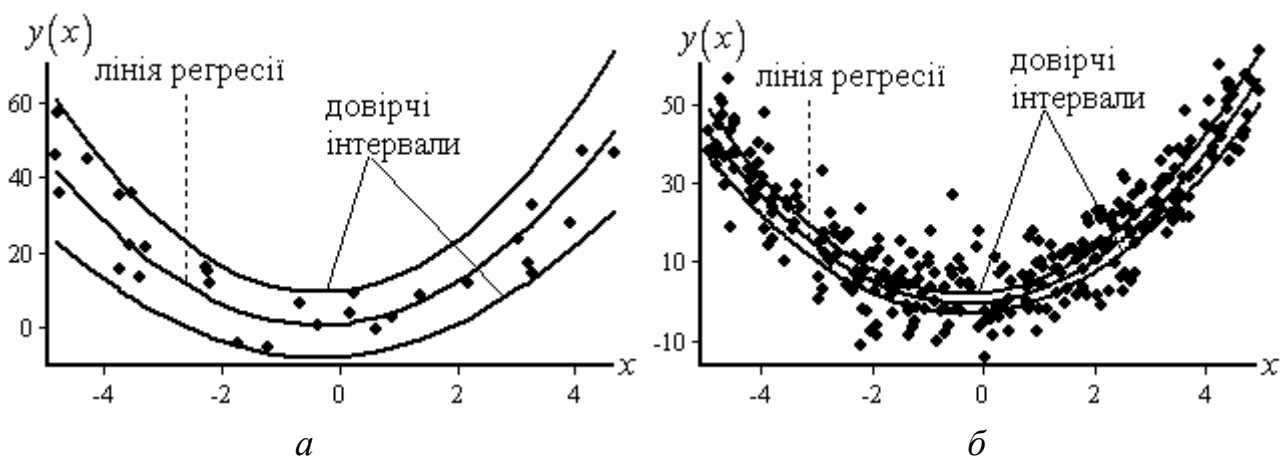


Рис. 3.15. Графічне зображення інтервального оцінювання параболічної регресії:
а – $N = 30$; б – $N = 300$

Порівняльний аналіз наведених меж із довірчими межами лінійної моделі показує, що чим вищий порядок регресійної кривої, тим більше розходження довірчих меж за віддалення від середнього \bar{x} .

Побудова довірчого інтервалу для прогнозу нового спостереження здійснюється з урахуванням величини $S_{(y|x_0)}$ стандартної похибки у при заданому x_0 :

$$S_{(y|x_0)} = \sqrt{\hat{\sigma}_\varepsilon^2 \left(1 + \frac{1}{N}\right) + S_{b_1}^2 \varphi_1^2(x) + S_{c_1}^2 \varphi_2^2(x)} =$$

$$= \frac{S_{\text{Зал}(2)}}{\sqrt{N}} \sqrt{N + 1 + \frac{\varphi_1^2(x)}{\sigma_x^2} + \frac{\varphi_2^2(x)}{\varphi_2^2(x)}}.$$

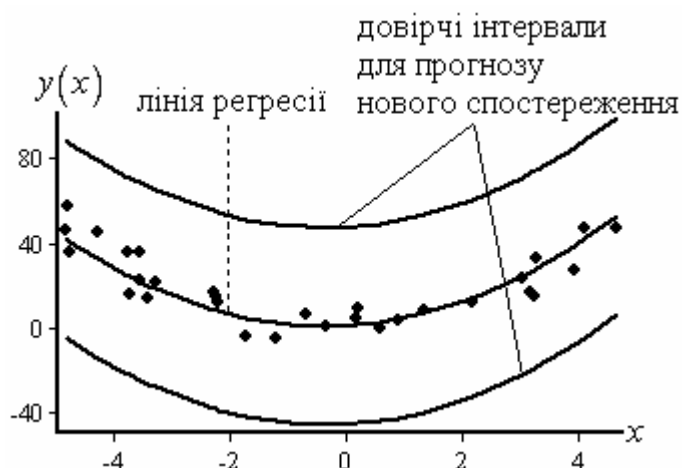


Рис. 3.16. Графічне зображення довірчого інтервалу для прогнозу нового спостереження у випадку параболічної регресії

Відповідний довірчий інтервал для нового спостереження у при заданому x_0 (рис. 3.16) такий:

$$\hat{y}(x_0) - t_{\alpha/2, \nu} S_{(y|x_0)} \leq y \leq \hat{y}(x_0) + t_{\alpha/2, \nu} S_{(y|x_0)}, \quad \nu = N - 3.$$

Нижче для наочності показані толерантні межі, довірчі інтервали для лінії регресії та прогнозного значення (рис. 3.17).

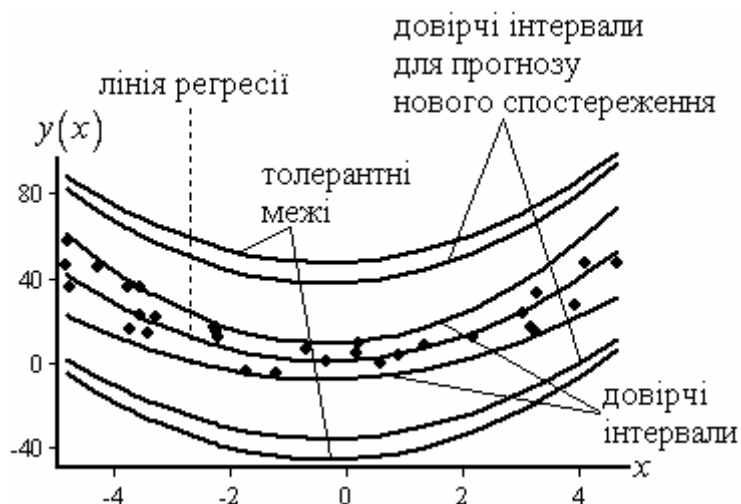


Рис. 3.17. Графічне зображення довірчого оцінювання параболічної регресії

Визначення коефіцієнта детермінації R^2 (частки варіабельності ознаки Y , поясненої за нелінійною моделлю) здійснюється на основі оцінки коефіцієнта кореляційного відношення

$$R^2 = \hat{\rho}_{\eta/\xi}^2 \cdot 100\%.$$

На завершення підрозділу слід зауважити, що перевірка адекватності відтворення параболічної моделі здійснюється аналогічно, як і у випадку з лінійною моделлю.

Контрольні запитання та завдання

1. Перерахувати складові частини первинного статистичного аналізу на основі двовимірного масиву спостережень.
2. Записати функцію щільності двовимірного нормального закону розподілу.
3. Навести статистику χ^2 для оцінки адекватності відтворення двовимірного нормального розподілу.
4. Дати визначення коефіцієнта кореляції та його оцінки.
5. Показати геометричну інтерпретацію оцінки парного коефіцієнта кореляції.
6. Перевірити значущість коефіцієнта кореляції $\hat{r} = 0,12$; $N = 18$; $\alpha = 0,05$.
7. Що таке кореляційне відношення? Які його властивості?
8. Яким чином визначають оцінку рангового коефіцієнта кореляції Спірмена?
9. Як визначають оцінку рангового коефіцієнта кореляції Спірмена у випадку зв'язаних рангів? У який спосіб перевіряють його значущість?
10. Яким співвідношенням зв'язані рангові коефіцієнти Спірмена та Кендалла?
11. В якій задачі має застосування індекс Фехнера?
12. Як перевірити значущість коефіцієнта зв'язку Юла?
13. Для якого з випадків застосовують коефіцієнт сполучень Кендалла, а для якого коефіцієнт сполучень Стюарта?
14. Сформулювати постановку задачі на проведення лінійного регресійного аналізу. Перерахувати початкові умови регресійного аналізу.
15. Описати процедуру відтворення лінійної регресії за МНК.
16. Навести дисперсії оцінок параметрів лінійної моделі регресії.
17. Яка оцінка дозволяє вказати стандартну похибку регресійної оцінки?
18. У чому полягає різниця в побудові довірчих інтервалів для лінії регресії та прогнозу нового спостереження?

19. На основі якої статистики здійснюється перевірка адекватності відтворення моделі регресії?
20. Визначити довірчий інтервал для параболічної регресії.
21. Як перевіряється значущість вільного члена параболічної моделі регресії?
22. Перевірити гіпотезу про збіг двох регресійних прямих:
 $\bar{y}_1(x) = 7,2 + 0,52x$; $N = 65$; $S_{1, \text{зал}} = 11$; $S_{x_1} = 4$; $\bar{x}_1 = 4,8$; $\bar{y}_1 = 12$;
 $\bar{y}_2(x) = 7,7 + 0,38x$; $N = 100$; $S_{2, \text{зал}} = 9$; $S_{x_2} = 5$; $\bar{x}_2 = 5$; $\bar{y}_2 = 13$.

Розділ 4. ОБРОБКА Й АНАЛІЗ БАГАТОВИМІРНИХ ДАНИХ

Багатовимірні дані на додаток до інформації про кожну окрему змінну (як набір одновимірних даних) та взаємозв'язки між парами змінних (як під час аналізу двовимірних даних) дозволяють:

- 1) встановлення наявності стохастичного зв'язку між усіма змінними;
- 2) за наявності зв'язку – відтворення багатовимірної регресії;
- 3) вивчення факторів, що породжують даний зв'язок.

Розглянемо основні етапи статистичного аналізу багатовимірних даних, які передбачають проведення первинного аналізу, знаходження взаємозв'язків між змінними на основі коефіцієнтів кореляції, відтворення форми зв'язку методами регресійного аналізу та здійснення лінійних перетворень над даними методами факторного аналізу.

4.1. Первинний аналіз

Будемо припускати, що n -вимірні дані є реалізаціями деяких функцій $\bar{\xi} = (\xi_1(\omega), \dots, \xi_n(\omega))$ – випадкових величин, що відбивають простір результатів спостережень Ω у простір дійсних чисел R_n , $n \geq 1$ (мова йде про неперервні випадкові величини). Отже, кожне спостереження об'єкта дослідження є точкою n -вимірного дійсного простору.

Головна характеристика n -вимірної випадкової величини є **функція розподілу ймовірностей**:

$$\begin{aligned} F(x_1, \dots, x_n) &= P\{\omega : -\infty < \xi_1(\omega) < x_1, \dots, -\infty < \xi_n(\omega) < x_n\} = \\ &= P\left\{\bigcap_{k=1}^n \{-\infty < \xi_k(\omega) < x_k\}\right\}. \end{aligned} \quad (4.1)$$

Якщо існує

$$\begin{aligned} \lim_{x_k \rightarrow x_k^0 + 0, k=1, n} F(x_1, \dots, x_k, \dots, x_n) &=, \\ &= P\{\xi_1 = x_1^0, \dots, \xi_k = x_k^0, \dots, \xi_n = x_n^0\} + F(x_1^0, \dots, x_k^0, \dots, x_n^0), \end{aligned}$$

то функція $F(x_1, \dots, x_n)$ неперервна праворуч, і при цьому існує функція щільності розподілу ймовірностей

$$f(x_1, \dots, x_n) = \frac{\partial^n F(x_1, \dots, x_n)}{\partial x_1 \dots \partial x_n},$$

отже

$$F(x_1, \dots, x_n) = \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_n} f(\tau_1, \dots, \tau_n) d\tau_1 \dots d\tau_n.$$

Якщо ξ_1, \dots, ξ_n незалежні випадкові величини

$$\left\{ \omega : \bigcap_{k=1}^n \{ -\infty < \xi_k(\omega) < x_k \} \right\} = \bigcap_{k=1}^n \{ \omega_k : -\infty < \xi_k(\omega) < x_k \},$$

то

$$F(x_1, \dots, x_n) = \prod_{k=1}^n F(x_k)$$

або

(4.2)

$$f(x_1, \dots, x_n) = \prod_{k=1}^n f(x_k).$$

Із властивості (4.2) випливає, що процедура знаходження оцінки функції розподілу ймовірностей $\hat{F}(x_1, \dots, x_n)$ (оцінки функції щільності $\hat{f}(x_1, \dots, x_n)$) багатовимірної випадкової величини, на разі незалежності одновимірних складових вектора $\vec{\xi}$ може бути одержана на підставі оцінок маргінальних розподілів: $\hat{F}(x_k)$ (або $\hat{f}(x_k)$), $k = \overline{1, n}$.

Розглянемо забезпечення проведення ймовірнісної оцінки за масивом

$$\Omega_{n,N} = \{ (x_{1,l}, \dots, x_{n,l}); l = \overline{1, N} \}$$

реалізацій n -вимірної випадкової величини $\vec{\xi} = (\xi_1(\omega), \dots, \xi_n(\omega))$ з функцією розподілу ймовірностей (4.1). Зважаючи на можливість проведення ймовірнісної оцінки відповідних масивів реалізацій

$$\Omega_{1,N}^{(k)} = \{ x_{k,l}; l = \overline{1, N} \}, \quad k = \overline{1, n},$$

шляхом уведення низки рівномірних розбиттів за осями спостережень

$$x_k : \Delta_{h_{x_k}}, \quad h_{x_k} > 0,$$

можемо розглядати й можливість ймовірнісної обробки n -вимірного **варіацій-**

ного ряду, розбитого на класи, згідно рівномірного розбиття $\Delta_{h_{x_1}, \dots, h_{x_n}}$:

$$\left\{ (x_{1,i_1}, \dots, x_{n,i_n}), n_{i_1, \dots, i_n}, p_{i_1, \dots, i_n}; i_k = \overline{1, M_k}, k = \overline{1, n} \right\}, \quad (4.3)$$

де $(x_{1,i_1}, \dots, x_{n,i_n})$ – варіанта, яка визначає центральну (або мінімальну) точку (i_1, \dots, i_n) -го елементу розбиття $\Delta_{h_{x_1}, \dots, h_{x_n}}$; M_k – кількість елементів (класів) розбиття $\Delta_{h_{x_1}, \dots, h_{x_n}}$ за напрямками x_k , $k = \overline{1, n}$, які визначаються для реалізацій відповідних одновимірних випадкових величин згідно відповідних гістограмних оцінок; n_{i_1, \dots, i_n} – частота (кількість потраплянь точок з масиву $\Omega_{n,N}$ в (i_1, \dots, i_n) -й елемент розбиття $\Delta_{h_{x_1}, \dots, h_{x_n}}$); p_{i_1, \dots, i_n} – **відносна частота** варіанти n -вимірного варіаційного ряду:

$$p_{i_1, \dots, i_n} = \frac{n_{i_1, \dots, i_n}}{N}, \quad \sum_{i_1=1}^{M_1} \dots \sum_{i_n=1}^{M_n} p_{i_1, \dots, i_n} = 1, \quad (4.4)$$

причому

$$p_{i_1, \dots, i_n} = P \left\{ \omega : x_{1,i_1} - 0,5h_{x_1} \leq \xi_1(\omega) < x_{1,i_1} + 0,5h_{x_1}, \dots, \right.$$

$$\left. x_{n,i_n} - 0,5h_{x_n} \leq \xi_n(\omega) < x_{n,i_n} + 0,5h_{x_n} \right\} = \bar{f}_{i_1, \dots, i_n} h_{x_1} \cdot \dots \cdot h_{x_n},$$

де $\bar{f}_{i_1, \dots, i_n}$ – усереднене значення функції щільності розподілу ймовірностей $f(x_1, \dots, x_n)$ величини $\bar{\xi}$ на (i_1, \dots, i_n) -му елементі розбиття $\Delta_{h_{x_1}, \dots, h_{x_n}}$:

$$\bar{f}_{i_1, \dots, i_n} = \frac{1}{h_{x_1} \cdot \dots \cdot h_{x_n}} \int_{x_{1,i_1}-0,5h_{x_1}}^{x_{1,i_1}+0,5h_{x_1}} \dots \int_{x_{n,i_n}-0,5h_{x_n}}^{x_{n,i_n}+0,5h_{x_n}} f(x_1, \dots, x_n) dx_1 \dots dx_n,$$

Емпірична функція розподілу $F_{n,N_{i_1, \dots, i_n}}$ за варіаційним рядом (4.3) визначається із співвідношення

$$F_{n,N_{i_1, \dots, i_n}} = \sum_{j_1=1}^{i_1} \dots \sum_{j_n=1}^{i_n} p_{j_1, \dots, j_n}, \quad i_k = \overline{1, M_k}, k = \overline{1, n}.$$

Відмітимо, що для незалежних випадкових величин $\xi_1(\omega), \dots, \xi_n(\omega)$

$$\begin{aligned} F_{n,N_{i_1, \dots, i_n}} &= \sum_{j_1=1}^{i_1} \dots \sum_{j_n=1}^{i_n} p_{j_1, \dots, j_n} = \\ &= \sum_{j_1=1}^{i_1} p_{j_1} \cdot \dots \cdot \sum_{j_n=1}^{i_n} p_{j_n} = F_{1,N_{i_1}} \cdot \dots \cdot F_{1,N_{i_n}}, \quad i_k = \overline{1, M_k}, k = \overline{1, n}. \end{aligned}$$

Визначення аномальних результатів спостережень у вибірці $\Omega_{n,N}$ відбувається на підставі перевірки умови

$$p_{i_1, \dots, i_n} \leq \gamma, \quad l = \overline{1, N}, \quad (4.5)$$

де

p_{i_1, \dots, i_n} , $i_k = \overline{1, M_k}$, $k = \overline{1, n}$; – відносна частота варіанти розбитого на класи ряду;

γ – похибка у прийнятті рішення про малоймовірність l -ої реалізації.

Отже, коли для довільного індексу l варіанти масиву даних $\Omega_{n,N}$, справедливо виконання нерівності (4.5) (зважаючи, що варіанта належить відповідному класу), спостереження має бути вилучено з процесу подальшої обробки.

Підготовка даних для аналізу на підставі ймовірнісних оцінок може як містити, так і ні наступні заходи: стандартизація (нормування) спостережень; при необхідності, зменшення асиметрії в наборах даних; координатні перетворення, що приводять до незалежності окремих складових реалізацій дво- та багатовимірних випадкових величин.

Найпростішими **точковими оцінками** реалізацій $\Omega_{n,N}$ випадкової величини $\vec{\xi}$ є оцінка вектора математичного сподівання

$$\hat{E}\{\vec{\xi}\} = (\bar{x}_1, \dots, \bar{x}_n),$$

де

$$\bar{x}_k = \frac{1}{N} \sum_{l=1}^N x_{k,l}, \quad k = \overline{1, n},$$

який характеризує геометричний центр тяжіння однорідної сукупності спостережень, та оцінка дисперсійно-коваріаційної матриці (ДК-матриці)

$$\begin{aligned} \hat{DC}\{\vec{\xi}\} &= \begin{pmatrix} \hat{D}\{\xi_1\} & \text{cov}\{\xi_1, \xi_2\} & \dots & \text{cov}\{\xi_1, \xi_n\} \\ \text{cov}\{\xi_2, \xi_1\} & \hat{D}\{\xi_2\} & \dots & \text{cov}\{\xi_2, \xi_n\} \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}\{\xi_n, \xi_1\} & \text{cov}\{\xi_n, \xi_2\} & \dots & \hat{D}\{\xi_n\} \end{pmatrix} = \\ &= \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_1 \hat{\sigma}_2 \hat{r}_{1,2} & \dots & \hat{\sigma}_1 \hat{\sigma}_n \hat{r}_{1,n} \\ \hat{\sigma}_2 \hat{\sigma}_1 \hat{r}_{2,1} & \hat{\sigma}_2^2 & \dots & \hat{\sigma}_2 \hat{\sigma}_n \hat{r}_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\sigma}_n \hat{\sigma}_1 \hat{r}_{n,1} & \hat{\sigma}_n \hat{\sigma}_2 \hat{r}_{n,2} & \dots & \hat{\sigma}_n^2 \end{pmatrix}, \end{aligned}$$

де

$\hat{\sigma}_k^2$ – незсунена оцінка дисперсії k -ої ознаки

$$\sigma_k^2 = \frac{1}{N-1} \sum_{l=1}^N (x_{k,l} - \bar{x}_k)^2, \quad k = \overline{1, n};$$

$\hat{r}_{k,v}$ – оцінка парного коефіцієнта кореляції поміж k -ою та v -ою ознаками.

Якщо ознаки об'єкту спостереження розподілені за розподілами, длизькими до нормального та вимірюються в різних одиницях масштабу, проводять стандартизацію даних у вигляді

$$x_{k,l}^* = \frac{x_{k,l} - \bar{x}_k}{\hat{\sigma}_k}, \quad k = \overline{1, n}, \quad l = \overline{1, N},$$

причому, отримані шляхом такої операції дані $\Omega_{n,N}^* = \left\{ (x_{1,l}^*, \dots, x_{n,l}^*); l = \overline{1, N} \right\}$ мають співставний рівень виміру та виявляється, що

$$\bar{x}_k^* = 0, \quad \hat{\sigma}_k^* = 1, \quad k = \overline{1, n},$$

а замість дисперсійно-коваріаційної, оцінці підлягає кореляційна матриця

$$\hat{R}\{\bar{\xi}\} = \begin{pmatrix} 1 & \hat{r}_{1,2} & \dots & \hat{r}_{1,n} \\ \hat{r}_{2,1} & 1 & \dots & \hat{r}_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{r}_{n,1} & \hat{r}_{n,2} & \dots & 1 \end{pmatrix}.$$

Зауваження 4.1. Оцінка матриці кореляції може проводитись і для нестандартизованих величин, проте в цьому разі така оцінка є визначальною лише для виявлення наявності лінійного зв'язку між окремими складовими вектору $\bar{\xi}$.

Відтворення багатовимірних параметричних розподілів є задачею нетривіальною, як за методами визначення оцінок параметрів, так і при перевірці адекватності результатів відтворення. Тому, за всяк час, при опрацюванні багатовимірних масивів обмежуються перевіркою адекватності **багатовимірного нормального розподілу**, яку здійснюють шляхом аналізу одновимірних складових реалізацій випадкового вектору $\bar{\xi}$.

Масив $\Omega_{n,N} = \{X_l; l = \overline{1, N}\} = \{(x_{1,l}, \dots, x_{n,l}); l = \overline{1, N}\}$ змодельованих нормально розподілених випадкових векторів із параметрами $E\{\bar{\xi}\} = (m_1, \dots, m_n)$, $DC\{\bar{\xi}\} = \|\text{cov}_{i,j}; i, j = \overline{1, n}\|$ можна одержати за формулами:

$$X_l = E\{\bar{\xi}\} + AU_l, \quad l = \overline{1, N}$$

де $U_l = (u_{1,l}, \dots, u_{n,l})$, $u_{k,l}$ – реалізації одновимірних стандартизованих нормально розподілених випадкових величин $N(u; 0, 1)$; $A = \{a_{i,j}; i, j = \overline{1, n}\}$ – матриця $n \times n$,

елементи якої обчислюється з умови $AA^T = DC\{\bar{\xi}\}$ у такий спосіб:

$$a_{i,j} = \begin{cases} \sqrt{\text{cov}_{i,i} - \sum_{w=1}^{i-1} a_{i,w}^2}, & i = j, \\ \frac{\text{cov}_{i,j} - \sum_{w=1}^{j-1} a_{i,w}a_{j,w}}{a_{j,j}}, & i > j, \\ 0, & i < j. \end{cases}$$

Зауваження 4.2. Наведений спосіб моделювання придатний лише у разі, коли усі елементи матриці $DC\{\bar{\xi}\}$ невід'ємні.

4.2. Перевірка гіпотез про збіг параметрів багатовимірних даних

Серед задач, що безпосередньо базуються на оцінках вектора математичного сподівання та дисперсійно-коваріаційної матриці, є **перевірка збігу векторів середніх і дисперсійно-коваріаційних матриць нормально розподілених випадкових величин**, заснована на реалізації однойменного критерію.

Для перевірки головної гіпотези H_0 про **однорідність двох та більше нормально розподілених сукупностей** необхідно послідовно перевірити гіпотези про збіг середніх і коваріаційних матриць.

Нехай $\bar{\xi} = \{\xi_1, \dots, \xi_n\}$ та $\bar{\eta} = \{\eta_1, \dots, \eta_n\}$ – дві незалежні n -вимірні нормально розподілені випадкові величини. Над кожною випадковою величиною проведено відповідно по N_1 та N_2 спостережень

$$\Omega_{n,N_1} = \{(x_{1,l}, \dots, x_{n,l}); l = \overline{1, N_1}\},$$

$$\Omega_{n,N_2} = \{(y_{1,l}, \dots, y_{n,l}); l = \overline{1, N_2}\},$$

обчислено оцінки векторів математичного сподівання

$$\hat{E}\{\bar{\xi}\} = \{\bar{x}_1, \dots, \bar{x}_n\}, \quad \hat{E}\{\bar{\eta}\} = \{\bar{y}_1, \dots, \bar{y}_n\},$$

та оцінки ДК-матриць: $\hat{DC}\{\bar{\xi}\}, \hat{DC}\{\bar{\eta}\}$.

Наступна процедура передбачає перевірку низки статистичних гіпотез.

Перевірка гіпотези про **рівність двох багатовимірних середніх у разі рівних дисперсійно-коваріаційних матриць**. За вибірковими даними треба перевірити гіпотезу

$$H_0 : E\{\bar{\xi}\} = E\{\bar{\eta}\}$$

при альтернативі

$$H_1 : E\{\bar{\xi}\} \neq E\{\bar{\eta}\}.$$

Розглянемо критерій для перевірки гіпотези при умові, що

$$\hat{DC}\{\bar{\xi}\} = \hat{DC}\{\bar{\eta}\} = \hat{DC},$$

де \hat{DC} – деяка оцінка ДК-матриці. Для перевірки гіпотези H_0 застосовують статистику

$$V = -\left(N_1 + N_2 - 2 - \frac{n}{2}\right) \ln \frac{|S_1|}{|S_0|},$$

яка розподіленої за χ^2 -розподілом з n степенями вільності. Отже, гіпотеза H_0 правильна, якщо $V \leq \chi_{\alpha, n}^2$, де α – критичний рівень значущості.

Елементи матриць

$$S_0 = \|s_{i,j}^0; i, j = \overline{1, n}\|$$

і

$$S_1 = \|s_{i,j}^1; i, j = \overline{1, n}\|$$

обчислюються за формулами:

$$\begin{aligned} s_{i,j}^0 &= \frac{1}{N_1 + N_2 - 2} \left(\sum_{l=1}^{N_1} x_{i,l} x_{j,l} + \sum_{l=1}^{N_2} y_{i,l} y_{j,l} - \right. \\ &\quad \left. - \frac{1}{N_1 + N_2} \left(\sum_{l=1}^{N_1} x_{i,l} + \sum_{l=1}^{N_2} y_{i,l} \right) \left(\sum_{l=1}^{N_1} x_{j,l} + \sum_{l=1}^{N_2} y_{j,l} \right) \right), \\ s_{i,j}^1 &= \frac{1}{N_1 + N_2 - 2} \left(\sum_{l=1}^{N_1} x_{i,l} x_{j,l} + \sum_{l=1}^{N_2} y_{i,l} y_{j,l} - \right. \\ &\quad \left. - \frac{1}{N_1} \left(\sum_{l=1}^{N_1} x_{i,l} \right) \left(\sum_{l=1}^{N_1} x_{j,l} \right) - \frac{1}{N_2} \left(\sum_{l=1}^{N_2} y_{i,l} \right) \left(\sum_{l=1}^{N_2} y_{j,l} \right) \right). \end{aligned}$$

Слід відзначити, що елементи $s_{i,j}^0$ вибіркової ДК-матриці S_0 обчислюються так, неначе обидві вибірки виявляються об'єднаними в одну, тим самим враховується припущення про збіг середніх. Елементи ж $s_{i,j}^1$ матриці S_1 , навпаки, обчислюються так, що можливий збіг між середніми виключається, чим

досягається відповідність цих оцінок відносно гіпотези H_1 .

Перевірка гіпотези про **збіг k n -вимірних середніх при розбіжності дисперсійно-коваріаційних матриць**. Нехай $\bar{\xi}_1, \bar{\xi}_2, \dots, \bar{\xi}_k$ – набір n -вимірних незалежних нормально розподілених випадкових величин, кожній з яких відповідає n -вимірний вектор математичного сподівання $E\{\bar{\xi}_d\}$, $d = \overline{1, k}$ та ДК-матриці $DC\{\bar{\xi}_d\}$, $d = \overline{1, k}$, що відрізняються між собою.

Над кожною з величин проведено N_d , $d = \overline{1, k}$ спостережень $\Omega_{n, N_d}^{(d)} = \left\{ \left(x_{1,l}^{(d)}, \dots, x_{n,l}^{(d)} \right); l = \overline{1, N_d} \right\}$ та обчислені оцінки векторів математичного сподівання $\hat{E}\{\bar{\xi}_d\} = \bar{x}^{(d)} = \left\{ \bar{x}_1^{(d)}, \dots, \bar{x}_n^{(d)} \right\}$ та оцінки ДК-матриць $\hat{DC}\{\bar{\xi}_d\}$. Треба перевірити гіпотезу

$$H_0 : E\{\bar{\xi}_1\} = E\{\bar{\xi}_2\} = \dots = E\{\bar{\xi}_k\}$$

при альтернативі

$$H_1 : E\{\bar{\xi}_i\} \neq E\{\bar{\xi}_j\}$$

хоча б для однієї двійки (i, j) , $i, j = \overline{1, k}$.

Критерій перевіряється за статистичною характеристикою

$$V = \sum_{d=1}^k N_d \left(\bar{x}^{(d)} - \bar{x} \right) S_d^{-1} \left(\bar{x}^{(d)} - \bar{x} \right)^T,$$

де

$$\bar{x}^{(d)} = \frac{1}{N_d} \sum_{l=1}^{N_d} X_l^{(d)}, \quad X_l^{(d)} = \left\{ x_{1,l}^{(d)}, \dots, x_{n,l}^{(d)} \right\};$$

$$S_d = \frac{1}{N_d - 1} \sum_{l=1}^{N_d} \left(X_l^{(d)} - \bar{x}^{(d)} \right) \left(X_l^{(d)} - \bar{x}^{(d)} \right)^T.$$

Узагальнене вибіркове n -вимірне середнє \bar{x} обчислюється за формулою

$$\bar{x} = \left(\left(\sum_{d=1}^k N_d S_d^{-1} \right)^{-1} \left(\sum_{d=1}^k N_d S_d^{-1} \bar{x}^{(d)T} \right) \right)^T.$$

Якщо гіпотеза, яку перевіряють, має місце, то статистика V буде значенням випадкової величини, розподіленої асимптотично, за χ^2 з $\nu = n(k-1)$ степенями вільності. У результаті головну гіпотезу буде прийнято, якщо $V \leq \chi_{\alpha, \nu}^2$.

Перевірка гіпотези про **збіг дисперсійно-коваріаційних матриць**. Нехай $\bar{\xi}_1, \bar{\xi}_2, \dots, \bar{\xi}_k$ – набір n -вимірних незалежних нормально розподілених випадко-

вих величин, кожній з яких відповідає n -вимірний вектор математичного сподівання $E\{\bar{\xi}_d\}$, $d = \overline{1, k}$ та ДК-матриці $DC\{\bar{\xi}_d\}$, $d = \overline{1, k}$. За результатами спостережень необхідно перевірити гіпотезу

$$H_0 : DC\{\bar{\xi}_1\} = DC\{\bar{\xi}_2\} = \dots = DC\{\bar{\xi}_k\}$$

при альтернативі $H_1 : DC\{\bar{\xi}_i\} \neq DC\{\bar{\xi}_j\}$ хоча б для однієї двійки (i, j) , $i, j = \overline{1, k}$.

Критерій для перевірки гіпотези H_0 базується на статистиці

$$V = \sum_{d=1}^k \frac{N_d - 1}{2} \ln \frac{|S|}{|S_d|},$$

де

$$S_d = \frac{1}{N_d - 1} \sum_{l=1}^{N_d} \left(X_l^{(d)} - \bar{x}^{(d)} \right)^T \left(X_l^{(d)} - \bar{x}^{(d)} \right),$$

$$S = \frac{1}{N - k} \sum_{d=1}^k (N_d - 1) S_d,$$

$$N = \sum_{d=1}^k N_d.$$

З останнього випливає, що S_d – оцінка ДК-матриці у вибірці з номером d , а S – узагальнена вибіркова ДК-матриця, яку обчислено у припущенні, що головна гіпотеза H_0 має місце.

Якщо гіпотеза H_0 правильна, то статистика V буде значенням випадкової величини, розподіленої асимптотично за χ^2 -розподілом з $v = n(n+1)(k-1)/2$ степенями вільності. Тоді, якщо $V \leq \chi_{\alpha, v}^2$, то головну гіпотезу приймаємо.

Слід зауважити, що, окрім наведених параметричних критеріїв, для багатовимірних величин застосовують ще й непараметричні, наприклад, наступний.

H -критерій є узагальненням для $k > 2$ критерію Вілкоксона. Будують загальний варіаційний ряд, який містить $N_1 + N_2 + \dots + N_k = N$ елементів, де N_i – кількість спостережень у i -й вибірці, і обчислюють статистику \bar{W}_i для кожної вибірки. Зазначимо, що

$$\bar{W}_i = \frac{W_i}{N_i}, \quad E\{\bar{W}_i\} = \frac{N+1}{2},$$

$$D\{\bar{W}_i\} = \frac{(N+1)(N-N_i)}{12N_i}.$$

Якщо $N_i \rightarrow \infty$, то величина

$$U = \frac{\bar{W}_i - E\{\bar{W}_i\}}{\sqrt{D\{\bar{W}_i\}}}$$

має стандартний нормальний розподіл. Цей факт був використаний Крускалом та Уолісом для побудови критерію зі статистичною характеристикою

$$H = \sum_{i=1}^k \frac{(\bar{W}_i - (N+1)/2)^2}{(N+1)(N-N_i)/(12N_i)} \left(1 - \frac{N_i}{N}\right).$$

Крускал та Уоліс довели, що асимптотично статистика H має χ^2 -розподіл з $\nu = k - 1$ степенями вільності, де k – кількість вибірок.

4.3. Часткові та множинні коефіцієнти кореляції

Під час проведення кореляційного аналізу за багатовимірними даними припускають, що задано вибірку $\Omega_{n,N} = \{x_{k,l}; k = \overline{1,n}, l = \overline{1,N}\}$, отже, задано N реалізацій масиву n змінних (ознак об'єкта спостереження)

$$X = \{x_1, \dots, x_n\}.$$

Припускається, що $\Omega_{n,N}$ є результатом реалізації n -вимірної нормально розподіленої випадкової величини

$$\vec{\xi} = (\xi_1(\omega), \dots, \xi_n(\omega)).$$

Можливі три типи зв'язку між складовими даного випадкового вектора:

1) лінійний зв'язок між двома змінними, який вимірюється за допомогою **парного коефіцієнта кореляції** (див. підрозд. 3.2.1);

2) лінійний зв'язок між двома змінними після усунення частини зв'язку, що обумовлена взаємозв'язком цих змінних з іншими, – вимірюється за допомогою **часткового коефіцієнта кореляції**;

3) лінійний зв'язок однієї змінної одночасно з усіма іншими, який визначається через **множинний коефіцієнт кореляції**.

Наведемо процедуру оцінювання часткових та множинних коефіцієнтів кореляції за вхідною вибіркою. Нехай i та j – індекси двох випадкові величини з $\vec{\xi}$, c – деяка непуста підмножина інших $n - 2$ випадкових величин.

Оцінка часткового коефіцієнту кореляції $\hat{r}_{i,j,c}$ визначає міру лінійної залежності поміж i -ої та j -ої ознаки, без урахування залежності цих двох змінних з елементами набору c (іншими словами – коли величини змінних із c фік-

совані). Значення $\hat{r}_{i,j-c}$ завжди містяться поміж -1 та 1 , якщо

$$\hat{r}_{i,j-c} = 0,$$

то останнє є слідством незалежності ознак i та j , коли величини змінних в c фіксовані.

Зазначають, що квадрат оцінки часткового коефіцієнта кореляції $\hat{r}_{i,j-c}^2$ є часткою дисперсії змінної i , поясненої j після вилучення ефекту змінних з c .

Обчислення оцінок часткових коефіцієнтів кореляції виконують за рекурентними співвідношеннями. Якщо i , j і d – три різні змінні з множини X , то усі оцінки часткових коефіцієнтів кореляції першого порядку визначають за виразом

$$\hat{r}_{i,j-d} = \frac{\hat{r}_{i,j} - \hat{r}_{i,d}\hat{r}_{j,d}}{\sqrt{(1-\hat{r}_{i,d}^2)(1-\hat{r}_{j,d}^2)}},$$

де $\hat{r}_{i,j}$, $\hat{r}_{i,d}$ та $\hat{r}_{j,d}$ – значення оцінок парної кореляції для відповідних змінних.

Надалі послідовно застосовуючи рекурентну формулу

$$\hat{r}_{i,j-cd} = \frac{\hat{r}_{i,j-c} - \hat{r}_{i,d-c}\hat{r}_{j,d-c}}{\sqrt{(1-\hat{r}_{i,d-c}^2)(1-\hat{r}_{j,d-c}^2)}},$$

де c – будь-яка підмножина інших змінних,

можна одержувати оцінки часткових коефіцієнтів будь-якого порядку.

Приклад 4.1. Нехай на основі спостережень над 4-вимірними даними одержана оцінка матриці кореляції

$$\hat{R}\{\bar{\xi}\} = \begin{pmatrix} 1 & 0,5 & 0,2 & 0,7 \\ 0,5 & 1 & 0,6 & 0,9 \\ 0,2 & 0,6 & 1 & 0,3 \\ 0,7 & 0,9 & 0,3 & 1 \end{pmatrix}.$$

Тоді оцінка часткового коефіцієнта кореляції $\hat{r}_{1,3\{4,2\}}$ дорівнює

$$\hat{r}_{1,3\{4,2\}} = \frac{\hat{r}_{1,3} - \hat{r}_{1,4,2}\hat{r}_{3,4,2}}{\sqrt{(1-\hat{r}_{1,4,2}^2)(1-\hat{r}_{3,4,2}^2)}},$$

де

$$\hat{r}_{1,3} = \frac{\hat{r}_{1,3} - \hat{r}_{1,2}\hat{r}_{3,2}}{\sqrt{(1-\hat{r}_{1,2}^2)(1-\hat{r}_{3,2}^2)}} = \frac{0,2 - 0,5 \cdot 0,6}{\sqrt{(1-0,5^2)(1-0,6^2)}} \approx -0,144,$$

$$\hat{r}_{1,4,2} = \frac{\hat{r}_{1,4} - \hat{r}_{1,2}\hat{r}_{4,2}}{\sqrt{(1-\hat{r}_{1,2}^2)(1-\hat{r}_{4,2}^2)}} = \frac{0,7 - 0,5 \cdot 0,9}{\sqrt{(1-0,5^2)(1-0,9^2)}} \approx 0,662,$$

$$\hat{r}_{3,4,2} = \frac{\hat{r}_{3,4} - \hat{r}_{3,2}\hat{r}_{4,2}}{\sqrt{(1-\hat{r}_{3,2}^2)(1-\hat{r}_{4,2}^2)}} = \frac{0,3 - 0,6 \cdot 0,9}{\sqrt{(1-0,6^2)(1-0,9^2)}} \approx -0,688,$$

отже

$$\hat{r}_{1,3,\{4,2\}} \approx 0,573.$$

Для перевірки гіпотези про **значущість** часткового коефіцієнта кореляції

$$H_0 : r_{i,j,c} = 0$$

використовується статистика

$$t = \frac{\hat{r}_{i,j,c} \sqrt{N-w-2}}{\sqrt{1-\hat{r}_{i,j,c}^2}},$$

де w – кількість змінних у наборі c ,

яка має t -розподіл Стюдента з $N-w-2$ степенями вільності.

Відповідний $100 \cdot (1-\alpha)\%$ -й довірчий інтервал для $r_{i,j,c}$ можна отримати, використовуючи перетворення Фішера:

$$\frac{1}{2} \ln \frac{1+\hat{r}_{i,j,c}}{1-\hat{r}_{i,j,c}} - \frac{u_{\alpha/2}}{N-w-3} < \frac{1}{2} \ln \frac{1+r_{i,j,c}}{1-r_{i,j,c}} < \frac{1}{2} \ln \frac{1+\hat{r}_{i,j,c}}{1-\hat{r}_{i,j,c}} + \frac{u_{\alpha/2}}{N-w-3},$$

де

$u_{\alpha/2}$ – квантиль стандартного нормального розподілу $N(u; 0, 1)$.

Позначаючи

$$v_1 = \frac{1}{2} \ln \frac{1+\hat{r}_{i,j,c}}{1-\hat{r}_{i,j,c}} - \frac{u_{\alpha/2}}{N-w-3},$$

$$v_2 = \frac{1}{2} \ln \frac{1+\hat{r}_{i,j,c}}{1-\hat{r}_{i,j,c}} + \frac{u_{\alpha/2}}{N-w-3},$$

одержують

$$\frac{\exp(2v_1)-1}{\exp(2v_1)+1} < r_{i,j,c} < \frac{\exp(2v_2)-1}{\exp(2v_2)+1}.$$

Оцінка множинного коефіцієнта кореляції $\hat{r}_{x_k, x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n}$ є мірою лі-

нійної залежності поміж ознакою x_k та набором $x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n$, причому

$$0 \leq \hat{r}_{x_k \cdot x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n} \leq 1.$$

Якщо

$$\hat{r}_{x_k \cdot x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n} = 0,$$

то говорять про відсутність залежності x_k від інших змінних з множини X . На разі, коли

$$\hat{r}_{x_k \cdot x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n} = 1,$$

має місце лінійна залежність, при якій змінна x_k визначається лінійною комбінацією змінних $x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n$:

$$x_k = \beta_0 + \beta_1 x_1 + \dots + \beta_{k-1} x_{k-1} + \beta_{k+1} x_{k+1} + \dots + \beta_n x_n.$$

Квадрат оцінки множинної кореляції $\hat{r}_{x_k \cdot x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n}^2$ визначає частку дисперсії x_k , яка пояснюється лінійною регресією x_k за $x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n$.

Оцінку коефіцієнта множинної кореляції зручно обчислювати за формулою

$$\hat{r}_{x_k \cdot x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n} = \sqrt{\Delta^* / \Delta},$$

де

$$\Delta = (-1)^n \begin{vmatrix} 1 & \hat{r}_{1,2} & \cdots & \hat{r}_{1,k-1} & \hat{r}_{1,k+1} & \cdots & \hat{r}_{1,n} \\ \hat{r}_{2,1} & 1 & \cdots & \hat{r}_{2,k-1} & \hat{r}_{2,k+1} & \cdots & \hat{r}_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \hat{r}_{k-1,1} & \hat{r}_{k-1,2} & \cdots & 1 & \hat{r}_{k-1,k+1} & \cdots & \hat{r}_{k-1,n} \\ \hat{r}_{k+1,1} & \hat{r}_{k+1,2} & \cdots & \hat{r}_{k+1,k-1} & 1 & \cdots & \hat{r}_{k+1,n} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ \hat{r}_{n,1} & \hat{r}_{n,2} & \cdots & \hat{r}_{n,k-1} & \hat{r}_{n,k+1} & \cdots & 1 \end{vmatrix};$$

$$\Delta^* = (-1)^n \begin{vmatrix} \hat{r}_{0,1} & \hat{r}_{0,2} & \cdots & \hat{r}_{0,k-1} & \hat{r}_{0,k+1} & \cdots & \hat{r}_{0,n} & 0 \\ 1 & \hat{r}_{1,2} & \cdots & \hat{r}_{1,k-1} & \hat{r}_{1,k+1} & \cdots & \hat{r}_{1,n} & \hat{r}_{1,0} \\ \hat{r}_{2,1} & 1 & \cdots & \hat{r}_{2,k-1} & \hat{r}_{2,k+1} & \cdots & \hat{r}_{2,n} & \hat{r}_{2,0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \hat{r}_{k-1,1} & \hat{r}_{k-1,2} & \cdots & 1 & \hat{r}_{k-1,k+1} & \cdots & \hat{r}_{k-1,n} & \hat{r}_{k-1,0} \\ \hat{r}_{k+1,1} & \hat{r}_{k+1,2} & \cdots & \hat{r}_{k+1,k-1} & 1 & \cdots & \hat{r}_{k+1,n} & \hat{r}_{k+1,0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \hat{r}_{n,1} & \hat{r}_{n,2} & \cdots & \hat{r}_{n,k-1} & \hat{r}_{n,k+1} & \cdots & 1 & \hat{r}_{n,0} \end{vmatrix}.$$

Визначник Δ^* одержують з Δ додаванням першого рядка та останнього стовпчика, які складаються з оцінок коефіцієнтів кореляції поміж t_0 нормованої залежної змінної

$$t_0 = \frac{x_k - \bar{x}_k}{\sigma_k}$$

і t_q нормованими незалежними змінними:

$$t_q = \frac{x_q - \bar{x}_q}{\sigma_q}, \quad q = \overline{1, k-1, k+1, n}, \quad x_k \neq x_q,$$

$\hat{r}_{q,g}$ – оцінка парного коефіцієнта кореляції, $q, g = \overline{1, k-1, k+1, n}$.

Для перевірки гіпотези

$$H_0 : r_{x_k \cdot x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n} = 0$$

використовують статистику

$$f = \frac{N-n-1}{n} \cdot \frac{\hat{r}_{x_k \cdot x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n}^2}{1 - \hat{r}_{x_k \cdot x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n}^2},$$

яка має F -розподіл з $v_1 = n$ і $v_2 = N - n - 1$ степенями вільності. Якщо відповідна ймовірність отриманої статистики f є меншою, ніж рівень значущості α , то гіпотеза H_0 має бути відкинута.

4.4. Основи багатовимірного регресійного аналізу

Багатовимірний статистичний аналіз визначає причинно-наслідкові зв'язки об'єкта дослідження і його показників (характеристик)

$$O \leftrightarrow \{\vec{\eta}, \vec{\xi}\}, \quad (4.6)$$

де

$\vec{\xi} = (\xi_1, \xi_2, \dots, \xi_n)$ – вхідні показники;

$\vec{\eta} = (\eta_1, \eta_2, \dots, \eta_m)$ – вихідні показники.

При проведенні експерименту або спостережень реєструється масив даних $\{x_{k,l}, y_{c,l}; c = \overline{1, m}, k = \overline{1, n}, l = \overline{1, N}\}$, за яким необхідно оцінити причинно-наслідковий зв'язок (4.6).

Нехай розглядається випадок $m = 1$. Тоді задачею регресійного аналізу є дослідження зв'язку поміж випадковою величиною η і випадковим вектором $\vec{\xi} = (\xi_1, \xi_2, \dots, \xi_n)$. Наведений нижче аналіз базується на лінійній моделі зв'язку

поміж випадковими величинами:

$$\eta = \theta_0 + \theta_1 \xi_1 + \theta_2 \xi_2 + \dots + \theta_n \xi_n + E \quad (4.7)$$

де $E^T = \{\varepsilon_1, \dots, \varepsilon_n\}$ – випадкова похибка з властивостями:

1) похибки ε_k , $k = \overline{1, n}$ **некорельовані** та не залежать від реалізацій випадкової величини

$$\vec{\xi}^* = (\xi_1, \dots, \xi_n, \theta_1, \dots, \theta_n);$$

2) похибки ε_k – **нормально розподілені** з нульовим математичним сподіванням і однаковою дисперсією σ^2 :

$$E : N_N(E; 0, \sigma^2 I_N),$$

де

I_N – одинична матриця розмірності $N \times N$;

$\sigma^2 I_N$ – коваріаційна матриця вектора E .

На разі, коли обробці підлягає масив даних $\Omega_{n+1, N} = \{(x_{1,l}, \dots, x_{n,l}, y_l); l = \overline{1, N}\}$ лінійна регресійна модель має вигляд

$$\bar{y}(X) = \sum_{k=1}^n a_k x_k, \quad (4.8)$$

або

$$\bar{y}(X) = a_0 + \sum_{k=1}^n a_k x_k, \quad (4.9)$$

де a_0, a_1, \dots, a_n – невідомі параметри, що підлягають визначенню.

Ідентифікація багатовимірної регресії є більш складна процедура порівно із двовимірним випадком. Візуальна ідентифікації обмежена можливостями зображення кореляційного поля для даних, вимірність яких вища трьох. Для 3-вимірних даних кореляційне поле у вигляді сфери (рис. 4.1, а) свідчить про відсутність зв'язку. Якщо поле вписується в еліпсоїд (рис. 4.1, б), то має місце лінійна регресійна залежність. Поле складної конфігурації дає можливість говорити про нелінійний зв'язок (рис. 4.1, в) або неоднорідність даних (рис. 4.1, г).

Оцінювання коефіцієнтів регресії. Нехай функція щільності $f(\eta, \xi_1, \dots, \xi_n, a_1, \dots, a_n)$ сумісного розподілу випадкових величин η і

$\vec{\xi} = (\xi_1, \xi_2, \dots, \xi_n)$ є багатовимірною нормальною. З урахуванням (4.7), модель (4.8) стосовно до вибірки $\Omega_{n+1,N}$ можна записати у матричній формі

$$Y = XA + E,$$

де $A^T = (a_1, a_2, \dots, a_n)$ – вектор коефіцієнтів.

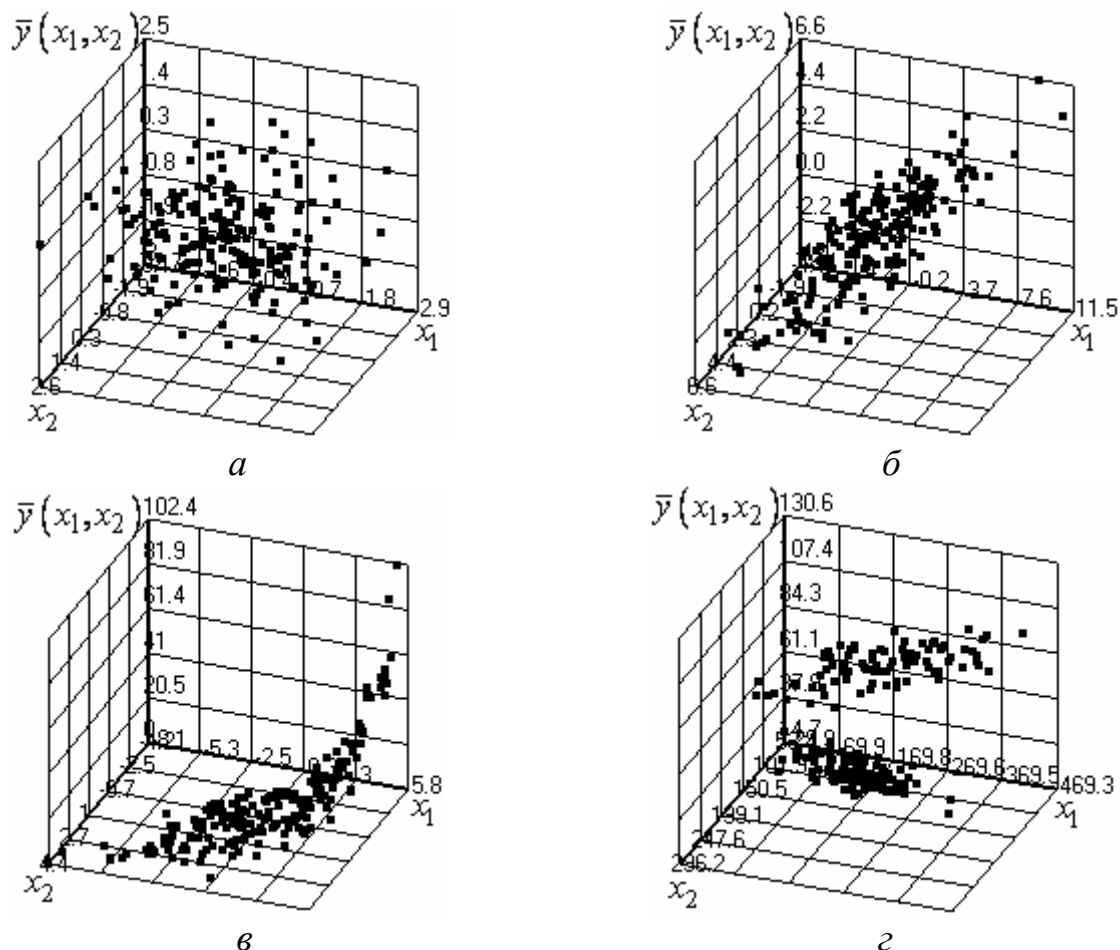


Рис. 4.1. Кореляційні поля 3-вимірних даних: а – зв'язок відсутній; б – лінійний зв'язок; в – нелінійний зв'язок; г – випадок неоднорідних даних

Припускається, що матриця X не містить лінійно залежних стовпців і її ранг дорівнює n : $rg(X) = n$. Тоді на підставі введених умов про нормальний розподіл η і $\vec{\xi} = (\xi_1, \xi_2, \dots, \xi_n)$ є вірне:

$$\begin{aligned} E\{Y\} &= E\{XA + E\} = E\{XA\} + E\{E\} = E\{XA\}, \\ \text{cov}\{Y\} &= E\left\{(Y - E\{Y\})(Y - E\{Y\})^T\right\} = \sigma^2 I_N. \end{aligned} \quad (4.10)$$

З (4.10) випливає, що вектор Y має нормальний розподіл вигляду

$$Y : N_N(Y; XA, \sigma^2 I_N).$$

Зауваження 4.3. Не слід плутати дисперсію випадкової величини η у l -му експерименті, яка дорівнює σ^2 -дисперсії похибки, з дисперсією відносно середнього η , отже, з величиною σ_y^2 .

Якщо позначити $\hat{\sigma}^2 = S_{\text{зал}}^2$, то є можливість отримати оцінки \hat{A} , $\hat{\sigma}^2$, одержані за методом найменших квадратів, відповідно до якого мінімізується сума квадратів похибок:

$$S_{\text{зал}}^2 = \frac{1}{N-n} \sum_{l=1}^N \varepsilon_l^2 = E^T E = (Y - X\hat{A})^T (Y - X\hat{A}).$$

Значення вектора \hat{A} , що мінімізує останній вираз, визначається з умови

$$\frac{\partial}{\partial \hat{A}} (E^T E) = 0$$

або

$$2X^T Y - 2X^T X\hat{A} = 0,$$

звідки

$$\hat{A} = (X^T X)^{-1} (X^T Y),$$

причому оцінки \hat{A} є незсуненими, спроможними та ефективними.

Відзначимо, що додавання вільного члена a_0 в регресійну модель (4.8) є рівнозначним введенню змінної $\xi_0 = 1$. Тоді

$$y - \bar{y} = a_1(x_1 - \bar{x}_1) + \dots + a_n(x_n - \bar{x}_n) + E,$$

або

$$(X - \bar{x})^T (X - \bar{x}) A = (X - \bar{x})^T Y_0,$$

де

\bar{x} – вектор середніх;

$Y_0 = Y - \bar{y}$ – вектор-стовпець з елементами $y_l - \bar{y}$, $l = \overline{1, N}$.

Для визначення оцінок $\hat{a}_1, \hat{a}_2, \dots, \hat{a}_n$ використовують розв'язок рівняння

$$\hat{A} = \left((X - \bar{x})^T (X - \bar{x}) \right)^{-1} (X - \bar{x})^T Y_0,$$

отже

$$\hat{a}_0 = \bar{y} - \sum_{k=1}^n \hat{a}_k \bar{x}_k.$$

Нижче у якості прикладу наведено графік відтвореної лінійної регресій-

ної моделі (4.9) (рис. 4.2).

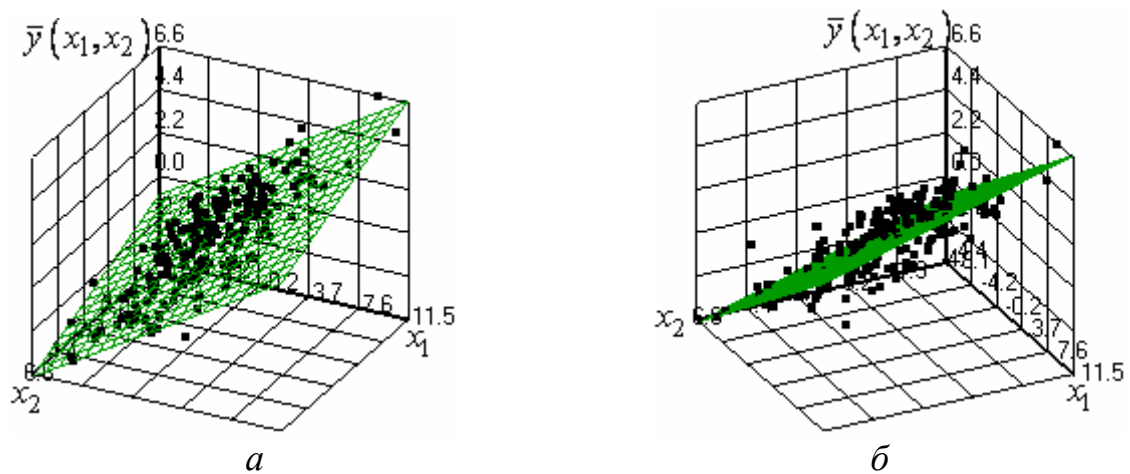


Рис. 4.2. Кореляційне поле та відтворена лінійна регресія під різними кутами

Дослідження якості відтворення регресії. Дослідження якості відтворення лінійної регресії зумовлює реалізацію процедур:

- 1) обчислення коефіцієнта детермінації R^2 ;
- 2) дослідження значущості і точності оцінок параметрів \hat{A} ;
- 3) з'ясування, які змінні здійснюють більший вплив на залежну змінну;
- 4) оцінка відхилень окремих значень y_l , $l = \overline{1, N}$ залежної змінної від емпіричної регресії $\hat{y}(X)_l = \hat{y}(x_{1,l}, \dots, x_{n,l})$;

- 5) побудова довірчого інтервалу для лінії регресії $\bar{y}(x_{1,l}, \dots, x_{n,l})$ з урахуванням її оцінки $\hat{y}(x_{1,l}, \dots, x_{n,l})$.

Поміж залишковою дисперсією $\hat{\sigma}^2$ і вибірковою дисперсією $\hat{\sigma}_y^2$ існує залежність

$$\hat{\sigma}^2 = \frac{N-1}{N-n} \hat{\sigma}_y^2 (1 - \hat{r}_{y \cdot x_1, \dots, x_n}^2),$$

де $\hat{r}_{y \cdot x_1, \dots, x_n}^2$ – квадрат оцінки вибіркового коефіцієнту множинної кореляції.

За аналогією з коефіцієнтом детермінації одновимірної моделі коефіцієнт детермінації багатовимірної моделі

$$R^2 = \hat{r}_{y \cdot x_1, \dots, x_n}^2$$

дозволяє оцінити яку частку варіації Y пояснено впливом усіх x -змінних.

Для перевірки значущості відтвореної регресії або, іншими словами,

для перевірки, чи може вплив усіх x -змінних як групи вважатися не випадковим, реалізують F -тест. Висувають головну гіпотезу

$$H_0 : a_1 = a_2 = \dots = a_n = 0,$$

при альтернативі

$$H_1 : a_k \neq 0, \forall k, k = \overline{1, n}.$$

Статистична характеристика для H_0

$$f = \frac{N-n-1}{n} \left(\frac{1}{1-R^2} - 1 \right)$$

має F -розподіл з кількістю ступенів вільності $v_1 = n$, $v_2 = N - n - 1$. Якщо виконується нерівність

$$f > f_{\alpha, v_1, v_2},$$

то слідує висновок про значущість регресійної моделі.

Зауваження 4.4. Гіпотеза $H_0 : a_1 = a_2 = \dots = a_n = 0$ є еквівалентною гіпотезі про значущість множинного коефіцієнта кореляції $H_0 : r_{y \cdot x_1, \dots, x_n} = 0$.

Дослідження значущості і точності оцінок параметрів \hat{A} . Якщо вектор \hat{A} має нормальний розподіл, то величина

$$u = \frac{\hat{\sigma}^2 (N - n)}{\sigma^2}$$

має нецентральний χ^2 -розподіл з $v = N - n$, а вектори \hat{A} і $\hat{\sigma}^2$ – статистично незалежні. Коваріаційна матриця для \hat{A} визначається у вигляді

$$\text{cov}\{\hat{A}\} = \hat{\sigma}^2 (X^T X)^{-1},$$

отже, вектор \hat{A} буде мати нормальний розподіл

$$\hat{A} : N_n \left(\hat{A}; A, \sigma^2 (X^T X)^{-1} \right).$$

Тоді кожна оцінка \hat{a}_k буде підпорядкована нормальному закону

$$\hat{a}_k : N_k \left(\hat{a}_k; a_k, \sigma^2 c_{k,k} \right), k = \overline{1, n},$$

з дисперсією

$$D\{\hat{a}_k\} = \sigma^2 c_{k,k},$$

де $c_{k,k}$ – діагональний елемент матриці $C = (X^T X)^{-1}$

та коваріацією, яка для двох будь-яких оцінок \hat{a}_k і \hat{a}_j дорівнює

$$\text{cov}\{\hat{a}_k, \hat{a}_j\} = \sigma^2 c_{k,j}.$$

Тоді при перевірці гіпотез

$$H_0 : a_k = \hat{a}_k, \quad k = \overline{1, n}$$

реалізують статистичну характеристику вигляду

$$t = \frac{\hat{a}_k - a_k}{\hat{\sigma} \sqrt{c_{k,k}}},$$

яка має t -розподіл з $v = N - n$ степенями вільності.

Інтервальна оцінка параметрів a_k відбувається на основі нерівностей

$$\hat{a}_k - t_{\alpha/2, v} \hat{\sigma} \sqrt{c_{k,k}} \leq a_k \leq \hat{a}_k + t_{\alpha/2, v} \hat{\sigma} \sqrt{c_{k,k}}, \quad k = \overline{1, n}.$$

При **інтерпретації** значення оцінок \hat{a}_0 та \hat{a}_k , $k = \overline{1, n}$ враховують наступне. Зсув на сталу \hat{a}_0 визначає очікуване значення $\bar{y}(X)$, коли всі змінні x дорівнюють нулю. Коефіцієнт регресії \hat{a}_k для кожної x_k -змінної визначає вплив цієї x_k -змінної на y при умові, що усі інші x -змінні не змінюються. Інакше кажучи: оцінка \hat{a}_k визначає, яке збільшення (чи зменшення) $\bar{y}(X)$ слід очікувати, коли усі x -змінні залишаються незмінними, за виключенням k -ої, яка збільшується на одну одиницю.

Перевірка значущості параметрів регресії на підставі гіпотез

$$H_0 : a_k = 0, \quad k = \overline{1, n}$$

здійснюється лише при умові значущості F -теста. На разі, якщо для статистики

$$t = \frac{\hat{a}_k}{\hat{\sigma} \sqrt{c_{k,k}}}$$

виконується нерівність

$$|t| \leq t_{\alpha/2, v}$$

говорять про незначущість в моделі відповідної x_k -змінної.

Оскільки параметри регресії a_1, a_2, \dots, a_n можуть бути виражені в різних одиницях вимірювання (як відповідні x_k -змінні), безпосереднє порівняння їх оцінок для з'ясування, яка з x_k -змінних має найбільший вплив на y , може викликати затруднення. В подібній ситуації навіть мале значення оцінки може виявлятися більш важливим, ніж велике, визначене в інших умовних одиницях. Вирішенням проблеми є **порівняння стандартизованих оцінок параметрів регресії**:

$$\hat{a}_k^* = \frac{\hat{a}_k \hat{\sigma}_{x_k}}{\hat{\sigma}_y},$$

де $\hat{\sigma}_{x_k}$ – оцінка середньоквадратичного відхилення x_k -змінної.

Саме абсолютні значення

$$|\hat{a}_k^*|, \quad k = \overline{1, n},$$

шляхом порівняння між собою, дозволяють отримати уяву про відносну важливість відповідних x_k -змінних.

Якщо ж при аналізі додатково цікавить вплив на y кожної x_k -змінної, при умові повної або часткової незмінності інших, аналізу підлягають відповідні оцінки парної та часткової кореляції.

Як і в одновимірній моделі, оцінка відхилень окремих значень спостережень y_i від регресії дозволяє вказати **стандартну похибку регресійної оцінки** – величина $S_{\text{зал}} = \hat{\sigma}$ приблизно вказує величину похибки оцінювання.

Довірчий інтервал для σ^2 (толерантні межі) будується на основі перевірки гіпотези

$$H_0 : \sigma^2 = \hat{\sigma}^2$$

при будь-якій альтернативі. Для перевірки гіпотези H_0 вводиться статистична характеристика вигляду

$$u = \frac{(N - n) \hat{\sigma}^2}{\sigma^2},$$

яка має нецентральний χ^2 -розподіл з $\nu = N - n$ степенями вільності. Враховуючи, що критерій двобічний, можна прийняти гіпотезу H_0 , якщо виконується

$$P \left\{ \frac{\hat{\sigma}^2 (N - n)}{\chi_{\alpha_2, (N - n)}^2} \leq \sigma^2 \leq \frac{\hat{\sigma}^2 (N - n)}{\chi_{\alpha_1, (N - n)}^2} \right\} = 1 - \alpha,$$

де

$$\alpha_{1,2} = \frac{1 \pm (1 - \alpha)}{2}.$$

Задача дослідження точності відхилень емпіричної лінії регресії від теоретичної зводиться до перевірки гіпотези

$$H_0 : \bar{y}(X) = \hat{\bar{y}}(X),$$

при будь-якій альтернативі. Як статистичну характеристику для перевірки гіпотези H_0 використовують статистику

$$t = \frac{X\hat{A} - \bar{y}(X)}{\hat{\sigma}\sqrt{1 + XCH^T}}.$$

Як завжди, гіпотезу H_0 приймають, якщо $|t| \leq t_{\alpha/2, v}$, $v = N - n$.

Довірчий інтервал для значення регресії визначають так:

$$X\hat{A} - t_{\alpha/2, v}\hat{\sigma}\sqrt{1 + XCH^T} \leq \bar{y}(X) \leq X\hat{A} + t_{\alpha/2, v}\hat{\sigma}\sqrt{1 + XCH^T}.$$

При відтворенні та аналізі багатовимірної лінійної регресії виникають три основні **різновиди проблем**: мільтиколінеарність, проблема вибору змінних та неправильний вибір моделі.

При **мультіколінеарності** окремі з x_k -змінних суттєво корелюють між собою, що може призвести до неприйнятних статистичних та обчислювальних наслідків. Статистичні наслідки полягають в тому, що при великій кореляції перевірка на значущість окремих оцінок параметрів може не відбутися, за рахунок збільшення величини $\sqrt{c_{k,k}}$, $k = \overline{1, n}$ у відповідних t -тестах. Отже, x_k -змінна може насправді мати вплив на y , але виявити це при автоматизації розрахунків стає практично неможливо. Обчислювальні проблеми пов'язані з нестійкістю обчислень у випадку мультіколінеарності – накопичення обчислювальної похибки при оцінюванні та аналізі параметрів може призвести до отримання безглузвих та помилкових результатів.

Усунення проблеми мультіколінеарності ознак здійснюють шляхом вилучення з набору змінних, що дублюють інформацію, об'єднання декількох «схожих» змінних в одну (шляхом усереднення, ділення, тощо), використання відносних величин (наприклад $1/x_k$). В будь-якому разі слід пам'ятати, що найкращі результати аналізу можливо отримати в умовах максимально можливої незалежності окремих ознак.

Проблема вибору змінних полягає в тому, що включення в модель великої кількості змінних ускладнює аналіз. Не зважаючи на те, що при збільшенні кількості ознак зростає величина коефіцієнта детермінації R^2 , стандартна похибка S_{3al}^2 – зменшується, що може призвести до відхилення F -тесту. З іншого боку ігнорування окремих ознак в моделі регресії впливає на результат аналізу, роблячи його менш точним. Тому при визначенні, які з ознак слід включити до моделі, необхідно враховувати зроблені зазначення.

Проблеми, пов'язані з неадекватністю лінійної моделі можна усувати як шляхом підбору відповідної, так і рядом суто технічних процедур, зокрема нелінійним перетворенням окремих складових, вилученням аномальних спостережень, введенням заміни (наприклад, вважати деяку $x_k^* = x_k^2$), тощо. Спосіб виявлення проблеми неадекватності відтворення багатовимірної регресії є візуальний аналіз **діагностичної діаграми**. На діагностичній діаграмі відкладаються за віссю абсцис y_l , $l = \overline{1, N}$, а за віссю ординат залишки, тобто, величини

$$\varepsilon_l = y_l - \hat{a}_0 - \sum_{k=1}^n \hat{a}_k x_{k,l}, \quad l = \overline{1, N}.$$

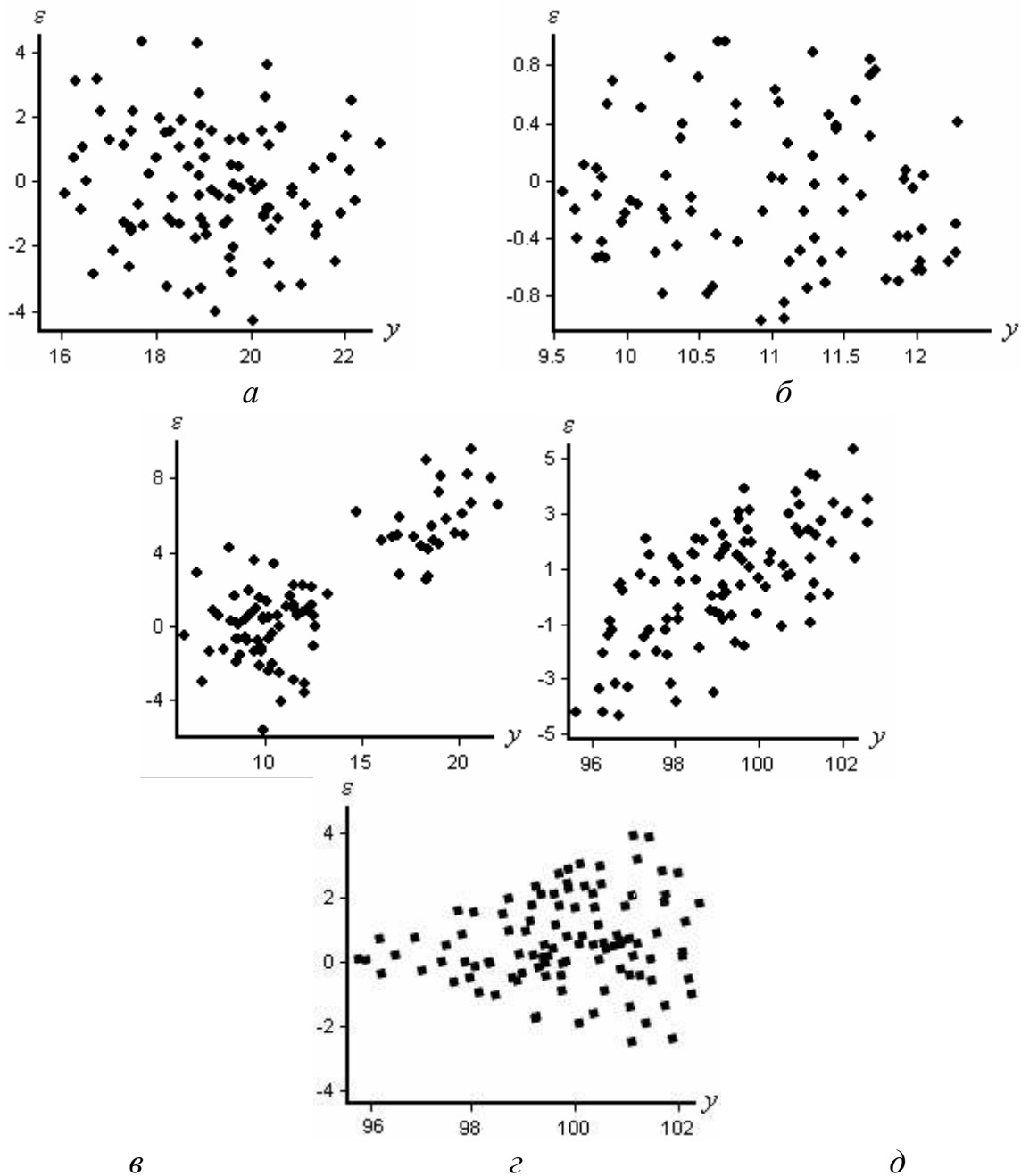


Рис. 4.3. Діагностичні діаграми: *a*, *б* – модель регресії адекватна;
в, *г*, *д* – модель регресії неадекватна

Залишки являють собою «непояснені» похибки оцінювання y за допомогою x_k -змінних. З аналізу діагностичної діаграми можуть впливати такі висновки. Якщо поле залишків колоподібне, або овалоподібне, без кутового нахилу (рис. 4.3, *a*, *б*), то модель регресії обрано адекватну. Якщо ж виявлено окремі кластери (рис. 4.3, *в*), або форма поля діаграми відмінна від зазначеної (рис. 4.3, *г*, *д*), дослідник має зосередитись на пошуку прийнятного перетворення окре-

мих x_k -змінних, на вилученні аномальних спостережень, тощо.

4.5. Компонентний та факторний аналіз

З попереднього викладення основ багатовимірної статистичної аналізу випливає, що процедура **оцінювання функцій щільності та розподілу ймовірностей**, в силу маргінальної властивості, значно **спрощується на разі незалежності складових випадкового вектору** $\vec{\xi} = (\xi_1(\omega), \dots, \xi_n(\omega))$. Адекватність та інформативність багатовимірної регресії також, певною мірою залежить від фактору ортогональності змінних-аргументів. Саме тому, серед процедур статистичного аналізу, окреме відмітне місце посідають процедури перетворення початкових масивів спостережень, що визначають **перехід до нових ортогональних систем ознак**. За всяк час реалізують лінійні перетворення випадкових величин, що надає можливість проведення аналізу в термінах нових, незалежних координат, з подальшим поверненням до початкового простору та вже у ньому інтерпретацією результатів аналізу.

4.5.1. Основи компонентного аналізу

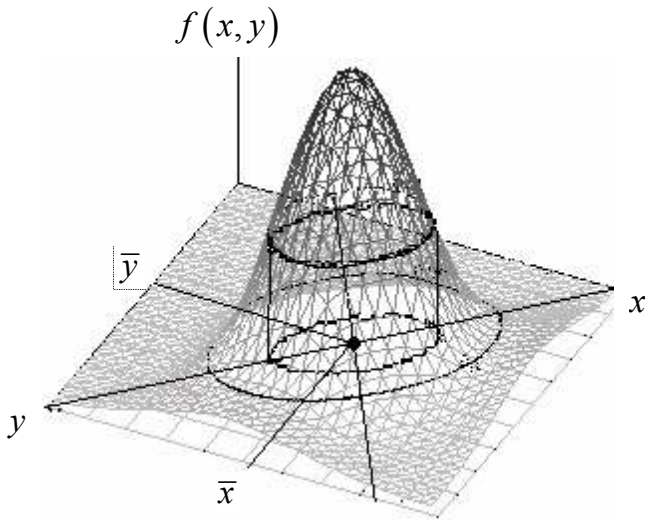
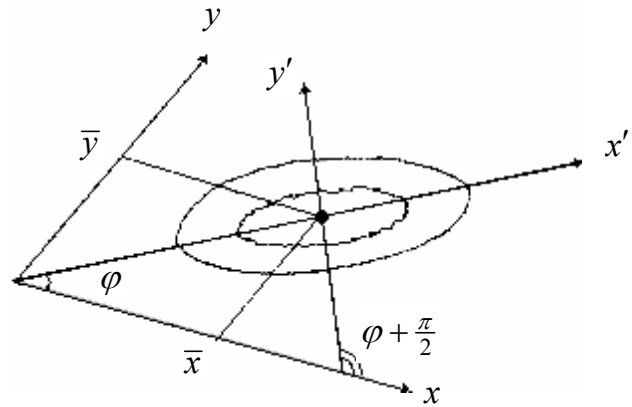
Найпростішою з обчислювальних процедур переходу до незалежних координат є процедура, що реалізується у двовимірному випадку. Так для реалізації $\Omega_{2,N} = \{(x_l, y_l); l = \overline{1, N}\}$ двовимірний випадковий вектор $\vec{\zeta} = (\xi(\omega), \eta(\omega))$, на разі, коли існує залежність між складовими $\vec{\zeta}$, можна виконати лінійне перетворення величин $\xi(\omega)$, $\eta(\omega)$, яке приводить до одержання незалежних випадкових величин $\xi'(\omega)$, $\eta'(\omega)$. Іншими словами, точки двовимірного простору, визначені масивом $\Omega_{2,N}$ (не зменшуючи загальності будемо вважати, що дані нормовані відносно середнього) необхідно представити в термінах двох головних ортогональних осей (компонент). Суть процедури полягає у використанні формул повороту ортогональної системи на деякий кут φ (рис. 4.4; 4.5):

$$\xi'(\omega) = \xi(\omega) \cos \varphi + \eta(\omega) \sin \varphi, \quad (4.11)$$

$$\eta'(\omega) = -\xi(\omega) \sin \varphi + \eta(\omega) \cos \varphi.$$

Величина кута φ , при повороті на який приходимо до незалежності $\xi'(\omega)$ і $\eta'(\omega)$, визначається із співвідношення

$$\operatorname{tg} 2\varphi = \frac{2\hat{r}_{x,y}\hat{\sigma}_x\hat{\sigma}_y}{\hat{\sigma}_x^2 - \hat{\sigma}_y^2}.$$

Рис. 4.4. Функція щільності $f(x, y)$ Рис. 4.5. Проекція розтину функції щільності $f(x, y)$

Таким чином, для подальшої обробки можемо використовувати масив $\Omega'_{2,N} = \{(x'_l, y'_l); l = \overline{1, N}\}$ реалізацій двовимірної випадкової величини $\bar{\zeta}' = (\xi'(\omega), \eta'(\omega))$, одновимірні складові якої незалежні (рис. 4.6).

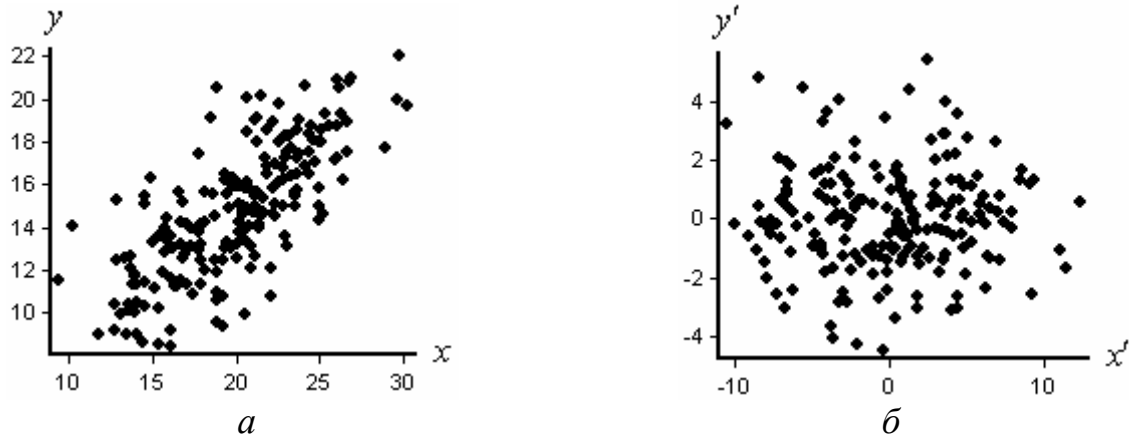


Рис. 4.6. Кореляційне поле реалізацій двовимірної випадкової величини: а – до застосування перетворення (4.11); б – після переходу до нових координат

У загальному випадку осі нової системи координат задаються лініями, для яких сума квадратів відстаней до всіляких точок масиву $\Omega_{2,N}$ мінімальна. Отже, є можливість говорити про дисперсії проекцій точок на напрямки компоненти (вісі). Розв'язання цієї задачі носить назву «метод головних компонент» (МГК). Таким чином, головні компоненти – це такі комбінації початкових ознак:

$$\begin{aligned} x'_l &= \alpha_{x',x} x_l + \alpha_{x',y} y_l, \\ y'_l &= \alpha_{y',x} x_l + \alpha_{y',y} y_l, \quad l = \overline{1, N}, \end{aligned} \quad (4.12)$$

де

$$\begin{aligned} \alpha_{x',x}^2 + \alpha_{x',y}^2 &= 1; & \alpha_{y',x}^2 + \alpha_{y',y}^2 &= 1; \\ \alpha_{x',x}, \alpha_{x',y}, \alpha_{y',x}, \alpha_{y',y} &- \text{відповідно косинуси кутів між осями} \\ x' \text{ і } x, & \quad x' \text{ і } y, & \quad y' \text{ і } x, & \quad y' \text{ і } y. \end{aligned}$$

Положення вісі x' визначається умовою максимуму дисперсії вибірки $\Omega_{2,N}$ вздовж неї; вісі y' — умовою її ортогональності до x' .

Щодо величин $\alpha_{x',x}$, $\alpha_{x',y}$, $\alpha_{y',x}$, $\alpha_{y',y}$ можна сказати наступне: метод одержання напрямків головних осей x' , y' ґрунтується на знаходженні власних чисел і власних векторів дисперсійно-коваріаційної матриці $\hat{DC}\{\vec{\zeta}\}$ або ж кореляційної матриці $\hat{R}\{\vec{\zeta}\}$, на разі, коли дані стандартизовані. В останньому випадку власний вектор з індексом x' являє собою набір коефіцієнтів

$$\vec{A}_{x'} = (\alpha_{x',x}, \alpha_{x',y}),$$

а відповідне йому власне число

$$\lambda_{x'} = 1 + |\hat{r}_{x,y}|$$

дорівнює дисперсії компоненти x' . Тоді як при використанні матриці $\hat{R}\{\vec{\zeta}\}$ сума власних чисел дорівнює кількості ознак, то знайшовши відношення $\frac{\lambda_{x'}}{2}$, — одержують частку дисперсії вибірки, що відповідає x' -му напрямку. Те ж саме, відповідно, стосується і y' -го напрямку (тут маємо $\lambda_{y'} = 1 - |\hat{r}_{x,y}|$). У будь-якому разі, як для нормованих, так и для стандартизованих даних, якщо позначити дисперсії головних компонент $\sigma_{x'}^2$, $\sigma_{y'}^2$, виявляється, що

$$\sigma_{x'}^2 + \sigma_{y'}^2 = \sigma_x^2 + \sigma_y^2,$$

отже, початково закладена в спостереженнях сумарна варіабельність при переході до нових змінних не змінюється, а лише перерозподіляється. Окрім того, нові змінні (компоненти), на відміну від початкових ознак, набувають таку цінну властивість, як відсутність кореляції поміж собою.

Зворотне перетворення для повернення до вихідних ознак наступне

$$x_l = \alpha_{x,x'}x'_l + \alpha_{y,x'}y'_l, \quad (4.13)$$

$$y_l = \alpha_{x,y'}x'_l + \alpha_{y,y'}y'_l, \quad l = \overline{1, N},$$

де $\alpha_{x,x'}$, $\alpha_{y,x'}$, $\alpha_{x,y'}$, $\alpha_{y,y'}$ — косинуси кутів між відповідними осями.

Так само, як і в двовимірному випадку, при обробці вибірки $\Omega_{n,N} = \{x_{k,l}; k = \overline{1,n}, l = \overline{1,N}\}$ є можливість застосування **методу головних компонент** задля переходу до нової ортогональної (незалежної) системи ознак x'_k , $k = \overline{1,n}$. Має місце наступне подання:

$$x'_k = \sum_{v=1}^n \alpha_{k,v} x_v, \quad \sum_{v=1}^n \alpha_{k,v}^2 = 1, \quad (4.14)$$

де $\alpha_{k,v}$ – косинус кута між осями x'_k та x_v .

Отже, шляхом **лінійного перетворення** (4.14) з вихідного масиву $\Omega_{n,N}$ отримують

$$\Omega'_{n,N} = \{x'_{k,l}; k = \overline{1,n}, l = \overline{1,N}\},$$

де

$$x'_{k,l} = \sum_{v=1}^n \alpha_{k,v} x_{v,l}, \quad k = \overline{1,n}, \quad l = \overline{1,N}.$$

Слід наголосити, що вісь x'_1 визначається умовою максимуму дисперсії даних $\Omega_{n,N}$ уздовж її напрямку; компонента x'_2 – умовою максимуму дисперсії серед комбінацій x_k , $k = \overline{1,n}$, що не корелюють з x'_1 (умовою максимуму дисперсії даних, без урахування дисперсії вздовж x'_1 , отже, осі x'_1 та x'_2 визначають площину, вздовж якої розсіювання даних $\Omega_{n,N}$ максимальне); x'_3 – умовою максимуму дисперсії серед комбінацій x_k , $k = \overline{1,n}$, що не корелюють з x'_1 та x'_2 (x'_1 , x'_2 та x'_3 визначають тримірну гіперплощину, вздовж якої дисперсія початкових даних максимальна) і т. д.

Зворотне перетворення для визначення вихідних даних через головні компоненти таке:

$$x_k = \sum_{v=1}^n \alpha_{v,k} x'_v, \quad k = \overline{1,n}, \quad (4.15)$$

або (що еквівалентно)

$$x_{k,l} = \sum_{v=1}^n \alpha_{v,k} x'_{v,l}, \quad k = \overline{1,n}, \quad l = \overline{1,N}.$$

Як і в двовимірному випадку, виконується

$$\sum_{k=1}^n \sigma_k'^2 = \sum_{k=1}^n \sigma_k^2,$$

де $\sigma_k'^2$, $k = \overline{1,n}$ – дисперсії ознак, після виконання перетворення (4.14),

отже, має місце **перерозподіл варіабельності**, початково закладеної в спостереженнях, окрім того

$$\sigma_1'^2 \geq \sigma_2'^2 \geq \dots \geq \sigma_n'^2.$$

Остання нерівність може бути використана при прийнятті рішення про ігнорування для подальшої обробки сукупності новоутворених ознак

$$x'_k, \quad k = \overline{w+1, n}, \quad w < n,$$

дисперсії котрих мало відрізняються від нуля. Подібна ситуація має місце при наявності суттєвої кореляції між окремими ознаками вихідного масиву спостережень. Тоді, якщо насправді в даних існує лінійна залежність, а, отже, деякі з головних компонент є малоінформативними, модель (4.14) можна подати так:

$$x'_k = \sum_{v=1}^n \alpha_{k,v} x_v, \quad k = \overline{1, w}, \quad w < n. \quad (4.16)$$

Зазначимо, що на разі використання лише w перших головних компонент, зворотне перетворення

$$x_k = \sum_{v=1}^w \alpha_{v,k} x'_v, \quad k = \overline{1, n}, \quad w < n \quad (4.17)$$

фактично визначає результат **фільтрації вихідного масиву** $\Omega_{n,N}$, шляхом вилучення високочастотної складової варіабельності даних.

Якщо $w \leq 3$, то є можливість візуального представлення даних $\Omega_{n,N}$ у термінах головних компонент, що, звичайно, сприяє їхньому аналізу та опрацюванню безпосередньо дослідником в інтерактивному режимі.

Для визначення величини w використовують наступне правило. Якщо вихідні дані є стандартизованими, то, тоді як $\hat{\sigma}_k^2 = 1$, $k = \overline{1, n}$, то

$$\sum_{k=1}^n \hat{\sigma}_k'^2 = n$$

і тоді в (4.16), (4.17) за w приймають кількість перших компонент, для яких виконується нерівність

$$\hat{\sigma}_k'^2 \geq 1.$$

Відносно методу визначення величин $\alpha_{k,v}$, $k, v = \overline{1, n}$ із моделі (4.14), то, як і для двовимірного випадку, відшукують власні значення та відповідні власні вектори дисперсійно-коваріаційної матриці $\hat{DC}\{\vec{\xi}\}$ нормованих відносно середнього даних спостережень (матриці $\hat{R}\{\vec{\xi}\}$ – для стандартизованих даних).

Власний вектор матриці $\hat{R}\{\vec{\xi}\}$ являє собою набір коефіцієнтів

$$\vec{A}_{x'_k} = (\alpha_{k,1}, \dots, \alpha_{k,n}), \quad k = \overline{1, n},$$

а відповідні власні числа

$$\lambda_k, \quad k = \overline{1, n},$$

такі, що

$$\sum_{k=1}^n \lambda_k = n,$$

визначають **частку** $\frac{\lambda_k}{n}$ **дисперсії**, яка припадає на відповідний напрямок, причому для нестандартизованих масивів виконується

$$\frac{\lambda_k}{n} = \sigma_k'^2, \quad k = \overline{1, n}.$$

На разі, коли реалізується модель (4.16), величина

$$\delta_k = \sum_{v=1}^w \alpha_{v,k}^2 \cdot 100\%, \quad k = \overline{1, n} \quad (4.18)$$

встановлює, який відсоток дисперсії k -ої вихідної ознаки x_k пояснено першими w головними компонентами.

Для безпосереднього визначення величин λ_k та векторів $\vec{A}_{x'_k}$, $k = \overline{1, n}$ можна застосовувати будь-який відомий з курсу обчислювальної математики метод відшукування власних значень та власних векторів матриць (метод Крилова, обертань, ітераційні тощо).

Підводячи підсумок, можна зазначити: якщо $\Omega_{n,N} = \{x_{k,l}; k = \overline{1, n}, l = \overline{1, N}\}$ є масив реалізацій n -вимірної випадкової величини $\vec{\xi} = (\xi_1(\omega), \dots, \xi_n(\omega))$, окремі одновимірні складові якої можуть бути лінійно резресіно залежними між собою, то за використанням перетворення (4.14) відбувається перехід до ортогональної системи ознак, отже обробці підлягає масив $\Omega'_{n,N} = \{x'_{k,l}; k = \overline{1, n}, l = \overline{1, N}\}$ реалізацій випадкової величини $\vec{\xi}' = (\xi'_1(\omega), \dots, \xi'_n(\omega))$ з незалежними одновимірними складовими. Тоді за масивом $\Omega'_{n,N}$ є можливість проведення ймовірнісної оцінки реалізацій n -вимірної випадкової величини. Зворотне перетворення (4.15) дозволяє застосування результатів подібної оцінки до вихідного масиву $\Omega_{n,N}$. Окрім того, використання моделі (4.16) дає змогу зменшення розмірності з n до w вихідного простору ознак, а зворотне перетворення (4.17), окрім іншого, – проведення фільтрації даних, причому співвідношення (4.18) дозволяє оцінку частки дисперсії даних після фільтрації.

4.5.2. Розвідницький факторний аналіз

На сьогоднішній день аналіз однорідних багатовимірних нормально розподілених статистичних даних у більшості випадків розв'язку прикладних задач дозволяє: знаходити інтенсивність зв'язку між змінними (методи кореляції), одержувати аналітичну форму зв'язку (методи регресії) і **визначати фактори, що породжують цей зв'язок** (теорія та методи факторного аналізу).

Математично факторний аналіз займається вивченням внутрішньої структури кореляційних матриць. У факторному аналізі припускається, що змінні, які спостерігаються, являють собою лінійну комбінацію деяких латентних (гіпотетичних) факторів.

Постановка задачі факторного аналізу

Нехай маємо масив спостережень $\Omega_{n,N} = \{x_{k,l}; k = \overline{1,n}, l = \overline{1,N}\}$, причому, стосовно даних бажаним є **виконання** ряду наступних **рекомендацій**:

- 1) ознаки об'єкту спостережень, що вимірюються є рівноправними поміж себе;
- 2) ознаки об'єкту, що вимірюються є неперервними випадковими нормально розподіленими величинами;

3) розгляду та аналізу підлягають лише змінні, що вимірюються. Список змінних не збільшують за рахунок додавання сум, добутків та інших аналогічних функцій від вимірних ознак.

На відміну від підходу, пов'язаного з максимізацією загальної дисперсії в МГК, у моделі факторного аналізу потребується якнайкраще апроксимувати кореляції. **Головну модель факторного аналізу** можна записати у вигляді:

$$x_k = a_{k,1}F_1 + a_{k,2}F_2 + \dots + a_{k,w}F_w + d_kU_k, \quad k = \overline{1,n}, \quad w < n. \quad (4.19)$$

В даній моделі параметр x_k лінійно залежить від w **загальних факторів** F_v , $v = \overline{1,w}$ і **характерного фактора** U_k . Загальні фактори враховують кореляції між ознаками, характерний фактор ураховує залишену (в тому числі й пов'язану з різноманітними похибками) дисперсію. Коефіцієнти при загальних факторах називають навантаженнями. **Головною задачею факторного аналізу** є **знаходження** матриці навантажень (**матриці факторного відображення**)

$$A = \|a_{k,v}\|, \quad k = \overline{1,n}, \quad v = \overline{1,w}.$$

Використовуючи (4.19), координату кожного вектора спостережень x_k можна подати у вигляді

$$x_{k,l} = \sum_{v=1}^w a_{k,v}f_{v,l} + d_k u_{k,l}, \quad k = \overline{1,n}, \quad l = \overline{1,N}.$$

Тоді дисперсія k -ї стандартизованої ознаки може бути виражена через фактори так:

$$S_k^2 = \frac{1}{N} \sum_{l=1}^N x_{k,l}^2 = \sum_{v=1}^w a_{k,v}^2 \frac{1}{N} \sum_{l=1}^N f_{v,l}^2 + \frac{d_k^2}{N} \sum_{l=1}^N u_{k,l}^2 + \\ + 2 \sum_{v \neq q=1}^w a_{k,v} a_{k,q} \frac{1}{N} \sum_{l=1}^N (f_{v,l} f_{q,l}) + 2 d_k \sum_{v=1}^w a_{k,v} \frac{1}{N} \sum_{l=1}^N (f_{v,l} u_{k,l}).$$

Оскільки дисперсія параметра, заданого у стандартному вигляді, дорівнює одиниці, всі параметри (і фактори також) припускаються заданими у стандартному вигляді, можна записати:

$$S_k^2 = 1 = \sum_{v=1}^w a_{k,v}^2 + d_k^2 + 2 \sum_{v \neq q=1}^w a_{k,v} a_{k,q} r_{F_v, F_q} + 2 d_k \sum_{v=1}^w a_{k,v} r_{F_v, U_k}. \quad (4.20)$$

Характерні фактори завжди **некорельовані** з загальними, тому $r_{F_v, U_k} = 0$ для $\forall v, k$. А якщо прийняти ортогональність загальних факторів, то і $r_{F_v, F_q} = 0$ для $\forall v, q, v \neq q$. Отже, (4.20) перепишеться у вигляді:

$$S_k^2 = 1 = a_{k,1}^2 + a_{k,2}^2 + \dots + a_{k,w}^2 + d_k^2, \quad (4.21)$$

де $a_{k,v}^2$ – внесок фактора F_v в дисперсію параметра x_k ;
 d_k^2 – внесок характерного фактора в дисперсію параметра.

Загальний внесок v -го загального фактора у підсумкову дисперсію параметра x_k визначається за формулою

$$V_v = \sum_{k=1}^n a_{k,v}^2, \quad v = \overline{1, w},$$

а повний внесок усіх загальних факторів у підсумкову дисперсію дорівнює

$$V = \sum_{v=1}^w V_v.$$

Із виразу (4.21) випливають два важливих поняття факторного аналізу:

1) **загальність** (communality) параметра x_k , яка є часткою його дисперсії, яку можна описати загальними факторами:

$$h_k^2 = \sum_{v=1}^w a_{k,v}^2, \quad k = \overline{1, n};$$

2) **характерність** (uniqueness) параметра x_k , а саме

$$d_k^2, \quad k = \overline{1, n}.$$

Повсякчас з характерності виділяють дві складові: специфічність – частку дисперсії, яка дійсно пов’язана зі специфікою параметрів, що вивчаються, і ту частку, що пов’язана з похибками вимірювань. У цьому випадку модель (4.19) записують у вигляді

$$x_k = a_{k,1}F_1 + a_{k,2}F_2 + \dots + a_{k,w}F_w + b_kB_k + e_kE_k, \quad k = \overline{1, n}, \quad (4.22)$$

де

B_k – фактор специфічності;

E_k – фактор похибки.

Тоді дисперсія k -ї ознаки об’єкта спостережень набуває вигляду

$$S_k^2 = 1 = h_k^2 + d_k^2 = h_k^2 + b_k^2 + e_k^2.$$

Якщо відома вибіркова дисперсія похибки вимірювання k -го параметра e_k^2 , можна визначити надійність змінної (reliability):

$$rel_k = 1 - e_k^2$$

(аналогічно $rel_k = h_k^2 + b_k^2$). Із визначень загальності і надійності k -го параметра випливає поняття повноти факторизації C_k (частки загальності у відсотках від дисперсії, що враховується загальними факторами):

$$C_k = \frac{h_k^2}{rel_k} \cdot 100\%.$$

Моделі (4.19), (4.22) називають **факторним відображенням** або просто – відображенням. Факторний аналіз дозволяє одержати не тільки відображення, а й значення коефіцієнтів кореляції між параметрами й факторами. Таблиця таких коефіцієнтів кореляції називається **факторною структурою** або просто структурою. Для виконання повного факторного аналізу необхідні і відображення, і структура.

Якщо помножити кожне з рівнянь (4.19) на відповідні фактори, провести додавання за усіма N спостереженнями та поділити на N , одержують:

$$r_{x_k, F_1} = a_{k,1} + a_{k,2}r_{F_1, F_2} + \dots a_{k,v}r_{F_1, F_v} + \dots + a_{k,w}r_{F_1, F_w},$$

...

$$r_{x_k, F_v} = a_{k,1}r_{F_v, F_1} + a_{k,2}r_{F_v, F_2} + \dots a_{k,v} + \dots + a_{k,w}r_{F_v, F_w},$$

...

$$r_{x_k, F_w} = a_{k,1}r_{F_w, F_1} + a_{k,2}r_{F_w, F_2} + \dots a_{k,v}r_{F_w, F_v} + \dots + a_{k,w},$$

та

$$r_{x_k, U_k} = d_k.$$

Якщо прийняти ортогональність загальних факторів, маємо, що $r_{F_v, F_q} = 0$ для $\forall v, q, v \neq q$. Тоді $a_{k,v}$ – коефіцієнт матриці навантажень – дорівнює коефіцієнту кореляції k -го параметра з v -м фактором.

Зауваження 4.5. Якщо покласти $h_k^2 = 1, k = \overline{1, n}$ – як частковий випадок одержуємо МГК.

У припущенні ортогональності загальних факторів кореляція k -го і v -го параметрів визначається за виразом

$$r'_{k,v} = a_{k,1}a_{v,1} + a_{k,2}a_{v,2} + \dots + a_{k,w}a_{v,w}, \quad k, v = \overline{1, n}, k \neq v. \quad (4.23)$$

Як кореляцію параметра з самим собою вибирають значення загальності, яка при зроблених припущеннях дорівнює квадрату коефіцієнта множинної кореляції між параметром і загальними факторами.

Система (4.23) у матричному вигляді може бути представлена так:

$$R_h = AA^T, \quad (4.24)$$

де

R_h – кореляційна матриця ознак, на діагоналі якої замість одиниць розташовані загальності h_k^2 , так звана **редуційна матриця**;

A^T – транспонована матриця коефіцієнтів навантажень.

Останнє рівняння презентує математичну постановку задачі розвідницького факторного аналізу: за відомою матрицею парних кореляцій ознак об'єкта спостережень знайти матрицю A і набір загальностей, що задовольняє (4.24).

Обчислювальні схеми факторного аналізу

В основу наведеного нижче ітераційного методу розв'язування задачі факторного аналізу покладено метод головних факторів і метод мінімальних залишків, запропонований Г.Харманом. Відмінність МГФ від МГК полягає в тому, що фактори визначаються власними векторами і власними числами редуційної кореляційної матриці R_h . Таким чином, попередньо необхідно одержати оцінки загальностей і кількість загальних факторів w .

Іншими словами, факторний аналіз проводиться на підставі **характеристичного рівняння** матриці R_h , яке має вигляд

$$\det(R_h - \lambda I) = 0.$$

Аналітично загальність параметра дорівнює квадрату коефіцієнта множинної кореляції між самим параметром і загальними факторами:

$$h_k^2 = R_{x_k \cdot F_1, F_2, \dots, F_w}^2 \quad (4.25)$$

На практиці реалізується спроба наблизити **попередню оцінку загальності** до виразу (4.25). Нижче розглядається два найпоширеніші методи такого оцінювання.

1. Метод максимальних кореляцій. За k -у загальність приймається максимальний за модулем із позадіагональних елементів k -го рядка матриці обчислених коефіцієнтів парної кореляції R

$$h_k^2 = \max_{\substack{v=1, w, \\ v \neq k}} |r_{k,v}|, \quad k = \overline{1, n}, \quad v = \overline{1, w}, \quad k \neq v. \quad (4.26)$$

Теоретично метод не обґрунтовано, проте він дає непогані результати, особливо для матриць зі значущими коефіцієнтами кореляції.

2. Метод, що ґрунтується на підсумках МГК: загальність k -го параметра визначається із співвідношення

$$h_k^2 = \sum_{v=1}^w \alpha_{k,v}^2, \quad k = \overline{1, n}, \quad (4.27)$$

де $\alpha_{k,v}$ – коефіцієнти при головних осях моделі компонентного аналізу; w – кількість головних компонент.

Оцінка кількості загальних факторів для моделі факторного аналізу (4.19) проводиться неоднозначно, проте є можливість указати деякі загальні правила такого визначення. Аналізуючи власні числа матриці коефіцієнтів парної кореляції R , за кількість загальних факторів w зазвичай беруть число власних значень, що є більшими за одиницю. Для матриці R , що відноситься до генеральної сукупності, це дає нижню оцінку кількості загальних факторів. При використанні редуційної матриці R_h за число факторів беруть кількість, яка дорівнює кількості власних значень, більших від їх середнього значення (але лише у тому разі, якщо така кількість є більшою від одержаної при розгляді матриці R).

Одержана кількість факторів за наведеним критерієм узгоджується з приведеними Г.Харманом значеннями мінімального рангу редуційної матриці у залежності від кількості параметрів.

Так само, як і у випадку оцінки загальностей, існує багато методів виділення факторів. У статистичних пакетах обробки даних повсякчасно реалізують такі: метод головних факторів (МГФ), метод мінімальних залишків (ММЗ) та метод максимальної правдоподібності.

Зокрема, МГФ, є схожим на МГК, але з тою відмінністю, що фактори визначаються власними векторами і власними числами редуційної кореляційної матриці R_h . Перший фактор, одержаний за МГФ, дає максимальний внесок у сумарну загальність (сума $a_{1,1}^2 + a_{2,1}^2 + \dots + a_{w,1}^2$ максимальна). Другий фактор ура-

ховує максимум залишкової загальності і не корелює з першим і т.ін.

При реалізації ММЗ попередньо необхідно задати розмірність простору загальних факторів. Надалі факторне відображення одержують ітеративно таким чином, щоб мінімізувати суму квадратів позадіагональних елементів залишкової матриці.

Під залишковою матрицею $R_{зал}$ розуміють матрицю

$$R_{зал} = R_h - AA^T.$$

Отже, метод зводиться до мінімізації функціоналу

$$f = \sum_{v=1}^w \sum_{q=1}^w r_{залv,q}^2, \quad v \neq q. \quad (4.28)$$

Зауваження 4.6. При використанні методу накладають обмеження, щоб оцінка загальностей не перевищувала одиницю.

Для реалізації в програмному середовищі пропонується наступний **ітераційний метод**, що показав більшу універсальність при обробці даних від наведених вище. Метод ґрунтується на повторному застосуванні МГФ до редуційної матриці R_h . Спочатку на головну діагональ R_h розміщують загальності, що були одержані в підсумку МГФ при умові мінімуму функції f (4.28) для набору попередніх оцінок загальностей (4.26), (4.27), або будь-яких інших оцінок загальностей, отриманих на основі відомих методів. Отже, з усіх попередніх оцінок загальностей обирається та єдина, використання якої з конкретною матрицею кореляції найоптимальніше (кожна з них є найоптимальнішою лише для певної структури матриці R).

У подальшому на головну діагональ R_h розміщують нові одержані загальності і процес факторизації продовжують ітераційно до зупинки. Для роботи методу необхідно одночасне виконання трьох умов:

$$1) \quad f^{(i+1)} < f^{(i)},$$

де i – номер ітерації;

$$2) \quad \sum_{k=1}^n \sum_{g=1}^n \left(a_{k,g}^{(i+1)} - a_{k,g}^{(i)} \right)^2 > \varepsilon,$$

де ε – будь-яке наперед задане число;

$$3) \quad h_k^2 \leq 1, \quad k = \overline{1, n}.$$

Порушення першої умови свідчить про досягнення мінімуму функції f . Друга умова використовується для завершення процесу, коли при переході від ітерації до ітерації елементи шуканої матриці факторних навантажень A мало змінюються. Як правило, умова добре працює при обробці матриць з великою кількістю значущих коефіцієнтів кореляції. Остання третя умова вводиться для неможливості ситуації, коли загальність перевищує одиницю (усякчас це

пов'язано з від'ємними власними значеннями). Нижче наведено алгоритм.

1. Знаходження матриці кореляцій R за даними спостережень.
2. Знаходження власних значень і власних векторів матриці R .
3. Визначення кількості власних факторів w .
4. Знаходження попередніх оцінок загальностей параметрів.
5. Виконання кроків 2–4 для (4.26), (4.27) (або у вигляді іншої оцінки) вектора оцінок загальностей і відповідної редуційної матриці R_h . Знаходження значення f .
6. Вибір вектора загальностей, що відповідає мінімуму функції f . Підготовка редуційної матриці для початку процесу ітерацій.
7. Допоки одночасно виконуються умови продовження процесу ітерацій 1–3, виконувати кроки 2–3.

Приклад 4.2. Аналізувались дані варіабельності структури врожаю озими, одержані при проведенні польового дослідження на 39 ділянках, в яких вивчались 11 ознак, що характеризують середнє значення показника по ділянці [11]:

- X_1 кількість кущів з 1 кв.м;
- X_2 кількість продуктивних стебел з 1 кв.м;
- X_3 кількість непродуктивних стебел з 1 кв.м;
- X_4 висота рослини, см;
- X_5 довжина колоса, см;
- X_6 кількість зерен у колосі;
- X_7 вага снопа з 1 кв.м, г;
- X_8 маса 1000 зерен, г;
- X_9 вага зерна 100 рослин, г;
- X_{10} вага соломи 100 рослин, г;
- X_{11} врожай зерна ц/га.

При проведенні факторного аналізу, на підставі кореляційної матриці ознак (табл.4.1), обмежились трьома загальними факторами, які пояснюють 76% варіабельності початкових даних (табл.4.2).

Таблиця 4.1

**Кореляційна матриця ознак, що характеризують
структуру врожаю озимини**

№	1	2	3	4	5	6	7	8	9	10	11
1	1										
2	,559	1									
3	,165	,076	1								
4	,089	-,083	-,211	1							
5	-,035	,016	-,278	,227	1						
6	,066	,201	-,286	-,053	,796	1					
7	,293	,613	-,086	,376	,503	,508	1				
8	-,149	-,223	-,192	,503	,184	-,137	,096	1			
9	-,071	,016	-,349	,128	,832	,872	,495	,261	1		
10	,118	-,057	-,308	,291	,738	,675	,400	,252	,761	1	
11	,396	,783	-,205	,035	,505	,686	,783	-,005	,614	,405	1

Таблиця 4.2

**Матриця факторного відображення, побудована
за даними про структуру врожаю озимини**

№	F_1	F_2	F_3
X_1	-0,140	0,746	0,013
X_2	-0,021	0,927	-0,175
X_3	-0,447	0,149	-0,242
X_4	0,080	0,128	0,885
X_5	0,894	0,085	0,127
X_6	0,915	0,226	-0,235
X_7	0,451	0,730	0,244
X_8	0,128	-0,149	0,818
X_9	0,951	0,075	0,089
X_{10}	0,808	0,063	0,251
X_{11}	0,560	0,771	0,061

Відповідно до результатів (табл.4.2), перший фактор F_1 має навантаження (що перевищують 0,7) на ознаки: 5, 6, 9, 10; аналогічні ознаки для другого фактора F_2 : 1, 2, 7, 11; нарешті, для третього фактора F_3 : 4, 8.

Є всі підстави вважати перший фактор узагальненою ознакою, що описує продуктивність рослин, другий фактор – узагальненою ознакою, що описує продуктивність 1 кв.м ділянки, третій фактор – ознакою, що характеризує морфометричну структуру рослини.

Контрольні запитання та завдання

1. Перерахувати складові первинного статистичного аналізу реалізацій багатовимірних випадкових величин.
2. Дати визначення емпіричної функції розподілу за багатовимірним варіаційним рядом.
3. Як здійснюють перевірку гіпотези про рівність двох багатовимірних середніх у разі рівності дисперсійно-коваріаційних матриць?
4. Навести процедуру реалізації H -критерію.
5. Дати визначення оцінки часткового коефіцієнта кореляції. Яким чином перевіряють його значущість?
6. Яка інтерпретація множинного коефіцієнта кореляції?
7. Скільки степенів вільності має статистика для перевірки гіпотези про значущість множинного коефіцієнта кореляції?
8. Які властивості похибки в лінійній моделі багатовимірної регресії?
9. Записати залишкову дисперсію лінійної багатовимірної регресії.
10. Яка залежність існує між залишковою дисперсією та дисперсією залежної змінної в лінійній багатовимірній регресійній моделі?
11. Записати рівняння для визначення оцінок параметрів лінійної багатовимірної регресії.
12. Як перевіряють значущість відтворення багатовимірної регресії?
13. У який спосіб здійснюють інтервальну оцінку параметрів лінійної моделі багатовимірної регресії?
14. Для чого використовують стандартизовані оцінки параметрів регресії? Як здійснюють стандартизацію оцінок параметрів?
15. Прокоментувати: які проблеми виникають при відтворенні та аналізі багатовимірної лінійної регресії?
16. Що таке діагностична діаграма, для чого її використовують?
17. Як здійснюють перехід до незалежних ознак у двовимірному випадку?
18. Чому дорівнює величина кута φ , при повороті на який, приходимо до незалежності двох випадкових величин?
19. Записати головні компоненти як комбінації початкових ознак.
20. Яким співвідношенням пов'язані дисперсії початкових ознак та головних компонент?

21. Записати зворотне перетворення для визначення вихідних даних через головні компоненти.
22. Сформулювати обчислювальну процедуру методу головних компонент.
23. Чому дорівнює сума власних чисел кореляційної матриці?
24. Встановити, який відсоток дисперсії k -ої вихідної ознаки x_k пояснено першими w головними компонентами.
25. У який спосіб здійснюють фільтрацію багатовимірних даних на основі методу головних компонент?
26. Сформулювати постановку задачі розвідницького факторного аналізу.
27. Дати інтерпретацію матриці факторного відображення.
28. Яке пояснення має дисперсія k -ої вихідної ознаки x_k при реалізації моделі факторного аналізу?
29. Який загальний внесок v -го загального фактора у підсумкову дисперсію параметра x_k ?
30. Чому дорівнює повний внесок усіх загальних факторів у загальну дисперсію багатовимірної вибірки?
31. Дати визначення загальності, характерності та редукційної матриці.
32. Сформулювати математичну постановку задачі розвідницького факторного аналізу.
33. Навести приклади проведення оцінки загальності.
34. Як оцінюють кількість загальних факторів?
35. Сформулювати обчислювальну процедуру методу головних факторів.
36. Навести обчислювальну процедуру ітераційного методу побудови моделі розвідницького факторного аналізу.

Розділ 5. ОСНОВИ РОЗПІЗНАВАННЯ ОБРАЗІВ

У попередніх розділах припускали, що дані, які підлягають обробці та аналізу, є однорідні. Під **неоднорідними** будемо розуміти дані, сформовані з різних генеральних сукупностей. У такому разі гістограма зазвичай багатомодальна, а дані утворюють у просторі ознак декілька сукупностей, конфігурація та властивості яких можуть значно різнитися (рис 5.1).

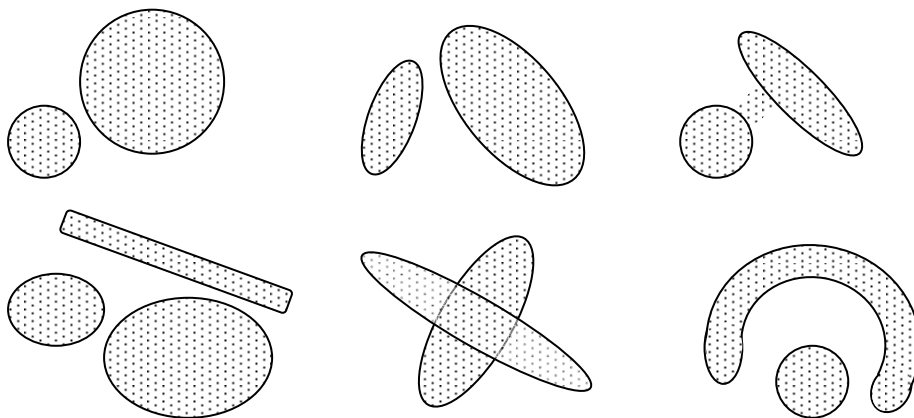


Рис. 5.1. Форми неоднорідних сукупностей

У випадку обробки неоднорідних даних постають **два типи задач**. Перша пов'язана із виділенням у сукупності однорідних груп (класів, або кластерів) для подальшої обробки, друга – із розпізнаванням, до якої з неоднорідних груп відноситься новий об'єкт. Для їх вирішення застосовують методи теорії розпізнавання образів.

Перша задача в теорії розпізнавання образів має назву «**розпізнавання без вчителя**». Припускається що задано результати спостережень над N об'єктами, кожен з яких характеризується n ознаками; необхідно розбити множину об'єктів на однорідні групи на основі подібності значень їх ознак. Синонімом розпізнавання є кластерний аналіз, таксономія, авоматична класифікація. Надалі будемо використовувати термін «кластерний аналіз» і розглянемо два найбільш поширені методи кластерного аналізу – ієрархічний та K -середніх.

Друга задача – це **розпізнавання з вчителем**. Припускається, що задано результати спостережень над N об'єктами, кожен з яких характеризується n ознаками та «вчителем» віднесений до одного з K класів; необхідно на основі цих даних (навчаючої вибірки) навчитися відносити новий об'єкт до одного із

заданих класів. Задачу розпізнавання з вчителем будемо називати задачею класифікації і розглянемо для її розв'язку метод дискримінантного аналізу.

5.1. Кластерний аналіз

Кластеризація – це є розбиття множини об'єктів на однорідні підмножини за схожістю опису ознак об'єктів. При цьому вимагають, щоб об'єкти однієї підмножини були більш схожі між собою, ніж з об'єктами інших підмножин. Одержані в результаті розбиття підмножини називають кластерами.

Нехай задана множина об'єктів

$$\Omega_{n,N} = \{X_l; l = \overline{1, N}\} = \{(x_{1,l}, \dots, x_{n,l}); l = \overline{1, N}\},$$

яку треба кластеризувати. Формально кластеризація (розбиття)

$$S = \{S_j; j = \overline{1, K}\}$$

являє собою сукупність непустих кластерів

$$S_j = \{X_l^{(j)}; l = \overline{1, N_j}\}, \quad j = \overline{1, K},$$

таких що

$$S_i \cap S_j = \emptyset,$$

коли

$$i \neq j, \quad i, j = \overline{1, K}, \quad \Omega_{n,N} = \bigcup_{j=1}^K S_j.$$

Кількість кластерів K заздалегідь може бути невідома.

Розв'язання задачі кластеризації вимагає визначення поняття однорідності об'єктів та кластерів.

5.1.1. Відстані між об'єктами та кластерами

У загальному випадку поняття однорідності об'єктів задається або введенням правила обчислення **відстані** $d(X_l, X_h)$ між об'єктами X_l та X_h , або функції $s(X_l, X_h)$, що характеризує ступінь схожості об'єктів.

Невід'ємну функцію $d(X_l, X_h)$ називають функцією відстані або метрикою відстані, якщо вона задовольняє умовам:

1) невід'ємності:

$$d(X_l, X_h) \geq 0 \quad \text{для } \forall X_l, X_h;$$

2) максимальної схожості об'єкта із самим собою:

$$d(X_l, X_h) = 0, \text{ коли } X_l = X_h;$$

3) симетрії:

$$d(X_l, X_h) = d(X_h, X_l) \text{ для } \forall X_l, X_h;$$

4) нерівності трикутника:

$$d(X_l, X_h) \leq d(X_l, X_q) + d(X_q, X_h) \text{ для } \forall X_l, X_h, X_q.$$

Найбільш розповсюджені **метрики відстані** для випадку неперервних даних:

1) евклідова відстань

$$d(X_l, X_h) = \sqrt{\sum_{k=1}^n (x_{k,l} - x_{k,h})^2};$$

2) зважена евклідова відстань

$$d(X_l, X_h) = \sqrt{\sum_{k=1}^n \omega_k (x_{k,l} - x_{k,h})^2},$$

де $\omega_k > 0$ – ваговий коефіцієнт k -ї ознаки;

3) манхетенська відстань (або відстань міських кварталів)

$$d(X_l, X_h) = \sum_{k=1}^n |x_{k,l} - x_{k,h}|;$$

дана відстань також може бути застосована для якісних ознак, у такому разі її називають Хемінговою відстанню;

4) відстань Чебишева:

$$d(X_l, X_h) = \max_{1 \leq k \leq n} |x_{k,l} - x_{k,h}|;$$

5) узагальнена степенева відстань Мінковського

$$d(X_l, X_h) = \sqrt[m]{\sum_{k=1}^n |x_{k,l} - x_{k,h}|^m},$$

де m – показник степеня; як правило, використовують значення $m = 1, 2, \infty$, які приводять до трьох попередніх відстаней: при $m = 2$ має місце евклідова відстань, коли $m = 1$ – манхетенська, при $m = \infty$ – Чебишева;

6) відстань Махаланобіса

$$d(X_l, X_h) = \sqrt{(X_l - X_h)^T V^{-1} (X_l - X_h)},$$

де $V = \{v_{k,p}, k, p = \overline{1, n}\}$ – матриця коваріацій між показниками, елементи якої обчислюються за формулою

$$v_{k,p} = \sum_{i=1}^N (x_{k,i} - \bar{x}_k)(x_{p,i} - \bar{x}_p);$$

при цьому якщо кореляції між змінними відсутні, відстань Махаланобіса еквівалентна квадрату евклідової відстані.

Поняттям, протилежним відстані, є **міра схожості**

$$s(X_l, X_h)$$

між об'єктами X_l та X_h . Якщо більші значення метрики відстані говорять про меншу схожість об'єктів, то для міри схожості навпаки, більші значення свідчать про більшу схожість. Зазвичай дуже легко перейти від метрик відстаней до міри схожості і навпаки. Наприклад, міру схожості можна обчислювати так:

$$s(X_l, X_h) = \exp(-d(X_l, X_h)).$$

Перед обчисленням міри відстані або схожості часто виконуються перетворення змінних. Їх мета – приведення змінних до єдиного масштабу для осмисленого порівняння об'єктів по різних змінних. Найбільш вживана стандартизація

$$x_{k,l}^* = \frac{x_{k,l} - \bar{x}_k}{\hat{\sigma}_k}, \quad k = \overline{1, n}, \quad l = \overline{1, N},$$

де

$$\bar{x}_k = \frac{1}{N} \sum_{l=1}^N x_{k,l};$$

$$\hat{\sigma}_k = \frac{1}{N-1} \sum_{l=1}^N (x_{k,l} - \bar{x}_k)^2,$$

окрім якої можна застосовувати й такі типи перетворень:

$$x_{k,l}^* = \frac{x_{k,l}}{\bar{x}_k},$$

$$x_{k,l}^* = \frac{x_{k,l}}{x_{k,\max}},$$

$$x_{k,l}^* = \frac{x_{k,l}}{x_{k,\max} - x_{k,\min}}, \quad k = \overline{1, n}, \quad l = \overline{1, N},$$

де

$$x_{k,\max} = \max \{x_{k,1}, \dots, x_{k,N}\}; \quad x_{k,\min} = \min \{x_{k,1}, \dots, x_{k,N}\}.$$

Поряд із відстанями між об'єктами доцільно ввести поняття **відстані між кластерами**, тобто між цілими групами об'єктів. Нехай маємо два кластери: $S_1 = \{X_l^{(1)}, l = \overline{1, N_1}\}$ та $S_2 = \{X_l^{(2)}, l = \overline{1, N_2}\}$. Відстань $D(S_1, S_2)$ між ними можна визначати як

1) **відстань найближчого сусіда** – це є відстань між найближчими об'єктами кластерів (рис. 5.2)

$$D(S_1, S_2) = \min_{\substack{l_1 = \overline{1, N_1}; \\ l_2 = \overline{1, N_2}}} d(X_{l_1}^{(1)}, X_{l_2}^{(2)});$$

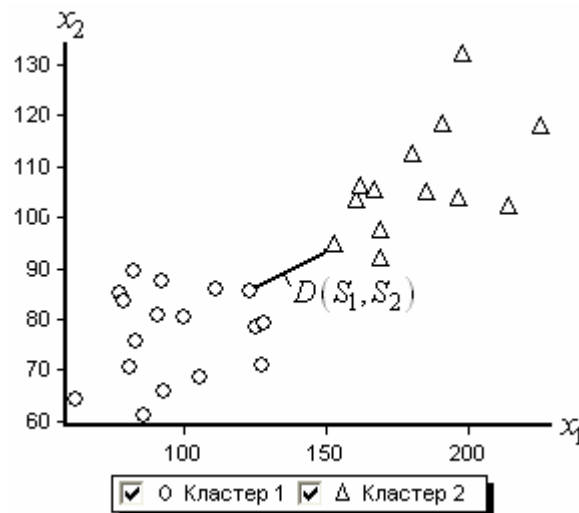


Рис. 5.2. Відстань між кластерами найближчого сусіда
(для випадку евклідової відстані між об'єктами)

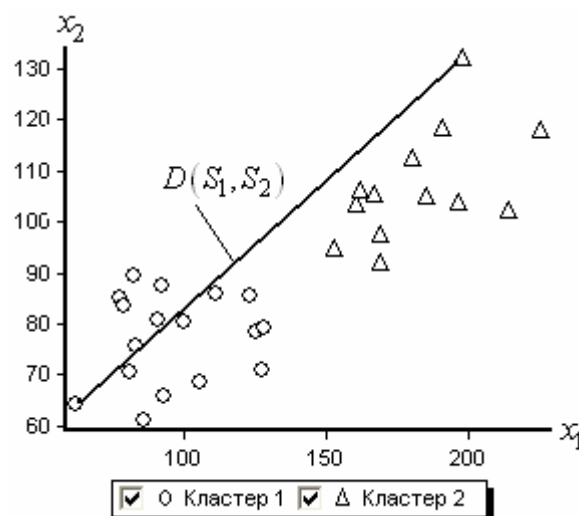


Рис. 5.3. Відстань між кластерами найвіддаленішого сусіда
(для випадку евклідової відстані між об'єктами)

2) **відстань найвіддаленішого сусіда**, яка дорівнює відстані між найбільш віддаленими об'єктами кластерів (рис. 5.3)

$$D(S_1, S_2) = \max_{\substack{l_1=1, N_1; \\ l_2=1, N_2}} d(X_{l_1}^{(1)}, X_{l_2}^{(2)});$$

3) **середню зважену відстань**, що дорівнює середньому значенню попарних відстаней між об'єктами різних кластерів

$$D(S_1, S_2) = \frac{1}{N_1 N_2} \sum_{l_1=1}^{N_1} \sum_{l_2=1}^{N_2} d(X_{l_1}^{(1)}, X_{l_2}^{(2)});$$

4) **середню незважену відстань**, яка відрізняється від попередньої тим, що в ній не враховуються розміри кластерів

$$D(S_1, S_2) = \frac{1}{4} \sum_{l_1=1}^{N_1} \sum_{l_2=1}^{N_2} d(X_{l_1}^{(1)}, X_{l_2}^{(2)});$$

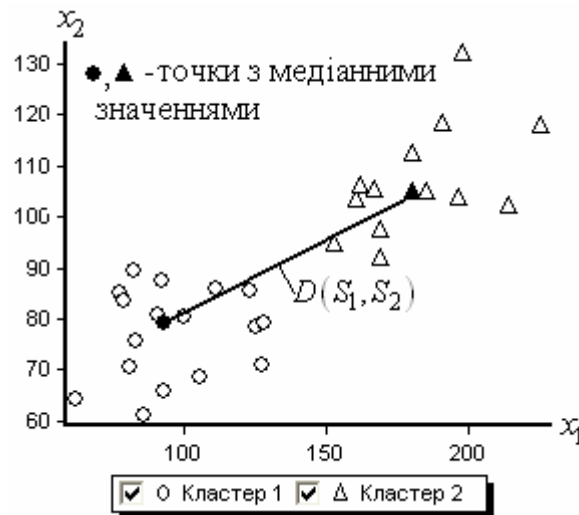


Рис. 5.4. Відстань між кластерами як медіанна відстань (для випадку евклідової відстані між об'єктами)

5) **медіанну відстань** (рис. 5.4)

$$D(S_1, S_2) = \frac{1}{2} d(Me^{(1)}, Me^{(2)}),$$

де

$$Me^{(q)} = (me_1^{(q)}, me_2^{(q)}, \dots, me_n^{(q)})$$

– об'єкт із медіанними значеннями ознак у q -му кластері, $q = 1, 2$;

значення $me_k^{(q)}$, $k = \overline{1, n}$, – визначаються за відсортованою вибіркою

$\{x_{1,k}^{(q)}, x_{2,k}^{(q)}, \dots, x_{N_q,k}^{(q)}\}$ за формулою

$$me_k^{(q)} = \begin{cases} x_{(N_q+1)/2,k}^{(q)}, & \text{коли } N_q - \text{ не парне,} \\ \frac{1}{2} \left(x_{N_q/2,k}^{(q)} + x_{N_q/2+1,k}^{(q)} \right), & \text{коли } N_q - \text{ парне;} \end{cases}$$

6) **відстань між центрами**, що дорівнює відстані між об'єктами із середніми значеннями усіх ознак (рис. 5.5)

$$D(S_1, S_2) = d(\bar{X}^{(1)}, \bar{X}^{(2)}),$$

де

$$\bar{X}^{(q)} = (\bar{x}_1^{(q)}, \bar{x}_2^{(q)}, \dots, \bar{x}_n^{(q)}),$$

$$\bar{x}_k^{(q)} = \frac{1}{N_q} \sum_{l=1}^{N_q} x_{l,k}^{(q)}, \quad q = 1, 2;$$

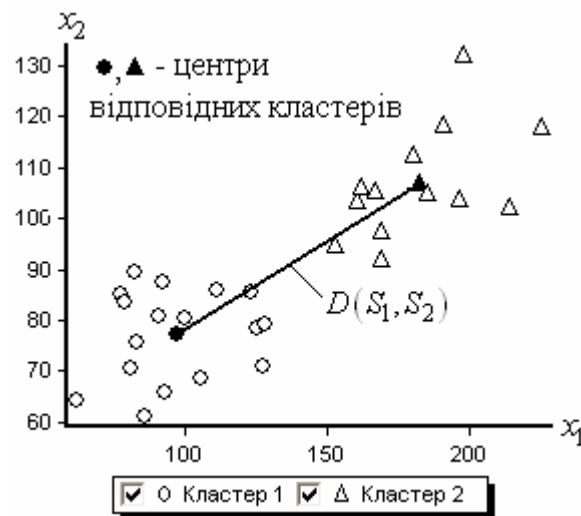


Рис. 5.5. Відстань між кластерами як відстань між їх центрами (для випадку евклідової відстані між об'єктами)

7) відстань Уорда

$$D(S_1, S_2) = \frac{N_1 N_2}{N_1 + N_2} d^2(\bar{X}^{(1)}, \bar{X}^{(2)}).$$

Поняття відстані між кластерами особливо важливе у ієрархічних методах кластерного аналізу.

5.1.2. Ієрархічні методи кластеризації

Ієрархічні методи кластеризації бувають двох видів: агломеративні (об'єднувальні) та дивізімні (розділювальні). В **агломеративних ієрархічних методах** спочатку кожен об'єкт розглядають як окремий кластер, далі знаходять два найбільш близько розташовані кластери та об'єднують їх в один. Процес об'єднання продовжують доки усі об'єкти не утворять один кластер. Логічна протилежність – є **дивізімні методи**. В них на початковому етапі усі об'єкти вважають належними одному кластеру, який ділять на складові частини доки кожен об'єкт на утворе окремий кластер.

На практиці найбільш вживані агломеративні методи, тому зупинимось на них детальніше. Нехай обрано метрику відстані й обчислено матрицю відстаней $D = \|d_{l,h} = d(X_l, X_h); l, h = \overline{1, N}\|$. Тоді процес розбиття об'єктів на кластери відбувається послідовно за $N-1$ крок. На першому кроці кожен об'єкт вважають окремим кластером, тобто $S_j = \{X_j\}$, $j = \overline{1, N}$. В матриці відстаней D знаходять мінімальний елемент $d_{l,h}$ і кластери S_l та S_h об'єднують в один кластер $S_{l+h} = S_l \cup S_h$, що складається вже з двох об'єктів. Після цього матриця D змінюють. З неї викидають два рядки і два стовпчика, що містять відстані від S_l та S_h до інших кластерів, але додають один рядок і один стовпець з відстанями від кластера S_{l+h} до інших. Далі на кожному кроці процедура повторюють, тобто знаходять мінімальний елемент у перетвореній матриці відстаней і відповідні кластери об'єднують в один.

При цьому відстані від нового кластера до інших можна перераховувати з використанням вже відомих відстаней за формулою **Ланса–Уільямса**:

$$D(S_{l+h}, S_m) = \alpha_l D(S_l, S_m) + \alpha_h D(S_h, S_m) + \\ + \beta D(S_l, S_h) + \gamma |D(S_l, S_m) - D(S_h, S_m)|,$$

де

$\alpha_l, \alpha_h, \beta, \gamma$ – числові параметри.

Різні сполучення параметрів $\alpha_l, \alpha_h, \beta, \gamma$ відповідають різним способам обчислення відстані між кластерами та породжують різні види агломеративних ієрархічних методів:

1) **найближчого сусіда** (або одного зв'язку):

$$\alpha_l = 0,5;$$

$$\alpha_h = 0,5;$$

$$\beta = 0; \gamma = -0,5;$$

2) найвіддаленішого сусіда (або повного зв'язку):

$$\alpha_l = 0,5;$$

$$\alpha_h = 0,5;$$

$$\beta = 0; \gamma = 0,5;$$

3) середнього зваженого зв'язку:

$$\alpha_l = \frac{N_l}{N_l + N_h};$$

$$\alpha_h = \frac{N_h}{N_l + N_h};$$

$$\beta = 0; \gamma = 0,$$

де

N_l, N_h – кількість об'єктів у кластерах S_l та S_h , які об'єднують;

4) простого середнього зв'язку:

$$\alpha_l = 0,5;$$

$$\alpha_h = 0,5;$$

$$\beta = 0;$$

$$\gamma = 0;$$

5) медіанного зв'язку:

$$\alpha_l = 0,5;$$

$$\alpha_h = 0,5;$$

$$\beta = -0,25;$$

$$\gamma = 0;$$

6) центроїдний:

$$\alpha_l = \frac{N_l}{N_l + N_h};$$

$$\alpha_h = \frac{N_h}{N_l + N_h};$$

$$\beta = -\frac{N_l N_h}{(N_l + N_h)^2};$$

$$\gamma = 0;$$

7) Уорда:

$$\alpha_i = \frac{N_m + N_l}{N_m + N_l + N_h};$$

$$\alpha_j = \frac{N_m + N_h}{N_m + N_l + N_h};$$

$$\beta = -\frac{N_m}{N_m + N_l + N_h};$$

$$\gamma = 0,$$

де

N_m – кількість об'єктів у кластері S_m , відстань до якого обчислюють.

Недоліком методу найближчого сусіда є наявність ланцюгового ефекту. Він об'єднує в один кластер навіть дуже далеко розташовані об'єкти, якщо існує з'єднуючий їх ланцюг об'єктів. Метод найвіддаленішого сусіда, навпаки, має тенденцію до виявлення компактних гіперсферичних кластерів.

Методи середнього зваженого та простого середнього зв'язку вважаються проміжними за своїми властивостями між першими двома.

Методи медіанного зв'язку та центроїдний рідко використовуються на практиці. На це є дві причини. По-перше, одержана за ними кластеризація має інверсії, тобто у процесі послідовного об'єднання кластерів відстань на кожному кроці не обов'язково збільшується. Як наслідок, дендрограма (див. далі) обов'язково має самоперетини. По-друге, в них міжкластерні відстані не є редуковані. Властивість редукованості, введена М. Брюїношем, полягає у тому, що для $\forall \delta > 0$ δ -окіл кластера, одержаного об'єднанням двох інших кластерів, має знаходитися всередині δ -околів початкових кластерів. Якщо дана властивість має місце, то можна прискорити процедуру пошуку кластерів-кандидатів для об'єднання на кожному кроці, шукаючи їх лише серед кластерів, що потрапили до δ -околу розглянутих на попередніх кроках алгоритму кластерів.

Метод Уорда намагається мінімізувати суму квадратів відстаней між двома гіпотетичними кластерами, що можуть бути сформовані на кожному кроці. Його недоліком є тенденція до утворення кластерів малого розміру

Перевага усіх зазначених методів – є наочність проведеного аналізу. Результати можна представити у вигляді дендрограми, яка графічно зображує ієрархічну структуру даних. **Дендрограма** – це графік, у якому за вертикальною віссю відкладаються номери об'єктів, а за горизонтальною – міжкластерні відстані, за яких відбувалося об'єднання двох кластерів (рис. 5.6). Слід зазначити, що під час побудови дендрограми об'єкти краще відкладати у порядку, в якому вони виявилися розташованими на останньому кроці роботи алгоритму. Це дозволить одержати дендрограму без самоперетинів.

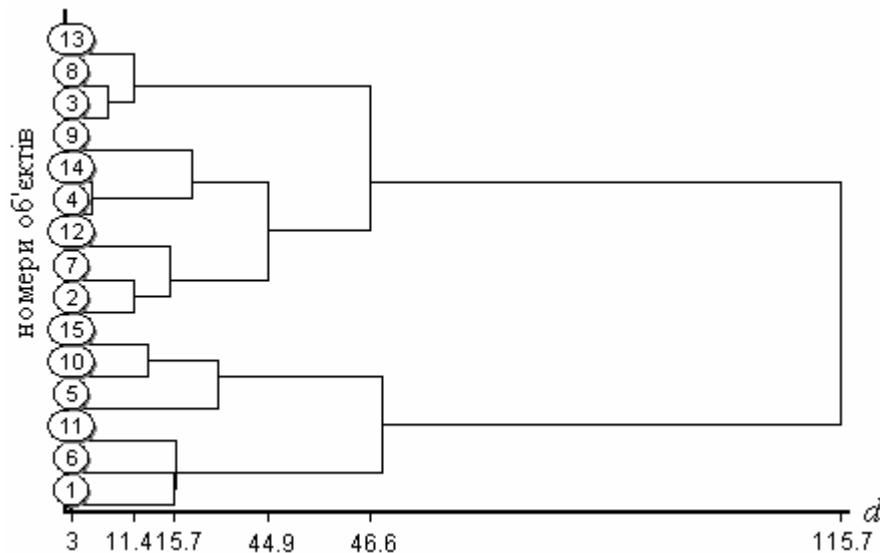


Рис. 5.6. Дендрограма

Головний недолік агломеративних методів – потреба у великому обсязі пам'яті та значній кількості обчислювальних операцій. В меншій мірі такий недолік стосується методу *К*-середніх.

5.1.3. Метод *К*-середніх

Метод *К*-середніх (*K*-means) є найбільш поширений серед неієрархічних методів кластеризації. Для його роботи попередньо необхідно задавати кількість кластерів *K*. Даний метод існує у двох варіантах: Болла–Холла та Мак-Кіна.

Нижче наведений **алгоритм методу *К*-середніх** у варіанті Болла–Холла.

1. Обирається *K* точок, які вважаються центрами кластерів $\bar{X}^{(j)} = (\bar{x}_1^{(j)}, \bar{x}_2^{(j)}, \dots, \bar{x}_n^{(j)})$, $j = \overline{1, K}$. Вони можуть бути визначені одним із способів:

- 1) як *K* найбільш віддалених точок вихідної вибірки;
- 2) випадкові *K* точок;
- 3) перші *K* точок.

2. Кожна точка X_l , $l = \overline{1, N}$, приєднується до кластеру, центр якого ближчий. Номер кластеру y_l , до якого відноситься X_l , визначається так:

$$y_l = h: \quad d(X_l, \bar{X}^{(h)}) = \min_{j=1, K} d(X_l, \bar{X}^{(j)}).$$

3. Центри кластерів $\bar{X}^{(j)} = (\bar{x}_1^{(j)}, \bar{x}_2^{(j)}, \dots, \bar{x}_n^{(j)})$ перераховуються як центри мас:

$$\bar{x}_k^{(j)} = \frac{1}{N_j} \sum_{\substack{l=1, N, \\ y_l=j}} x_{k,l}, \quad k = \overline{1, n}, \quad j = \overline{1, K},$$

де

$$N_j = \sum_{\substack{l=1, N, \\ y_l=j}} 1 \text{ — кількість точок у } j\text{-му кластері.}$$

4. Кроки 2–3 повторюються для нових центрів кластерів доки не буде виконана одна з умов:

1) центри кластерів не стабілізуються, тобто на двох послідовних ітераціях не буде виконуватися умова

$$|\bar{x}_k^{(j,t+1)} - \bar{x}_k^{(j,t)}| \leq \varepsilon, \quad k = \overline{1, n}, \quad j = \overline{1, K},$$

де

t — номер ітерації;

$\varepsilon > 0$ — будь-яке наперед задане число;

2) кількість ітерацій не стане рівною заданій максимальній кількості.

Метод K -середніх у варіанті Мак-Кіна відрізняється тим, що центри кластерів перераховуються після віднесення кожної точки до найближчого кластеру, тобто кроки 2 та 3 алгоритму виконуються не послідовно, а паралельно. Мак-Кін показав, що такий алгоритм приводить до мінімізації суми внутрішньокластерних дисперсій.

Зауваження 5.1. Назва «метод K -середніх» запропонована Дж. Мак-Кіном. Але фактично алгоритм у варіанті Болла та Холла реалізує ідею того ж методу з іншим порядком стабілізації, не послідовно для кожного класу, а паралельно для усіх класів. Тому його також називають методом K -середніх (І.Д. Мандель, 1988).

Даний метод вирізняється простотою реалізації, наочністю та швидкістю роботи, що дозволяє його застосування для дуже великих за обсягом вибірок. Головний недолік полягає у тому, що метод дуже чутливий до початкового вибору центрів кластерів та їх кількості.

5.1.4. Оцінка якості кластеризації

Розбиття вихідної вибірки на кластери, одержані різними методами або при різних значеннях параметрів, можуть відрізнятися. Кількісним критерієм, що дозволяє віддати перевагу одному із розбиттів, є функціонал якості.

Нехай у результаті застосування методу кластеризації одержане розбиття

$$S = \{S_1, S_2, \dots, S_K\}$$

вихідної вибірки на K кластерів

$$S_j = \{X_l^{(j)}; l = \overline{1, N_j}\} = \{(x_{1,l}^{(j)}, \dots, x_{n,l}^{(j)}); l = \overline{1, N_j}\}, \quad j = \overline{1, K},$$

де

N_j – кількість об'єктів у кластері S_j ;

$X_l^{(j)} = (x_{1,l}^{(j)}, x_{2,l}^{(j)}, \dots, x_{n,l}^{(j)})$ – l -й об'єкт кластера S_j , $l = \overline{1, N_j}$;

$x_{k,l}^{(j)}$ – значення k -ї ознаки об'єкта $X_l^{(j)}$, $k = \overline{1, n}$.

Для оцінки якості розбиття S відомо близько 50 функціоналів якості. Найбільш поширені такі:

1. Сума («зважена») внутрішньокластерних дисперсій:

$$Q_1(S) = \sum_{j=1}^K \sum_{l=1}^{N_j} d^2(X_l^{(j)}, \bar{X}^{(j)}),$$

де

$\bar{X}^{(j)}$ – центр кластера S_j .

2. Сума попарних внутрішньокластерних відстаней, що також має бути мінімальна:

$$Q_2(S) = \sum_{j=1}^K \sum_{l=1}^{N_j-1} \sum_{h=l+1}^{N_j} d(X_l^{(j)}, X_h^{(j)}).$$

Зручність даного функціоналу у тому, що його мінімізація автоматично забезпечує максимізацію суми міжкластерних відстаней.

3. Загальна внутрішньокластерна дисперсія має бути мінімальна:

$$Q_3(S) = \det \left(\sum_{j=1}^K N_j V_j \right)$$

або

$$Q'_3(S) = \prod_{j=1}^K \det(V_j)^{N_j},$$

де

$$V_j = \left\| v_{k,p}^{(j)}, k, p = \overline{1, n} \right\|$$

– матриця коваріацій кластера S_j , елементи якої обраховуються за формулою

$$v_{k,p}^{(j)} = \frac{1}{N_j} \sum_{l=1}^{N_j} \left(x_{k,l}^{(j)} - \bar{x}_k^{(j)} \right) \left(x_{p,l}^{(j)} - \bar{x}_p^{(j)} \right).$$

4. Відношення функціоналів, що повинно бути мінімальне:

$$Q_4(S) = \frac{Q'_4(S)}{Q''_4(S)},$$

де

$Q'_4(S)$ – середня внутрішньокластерна відстань

$$Q'_4(S) = \frac{1}{\sum_{j=1}^K \frac{N_j(N_j-1)}{2}} \cdot \sum_{j=1}^K \sum_{l=1}^{N_j-1} \sum_{h=l+1}^{N_j} d(X_l^{(j)}, X_h^{(j)});$$

$Q''_4(S)$ – середня міжкластерна відстань

$$Q''_4(S) = \frac{1}{\prod_{j=1}^K N_j} \cdot \sum_{j=1}^{K-1} \sum_{l=1}^{N_j} \left(\sum_{m=j+1}^K \sum_{h=1}^{N_m} d(X_l^{(j)}, X_h^{(m)}) \right).$$

Оскільки мінімізація $Q'_4(S)$ не гарантує максимізації $Q''_4(S)$, щоб урахувати як внутрішньокластерну, так і міжкластерну відстань застосовують відношення у вигляді $Q_4(S)$.

Якість багатьох методів кластеризації значною мірою залежить від вибору кількості кластерів K . Задача вибору оптимальної кількості кластерів поки не має однозначного розв'язку. Стандартна рекомендація полягає у проведенні кластеризації при різних значеннях K та виборі розбиття, при якому досягається екстремальне значення заданого функціонала якості.

5.2. Дискримінантний аналіз

Імовірнісна постановка **задачі класифікації** припускає, що кожен об'єкт

$$X_l = (x_{1,l}, \dots, x_{n,l}), \quad l = \overline{1, N},$$

є реалізація однієї з K n -вимірних випадкових величин

$$\bar{\xi}_j, \quad j = \overline{1, K},$$

що відповідають класам

$$S_1, \dots, S_K.$$

Для кожного класу припускаються заданими:

$$f_j(X) = f_j(x_1, \dots, x_n), \quad j = \overline{1, K},$$

– функції щільності розподілу ймовірностей, які характеризують імовірність появи об'єкта

$$X = (x_1, \dots, x_n)$$

з класу S_j ;

$$p_j, \quad j = \overline{1, K},$$

– апіорні імовірності появи об'єктів з класу S_j . Треба побудувати правило класифікації, що мінімізує імовірність помилкової класифікації.

Користуючись формулою Байеса, можна для будь-якого об'єкта $X = (x_1, \dots, x_n)$ обчислити апостеріорну імовірність належності його до класу S_j :

$$P(S_j|X) = \frac{p_j f_j(X)}{\sum_{w=1}^K p_w f_w(X)}. \quad (5.1)$$

Тоді природно запропонувати таке **правило класифікації**: відносити черговий об'єкт X до класу, для якого максимальна його апостеріорна імовірність, тобто

$$X \in S_j,$$

якщо

$$P(S_j|X) > P(S_h|X) \quad (5.2)$$

для

$$\forall j, h = \overline{1, K}, \quad j \neq h.$$

Правило (5.2) називають **баєсівським вирішальним правилом**. Оскільки знаменник у формулі Байеса (5.1) для усіх класів однаковий, то для прийняття рішення щодо віднесення об'єкта до класу достатньо порівнювати лише чисельники

$$p_j f_j(x_1, \dots, x_n), \quad j = \overline{1, K},$$

тобто правило (5.2) еквівалентне такому:

$$X \in S_j,$$

якщо

$$p_j f_j(X) > p_h f_h(X) \quad (5.3)$$

для

$$\forall j, h = \overline{1, K}, \quad j \neq h$$

або, зважаючи, що функція логарифма є монотонно зростаюча, такому:

$$X \in S_j,$$

якщо

$$\ln p_j + \ln f_j(X) > \ln p_h + \ln f_h(X) \quad (5.4)$$

для

$$\forall j, h = \overline{1, K}, \quad j \neq h.$$

Канонічною формою запису правил (5.2)–(5.4) вважається їх подання за допомогою системи дискримінантних функцій

$$g_j(X), \quad j = \overline{1, K}.$$

Дискримінантними називають такі **функції** ознак $g_j(X)$, що для усіх

$$X \in S_j$$

$$g_j(X) > g_h(X) \quad \text{для } \forall j, h = \overline{1, K}, \quad j \neq h.$$

Для правил (5.2)–(5.4) дискримінантні функції мають вигляд:

$$g_j(X) = P(S_j | X),$$

$$g_j(X) = p_j f_j(X), \quad (5.5)$$

$$g_j(X) = \ln p_j + \ln f_j(X).$$

Наведені варіанти дискримінантних функцій приводять до однакових результатів, лише деякі є зручніші у використанні.

Дія кожного з вирішальних правил полягає у тому, щоб розбити простір ознак на K областей прийняття рішень

$$R_1, R_2, \dots, R_K,$$

таких що R_j містить об'єкти, які на основі вирішального правила були віднесені до класу S_j . Области R_j та R_h розділяються межами областей рішень – такими поверхнями у просторі ознак, уздовж яких

$$g_j(X) = g_h(X), \quad \forall j, h = \overline{1, K}, \quad j \neq h.$$

Застосування баєсівського вирішального правила проілюстровано для випадку трьох класів та одновимірних даних (рис. 5.7).

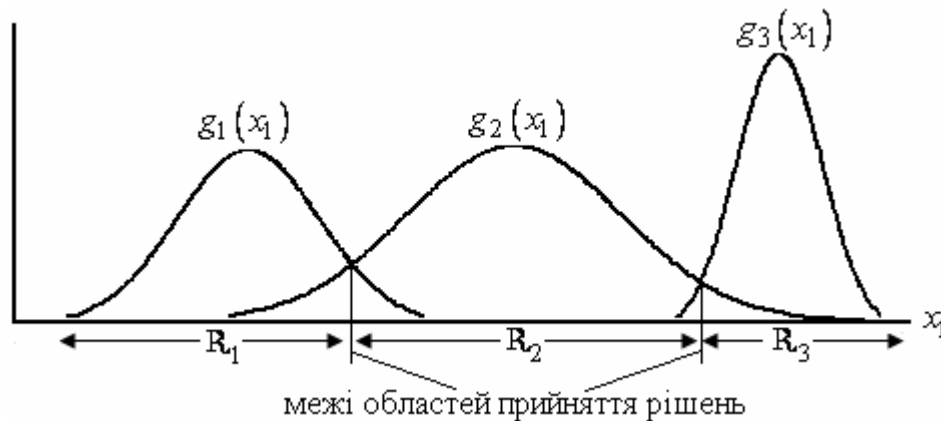


Рис. 5.7. Межі областей прийняття рішень під час застосування баєсівського вирішального правила ($K = 3, n = 1$)

Можна довести, що баєсівське вирішальне правило забезпечує мінімізацію імовірності помилкової класифікації, і у такому розумінні є оптимальне. Для цього розглянемо випадок двох класів (випадок багатьох класів можна розглядати за аналогією). Скористаємось правилом (5.3). Воно розбиває простір ознак на дві області R_1 та R_2 . Можливі **два типи помилок**:

- 1) об'єкт X потрапляє до області R_2 , коли в дійсності $X \in S_1$;
- 2) об'єкт X потрапляє до області R_1 , коли в дійсності $X \in S_2$.

Оскільки ці події несумісні та утворюють повну множину подій, то імовірність помилки класифікації дорівнює

$$\begin{aligned} P_{\text{пом}} &= P\{x_1 \in R_2, S_1\} + P\{x_1 \in R_1, S_2\} = \\ &= \int_{R_2} p_1 f_1(X) dX + \int_{R_1} p_2 f_2(X) dX = \end{aligned}$$

$$\begin{aligned}
&= p_1 \left(1 - \int_{R_1} f_1(X) dX \right) + \int_{R_1} p_2 f_2(X) dX = \\
&= p_1 - \int_{R_1} (p_1 f_1(X) - p_2 f_2(X)) dX.
\end{aligned}$$

Отже, мінімуму $P_{\text{пом}}$ досягається, коли

$$R_1 = \{X : p_1 f_1(X) > p_2 f_2(X)\},$$

а область R_2 утворюють усі інші точки.

Цей результат для одновимірного випадку ілюстровано (рис. 5.8, а, б). Очевидно, що в силу довільного вибору областей R_1 , R_2 імовірність помилки $P_{\text{пом}}$ виявляється не такою малою, як могла би бути. Зміщуючи межу області прийняття рішення вліво (рис. 5.8, а) або вправо (рис. 5.8, б), можна звести до нуля площу чорного трикутника і тим самим зменшити імовірність помилки. Отже, мінімуму $P_{\text{пом}}$ досягається, коли $R_1 = \{x_1 : g_1(x_1) > g_2(x_1)\}$, що відповідає результату застосування баєсівського вирішального правила.

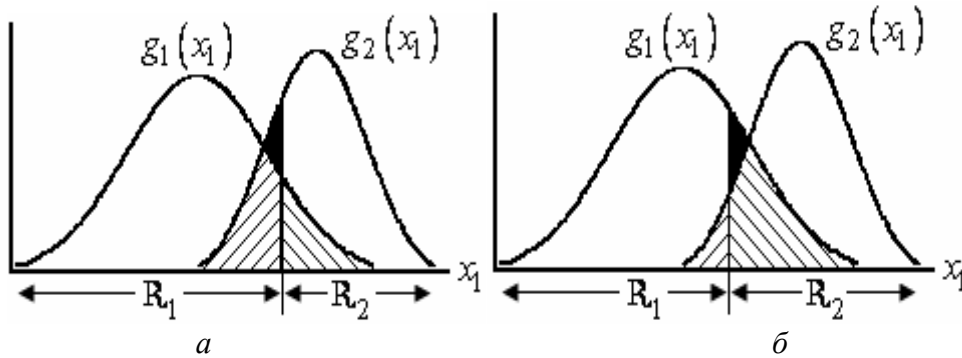


Рис. 5.8. Складові імовірності помилки ($K = 2, n = 1$)

Далі будемо вважати, що кожен клас S_j , $j = \overline{1, K}$, описується багатовимірним нормальним розподілом з функцією щільності

$$f_j(X) = \frac{1}{\sqrt{(2\pi)^n |\Sigma_j|}} \exp \left(-\frac{1}{2} (X - M_j) \Sigma_j^{-1} (X - M_j)^T \right), \quad (5.6)$$

де

$$M_j = E\{\tilde{\xi}_j\} = (m_1^{(j)}, \dots, m_n^{(j)})$$

– вектор математичного сподівання j -го класу;

$$\Sigma_j = DC\{\tilde{\xi}_j\} = \left\| \text{cov}_{k,v}^{(j)}; k, v = \overline{1, n} \right\|$$

– коваріаційна матриця j -го класу;

– матриця, обернена до Σ_j ;

$|\Sigma_j|$ – визначник матриці Σ_j .

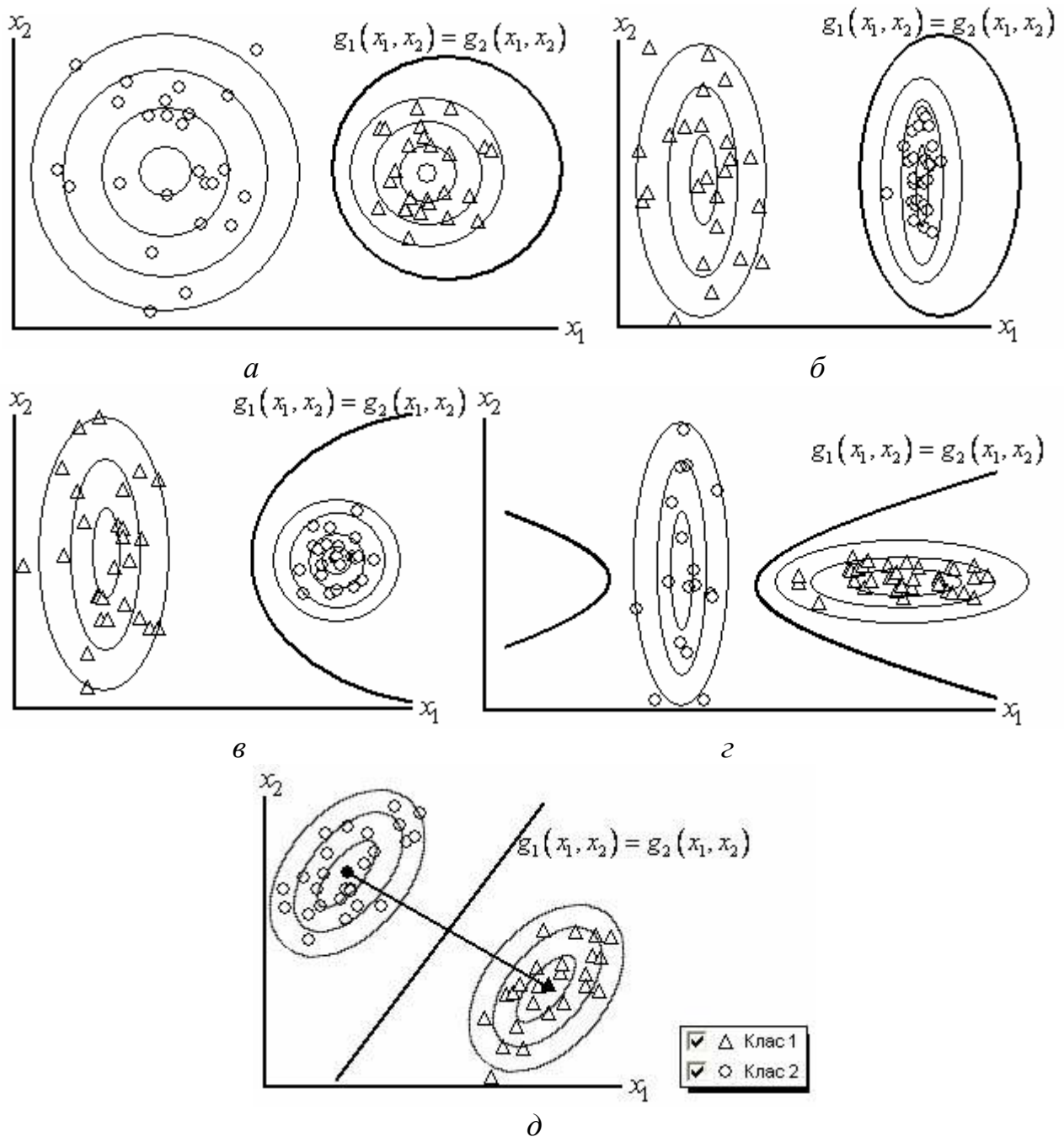


Рис. 5.9. Види меж областей прийняття рішень у випадку двовимірного нормального розподілу: а – коло; б – еліпс; в – парабола; г – гіпербола; д - пряма

Підставляючи (5.6) у (5.5) та відкидаючи спільні для усіх класів константи, одержуємо дискримінантні функції у вигляді

$$g_j(X) = \ln p_j - \frac{1}{2} \ln |\Sigma_j| - \frac{1}{2} (X - M_j) \Sigma_j^{-1} (X - M_j)^T, \quad j = \overline{1, K}. \quad (5.7)$$

Отже, оптимальні дискримінантні функції для нормальних об'єктів є квадратичними. Межі областей прийняття рішення можуть приймати форму гіперсфери, гіпереліпсоїда, гіперпараболоїда або різного роду гіперболоїда (рис. 5.9).

В окремому випадку, коли відомо, що коваріаційні матриці усіх класів однакові, дискримінантні функції можна переписати

$$g_j(X) = \log p_j + X \Sigma^{-1} M_j^T - \frac{1}{2} M_j \Sigma^{-1} M_j^T, \quad j = \overline{1, K}, \quad (5.8)$$

де

Σ – коваріаційна матриця, спільна для усіх класів.

З (5.8) видно, що даний випадок призводить до лінійних дискримінантних функцій (рис. 5.9, д). при цьому межа прийняття рішення проходить через точку, що з'єднує середні значення класів. Якщо у кожному класі ознаки будуть незалежні й з однаковими дисперсіями, тобто $\Sigma = \sigma^2 I$, то межа буде ще й ортогональна прямій, що з'єднує середні значення класів.

Слід зазначити, що баєсівське вирішальне правило є оптимальне, коли повністю відомі імовірнісні характеристики

$$f_j(X), \quad p_j, \quad j = \overline{1, K}.$$

На практиці доводиться мати справу з оцінками

$$\hat{f}_j(X), \quad \hat{p}_j, \quad j = \overline{1, K},$$

одержаними на основі масиву спостережень $\Omega_{n+1, N} = \{(x_{1,l}, \dots, x_{n,l}), y_l; l = \overline{1, N}\}$, де y_l вказує на належність l -го об'єкта до одного з K класів (можна вважати, що y_l приймає значення $1, 2, \dots, K$). В силу похибок оцінювання баєсівське вирішальне правило перестає бути оптимальним.

Оцінка \hat{p}_j визначається за масивом $\Omega_{n+1, N}$ із співвідношення

$$\hat{p}_j = \frac{N_j}{N}, \quad j = \overline{1, K},$$

де

N_j – кількість об'єктів j -го класу

$$N_j = \sum_{l=1}^N I_l,$$

$$I_l = \begin{cases} 1, & y_l = j, \\ 0, & y_l \neq j. \end{cases}$$

Оцінка вектора математичного сподівання є

$$\hat{M}_j = \hat{E}\{\xi_j\} = (\bar{x}_1^{(j)}, \dots, \bar{x}_n^{(j)}), \quad j = \overline{1, K},$$

де

$$\bar{x}_k^{(j)} = \frac{1}{N_j} \sum_{l=1}^N I_l x_{k,l}, \quad k = \overline{1, n},$$

а оцінка коваріаційної матриці

$$\hat{\Sigma}_j = \left\| \text{cov}_{k,v}^{(j)}; k, v = \overline{1, n} \right\|,$$

де $\text{cov}_{k,v}^{(j)}$ – оцінка коваріації поміж k -ою та v -ою ознаками у j -му класі

$$\text{cov}_{k,v}^{(j)} = \frac{1}{N_j - 1} \sum_{l=1}^N I_l (x_{k,l} - \bar{x}_k^{(j)}) (x_{v,l} - \bar{x}_v^{(j)}), \quad k = \overline{1, n}.$$

Коли коваріаційні матриці усіх класів рівні, тобто у разі прийняття гіпотези

$$H_0 : \Sigma_1 = \Sigma_2 = \dots = \Sigma_K$$

(див. підрозд. 4.2), оцінка загальної коваріаційної матриці може бути одержана як

$$\hat{\Sigma} = \frac{1}{N - K} \sum_{j=1}^K (N_j - 1) \hat{\Sigma}_j.$$

Тоді класифікація нового об'єкта X проводиться із використанням оцінок дискримінантних функцій (5.7) або (5.8):

$$X \in S_j,$$

якщо

$$\hat{g}_j(X) > \hat{g}_h(X)$$

для

$$\forall j, h = \overline{1, K}, \quad j \neq h.$$

Імовірність помилки такої класифікації $P_{\text{пом}}$ можна оцінити декількома способами. Згідно першого масив $\Omega_{n+1, N}$ розбивають на дві частини: контрольну та екзаменаційну. На основі контрольної частини будують вирішальне правило, а потім на екзаменаційній частині обчислюють імовірність помилки як

$$\hat{P}_{\text{пом}} = \frac{N_{\text{пом}}}{N_{\text{екзам}}},$$

де $N_{\text{пом}}$ – кількість помилок;
 $N_{\text{екзам}}$ – об'єм екзаменаційної вибірки.

Інший спосіб – ковзний контроль. Масив $\Omega_{n+1,N}$ розбивають на контрольну та екзаменаційну частини так, щоб контрольна вибірка містила $N-1$ об'єкт, а екзаменаційна – один. За допомогою контрольної вибірки будують вирішальне правило, за яким класифікують контрольний об'єкт. Процедура повторюють N разів, пред'являючи у якості контрольного різні об'єкти, після чого оцінюють імовірність помилки.

Контрольні запитання та завдання

1. Що таке неоднорідні дані?
2. У чому полягає задача розпізнавання?
3. Як визначають однорідність об'єктів спостережень?
4. Що називають метрикою відстані?
5. У який спосіб обчислюють відстань Мінковського? Які відстані є її окремими випадками?
6. Як пов'язані між собою відстані евклідова та Махаланобіса?
7. Як і з якою метою здійснюють стандартизацію даних?
8. Яка ідея алгомеративних та дивізімних ієрархічних методів кластерного аналізу?
9. Навести формулу Ланса–Уільямса.
10. У чому полягає агломеративний метод найближчого сусіда?
11. Які параметри формули Ланса–Уільямса для агломеративного методу найвіддаленішого сусіда?
12. Як обчислюють відстань між кластерами в агломеративному методі Уорда?
13. Прокоментувати переваги та недоліки алгомеративних методів кластерного аналізу.
14. Що таке дендрограма?
15. Для яких методів дендрограму можна побудувати без самоперетинів,

чим це пояснюється?

16.Що таке редуktivна властивість? Які метрики мають дану властивість?

17.Яка різниця між варіантами методу K -середніх Мак-Кіна та Болла–Холла?

18.Навести обчислювальну процедуру методу K -середніх у варіанті Мак-Кіна.

19.Охарактеризувати переваги та недоліки методу K -середніх.

20.Як оцінюють якість кластеризації?

21.У який спосіб оцінюють потрібну кількість кластерів?

22.Яке вирішальне правило називають оптимальним?

23.Сформулювати баєсівське вирішальне правило.

24.Довести, що баєсівське вирішальне правило є оптимальне.

25.Що таке дискримінантна функція?

26.Навести функцію щільності багатовимірного нормального розподілу.

27.Одержати дискримінантну функцію для випадку нормального розподілу.

28.Який вигляд має дискримінантна функція, коли кожен клас описується нормальним розподілом з однаковими коваріаційним матрицями?

29.Як оцінюють імовірність помилки класифікації?

РОЗДІЛ 6. ОСНОВИ АНАЛІЗУ ВИПАДКОВИХ ПРОЦЕСІВ ТА ЧАСОВИХ РЯДІВ

Поняття випадкового процесу узагальнює визначення випадкової величини на випадок, коли випадкова величина може змінюватися з часом.

Випадковим процесом $\xi(t)$ називають процес, значення якого за $\forall t \in T$, $t = t_0$ є випадкова величина $\xi(t_0)$. Формально, випадковий процес є функція від двох аргументів – часу t та елементарної події ω :

$$\xi(t) = \varphi(t, \omega), \quad \omega \in \Omega, \quad t \in T,$$

яка за фіксованого $t = t_0$ є випадкова величина $\xi(t_0)$.

Випадкову величину $\xi(t_0)$, у яку перетворюється випадковий процес за $t = t_0$, називають **перерізом випадкового процесу**, який відповідає даному значенню аргументу. Тобто на випадковий процес можна дивитися як на сукупність залежних від часу випадкових величин (перерізів випадкового процесу).

Конкретний вигляд, що приймає випадковий процес в результаті спостереження, називають його реалізацією (рис. 6.1). Реалізацію випадкового процесу $\xi(t)$ будемо позначати $x(t)$.

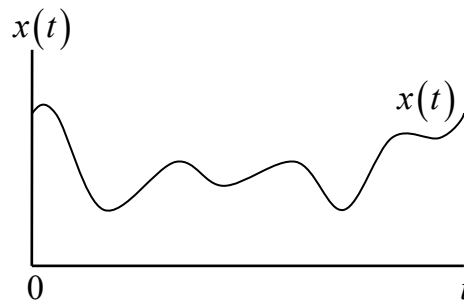


Рис. 6.1. Реалізація випадкового процесу

Частковим випадком випадкового процесу є **часовий ряд**. Для часових рядів, на відміну від даних про часовий переріз, сама послідовність спостережень несе в собі важливу інформацію. Головною метою аналізу часових рядів є прогнозування. Розгляд методів прогнозування виходить за межі даного підручника, проте в даному розділі подано важливі складові аналізу, що при належній реалізації забезпечують подальший якісний прогноз. Мова йде про тренд-аналіз та редагування.

6.1. Характеристики випадкового процесу

Розглянемо імовірнісні характеристики випадкового процесу $\xi(t)$. Зважаючи, що за будь-якого значення аргументу t , $\xi(t)$ є випадкова величина, яка має закон розподілу, визначають одновимірну функцію розподілу випадкового процесу

$$F(t, x) = P\{\xi(t) < x\}.$$

Функція $F(t, x)$ характеризує властивості конкретного перерізу процесу і може змінюватися при різних t . Тому більш повною характеристикою процесу є N -вимірна функція розподілу

$$F(t_1, \dots, t_N, x_1, \dots, x_N) = P\{\xi(t_1) < x_1, \dots, \xi(t_N) < x_N\},$$

що представляє собою закон розподілу системи N випадкових величин

$$\xi(t_1), \dots, \xi(t_N),$$

тобто N довільних перерізів випадкового процесу $\xi(t)$.

На практиці зазвичай обмежуються одно- та двовимірними законами розподілу. Існують певні класи процесів (наприклад, марківські та гаусівські), для яких такі закони є вичерпними характеристиками. Але, здебільшого, під час практичного дослідження відмовляються від законів розподілу випадкового процесу на користь його основних характеристик.

До **основних характеристик випадкового процесу** відносять математичне сподівання, дисперсію, коваріаційну та кореляційну функції. Під час їх розгляду будемо припускати, що переріз випадкового процесу $\xi(t)$ за конкретного t є неперервна випадкова величина з функцією щільності $f(t, x)$.

Математичне сподівання випадкового процесу $\xi(t)$ є функція $m(t)$, яка при будь-якому значенні аргументу t дорівнює математичному сподіванню відповідного перерізу процесу

$$m(t) = E\{\xi(t)\} = \int_{-\infty}^{+\infty} xf(t, x) dx.$$

Дисперсія випадкового процесу $\xi(t)$ – це функція $D(t)$, значення якої для кожного t дорівнює дисперсії відповідного перерізу

$$D(t) = D\{\xi(t)\} = E\{(\xi(t) - m(t))^2\} = \int_{-\infty}^{+\infty} (x - m(t))^2 f(t, x) dx.$$

Дисперсія характеризує розсіювання можливих реалізацій випадкового процесу відносно середнього.

Корінь квадратний із дисперсії називають **середньоквадратичним відхиленням випадкового процесу**

$$\sigma(t) = \sigma\{\xi(t)\} = \sqrt{D(t)},$$

розмірність функції $\sigma(t)$ така ж, як і у випадкового процесу $\xi(t)$.

Коваріаційна функція $K(t_1, t_2)$ **випадкового процесу** $\xi(t)$ – це функція, яка за кожної пари значень t_1, t_2 дорівнює коваріаційному моменту відповідних перерізів процесу

$$\begin{aligned} K(t_1, t_2) &= \text{cov}(t_1, t_2) = E\{(\xi(t_1) - m(t_1))(\xi(t_2) - m(t_2))\} = \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x_1 - m(t_1))(x_2 - m(t_2)) f(t_1, t_2, x_1, x_2) dx_1 dx_2. \end{aligned}$$

Вона характеризує ступінь лінійної залежності між перерізами $\xi(t_1)$ та $\xi(t_2)$, а також розсіювання цих перерізів відносно математичного сподівання $m(t)$. Коли $t_1 = t_2 = t$ коваріаційна функція перетворюється у дисперсію випадкового процесу:

$$K(t, t) = E\{(\xi(t) - m(t))^2\} = D(t).$$

Очевидно, що при близьких значеннях t_1 і t_2 величини $\xi(t_1)$ та $\xi(t_2)$ зв'язані тісною залежністю: якщо величина $\xi(t_1)$ прийняла деяке значення, то і величина $\xi(t_2)$ з великою імовірністю прийме близьке до нього значення (рис.6.2). У разі збільшення інтервалу між перерізами за t_1 і t_2 залежність між $\xi(t_1)$ та $\xi(t_2)$ зменшується (окрім часткових випадків періодичних процесів).

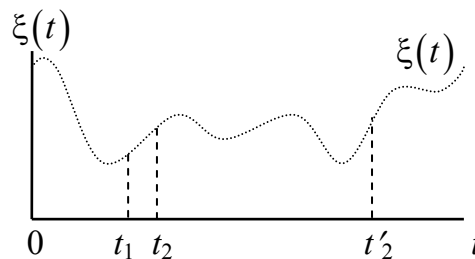


Рис. 6.2. Перерізи випадкового процесу

Кореляційна функція **випадкового процесу** являє собою нормовану коваріаційну функцію

$$R(t_1, t_2) = \frac{K(t_1, t_2)}{\sigma(t_1)\sigma(t_2)},$$

називається коефіцієнтом кореляції та характеризує ступінь лінійної залежності між перерізами $\xi(t_1)$ та $\xi(t_2)$.

Кореляційна функція має такі властивості:

- 1) у разі рівності аргументів $t_1 = t_2 = t$ дорівнює 1:

$$R(t, t) = 1;$$

- 2) симетрична відносно своїх аргументів:

$$R(t_1, t_2) = R(t_2, t_1);$$

- 3) за модулем не перевищує 1:

$$|R(t_1, t_2)| \leq 1.$$

Важливий клас випадкових процесів складають **стаціонарні** випадкові процеси, визначення яких дамо, базуючись на наведених вище характеристиках.

Для стаціонарних випадкових процесів характерна незалежність від початку відліку часу певних характеристик: законів розподілу, моментів тощо. Такий процес відбувається в приблизно однорідних умовах та має вигляд неперервних випадкових коливань навколо деякого середнього значення (рис. 6.3, *а*). При цьому ні середня амплітуда, ні його частота з часом суттєво не змінюються. Кожен стаціонарний процес можна розглядати як такий, що продовжується у часі нескінченно довго. Для його дослідження у якості початку відліку можна обрати довільний момент часу, і на будь-якому проміжку часу одержимо однакові характеристики.

Нестаціонарний процес характеризується наявністю певної тенденції розвитку у часі (рис. 6.3, *б*). Характеристики такого процесу залежать від початку відліку та від часу.

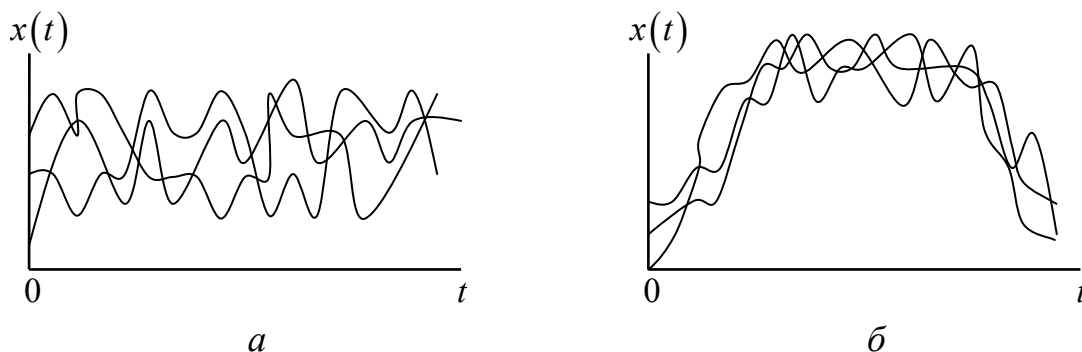


Рис. 6.3. Реалізації стаціонарного (*а*) та нестаціонарного (*б*) випадкових процесів

Зазначимо, що далеко не всі нестаціонарні випадкові процеси суттєво нестаціонарні протягом усього розвитку. Існують нестаціонарні процеси, які на певних відрізках часу та з певним наближенням можна прийняти за стаціонарні.

Зупинимось на понятті стаціонарності детальніше і дамо більш формальне визначення. Зазначимо, що в залежності від того, які характеристики вважають незалежними від початку відліку часу, розрізняють стаціонарність у вузькому або широкому розумінні.

Випадковий процес $\xi(t)$ є стаціонарний у **вузькому розумінні**, якщо його N -вимірна функція розподілу не змінюється під час зсуву усіх його часових аргументів на однакову величину τ

$$F(t_1, \dots, t_N, x_1, \dots, x_N) = F(t_1 + \tau, \dots, t_N + \tau, x_1, \dots, x_N).$$

Випадковий процес $\xi(t)$ є стаціонарний у **широкому розумінні**, якщо його математичне сподівання постійне, а коваріаційна функція залежить не від положення першого аргументу t_1 , а від проміжку τ між першим та другим аргументами:

$$m(t) = \text{const},$$

$$K(t_1, t_1 + \tau) = k(\tau). \quad (6.1)$$

Відзначимо, що функцію $k(\tau)$ називають автоковаріаційною, а величину τ називають зсувом або запізненням.

Вочевидь, якщо випадковий процес стаціонарний у вузькому розумінні, то він стаціонарний і у широкому розумінні. Обернене твердження, взагалі кажучи, не вірне. Тобто стаціонарний процес у широкому розумінні (СПШР) є підмножина стаціонарного процесу у вузькому розумінні (СПВР) (рис. 6.4).

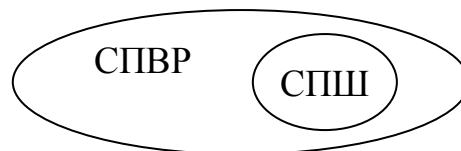


Рис. 6.4. Співвідношення між випадковими процесами у вузькому та широкому розуміннях

У будь-якому розумінні стаціонарний випадковий процес має постійні математичне сподівання та дисперсію, а його коваріаційна функція є функція лише одного аргументу – зсуву τ .

Зауваження 6.1. Звернемо увагу, що умова постійності математичного сподівання стаціонарного випадкового процесу не дуже суттєва. Завжди є можливість перейти від процесу $\xi(t)$ до $\eta(t) = \xi(t) - m_\xi(t)$, для якого математичне сподівання тотожно дорівнює нулю. Тому якщо процес нестаціонарний лише за рахунок змінного математичного сподівання, це не заважає вивчати його як стаціонарний.

Коваріаційна функція стаціонарного процесу має властивості:

1) парна функція зсуву τ між двома перерізами процесу:

$$k(\tau) = k(-\tau);$$

2) $k(0) = D$, оскільки

$$D = K(t, t) = k(t - t) = k(0);$$

3) $k(0) \geq 0$, оскільки $D \geq 0$;

4) $|k(\tau)| \leq k(0)$;

5) додатна визначена, тобто

$$\int_{(B)} \int_{(B)} k(t - t') \varphi(t) \varphi(t') dt dt' \geq 0,$$

де $\varphi(t)$ – будь-яка функція аргументу t ;

B – будь-яка область зміни аргументу t .

Як і коваріаційна, кореляційна функція стаціонарного випадкового процесу залежить лише від зсуву між перерізами процесу

$$r(t_1, t_1 + \tau) = r(\tau) = \frac{k(\tau)}{D} = \frac{k(\tau)}{k(0)}$$

і має практично ті самі властивості:

1) $r(\tau) = r(-\tau)$;

2) $r(0) = 1$;

3) $|r(\tau)| \leq 1$;

4) $\int_{(B)} \int_{(B)} r(t - t') \varphi(t) \varphi(t') dt dt' \geq 0,$

саму ж функцію $r(\tau)$ називають автокореляційною.

Стаціонарні випадкові процеси бувають **ергодичні та неергодичні**. Ергодичними називають процеси, що мають властивість ергодичності.

Ергодична властивість полягає у тому, що кожна окрема реалізація випадкового процесу є «повноважним» представником усієї сукупності реалізацій; одна реалізація достатньої довжини може замінити під час автоматизованої обробки будь-яку кількість реалізацій тієї ж довжини.

Про ергодичність або неергодичність випадкового процесу може свідчити вигляд автоковаріаційної (автокореляційної) функції. Якщо автоковаріаційна

функція стаціонарного процесу зі збільшенням τ прямує до деякого сталого значення, то процес не є ергодичний. Її прямування до нуля за $\tau \rightarrow \infty$ говорить на користь ергодичного процесу. Якщо стаціонарний випадковий процес має властивість ергодичності, то його характеристики (математичне сподівання, дисперсія, коваріаційна, кореляційна функції) можна визначити за однією достатньо довгою реалізацією як середні за часом.

Математичне сподівання ергодичного випадкового процесу визначається за такою формулою:

$$m = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T x(t) dt \approx \frac{1}{T} \int_0^T x(t) dt.$$

Дисперсія ергодичного випадкового процесу визначається як

$$D = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T (x(t) - m)^2 dt \approx \frac{1}{T} \int_0^T (x(t) - m)^2 dt.$$

Автоковаріаційну функцію ергодичного випадкового процесу знаходять як математичне сподівання випадкових величин $\xi(t) - m$ та $\xi(t + \tau) - m$:

$$\begin{aligned} k(\tau) &= \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T (x(t) - m)(x(t + \tau) - m) dt \approx \\ &\approx \frac{1}{T - \tau} \int_0^{T - \tau} (x(t) - m)(x(t + \tau) - m) dt. \end{aligned}$$

Відповідно кореляційна функція визначається як

$$r(\tau) = \frac{k(\tau)}{D}.$$

6.2. Визначення та оцінка спектральної щільності

Математичною основою, яка пов'язує часовий ряд з його представленням в частотній області, є перетворення Фур'є. В рядах Фур'є використовуються синусоїди та косинусоїди з кратними частотами, що отримали назву гармоніки. Знаходження гармонік сигналу називається спектральним аналізом. В його результаті представлення сигналу переводиться в частотне представлення, тобто він представляється сукупністю гармонік з амплітудами та фазами. Задача перетворення сигналу, представленого множиною гармонік, в часову область називається гармонічним синтезом.

Якщо коливальний процес представляється у вигляді суми гармонічних коливань різної частоти (гармонік), то спектром коливального процесу назива-

ється функція, що описує розподіл амплітуд за різними частотами. Спектр показує якого роду коливання переважають у процесі, що вивчається, якою є його внутрішня структура. Для випадкового процесу амплітуди коливань будуть випадковими величинами. Таким чином, спектр стаціонарного випадкового процесу буде описувати розподіл дисперсій за різними частотами.

З точки зору спектрального аналізу найбільш важливою характеристикою випадкового процесу є **спектральна щільність** (спектр потужності), яка може бути визначеною за формулою

$$S(f) = \frac{2}{\pi} \int_0^{\infty} k(\tau) \cos(2\pi f \tau) d\tau,$$

де $k(\tau)$ – автоковаріаційна функція (6.1).

Крива $S(f)$ зображує щільність розподілу дисперсій за частотами неперервного спектру, а площа, обмежена цією кривою дорівнює дисперсії стаціонарного випадкового процесу $x(t)$.

Автоковаріаційна функція та спектральна щільність можуть мати вираз одна через одну за допомогою перетворення Фур'є подібно до того, як ряд Фур'є виражає функцію, що розкладається через коефіцієнти ряду, які в свою чергу виражаються через коефіцієнти функції, що розкладається. Тобто автоковаріаційна функція може бути представлена у вигляді

$$k(\tau) = \frac{2}{\pi} \int_0^{\infty} S(f) \cos(2\pi f \tau) df.$$

На практиці часто замість спектральної щільності користуються нормованою спектральною щільністю

$$s(f) = S(f)/D,$$

де D – дисперсія випадкового процесу.

Автокореляційна функція $r(\tau)$ та нормована спектральна щільність також пов'язані перетворенням Фур'є

$$r(\tau) = \int_0^{\infty} s(f) \cos(2\pi f \tau) df, \quad (6.2)$$

$$s(f) = \int_0^{\infty} r(\tau) \cos(2\pi f \tau) d\tau. \quad (6.3)$$

Поклавши в рівнянні (6.2) $\tau = 0$ та враховуючи, що $r(0) = 1$, маємо

$$\int_0^{\infty} s(f) df = 1,$$

тобто повна площа, яка обмежена графіком нормованої спектральної щільності, дорівнює одиниці.

На практиці реальний стаціонарний процес фіксують на короткому проміжку $[0; T]$, тобто поза цим проміжком $k(\tau)$ (а отже й $r(\tau)$) дорівнює нулю. У цьому випадку оцінка спектральної щільності

$$\hat{S}(f) = \frac{2}{\pi} \int_0^T k(\tau) \cos(2\pi f \tau) d\tau$$

не є спроможна. Тому виникає задача знаходження такої функції $\lambda(\tau)$ (яку називають спектральним вікном), щоб забезпечити

$$\min E \left\{ \int_0^{\infty} (\hat{S}(f) - S(f))^2 df \right\}.$$

Тим самим розв'язується задача згладжування спектральних функцій на кінцевому відрізку часу. Нижче наведені спектральні вікна, що найчастіше використовуються на практиці (табл. 6.1).

Таблиця 6.1

Кореляційні вікна

Назва вікна	Функція вікна
Функція Бартлета	$\lambda(\tau) = \begin{cases} 1 - \tau /\tau_m, & \tau \leq \tau_m, \\ 0, & \tau > \tau_m. \end{cases}$
Функція Тьюкі	$\lambda(\tau) = \begin{cases} 1 - 2a + 2a \cos(\pi\tau/\tau_m), & \tau \leq \tau_m, \\ 0, & \tau > \tau_m, \end{cases}$ де $a \approx 0,25$
Функція Парзена	$\lambda(\tau) = \begin{cases} 1 - (\tau /\tau_m)^q, & \tau \leq \tau_m, \\ 0, & \tau > \tau_m, \end{cases}$ де $q \approx 2$
Функція Лемінга	$\lambda(\tau) = \begin{cases} 0,54 + 0,46 \cos(\pi\tau/\tau_m), & \tau \leq \tau_m, \\ 0, & \tau > \tau_m. \end{cases}$

За масивом $\Omega_{1,N} = \{x(t_i); i = \overline{1, N}\}$, що є результатом спостереження над ергодичним випадковим процесом з інтервалом h , одержують згладжену вибірку **оцінку спектральної щільності** у вигляді

$$\hat{S}(f) = 2h \left(\hat{k}(0) + 2 \sum_{\tau=1}^{L-1} \hat{k}(\tau) \lambda(\tau) \cos(2\pi f \tau) \right), 0 \leq f \leq \frac{1}{2h},$$

де $L = \frac{\tau_m}{h}$;

$\lambda(\tau)$ – кореляційне вікно з точкою відсікання τ_m ;

$\hat{k}(\tau), \tau = \overline{0, L-1}$ – оцінка автоковаріаційної функції (6.1).

Згладжена вибірка оцінка нормованої спектральної щільності обчислюється за формулою

$$\hat{s}(f) = 2h \left[1 + 2 \sum_{\tau=1}^{L-1} \hat{r}(\tau) \lambda(\tau) \cos(2\pi f \tau) \right], 0 \leq f \leq \frac{1}{2h},$$

де $\hat{r}(\tau), \tau = \overline{0, L-1}$ – оцінка автокореляційної функції.

Оцінювання автоковаріаційної функції $k(\tau)$ можна провести за масивом $\Omega_{1,N}$ у такий спосіб. Величину τ вважають цілочисловою, що змінюється від 0 до $(L-1) \leq N$. Тоді обчислюють коваріацію між елементами масиву $\Omega_{1,N}$, що розташовані на відстані τ одиниць одне від одного

$$\hat{k}(0) = \text{cov} \{ (x(t_1), x(t_2), \dots, x(t_N)), (x(t_1), x(t_2), \dots, x(t_N)) \},$$

$$\hat{k}(1) = \text{cov} \{ (x(t_1), x(t_2), \dots, x(t_{N-1})), (x(t_2), x(t_3), \dots, x(t_N)) \},$$

...

$$\hat{k}(L-1) = \text{cov} \{ (x(t_1), x(t_2), \dots, x(t_{N-L+1})), (x(t_L), x(t_{L+1}), \dots, x(t_N)) \},$$

тобто

$$\hat{k}(\tau) = \frac{1}{N-\tau} \sum_{i=1}^{N-\tau} (x(t_i) - \hat{m}_1)(x(t_{i+\tau}) - \hat{m}_2), \quad \tau = \overline{0, L-1},$$

де

$$\hat{m}_1 = \frac{1}{N-\tau} \sum_{i=1}^{N-\tau} x(t_i),$$

$$\hat{m}_2 = \frac{1}{N-\tau} \sum_{i=1}^{N-\tau} x(t_{i+\tau}).$$

Відповідно оцінка автокореляційної функції є

$$\hat{r}(\tau) = \frac{\hat{k}(\tau)}{\sqrt{\hat{D}_1 \hat{D}_2}}, \quad \tau = \overline{0, L-1},$$

де

$$\hat{D}_1 = \frac{1}{N-\tau-1} \sum_{i=1}^{N-\tau} (x(t_i) - \hat{m}_1)^2,$$

$$\hat{D}_2 = \frac{1}{N-\tau-1} \sum_{i=1}^{N-\tau} (x(t_{i+\tau}) - \hat{m}_2)^2.$$

Частота $\frac{1}{2h}$ називається найквістовою, це найвища з частот, яку можна виявити за даними, відлік яких відбувається через h секунд.

На рис. 6.5 наведена нормована спектральна щільність для сигналу заданого дискретно $\{\sin(10t); t = \overline{1, 100}\}$ з кореляційним вікном Тьюкі з $a = 0,23$ з різною шириною вікна.

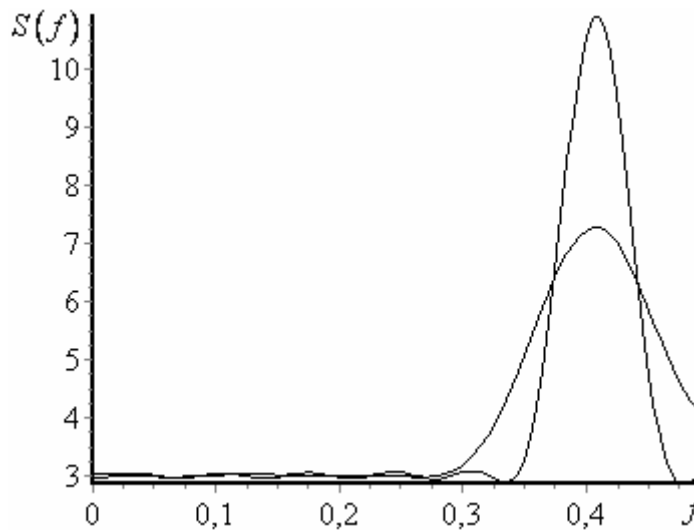


Рис. 6.5. Спектральна щільність з $\tau_m = 8$ та $\tau_m = 15$

Значення згладженої оцінки спектральної щільності залежить як від вибору $\lambda(\tau)$, так і від точки зрізу τ_m . При малому значенні τ_m смуга спектру буде широкою, що призводить до згладжування піків спектральної щільності. Зі збільшенням τ_m будуть проявлятися тенденції до зашумлення процесу.

6.3. Первинний аналіз часових рядів та випадкових процесів

Автоматизована обробка спостережень над випадковим процесом має передбачати імовірісно-статистичне редагування зібраної інформації. Задачами такого редагування є оцінка стаціонарності процесу; виявлення та вилучення тренду з попереднім згладжуванням даних; вилучення аномальних спостережень.

Розглянемо обчислювальні процедури вирішення даних задач. Будемо припускати, що досліджуваний випадковий процес заданий однією реалізацією, яка є його повноважний представник, а результати спостереження над процесом подані масивом $\Omega_{1,N} = \{x(t_i), i = \overline{1, N}\}$. Нехай спостереження над випадковим процесом проводилися рівномірно з кроком $h = t_{i+1} - t_i$. Тоді можна вважати, що $t_i = ih$, а у простішому випадку $t_i = i$, та записувати масив у вигляді $\Omega_{1,N} = \{x_i, i = \overline{1, N}\}$.

6.3.1. Вилучення аномальних спостережень

Якщо має місце стаціонарний процес, або процес, що зводиться до стаціонарного шляхом вилучення тренду, виявлення аномальних значень здійснюється за наведеними нижче процедурами.

Процедура 6.1. За масивом $\Omega_{1,N} = \{x_i; i = \overline{1, N}\}$ для кожного i обчислюють

$$\bar{x}_i = \frac{1}{N} \sum_{j=1}^i x_j, \quad S_i^2 = \frac{1}{N-1} \sum_{j=1}^i (x_j - \bar{x}_i)^2.$$

Тоді для $(i+1)$ -го вимірювання перевіряють умову

$$\bar{x}_i - kS_i < x_{i+1} < \bar{x}_i + kS_i,$$

Якщо вказана умова не виконується, то x_{i+1} вважають аномальним значенням і замінюють на

$$\tilde{x}_{i+1} = 2(x_i - x_{i-1}),$$

отримане, по суті, шляхом лінійної екстраполяції. Рекомендують обирати $k = 3 \div 9$ (краще 6).

Дана процедура має певний недолік. Послідовна екстраполяція для підряд розташованих аномальних значень може привести до того, що сформульованій умові не будуть задовольняти «добрі» дані, які в решті решт з'являться.

Поряд із наведеною, можуть бути реалізовані більш точні процедури, які аналогічні тим, що використовуються для випадку нормально розподілених випадкових величин.

Процедура 6.2. Значення x_i вважають аномальним (грубим, x_{gp}), якщо

$$x_i \leq a \quad \text{або} \quad x_i \geq b.$$

Значення a та b визначають за співвідношеннями:

$$a = \bar{x} - t_1 S, \quad b = \bar{x} + t_1 S, \quad \text{якщо } |\bar{A}| < 0,2,$$

$$a = \bar{x} - t_2 S, \quad b = \bar{x} + t_1 S, \quad \text{якщо } \bar{A} < -0,2,$$

$$a = \bar{x} - t_1 S, \quad b = \bar{x} + t_2 S, \quad \text{якщо } \bar{A} > 0,2$$

де \bar{x} та S^2 – середнє та дисперсія вибірки; \bar{A} – незсунений коефіцієнт асиметрії; \bar{E} – незсунений коефіцієнт ексцесу (розд.1);

$$t_1 = 2 + 0,2 \lg(0,04N); \quad t_2 = \left(19(\bar{E} + 2)^{0,5} + 1\right)^{0,5}.$$

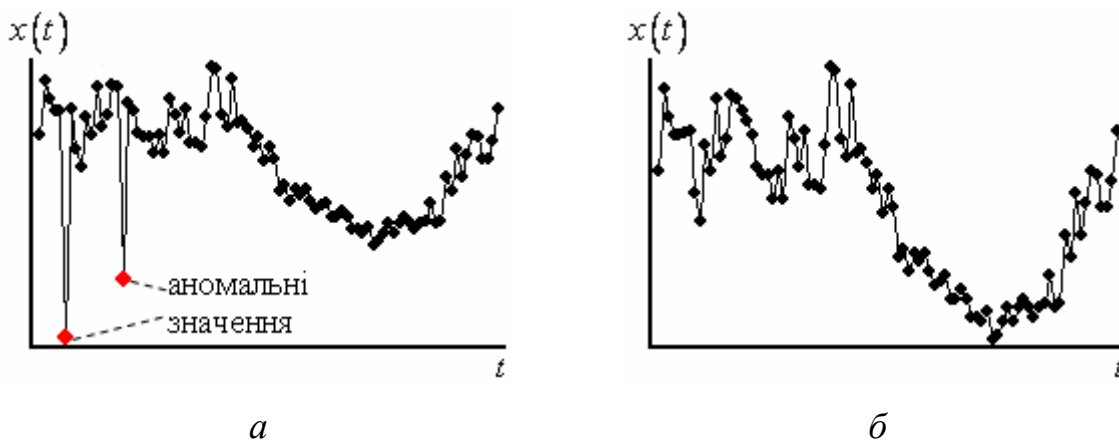


Рис. 6.6. Реалізація випадкового процесу до (а) та після (б) вилучення аномальних значень

На графіку приклад застосування наведених процедур (рис. 6.6). При цьому аномальні значення замінені на середнє арифметичне сусідніх значень.

6.3.2. Ідентифікація тренду процесу

Задача дослідження стаціонарності процесу може бути визначена відносно функції математичного сподівання $m(t)$, яка функціонально визначає тренд. **Тренд** – це функція, що показує глобальні зміни у досліджуваному процесі. Коли $m(t) = m = \text{const}$, говорять про відсутність тренду (рис.6.7, а). Інакше процес вважають нестационарним за рахунок змінного математичного сподівання (рис. 6.7, б, в). На разі виявлення тренду $m(t)$, процес може бути зведений до стаціонарного шляхом вилучення тренду.

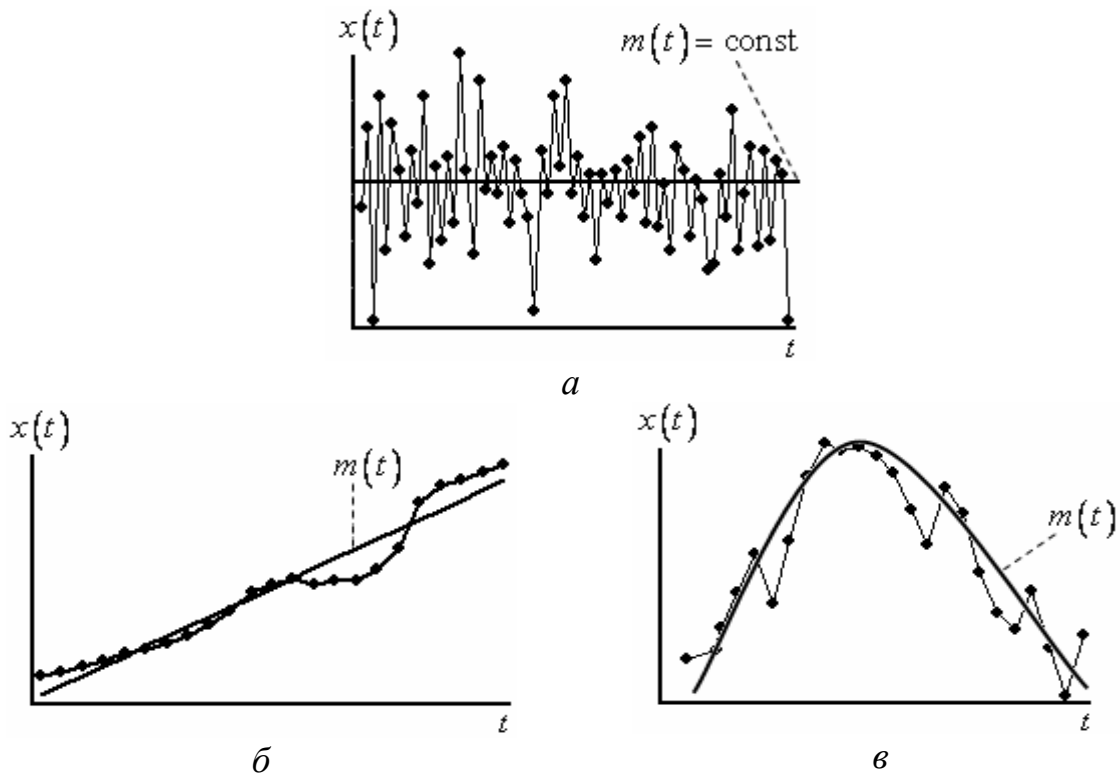


Рис. 6.7. Графіки тренду випадкового процесу:
a – тренд відсутній; *б* – лінійний тренд; *в* – нелінійний тренд

Отже, задачу дослідження стаціонарності процесу можна сформулювати як задачу тренд-аналізу, складовими якого є ідентифікація тренду, згладжування даних та видалення тренду.

Наведемо непараметричні процедури ідентифікації тренду за масивом $\Omega_{1,N} = \{x_i; i = \overline{1, N}\}$. В їх основі лежить той факт, що для стаціонарного процесу спостереження x_i статистично незалежні. Інакше кажучи, $\Omega_{1,N}$ є випадкова вибірка.

Критерій екстремальних точок є найбільш простий, він полягає у підрахунку «піків» та «ям» у реалізації випадкового процесу (рис. 6.8). «Піком» називається значення, яке більше двох сусідніх. Якщо є два та більше розташовані підряд рівні значення, що більші за попередні та наступні, то їх розглядають як один «пік». Так само, «яма» – це значення, що менше двох сусідніх. Максимальна кількість «піків» та «ям», інакше кажучи, екстремальних точок, на одиницю менша кількості інтервалів монотонності.

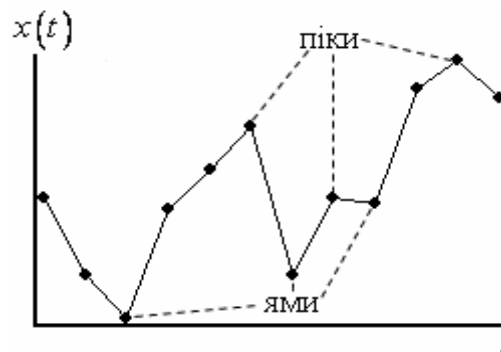


Рис. 6.8. Ілюстрація «піків» та «ям» у реалізації випадкового процесу

Для визначення екстремальної точки потрібні три послідовні спостереження. Коли вибірка випадкова, ці три величини могли б з'явитися в будь-якому порядку: всього шість варіантів. Тільки у чотирьох варіантах є екстремальна точка (коли найбільше і найменше значення в середині). Отже, імовірність появи екстремальної точки у випадку трьох спостережень є $2/3$.

Для $\{x_i; i = \overline{1, N}\}$ визначають індикатор y_i , як

$$y_i = \begin{cases} 1, & \text{якщо } x_{i-1} < x_i > x_{i+1} \text{ або } x_{i-1} > x_i < x_{i+1}, \\ 0, & \text{у протилежному випадку,} \end{cases} \quad i = \overline{2, N-1}.$$

Тоді кількість екстремальних точок дорівнює

$$p = \sum_{i=1}^{N-2} y_i.$$

Величина p має математичне сподівання та дисперсію

$$E\{p\} = \sum_{i=1}^{N-2} E\{y_i\} = \frac{2}{3}(N-2),$$

$$D\{p\} = E\{p^2\} - E^2\{p\} = \frac{1}{90}(16N-29),$$

де

$$E\{p^2\} = \frac{1}{90}(40N^2 - 144N + 131),$$

отже, p швидко прагне до нормальності при збільшенні N . Надалі обчислюють статистику

$$S = \frac{p - E\{p\}}{\sqrt{D\{p\}}},$$

значення якої порівнюють із квантилем нормального розподілу $u_{\alpha/2}$. При виконанні умови

$$|S| < u_{\alpha/2},$$

процес вважають стаціонарним. Якщо $S < -u_{\alpha/2}$, має місце тренд до спадання, за $S > u_{\alpha/2}$ – присутній тренд до зростання.

Критерій знаків для визначення наявності тренду полягає у підрахунку кількості позитивних різниць першого порядку в $\Omega_{1,N}$, інакше кажучи, кількості точок зростання значень процесу. При цьому ігнорують точки, де не відбувається ні збільшення, ні зменшення значень. Для масиву з N елементів

отримують $(N-1)$ -у різницю і визначають індикатор

$$y_i = \begin{cases} 1, & x_{i+1} \geq x_i, \\ 0, & x_{i+1} < x_i. \end{cases}$$

Кількість точок зростання дорівнює величині

$$c = \sum_{i=1}^{N-1} y_i,$$

для якої

$$E\{c\} = (N-1)E\{y_i\} = \frac{1}{2}(N-1),$$

$$E\{c^2\} = \frac{1}{2}(N-1) + \frac{1}{3}(N-2) + \frac{1}{4}(N-2)(N-3),$$

$$D\{c\} = \frac{1}{2}(N-1) + \frac{1}{3}(N-2) + \frac{1}{4}(N-2)(N-3) - \frac{1}{4}(N-1)^2 = \frac{1}{12}(N+1).$$

Розподіл величини c досить швидко збігається до нормального. Тому можна порівнювати статистику

$$S = \frac{c - E\{c\}}{\sqrt{D\{c\}}}$$

із квантилем нормального розподілу $u_{\alpha/2}$. За $|S| < u_{\alpha/2}$ процес є стаціонарний. Коли $S < -u_{\alpha/2}$, процес має тенденцію до спадання, а якщо $S > u_{\alpha/2}$, присутня тенденція до зростання.

В критерії Манна на основі $\{x_i; i = \overline{1, N}\}$ обчислюють величину T , яка дорівнює кількості випадків, коли $x_i < x_j$ при $i < j$. Якщо значення рівні, то до величини T додають 0,5. Величина T має розподіл, близький до нормального з математичним сподіванням та дисперсією

$$E\{T\} = \frac{1}{4}N(N-1),$$

$$D\{T\} = \frac{1}{72}(2N+5)(N-1)N.$$

Тоді визначають статистику

$$u = \frac{T + 0,5 - E\{T\}}{\sqrt{D\{T\}}}.$$

Для заданого критичного рівня α , якщо $|u| \leq u_{\alpha/2}$, то говорять, що досліджуваний процес стаціонарний. Інакше існує тенденція зміни процесу. Коли $u > u_{\alpha/2}$, є тенденція до збільшення, а за $u < -u_{\alpha/2}$ – до зменшення.

Критерій серій передбачає побудову бінарного ряду спостережень y , що приймають значення

$$y_i = \begin{cases} 1, & x'_i \geq x'_m, \\ -1, & x'_i < x'_m, \end{cases}$$

де x'_m – медіана відсортованого ряду $\{x'_i; i = \overline{1, N}\}$

$$x'_m = \begin{cases} x'_{(N+1)/2}, & \text{коли } N - \text{ непарне,} \\ \frac{1}{2}(x'_{N/2} + x'_{N/2+1}), & \text{коли } N - \text{ парне.} \end{cases}$$

Сформований ряд $\{y_i; i = \overline{1, N}\}$ характеризується послідовностями серій. Серія є сукупність розташованих підряд «1» або «-1». Позначимо $v(N)$ – загальну кількість серій в ряді, а $d(N)$ – довжину найбільшої серії. Тоді гіпотеза про стаціонарність процесу приймається, якщо одночасно виконуються умови

$$v(N) > \left\lceil \frac{1}{2}(N+1-1,96\sqrt{N-1}) \right\rceil,$$

$$d(N) < \lceil 3,3 \cdot \lg(N+1) \rceil,$$

де $\lceil \bullet \rceil$ – позначає цілу частину.

У протилежному разі робиться висновок про наявність залежності між спостереженнями й існування тренду.

Критерій «зростаючих» і «спадаючих» серій базується на дослідженні ряду з «1» і «-1», утвореного за правилом

$$y_i = \begin{cases} 1, & x_{i+1} - x_i \geq 0, \\ -1, & x_{i+1} - x_i < 0. \end{cases}$$

Статистики $v(N)$ і $d(N)$ обчислюються аналогічно попередній процедурі, але мають задовольняти умовам

$$v(N) > \left\lceil \frac{1}{3}(2N-1) - 1,96\sqrt{\frac{1}{90}(16N-29)} \right\rceil,$$

$$d(N) > d_0(N),$$

де

$$d_0(N) = \begin{cases} 5, & N \leq 26, \\ 6, & 26 < N \leq 153, \\ 7, & N > 153. \end{cases}$$

Ранговий критерій базується на обчисленні кількості випадків p , у яких $x_j > x_i$ при $j > i$. Встановлено, що p зв'язано простим співвідношенням із коефіцієнтом рангової кореляції Кендалла

$$\rho = \frac{4p}{N(N-1)} - 1.$$

Статистична характеристика

$$z = \frac{3\rho\sqrt{N(N-1)}}{\sqrt{2(2N+5)}}$$

коефіцієнта ρ має близький до нормального розподіл. За заданого рівня значущості α виконання нерівності $|z| \leq u_{\alpha/2}$ свідчить про відсутність тренду процесу.

Критерій Аббе є критерій квадратів послідовних різниць і вимагає підрахування значення

$$\gamma = \frac{q^2}{2s^2},$$

де

$$q^2 = \frac{1}{(N-1)} \sum_{i=1}^{N-1} (x_i - x_{i+1})^2;$$

$$s^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2.$$

За $N \rightarrow \infty$ статистика

$$u = (\gamma - 1) \sqrt{\frac{N^2 - 1}{N - 2}}$$

асимптотично має нормальний розподіл. Тому за виконання $|u| \leq u_{\alpha/2}$, гіпотеза про стохастичну незалежність результатів спостережень приймається.

6.3.3. Згладжування даних

Якщо процедури ідентифікації тренду показали, що має місце нестационарний процес, то наступний аналіз пов'язаний з побудовою гладкої функції тренду. Враховуючи, що тренд практично «зашумлено» неідентифікованими похибками, виникає задача згладжування початкових даних з наступним описанням тренду.

Процедура згладжування може бути реалізована шляхом фільтрації вихідного масиву даних. Поряд з цифровими фільтрами, що застосовуються при відомій природі шуму, реалізують обчислювальні процедури згладжування. До найбільш поширених у використанні відносяться **процедури ковзного середнього і медіанного згладжування**.

Розглянемо процедуру ковзного середнього, яка базується на методі найменших квадратів.

Для методу найменших квадратів обирають непарну кількість точок ковзання $k = 2m + 1$. Тоді знаходження згладженого значення точки $x_i = x(t_i)$ за його допомогою передбачає залучення k точок: $x_{i-m}, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_{i+m}$. З метою спрощення обчислювального процесу вводять нову систему координат: індекси $i-m, \dots, i-1, i, i+1, \dots, i+m$ заміняють на $-m, \dots, -1, 0, 1, \dots, m$ (рис. 6.9).

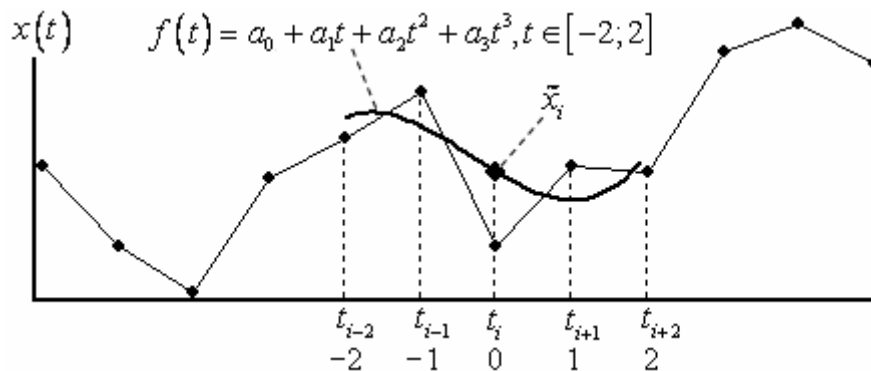


Рис. 6.9. Ілюстрація застосування методу найменших квадратів ($k = 5$)

На кожному проміжку $[-m; m]$ початкові значення згладжують за допомогою полінома порядку p ($p < m$)

$$f(t) = \sum_{j=0}^p a_j t^j, \quad t \in [-m; m],$$

оцінки $\hat{\Theta} = \{\hat{a}_0, \hat{a}_1, \dots, \hat{a}_p\}$ якого визначають із умови мінімуму функціонала S^2 :

$$S^2 = \sum_{t=-m}^m \left(x_{i+t} - \sum_{j=0}^p a_j t^j \right)^2, \quad (6.4)$$

яка еквівалентна розв'язуванню системи лінійних рівнянь

$$\frac{\partial S^2}{\partial a_j} = 0, \quad j = \overline{0, p}. \quad (6.5)$$

Як згладжене \tilde{x}_i приймають значення поліному у точці $t = 0$, тобто $\tilde{x}_i = \hat{a}_0$.

На практиці у якості функції $f(t)$ реалізують поліноми порядку $p \leq 5$. Розглянемо поліном третього порядку

$$f(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3, \quad t \in [-m; m]. \quad (6.6)$$

Для нього мінімальна необхідна кількість точок ковзання дорівнює $k = 5$. За даного значення величина (6.4) приймає вигляд:

$$S^2 = \sum_{t=-2}^2 \left(x_{i+t} - a_0 - a_1 t - a_2 t^2 - a_3 t^3 \right)^2,$$

а система лінійних алгебраїчних рівнянь (6.5), з урахуванням того, що

$$\begin{aligned} \sum_{i=-2}^2 t &= 0, & \sum_{i=-2}^2 t^2 &= 10h^2, & \sum_{i=-2}^2 t^3 &= 0, \\ \sum_{i=-2}^2 t^4 &= 34, & \sum_{i=-2}^2 t^5 &= 0, & \sum_{i=-2}^2 t^6 &= 130, \end{aligned}$$

набуває вигляду:

$$\begin{pmatrix} 5 & 0 & 10 & 0 \\ 0 & 10 & 0 & 34 \\ 10 & 0 & 34 & 0 \\ 0 & 34 & 0 & 130 \end{pmatrix} \cdot \begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \end{pmatrix} = \begin{pmatrix} \sum_{t=-2}^2 x_{i+t} \\ \sum_{t=-2}^2 t x_{i+t} \\ \sum_{t=-2}^2 t^2 x_t \\ \sum_{t=-2}^2 t^3 x_{i+t} \end{pmatrix}. \quad (6.7)$$

Розв'язок системи (6.7) має такий вигляд:

$$\begin{aligned}
\hat{a}_0 = \tilde{x}_i &= \frac{1}{35}(-3x_{i-2} + 12x_{i-1} + 17x_i + 12x_{i+1} - 3x_{i+2}), \\
\hat{a}_1 &= \frac{1}{72} \left(65 \sum_{t=-2}^2 tx_{i+t} - 17 \sum_{t=-2}^2 t^3 x_{i+t} \right), \\
\hat{a}_2 &= \frac{1}{14} \left(-2 \sum_{t=-2}^2 x_{i+t} + \sum_{t=-2}^2 t^2 x_{i+t} \right), \\
\hat{a}_3 &= \frac{1}{72} \left(-17 \sum_{t=-2}^2 tx_{i+t} + 5 \sum_{t=-2}^2 t^3 x_{i+t} \right).
\end{aligned} \tag{6.8}$$

Із аналізу виразів (6.8) випливає, що значення x_1, x_2, x_{N-1}, x_N не підлягають згладжуванню, отже має місце урізання вихідного масиву даних. Для відтворення його розмірності необхідне обчислення значень поліному (6.6) за $t = -2; -1; 1; 2$ з урахуванням параметрів (6.8). Згладжені значення у цих точках одержують за формулами:

$$\begin{aligned}
f(-2) &= \frac{1}{70}(69x_{i-2} + 4x_{i-1} - 6x_i + 4x_{i+1} - x_{i+2}), \\
f(-1) &= \frac{2}{70}(2x_{i-2} + 27x_{i-1} + 12x_i - 8x_{i+1} + 2x_{i+2}), \\
f(1) &= \frac{2}{70}(2x_{i-2} - 8x_{i-1} + 12x_i + 27x_{i+1} + 2x_{i+2}), \\
f(2) &= \frac{1}{70}(-x_{i-2} + 4x_{i-1} - 6x_i + 4x_{i+1} + 69x_{i+2}).
\end{aligned} \tag{6.9}$$

Приймаючи $i = 3$ та $N - 2$, на підставі виразів (6.9), одержують

$$\begin{aligned}
\tilde{x}_1 = f(-2) &= \frac{1}{70}(69x_1 + 4x_2 - 6x_3 + 4x_4 - x_5), \\
\tilde{x}_2 = f(-1) &= \frac{2}{70}(2x_1 + 27x_2 + 12x_3 - 8x_4 + 2x_5), \\
\tilde{x}_{N-1} = f(1) &= \frac{2}{70}(2x_{N-4} - 8x_{N-3} + 12x_{N-2} + 27x_{N-1} + 2x_N), \\
\tilde{x}_N = f(2) &= \frac{1}{70}(-x_{N-4} + 4x_{N-3} - 6x_{N-2} + 4x_{N-1} + 69x_N).
\end{aligned} \tag{6.10}$$

Зауважимо, що значення $\hat{a}_0 = \tilde{x}_i$ можна записати символічно

$$[5] \quad \hat{a}_0 = \frac{1}{35}[-3, 12, 17].$$

Отже, згладжування **кубічним поліномом** (6.6) базується на обчисленні $\tilde{x}_i = \hat{a}_0$ згідно (6.8) для $i = \overline{3, N-2}$ та визначенні $\tilde{x}_1, \tilde{x}_2, \tilde{x}_{N-1}, \tilde{x}_N$ за формулами (6.10). Результат згладжування проілюстровано (рис. 6.10).

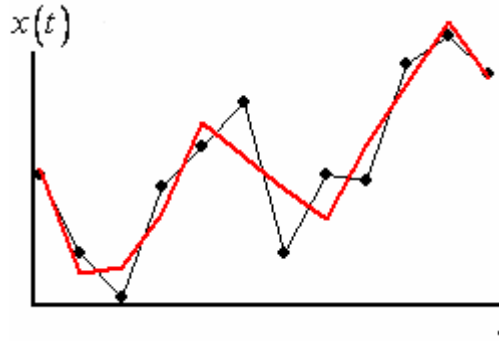


Рис. 6.10. Реалізація випадкового процесу та її згладжування методом найменших квадратів

Наведемо вирази, що визначають процедуру згладжування, для поліному третього порядку (6.6):

1) для 7-и точок ковзання

$$[7]\hat{a}_0 = \frac{1}{21}[-2, 3, 6, 7];$$

$$\hat{a}_1 = \frac{1}{1512} \left(397 \sum_{t=-3}^3 tx_{i+t} - 49 \sum_{t=-3}^3 t^3 x_{i+t} \right);$$

$$\hat{a}_2 = \frac{1}{84} \left(-4 \sum_{t=-3}^3 x_{i+t} + \sum_{t=-3}^3 t^2 x_{i+t} \right);$$

$$\hat{a}_3 = \frac{1}{216} \left(-7 \sum_{t=-3}^3 tx_t + \sum_{t=-3}^3 t^3 x_{i+t} \right);$$

2) для 9-и точок ковзання

$$[9]\hat{a}_0 = \frac{1}{231}[-21, 14, 39, 54, \mathbf{59}];$$

$$\hat{a}_1 = \frac{1}{12} \left(815 \sum_{t=-4}^4 tx_{i+t} - 59 \sum_{t=-4}^4 t^3 x_{i+t} \right);$$

$$\hat{a}_2 = \frac{1}{14} \left(-20 \sum_{t=-4}^4 x_{i+t} + 3 \sum_{t=-4}^4 t^2 x_{i+t} \right);$$

$$\hat{a}_3 = \frac{1}{1512} \left(-59 \sum_{t=-4}^4 tx_{i+t} + 5 \sum_{t=-4}^4 t^3 x_{i+t} \right).$$

Найпростішою й ефективною є процедура «Т'юкі 53х» (**процедура медіанного згладжування**) на основі якої побудований згладжуючий фільтр Хеммінга. Враховуючи стохастичність вихідних даних, медіану визначають як середню за ймовірністю точку ковзання. Її геометрична інтерпретація є точка перетину медіан трикутника, побудованого на вікні ковзання, вершинами якого є три точки експериментальних даних. Останнє покладено в основу нижченаведеної процедури.

Будемо характеризувати вхідний масив $\{x(t_i); i = \overline{1, N}\}$ набором точок

$$\{M_i(t_i, x_i); i = \overline{1, N}\},$$

згладжений масив – точками

$$\{\tilde{M}_i(\tilde{t}_i, \tilde{x}_i); i = \overline{2, N-1}\},$$

де \tilde{M}_i – точки, що визначають перетин медіан $\Delta M_{i-1} M_i M_{i+1}$, $i = \overline{2, N-1}$:

$$\tilde{M}_i = \frac{1}{3}(M_{i-1} + M_i + M_{i+1}),$$

де

$$\tilde{t}_i = \frac{1}{3}(t_{i-1} + t_i + t_{i+1});$$

$$\tilde{x}_i = \frac{1}{3}(x_{i-1} + x_i + x_{i+1})$$

або

$$\tilde{t}_i = t_i + \frac{1}{3}\Delta^2 t_i;$$

$$\tilde{x}_i = x_i + \frac{1}{3}\Delta^2 x_i, \quad i = \overline{2, N-1},$$

де

$$\Delta^2 t_i = t_{i-1} - 2t_i + t_{i+1},$$

$$\Delta^2 x_i = x_{i-1} - 2x_i + x_{i+1}.$$

Для усунення крайового ефекту визначають точки M_1, M_N одним зі способів, що базуються на умовах:

$$M_0 M_1 = M_1 M_2, \quad M_N M_{N-1} = M_{N+1} M_N \quad (6.11)$$

або

$$M_3 M_0 = M_3 \tilde{M}_2 + M_3 M_1, \quad M_{N-2} M_{N+1} = M_{N-2} \tilde{M}_{N-1} + M_{N-2} M_N. \quad (6.12)$$

Із наведених умов одержують координати точок $M_0(x_0, y_0)$, $M_{N+1}(t_{N+1}, x_{N+1})$:

1) з умови (6.11)

$$\begin{aligned} t_0 &= 2t_1 - t_2, & t_{N+1} &= 2t_N - t_{N-1}, \\ x_0 &= 2x_1 - x_2, & x_{N+1} &= 2x_N - t_{N-1}; \end{aligned} \quad (6.13)$$

2) з умови (6.12)

$$\begin{aligned} t_0 &= \frac{1}{3}(4t_1 + t_2 - 2t_3), & t_{N+1} &= \frac{1}{3}(4t_N + t_{N-1} - 2t_{N-2}), \\ x_0 &= \frac{1}{3}(4x_1 + x_2 - 2x_3), & x_{N+1} &= \frac{1}{3}(4x_N + x_{N-1} - 2x_{N-2}). \end{aligned} \quad (6.14)$$

Згладжені значення визначаються за формулами:

$$\tilde{t}_i = t_i + \alpha \Delta^2 t_i, \quad \tilde{x}_i = x_i + \alpha \Delta^2 x_i, \quad i = \overline{1, N}. \quad (6.15)$$

Коли $\alpha = 1/3$, має місце медіанне згладжування, при $\alpha = 1/8$ алгоритм тісно пов'язаний зі згладжуванням параболічними сплайнами, при $\alpha = 1/6$ – кубічними сплайнами. Процедура згладжування на основі (6.15) є ітераційна. Умова закінчення згладжування є оцінка

$$|L - \tilde{L}| \leq A\alpha(\tilde{L}/N)^2, \quad (6.16)$$

де $0 \leq \alpha \leq 1/3$; L, \tilde{L} – довжини ламаних до та після згладжування, проведені через точки $M_i, \tilde{M}_i, i = \overline{1, N}$; A визначається за нижченаведеним виразом (6.17).

Формулюється наступна процедура згладжування стосовно масиву $\{M_i(t_i, x_i); i = \overline{1, N}\}$:

1. Визначають точки $M_0(t_0, x_0), M_{N+1}(t_{N+1}, x_{N+1})$ за (6.13) або (6.14).
2. Для $\alpha \in [0; 1/3]$ обчислюють згладжені значення \tilde{t}_i, \tilde{x}_i згідно (6.15).
3. Обчислюють довжини ламаних за такими виразами:

$$L = \sum_{i=2}^N l_i,$$

$$\tilde{L} = \sum_{i=2}^N \tilde{l}_i,$$

де

$$l_i = \sqrt{(t_i - t_{i-1})^2 + (x_i - x_{i-1})^2},$$

$$\tilde{l}_i = \sqrt{(\tilde{t}_i - \tilde{t}_{i-1})^2 + (\tilde{x}_i - \tilde{x}_{i-1})^2}.$$

4. Обчислюють величини, пов'язані з кінцевими різницями:

$$qt_i = t_i - t_{i-1},$$

$$pt_i = qt_{i+1} - 2qt_i + qt_{i-1},$$

$$qx_i = x_i - x_{i-1},$$

$$px_i = qx_{i+1} - 2qx_i + qx_{i-1}, \quad i = \overline{2, N}.$$

5. Визначають величину

$$A = \sum_{i=2}^N \tilde{l}_i T_i, \quad T_i = \frac{1}{2\tilde{l}_i} (qt_i pt_i - qx_i px_i). \quad (6.17)$$

6. Перевіряють умову (6.16). У разі її виконання процес згладжування вважають завершеним, інакше продовжують згладжування.

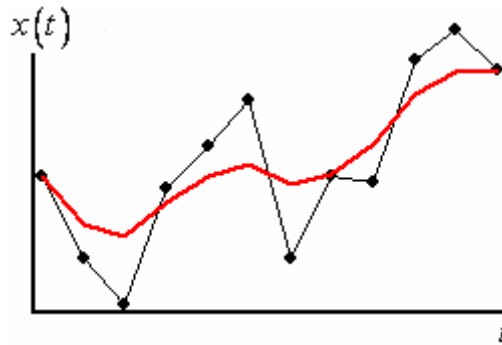


Рис. 6.11. Реалізація випадкового процесу та її медіанне згладжування ($\alpha = 1/3$)

Приклад застосування поданої процедури наведено на гарфіку (рис. 6.11).

6.3.4. Вилучення поліноміального тренду

Для проведення якісного спектрального аналізу випадкових процесів необхідно при попередній обробці даних виявляти і вилучати тренд. **Природою тренда можуть бути:**

1) детермінована складова фізичного процесу, який має деградацію (наприклад, процеси зносу, старіння, виходу з використання тощо);

- 2) похибка, що накопичується за рахунок складової (наприклад, похибка калібрування при вимірюваннях, яка має лінійний тренд);
- 3) наявність низькочастотних шумів при фіксації даних у часі;
- 4) сезонні коливання;
- 5) інші фактори.

Тренд може бути вилучений, і отриманий процес стане стаціонарний або функціонально описаний для подальшого аналізу.

Розглянемо функцію тренду у вигляді алгебраїчного поліному

$$m(t) = \sum_{v=0}^k a_v t^v. \quad (6.18)$$

Оцінка поліноміального тренду (6.18) для дискретно заданого процесу $\Omega_{1,N} = \{x_i, i = \overline{1, N}\}$ визначається із умови мінімуму залишкової дисперсії

$$S_{\text{зали}}^2 = \frac{1}{N-k-1} \sum_{i=1}^N \left(x_i - \sum_{v=0}^k \hat{a}_v t_i^v \right)^2,$$

що еквівалентна розв'язуванню системи лінійних рівнянь:

$$\sum_{i=1}^N \left(x_i - \sum_{v=0}^k \hat{a}_v t_i^v \right)^2 (-t_i^j) = 0, \quad j = \overline{0, k}$$

або

$$\sum_{v=0}^k \hat{a}_v \sum_{i=1}^N t_i^{j+v} = \sum_{i=1}^N x_i t_i^j, \quad j = \overline{0, k}. \quad (6.19)$$

Практично реалізують поліноми ступеня $k < 3$.

Розв'язуючи систему (6.19), знаходять оцінки параметрів \hat{a}_v , $v = \overline{0, k}$. Так, для $k = 0$ отримують

$$\hat{a}_0 = \frac{1}{N} \sum_{i=1}^N x_i.$$

При $k = 1$ (лінійний тренд) система (6.19) набуває такого вигляду:

$$\begin{pmatrix} 1 & \bar{t} \\ \bar{t} & \bar{t}^2 \end{pmatrix} \begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \end{pmatrix} = \begin{pmatrix} \bar{x} \\ \bar{tx} \end{pmatrix}, \quad (6.20)$$

де

$$\bar{t} = \frac{1}{N} \sum_{i=1}^N t_i; \quad \overline{t^2} = \frac{1}{N} \sum_{i=1}^N t_i^2;$$

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i, \quad \overline{tx} = \frac{1}{N} \sum_{i=1}^N t_i x_i.$$

Із розв'язання системи (6.20) одержують

$$\hat{a}_0 = \frac{\overline{xt^2} - \bar{t} \overline{tx}}{\overline{t^2} - \bar{t}^2}, \quad \hat{a}_1 = \frac{\overline{tx} - \bar{t} \cdot \bar{x}}{\overline{t^2} - \bar{t}^2}.$$

У якості прикладу наведемо результати побудови та вилучення лінійного тренда $m(t) = a_0 + a_1 t$ (рис. 6.12).

При $k = 2$, оцінки \hat{a}_0 , \hat{a}_1 , \hat{a}_2 параметрів параболічного тренду знаходять із системи рівнянь

$$\begin{pmatrix} 1 & \bar{t} & \overline{t^2} \\ \bar{t} & \overline{t^2} & \overline{t^3} \\ \overline{t^2} & \overline{t^3} & \overline{t^4} \end{pmatrix} \cdot \begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \\ \hat{a}_2 \end{pmatrix} = \begin{pmatrix} \bar{x} \\ \overline{tx} \\ \overline{t^2 x} \end{pmatrix}$$

де

$$\overline{t^3} = \frac{1}{N} \sum_{i=1}^N t_i^3; \quad \overline{t^4} = \frac{1}{N} \sum_{i=1}^N t_i^4;$$

$$\overline{t^2 x} = \frac{1}{N} \sum_{i=1}^N t_i^2 x_i.$$

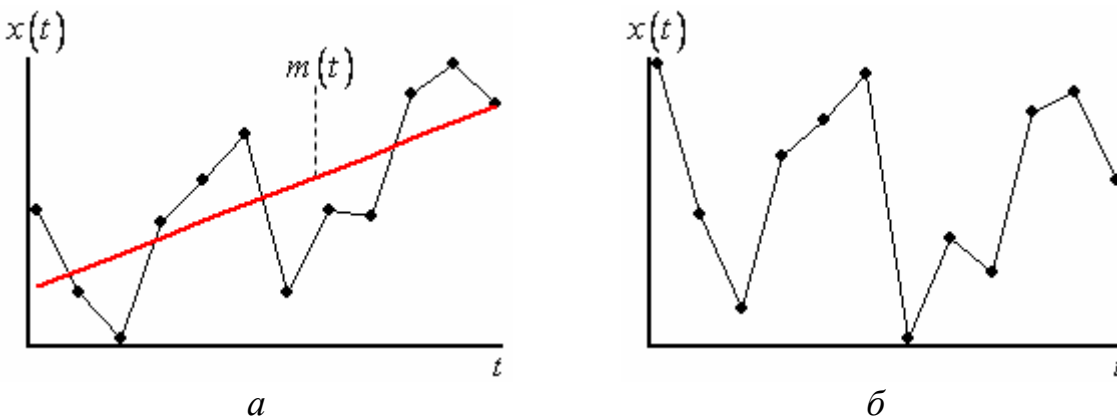


Рис. 6.12. Реалізація випадкового процесу до (а) та після (б) вилучення тренду

Реалізація поліномів ступеня $k \geq 3$ у більшості випадків неефективна,

оскільки їх використання призводить до зміни характеру тренду.

Контрольні запитання та завдання

1. Дати визначення випадкового процесу та вказати відміну від детермінованого процесу.
2. Визначити характеристики випадкового процесу.
3. Чи є ергодичний процес стаціонарний?
4. У чому полягає різниця між стаціонарним процесом у широкому та вузькому розумінні?
5. Що таке частота Найквіста?
6. Визначити, яке із тверджень для автокореляційної функції правильне:
 - зменшується із зростанням $|\tau|$;
 - залежить тільки від зсуву τ ;
 - прямує до сталої величини із зростанням $|\tau|$;
 - завжди від'ємна.
7. Показати, що автокореляційна функція не зміниться при додаванні до випадкового процесу детермінованої сталої.
8. Встановити, чи є щільність розподілу $f(x)$, автокореляційна функція $R(\tau)$ та спектральна щільність $S(\omega)$ стаціонарного випадкового процесу: парними; невід'ємними та необмеженими в нулі функціями.
9. Навести процедуру перевірки стаціонарності процесу, що базується на дослідженні наявності тренду.
10. Подати процедури згладжування за методом найменших квадратів для 5, 7 і 9 точок.
11. Яка умова закінчення медіанного згладжування?
12. Знайти оцінки параметрів параболічного та кубічного трендів за $k = 1; 2; 3$.

ДОДАТОК А

Процедури знаходження квантилів

У статистичному аналізі, а саме в задачах перевірки статистичних гіпотез, виникає необхідність знаходження квантилів розподілів. Найбільш поширені є квантилі розподілів: нормального, Стюдента, Пірсона та Фішера. Визначення квантиля нормального розподілу подане в п.1.2.4. Для квантилів інших розподілів нижче вказані найбільш ефективні й водночас прості в реалізації процедури знаходження.

Квантиль $t_{\alpha/2, \nu}$ розподілу Стюдента (табл. Б.2) обчислюється на основі розвинення в ряд:

$$t_{\alpha/2, \nu} \approx u_{\alpha/2} + \frac{1}{\nu} g_1(u_{\alpha/2}) + \frac{1}{\nu^2} g_2(u_{\alpha/2}) + \frac{1}{\nu^3} g_3(u_{\alpha/2}) + \frac{1}{\nu^4} g_4(u_{\alpha/2}),$$

де $u_{\alpha/2}$ – квантиль нормального розподілу;

$$g_1(u_{\alpha/2}) = \frac{1}{4}(u_{\alpha/2}^3 + u_{\alpha/2});$$

$$g_2(u_{\alpha/2}) = \frac{1}{96}(5u_{\alpha/2}^5 + 16u_{\alpha/2}^3 + 3u_{\alpha/2});$$

$$g_3(u_{\alpha/2}) = \frac{1}{384}(3u_{\alpha/2}^7 + 19u_{\alpha/2}^5 + 17u_{\alpha/2}^3 - 15u_{\alpha/2});$$

$$g_4(u_{\alpha/2}) = \frac{1}{92160}(79u_{\alpha/2}^9 + 779u_{\alpha/2}^7 + 1482u_{\alpha/2}^5 - 1920u_{\alpha/2}^3 - 945u_{\alpha/2}).$$

Квантиль $\chi_{\alpha, \nu}^2$ розподілу χ^2 (Пірсона) (табл. Б.3) може бути визначений на основі формули

$$\chi_{\alpha, \nu}^2 \approx \nu \left(1 - \frac{2}{9\nu} + u_{\alpha} \sqrt{\frac{2}{9\nu}} \right)^3,$$

де u_{α} – квантиль нормального розподілу.

Як апроксимацію квантиля f_{α, ν_1, ν_2} розподілу Фішера (табл. Б.4) можна застосовувати такий вираз:

$$f_{\alpha, v_1, v_2} = \exp(2z),$$

де

$$\begin{aligned} z = & u_{\alpha} \sqrt{\frac{\sigma}{2}} - \frac{1}{6} \delta (u_{\alpha}^2 + 2) + \sqrt{\frac{\sigma}{2}} \left(\frac{\sigma}{24} (u_{\alpha}^2 + 3u_{\alpha}) + \frac{1}{72} \frac{\delta^2}{\sigma} (u_{\alpha}^3 + 11u_{\alpha}) \right) - \\ & - \frac{\delta \sigma}{120} (u_{\alpha}^4 + 9u_{\alpha}^2 + 8) + \frac{\delta^3}{3 \cdot 240 \sigma} (3u_{\alpha}^4 + 7u_{\alpha}^2 - 16) + \sqrt{\frac{\sigma}{2}} \left(\frac{\sigma^2}{1920} (u_{\alpha}^5 + 20u_{\alpha}^3 + 15u_{\alpha}) + \right. \\ & \left. + \frac{\delta^4}{2 \cdot 880} (u_{\alpha}^5 + 44u_{\alpha}^3 + 183u_{\alpha}) + \frac{\delta^4}{155 \cdot 520 \sigma^2} (9u_{\alpha}^5 - 284u_{\alpha}^3 - 1513u_{\alpha}) \right), \end{aligned}$$

де

$$\sigma = \frac{1}{v_1} + \frac{1}{v_2};$$

$$\delta = \frac{1}{v_1} - \frac{1}{v_2};$$

u_{α} — квантиль нормального розподілу.

Закінчення табл. Б.1

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6404	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7703	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9014
1,3	0,9032	0,9049	0,9065	0,9082	0,9098	0,9114	0,9130	0,9146	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9250	0,9264	0,9278	0,9292	0,9305	0,9318
1,5	0,9331	0,9344	0,9357	0,9369	0,9382	0,9394	0,9406	0,9417	0,9429	0,9440
1,6	0,9452	0,9463	0,9473	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9544
1,7	0,9554	0,9563	0,9572	0,9581	0,9590	0,9599	0,9608	0,9616	0,9624	0,9632
1,8	0,9640	0,9648	0,9656	0,9663	0,9671	0,9678	0,9685	0,9692	0,9699	0,9706
1,9	0,9712	0,9719	0,9725	0,9732	0,9738	0,9744	0,9750	0,9755	0,9761	0,9767
2,0	0,9772	0,9777	0,9783	0,9788	0,9793	0,9798	0,9803	0,9807	0,9812	0,9816
2,1	0,9821	0,9825	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9853	0,9857
2,2	0,9861	0,9864	0,9867	0,9871	0,9874	0,9877	0,9880	0,9884	0,9887	0,9889
2,3	0,9892	0,9895	0,9898	0,9900	0,9903	0,9906	0,9908	0,9911	0,9913	0,9915
2,4	0,9918	0,9920	0,9922	0,9924	0,9926	0,9928	0,9930	0,9932	0,9934	0,9936
2,5	0,9937	0,9939	0,9941	0,9942	0,9944	0,9946	0,9947	0,9949	0,9950	0,9952
2,6	0,9953	0,9954	0,9956	0,9957	0,9958	0,9959	0,9960	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9971	0,9972	0,9973
2,8	0,9974	0,9975	0,9975	0,9976	0,9977	0,9978	0,9978	0,9979	0,9980	0,9980
2,9	0,9981	0,9981	0,9982	0,9983	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986
3,0	0,9986	0,9986	0,9987	0,9987	0,9988	0,9988	0,9988	0,9989	0,9989	0,9989
3,1	0,9990	0,9990	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992
3,2	0,9993	0,9993	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9994
3,3	0,9995	0,9995	0,9995	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996
3,4	0,9996	0,9996	0,9996	0,9996	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997
3,5	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998	0,9998	0,9998	0,9998	0,9998
3,6	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998

Таблиця Б.2

Квантилі t -розподілу Стюдента

ν	$\alpha=0,50$	0,25	0,10	0,05	0,02	0,01
1	1,00	2,41	6,31	12,7	31,82	63,7
2	0,816	1,60	2,92	4,30	6,97	9,92
3	0,765	1,42	2,35	3,18	4,54	5,84
4	0,741	1,34	2,13	2,78	3,75	4,60
5	0,727	1,30	2,01	2,57	3,37	4,03
6	0,718	1,27	1,94	2,45	3,14	3,71
7	0,711	1,25	1,89	2,36	3,00	3,50
8	0,706	1,24	1,86	2,31	2,90	3,36
9	0,703	1,23	1,83	2,26	2,82	3,25
10	0,700	1,22	1,81	2,23	2,76	3,17
11	0,697	1,21	1,80	2,20	2,72	3,11
12	0,695	1,21	1,78	2,18	2,68	3,05
13	0,694	1,20	1,77	2,16	2,65	3,01
14	0,692	1,20	1,76	2,14	2,62	2,98
15	0,691	1,20	1,75	2,13	2,60	2,95
16	0,690	1,19	1,75	2,12	2,58	2,92
17	0,689	1,19	1,74	2,11	2,57	2,90
18	0,688	1,19	1,73	2,10	2,55	2,88
19	0,688	1,19	1,73	2,09	2,54	2,86
20	0,687	1,18	1,73	2,09	2,53	2,85
21	0,686	1,18	1,72	2,08	2,52	2,83
22	0,686	1,18	1,72	2,07	2,51	2,82
23	0,685	1,18	1,71	2,07	2,50	2,81
24	0,685	1,18	1,71	2,06	2,49	2,80
25	0,684	1,18	1,71	2,06	2,49	2,79
26	0,684	1,18	1,71	2,06	2,48	2,78
27	0,684	1,18	1,71	2,05	2,47	2,77
28	0,683	1,17	1,70	2,05	2,47	2,76
29	0,683	1,17	1,70	2,05	2,46	2,76
30	0,683	1,17	1,70	2,04	2,46	2,75
40	0,681	1,17	1,68	2,02	2,42	2,70
60	0,679	1,16	1,67	2,00	2,39	2,66
120	0,677	1,16	1,66	1,98	2,36	2,62
∞	0,674	1,15	1,64	1,96	2,33	2,58
F	$\alpha/2=0,25$	0,125	0,05	0,025	0,01	0,005

Таблиця Б.3

Квантилі розподілу χ^2

ν	$\alpha=0,99$	0,95	0,90	0,50	0,10	0,05	0,01
1	0,0001	0,0039	0,0158	0,455	2,71	3,84	6,64
2	0,0201	0,103	0,211	1,39	4,61	5,99	9,21
3	0,115	0,352	0,584	2,37	6,25	7,81	11,3
4	0,297	0,711	1,06	3,36	7,78	9,49	13,3
5	0,554	1,15	1,61	4,35	9,24	11,1	15,1
6	0,872	1,64	2,20	5,35	10,6	12,6	16,8
7	1,24	2,17	2,83	6,35	12,0	14,1	18,5
8	1,65	2,73	3,49	7,34	13,4	15,5	20,1
9	2,09	3,33	4,17	8,34	14,7	16,9	21,7
10	2,56	3,94	4,87	9,34	16,0	18,3	23,2
11	3,05	4,57	5,58	10,3	17,3	19,7	24,7
12	3,57	5,23	6,30	11,3	18,5	21,0	26,2
13	4,11	5,89	7,04	12,3	19,8	22,4	27,7
14	4,66	6,57	7,79	13,3	21,1	23,7	29,1
15	5,23	7,26	8,55	14,3	22,3	25,0	30,6
16	5,81	7,96	9,31	15,3	23,5	26,3	32,0
17	6,41	8,67	10,1	16,3	24,8	27,6	33,4
18	7,01	9,39	10,9	17,3	26,0	28,9	34,8
19	7,63	10,1	11,7	18,3	27,2	30,1	36,2
20	8,26	10,9	12,4	19,3	28,4	31,4	37,6
21	8,90	11,6	13,2	20,3	29,6	32,7	38,9
22	9,54	12,3	14,0	21,3	30,8	33,9	40,3
23	10,2	13,1	14,8	22,3	32,0	35,2	41,6
24	10,9	13,8	15,7	23,3	33,2	36,4	43,0
25	11,5	14,6	16,5	24,3	34,4	37,7	44,3
26	12,2	15,4	17,3	25,3	35,6	38,9	45,6
27	12,9	16,2	18,1	26,3	36,7	40,1	47,0
28	13,6	16,9	18,9	27,3	37,9	41,3	48,3
29	14,3	17,7	19,8	28,3	39,1	42,6	49,6
30	15,0	18,5	20,6	29,3	40,3	43,8	50,6
40	22,2	26,5	29,1	39,3	51,8	55,6	63,7
50	29,7	34,8	37,7	49,3	63,2	67,5	76,1
60	37,5	43,2	46,5	59,3	74,4	79,1	88,4
70	45,4	51,7	55,3	69,3	85,5	90,5	100,4
80	53,5	60,4	64,3	79,3	96,6	101,9	112,3
90	61,8	69,1	73,3	99,3	107,5	113,3	124,1

Таблиця Б.4

Квантилі F -розподілу Фішера ($\alpha = 0,05$)

$v_2 \backslash v_1$	1	2	3	4	5	6	7	8	9	10
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9
2	18,51	19,00	19,16	19,25	19,3	19,33	19,35	19,37	19,38	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83

Закінчення табл. Б.4

$v_2 \backslash v_1$	12	15	20	24	30	40	60	120	∞
1	243,9	245,9	248,0	249,1	250,1	251,1	252,2	253,3	254,3
2	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	8,78	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

ДОДАТОК В

Приклади завдань до лабораторних робіт

Лабораторна робота 1

Первинний статистичний аналіз та відтворення розподілів

Постановка задачі

Написати програму, яка б дозволила користувачу провести аналіз статистичних даних, що передбачає реалізацію таких обчислювальних процедур:

1) первинного статистичного аналізу, складовими частинами якого є:

- формування варіаційного ряду;
- проведення гістограмної оцінки (кількість класів має визначатися автоматично та за вимогою користувача);
- підрахунок незсунених кількісних характеристик (середнього арифметичного, середньоквадратичного, коефіцієнтів асиметрії, ексцесу, контрексцесу, варіації Пірсона), їх середньоквадратичних і довірчих інтервалів;
- вилучення аномальних значень (після підтвердження користувача);
- побудова графіка емпіричної функції розподілу;
- ідентифікація типу розподілу.

2) відтворення розподілів (нормального, експоненціального, Вейбулла та рівномірного), що включає:

- знаходження оцінок параметрів розподілів та оцінку їх точності;
- довірче оцінювання теоретичної функції розподілу;
- перевірку вірогідності відтворення на основі критеріїв згоди (Пірсона та Колмогорова).

Провести тестування програмного забезпечення на реальних даних.

За результатами виконання лабораторної роботи оформити звіт.

Загальні вимоги до програми

1. Програма повинна бути незалежна від даних. Вхідний файл має обиратися в діалозі з користувачем. Передбачається, що вхідні дані знаходяться в текстовому файлі, обсяг даних не відомий. Потрібно забезпечити можливість модифікації та збереження даних.

2. Слід уможливити перетворення даних (логарифмування, стандартизація, зсув).

3. Після перетворення або вилучення аномальних значень користувач повинен мати можливість повернутися до початкових даних.

4. Необхідно нанести на одну площину з гістограмою графік статистичної функції щільності, а на площину з графіком емпіричної функції розподілу – графік статистичної функції розподілу разом із її довірчими інтервалами.

5. Результатом використання критерію згоди повинні бути як проміжні результати (статистика критерію та її критичне значення), так і висновок (чи є відтворення розподілу достовірне).

6. Результати виконання всіх обчислень мають виводитись у вигляді таблиць, графіків і текстових коментарів.

7. Для кожного графіка слід виконати автоматичне масштабування, зобразити шкалу й показати одиниці виміру.

8. Відображення результатів повинне відповідати точності обчислень.

Загальні вимоги до звіту

Звіт із лабораторної роботи складається з таких частин:

1. Постановка задачі.
2. Теоретична частина.
3. Опис програми (програмні модулі, основні об'єкти, схема взаємодії модулів, інтерфейс, порядок роботи з програмою, опис формату вхідних даних та додаткових можливостей програми).
4. Реалізація (вхідні дані повністю, вихідні результати у вигляді графіків і таблиць, коментарі та пояснення щодо отриманих результатів).
5. Висновки.

Звіт здається в письмовій формі українською мовою.

Лабораторна робота 2 **Критерії однорідності**

Постановка задачі

1. Організувати роботу з вхідними даними таким чином, щоб уможливити подальшу обробку однієї або кількох вибірок, які характеризують одновимірні або багатовимірні об'єкти спостережень. Для цього передбачити можливість прямого та поекторного зчитування даних із файлу.

2. Лабораторну роботу 2 виконати на основі лабораторної роботи 1 в рамках єдиної автоматизованої системи аналізу статистичних даних.

3. Реалізувати обчислювальні процедури перевірки однорідності двох вибірок, що характеризують одновимірні об'єкти спостережень:

- перевірку збігу дисперсій та середніх для вибірок, розподілених за нормальним законом;
- критерій Вілкоксона, Манна–Уїтні або різниці середніх рангів (на вибір).

4. Реалізувати обчислювальні процедури перевірки однорідності множин вибірок, які характеризують одновимірні об'єкти спостережень:

- критерій Бартлетта та однофакторний дисперсійний аналіз для вибірок, розподілених за нормальним законом;
 - Н-критерій.
5. Провести тестування програмного забезпечення на реальних даних.
 6. За результатами виконання лабораторної роботи оформити звіт.

Вимоги до програмного забезпечення та звіту аналогічні тим, що ставляться в лабораторній роботі 1.

Лабораторна робота 3

Аналіз двовимірних об'єктів спостережень.

Кореляційний та регресійний аналіз

Постановка задачі

На основі лабораторних робіт 1, 2 в рамках єдиної автоматизованої системи аналізу статистичних даних реалізувати такі обчислювальні процедури:

- 1) аналіз двовимірних об'єктів спостережень:
 - проведення первинного статистичного аналізу двовимірних даних;
 - відтворення двовимірного нормального розподілу;
 - перевірку достовірності відтворення на основі критерію згоди χ^2 ;
 - 2) перевірку наявності стохастичного зв'язку між окремими ознаками об'єкта:
 - знаходження оцінки коефіцієнта кореляції, перевірку його значущості та призначення довірчого інтервалу (у випадку значущості);
 - обчислення коефіцієнта кореляційного відношення та перевірку його значущості;
 - 3) за наявності стохастичного зв'язку між ознаками об'єкта – відтворення моделей лінійної та параболічної регресій, що включає:
 - знаходження оцінок параметрів регресій та дослідження їх значущості й точності;
 - визначення коефіцієнта детермінації;
 - побудову толерантних та довірчих інтервалів для кожної з ліній регресії, а також довірчих інтервалів для прогнозного значення;
 - перевірку адекватності відтворених моделей.
- Провести тестування програмного забезпечення на реальних даних.
За результатами виконання лабораторної роботи оформити звіт.

Вимоги до програмного забезпечення та звіту аналогічні викладеним у завданні до лабораторної роботи 1.

Лабораторна робота 4

Аналіз багатовимірних об'єктів спостережень

Постановка задачі

На основі лабораторних робіт 1, 2, 3 в рамках єдиної автоматизованої системи аналізу статистичних даних реалізувати такі обчислювальні процедури:

1) первинний статистичний аналіз багатовимірних об'єктів спостережень (знаходження векторів середніх та середньоквадратичних, матриці значущих парних коефіцієнтів кореляції);

2) стандартизацію та перехід до незалежних даних за (вимогою користувача);

3) перевірку збігу векторів середніх та коваріаційних матриць для множини вибірок у припущенні нормального закону розподілу випадкової величини;

4) перевірку наявності стохастичного зв'язку між ознаками об'єкту на основі часткових та множинних коефіцієнтів кореляції (за вимогою користувача):

- знаходження оцінок коефіцієнтів кореляції;
- перевірку їх значущості та для часткових коефіцієнтів кореляції призначення довірчих інтервалів (у випадку значущості);

5) за наявності стохастичного зв'язку між ознаками об'єкта – відтворення моделі багатовимірної лінійної регресії, що включає:

- знаходження оцінок параметрів регресії та дослідження їх значущості та точності;
- знаходження стандартизованих оцінок параметрів регресії;
- знаходження коефіцієнта детермінації;
- побудову толерантних меж для залишкової дисперсії;
- побудову довірчих інтервалів для лінії регресії;
- перевірку значущості відтвореної моделі (F-тест);
- побудову діагностичної діаграми.

Провести тестування програмного забезпечення на реальних даних.

За результатами виконання лабораторної роботи оформити звіт.

Вимоги до програмного забезпечення та звіту аналогічні тим, що ставляться в лабораторній роботі 1.

Лабораторна робота 5

Кластерний аналіз та класифікація

Постановка задачі

1. Лабораторну роботу 5 виконати на основі лабораторних робіт 1–4 в рамках єдиної автоматизованої системи аналізу статистичних даних.

2. Реалізувати обчислювальні процедури кластерного аналізу:

- проведення кластеризації методами агломеративним ієрархічним (один із методів на вибір) та K-середніх;

- обчислення для кожного виділеного кластеру середнього арифметичного та середньоквадратичного відхилення;
 - порівняння результатів кластеризації за допомогою функціоналу якості (один із функціоналів на вибір).
3. Реалізувати обчислювальні процедури класифікації:
 - проведення класифікації за допомогою лінійного та квадратичного класифікаторів;
 - перевірку якості класифікації за кожним із класифікаторів.
 4. Реалізувати метрики відстаней евклідову, манхетенську, Чебишева та Махаланобіса і надати користувачу можливість вибору однієї з них.
 5. Забезпечити можливості стандартизації даних (за вимогою користувача).
 6. Провести тестування програмного забезпечення на реальних даних.
 7. За результатами виконання лабораторної роботи оформити звіт.

Загальні вимоги до програми

1. Програма повинна бути незалежна від даних. Вхідний файл має обиратися в діалозі з користувачем. Передбачається, що вхідні дані знаходяться в текстовому файлі, обсяг даних не відомий. Потрібно забезпечити можливість модифікації та збереження даних.
2. Результати кластеризації та класифікації представити у вигляді таблиці та графіку. Графічне представлення результатів передбачає відображення значень двох ознак у вигляді сукупності точок: ознаки для відображення має обирати користувач; точки різних класів зобразити різним кольором або різної форми.
3. Для агломеративного ієрархічного методу подати дендрограму.
4. Результати виконання всіх обчислень мають виводитись у вигляді таблиць, графіків і текстових коментарів.
5. Для кожного графіка слід виконати автоматичне масштабування, зобразити шкалу й показати одиниці виміру.
6. Відображення результатів повинне відповідати точності обчислень.

Вимоги до звіту аналогічні поставленим у лабораторній роботі 1.

Лабораторна робота 6 **Редагування випадкових процесів**

Ціль роботи – вивчення методів, створення алгоритмів та програмного середовища редагування випадкових процесів, надбання навичок розв’язання конкретних задач вказаного типу у діалоговому режимі, проведення обчислювальних експериментів.

Постановка задачі

Задані результати спостереження над випадковим процесом у вигляді масиву $\Omega_{1,N} = \{x_i; i = \overline{1, N}\}$.

Написати програмне середовище редагування випадкового процесу, що дозволяє вирішувати такі задачі:

1. Вилучення аномальних значень.
2. Оцінка стаціонарності процесу на основі автокореляційної функції.
3. Ідентифікація тренда процесу (однією процедурою за варіантом).
4. Згладжування даних за допомогою методів ковзного середнього (5, 7, 9 точок) та медіанного згладжування.

Загальні вимоги до програми

1. Програма повинна бути незалежна від даних, файл з вхідними даними – обиратися діалогово (припускається, що вихідні данні знаходяться в DBF- або в текстовому файлі).
2. Організувати роботу з даними за допомогою меню.
3. Результати усіх обчислень повинні виводитись у вигляді таблиць, графіків та текстових коментарів. Для кожного графіка виконувати автоматичне масштабування, відображати шкалу й одиниці виміру.
4. Побудувати графіки досліджуваного процесу та автокореляційної функції.
5. Результати згладжування вивести на графік поряд із досліджуваним процесом та у таблицю вигляду

i	x_i	\tilde{x}_i	$ x_i - \tilde{x}_i $
1			
...
N			

СПИСОК РЕКОМЕНДОВАНОЇ ЛІТЕРАТУРИ

1. Айвазян, С.А. Классификация многомерных наблюдений [Текст] / С.А. Айвазян, З.И. Бежаева, О.В. Староверов. – М.: Статистика, 1974. – 240 с.
2. Андерсон, Т. Введение в многомерный статистический анализ [Текст] / Т. Андерсон. – М.: Наука, 1963. – 500 с.
3. Большев, Л.Н. Таблицы математической статистики [Текст] / Л.Н. Большев, Н.В. Смирнов. – М.: Наука, 1965. – 464 с.
4. Браунли, К.А. Статистическая теория и методология в науке и технике [Текст] / К.А. Браунли. – М.: Наука, 1977. – 407 с.
5. Ван-дер-Варден, Б.П. Математическая статистика [Текст] / Б.П. Ван-дер-Варден. – М.: Иностр. лит., 1960. – 434 с.
6. Деврой, Л. Непараметрическое оценивание плотности. L_1 -подход [Текст] / Л. Деврой, Л. Дьерфи. – М.: Мир, 1988. – 407 с.
7. Дуда, Р. Распознавание образов и анализ сцен [Текст] / Р. Дуда, П. Харт. – М.: Мир, 1976. – 512 с.
8. Кендалл, М. Теория распределений [Текст] / М. Кендалл, А. Стюарт. – М.: Наука, 1966. – 588 с.
9. Коваленко, И.Н. Теория вероятностей [Текст] / И.Н. Коваленко, Б.В. Гнеденко. – К.: Выща шк., 1990. – 328 с.
10. Коваленко, И.Н. Теория вероятностей и математическая статистика [Текст] / И.Н. Коваленко, А.А. Филиппова. – 2-е изд. – М.: Высш. шк., 1982. – 256 с.
11. Компьютерная биометрика [Текст] / Под ред. В.Н. Носова. – М.: Изд-во МГУ, 1990. – 232 с.
12. Мандель, И.Д. Кластерный анализ [Текст] / И.Д. Мандель. – М.: Финансы и статистика, 1988. – 176 с.

-
13. Основи теорії ймовірностей та математичної статистики [Текст] / В.П. Бабак, А.Я. Білецький, О.П. Приставка, П.О. Приставка. – К.: КВІЦ, 2003. – 432 с.
 14. Сигел, Э. Практическая бизнес-статистика [Текст]: пер. с англ. / Э. Сигел. – 4-е изд. – М.: Издат. дом «Вильямс», 2002. – 1056 с.
 15. Статистична обробка даних [Текст] / В.П. Бабак, А.Я. Білецький, О.П. Приставка, П.О. Приставка. – К.: МІВВЦ, 2001. – 388 с.
 16. Уилкс, С. Математическая статистика [Текст] / С. Уилкс. – М.: Наука, 1967. – 632 с.
 17. Харман, Г. Современный факторный анализ [Текст] / Г. Харман. – М.: Статистика, 1972. – 486 с.
 18. Эноксон Л. Прикладной анализ временных рядов. Основные методы [Текст] / Р. Отнес, Л. Эноксон. – М.: Мир, 1982. – 428 с.
 19. Ядренко М.И. Теория вероятностей и математичес: Учебник [Текст] / И. И. Гихман, А. В. Скороход, М. И. Ядренко. – К.: Выщ. шк, 1979. – 480 с.