

**Міністерство освіти і науки України
Дніпропетровський національний університет**

90-річчю ДНУ присвячується

П.О. Приставка, О.М. Мацуга

АНАЛІЗ ДАНИХ

*Рекомендовано
Міністерством освіти і науки України
як навчальний посібник для студентів
вищих навчальних закладів*

**Дніпропетровськ
РВВ ДНУ
2008**

ББК 22.172я73
УДК 519.25 (075.8)
П 77

Рецензенти:

д-р техн. наук, проф. І.Г. Прокопенко

д-р техн. наук, проф. Б.І. Мороз

д-р техн. наук, проф. Б.С. Бусигін

П 77 Приставка, П.О. Аналіз даних [Текст]: Навч. посіб. / МОН України /
П.О. Приставка, О.М. Мацуга. – Д.: РВВ ДНУ, 2008. – 92 с.

Систематично викладені основи статистичного аналізу одновимірних та двовимірних даних стосовно задач автоматизованої обробки результатів спостережень. Подані обчислювальні процедури первинного статистичного аналізу, відтворення класичних параметричних розподілів, перевірки статистичних гіпотез, кореляційного та регресійного аналізу.

Для студентів і аспірантів факультету прикладної математики, а також спеціальностей інженерного й технічного напрямків ДНУ.

Темплан 2008, поз. 34

*Гриф надано Міністерством освіти і науки України
Лист № 1.4/18-Г-2663 від 15.12.2008*

Навчальне видання

Пилип Олександрович Приставка
Ольга Миколаївна Мацуга

Аналіз даних
Навчальний посібник

Редактор О.В. Бец
Коректор Т.А. Андрєєва
Техредактор Л.П. Замятіна

Підписано до друку 04.07.08. Формат 60x84/16. Папір друкарський. Друк плоский.
Ум. друк. арк. 5,34. Ум. фарбовідб. 5,34. Обл.-вид. арк. 5,3. Тираж 200 пр. Зам. №
РВВ ДНУ, пр. Гагаріна, 72, м. Дніпропетровськ, 49010.
Друкарня ДНУ, вул. Наукова, 5, м. Дніпропетровськ, 49050.

© Приставка П.О., Мацуга О.М., 2008

ЗМІСТ

Вступ	4
1. Обробка й аналіз одновимірних даних.....	6
1.1. Первинний статистичний аналіз	6
1.1.1. Формування варіаційного ряду.....	6
1.1.2. Гістограмна оцінка.....	7
1.1.3. Точкові та інтервальні оцінки.....	13
1.2. Відтворення розподілів	20
1.2.1. Методи оцінки параметрів розподілу	20
1.2.2. Оцінювання точності оцінок параметрів.....	24
1.2.3. Інтервальне оцінювання теоретичної функції розподілу.....	25
1.2.4. Параметричні розподіли.....	25
Контрольні запитання та завдання.....	32
2. Перевірка статистичних гіпотез.....	34
2.1. Головні поняття та визначення	34
2.2. Оцінка згоди відтворення розподілів	40
2.3. Задача двох вибірок	42
2.4. Перевірка збігу середніх	42
2.5. Перевірка збігу дисперсій.....	44
2.6. Однофакторний дисперсійний аналіз.....	45
2.7. Критерії порядкових статистик.....	46
Контрольні запитання та завдання.....	49
3. Обробка й аналіз двовимірних даних.....	50
3.1. Первинний аналіз.....	50
3.2. Кореляційний аналіз.....	55
3.2.1. Парна кореляція.....	55
3.2.2. Кореляційне відношення	58
3.2.3. Парна рангова кореляція	59
3.3. Одновимірний регресійний аналіз.....	61
3.3.1. Лінійний регресійний аналіз	61
3.3.2. Нелінійний регресійний аналіз	75
Контрольні запитання та завдання.....	81
Додаток А. Процедури знаходження квантилів	82
Додаток Б. Статистичні таблиці.....	83
Додаток В. Приклади завдань до лабораторних робіт.....	89
Список використаної літератури.....	92

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. Дало 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 типове значення та характеристику 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 чисел (0 чи 1 або 1, 2, 3, ...). Кількісну змінну, що не є дискретна, 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, де мешкають студенти факультету (Жовтневий, Ленінський, Бабушкінський); 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 величин та функцій, яка базується на вихідних даних;
- 4) перевірка статистичних гіпотез, що дозволяє на основі вибіркових 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.

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. Дамо 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 гістограми 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

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

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 $\Omega_{1,N}$:

$$\begin{aligned} x_{\min} &\in [x_{i_{\min}} h; x_{i_{\min} + 1} h), \quad x_{\max} \in [x_{i_{\max}} h; x_{i_{\max} + 1} h), \\ i_{\min}, i_{\max} &\in Z, \quad i_{\min} < i_{\max}, \end{aligned}$$

можливе оцінювання функції розподілу $F(x)$ у вигляді емпіричної функції розподілу $F_{1,N}(x)$:

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

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

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

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.

М.М. Ченцов доводить, що за існування для $f(x)$ обмеженої другої похідної слушне таке співвідношення:

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

Тому при $N \geq 100$ можна застосовувати формулу

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

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

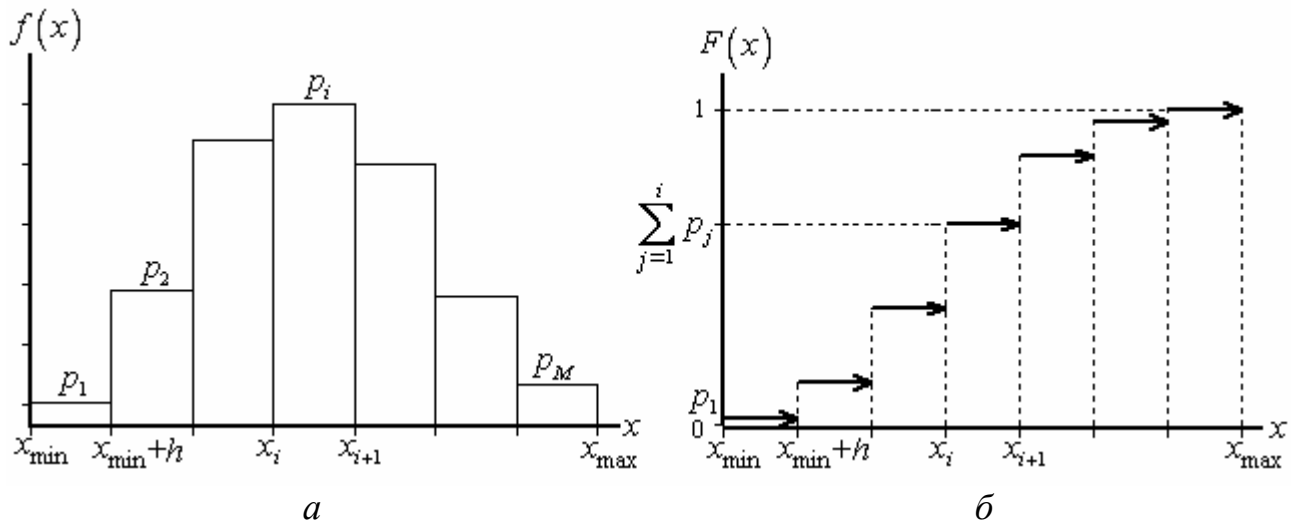


Рис. 1.1. Графічне подання результатів гістограмної оцінки:
a – гістограма відносних частот; *б* – графік емпіричної функції розподілу

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:

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

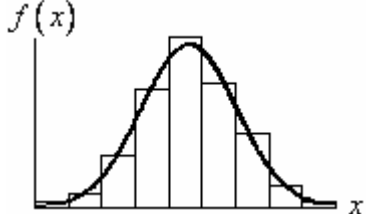

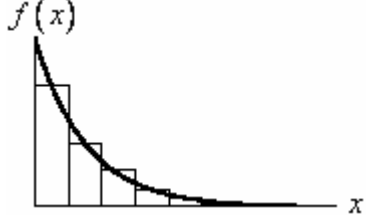

де $\tilde{f}(x)$ – нормована функція щільності. Надалі, говорячи про нормовану функцію щільності, знак « \sim » будемо опускати.

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.

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

Розподіл	Аналітичне подання	Вигляд гістограми та графіка нормованої функції щільності
Нормальний	$F(x; m, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{(u-m)^2}{2\sigma^2}\right) du,$ $-\infty < x < \infty$	 <p>The figure shows a histogram with approximately 10 bars of varying heights, centered around a mean value. A smooth, symmetric bell-shaped curve is overlaid on the histogram, representing the normal distribution function. The vertical axis is labeled $f(x)$ and the horizontal axis is labeled x.</p>
Рівномірний	$F(x; a, b) = \begin{cases} 0, & -\infty < x < a, \\ \frac{x-a}{b-a}, & a \leq x < b, \\ 1, & b \leq x < \infty \end{cases}$	 <p>The figure shows a histogram with approximately 10 bars of equal height, indicating a uniform distribution. A horizontal line is drawn across the top of the bars, representing the constant probability density function. The vertical axis is labeled $f(x)$ and the horizontal axis is labeled x.</p>
Експоненціальний	$F(x; \lambda) = 1 - \exp(-\lambda x),$ $0 \leq x < \infty$	 <p>The figure shows a histogram with approximately 10 bars, where the height of the bars decreases as they move to the right, characteristic of an exponential distribution. A smooth curve that starts high on the y-axis and decays towards the x-axis is overlaid. The vertical axis is labeled $f(x)$ and the horizontal axis is labeled x.</p>
Вейбулла	$F(t; \alpha, \beta) = 1 - \exp\left(-\frac{t^\beta}{\alpha}\right),$ $0 \leq x < \infty$	 <p>The figure shows a histogram with approximately 10 bars. A smooth curve is overlaid that starts at the origin, rises to a peak, and then decays towards the x-axis, representing the Weibull distribution. The vertical axis is labeled $f(x)$ and the horizontal axis is labeled x.</p>

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.

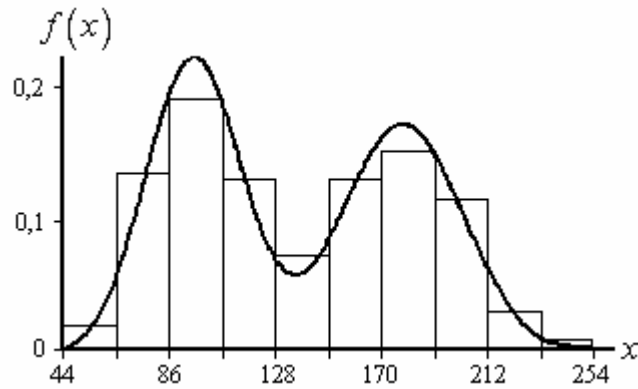


Рис. 1.2. Гістограма та графік нормованої функції щільності розподілу у випадку неоднорідних даних

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):

$$x_{2p} \leq x_{\gamma_1} \quad \text{або} \quad x_{2p} \geq x_{\gamma_2},$$

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_{2p} ;

2) має місце випадкове порушення умов експерименту.

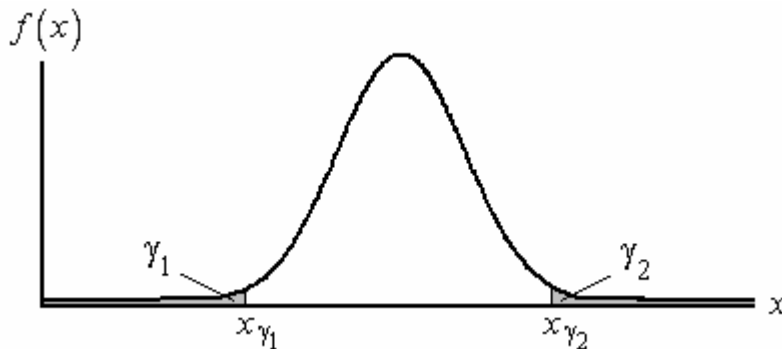


Рис. 1.3. Области малоймовірних спостережень на графіку функції щільності

У будь-якому разі за достатнього обсягу вибірки доцільно вилучати такі значення перед подальшою обробкою. Наприклад, оцінка x_{2p} 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

$$p_i = \int_{x_i-0,5h}^{x_i+0,5h} f(u) du < \gamma_1, \quad i = \overline{1, s_1}$$

або

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

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

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.

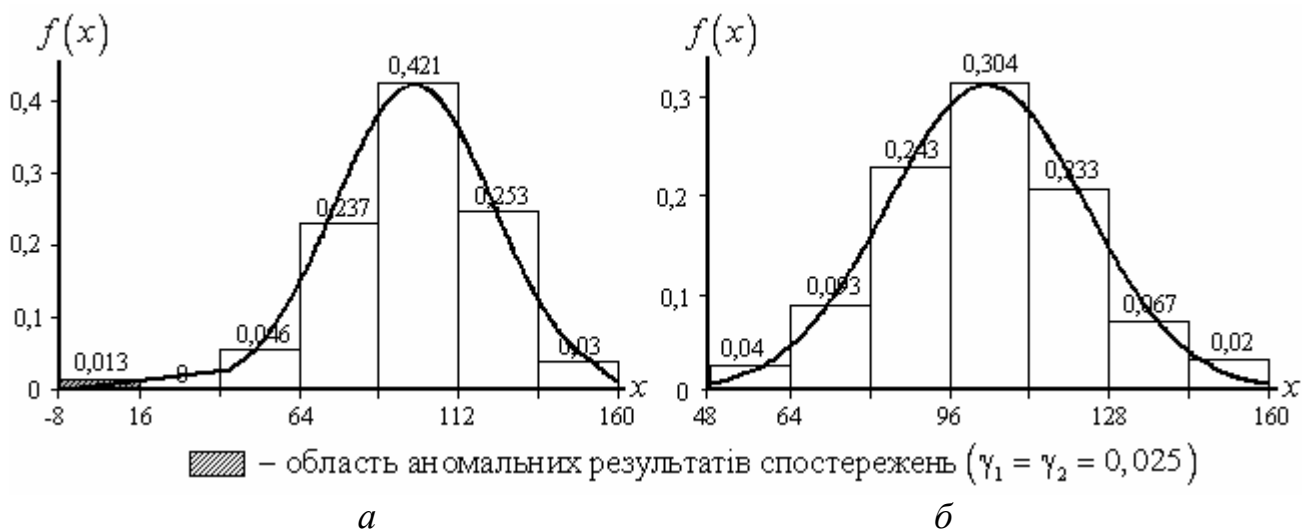


Рис. 1.4. Гістограма та графік нормованої функції щільності за наявності аномальних результатів спостережень: *a* – вихідний вигляд; *б* – після вилучення аномальних значень

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:

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

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).

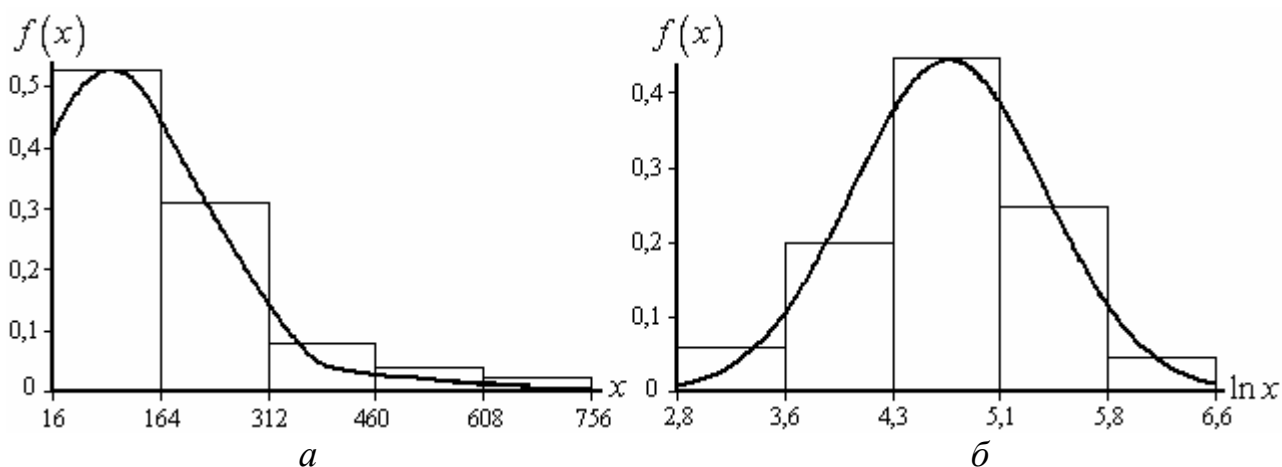


Рис. 1.5. Вигляд гістограми та графіка нормованої функції щільності асиметричного розподілу: a – вихідний; b – після застосування логарифмічного перетворення

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

$$x_l^* = x_l^c, \quad x_l^* = \frac{1}{x_l^c}, \quad x_l^* = \exp(-x_l) \text{ і т.д.},$$

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 характеристику генеральної сукупності. Параметр є невідома та фіксована величина.

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

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

яка оцінює параметр θ . При цьому вибір функції $\varphi(\cdot)$ залежить від методу знаходження оцінок параметра. Реалізація процедур, що визначають $\varphi(\cdot)$, 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

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

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:

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

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 [5].

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

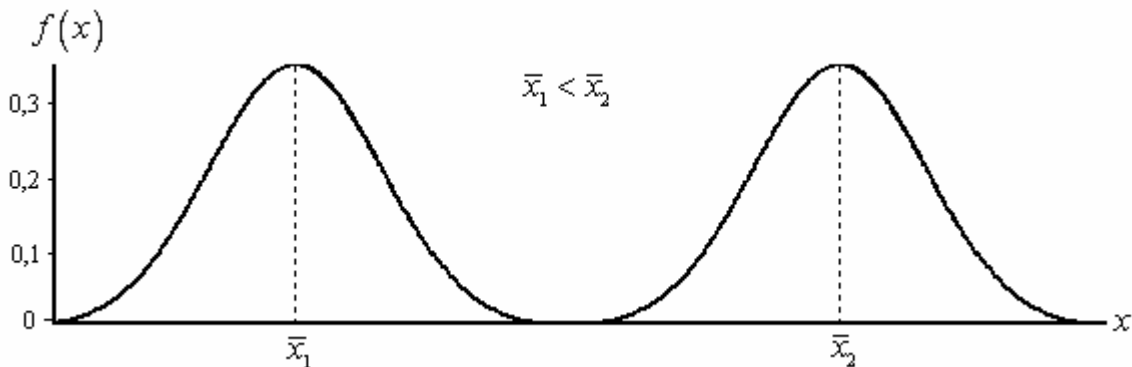


Рис. 1.6. Графік функції щільності залежно від середнього арифметичного

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

– зсунені:

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

– незсунені:

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

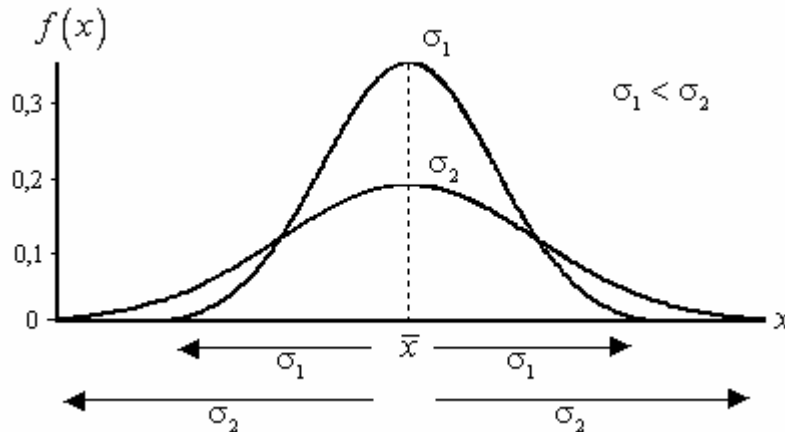


Рис. 1.7. Графік функції щільності залежно від σ

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:

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

– 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).

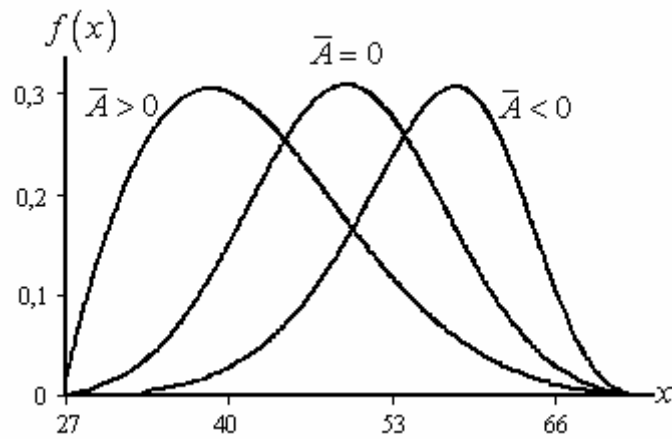


Рис. 1.8. Графік функції щільності залежно від \bar{A}

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

– зсунений:

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

– незсунений:

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

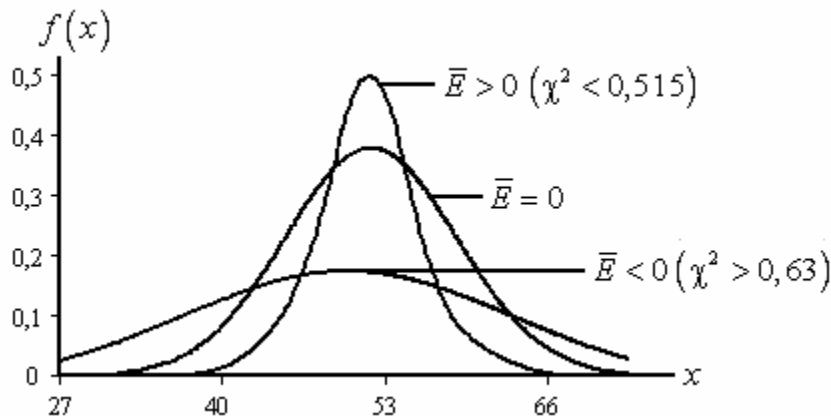


Рис. 1.9. Графік функції щільності залежно від \bar{E} та $\hat{\chi}$

Коефіцієнт контрексцесу

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

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

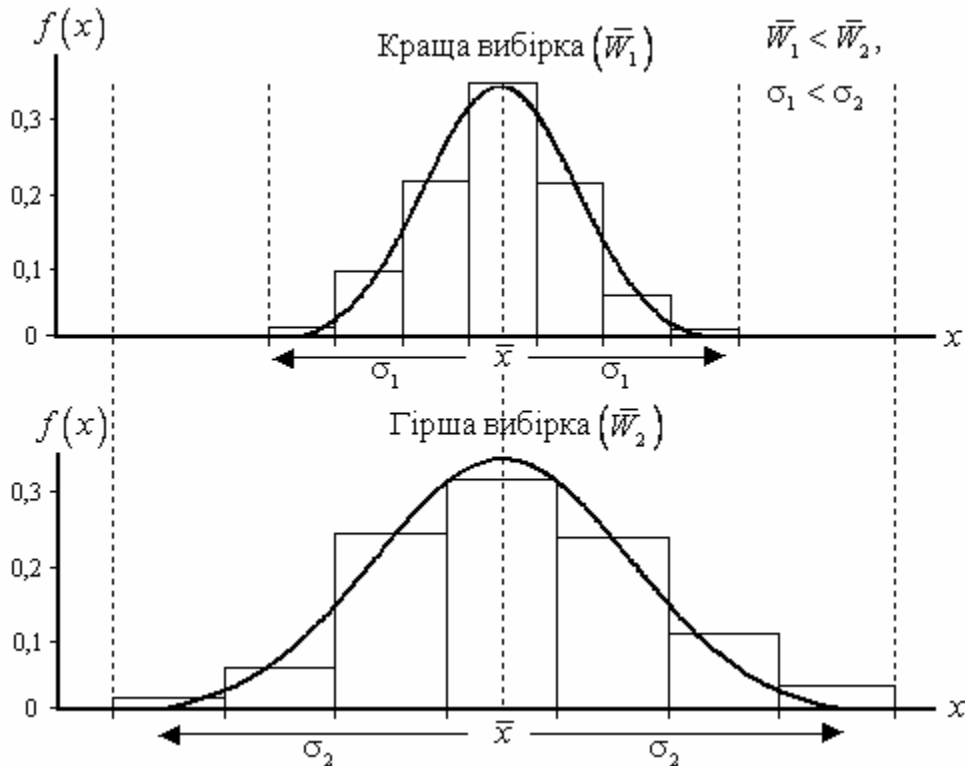


Рис. 1.10. Графік нормованої функції щільності та гістограм залежно від \bar{W}

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

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 вибірковий 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 сукупності 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	Середнє $\mu_{\bar{x}} = \mu$	Загальна сума $\mu_{\text{sum}} = \mu$
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 $\hat{\theta}_B$:

$$P\{\hat{\theta}_H < \theta < \hat{\theta}_B\} = \gamma, \quad \theta \in (\hat{\theta}_H; \hat{\theta}_B),$$

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

$$\begin{aligned} \hat{\theta}_H &= \hat{\theta} - t_{\alpha/2, v} \sqrt{D\{\hat{\theta}\}} = \hat{\theta} - t_{\alpha/2, v} \sigma\{\hat{\theta}\}, \\ \hat{\theta}_B &= \hat{\theta} + t_{\alpha/2, v} \sqrt{D\{\hat{\theta}\}} = \hat{\theta} + t_{\alpha/2, v} \sigma\{\hat{\theta}\}, \end{aligned}$$

де $t_{\alpha/2, v}$ – квантиль t -розподілу Стюдента; $v = N - 1$ (при $N > 60$ замість $t_{\alpha/2, v}$ використовують квантиль $u_{\alpha/2}$ стандартного 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

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

або шляхом призначення односторонніх довірчих інтервалів

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

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

$$\begin{aligned}\sigma\{\bar{x}\} &= \frac{S}{\sqrt{N}}, & \sigma\{S\} &= \frac{S}{\sqrt{2N}}, \\ \sigma\{\bar{A}\} &= \sqrt{\frac{6}{N}\left(1 - \frac{12}{2N+7}\right)} \quad \text{або} \quad \sigma\{\bar{A}\} = \sqrt{\frac{6(N-2)}{(N+1)(N+3)}}, \\ \sigma\{\bar{E}\} &= \sqrt{\frac{24}{N}\left(1 - \frac{225}{15N+124}\right)} \quad \text{або} \quad \sigma\{\bar{E}\} = \sqrt{\frac{24N(N-2)(N-3)}{(N+1)^2(N+3)(N+5)}}, \\ \sigma\{\hat{\chi}\} &= \sqrt{\frac{|\hat{E}|}{29N}} \sqrt[4]{|\hat{E}^2 - 1|^3}, & \sigma\{\bar{W}\} &= \bar{W} \sqrt{\frac{1+2\bar{W}^2}{2N}}.\end{aligned}$$

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.

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 – ймовірність «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: $[x_{0,05}; x_{0,95}]$.

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 невизначеності 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 ймовірністю $1 - \alpha$ 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 відрізнятись 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 $F(x; \bar{\Theta})$;
- 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 як від методу, на основі якого здійснюється обчислювальна процедура, так і 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

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

еквівалентної

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

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 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

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

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

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

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

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

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

тобто оцінка параметра є деякою функцією моментів.

Нагадаємо, що

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

Приклад 1.4. Для експоненціального розподілу є правильне

$$\nu_1 = \frac{1}{\lambda},$$

отже,

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

Приклад 1.5. Для нормального закону розподілу є слушне

$$\nu_1 = m, \quad \mu_2 = \sigma^2,$$

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 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

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

тобто з розв'язку системи рівнянь

$$\frac{\partial S_{3\text{ап}}^2}{\partial \hat{\theta}_1} = 0; \quad \frac{\partial S_{3\text{ап}}^2}{\partial \hat{\theta}_2} = 0,$$

одержують систему лінійних алгебричних рівнянь

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

де

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

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

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

Приклад 1.6. Експоненціальний розподіл

$$F(x) = 1 - \exp(-\lambda x)$$

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

$$\frac{\partial S_{\text{Зал}}^2}{\partial \lambda} = \frac{-2}{N-2} \sum_{l=1}^{N-1} (z_l - \hat{\lambda} t_l) t_l = 0,$$

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

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

визначають

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

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 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 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 здійсненна: 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

$$D\{\theta_i\} = \left(\frac{\partial H_i}{\partial \bar{x}}\right)^2 D\{\bar{x}\} + \left(\frac{\partial H_i}{\partial x^2}\right)^2 D\{\bar{x}^2\} + 2\left(\frac{\partial H_i}{\partial \bar{x}}\right)\left(\frac{\partial H_i}{\partial x^2}\right) \text{cov}\{\bar{x}, \bar{x}^2\}, \quad i=1,2,$$

де

$$D\{\bar{x}\} = \frac{v_2 - v_1^2}{N}; \quad D\{\bar{x}^2\} = \frac{v_4 - v_2^2}{N};$$

$$\text{cov}\{\bar{x}, \bar{x}^2\} = \frac{v_3 - v_1 v_2}{N}.$$

Коваріація оцінок параметрів у ММ обчислюється на основі виразу

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

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

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

де $D\{F(x; \hat{\Theta})\}$ – оцінка дисперсії відтвореного розподілу; $u_{\alpha/2}$ – квантиль нормального розподілу.

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

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

Для **двопараметричного розподілу** має місце

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

Значення дисперсій оцінок $D\{\hat{\theta}\}$, $D\{\hat{\theta}_1\}$, $D\{\hat{\theta}_2\}$, $\text{cov}\{\hat{\theta}_1, \hat{\theta}_2\}$ обчислюють відповідно до третього пункту загальної схеми відтворення розподілу (див. с. 20).

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

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

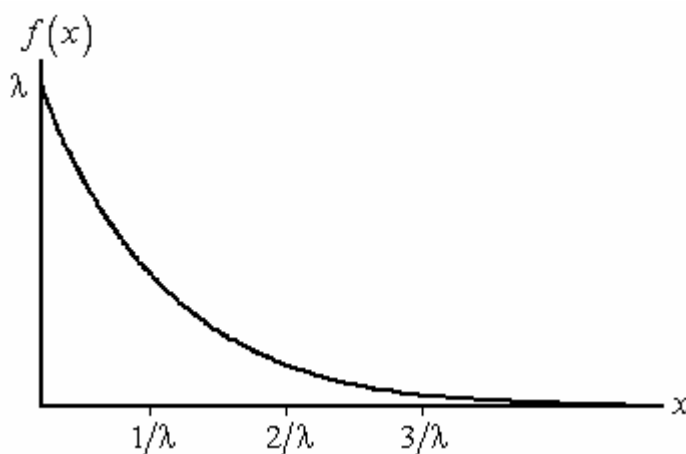


Рис. 1.11. Графік функції щільності експоненціального розподілу

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:

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

де

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

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

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

Нормальний розподіл

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

Головні характеристики нормального розподілу такі:

1) функція щільності розподілу ймовірностей (рис. 1.12, 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. 1.12, 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:

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

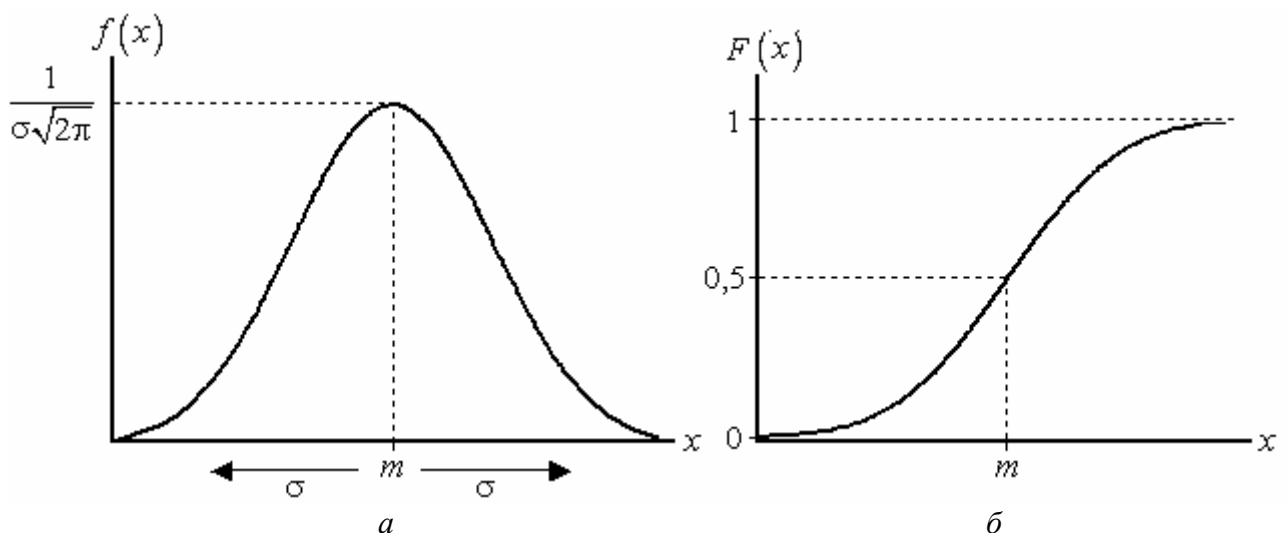


Рис. 1.12. Графіки функцій нормального розподілу:
а – функції щільності розподілу; б – функції розподілу

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

$$D\{\hat{m}\} = \frac{\hat{\sigma}^2}{N}, \quad D\{\hat{\sigma}\} = \frac{\hat{\sigma}^2}{2N}, \quad \text{cov}\{\hat{m}, \hat{\sigma}\} = 0.$$

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

$$u \geq 0; \quad t = \frac{1}{1 + \rho u}; \quad \rho = 0,231\,641\,9; \quad |\varepsilon(u)| \leq 7,8 \cdot 10^{-8};$$

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

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

У разі, якщо $u < 0$, слухне 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 $\alpha > 0,5$

$$u_\alpha = u_p, \quad \text{де } p = 1 - \alpha.$$

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, правоасиметричною або лівоасиметричною. Якщо $\beta = 1$ та $\alpha = 1/\lambda$, то розподіл Вейбулла зводиться до експоненціального.

Розподіл Вейбулла характеризують такі функції:

1) щільності розподілу ймовірностей (рис. 1.13)

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

2) розподілу ймовірностей

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

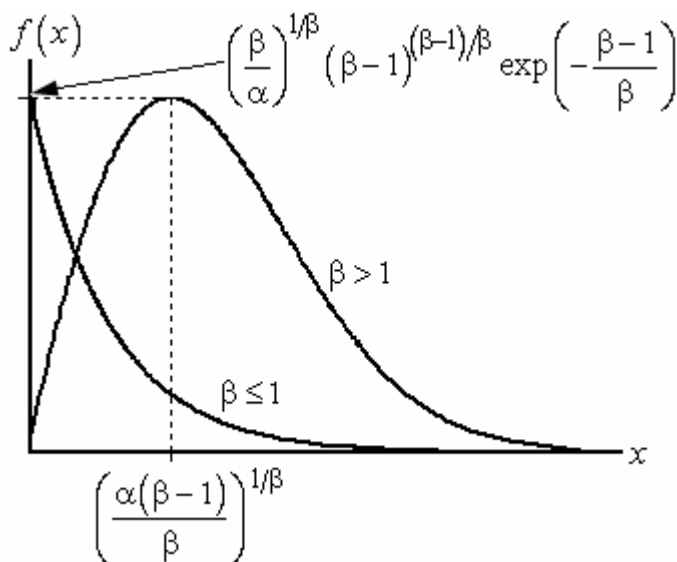


Рис. 1.13. Графік функції щільності розподілу Вейбулла

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

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

де

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

звідси

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

З урахуванням умов мінімуму необхідне розв'язання системи рівнянь

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

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:

$$D\{\hat{A}\} = \frac{a_{22}S_{\text{Зал}}^2}{a_{11}a_{22} - a_{12}a_{21}}, \quad D\{\hat{\beta}\} = \frac{a_{11}S_{\text{Зал}}^2}{a_{11}a_{22} - a_{12}a_{21}}.$$

Коваріацію визначаємо за співвідношенням

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

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

$$D\{\hat{\alpha}\} = \exp(-2\hat{A}) \cdot D\{\hat{A}\}, \quad \text{cov}\{\hat{\alpha}, \hat{\beta}\} = -\exp(\hat{A}) \cdot \text{cov}\{\hat{A}, \hat{\beta}\}.$$

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

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

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 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)

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

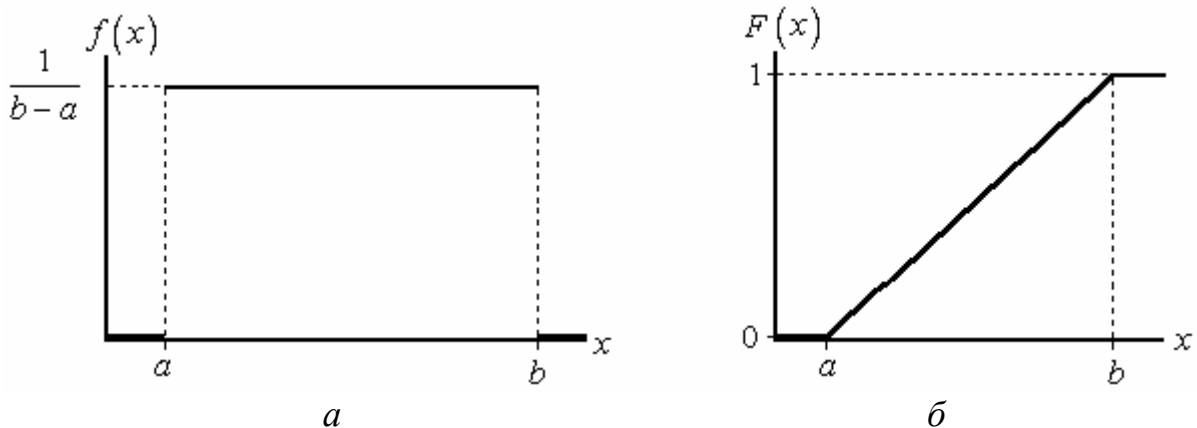


Рис. 1.14. Графік функцій рівномірного розподілу:
а – функції щільності розподілу; б – функції розподілу

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 :

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

Отже,

$$\hat{a} = \bar{x} - \sqrt{3(x^2 - \bar{x}^2)} = H_1(\bar{x}, x^2), \quad \hat{b} = \bar{x} + \sqrt{3(x^2 - \bar{x}^2)} = H_2(\bar{x}, x^2).$$

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

$$\begin{aligned} \frac{\partial H_1}{\partial \bar{x}} &= 1 + 3 \frac{\hat{a} + \hat{b}}{\hat{b} - \hat{a}}, & \frac{\partial H_1}{\partial x^2} &= -\frac{3}{\hat{b} - \hat{a}}, \\ \frac{\partial H_2}{\partial \bar{x}} &= 1 - 3 \frac{\hat{a} + \hat{b}}{\hat{b} - \hat{a}}, & \frac{\partial H_2}{\partial x^2} &= \frac{3}{\hat{b} - \hat{a}}, \\ D\{\bar{x}\} &= \frac{(\hat{b} - \hat{a})^2}{12N}, & \text{cov}\{\bar{x}, x^2\} &= \frac{(\hat{a} + \hat{b})(\hat{b} - \hat{a})^2}{12N}, \\ D\{x^2\} &= \frac{1}{180N} \left((\hat{b} - \hat{a})^4 + 15(\hat{a} + \hat{b})^2 (\hat{b} - \hat{a})^2 \right). \end{aligned}$$

Призначаючи довірчі інтервали для функції розподілу, беруть до уваги залежність

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

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 гістограмної оцінки?
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 відхилення 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. Записати функцію вибірки та умови досягнення максимуму функції вибірки для розподілу Вейбулла та нормального розподілу.
26. Записати функцію щільності нормального розподілу з параметрами $(0;1)$.
27. Схарактеризувати модель експоненціального закону розподілу та її властивості, указати приклади застосування.
28. Звести до лінійної форми функцію розподілу Вейбулла.
29. Навести модель рівномірного розподілу, графіки функцій щільності та розподілу.
30. Визначити оцінку параметра експоненціального закону розподілу за методом найменших квадратів.
31. Реалізувати метод максимальної правдоподібності для оцінки параметрів експоненціального розподілу.
32. За допомогою методу найменших квадратів знайти оцінки параметрів розподілу Вейбулла.

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 будемо говорити про неперервні випадкові 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 $\vec{\Theta}$. 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

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

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 правильною.

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 правильна (помилка другого роду);
- 3) має місце гіпотеза H_0 , проте приймається гіпотеза H_1 , яка не є правильна (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. По-справжньому 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

$$H_0 : \vec{\Theta} \in \omega_0 \text{ або } H_0 : \theta_1 = \hat{\theta}_1, \dots, \theta_s = \hat{\theta}_s$$

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 $\vec{\Theta} \in \mathfrak{w}_0$ або $\vec{\Theta} \in \mathfrak{w}_1$, 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).

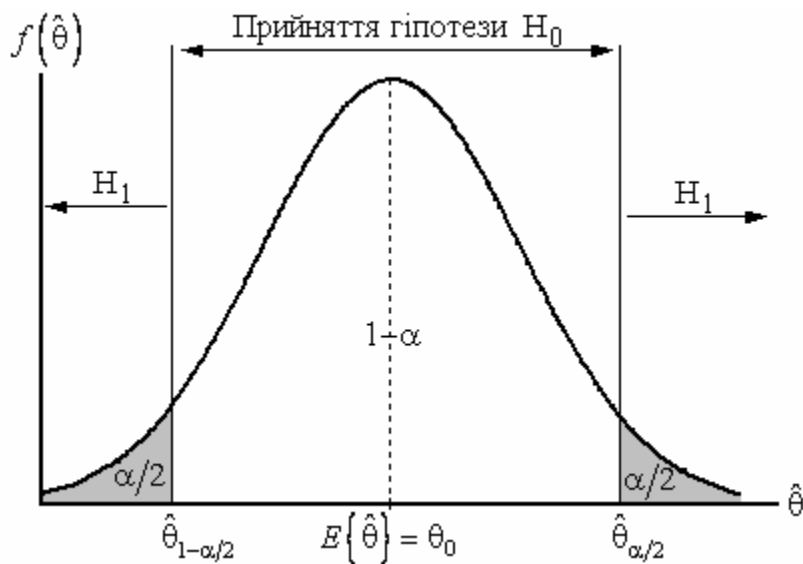


Рис. 2.1. Графік функції щільності розподілу оцінки параметра $\hat{\theta}$ за умови правильності гіпотези H_0

Тоді помилка першого роду визначається згідно з таким виразом:

$$\alpha = P\{\Omega_n \in \Omega_1 / \theta_0\} = P\{\mathfrak{w}_1 | H_0\} = P\{\hat{\theta} < \hat{\theta}_{1-\alpha/2}\} + P\{\hat{\theta} > \hat{\theta}_{\alpha/2}\} =$$

$$= \int_{-\infty}^{\hat{\theta}_{1-\alpha/2}} f(\hat{\theta}) d\hat{\theta} + \int_{\hat{\theta}_{\alpha/2}}^{\infty} f(\hat{\theta}) d\hat{\theta}.$$

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

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

Ймовірність 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 Стьюдента з кількістю степенів вільності $\nu = N - 1$ (табл. Б.2). В останньому випадку довірча ймовірність того, що оцінка $\hat{\theta}$ збігається з величиною параметра θ_0 , визначається так:

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

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

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

одержують

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

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 правильною.

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 ін-

тервал, 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

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

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 гіпотези, визначається так:

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

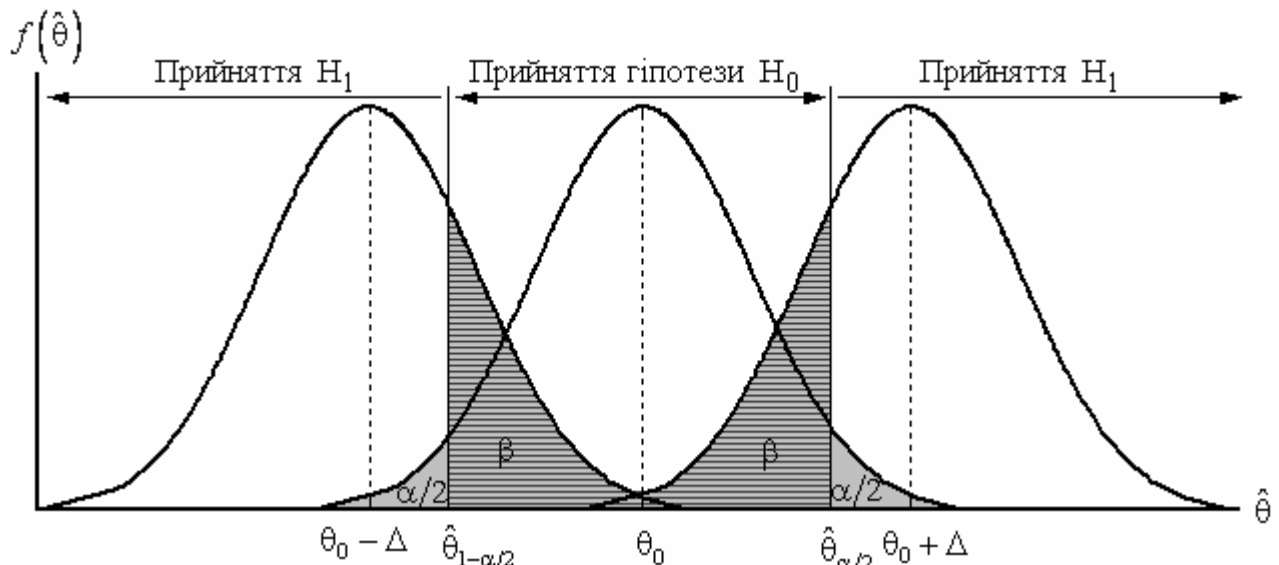


Рис. 2.2. Визначення помилок під час перевірки гіпотез

Якщо α , β – помилки 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 обгрунтованим.

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

$$H_0 : \theta < \hat{\theta} \quad \text{або} \quad H_0 : \theta > \hat{\theta}.$$

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

$$K(z) = 1 + 2 \sum_{k=1}^{\infty} (-1)^k \exp(-2k^2 z^2) \left(1 - \frac{2k^2 z}{3\sqrt{N}} - \frac{1}{18N} \left((f_1 - 4(f_1 + 3))k^2 z^2 + 8k^4 z^4 \right) + \frac{k^2 z}{27\sqrt{N^3}} \left(\frac{f_2^2}{5} - \frac{4(f_2 + 45)k^2 z^2}{15} + 8k^4 z^4 \right) \right) + O\left(\frac{z^{13}}{N^2}\right),$$

де

$$f_1 = k^2 - 0,5(1 - (-1)^k); \quad f_2 = 5k^2 + 22 - 7,5(1 - (-1)^k);$$

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

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

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

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

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 $\Omega_{1,N_2} = \{y_1, \dots, y_{N_2}\}$. 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:

$$f(x) = \frac{1}{\sigma_1 \sqrt{2\pi}} \exp\left(-\frac{(x-m_1)^2}{2\sigma_1^2}\right), \quad g(y) = \frac{1}{\sigma_2 \sqrt{2\pi}} \exp\left(-\frac{(y-m_2)^2}{2\sigma_2^2}\right).$$

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 $\Omega_{1,N_1}, \Omega_{1,N_2}$.

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

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

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 $|t| > t_{\alpha/2, \nu}$ 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 відрізнятись. 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 Стьюдента з кількістю степенів вільності $\nu = N_1 + N_2 - 2$.

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

$$S_x^2 = \frac{S_x^2}{N_1}, \quad S_y^2 = \frac{S_y^2}{N_2}.$$

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 Стьюдента з кількістю степенів вільності $\nu = N_1 + N_2 - 2$. Подальша процедура перевірки не становить труднощів.

Приклад 2.1. 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 $v_1 = N_1 - 1$ та $v_2 = N_2 - 1$. Враховуючи, що $f > 0$, за відомого α обчислюють критичне значення f_{α, v_1, v_2} (табл. Б.4) і, якщо $f \leq f_{\alpha, v_1, v_2}$, приймають головну гіпотезу.

Слід зауважити, що 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 про **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:

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

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

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

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.

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

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

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

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

$$\text{де } N = \sum_{i=1}^k N_i.$$

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 використовуваними. 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 2.2. 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 2.1. 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 2.3. 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

$$E\{W\} = \frac{N_1(N+1)}{2}, \quad D\{W\} = \frac{N_1N_2(N+1)}{12}.$$

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 передують y , то таке явище називають однією інверсією, якщо ж певному x передують 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:

$$U = \sum_{j=1}^{N_2} \sum_{i=1}^{N_1} z_{i,j}, \quad z_{i,j} = \begin{cases} 1, & \text{якщо } x_i > y_j, \\ 0, & \text{якщо } x_i \leq y_j. \end{cases}$$

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

$$E\{U\} = \frac{N_1 N_2}{2}, \quad D\{U\} = \frac{N_1 N_2 (N + 1)}{12}.$$

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 2.2. 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

$$\bar{r}_x = \frac{1}{N_1} \sum_{i=1}^{N_1} r(x_i), \quad \bar{r}_y = \frac{1}{N_2} \sum_{j=1}^{N_2} r(y_j),$$

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. У припущенні про нормальний закон розподілу результатів спостережень визначити мінімальний обсяг даних для «якісного» формування вибірки, якщо $\bar{x} = 8$; $S = 2$.
9. Навести статистику та обчислювальну схему реалізації уточненого критерію згоди Колмогорова.
10. Призначити 95%-й довірчий інтервал для одновимірної функції розподілу на основі реалізації критерію згоди Колмогорова.
11. Подати статистику та обчислювальну схему реалізації критерію згоди Пірсона. До яких варіаційних рядів застосовують цей критерій?
12. У чому полягає відмінність між залежними та незалежними вибірками в задачі перевірки однорідності двох вибірок?
13. На основі якої статистики перевіряють збіг двох дисперсій?
14. Яка гіпотеза перевіряється в однофакторному дисперсійному аналізі? Що таке міжгрупова варіація?
15. Навести процедуру реалізації критерію Бартлетта.
16. Чим відрізняються непараметричні критерії від параметричних?
17. Визначити статистики Вілкоксона та Манна–Уїтні.

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 двовимірною 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 $\Omega_{2,N} = \{(x_l, y_l); l = \overline{1, N}\}$ двовимірною випадкового вектора $\vec{\xi} = (\xi(\omega), \eta(\omega))$ 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

$$\xi(\omega): \{x_l; l = \overline{1, N}\} \quad \text{та} \quad \eta(\omega): \{y_l; l = \overline{1, N}\},$$

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 $\vec{\xi}$.

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).

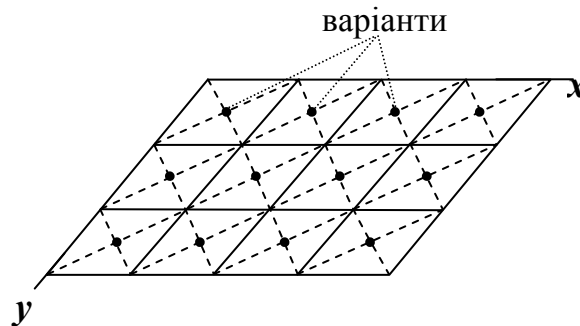


Рис. 3.1. Розбиття Δ_{h_x, h_y} площини реалізації $\vec{\zeta}$

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 $\vec{\zeta}$ у зазначеній області й має місце такий зв'язок із відносною частотою варіанти:

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

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 $\vec{\zeta}$.

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

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

Побудований таким чином варіаційний ряд можна зобразити у вигляді двовимірної гістограми випадковостей (рис. 3.2). У разі практичної реалізації достат-

ньо подавати вид зверху на гістограму випадковостей (рис. 3.3).

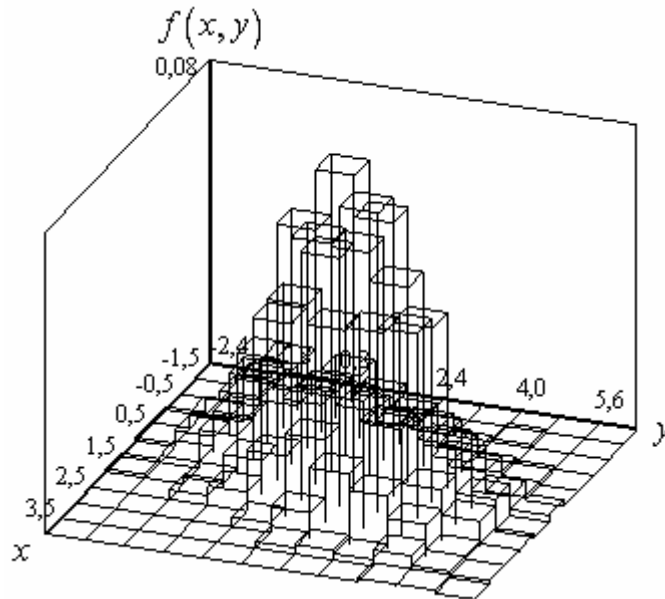


Рис. 3.2. Двовимірна гістограма випадковостей

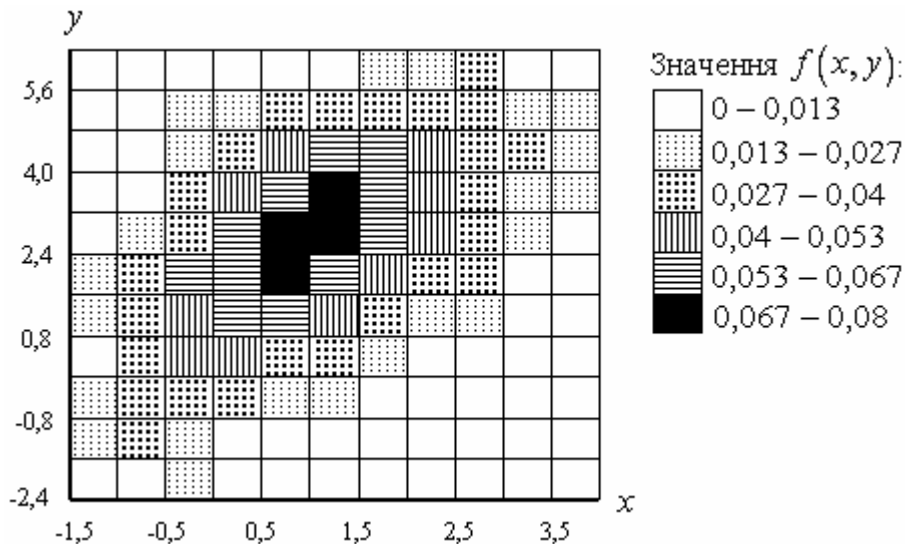


Рис. 3.3. Вид зверху на двовимірну гістограму випадковостей

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 вилучення 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 – імовірність появи аномального значення, відповідні спостереження, що входять до (i, j) -го класу, вилучаються 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 $\hat{E}\{\bar{\xi}\} = (\bar{x}, \bar{y})$, де

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

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

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

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

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

$\hat{r}_{x,y}$ – 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 $\bar{\xi} = (\xi(\omega), \eta(\omega))$ 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:

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

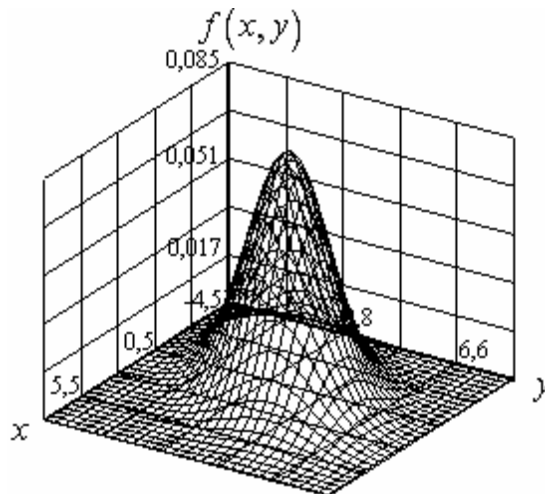


Рис. 3.4. Графік функції щільності двовимірного нормального розподілу

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 $\chi_{\alpha, \nu}^2$ розподілу χ^2 , можна говорити про адекватність відтворення двовимірного нормального розподілу як за окремими розрізами, так і за всією областю визначення в цілому.

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

$$x = m_1 + \sigma_1 z_1, \quad y = m_2 + \sigma_2 \left(z_2 \sqrt{1 - r_{x,y}^2} + z_1 r_{x,y} \right),$$

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

$$z_1 = \frac{\xi(\omega) - m_1}{\sigma_1}, \quad z_2 = \frac{\frac{\eta(\omega) - m_2}{\sigma_2} - r_{x,y} \frac{\xi(\omega) - m_1}{\sigma_1}}{\sqrt{1 - r_{x,y}^2}} \xi(\omega).$$

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 має властивості:

1) $|r| \leq 1$;

2) якщо $r = 0$, то $\xi(\omega)$ та $\eta(\omega)$ – незалежні випадкові величини;

3) при $r = 1$ 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

$$\vec{X} = \{x_l; l = \overline{1, N}\} \quad \text{та} \quad \vec{Y} = \{y_l; l = \overline{1, N}\}.$$

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, *a*), 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.

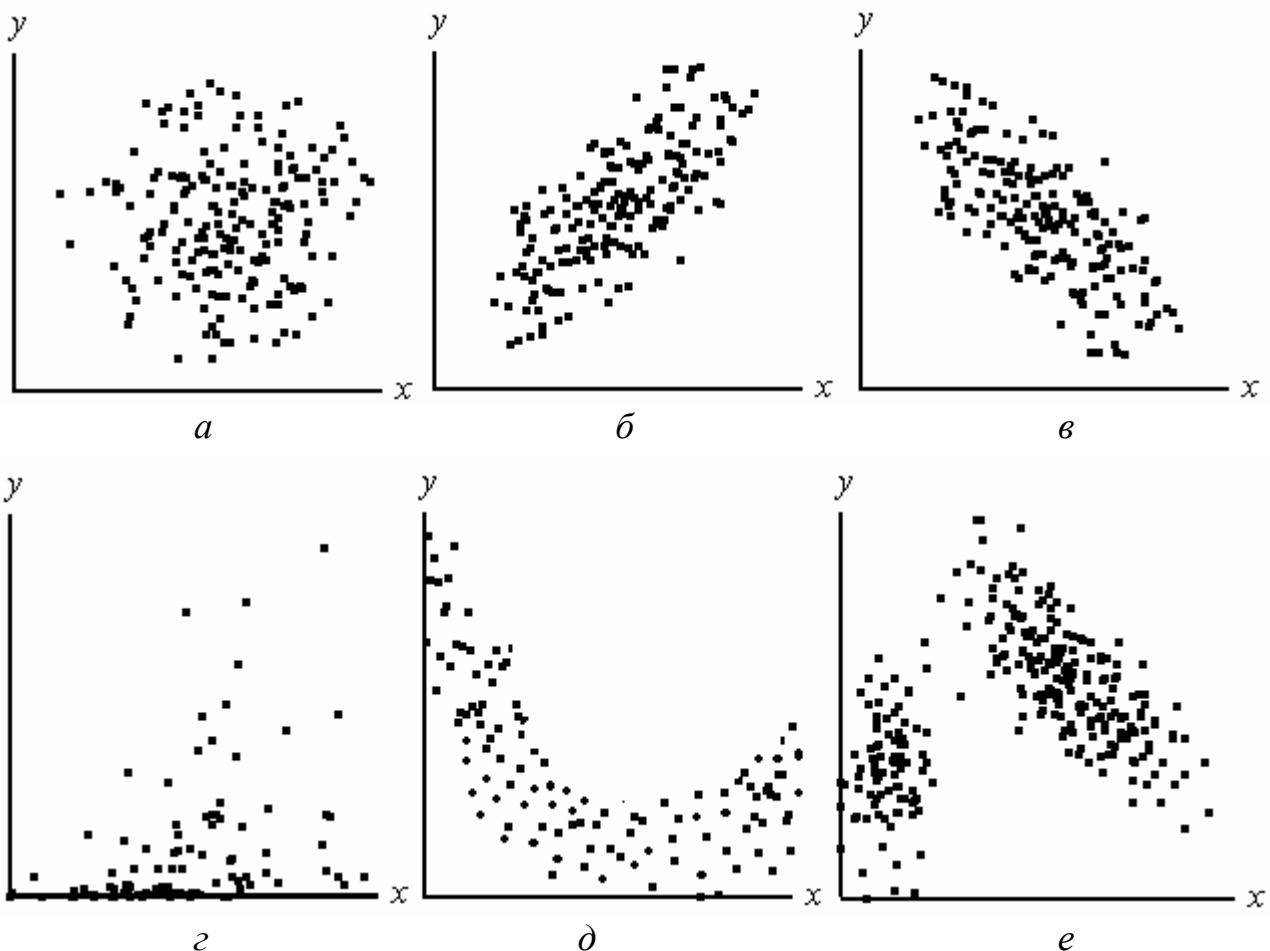


Рис. 3.5. Кореляційні поля: *a* – зв'язок відсутній; *б* – додатний лінійний зв'язок; *в* – від'ємний лінійний зв'язок; *г, д* – нелінійний зв'язок; *е* – випадок неоднорідних даних

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

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

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

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

з огляду на статистичну характеристику

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

де

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

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 правильною 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 $\{x_i, y_{i,j}; j = \overline{1, m_i}, i = \overline{1, k}\}$ 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) $\rho_{\eta/\xi} = \rho_{\xi/\eta}$;
- 3) $\rho_{\eta/\xi} \geq |r_{x,y}|$, $\rho_{\xi/\eta} \geq |r_{x,y}|$;
- 4) якщо $\rho_{\eta/\xi} = \rho_{\xi/\eta} = 0$, то кореляційний зв'язок відсутній;
- 5) якщо $\rho_{\eta/\xi} = \rho_{\xi/\eta} = |r_{x,y}|$, то поміж η та ξ існує лінійний регресійний зв'язок.

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

має нормальний розподіл із параметрами $E\{\hat{\rho}\} = \rho$, $D\{\hat{\rho}\} = \frac{1-\rho^2}{N-2}$. 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 $\{x_i, y_{i,j}; j = \overline{1, m_i}, i = \overline{1, k}\}$ 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 $A = \frac{1}{12} \sum_{j=1}^z (A_j^3 - A_j)$; $B = \frac{1}{12} \sum_{k=1}^p (B_k^3 - B_k)$;

де 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 $\tau_c = 0$ кореляція відсутня.

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;

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

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

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 $\tau_c > \tau_k$, 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.

3.3. 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 лінійної 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

$$E\{\varepsilon\} = 0; D\{\varepsilon\} = \sigma_\varepsilon^2 = \text{const}.$$

Якщо обробці підлягає масив даних $\Omega_{2,N} = \{(x_l, y_l); l = \overline{1, N}\}$, лінійна регресійна модель має вигляд

$$\bar{y}(x) = a + bx, \tag{3.2}$$

тоді оцінкою наведеної залежності є

$$\hat{y}(x) = \hat{a} + \hat{b}x,$$

де \hat{a}, \hat{b} – оцінки вектора параметрів регресії $\bar{\Theta} = \{\theta_1, \theta_2\}$ (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), уздовж якої розсіювання даних мінімальне (рис. 3.6). З огляду на це лінія, позначена пунктиром (рис. 3.6), не може бути лінією регресії.

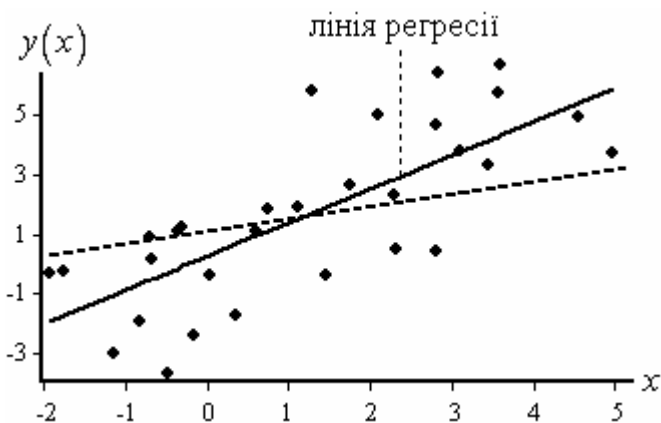


Рис. 3.6. Графік лінійної регресійної залежності

Проведення регресійного аналізу не обмежується відтворенням лінійної залежності. Можлива оцінка залежностей

$$\eta = \sum_{i=0}^s \theta_i \xi^i + \varepsilon, \tag{3.3}$$

чи будь-яких інших нелінійних залежностей:

$$\eta = \varphi(\xi; \bar{\Theta}), \bar{\Theta} = \{\theta_i; i = \overline{0, s}\}.$$

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.

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,

$$D\{y/x\} = \sigma_y^2 = const \quad (3.4)$$

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_i , y_i стохастично незалежні, таким чином, результати, одержані на i -му кроці експерименту, не пов'язані з попереднім $(i-1)$ -м кроком і не містять інформації для $(i+1)$ -го кроку.

Нижче подана ілюстрація зазначених вимог (рис. 3.7).

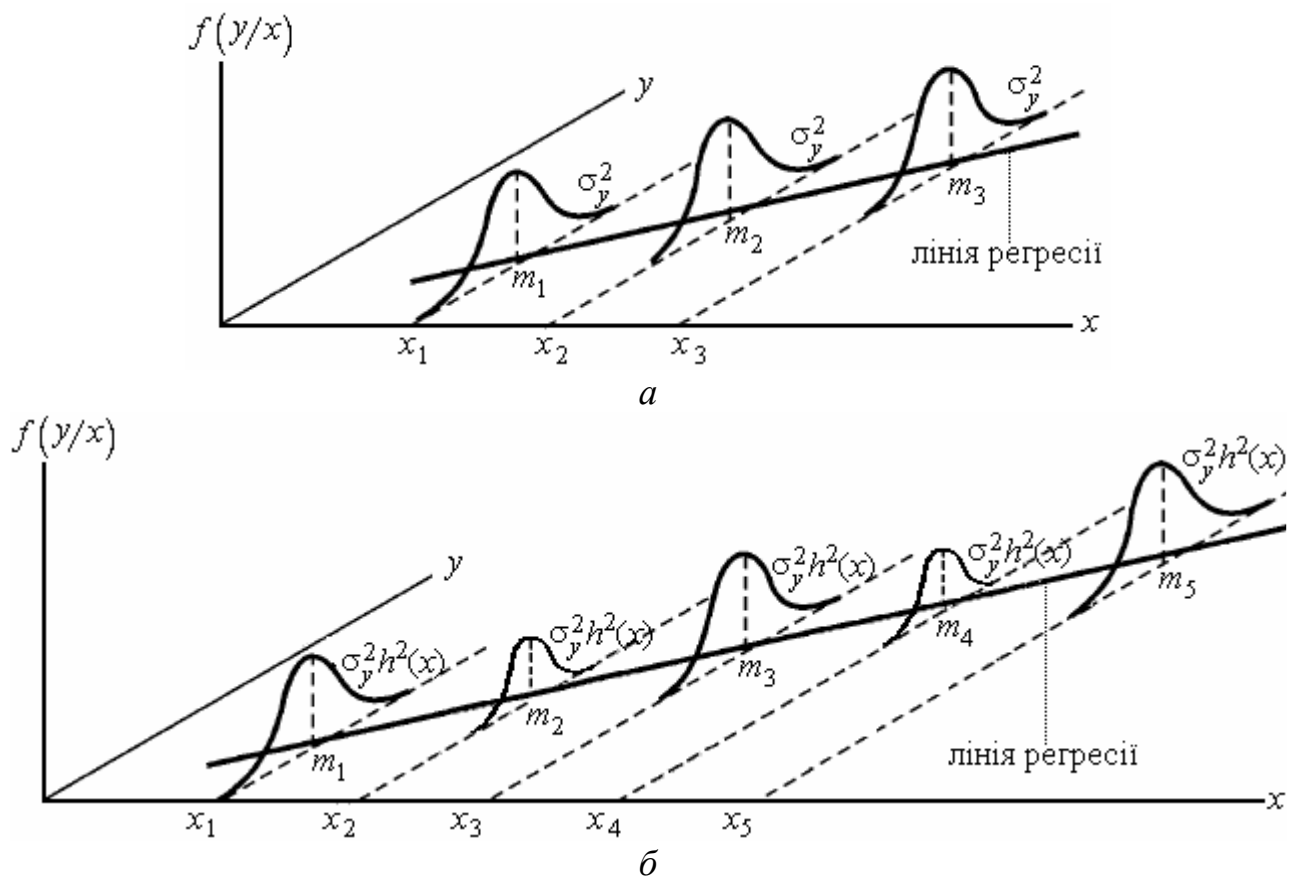


Рис. 3.7. Графічне зображення початкових умов регресійного аналізу:
а – дисперсія y стала; б – дисперсія y пропорційна $h(x)$

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 $Y = \{y_i; i = \overline{1, k}, j = \overline{1, m_i}\}$ 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

$$S'_{\bar{y}(x_i)}{}^2 = \frac{S_{\bar{y}(x_i)}^2}{h^2(x_i)}; \quad S'^2 = \frac{1}{N-k} \sum_{i=1}^k (m_i - 1) S'_{\bar{y}(x_i)}{}^2.$$

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 $\hat{r}_{x,y}$, $\hat{\rho}$ 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 ідентифікації 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

$$S_{3ал}^2 = \frac{1}{N-2} \sum_{l=1}^N (y_l - \hat{y}(x_l))^2 = \frac{1}{N-2} \sum_{l=1}^N (y_l - \hat{a} - \hat{b}x_l)^2,$$

що формується як сума квадратів відхилень результатів спостережень від лінії регресії (рис. 3.8).

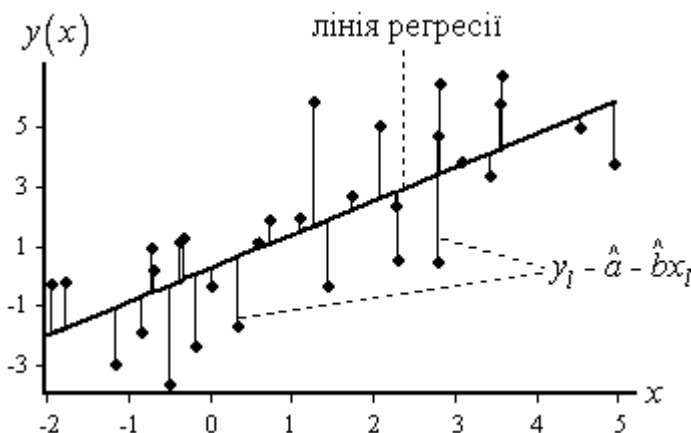


Рис. 3.8. Графічне зображення відхилення результатів спостережень від лінії регресії

Необхідна та достатня умова $\min_{a,b} S_{3ал}^2$ визначається системою лінійних рівнянь

$$\begin{cases} \frac{\partial S_{3ал}^2}{\partial \hat{a}} = 0, \\ \frac{\partial S_{3ал}^2}{\partial \hat{b}} = 0 \end{cases}$$

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

$$\hat{a} = \frac{\overline{yx^2} - \bar{x}\overline{xy}}{x^2 - \bar{x}^2}, \quad \hat{b} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{x^2 - \bar{x}^2},$$

тобто

$$\hat{a} = \bar{y} - \hat{b}\bar{x}, \quad \hat{b} = r_{x,y} \frac{\hat{\sigma}_y}{\hat{\sigma}_x}.$$

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

$$\min_{\hat{a}, \hat{b}} S_{3ал}^2 = \min_{\hat{a}, \hat{b}} \frac{1}{N-2} \sum_{i=1}^k \sum_{j=1}^{m_i} (y_{i,j} - \hat{a} - \hat{b}x_i)^2,$$

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

$$\hat{a} = \bar{y} - \hat{b} \bar{x}, \quad \hat{b} = \hat{r} \frac{\sigma_y}{\sigma_x},$$

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

$$\min_{\hat{a}, \hat{b}} S_{\text{Зал}}^2 = \min_{\hat{a}, \hat{b}} \frac{1}{N-2} \sum_{i=1}^k \sum_{j=1}^{m_i} \omega_i (y_{i,j} - \hat{a} - \hat{b} x_i)^2.$$

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 $D\{y/x\} = \sigma_y^2 = const$ 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 $r^2 \in [0;1]$ 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

$$H_0 : a = 0, \quad H_0 : b = 0$$

та гіпотез про рівність оцінок деяким значенням параметрів

$$H_0 : a = \hat{a}, \quad H_0 : b = \hat{b}.$$

Зазначені гіпотези перевіряються на основі t -тесту з урахуванням середньоквадратичних оцінок параметрів \hat{a} , \hat{b} :

$$S_a = S_{\text{Зал}} \sqrt{\frac{1}{N} + \frac{\bar{x}^2}{\sigma_x^2 (N-1)}}, \quad S_b = \frac{S_{\text{Зал}}}{\sigma_x \sqrt{N-1}}.$$

Тоді відповідні t -статистики, як завжди, дорівнюють

$$t_a = \frac{\hat{a} - a}{S_a}, \quad t_b = \frac{\hat{b} - b}{S_b}.$$

Слід відзначити, що в разі спростування гіпотези $H_0 : b = 0$ (невиконання нерівності $|t_b| \leq t_{\alpha/2, \nu}$, $\nu = N - 2$) говорять про значущість регресійного зв'язку.

Інтервальне оцінювання параметрів лінійної регресії здійснюють, виходячи з нерівностей ($\nu = N - 2$)

$$\begin{aligned} \hat{a} - t_{\alpha/2, \nu} S_a &\leq a \leq \hat{a} + t_{\alpha/2, \nu} S_a, \\ \hat{b} - t_{\alpha/2, \nu} S_b &\leq b \leq \hat{b} + t_{\alpha/2, \nu} S_b. \end{aligned}$$

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).

Толерантні межі $\hat{y}_{\min}(x)$, $\hat{y}_{\max}(x)$ для y_l , $l = \overline{1, N}$ визначаються за виразами:

$$\begin{aligned} \hat{y}_{\min}(x) &= \hat{y}(x) - t_{\alpha/2, \nu} \hat{\sigma}_\varepsilon, \\ \hat{y}_{\max}(x) &= \hat{y}(x) + t_{\alpha/2, \nu} \hat{\sigma}_\varepsilon \end{aligned}$$

де $\nu = N - 2$.

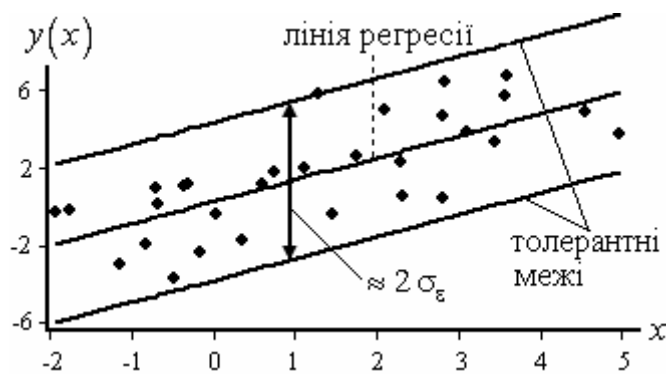


Рис. 3.9. Графічне зображення толерантних меж для лінійної регресії

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, оцінюються середні 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

$$\hat{y}(x) - t_{\alpha/2, v} S_{(\bar{y}|x)} \leq \bar{y}(x) \leq \hat{y}(x) + t_{\alpha/2, v} S_{(\bar{y}|x)},$$

де $v = N - 2$.

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.

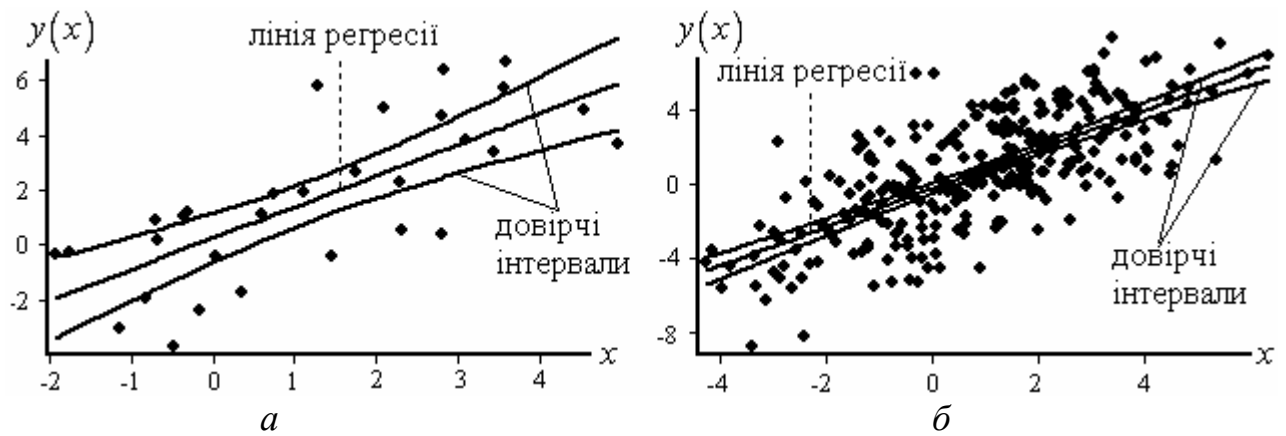


Рис. 3.11. Графічне зображення інтервального оцінювання лінійної регресії:
 $a - N = 30$; $b - N = 300$

Для наочності нижче наведені толерантні межі, довірчі інтервали для лінії регресії та прогнозного значення (рис. 3.12).

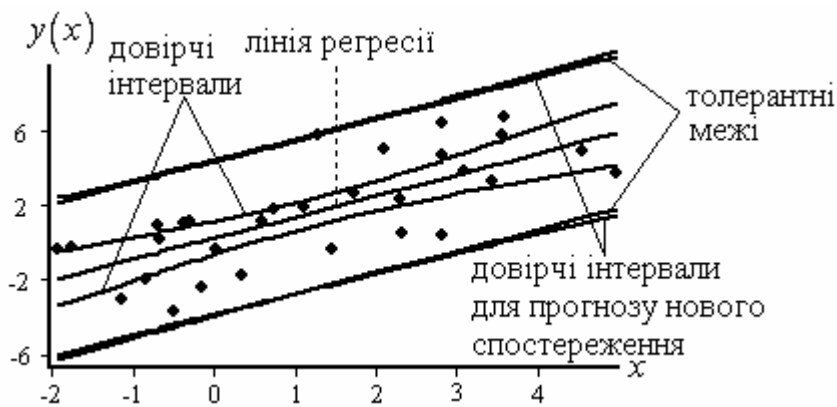


Рис. 3.12. Графічне зображення довірчого оцінювання лінійної регресії

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 $D\{y/x\} = \sigma^2 = const$ 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 $\nu_1 = k-s-1$, $\nu_2 = N-k$. Процедура перевірки гіпотези є еквівалентна вищенаведеним.

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:

$$\hat{y}_1(x) = \hat{a}_1 + \hat{b}_1(x - \bar{x}_1), \quad \hat{y}_2(x) = \hat{a}_2 + \hat{b}_2(x - \bar{x}_2),$$

залишкова дисперсія для яких відповідно визначається так:

$$S_{1, \text{Зал}}^2 = \frac{1}{N_1 - 2} \sum_{l=1}^{N_1} (y_{1,l} - \hat{a}_1 - \hat{b}_1(x_{1,l} - \bar{x}_1))^2,$$

$$S_{2, \text{Зал}}^2 = \frac{1}{N_2 - 2} \sum_{l=1}^{N_2} (y_{2,l} - \hat{a}_2 - \hat{b}_2(x_{2,l} - \bar{x}_2))^2.$$

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

$$f = \begin{cases} \frac{S_{1,Зал}^2}{S_{2,Зал}^2}, & \text{якщо } S_{1,Зал}^2 > S_{2,Зал}^2, \\ \frac{S_{2,Зал}^2}{S_{1,Зал}^2}, & \text{якщо } S_{1,Зал}^2 < S_{2,Зал}^2, \end{cases}$$

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:

$$S^2 = \frac{(N_1 - 2)S_{1,Зал}^2 + (N_2 - 2)S_{2,Зал}^2}{N_1 + N_2 - 4}.$$

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 $S_{1, \text{Зал}}^2$, $S_{2, \text{Зал}}^2$ 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

$$t = \frac{\hat{b}_1 - \hat{b}_2}{S \sqrt{\frac{S_{1, \text{Зал}}^2}{N_1 \hat{\sigma}_{x_1}^2} + \frac{S_{2, \text{Зал}}^2}{N_2 \hat{\sigma}_{x_2}^2}}},$$

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

$$C_0 = \frac{S_{1, \text{Зал}}^2}{N_1 \hat{\sigma}_{x_1}^2} \left/ \left(\frac{S_{1, \text{Зал}}^2}{N_1 \hat{\sigma}_{x_1}^2} + \frac{S_{2, \text{Зал}}^2}{N_2 \hat{\sigma}_{x_2}^2} \right) \right.; \quad [\cdot] - \text{ціла частина.}$$

Якщо $|t| \leq t_{\alpha/2, v}$, то правильна гіпотеза про збіг кутових коефіцієнтів ліній регресій.

Нижчезрозглянута 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

$$\hat{b} = \left(\hat{b}_1 \frac{N_1 \hat{\sigma}_{x_1}^2}{S_{1,3ал}^2} + \hat{b}_2 \frac{N_2 \hat{\sigma}_{x_2}^2}{S_{2,3ал}^2} \right) / \left(\frac{N_1 \hat{\sigma}_{x_1}^2}{S_{1,3ал}^2} + \frac{N_2 \hat{\sigma}_{x_2}^2}{S_{2,3ал}^2} \right); \quad \hat{b}_0 = \frac{\bar{y}_1 - \bar{y}_2}{\bar{x}_1 - \bar{x}_2};$$

$$S_{10}^2 = \frac{N_2 S_{1,3ал}^2 + N_1 S_{2,3ал}^2}{N_1 N_2 (\bar{x}_1 - \bar{x}_2)^2} + \frac{S_{1,3ал}^2 S_{2,3ал}^2}{N_1 \hat{\sigma}_{x_1}^2 S_{2,3ал}^2 + N_2 \hat{\sigma}_{x_2}^2 S_{1,3ал}^2}.$$

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:

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.

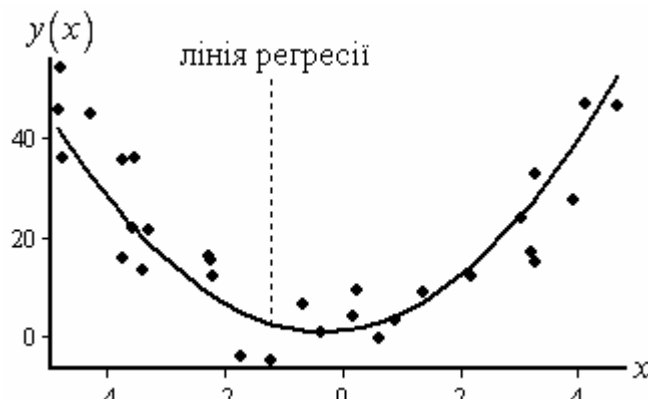


Рис. 3.13. Графік поліноміальної регресійної залежності другого порядку

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.

Обчислювальні процедури, що ґрунтуються на МНК, вводяться для регресії (3.9), яку подають у вигляді

$$\bar{y}(x) = a_1 + b_1\varphi_1(x) + c_1\varphi_2(x), \quad (3.11)$$

де

$$\varphi_1(x) = x - \bar{x};$$

$$\varphi_2(x) = x^2 - \frac{\sum_{l=1}^N x_l^3 - \bar{x} \sum_{l=1}^N x_l^2}{\sum_{l=1}^N x_l^2 - N\bar{x}^2} (x - \bar{x}) + \frac{1}{N} \sum_{l=1}^N x_l^2 = x^2 - \frac{\overline{x^3} - \bar{x}^2 \bar{x}}{\sigma_x^2} (x - \bar{x}) - \bar{x}^2.$$

Нижче наведені процедури відтворення залежностей (3.9), (3.11) 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

$$\frac{\partial S_{3ал(1)}^2}{\partial \hat{a}} = 0, \quad \frac{\partial S_{3ал(1)}^2}{\partial \hat{b}} = 0, \quad \frac{\partial S_{3ал(1)}^2}{\partial \hat{c}} = 0,$$

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:

$$\begin{pmatrix} \hat{\sigma}_x^2 & (\overline{x^3 - x^2\bar{x}}) \\ (\overline{x^3 - x^2\bar{x}}) & (\overline{x^4 - (x^2)^2}) \end{pmatrix} \begin{pmatrix} \hat{b} \\ \hat{c} \end{pmatrix} = \begin{pmatrix} \hat{r}_{x,y} \hat{\sigma}_x \hat{\sigma}_y \\ \overline{(y - \bar{y})(x^2 - \bar{x}^2)} \end{pmatrix}, \quad (3.12)$$

де

$$\overline{x^k} = \frac{1}{N} \sum_{i=1}^N x_i^k, \quad k = 1, 2, 3, 4;$$

$$\overline{(y - \bar{y})(x^2 - \bar{x}^2)} = \frac{1}{N} \sum_{l=1}^N (x_l^2 - \bar{x}^2)(y_l - \bar{y}).$$

Із розв'язку системи (3.12) знаходять оцінки параметрів регресії \hat{b} , \hat{c} :

$$\hat{b} = \frac{(\overline{x^4 - (x^2)^2}) \hat{r}_{x,y} \hat{\sigma}_x \hat{\sigma}_y - (\overline{x^3 - x^2\bar{x}}) \overline{(y - \bar{y})(x^2 - \bar{x}^2)}}{\hat{\sigma}_x^2 (\overline{x^4 - (x^2)^2}) - (\overline{x^3 - x^2\bar{x}})^2},$$

$$\hat{c} = \frac{\hat{\sigma}_x^2 \overline{(y - \bar{y})(x^2 - \bar{x}^2)} - (\overline{x^3 - x^2\bar{x}}) \hat{r}_{x,y} \hat{\sigma}_x \hat{\sigma}_y}{\hat{\sigma}_x^2 (\overline{x^4 - (x^2)^2}) - (\overline{x^3 - x^2\bar{x}})^2}.$$

Наведені вирази і визначають обчислювальну процедуру відтворення параболічної регресії у вигляді (3.9).

Найпростіша обчислювальна схема відтворення поліноміальної регресії ґрунтується на ортогональних поліномах Чебишева, окремим випадком якої є залежність (3.11). З умови

$$\min_{\hat{a}_1, \hat{b}_1, \hat{c}_1} S_{3ал(2)}^2 = \min_{\hat{a}_1, \hat{b}_1, \hat{c}_1} \frac{1}{N-3} \sum_{l=1}^N (y_l - \hat{a}_1 - \hat{b}_1 \varphi_1(x_l) - \hat{c}_1 \varphi_2(x_l))^2$$

знаходять оцінки параметрів:

$$\hat{a}_1 = \frac{1}{N} \sum_{l=1}^N y_l = \bar{y},$$

$$\hat{b}_1 = \frac{\sum_{l=1}^N (x_l - \bar{x}) y_l}{\sum_{l=1}^N (x_l - \bar{x})^2} = \frac{\overline{(x - \bar{x})y}}{\hat{\sigma}_x^2}, \quad (3.13)$$

$$\hat{c}_1 = \frac{\sum_{l=1}^N \varphi_2(x_l) y_l}{\sum_{l=1}^N \varphi_2^2(x_l)} = \frac{\overline{\varphi_2(x)y}}{\overline{\varphi_2^2(x)}}.$$

З аналізу формули (3.13) випливає, що оцінки \hat{a}_1 , \hat{b}_1 повністю збігаються з оцінками для лінійної регресії у вигляді

$$\bar{y}(x) = a + b(x - \bar{x}),$$

що визначається властивостями поліномів Чебишева. Іншими словами, підвищуючи степінь полінома, для кожної приєднаної функції $\varphi_k(x)$ обчислюють коефіцієнт регресії, зберігаючи одержані раніше параметри.

Оцінка точності та значущості параметрів \hat{a}_1 , \hat{b}_1 , \hat{c}_1 , як і для лінійної регресії, проводиться шляхом перевірки гіпотез

$$H_0 : a_1 = \hat{a}_1, \quad H_0 : b_1 = \hat{b}_1, \quad H_0 : c_1 = \hat{c}_1$$

на основі статистик

$$\begin{aligned} t_{a_1} &= \frac{\hat{a}_1 - a_1}{S_{\text{Зал}(2)}} \sqrt{N}, \\ t_{b_1} &= \frac{\hat{b}_1 - b_1}{S_{\text{Зал}(2)}} \sqrt{\sum_{l=1}^N \varphi_1^2(x_l)} = \frac{(\hat{b}_1 - b_1) \sigma_x}{S_{\text{Зал}(2)}} \sqrt{N}, \\ t_{c_1} &= \frac{\hat{c}_1 - c_1}{S_{\text{Зал}(2)}} \sqrt{\sum_{l=1}^N \varphi_2^2(x_l)} = \frac{(\hat{c}_1 - c_1)}{S_{\text{Зал}(2)}} \sqrt{N \varphi_2^2(x)}. \end{aligned} \quad (3.14)$$

Значущість оцінок параметрів перевіряють, вважаючи $a_1 = 0$, $b_1 = 0$, $c_1 = 0$, на основі умови $|t_{a_1}| \leq t_{\alpha/2, v}$, $|t_{b_1}| \leq t_{\alpha/2, v}$, $|t_{c_1}| \leq t_{\alpha/2, v}$. Якщо хоча б одна з нерівностей порушується, говорять про «втрату» відповідного члена параболі.

З урахуванням статистичних характеристик (3.14) проводять інтервальне оцінювання відповідних коефіцієнтів регресії:

$$\begin{aligned} a_{\text{H,B}} &= \hat{a}_1 \mp t_{\alpha/2, v} \frac{S_{\text{Зал}(2)}}{\sqrt{N}}, \\ b_{\text{H,B}} &= \hat{b}_1 \mp t_{\alpha/2, v} \frac{S_{\text{Зал}(2)}}{\sigma_x \sqrt{N}}, \\ c_{\text{H,B}} &= \hat{c}_1 \mp t_{\alpha/2, v} \frac{S_{\text{Зал}(2)}}{\sqrt{N \varphi_2^2(x)}}. \end{aligned}$$

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

$$\begin{aligned} \hat{y}_{\min}(x) &= \hat{a}_1 + \hat{b}_1 \varphi_1(x) + \hat{c}_1 \varphi_2(x) - t_{\alpha/2, v} S_{\text{Зал}(2)}, \\ \hat{y}_{\max}(x) &= \hat{a}_1 + \hat{b}_1 \varphi_1(x) + \hat{c}_1 \varphi_2(x) + t_{\alpha/2, v} S_{\text{Зал}(2)}, \end{aligned}$$

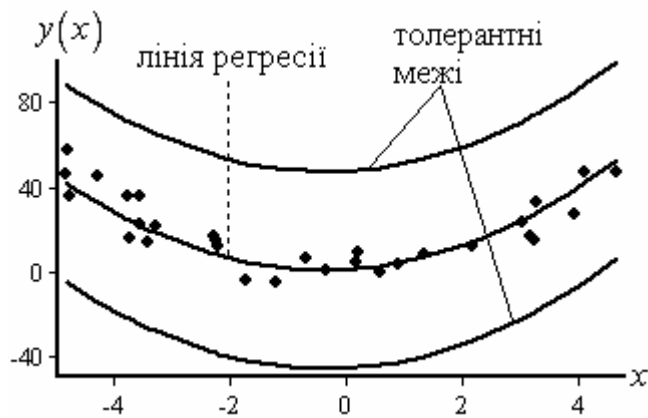


Рис. 3.14. Графічне зображення толерантних меж для параболічної регресії

Відхилення оцінки регресії $\hat{y}(x)$ від теоретичної оцінюють на основі статистичної характеристики

$$t(x) = \frac{\hat{y}(x) - \bar{y}(x)}{S_{(\bar{y}|x)}}$$

де (за повною аналогією з лінійною моделлю)

$$S_{(\bar{y}|x)} = \sqrt{\frac{1}{N} \hat{\sigma}_\varepsilon^2 + S_{b_1}^2 \varphi_1^2(x) + S_{c_1}^2 \varphi_2^2(x)} = \frac{S_{3ал(2)}}{\sqrt{N}} \sqrt{1 + \frac{\varphi_1^2(x)}{\sigma_x^2} + \frac{\varphi_2^2(x)}{\varphi_2^2(x)}};$$

$$\hat{\sigma}_\varepsilon^2 = S_{3ал(2)}^2; \quad S_{b_1}^2 = \frac{S_{3ал(2)}^2}{N\sigma_x^2}; \quad S_{c_1}^2 = \frac{S_{3ал(2)}^2}{N\varphi_2^2(x)}.$$

Якщо $|t(x)| \leq t_{\alpha/2, \nu}$, де $\nu = N - 3$, то правильна гіпотеза

$$H_0: \bar{y}(x) = \hat{y}(x)$$

і проводиться **інтервальне оцінювання параболічної регресії** (рис. 3.15). Межі довірчого інтервалу визначаються так:

$$\hat{y}_{н,в}(x) = \hat{y}(x) \mp t_{\alpha/2, \nu} S_{(\bar{y}|x)}.$$

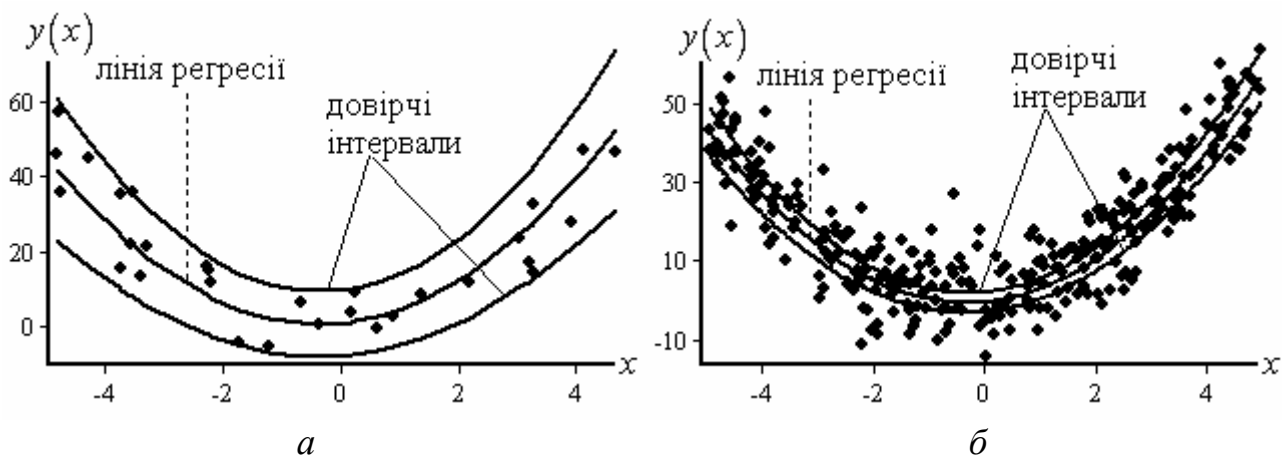


Рис. 3.15. Графічне зображення інтервального оцінювання параболічної регресії: а – $N = 30$; б – $N = 300$

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 :

$$S_{(y|x_0)} = \sqrt{\hat{\sigma}_\varepsilon^2 \left(1 + \frac{1}{N}\right) + S_{b_1}^2 \varphi_1^2(x) + S_{c_1}^2 \varphi_2^2(x) = \frac{S_{\text{Зал}(2)}}{\sqrt{N}} \sqrt{N + 1 + \frac{\varphi_1^2(x)}{\sigma_x^2} + \frac{\varphi_2^2(x)}{\varphi_2^2(x)}}.$$

Відповідний довірчий інтервал для нового спостереження у при заданому x_0 (рис. 3.16) такий:

$$\hat{y}(x_0) - t_{\alpha/2, v} S_{(y|x_0)} \leq y \leq \hat{y}(x_0) + t_{\alpha/2, v} S_{(y|x_0)}, \quad v = N - 3.$$

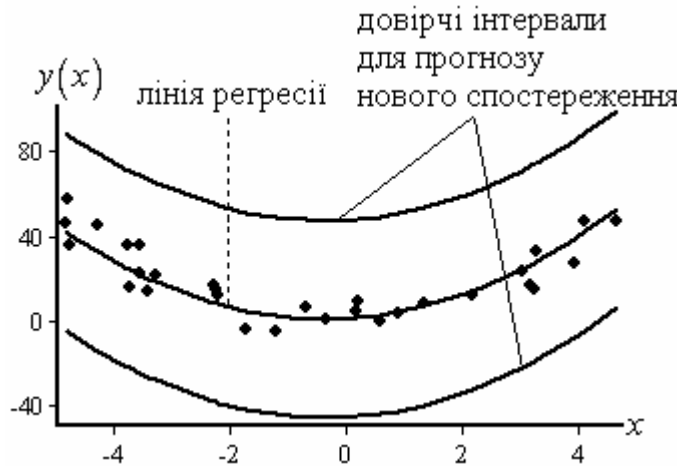


Рис. 3.16. Графічне зображення довірчого інтервалу для прогнозу нового спостереження у випадку параболічної регресії

Нижче для наочності показані толерантні межі, довірчі інтервали для лінії регресії та прогнозного значення (рис. 3.17).

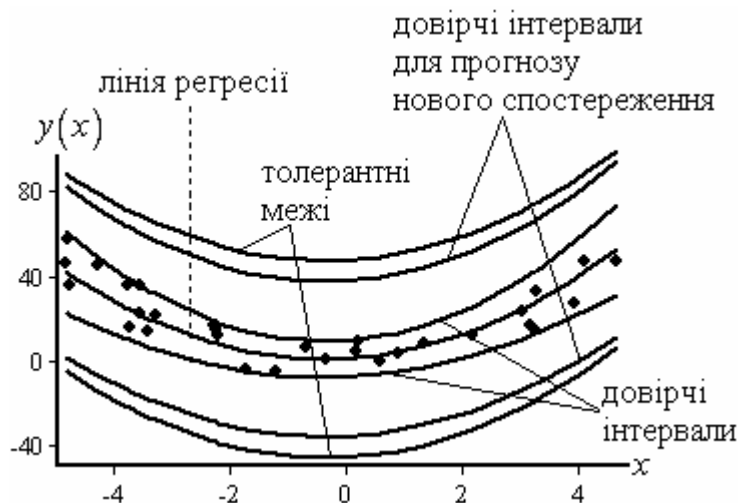


Рис. 3.17. Графічне зображення довірчого оцінювання параболічної регресії

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. Сформулювати постановку задачі на проведення лінійного регресійного аналізу. Перерахувати початкові умови регресійного аналізу.
12. Описати процедуру відтворення лінійної регресії за МНК.
13. Навести дисперсії оцінок параметрів лінійної моделі регресії.
14. Яка оцінка дозволяє вказати стандартну похибку регресійної оцінки?
15. У чому полягає різниця в побудові довірчих інтервалів для лінії регресії та прогнозу нового спостереження?
16. На основі якої статистики здійснюється перевірка адекватності відтворення моделі регресії?
17. Визначити довірчий інтервал для параболічної регресії.
18. Як перевіряється значущість вільного члена параболічної моделі регресії?
19. Перевірити гіпотезу про збіг двох регресійних прямих:
 $\bar{y}_1(x) = 7,2 + 0,52x$; $N = 65$; $S_{1,3ал} = 11$; $S_{x_1} = 4$; $\bar{x}_1 = 4,8$; $\bar{y}_1 = 12$;
 $\bar{y}_2(x) = 7,7 + 0,38x$; $N = 100$; $S_{2,3ал} = 9$; $S_{x_2} = 5$; $\bar{x}_2 = 5$; $\bar{y}_2 = 13$.

ДОДАТОК А

Процедури знаходження квантилів

У статистичному аналізі, а саме в задачах перевірки статистичних гіпотез, виникає необхідність знаходження квантилів розподілів. Найбільш поширені є квантілі розподілів: нормального, Стюдента, Пірсона та Фішера. Визначення квантіля 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;

$$g_1(u_{\alpha/2}) = \frac{1}{4}(u_{\alpha/2}^3 + u_{\alpha/2}); \quad g_2(u_{\alpha/2}) = \frac{1}{96}(5u_{\alpha/2}^5 + 16u_{\alpha/2}^3 + 3u_{\alpha/2});$$

$$g_3(u_{\alpha/2}) = \frac{1}{384}(3u_{\alpha/2}^7 + 19u_{\alpha/2}^5 + 17u_{\alpha/2}^3 - 15u_{\alpha/2});$$

$$g_4(u_{\alpha/2}) = \frac{1}{92160}(79u_{\alpha/2}^9 + 779u_{\alpha/2}^7 + 1482u_{\alpha/2}^5 - 1920u_{\alpha/2}^3 - 945u_{\alpha/2}).$$

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 квантіля f_{α, ν_1, ν_2} розподілу Фішера (табл. Б.4) можна застосовувати такий вираз:

$$f_{\alpha, \nu_1, \nu_2} = \exp(2z),$$

де

$$z = u_{\alpha} \sqrt{\frac{\sigma}{2}} - \frac{1}{6} \delta (u_{\alpha}^2 + 2) + \sqrt{\frac{\sigma}{2}} \left(\frac{\sigma}{24} (u_{\alpha}^2 + 3u_{\alpha}) + \frac{1}{72} \frac{\delta^2}{\sigma} (u_{\alpha}^3 + 11u_{\alpha}) \right) -$$

$$- \frac{\delta \sigma}{120} (u_{\alpha}^4 + 9u_{\alpha}^2 + 8) + \frac{\delta^3}{3240\sigma} (3u_{\alpha}^4 + 7u_{\alpha}^2 - 16) + \sqrt{\frac{\sigma}{2}} \left(\frac{\sigma^2}{1920} (u_{\alpha}^5 + 20u_{\alpha}^3 + 15u_{\alpha}) + \right.$$

$$\left. + \frac{\delta^4}{2880} (u_{\alpha}^5 + 44u_{\alpha}^3 + 183u_{\alpha}) + \frac{\delta^4}{155520\sigma^2} (9u_{\alpha}^5 - 284u_{\alpha}^3 - 1513u_{\alpha}) \right),$$

де $\sigma = \frac{1}{\nu_1} + \frac{1}{\nu_2}$; $\delta = \frac{1}{\nu_1} - \frac{1}{\nu_2}$;

u_{α} – квантиль нормального розподілу.

ДОДАТОК Б

Статистичні таблиці

Таблиця Б.1

Квантилі нормального розподілу

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	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
-0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
-0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
-0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
-0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
-0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
-0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
-0,7	0,2420	0,2389	0,2358	0,2327	0,2297	0,2266	0,2236	0,2206	0,2177	0,2148
-0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
-0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
-1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
-1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
-1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
-1,3	0,0968	0,0951	0,0934	0,0917	0,0901	0,0885	0,0869	0,0853	0,0837	0,0822
-1,4	0,0807	0,0792	0,0778	0,0763	0,0749	0,0735	0,0721	0,0707	0,0694	0,0681
-1,5	0,0668	0,0655	0,0642	0,0630	0,0617	0,0605	0,0593	0,0582	0,0570	0,0559
-1,6	0,0548	0,0537	0,0526	0,0515	0,0505	0,0494	0,0484	0,0474	0,0464	0,0455
-1,7	0,0445	0,0436	0,0427	0,0418	0,0409	0,0400	0,0392	0,0383	0,0375	0,0367
-1,8	0,0359	0,0351	0,0343	0,0336	0,0328	0,0321	0,0314	0,0307	0,0300	0,0293
-1,9	0,0287	0,0280	0,0274	0,0268	0,0261	0,0255	0,0250	0,0244	0,0238	0,0233
-2,0	0,0227	0,0222	0,0216	0,0211	0,0206	0,0201	0,0197	0,0192	0,0187	0,0183
-2,1	0,0178	0,0174	0,0170	0,0165	0,0161	0,0157	0,0153	0,0150	0,0146	0,0142
-2,2	0,0139	0,0135	0,0132	0,0128	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
-2,3	0,0107	0,0104	0,0101	0,0099	0,0096	0,0093	0,0091	0,0088	0,0086	0,0084
-2,4	0,0081	0,0079	0,0077	0,0075	0,0073	0,0071	0,0069	0,0067	0,0065	0,0063
-2,5	0,0062	0,0060	0,0058	0,0057	0,0055	0,0053	0,0052	0,0050	0,0049	0,0047
-2,6	0,0046	0,0045	0,0043	0,0042	0,0041	0,0040	0,0039	0,0037	0,0036	0,0035
-2,7	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0028	0,0027	0,0026
-2,8	0,0025	0,0024	0,0024	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019	0,0019
-2,9	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014	0,0013
-3,0	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010	0,0010
-3,1	0,0009	0,0009	0,0009	0,0008	0,0008	0,0008	0,0007	0,0007	0,0007	0,0007
-3,2	0,0006	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005	0,0005	0,0005	0,0005
-3,3	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003	0,0003	0,0003	0,0003
-3,4	0,0003	0,0003	0,0003	0,0003	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002
-3,5	0,0002	0,0002	0,0002	0,0002	0,0002	0,0001	0,0001	0,0001	0,0001	0,0001
-3,6	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001

<i>z</i>	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

Квантили 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

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

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

$v_2 \setminus v_1$	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 зазначених розподілів та оцінку їх точності;
- довічне 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 повекторного 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 множини вибірок, які характеризують одновимірні об'єкти спостережень:

– критерій Бартлетта та однофакторний дисперсійний аналіз для вибірок, розподілених за нормальним законом;

– H-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 регресії:
 - 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.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Большев, Л.Н. Таблицы математической статистики [Текст] / Л.Н. Большев, Н.В. Смирнов. – М.: Наука, 1965. – 464 с.
2. Браунли, К.А. Статистическая теория и методология в науке и технике [Текст] / К.А. Браунли. – М.: Наука, 1977. – 407 с.
3. Ван-дер-Варден, Б.П. Математическая статистика [Текст] / Б.П. Ван-дер-Варден. – М.: Иностран. лит., 1960. – 434 с.
4. Деврой, Л. Непараметрическое оценивание плотности. L_1 -подход [Текст] / Л. Деврой, Л. Дьерфи. – М.: Мир, 1988. – 407 с.
5. Кендалл, М. Теория распределений [Текст] / М. Кендалл, А. Стюарт. – М.: Наука, 1966. – 588 с.
6. Коваленко, И.Н. Теория вероятностей [Текст] / И.Н. Коваленко, Б.В. Гнеденко. – К.: Выща шк., 1990. – 328 с.
7. Коваленко, И.Н. Теория вероятностей и математическая статистика [Текст] / И.Н. Коваленко, А.А. Филиппова. – 2-е изд. – М.: Высш. шк., 1982. – 256 с.
8. Основи теорії ймовірностей та математичної статистики [Текст] / В.П. Бабак, А.Я. Білецький, О.П. Приставка, П.О. Приставка. – К.: КВІЦ, 2003. – 432 с.
9. Сигел, Э. Практическая бизнес-статистика [Текст]: пер. с англ. / Э. Сигел. – 4-е изд. – М.: Издат. дом «Вильямс», 2002. – 1056 с.
10. Статистична обробка даних [Текст] / В.П. Бабак, А.Я. Білецький, О.П. Приставка, П.О. Приставка. – К.: МІВВІЦ, 2001. – 388 с.
11. Уилкс, С. Математическая статистика [Текст] / С. Уилкс. – М.: Наука, 1967. – 632 с.