РОЗДІЛ 11

СТАТИСТИЧНЕ ОЦІНЮВАННЯ ПАРАМЕТРІВ РОЗПОДІЛУ

11.1. Статистичні оцінки

Оцінка параметра (*статистична оцінка*) – це випадкова величина, розрахована за вибіркою, яка дає підстави для прийняття обґрунтованих рішень щодо невідомого параметра генеральної сукупності.

Головне завдання математичної статистики в тому, щоб на підставі результатів випадкової вибірки якомога точніше оцінити значення параметрів генерального розподілу. Окреме спостереження здебільшого належить деякій гіпотетично необмеженій генеральній сукупності, що організована відповідно до певного закону, який називають *розподілом сукупності*. Кожне спостереження дає додаткову інформацію про параметри та вигляд розподілу. Зазвичай за вибіркою розраховують наближені значення (*оцінки*) таких параметрів розподілу, як середнє, дисперсія, асиметрія й ексцес.

Часто припускають, що вибірковий розподіл нормальний або близький до нормального, що дуже важливо з огляду на можливість реалізації деяких стандартних процедур, які стосуються, зокрема, оцінювання параметрів та перевірки різних гіпотез. Однак таке припущення можливе лише за умови, що розглядаються достатньо великі вибірки ($n \geq 30$). Якщо $n \leq 15$, то для застосування тестів, викладених у цій книжці, нема підстав, і вони мають бути модифіковані. Ці питання вивчає теорія малих вибірок.

Задачі оцінювання можна розбити на дві великі групи: *параметричні* і *непараметричні*. В теорії ймовірностей розглядаються моделі, які містять параметри: біномний, нормальний, показниковий розподіли, розподіл Пуассона тощо. Так, наприклад, параметрами нормального розподілу є математичне сподівання і дисперсія, а параметром показникового розподілу можуть бути або інтенсивність, або середнє значення. Якщо відомий вигляд розподілу й область зміни параметрів, то маємо параметричну задачу. Якщо ж невідомий навіть закон розподілу, не кажучи вже про можливі значення параметрів, то задача непараметрична.

Задачі параметричного оцінювання поділяють на два типи: відшукання *точкових* оцінок і відшукання *інтервальних* оцінок. У першому випадку слід знайти, яке саме значення, розраховане за вибіркою, треба взяти за наближене значення істинної характеристики генеральної сукупності, а в другому – вказати інтервал, який із заданою ймовірністю містить це істинне значення.

Позаяк склад вибірки випадковий, то висновки, зроблені за вибіркою, можуть виявитися хибними. Тому будь-якому висновкові треба поставити у відповідність імовірність, яка описує ступінь достовірності (*надійність*) цього висновку. У цьому й полягає суть побудови інтервальних оцінок.

Наразі мова про параметричне оцінювання. У багатьох випадках наперед відомо, до якого класу належить функція розподілу $F(x,\theta)$, котра залежить від одного або кількох параметрів $\theta = (\theta_1, \theta_2, ..., \theta_k)$. Тоді визначення невідомої функції зводиться до оцінювання невідомих параметрів за вибіркою.

Випадковий відбір n елементів розглядається як n повторних випробувань. При цьому конкретну вибірку x_1, x_2, …, x_n будемо тлумачити як деяку реалізацію послідовності X_1, X_2, …, X_n випадкових величин.

Якщо умови в процесі спостереження не змінюються і результат окремого спостереження не залежить від результатів попередніх і майбутніх випробувань, то ймовірнісні властивості i-го спостереження гіпотетичної вибірки визначаються законом розподілу генеральної сукупності. Позаяк спостереження взаємно незалежні, то послідовність випадкових величин X_1, X_2, …, X_n також незалежна. Так буде, зокрема, у разі повторного відбору або відбору з гіпотетичної генеральної сукупності нескінченного обсягу.

Зауважмо, що точність і надійність статистичних висновків визначається не стосовно якоїсь конкретної вибірки (реалізації x_1, x_2, …, x_n n випадкових величин X_1, X_2, …, X_n), а стосовно множини всіх можливих вибірок.

11.2. Властивості статистичних оцінок

Оцінка параметра θ – це така функція вибіркових значень

$$\hat{\theta} = f(X_1, X_2, .., X_n),$$

яка в певному статистичному сенсі досить близька до істинного значення цього параметра. Основними статистичними властивостями, які описують якість оцінки числового параметра, є *незміщеність* (*незсуненість*), *ефективність* і *спроможність* (*обґрунтованість*, *значущість*, *консистентність*). Дві перші властивості можуть справджуватися для скінченної вибірки, остання – лише за умови необмеженого збільшення обсягу вибірки.

Розгляньмо кожну з цих властивостей докладніше.

11.2.1. Незміщеність оцінок. Нехай θ – деякий параметр, а вибіркова статистика $\hat{\theta} = f(X_1, X_2, \ldots, X_n)$ – оцінка цього параметра. Статистика $\hat{\theta}$ як випадкова величина має свій вибірковий розподіл, тому можна розглядати середнє (математичне сподівання), дисперсію та середнє квадратичне відхилення (*стандартне відхилення, стандартна похибка*) цього розподілу.

Оцінку $\hat{\theta}$ параметра θ називають *незміщеною* (*незсуненою*), якщо за будь-якого скінченного обсягу вибірки її математичне сподівання існує й дорівнює істинному значенню параметра:

$$M(\hat{\theta}) = \theta, \text{ або } M(\hat{\theta}) - \theta = 0.$$

Якщо ця рівність не справджується, то оцінка параметра $\hat{\theta}$ може або завищувати значення параметра (тобто $M(\hat{\theta}) > \theta$), або занижувати його (тобто $M(\hat{\theta}) < \theta$). І в першому, і в другому випадках це призводить до систематичних помилок (помилок одного знаку) в оцінюванні параметра θ.

Оцінку $\hat{\theta}$, яка не є незміщеною, називають *зміщеною* (*зсуненою*). У цьому разі $M(\hat{\theta}) - \theta = M(\hat{\theta} - \theta) = \delta \neq 0$, $\delta = \text{const}$.

Величину δ називають *зміщенням* (*зсувом*).

11.2.2. Ефективність оцінок. Нехай $\hat{\theta}_1$ і $\hat{\theta}_2$ дві різні незміщені оцінки параметра θ: $M(\hat{\theta}_1) = \theta$, $M(\hat{\theta}_2) = \theta$. Якщо існують математичні сподівання оцінок, то існують і дисперсії $D(\hat{\theta}_1)$ та $D(\hat{\theta}_2)$. Природно з-поміж цих оцінок взяти ту, значення якої менше розсіяні навколо параметра θ. Так, якщо $D(\hat{\theta}_1) < D(\hat{\theta}_2)$, то за оцінку параметра θ треба взяти $\hat{\theta}_1$.

Незміщену оцінку, яка має найменшу дисперсію з-поміж усіх можливих незміщених оцінок параметра, що отримані за вибірками однакового обсягу, називають *ефективною оцінкою.*

11.2.3. Спроможність (значущість, обґрунтованість) оцінок. Досягти бажаної точності оцінки за малого обсягу вибірки неможливо. Але бажано, щоб зі збільшенням обсягу вибірки точність оцінювання зростала. Граничної точності буде досягнуто, якщо числові значення оцінок, знайдених за різними вибірками (однакового обсягу) будуть однаковими. Це означало би, що оцінка незміщена, а її вибіркова дисперсія дорівнює нулеві. У реальному житті таких ситуацій не буває, хоча зазвичай за необмеженого збільшення обсягу вибірки оцінка параметра відповідно до закону великих чисел прямує до значення самого параметра (збігаються за ймовірністю). Таку оцінку називають *спроможною* (*обґрунтованою, консистентною*).

Точніше, оцінку $\hat{\theta}$ параметра θ називають *спроможною*, якщо вона задовольняє закон великих чисел, тобто справджується рівність:

$$\lim_{n\to\infty} P\left(\left|\hat{\theta}-\theta\right|<\varepsilon\right)=1,$$

де $\varepsilon>0$ – будь-яке, навіть скільки завгодно мале, число.

Спроможність – це асимптотична властивість $(n\to\infty)$. Вона означає: що більший обсяг вибірки, то більша ймовірність того, що похибка оцінки не перевищує скільки завгодно малого додатного числа ε. Спроможна оцінка параметра прямує за ймовірністю до істинного значення параметра.

Ефективна оцінка – спроможна.

11.3. Оцінювання математичного сподівання й дисперсії

Математичне сподівання й дисперсія – найважливіші числові характеристики випадкової величини. У цьому параграфі з'ясовується, які саме вибіркові характеристики є найкращими оцінками генеральної середньої $\overline{X}_{Г}=a$ та генеральної дисперсії σ^2. За основу беруть ймовірнісну модель вибірки як системи n випадкових величин X_i, розподіл яких збігається з генеральним розподілом і для яких $M(X_j)=a$ і $D(X_j)=\sigma^2$.

Розгляньмо спочатку безповторну вибірку. Для оцінювання генеральної середньої a природно взяти вибіркову середню $\hat{\theta}=\overline{X}$.

Теорема 1. Вибіркова середня арифметична $\overline{X}$, отримана на підставі n спостережень над випадковою величиною X, яка має математичне сподівання a, є незміщеною оцінкою математичного сподівання.

Доведення. Нехай X_1, X_2, ..., X_n – n незалежних спостережень над випадковою величиною X. За умовою, математичне сподівання випадкової величини X дорівнює a, тобто

$$M\left(X_1\right)=a,\ M\left(X_2\right)=a, ..., M\left(X_n\right)=a. \tag{1}$$

З огляду на означення середнього арифметичного і властивості математичного сподівання дістанемо:

$$M(\overline{X})=M\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n}\sum_{i=1}^{n}M(X_i)=\frac{1}{n}\cdot n\cdot a=a.$$

Отже, $M(\overline{X})=a$. Теорему доведено. ▲

Теорема 2. Дисперсія вибіркової середньої арифметичної $\overline{X}$, знайденої на підставі n спостережень над випадковою величиною X з дисперсією σ^2, визначається формулою

$$D(\overline{X}) = \sigma^2/n. \tag{2}$$

Доведення. З огляду на те що випадкової величини X_1, X_2, ..., X_n незалежні і

$$D(X_1) = D(X_2) = \ldots = D(X_n) = \sigma^2, \tag{3}$$

отримуємо:

$$D(\overline{X}) = D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} D\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i),$$

тобто

$$D(\overline{X}) = \frac{1}{n^2}\left(\sum_{i=1}^{n}\sigma^2\right) = \frac{1}{n^2}\cdot n\cdot\sigma^2 = \frac{\sigma^2}{n},$$

що і треба було довести. ▲

Теорема 3. Вибіркова середня $\overline{X}$, знайдена на підставі n спостережень над випадковою величиною X з математичним сподіванням a і дисперсію σ^2, є спроможною оцінкою математичного сподівання, тобто

$$\lim_{n\to\infty} P\left(\left|\overline{X} - a\right| < \varepsilon\right) = 1. \tag{4}$$ ▲

Якщо генеральна сукупність має нормальний розподіл, то вибіркова середня $\overline{X}$ буде ще й ефективною оцінкою з дисперсією σ^2/n.

Отже, вибіркова середня $\overline{X}$ задовольняє вимоги незміщеності і спроможності, і може слугувати оцінкою генеральної середньої: $\overline{X} \approx a$. Про ступінь точності цієї наближеної рівності докладно йтиметься згодом, коли розглядатимемо інтервальне оцінювання параметрів.

Розгляньмо оцінку генеральної дисперсії. Для цього природно скористатися статистикою (вибірковою дисперсією)

$$\hat{D} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2. \tag{5}$$

Теорема 4. Якщо вибірка складається з n незалежних спостережень над випадковою величиною X з математичним сподіванням a та дисперсією σ^2, то вибіркова дисперсія (5) є зміщеною оцінкою генеральної дисперсії.

Доведення. Вибіркову дисперсію можна записати у вигляді:

$$\hat{D} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\left(X_i - a\right) - \left(\overline{X} - a\right)\right)^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(X_i - a\right)^2 - \frac{2}{n}\left(\overline{X} - a\right)\sum_{i=1}^{n}\left(X_i - a\right) + n \cdot \frac{1}{n}\left(\overline{X} - a\right)^2 .$$

Другий доданок в останньому рядку можна спростити. Справді, з огляду на те що $\overline{X} = \sum X_i / n$, тобто $n\overline{X} = \sum X_i$, дістанемо:

$$\frac{2}{n}\left(\overline{X} - a\right)\sum\left(X_i - a\right) = \frac{2}{n}\left(\overline{X} - a\right)\left(\sum X_i - na\right) =$$

$$= \frac{2}{n}\left(\overline{X} - a\right)\left(n\overline{X} - na\right) = 2\left(\overline{X} - a\right)^2 .$$

Отже, вираз для вибіркової дисперсії набирає вигляду:

$$\hat{D} = \frac{1}{n}\sum\left(X_i - a\right)^2 - 2\left(\overline{X} - a\right)^2 + \left(\overline{X} - a\right)^2,$$

або

$$\hat{D} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - a\right)^2 - \left(\overline{X} - a\right)^2 . \qquad (6)$$

Відшукаймо математичне сподівання від вибіркової дисперсії (6):

$$M(\hat{D}) = M\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - a\right)^2\right) - M(\overline{X} - a)^2.$$

Оскільки

$$M\left(\frac{1}{n}\sum(X_i - a)^2\right) = \sigma^2 \quad \text{і} \quad M\left(\overline{X} - a\right)^2 = D(\overline{X}) = \frac{\sigma^2}{n},$$

то

$$M(\hat{D}) = \sigma^2 - \frac{\sigma^2}{n}, \text{ тобто } M(\hat{D}) = \frac{n-1}{n}\sigma^2 . \qquad (7)$$

Отже, статистика (вибіркова дисперсія) $\hat{\theta} = \hat{D}$ є зміщеною оцінкою генеральної дисперсії. Як видно з формули (7), зміщення

$$\delta = M(\hat{D}) - \sigma^2 = -\frac{\sigma^2}{n} . \blacktriangle$$

Незміщена (*виправлена*) оцінка генеральної дисперсії

$$s^2 = \frac{1}{n-1}\sum_{n=1}^{n}\left(X_i - \overline{X}\right)^2 . \qquad (8)$$

Справді, з рівності (8) випливає, що

$$s^2 = \frac{n}{n-1}\hat{D},$$

звідки, з огляду на співвідношення (8), отримуємо:

$$M(s^2) = \frac{n}{n-1}M(\hat{D}) = \frac{n}{n-1}\cdot\frac{n-1}{n}\cdot\sigma^2 = \sigma^2 .$$

Множник $n/(n-1)$ називають *поправкою Беселя.*

Відмінність між зміщеною $\hat{D}$ і незміщеною s^2 оцінками генеральної дисперсії істотна за малих обсягів вибірки. Якщо ж $n > 50$, то множник Беселя мало відрізняється від одиниці і між оцінками $\hat{D}$ й s^2 різниці практично немає. Можна довести, що обидві ці оцінки є спроможними оцінками дисперсії σ^2. Оцінки $\hat{D}$ і s^2 не є ефективними.

Незміщена, спроможна й ефективна оцінка генеральної дисперсії

$$\tilde{s}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - a\right)^2 , \qquad (9)$$

для розрахунку якої потрібно знати математичне сподівання (генеральну середню) a випадкової величини X. Якщо значення генеральної середньої невідоме (а саме так зазвичай і буває), то на практиці генеральну дисперсію замінюють її найкращою доступною оцінкою s^2, визначену формулою (8), яка є незміщеною і спроможною, але не ефективною.

Ми розглянули побудову точкових оцінок генеральної середньої і дисперсії для повторної вибірки. Можна довести, що для безповторної вибірки точкова оцінка генеральної середньої така сама, як для повторної:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i .$$

Задачі до розділу 11

1. За вибіркою знайти незміщену оцінку генеральної середньої, вибіркову дисперсію, виправлену вибіркову дисперсію та коефіцієнт варіації:

а)

x_i	3	6	9
n_i	8	15	7

б)

x_i	13	15	17	20
n_i	9	16	18	7

2. Для вивчення рівня стабільності роботи локальної мережі компанії проведено обстеження випадково вибраних офісів. Упродовж десяти днів фіксувалася кількість збоїв у роботі мережі й отримали такий ряд: 2, 4, 3, 0, 3, 0, 3, 2, 0, 1.

Знайти незміщену оцінку середньої, вибіркову дисперсію, незміщену вибіркову дисперсію, вибіркові коефіцієнти варіації, асиметрії й ексцесу.

3. Дано вибірку з генеральної сукупності з неперервно розподіленою ознакою:

а)

z_{i-1}	z_i	n_i
10	13	9
13	16	13
16	19	19
19	22	33
22	25	37
25	28	30
28	31	19
31	34	7

б)

z_{i-1}	z_i	n_i
8	11	7
11	14	11
14	17	16
17	20	27
20	23	31
23	26	25
26	29	16
29	32	9

Знайти вибіркову середню, вибіркову дисперсію і виправлену дисперсію, вибіркове середнє квадратичне відхилення і виправлене середнє квадратичне відхилення. Визначити вибіркові коефіцієнти асиметрії й ексцесу і на підставі здобутих результатів порівняти розподіл із нормальним.

4. Статистичні дослідження зростання продуктивності праці підприємств регіону в поточному році у відсотках до відповідного періоду попереднього року описуються таким інтервальним розподілом:

$(z_{i-1}; z_i]$	(80; 90]	(90; 100]	(100; 110]	(110; 120]	(120; 130]
n_i	2	14	60	20	4

Визначити оцінки математичного сподівання, дисперсії і середнього квадратичного відхилення випадкової величини X – росту продуктивності праці підприємства у відсотках – до відповідного періоду попереднього року.

5. У таблиці подано кількість туристичних путівок, реалізованих 30 туристичними агенціями за один день у пік туристичного сезону:

9	3	4	12	11	9	11	10	15	10
8	10	3	13	14	13	9	4	10	9
5	6	11	8	3	2	14	9	11	10

Знайти незміщені оцінки математичного сподівання й дисперсії кількості путівок, проданих за один день.

РОЗДІЛ 13

ІНТЕРВАЛЬНІ ОЦІНКИ ПАРАМЕТРІВ РОЗПОДІЛУ

13.1. Довірчий інтервал. Довірча ймовірність

Уже мовилося про оцінювання математичного сподівання й дисперсії. Це приклади точкового оцінювання, які дають змогу знайти точку, поблизу якої знаходиться невідомий (оцінюваний) параметр.

Інтервальні методи оцінювання дають змогу знайти інтервал, який з певною ймовірністю містить невідоме значення параметра.

Нехай знайдено точкову оцінку $\widehat{\theta}$ параметра θ. Побудова інтервальної оцінки параметра виходять з умови $P(|\widehat{\theta}-\theta|<\delta)=p$. Величину δ називають *точністю оцінки*, а p – її *надійністю*. Тоді інтервальна оцінка для параметра θ має вигляд $(\widehat{\theta}-\delta;\, \widehat{\theta}+\delta)$. Довірчий інтервал – це проміжок, який із заданою ймовірність містить оцінюваний параметр. Цей інтервал випадковий, бо його межі залежать від випадкової величини $\widehat{\theta}$. Надійність p – це ймовірність того, що випадковий інтервал покриває дійсне значення невипадкового параметра θ.

Ймовірність p називають *довірчою ймовірністю*, або *надійністю*.

Довірча ймовірність – це “практична достовірність” того, що довірчий інтервал містить оцінку. Тоді $\alpha=1-p$ – це ймовірність того, що цей інтервал не містить оцінку. Цю імовірність називають рівнем значущості. *Рівень значущості* – це імовірність, за якої конкретна подія в конкретних умовах вважається “практично неможливою”. Рівень значущості обирає дослідник; зазвичай беруть $\alpha=0{,}05$, $\alpha=0{,}01$.

Збільшення довірчої ймовірності веде до зниження точності оцінки параметра, тобто до збільшення довірчого інтервалу. Отже, що точніша оцінка, то менший рівень її надійності, і навпаки, для збільшення надійності потрібно збільшувати довірчий інтервал, тобто знижувати точність оцінки. Цю неусувну конкуренцію точності й надійності дослідник розв’язує кожен раз сам, залежно від конкретної ситуації.

13.2. Побудова довірчого інтервалу для математичного сподівання за відомої дисперсії

Нехай випадкова величина X розподілена нормально, а її середнє квадратичне відхилення відоме і дорівнює σ. Точніше, використовуючи вибіркову середню $\bar{x}$, треба побудувати довірчий інтервал, який із заданою ймовірністю p міститиме невідомий параметр a.

Вибіркова середня арифметична

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1}$$

як лінійна функція від нормально розподілених випадкових величин також має нормальний розподіл із параметрами

$$M(\bar{X}) = a, \quad D(\bar{X}) = \frac{\sigma^2}{n}. \tag{2}$$

Потрібно з'ясувати, як далеко може бути випадкова (вибіркова) величина $\bar{X}$ від невідомого числа a, оцінкою якого вона є, тобто наскільки великим може бути відхилення $\left|\bar{X} - a\right|$?

Ймовірність того, що нормально розподілена величина X із математичним сподіванням a і дисперсією σ^2 відхиляється від свого математичного сподівання не більше ніж на δ, визначають за формулою

$$P\left(\left|X - a\right| < \delta\right) = 2\Phi\left(\frac{\delta}{\sigma}\right), \tag{3}$$

де $\Phi(t)$ – функція розподілу стандартної нормальної випадкової величини (функція Лапласа). Для нормально розподіленої випадкової величини $\bar{X}$ з параметрами a та σ^2/n рівність (3) набирає вигляду:

$$P\left(\left|\bar{X} - a\right| < \delta\right) = 2\Phi\left(\frac{\delta\sqrt{n}}{\sigma}\right).$$

Якщо ввести позначення $z = \dfrac{\delta\sqrt{n}}{\sigma}$, або $\delta = \dfrac{z\sigma}{\sqrt{n}}$, то дістанемо:

$$P\left(\left|\bar{X} - a\right| < \frac{z\sigma}{\sqrt{n}}\right) = 2\Phi(z).$$

З огляду на те що нерівність, яка стоїть під символом імовірності, еквівалентна подвійній нерівності

$$-\frac{z\sigma}{\sqrt{n}} < a - \overline{X} < \frac{z\sigma}{\sqrt{n}}, \text{ або } \overline{X} - \frac{z\sigma}{\sqrt{n}} < a < \overline{X} + \frac{z\sigma}{\sqrt{n}},$$

дістанемо остаточно:

$$P\left(\overline{X} - \frac{z\sigma}{\sqrt{n}} < a < \overline{X} + \frac{z\sigma}{\sqrt{n}}\right) = 2\Phi(z).$$

Тепер, задаючи певний рівень надійності $p = 1 - \alpha = \Phi(z)$, за таблицею значень функції $\Phi(z)$ (додаток 2) можна знайти відповідне значення довірчого коефіцієнта $z_{p/2}$.

Отже, довірчий інтервал, який з імовірністю p містить математичне сподівання випадкової величини $N(a, \sigma^2)$, має вигляд

$$\bar{x} - z_{p/2} \cdot \frac{\sigma}{\sqrt{n}} < a < \bar{x} + z_{p/2} \cdot \frac{\sigma}{\sqrt{n}}, \qquad (4)$$

де $\bar{x}$ – середня, знайдена за вибіркою обсягу n. Як бачимо, для заданої ймовірності p, тобто за фіксованого значення $z_{p/2}$, зі збільшенням n довжина інтервалу зменшується, тобто зростає точність оцінки. Збільшення надійності p оцінки веде до збільшення довірчого коефіцієнта $z_{p/2}$, а отже, до збільшення довірчого інтервалу (за фіксованого n), і отже, зменшує точність оцінки.

Приклад. Внаслідок вимірювання довжини 50 випадково відібраних деталей отримано $\bar{x} = 10$ см. Відшукаймо довірчий інтервал, який з імовірністю $p = 0{,}95$ містить середній розмір усієї партії деталей, якщо є підстави вважати, що генеральна дисперсія $\sigma^2 = 0{,}09$, а розміри деталей усієї партії розподілені нормально.

Розв'язання. З умови $p = 2\Phi(z) = 0{,}95$ за таблицею значень функції Лапласа знаходимо довірчий коефіцієнт $z_{p/2} = t_{0{,}475} = 1{,}96$. Тому відхилення (помилка)

$$\delta = \frac{z_{p/2}\sigma}{\sqrt{n}} = \frac{1{,}96 \cdot \sqrt{0{,}09}}{\sqrt{50}} \approx 0{,}083.$$

Отже,

$$10 - 0{,}083 \le a \le 10 + 0{,}083,$$

тобто проміжок $[9{,}917; 10{,}083]$ з імовірністю $p = 0{,}95$ містить середню довжину деталей партії. ►

13.3. Побудова довірчого інтервалу для математичного сподівання за невідомої дисперсії

Нехай кількісна ознака X генеральної сукупності розподілена за нормальним законом, причому середнє квадратичне відхилення σ розподілу невідоме. Треба побудувати довірчий інтервал для невідомого математичного сподівання a .

Як уже мовилося, найкращою точковою оцінкою математичного сподівання є вибіркова середня $\overline{X}$, а найкращою оцінкою дисперсії σ^2 є виправлена вибіркова дисперсія

$$s^2 = \frac{1}{n-1}\sum (X_i - \overline{X})^2 \ . \tag{1}$$

Довірчий інтервал будуємо на підставі цих двох оцінок.

Для розв'язання задачі розгляньмо статистику

$$t = \frac{\overline{X} - a}{s} \cdot \sqrt{n} \ , \tag{2}$$

яка має розподіл Стьюдента з $k = n-1$ ступенями вільності. Вибравши рівень надійності $p = 1-\alpha$ і знаючи обсяг вибірки n, за таблицею розподілу Стьюдента (додаток 4) можна відшукати таке $t_{k,\alpha}$, що

$$P\left(\frac{\left|\overline{X} - a\right|}{s}\sqrt{n} < t_{k,\alpha} \right) = 1-\alpha \ . \tag{3}$$

Після перетворення нерівності під знаком імовірності отримаємо:

$$P\left(\left|\overline{X} - a\right| < \frac{t_{k,\alpha}\, s}{\sqrt{n}} \right) = 1-\alpha \ , \tag{4}$$

або

$$P\left(\overline{X} - t_{k,\alpha}\frac{s}{\sqrt{n}} < a < \overline{X} + t_{k,\alpha}\frac{s}{\sqrt{n}} \right) = 1-\alpha \ .$$

Отже, з імовірністю $p = 1-\alpha$ інтервал

$$\overline{x} - t_{k,\alpha}\frac{s}{\sqrt{n}} < a < \overline{x} + t_{k,\alpha}\frac{s}{\sqrt{n}} \tag{5}$$

є довірчим інтервалом для оцінки математичного сподівання a .

Співвідношення (5) за формою аналогічні співвідношенню (4) з попереднього параграфа. Однак між ними є істотна відмінність: у співвідношеннях (3)–(5) довірчий коефіцієнт $t_{k,\alpha}$ залежить не лише від рівня надійності p (як у попередньому параграфі), а й від обсягу вибірки n (точніше, від кількості ступенів вільності k). Ця відмінність стає значною, якщо вибірка мала, оскільки тоді не діє закон великих чисел і не можна використовувати теорему Ляпунова.

Приклад 1. Для визначення середньомісячного прибутку a підприємств регіону, побудовано вибірку (x_i – прибуток (тис. грн.), n_i – частота):

x_i	42,2	43,3	45,1	48,4	50,1	53,5
n_i	1	1	2	3	2	1

Вважаючи, що випадкова величина X – рівень прибутків у підприємствах регіону – розподілена нормально, побудуймо довірчий, який з імовірністю $p = 0{,}95$ містить невідомий параметр a .

Розв'язання. Відшукаймо числові характеристики вибірки – середню $\overline{X}$ та незміщену дисперсію s^2 :

$$\overline{x} = \frac{42{,}2 + 43{,}3 + 45{,}1 \cdot 2 + 48{,}4 \cdot 3 + 50{,}1 \cdot 2 + 53{,}5}{10} = 47{,}46 \text{ (тис. грн.)}$$

$$s^2 = \frac{(42{,}2 - 47{,}46)^2 + \ldots + (53{,}5 - 47{,}46)^2}{9} \approx 12{,}13 ,$$

звідки $s \approx 3{,}48$ тис. грн.

За таблицями розподілу Стьюдента для надійності $p = 0{,}95$ ($\alpha = 0{,}05$) і $k = n - 1 = 9$ ступенів вільності знаходимо $t_{k,\,p} = 2{,}26$. Тоді

$$\delta = t_{k,\alpha} \frac{s}{\sqrt{n}} = 2{,}26 \frac{3{,}48}{\sqrt{10}} \approx 2{,}49 . \qquad \text{(а)}$$

Отже, інтервал, який з імовірністю $p = 0{,}95$ містить середній прибуток підприємств регіону, має вигляд:

$$(\overline{x} - \delta;\ \overline{x} + \delta) = (44{,}97;\ 49{,}95) . \blacktriangleright \qquad \text{(в)}$$

Якщо знехтувати залежністю коефіцієнта t від n і скористатися для його визначення додаток 2, то знайдемо $z_{p/2} = 1{,}96$, а відхилення

$$\tilde{\delta} = z_{p/2} \frac{s}{\sqrt{n}} = 1{,}96 \cdot \frac{3{,}48}{\sqrt{10}} \approx 2{,}16 .$$

Тепер дістанемо коротший довірчий інтервал:

$$(\bar{x}-\tilde{\delta};\ \bar{x}+\tilde{\delta})=(45{,}30;\ 49{,}62)\,. \qquad \text{(c)}$$

Порівнюючи формули (с) й (в), бачимо, що застосування нормального закону розподілу невиправдано звужує довірчий інтервал. Це звуження пов'язано з тим, що ми вважаємо, що нам відома дисперсія. Насправді ми не маємо жодної інформації про дисперсію, окрім тої, котру дає вибірка. Але за великого обсягу вибірки ($n>30$) відмінність між розподілом Стьюдента і нормальним розподілом незначна, і тоді заміна довірчого коефіцієнта з таблиці розподілу Стьюдента (додаток 4) $t_{k,\alpha}$ на довірчий коефіцієнт $z=z_{p/2}=z_{(1-\alpha)/2}$, що відповідає значенню $\Phi\left(\frac{p}{2}\right)=\Phi\left(\frac{1-\alpha}{2}\right)$ з таблиці значень функції Лапласа (додаток 2), не призводить до значної похибки.

Приклад 2. Щоб з'ясувати середній час (у годинах) справної роботи пристрою, з генеральної сукупності для обстеження відібрано 15 пристроїв. Результати обстежень подано в таблиці (у першому рядку – умовний номер пристрою. у другому – час його роботи).

i	1	2	3	4	5	6	7	8
x_i	1160	1120	1250	1170	1180	1280	1300	1320

i	9	10	11	12	13	14	15
x_i	1170	1250	1400	1350	1200	1360	1240

Є підстави вважати, що в генеральній сукупності випадкова величина X – час роботи пристроїв – розподілена нормально. Відшукаймо ймовірність того, що середній час роботи пристроїв у всій партії відрізняється від отриманого за вибіркою не більше ніж на 50 год.

Розв'язання. Вибіркова середня

$$\bar{x}=\frac{1}{15}\sum_{i=1}^{15}x_i=\frac{1}{15}(1160+1120+\dots+1240)=1250 \text{ (год.)}.$$

Оскільки генеральна дисперсія невідома, то слід скористатися її найкращою доступною оцінкою – незміщеною вибірковою дисперсією

$$s^2=\frac{1}{n-1}\sum\left(x_i-\bar{x}\right)^2\,.$$

Отримуємо:

$$s^2 = \frac{(1160-1250)^2 + \ldots + (1240-1250)^2}{14} = \frac{98200}{14} \approx 7014,$$

звідки оцінка середнього квадратичного відхилення $s \approx 83{,}75$ (год.).

За умовою задачі, максимальне відхилення δ істинного середнього часу a роботи пристрою від середнього часу роботи $\bar{x}$, знайденого за вибіркою, дорівнює 50 год. Отже, як зрозуміло з формул (4) і (5),

$$\delta = \frac{t_{k,\alpha} \cdot s}{\sqrt{n}} = 50, \text{ звідки } t_{k,\alpha} = \frac{50\sqrt{n}}{s} = \frac{50\sqrt{15}}{83{,}75} \approx 2{,}3.$$

За таблицею розподілу Стьюдента для $t_{k,\alpha} = 2{,}3$ і $k = n-1 = 14$ ступенів вільності знаходимо:

$$P\left(\left|\bar{X} - a\right| \le 50\right) \approx 0{,}96.$$

За точнішими таблицями можна знайти, що $p \approx 0{,}963$. ►

Розв'язування задач на відшукання довжини довірчого інтервалу за вибіркою з нормальної сукупності у разі відомої та невідомої дисперсій, як бачимо, формально однакове і відрізняється лише тим, що в першому випадку для відшукання довірчого коефіцієнта використовують таблицю значень функції Лапласа, а в другому – таблицю розподілу Стьюдента.

13.4. Побудова довірчого інтервалу для дисперсії

Тут розглянемо побудову довірчих інтервалів для генеральної дисперсії σ^2, використовуючи її незміщені оцінки — вибіркові дисперсії $\tilde{s}^2$ та s^2.

13.4.1. Довірчий інтервал для дисперсії за відомого математичного сподівання. Нехай випадкова величина X має нормальний розподіл із відомим математичним сподіванням a і невідомою дисперсією σ^2. Найкращою оцінкою дисперсії за відомого математичного сподівання є незміщена, ефективна і спроможна, оцінка

$$\tilde{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - a)^2. \quad (1)$$

Розгляньмо статистику

$$\chi^2 = \frac{n\tilde{s}^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - a}{\sigma}\right)^2. \quad (2)$$

Як відомо (п. 6.8), випадкова величина χ^2 має розподіл Пірсона з $k = n$ ступенями вільності. Укладено таблиці (додаток 3), де вказано критичні (граничні) значення $\chi^2_{\text{кр}} = \chi^2_{k,\alpha}$, які випадкова величина χ^2, визначена формулою (2), перевищить з імовірністю α:

$$P(\chi^2 > \chi^2_{\alpha,k}) = \alpha .$$

Для вибраної довірчої ймовірності $p = 1-\alpha$ можна записати:

$$P\left(\chi_1^2 < \frac{n\tilde{s}^2}{\sigma^2} < \chi_2^2\right) = 1-\alpha . \qquad (3)$$

Тепер за таблицею χ^2-розподілу треба вибрати такі два значення χ_1^2 і χ_2^2, щоб площа під графіком диференціальної функції розподілу випадкової величини χ^2 між χ_1^2 і χ_2^2 дорівнювала $1-\alpha$ (рис. 13.1).

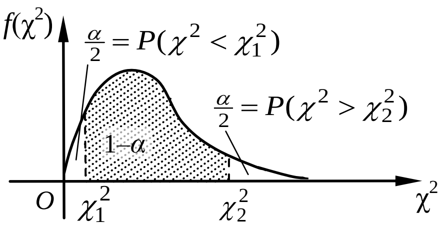


Рис. 13.1

З рис. 13.1 зрозуміло, що є безліч можливостей вибору χ_1^2 і χ_2^2 таких, що $S = 1-\alpha$. Логічно χ_1^2 і χ_2^2 вибрати так, щоб незаштриховані площі під кривою розподілу були однаковими й дорівнювали $\alpha/2$. Тоді

$$P(\chi^2 < \chi_1^2) = P(\chi^2 > \chi_2^2) = \alpha/2 . \qquad (4)$$

Позаяк у таблиці χ^2-розподілу вказано $P(\chi^2 > \chi^2_{k,\alpha})$, то для відшукання $P(\chi^2 < \chi_1^2)$ треба скористатися тотожністю

$$P(\chi^2 < \chi_1^2) = 1 - P(\chi^2 > \chi_1^2) . \qquad (5)$$

З геометричних міркувань зрозуміло (рис. 13.1), що

$$P\left(\chi_1^2 < \frac{n\tilde{s}^2}{\sigma^2} < \chi_2^2\right) = P(\chi^2 > \chi_1^2) - P(\chi^2 > \chi_2^2) .$$

Справді, ліва частина чисельно дорівнює площі заштрихованої фігури. Перший і другий доданки, що стоять праворуч, чисельно дорівнюють площам під кривою $f(\chi^2)$ правіше від χ_1^2 та χ_2^2 відповідно:

$$P(\chi^2 > \chi_1^2) = 1-\alpha/2 , \quad P(\chi^2 > \chi_2^2) = \alpha/2 .$$

Для того щоб оцінити σ^2, перетворімо подвійну нерівність:

$$\chi_1^2 < \frac{n\tilde{s}}{\sigma^2} < \chi_2^2, \quad \frac{\chi_1^2}{n\tilde{s}^2} < \frac{1}{\sigma^2} < \frac{\chi_2^2}{n\tilde{s}^2},$$

звідки отримуємо довірчий інтервал, який з імовірністю $p = 1 - \alpha$ містить невідому дисперсію:

$$\frac{n\tilde{s}^2}{\chi^2_{k,\alpha/2}} < \sigma^2 < \frac{n\tilde{s}^2}{\chi^2_{k,1-\alpha/2}},$$

де $\tilde{s}^2$ визначають за формулою (1), $k = n$ – кількість ступенів вільності.

13.4.2. Довірчий інтервал для дисперсії за невідомого математичного сподівання. Здебільшого генеральна середня μ невідома. Тоді, як ми знаємо, найкращою оцінкою дисперсії нормального розподілу є незсунена дисперсія (п. 11.3): $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$.

Міркуючи так, як у п. 14.4.1, дістанемо довірчий інтервал для дисперсії за невідомого математичного сподівання:

$$\frac{(n-1)s^2}{\chi^2_{k,\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{k,1-\alpha/2}}, \tag{9}$$

де $k = n - 1$ – кількість ступенів вільності.

Приклад. За вибіркою обсягом $n = 20$ отримано незміщену оцінку дисперсії $s^2 = 10$. Побудуймо довірчий інтервал, який з імовірністю $p = 0{,}96$ містить середнє квадратичне відхилення генеральної сукупності, якщо є підстави вважати, що значення ознаки в генеральній сукупності розподілені нормально.

Розв'язання. Тут $\alpha = 1 - p = 0{,}04$, $\alpha/2 = 0{,}02$, $k = n - 1 = 19$. За додатком 3 знаходимо: $\chi^2_{19;0{,}02} = 33{,}7$, $\chi^2_{19;0{,}98} = 8{,}6$.

Підставляючи числові дані в формулу (9), дістанемо:

$$\frac{19 \cdot 10}{33{,}7} < \sigma^2 < \frac{19 \cdot 10}{8{,}6}, \text{ тобто } 5{,}6 < \sigma^2 < 22{,}1,$$

звідки отримуємо довірчий інтервал для середнього квадратичного відхилення генеральної сукупності:

$$2{,}4 < \sigma < 4{,}7. \blacktriangleright$$

Задачі до розділу 13

1. Відомо, що випадкова величина X має нормальний розподіл. Побудувати довірчий інтервал, який з імовірністю p містить математичне сподівання, якщо:

1) $p = 0{,}95$, $\bar{x} = 15$, $\sigma^2 = 16$, $n = 25$;

2) $p = 0{,}95$, $\bar{x} = 10{,}5$, $\sigma = 3$, $n = 36$;

3) $p = 0{,}9$, $\bar{x} = 15$, $\sigma^2 = 16$, $n = 25$;

4) $p = 0{,}99$, $\bar{x} = 7$, $\sigma = 1$, $n = 49$.

2. Знайти довірчий інтервал для математичного сподівання, якщо

1) $n = 25$, $p = 0{,}95$, $\bar{x} = 15$, $s = 4$;

2) $n = 9$, $p = 0{,}95$, $\bar{x} = 6$, $s^2 = 9$.

3. З нормальної генеральної сукупності отримано таку вибірку:

а)

x_i	–3	–1	1	2
n_i	2	7	6	1

б)

x_i	4	6	7	9
n_i	4	7	8	6

Побудувати довірчий інтервал, який накриває математичне сподівання з імовірністю 1) $p = 0{,}95$; 2) $p = 0{,}9$; 3) $p = 0{,}99$. Побудувати довірчий інтервал для дисперсії, якщо рівень надійності $p = 0{,}96$.

4. Для випадкової величини, що має розподіл χ^2, вказати відхилення $\chi^2_{k;\alpha}$, ймовірність перевищення якого дорівнює $0{,}05$, якщо $k = 7$.

6. Відомо, що випадкова величина X має нормальний розподіл. З генеральної сукупності зроблено вибірку обсягом $n = 20$ й отримано виправлену вибіркову дисперсію $s^2 = 0{,}5$. Побудувати довірчий інтервал, який з імовірністю $0{,}96$ містить дисперсію σ^2 величини X.

7. Висота африканської жирафи – нормально розподілена випадкова величина з невідомою генеральною середньою і дисперсією $\sigma^2 = 0{,}3$. Обстеживши $n = 25$ випадкових особин, біолог отримав висоти x_1, x_2, ..., x_{25} і знайшов середню вибіркову (оцінку генеральної середньої) $\bar{x} = 5{,}3$ м. Побудувати $95\,\%$ довірчий інтервал для генеральної середньої. Скільки особин необхідно виміряти, щоб довірчий інтервал не перевищував $0{,}2$ м?

8. За вибіркою обсягу n із нормально розподіленої генеральної сукупності, математичне сподівання якої відоме, знайдено незміщену, спроможну й ефективну оцінку $\tilde{s}^2$ генеральної дисперсії. Побудувати довірчий інтервал, який зі ймовірністю p містить генеральне середнє квадратичне відхилення, якщо:

1) $p = 0{,}9$, $\tilde{s}^2 = 4$, $n = 15$;

2) $p = 0{,}96$, $\tilde{s}^2 = 9$, $n = 16$;

3) $p = 0{,}96$, $\tilde{s}^2 = 14$, $n = 10$;

4) $p = 0{,}98$, $\tilde{s}^2 = 1$, $n = 18$.

9. Розв'язати попередню задачу за умови, що генеральна середня невідома, і за вибіркою отримано виправлену вибіркову дисперсію (тобто в умові задачі 8 $\tilde{s}^2$ треба замінити на s^2).

10. З 200 працівників установи навмання вибрали 20. Середня зарплата, знайдена за вибіркою, становить 5000 гр. од., а виправлена дисперсія – 160 000 гр. од^2. Побудувати 95 %-й довірчий інтервал для середньої зарплати працівників установи, якщо є підстави вважати, що зарплата розподілена за нормальним законом.

11. З нормально розподіленої генеральної сукупності, середнє квадратичне відхилення якої дорівнює 1,2, зроблено вибірку обсягом 36. Вибіркова середня дорівнює 1,5. Побудувати довірчий інтервал для математичного сподівання, якщо рівень надійності $p = 0{,}96$.

12. Відомо, що генеральна сукупність має нормальний розподіл, а генеральна середня дорівнює 3. Статистичний розподіл вибірки із цієї сукупності має вигляд:

x_i	2	2,5	3,5	4
n_i	1	6	6	2

Знайти довірчий інтервал для генеральної дисперсії, якщо $p = 0{,}9$.

РОЗДІЛ 14

СТАТИСТИЧНА ПЕРЕВІРКА ГІПОТЕЗ

14. 1. Статистичні гіпотези

Закономірності, характерні для генеральної сукупності, можна виявити лише через статистичну перевірку спостережних даних, приймаючи або відхиляючи певні припущення – *гіпотези*.

Статистичні гіпотези – це припущення стосовно виду розподілу або окремих параметрів розподілу випадкової величини в генеральній сукупності.

Процедуру перевірки гіпотез за вибірковими статистичними даними називають *статистичною перевіркою статистичних гіпотез*. Суть перевірки гіпотези зводиться до того, щоб з'ясувати, узгоджуються чи не узгоджуються спостережні дані з гіпотезою. На підставі аналізу вибіркових даних гіпотезу можна прийняти або відхилити (обґрунтувати або спростувати).

Гіпотезу, яку необхідно перевірити, називають *нульовою* (*основною*) і позначають H_0, а другу гіпотезу називають *альтернативною* (*конкуруючою*) і позначають H . Альтернативна гіпотеза частково або повністю заперечує нульову.

Інформацію про випадкову величину, яка міститься в гіпотезі, називають *теоретичною* або *гіпотетичною*, а інформацію про випадкову величину, яку отримано за вибіркою, називають *статистичною* або *емпіричною*.

Перевіряючи гіпотези, можна допуститися помилки першого або другого роду. *Помилка першого роду* (*ризик I*) полягає в тому, що відхиляється нульова гіпотеза, хоча насправді вона правильна. *Помилка другого роду* (*ризик II*) полягає у відхиленні альтернативної гіпотези, хоча насправді вона правильна.

Уникнути помилок першого і другого роду неможливо. До того ж, як побачимо невдовзі, ці помилки є конкуруючими: зменшення ймовірності допуститися однієї з них веде до збільшення ймовірності допуститися другої. Тому, залежно від ситуації, доводиться вибирати компромісне рішення.

Зазвичай єдиним способом зниження ризику обох помилок є збільшення обсягу вибірки.

Розрізняють прості і складні гіпотези. Проста гіпотеза характеризує параметр однозначно, наприклад, $H_0 : x = a$. У складних гіпотезах вказують область імовірних значень параметра, наприклад, $H_0 : x > a$.

Слід мати на увазі те, що прийняття (точніше, невідхилення) тої чи іншої гіпотези не означає, що зроблене припущення є найкращим і єдино можливим. Насправді воно лише не суперечить наявним вибірковим даним, однак ці дані можуть не суперечити й іншим гіпотезам. Отож статистично підтверджену гіпотезу треба розглядати не як абсолютно правильний факт, а лише як твердження, що не суперечить одержаним дослідним даним і є досить правдоподібним.

14.2. Критерій перевірки гіпотези. Критична область

Висновок про сумісність або несумісність вибіркових даних із гіпотезою, яку перевіряють, роблять на підставі прийнятого статистичного критерію – однозначного правила (алгоритму), що встановлює умови, за яких гіпотеза має бути прийнята або відхилена. Критерій визначає, які вибіркові дані можна вважати такими, що не суперечать гіпотезі, а за яких даних вона має бути відхилена. Зазвичай критерій перевірки гіпотези реалізується за допомогою деякої статистики θ, яку будують на підставі вибіркових даних.

Гіпотезу відхиляють або не відхиляють залежно від того, наскільки вибіркові дані різняться від очікуваних (гіпотетичних). Отже, множину всіх можливих значень критерію θ можна поділити на дві підмножини, які не перетинаються. Перша підмножина w складається з тих значень критерію θ, за яких гіпотеза H_0 відхиляється. Вона називається *критичною областю*. Друга підмножина $\overline{w}$ називається *областю прийняття гіпотези* і складається з тих значень критерію θ, за яких гіпотеза H_0 не відхиляється.

Точки, які відокремлюють критичну область і область прийняття гіпотези, називають *критичними точками*. З огляду на випадковість вибірки ці точки не можуть бути вказані абсолютно точно, оскільки завжди існує можливість отримання невдалої вибірки, яка буде нести хибну інформацію про генеральну сукупність. Тому, не заперечуючи можливості допуститися помилки, слід уміти знаходити ймовірність тієї помилки. Якщо ця ймовірність незначна, то є підстави стверджувати, що критерій забезпечує малий ризик помилкових висновків.

Розгляньмо достатньо малу величину α, яку називають *рівнем значущості* критерію і яка дорівнює ймовірності помилки першого роду (ймовірності відхилення правильної основної гіпотези). Ця ймовірність має бути досить малою, щоб забезпечити високу *надійність критерію* $p = 1 - \alpha$, тобто високу ймовірність прийняття основної гіпотези, якщо вона правильна. Рівень значущості дорівнює ймовірності того, що випадкова статистика θ потрапить у критичну область за умови, що справедлива H_0.

Зазвичай рівень значущості α беруть рівним $0{,}05$ (5-відсотковий рівень значущості) або $0{,}01$ (1-відсотковий рівень значущості).

Що менше α, то менша ймовірність зробити помилку першого роду. Однак зі зменшенням α зменшується критична область, тому зменшується ймовірність потрапляння знайденого за вибіркою значення θ в цю область, навіть якщо H_0 неправильна (якщо $\alpha = 0$, то гіпотезу H_0 не можна відхилити за жодних значень θ). Отже, зменшення α веде до збільшення ймовірності β прийняти неправильну гіпотезу, тобто зробити помилку другого роду. Іншими словами, за великої кількості вибірок частка хибних висновків дорівнює α, якщо правильна H_0, і ця частка дорівнює β, якщо правильна H.

За фіксованого обсягу вибірки вибором відповідної критичної області w можна зробити скільки завгодно малим або α, або β.

Для перевірки гіпотези можна побудувати не один критерій. Ці критерії можуть давати різні результати щодо прийняття H_0 на підставі тієї самої вибірки. Тому для визначення кращого критерію вводять поняття потужності критерію. *Потужністю критерію* – це ймовірність $1 - \beta$ відхилення основної гіпотези, коли справедлива альтернативна.

14.3. Загальна схема перевіряння гіпотези

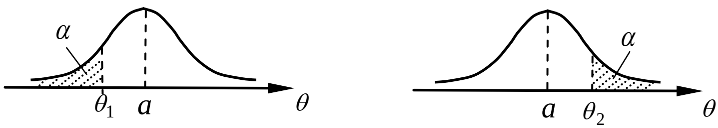
Задачу про перевірку гіпотези можна розбити на кілька етапів

Крок 1. Формулювання основної H_0 і альтернативної H гіпотез.

Крок 2. Вибір статистики для тестування гіпотези. Статистика θ, яку беруть за критерій перевірки гіпотези, має бути незміщеною, спроможною й ефективною; її розподіл за умови, що справедлива H_0, має бути відомим.

Крок 3. Визначення обсягу вибірки, вибір рівня значущості.

Крок 4. Визначення критичних точок. Нехай гіпотеза $H_0 : \tilde{a} = a$ перевіряється на противагу альтернативній $H : \tilde{a} < a$. Вибіркові результати θ будуть групуватися навколо центра розподілу $M(\theta) = a$. Але якщо вони виявляться значно меншими за a , то перевагу слід віддати гіпотезі H . При цьому критичну область логічно визначити нерівністю $\theta < \theta_1$, тобто критична область є лівобічною (рис. 14.1, а)). Вибравши рівень значущості α , дістанемо рівняння для відшукання критичної точки θ_1: $P(\theta < \theta_1) = \alpha$.



а) б)

Рис. 14.1

Аналогічно альтернативній гіпотезі $H : \tilde{a} > a$ відповідає правобічна критична область, яку визначають з умови (рис. 14.1, б)):

$$P(\theta > \theta_2) = \alpha . \tag{2}$$

Якщо нульова гіпотеза $H_0 : \tilde{a} = a$ перевіряється супроти альтернативної $H : \tilde{a} \neq a$, то дістанемо двобічну альтернативну область, критичні точки якої θ_1 та θ_2 визначають із рівняння

$$P(\theta < \theta_1) + P(\theta > \theta_2) = \alpha. \tag{3}$$

Останнє рівняння (одне рівняння з двома невідомими) має безліч розв'язків, тобто існує безліч пар чисел θ_1 й θ_2 , які задовольняють умову (3). Для побудови однозначного розв'язку слід задати додаткову умову. Зазвичай двобічну критичну область конструюють як симетричну (рис. 14.2), тобто критичні точки θ_1 і θ_2 визначають із рівнянь:

$$P(\theta < \theta_1) = \frac{\alpha}{2} \quad \text{та} \quad P(\theta > \theta_2) = \frac{\alpha}{2} .$$

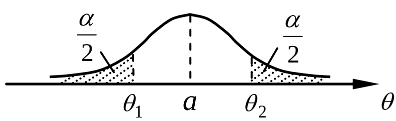


Рис. 14.2

Крок 5. Розрахунок емпіричного значення критерію θ^* за вибіркою запланованого обсягу n. Якщо значення θ^* потрапляє в допустиму область, то вважають, що гіпотеза H_0 не суперечить вибірковим даним (гіпотеза H_0 правдоподібна). Якщо θ^* потрапляє в критичну область, то це означає, що, ґрунтуючись на конкретних вибіркових даних, основну гіпотезу H_0 треба відхилити.

Для перевірки гіпотези кожного виду побудовано відповідні критерії. Найчастіше для побудови критеріїв використовуються z-статистика, t-статистика Стьюдента, F-статистика Фішера – Снедекора, χ^2-статистика Пірсона.

Далі розглянемо конкретні приклади перевірки статистичних гіпотез.

14.4. Перевіряння гіпотези про рівність математичних сподівань двох нормальних сукупностей

На практиці часто трапляються ситуації, коли середній результат в одній серії спостережень істотно різняться від середнього, отриманого на підставі іншої серії спостережень. Тоді постає запитання про те, чи можна вважати розбіжність у значеннях середніх випадковою, чи вона спричинена деякими неврахованими (або й невідомими) закономірностями. До такої задачі приводить, наприклад, вибірковий контроль якості однотипної продукції, що виготовлена на різному устаткуванні чи за різних технологічних режимів.

У загальному вигляді задача про порівняння двох центрів розподілів формулюється так. Нехай задано дві нормально розподілені сукупності X і Y. Внаслідок обстеження сукупностей X і Y отримано дві незалежні вибірки обсягами n_1 і n_2 відповідно, на підставі яких знайдено середні $\overline{X}$ і $\overline{Y}$. Потрібно перевірити нульову гіпотезу

$$H_0 : M(X) = M(Y) \qquad (1)$$

супроти альтернативної

$$H : \left| M(X) - M(Y) \right| > 0\,. \qquad (2)$$

Гіпотеза двобічна, бо умова (2) еквівалентна умові $M(X) \neq M(Y)$.

14.4.1. Перевіряння гіпотези про рівність генеральних середніх двох нормальних сукупностей за відомих дисперсій. Вважатимемо, що генеральні дисперсії σ_X^2 і σ_Y^2 відомі. Оскільки генеральні середні $M(X)$ і

$M(Y)$ невідомі, то для перевірки гіпотези скористаємося їх найкращими оцінками – вибірковими середніми $\overline{X}$ і $\overline{Y}$. Якщо X і Y мають нормальний розподіл, то, як ми знаємо, середні $\overline{X}$ і $\overline{Y}$ також нормально розподілені, до того ж

$$\overline{X} \sim N\Big(M(X);\ \sigma_X^2/n_1\Big), \quad \overline{Y} \sim N\Big(M(Y);\ \sigma_Y^2/n_2\Big).$$

Оскільки вибірки незалежні, то вибіркові середні $\overline{X}$ і $\overline{Y}$ є незалежними випадковими величинами. Різниця

$$Z = \overline{X} - \overline{Y}$$

також має нормальний розподіл. За властивістю дисперсії різниці

$$D(Z) = D(\overline{X} - \overline{Y}) = D(\overline{X}) + D(\overline{Y}) = \frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}.$$

Якщо гіпотеза H_0 справедлива, то для центрованої випадкової величини $Z - M(Z)$ матимемо:

$$Z - M(Z) = \overline{X} - \overline{Y} - M(\overline{X} - \overline{Y}) =$$
$$= \overline{X} - \overline{Y} - \Big(M(\overline{X}) - M(\overline{Y})\Big) = \overline{X} - \overline{Y}.$$

Отже, за умови справдження гіпотези гіпотеза H_0 статистика (нормована випадкова величина)

$$z = \frac{\left|\overline{X} - \overline{Y}\right|}{\sqrt{\dfrac{\sigma_X^2}{n_1} + \dfrac{\sigma_Y^2}{n_2}}} \qquad (3)$$

має нормальний розподіл із нульовим математичним сподіванням і дисперсією, що дорівнює одиниці: $z \sim N(0; 1)$.

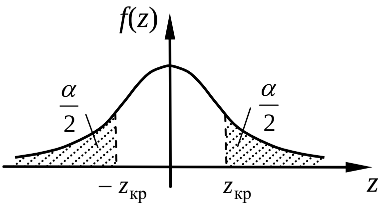


Рис. 14.3

Залежно від конкретних умов, вибираємо рівень значущості α і для надійності $p = 1 - \alpha$ за додатком 2 знаходимо критичне значення статистики: $z_{кр} = z_{p/2}$ у разі двобічної гіпотези (рис. 14.3), або $z_{кр} = z_p$ у разі однобічної гіпотези. Якщо

$$z < z_{кр},$$

де z визначено формулою (3), то вибіркові дані не дають підстави для від хилення гіпотези H_0: з імовірністю p можна вважати, що центри розподілу двох генеральних нормально розподілених сукупностей однакові.

Рівень значущості α – це ймовірність відхилення гіпотези H_0 за умови, що вона правильна. Зрозуміло: що менше α, то менша ймовірність помилки першого роду. Водночас зі зменшенням α збільшується область допустимих значень, отже, зростає ризик прийняти гіпотезу H_0 тоді, коли вона неправильна. У цьому разі дослідник контролює лише помилку першого роду, а про ступінь ризику, пов'язаного з прийняттям неправильної гіпотези, не можна зробити жодного висновку.

Приклад. З нормально розподілених генеральних сукупностей з однаковими дисперсіями $\sigma_X^2 = \sigma_Y^2 = 0{,}09$ зроблено вибірки обсягами $n_1 = 25$ і $n_2 = 50$. За вибірками отримано середні значення $\bar{x} = 9{,}07$ і $\bar{y} = 9{,}63$. Чи можна з надійністю $p = 0{,}95$ стверджувати, що розбіжності між середніми спричинені випадковими факторами?

Розв'язання. За формулою (3) вираховуємо емпіричне (дослідне) значення статистики z – нормованої різниці:

$$z = \frac{|\bar{x} - \bar{y}|}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}} = \frac{|\bar{x} - \bar{y}|}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{|9{,}07 - 9{,}63|}{0{,}3\sqrt{\frac{1}{25} + \frac{1}{50}}} \approx 2{,}41.$$

За додатком 2 знаходимо критичне значення статистики: $z_{\text{кр}} = z_{p/2} =$ $= z_{0{,}475} = 1{,}96$ (гіпотеза двобічна). Позаяк $z > z_{\text{кр}}$, то з імовірністю $0{,}95$ розбіжність між середніми є значущою (невипадковою). ►

14.4.2. Перевіряння гіпотези про рівність центрів розподілу двох нормально розподілених сукупностей за невідомої дисперсії. Потрібно перевірити гіпотезу (1) про рівність математичних сподівань двох сукупностей супроти альтернативної гіпотези (2), але, на відміну від попереднього випадку, дисперсії сукупностей невідомі. Для простоти вважатимемо, що ці дисперсії рівні: $\sigma_X^2 = \sigma_Y^2 = \sigma^2$.

За оцінки $M(X)$ і $M(Y)$ беруть вибіркові середні $\bar{X}$ і $\bar{Y}$, а за оцінки невідомих дисперсій – незміщені оцінки

$$s_X^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \bar{X})^2, \quad s_Y^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i - \bar{Y})^2, \tag{4}$$

де n_1 та n_2 – обсяги вибірок.

Оскільки дисперсії генеральних сукупностей рівні, то для оцінювання невідомої дисперсії σ^2 доцільно скористатися обома вибірками. Доведено, що найкраща оцінка дисперсії генеральної сукупності у цьому разі – це зважена середня арифметична дисперсій σ_X^2 та σ_Y^2, де вагами є кількості ступенів вільності:

$$s^2 = \frac{s_X^2(n_1 - 1) + s_Y^2(n_2 - 1)}{n_1 + n_2 - 2}, \tag{5}$$

або

$$s^2 = \frac{\sum_{i=1}^{n_1}(X_i - \overline{X})^2 + \sum_{i=1}^{n_2}(Y_i - \overline{Y})^2}{n_1 + n_2 - 2}. \tag{5a}$$

Якщо справедлива гіпотеза H_0, то випадкова величина $Z = \overline{X} - \overline{Y}$ має нормальний розподіл, до того ж ($\sigma_X^2 = \sigma_Y^2 = \sigma^2$)

$$M(Z) = 0, \quad D(Z) = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right).$$

Але дисперсія σ^2 генеральних сукупностей невідома, і постає питання про те, якою статистикою оцінити дисперсію $D(\overline{X} - \overline{Y})$. Зазвичай за таку оцінку беруть статистику

$$s_{\overline{X}-\overline{Y}}^2 = s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right), \tag{6}$$

де s^2 визначена формулою (5). Доведено, що $s_{\overline{X}-\overline{Y}}^2$ є незміщеною оцінкою дисперсії випадкової величини $\mathrm{Z} = \overline{X} - \overline{Y}$, тобто

$$M(s_{\overline{X}-\overline{Y}}^2) = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right). \tag{7}$$

Якщо випадкова величина $(\overline{X} - \overline{Y})$ має нормальний розподіл, то, як відомо, нормована випадкова величина

$$t = \frac{(\overline{X} - \overline{Y}) - M(\overline{X} - \overline{Y})}{s_{\overline{X}-\overline{Y}}}, \tag{8}$$

де $s_{\overline{X}-\overline{Y}}$ визначена формулою (6), має розподіл Стьюдента з $k = n_1 + n_2 - 2$ ступенями вільності.

Якщо справедлива гіпотеза $H_0 : M(X) = M(Y)$, то емпірична статистика (8) набирає вигляду:

$$t = \frac{\left|\overline{X} - \overline{Y}\right|}{\sqrt{\dfrac{n_1 + n_2}{n_1 n_2} \cdot \dfrac{(n_1 - 1)s_X^2 + (n_2 - 1)s_Y^2}{n_1 + n_2 - 2}}}, \qquad (8а)$$

де s_X^2 і s_Y^2 визначені формулами (4). Критичне значення статистики для вибраного рівня значущості α і $k = n_1 + n_2 - 2$ ступенів вільності знаходимо за таблицею розподілу Стьюдента (додаток 4).

Альтернативна гіпотеза має вигляд $H : \left|M(X) - M(Y)\right| > 0,$ тому критична область двобічна. Позаяк розподіл Стьюдента симетричний щодо нуля (п. 6.7), то критерій має найбільшу потужність тоді, коли праву й ліву критичні точки вибрано так, що ймовірність потрапляння критерію в кожен із двох інтервалів двосторонньої критичної області дорівнює $\alpha/2$. Тому

$$t_{\text{кр}} = t_{n_1 + n_2 - 2;\ \alpha}. \qquad (7)$$

Якщо альтернативна гіпотеза має вигляд

$$H : M(X) < M(Y) \text{ або } H : M(X) > M(Y), \qquad (8)$$

то, очевидно, критична область однобічна (рис. 14.1) і тоді критичне (табличне) значення критерію

$$t_{\text{кр}} = t_{n_1 + n_2 - 2;\ 2\alpha}. \qquad (9)$$

Приклад. Щоб з'ясувати, як запровадження нової технології впливає на продуктивність праці, відібрано дві групи працівників. Для першої групи обсягом $n_1 = 40$ осіб, яка працює за новою технологією, отримано такі дані: $\bar{x} = 84$ (одиниці), $s_X = 10{,}1$; для другої групи обсягом $n_2 = 54$ особи – $\bar{y} = 77{,}5$; $s_Y = 8{,}4$. Чи можна з рівнем значущості $\alpha = 0{,}05$ стверджувати, що нова технологія вплинула на середню продуктивність?

Розв'язання. Нульова гіпотеза $H_0 : M(X) = M(Y)$, альтернативна гіпотеза $H : \left|M(X) - M(Y)\right| > 0$.

За формулою (8а) знаходимо емпіричне значення статистики (критерію):

$$t = \frac{84 - 77{,}5}{\sqrt{\dfrac{(40 + 54)(39 \cdot 10{,}1^2 + 53 \cdot 8{,}4^2)}{40 \cdot 54(40 + 54 - 2)}}} \approx 3{,}364.$$

За таблицями розподілу Стьюдента знаходимо теоретичне значення критерію $t_{\text{кр}} = t_{92;\,0,05} \approx 1{,}99$. Оскільки $t > t_{\text{кр}}$, то розбіжності між вибірковими середніми не можна пояснити випадковим характером вибірок. Це означає, що, ґрунтуючись на вибіркових даних, нульову гіпотезу треба відхилити: спостереження дають підставу вважати, що нова технологія ефективна. ►

ЗАУВАГА. Вибірки обсягів 40 і 54 є немалими, і розподіл статистики (6) близький до нормального. Тому часто для великих вибірок ($n > 30$) при визначенні критичного значення статистики розподіл вважають нормальним і використовують таблицю значень функції Лапласа. Так, для $\alpha = 0{,}05$ ($p = (1-0{,}05)/2 = 0{,}475$) за цією таблицею знаходимо $z_{\text{кр}} = z_{0,475} = {}$ $= 1{,}96$, що відрізняється від $1{,}99$ на $1{,}5$ %. ☻

Зауважмо, що для значень $z \le z_{\text{кр}}$ чи $t \le t_{\text{кр}}$ не можна стверджувати, що гіпотеза H_0 справедлива; можна лише констатувати, що отримані вибіркові дані не дають підстав для відхилення гіпотези. За допомогою статистичної перевірки гіпотез можна лише відхилити початкову гіпотезу, але неможливо довести її справедливість.

14.5. Перевіряння гіпотези про рівність дисперсій двох нормальних генеральних сукупностей

Нехай

$$X_1,\, X_2,\, ...,\, X_{n_1} \quad \text{та} \quad Y_1,\, Y_2,\, ...,\, Y_{n_2}$$

– незалежні вибірки з першої та другої сукупностей обсягами n_1 та n_2 відповідно, а σ_X^2 та σ_Y^2 – генеральні дисперсії. Потрібно перевірити гіпотезу $H_0 : \sigma_X^2 = \sigma_Y^2$ супроти альтернативної гіпотези $H : \sigma_X^2 > \sigma_Y^2$. Інтуїтивно зрозуміло, що для цього потрібно в той чи інший спосіб оцінити генеральні дисперсії σ_X^2 та σ_Y^2 і порівняти здобуті оцінки.

14.5.1. Гіпотеза про рівність дисперсій за невідомих середніх. У цьому разі найкращі, точніше, незміщені і спроможні оцінки генеральних дисперсій, як ми знаємо, визначаються формулами:

$$s_X^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \bar{X})^2 \quad \text{та} \quad s_Y^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i - \bar{Y})^2\,, \qquad (1)$$

де $\overline{X}$ та $\overline{Y}$ – вибіркові середні. Позаяк дисперсії генеральних сукупностей невідомі, то перевірка гіпотези H_0 ґрунтується на зіставленні вибіркових дисперсії s_X^2 та s_Y^2. Для того щоб відхилити нульову гіпотезу про рівність дисперсій, потрібно довести значущість (невипадковість) розбіжності між s_X^2 та s_Y^2 за вибраного рівня значущості.

Для побудови критичної області, яка відповідає вибраному рівневі значущості, треба дослідити спільний закон розподілу статистик s_X^2 та s_Y^2. Таким спільним розподілом є F-розподіл (розподіл Фішера–Снедекора, додаток 6), який зазвичай і беруть за критерій у цьому разі.

Пригадаймо: випадкова величина $F = \dfrac{s_X^2}{\sigma_X^2} \Big/ \dfrac{s_Y^2}{\sigma_Y^2}$ має F-розподіл. Якщо справедлива нульова гіпотеза ($\dfrac{\sigma_X^2}{\sigma_Y^2} = 1$), то $F = \dfrac{s_X^2}{s_Y^2}$, і тоді F-розподіл можна використати для оцінювання відношення s_X^2 / s_Y^2. Якщо статистика

$$F = \frac{s_X^2}{s_Y^2}, \; s_X^2 > s_Y^2, \tag{2}$$

близька до 1, то, очевидно, нема підстав для відхилення нульової гіпотези. (Зауважмо: завжди можна ввести позначення так, що $s_X^2 \geq s_Y^2$, тобто $F \geq 1$; це потрібно для зменшення обсягу таблиць.)

Якщо відношення (2) істотно більше за 1, то це може бути підставою для відхилення гіпотези H_0. Питання в тому, за якого значення відношення вибіркових дисперсій це відхилення буде обґрунтованим.

F-розподіл залежить лише від кількостей ступенів вільності чисельника і знаменника $k_1 = n_1 - 1$ і $k_2 = n_2 - 1$, де n_1 і n_2 – обсяги вибірок, і не залежать від невідомих параметрів a та σ^2 (п. 6.8).

У додатку 6 для випадкової величини F вказано значення $F_{\alpha,\, k_1,\, k_2}$, що відповідають різним поєднанням кількостей ступенів вільності k_1 чисельника та k_2 знаменника, для яких справджується умова (рис. 14.4)

$$P(F > F_{\alpha,\, k_1,\, k_2}) = \alpha .$$

Перевіряння гіпотези H_0 зводиться до такого правила: якщо відношення вибіркових дисперсій більше від критичного, тобто

$$F > F_{\text{кр}} = F_{\alpha,\, k_1,\, k_2},$$

то гіпотезу H_0 відхиляють; якщо ж $F \le F_{\text{кр}}$, то гіпотезу H_0 не відхиляють.

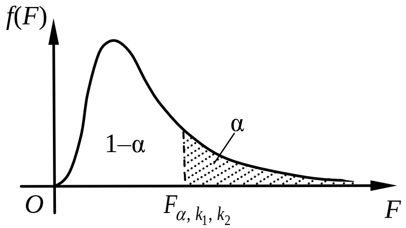


Рис. 14.4

Зрозуміло, якщо альтернативна гіпотеза має вигляд $H: \sigma_X^2 \ne \sigma_Y^2$, то використовують двобічний критерій; тоді

$$F_{\text{кр}} = F(\alpha/2,\, k_1,\, k_2).$$

Якщо ж альтернативна гіпотеза має вигляд $H: \sigma_X^2 > \sigma_Y^2$, то використовують однобічний критерій, тобто

$$F_{\text{кр}} = F(\alpha,\, k_1,\, k_2).$$

Приклад. Спостереження за якісним показником деякої продукції засвідчили, що якість залежить від тривалості обробки сировини. Після вдосконалення технології на другій виробничій лінії варіації часу попередньої обробки зменшилася і продукт став одноріднішим.

Вибіркові вимірювання варіації часу обробки сировини на першій і другій лініях дали $s_1^2 = 1{,}05$ хв^2 та $s_2^2 = 0{,}5$ хв^2 за однакового обсягу вибірок $n_1 = n_2 = 25$. Чи можна вважати істотною відмінністю між варіаціями тривалості процесу обробки сировини до і після покращення технології?

Розв'язання. Нульова гіпотеза $H_0: \sigma_1^2 = \sigma_2^2$, альтернативна гіпотеза $H: \sigma_1^2 > \sigma_2^2$ (тобто припущення про те, що коливання тривалості процесу обробки після вдосконалення технології зменшилися). Зрозуміло, що в цьому разі треба перевірити однобічну гіпотезу. Отже,

$$F = \frac{s_1^2}{s_2^2} = \frac{1{,}05}{0{,}5} = 2{,}1.$$

За таблицями розподілу Фішера – Снедекора для рівня значущості $\alpha = 0{,}05$ (додаток 6.2) і ступенів вільності чисельника та знаменника $k_1 = k_2 = 24$ знаходимо: $F_{\text{кр}} = F_{0{,}05;\,24;\,24} \approx 2$. Оскільки $F > F_{\text{кр}}$, (знайдене за вибіркою значення F потрапляє в критичну область), то гіпотезу

H_0 слід відхилити: на підставі вибіркових даних розбіжність між дисперсіями можна вважати значущою. Отже, впровадження нової технології призвело до підвищення однорідності продукції. ►

14.5.2. Перевіряння гіпотези про рівність дисперсії за відомих середніх. Цю гіпотезу перевіряють так само, як і попередню, але в цьому разі вибіркові оцінки дисперсій

$$\tilde{s}_X^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}\left(X_i - a_X\right)^2 \quad \text{та} \quad \tilde{s}_Y^2 = \frac{1}{n_2}\sum_{i=1}^{n_2}\left(Y_i - a_Y\right)^2$$

є незміщеними, спроможними і, на відміну від оцінок (1), ефективними. Якщо справедлива гіпотеза $H_0 : \sigma_X^2 = \sigma_Y^2 = \sigma^2$, то величини $\frac{n_1 \tilde{s}_X^2}{\sigma^2}$ та $\frac{n_2 \tilde{s}_Y^2}{\sigma^2}$ розподілені за законом χ^2 із $k_1 = n_1$ та $k_2 = n_2$ ступенями вільності відповідно. Отже, в цьому разі статистика $F = \frac{\tilde{s}_X^2}{\tilde{s}_Y^2}$ має розподіл Фішера з $k_1 = n_1$ та $k_2 = n_2$ ступенями вільності.

Треба пам'ятати, що критерії перевірки гіпотез про рівність середніх і дисперсій, є точними і найкращими. Тому в разі нормального закону розподілу генеральної сукупності ці критерії можна застосувати і для великих і для малих вибірок. Окрім того, для великих вибірок ці критерії (з певною мірою обачності) можна застосувати й тоді, коли генеральний розподіл не є нормальним, оскільки, як відомо, випадкові величини $\overline{X}$ та $\overline{Y}$ у такому разі розподілені асимптотично нормально.

14.6. Гіпотеза про закон розподілу

У багатьох практичних задачах точний закон розподілу невідомий. У таких випадках за допомогою оцінювання функції розподілу або щільності можна перевірити гіпотезу про те, наскільки добре вибіркові дані узгоджуються з теоретичними висновками про функцію (щільність) розподілу. Для перевірки гіпотез про закон розподілу використовують так звані *критерії згоди*, за допомогою яких можна з'ясувати, чи є розбіжність між дослідним і генеральним розподілами значущою, чи її можна пояснити випадковими впливами.

Найчастіше за критерієм згоди перевіряють гіпотезу про нормальний закон розподілу, позаяк припущення про нормальний розподіл генеральної сукупності покладено в основу багатьох критеріїв.

Нульова гіпотеза H_0 полягає в тому, що випадкова величина X із рівнем значущості α описується функцією розподілу $F(x)$.

Існують різні критерії згоди. Ми розглянемо один із них, а саме:

Критерій згоди χ^2**.** Цей критерій розподілу придатний для перевірки гіпотези про розподіл як дискретних, так і неперервних випадкових величин. Перевірка здійснюють за допомогою спеціально підібраної випадкової величини – *критерію згоди*. При цьому порівнюють емпіричні і теоретичні (знайдені за припущення, що випадкова величина X має розподіл $F(x)$, який залежить від параметра θ, значення якого невідоме) частоти і з'ясовують, значущою чи випадковою є розбіжність між тими частотами.

Розгляньмо докладніше критерій згоди хі-квадрат.

Нехай маємо n спостережень випадкової величини X. Розбиймо область зміни величини X на r інтервалів Δ_1, Δ_2, ..., Δ_r. Нехай n_i – кількість елементів вибірки, які потрапили в інтервал Δ_i:

$$\sum_{i=1}^{r} n_i = n, \quad i = 1,2,..., r.$$

За вибірковими даними знаходимо оцінки $\hat{\theta}$ невідомих параметрів θ теоретичного (гіпотетичного) розподілу $F(x)$. Тоді, вважаючи відомим теоретичний закон розподілу, завжди можна відшукати ймовірності p_i потрапляння випадкової величини X в інтервал Δ_i за формулою

$$p_i = p_i(\hat{\theta}) = P(X \in \Delta_i) = F(x_1) - F(x_{i-1}), \quad i = \overline{1, r}.$$

Тепер можна порівняти спостережувані частоти n_i потрапляння в кожен інтервал з очікуваними (теоретичними) частотами np_i.

Інтервали Δ_i	Δ_1	Δ_2	...	Δ_i	...	Δ_r
Емпіричні частоти n_i	n_1	n_2	...	n_i	...	n_r
Теоретичні частоти np_i	np_1	np_2	...	np_i	...	np_r

Якщо відмінність між емпіричними і теоретичними частотами незначна, то нема підстав для відхилення гіпотези H_0, у супротивному разі гіпотезу H_0 відхиляють. Випадкові розбіжності між частотами можуть бути спричинені мализною вибірки, невдалим способом групування даних тощо.

Невипадкові розбіжності породжені тим, що теоретичні частоти знайдено на підставі хибної гіпотези про закон розподілу.

За міру розбіжності між спостережними і гіпотетичними частотами беруть величину (критерій)

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i}, \tag{1}$$

яка є малою за доброї згоди.

Можна довести: якщо гіпотеза H_0 справедлива, то за достатньо великої вибірки статистика (1) має розподіл χ^2 із $k = r - l - 1$ ступенями вільності, де l – кількість параметрів розподілу, оцінених за вибіркою, а r – кількість інтервалів.

Якщо емпіричний і теоретичний розподіли цілком збігаються, то $\chi^2 = 0$, у супротивному разі $\chi^2 > 0$.

Нескладно зрозуміти алгоритм перевірки гіпотези. За таблицею розподілу χ^2 (додаток 3) для вибраного рівня значущості α і k ступенів вільності знаходимо критичне значення критерію $\chi^2_{\text{кр}} = \chi^2_{\alpha;k}$ і порівнюємо його з розрахованим за формулою (1).

Якщо

$$\chi^2 < \chi^2_{\text{кр}},$$

то гіпотезу H_0 не відхиляють; а якщо

$$\chi^2 > \chi^2_{\text{кр}},$$

то гіпотезу H_0 відхиляють.

Прийняття гіпотези H_0 не є доведенням її справедливості; це лише свідчення того, що вибіркові дані не суперечать гіпотезі.

Статистика (1) має розподіл χ^2 лише за умови, що $n \to \infty$. Тому необхідною умовою застосовності критерію Пірсона є дотримання таких умов: 1) дослідні мають бути незалежними; 2) обсяг вибірки має бути достатньо великим $(n > 50)$, до того ж у кожному інтервалі має бути щонайменше 5–10 спостережень. Якщо остання умова не справджується, то інтервали з малою кількістю спостережень треба попередньо об’єднати.

Розгляньмо приклади, що ілюструють застосування критерію згоди.

Приклад 1. Вимірювання діаметра 100 відібраних однотипних виробів дало такі відхилення діаметра від проектного.

Порядковий номер	Відхилення	Порядковий номер	Відхилення	Порядковий номер	Відхилення	Порядковий номер	Відхилення
1	0,1	26	0,3	51	1,3	76	–1,9
2	0,3	27	0,1	52	–2,8	77	4,2
3	–1,8	28	2,1	53	2,3	78	–0,8
4	1,3	29	–0,9	54	–1,9	79	–0,8
5	2,1	30	0,5	55	0,1	80	0,2
6	–0,9	31	2,1	56	–0,8	81	–3,0
7	1,4	32	1,3	57	3,1	82	1,1
8	$\boxed{-2{,}5}$	33	1,2	58	1,3	83	0,1
9	2,1	34	–0,8	59	0,1	84	0,2
10	3,1	35	0,3	60	2,1	85	2,3
11	2,1	36	1,2	61	3,1	86	–1,9
12	–1,8	37	0,1	62	1,1	87	1,2
13	1,1	38	–0,8	63	2,2	88	3,1
14	0,1	39	3,1	64	1,1	89	0,8
15	1,2	40	–1,9	65	$\boxed{5{,}0}$	90	0,2
16	0,7	41	2,2	66	–1,8	91	1,3
17	1,3	42	0,1	67	0,4	92	1,0
18	–0,9	43	–0,7	68	1,1	93	–0,8
19	1,4	44	1,2	69	0,2	94	–1,8
20	2,2	45	0,4	70	1,5	95	–0,8
21	1,1	46	1,2	71	–0,8	96	1,1
22	3,1	47	0,3	72	0,3	97	2,3
23	4,1	48	–1,9	73	0,1	98	–0,8
24	–0,9	49	–0,9	74	1,1	99	0,1
25	–1,8	50	1,3	75	2,1	100	3,0

Для рівня значущості $\alpha = 0{,}01$ перевірити гіпотезу H_0 про те, що відхилення від проєктного розміру підлягають нормальному законові розподілу.

Розв`язання. Випадкова величина X – відхилення діаметра деталі від проектного. Найменше значення $x_{\min} = -2{,}5$, а найбільше $x_{\max} = 5$. Область зміни випадкової величини X дорівнює $x_{\max} - x_{\min} =$ $= 5 - (-2{,}5) = 7{,}5$. Розбиймо цю область на 8 інтервалів, кожен з яких має довжину 1. Відповідний варіаційний ряд матиме вигляд:

$(z_{i-1};z_i]$	(–3;–2]	(–2; –1]	(–1;0]	(0;1]	(1;2]	(2;3]	(3;4]	(4;5]
n_i	3	10	15	24	25	13	7	3

Оскільки кількість спостережень у крайніх інтервалах менша 5, то їх можна об`єднати з сусідніми:

$(z_{i-1};z_i]$	(–3;–1]	(–1;0]	(0;1]	(1;2]	(2;3]	(3;5]
n_i	13	15	24	25	13	10

Для відшукання ймовірностей p_i потрібно знайти параметри розподілу – математичне сподівання й дисперсію. Позаяк ці параметри невідомі, то візьмімо їх найкращі оцінки – вибіркову середню $\overline{x}$ і вибіркову незміщену дисперсію s^2, взявши за варіанти середини інтервалів. Дістанемо:

$$\overline{x}=\frac{\frac{(-3-1)}{2}\cdot 13+\frac{(-1+0)}{2}\cdot 15+....+\frac{3+5}{2}\cdot 10}{100}\approx 0{,}9\,,$$

$$s^2=\frac{(-2-0{,}9)^2\cdot 13+(-0{,}5-0{,}9)^2\cdot 15+....+(0{,}4-0{,}9)^2\cdot 10}{99}\approx 2{,}55\,,\quad s\approx 1{,}6\,.$$

Тепер за формулою

$$p_i=P(z_{i-1}<X<z_i)=\Phi\left(\frac{z_i-\overline{x}}{s}\right)-\Phi\left(\frac{z_{i-1}-\overline{x}}{s}\right),$$

де Φ – функція Лапласа (додаток 2), знаходимо ймовірність потрапляння випадкової величини в кожен із шести інтервалів:

$$p_1=P(-3<X<-1)=\Phi\left(\frac{-1-0{,}9}{1{,}6}\right)-\Phi\left(\frac{-3-0{,}9}{1{,}6}\right)\approx$$

$$\approx\Phi(2{,}44)-\Phi(1{,}19)=0{,}4927-0{,}3830=0{,}1097\,;$$

$$p_2=P(-1<X<0)=\Phi\left(\frac{0-0{,}9}{1{,}6}\right)-\Phi\left(\frac{-1-0{,}9}{1{,}6}\right)=0{,}1707\,.$$

Аналогічно знаходимо:

$$p_3=\Phi(0{,}06)-\Phi(0{,}56)=0{,}0239+0{,}2123=0{,}2362\,,$$

$$p_4=\Phi(0{,}69)-\Phi(0{,}06)=0{,}2549-0{,}0239=0{,}2310\,,$$

$$p_5=\Phi(1{,}31)-\Phi(0{,}69)=0{,}4043-0{,}2549=0{,}1500\,,$$

$$p_6=\Phi(2{,}25)-\Phi(1{,}31)=0{,}4878-0{,}4049=0{,}0829\,.$$

Нескладно пересвідчитися: $\sum p_i=0{,}9805\approx 1$ (сума ймовірностей дорівнює 1 наближено внаслідок заокруглень під час обчислень).

На підставі здобутих результатів отримуємо:

$(z_{i-1};z_i]$	(–3;–1]	(–1;0]	(0;1]	(1;2]	(2;3]	(3;5]
n_i	13	15	24	25	13	10
np_i	10,97	17,07	23,62	23,10	15,00	8,29

Тепер за формулою (1) просто вирахувати статистику χ^2:

$$\chi^2 = \frac{(13-10{,}97)^2}{10{,}97} + \frac{(15-17{,}07)^2}{17{,}07} + ... + \frac{(10-8{,}29)^2}{8{,}29} =$$

$$= 0{,}38 + 0{,}25 + 0{,}01 + 0{,}16 + 0{,}27 + 0{,}35 = 1{,}42 .$$

Кількість інтервалів $r = 6$; кількість параметрів розподілу, знайдених за вибіркою, $l = 2$. Отже, кількість ступенів вільності $k = 6 - 2 - 1 = 3$. За таблицею розподілу Пірсона (додаток 3) знаходимо критичне значення статистики для рівня значущості $\alpha = 0{,}01$:

$$\chi^2_{\text{кр}} = \chi^2_{3;0{,}01} = 11{,}34 .$$

Позаяк $\chi^2 < \chi^2_{\text{кр}}$, то нема підстав для відхилення гіпотези, яка перевірялася: з імовірністю $0{,}99$ можна стверджувати, що відхилення діаметра деталей від проектного розподілено за нормальним законом. ▶

Приклад 2. Виробник реалізував 200 пристроїв. Кількість зауважень кожного зі споживачів щодо якості роботи пристрою упродовж гарантійного терміну подано в таблиці. На підставі критерію χ^2 Пірсона для рівня значущості $\alpha = 0{,}05$ перевірити гіпотезу про те, що випадкова величина – кількість зауважень споживачів – має розподіл Пуассона.

Кількість зауважень щодо якості роботи	Кількість споживачів, які зробили ці зауваження
0	110
1	66
2	20
3	3
4	1

Розв'язання. Треба перевірити гіпотезу

$$H_0 : P(X = m) = \frac{\lambda^m e^{-\lambda}}{m!} .$$

Визначмо середнє значення і незміщену вибіркову дисперсію:

$$\bar{x} = \frac{0 \cdot 110 + 1 \cdot 66 + 2 \cdot 20 + 3 \cdot 3 + 4 \cdot 1}{200} = \frac{119}{200} \approx 0{,}6 ;$$

$$s^2 = \frac{(0-0{,}6)^2 \cdot 110 + (1-0{,}6)^2 \cdot 66 + \ldots + (4-0{,}6)^2 \cdot 1}{199} \approx 0{,}6 .$$

Як ми знаємо, характерною рисою розподілу Пуассона є те, що параметри розподілу – математичне сподівання й дисперсія – мають однакові числові значення. Отже, необхідна умова для розподілу Пуассона $\bar{x} \approx s^2$ у цьому разі справджується.

Взявши за точкову оцінку параметра розподілу Пуассона вибіркову середню, тобто прийнявши $\lambda = 0{,}6$, нескладно знайти ймовірності p_i ($i = 0, 1, 2, 3, 4$), з якими випадкова величина набуває кожного зі своїх значень. За таблицею розподілу Пуассона для $\lambda = 0{,}6$ і $m = 0, 1, 2, 3, 4$ знаходимо:

$$p_0 = P(0; 0{,}6) = 0{,}5488, \quad p_1 = P(1; 0{,}6) = 0{,}3292,$$

$$p_2 = 0{,}0987, \quad p_3 = 0{,}0198, \quad p_4 = 0{,}0030.$$

Для зручності вихідні дані і результати розрахунків можна подати у вигляді таблиці:

i	0	1	2	3	4
x_i	0	1	2	3	4
n_i	110	66	20	3	1
np_i	109,7	65,8	19,7	3,96	0,6

Оскільки два останні стовпці містять мало спостережень, то їх можна об’єднати в один:

i	0	1	2	3
x_i	0	1	2	3 й 4
n_i	110	66	20	4
np_i	109,7	65,8	19,7	4,56

Значення статистики χ^2:

$$\chi^2 = \frac{(110 - 109{,}7)^2}{109{,}7} + \frac{(66 - 65{,}8)^2}{65{,}8} + \frac{(20 - 19{,}7)^2}{19{,}7} + \frac{(4 - 4{,}56)^2}{4{,}56} \approx 0{,}07.$$

За таблицею розподілу χ^2 для рівня значущості $\alpha = 0{,}05$ і $k = 4 - 1 - 1 = 2$ ступенів вільності знаходимо:

$$\chi^2_{\text{кр}} = \chi^2_{2;0{,}05} = 5{,}99.$$

Оскільки $\chi^2 < \chi^2_{\text{кр}}$, то дослідні дані не суперечать гіпотезі про те, що випадкова величина розподілена за законом Пуассона. ►

Задачі до розділу 14

1. З нормально розподілених генеральних сукупностей X і Y зроблено вибірки обсягами $n_1 = 40$ і $n_2 = 50$ відповідно. За вибірками розраховано середні значення $\bar{x} = 9{,}8$ і $\bar{y} = 9{,}4$. Відомо, що $\sigma_X = \sigma_Y = 0{,}3$. Чи можна вважати, що генеральні середні рівні, якщо $\alpha = 0{,}01$?

2. Підприємство розробило два способи виробництва продукції. Для того щоб з'ясувати, який зі способів більш матеріалозатратний, зібрано статистичні дані про витрати сировини на одиницю продукції для кожного способу виробництва, які подано в таблиці.

I спосіб	2,1	2,7	2,9	2,5	2,3	2,6
II спосіб	3,2	2,5	3,5	3,7	3,4	

Вважаючи дисперсії обох сукупностей однаковими, для рівня надійності $p = 0{,}95$ перевірити гіпотезу H_0 про те, що матеріалозатратність обох способів виробництва однакова.

3. Генеральні сукупності X і Y мають нормальний розподіл з однаковими дисперсіями $\sigma_X^2 = \sigma_Y^2$. За малими вибірками обсягами $n_1 = 5$ і $n_2 = 6$ отримано такі значення вибіркових середніх і виправлених дисперсій: $\overline{X} = 2{,}3$, $\overline{Y} = 2{,}48$, $s_X^2 = 0{,}25$, $s_Y^2 = 0{,}108$. Для рівня значущості $\alpha = 0{,}05$ перевірити гіпотезу $H_0 : M(X) = M(Y)$.

4. На двох верстатах виготовляються однотипні деталі. Для порівняння ступеня точності обробки деталей на двох верстатах обстежено вибірки обсягами $n_1 = 10$ деталей та $n_2 = 15$ деталей, за якими знайдено виправлені дисперсії: $s_1^2 = 9{,}6$ та $s_2^2 = 5{,}7$. Чи можна з рівнем надійності $p = 0{,}95$ стверджувати, що точність обробки деталей на двох верстатах різна?

5. За умовою задачі 2 перевірити для рівня значущості $\alpha = 0{,}05$ гіпотезу H_0 про те, що дисперсії величин витрат сировини на одиницю продукції однакові за обох способів виробництва.

6. Для порівняння продуктивностей двох технологій X і Y одержано такі вибіркові дані:

X	14,1	10,4	14,7	13,7	14,0
Y	14,0	14,5	13,0	12,7	14,8

Припустивши, що генеральні сукупності розподілені нормально,

1) перевірити гіпотезу про рівність дисперсій ($H:\sigma_X^2 \neq \sigma_Y^2$);

2) якщо дисперсії рівні, то перевірити гіпотезу про рівність продуктивностей $H: M(X) = M(Y)$.

7. Довготривалі обстеження засвідчили, що дисперсія часу обслуговування клієнта банку дорівнює 4 хв^2. Банк прийняв на роботу нового працівника. Спостереження за його роботою дало такі результати (t_i – час обслуговування клієнта, n_i – частота):

t_i	5,5	6	7,7	8,6	9
n_i	2	5	10	7	1

Чи можна з рівнем надійності $p = 0{,}95$ стверджувати, що новий співробітник працює ритмічно (в тому сенсі, що дисперсія витрат часу на обслуговування клієнта істотно не відрізняється від 4 хв^2)?

8. За критерієм Пірсона для рівня значущості $\alpha = 0{,}05$ перевірити, чи суттєві розбіжності між емпіричними частотами n_i і теоретичними частотами np_i, розрахованими за припущення, що ознака в генеральній сукупності має нормальний розподіл:

n_i	7	17	42	70	38	17	9
np_i	6	18	38	74	39	19	6

9. Для зросту 200 навмання вибраних чоловіків, що працюють на заводі, отримано такий інтервальний статистичний розподіл:

Зріст, см	Частота	Зріст, см	Частота
158 – 161	2	176 – 179	32
161 – 164	8	179 – 182	25
164 – 167	12	182 – 185	18
167 – 170	23	185 – 188	6
170 – 173	33	188 – 191	3
173 – 176	37	191 – 194	1

За критерієм хі-квадрат для рівня значущості $\alpha = 0{,}05$ з’ясувати, чи узгоджуються дані з гіпотезою про те, що зріст чоловіків є випадковою величиною з нормальним законом розподілу.

10. Для рівня значущості $\alpha = 0{,}01$ за критерієм χ^2 перевірити правдивість нульової гіпотези про нормальний закон розподілу ознаки генеральної сукупності, якщо відомі значення теоретичних та емпіричних частот.

Теоретичні частоти	5	9	16	26	13	6	2
Емпіричні частоти	4	12	13	30	15	8	3

11. Дослідження випадків смерті солдатів у пруській армії від удару копита коня засвідчили, що кількість таких випадків в окремих кавалерійських корпусах упродовж року мала розподіл, дуже подібний до розподілу Пуассона:

Кількість випадків у корпусі за рік	Кількість корпусів із даною кількістю випадків
0	119
1	65
2	22
3	3
4	1

Перевірити за критерієм χ^2 для рівня значущості $\alpha = 0{,}05$ припущення про те, що випадкова величина (кількість смертей) описується розподілом Пуасона.

РОЗДІЛ 15

ЕЛЕМЕНТИ ТОРІЇ КОРЕЛЯЦІЇ

15.1. Рівняння регресії. Кореляційне поле

Здебільшого зв'язок між випадковими величинами стохастичний (імовірнісний). *Стохастичний зв'язок* – це залежність умовного розподілу однієї величини від значень іншої (інших).

Центральною ідеєю опису стохастичної залежності є її опис у термінах строгих функціональних співвідношень. У разі двох залежних випадкових величин кожному значенню однієї з них ставлять у відповідність умовне математичне сподівання іншої, тобто

$$M(Y|X = x) = f(x).$$

Це рівняння називають *рівнянням регресії* Y *на* X .

На практиці розглядають спрощений випадок стохастичного зв'язку, коли кожному значенню однієї змінної ставлять у відповідність набір значень іншої, до того ж, кількість цих значень не є сталою, а самі значення не відображають певної закономірності. Такий зв'язок між випадковими змінними називають *статистичним зв'язком*. Тоді умовне математичне сподівання замінюють умовним середнім значенням, яке можна знайти за вибіркою: $M(Y|x) \approx \bar{y}(x)$.

Функцію, яка описує статистичний зв'язок між величинами, називають *вибірковим рівнянням регресії* Y *на* X (*кривою регресії*):

$$\bar{y}(x) = f(x).$$

На підставі цього рівняння за відомими значеннями однієї змінної можна знайти середні значення іншої.

Статистичні зв'язки поділяють на кореляційні та регресійні. Цей поділ ґрунтується на тому, що фактор X може випадковим або невипадковим.

Кореляційна залежність – це залежність між однією випадковою величиною й умовним середнім значенням іншої випадкової величини. Прикладом кореляційного зв'язку є статистична взаємозалежність між окремими частинами тіла людини, обумовлена їх взаємозв'язком і впливом певних первинних факторів, насамперед спадковістю.

На практиці часто розглядають не кореляційний, а *регресійний зв'язок*. *Регресія* – це залежність середнього значення однієї випадкової величини від однієї або більше невипадкових величин. Прикладом регресійного зв'язку є залежність урожайності деякої сільськогосподарської культури від кількості внесених добрив. Фактор X – кількість добрив – величина невипадкова, але врожайність є випадковою величиною, бо залежить, окрім фактора X, від випадкових чинників: якості ґрунту, погодних умов тощо.

Побудова рівняння регресії зводиться до встановлення форми залежності між ознаками (лінійна, квадратична, показникова тощо) і визначення параметрів рівняння за вибірковими даними. Для визначення параметрів найчастіше використовують метод найменших квадратів (МНК).

Після побудови рівняння регресії потрібно провести його статистичний аналіз, який складається з двох частин.

По-перше, потрібно оцінити похибку від заміни істинної регресії наближеною, перевірити значущість усіх складових рівняння регресії через порівняння їх із дисперсією спостережень. Цю частину дослідження називають *регресійним аналізом*.

По-друге, необхідно з'ясувати ступінь впливу фактора на результат, тобто оцінити частку регресії в загальному розсіянні Y. Ця частина дослідження називається *кореляційним аналізом* і полягає у визначенні різних показників щільності (тісноти) зв'язку. Кореляційний аналіз дає змогу встановити, значущою чи незначущою (випадковою чи невипадковою) є зміна спостережуваної випадкової величини внаслідок зміни фактора.

У разі двовимірної кореляції наближене уявлення про форму і щільність (тісноту) зв'язку між змінними X та Y дає *кореляційне поле*.

Кореляційним полем (або *діаграмою розсіяння*) називають діаграму, на якій зображено сукупність точок $(x_i; y_i)$ на площині, де x_i, y_i – це значення ознак в i-у спостереженні.

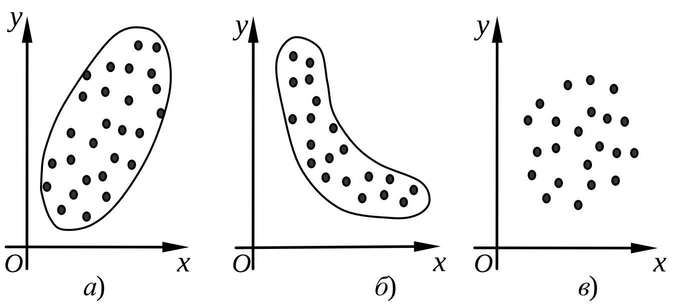


Рис. 15.1

Схематично приклади діаграм розсіяння зображено на рис. 15.1.

Якщо зі збільшенням змінної X змінна Y у середньому також збільшується, то кажуть, що між змінними існує *додатний зв'язок* (*додатна кореляція*) (рис. 15.1, *а*)), а якщо зі збільшенням X змінна Y у середньому зменшується, то маємо *від'ємний зв'язок* (*від'ємну кореляцію*) (рис. 15.1, *б*)).

Діаграма розсіяння, зображена на рис. 15.1, *в*) свідчить про відсутність зв'язку між змінними X та Y, тобто маємо *нульову кореляцію*.

На підставі першої діаграми розсіяння можна зробити висновок про наявність помірного лінійного зв'язку між ознаками, а друга діаграма засвідчує, що між величинами може бути сильний нелінійний зв'язок.

15.2. Кореляційна таблиця

За великої кількості спостережень те саме значення x може траплятися n_x разів, те саме значення $y - n_y$ разів, а та сама пара чисел $(x;\ y) - n_{xy}$ разів. У такому разі дані зручно згрупувати і подати у вигляді *кореляційної таблиці*. У першому рядку таблиці вказано спостережні значення X, а в першому стовпці – спостережні значення Y. На перетині рядків і стовпців вказано частоти n_{xy} спостереження відповідної пари значень (табл. 15.1).

В останньому рядку кореляційної таблиці вказано суми частот відповідного стовпця, а в останньому стовпці – суми частот відповідного рядка.

Так, у табл. 15.1 частота 22 означає, що пара чисел $(30;5)$ спостерігається у вибірці 22 рази. Число 24 в останньому рядку таблиці – це кількість спостережень, в яких $x = 30$, а число 2 в останньому стовпці – це кількість спостережень, в яких $y = 7$. Кореляційну таблицю називають також *вибірковим двовимірним розподілом*.

Табл. 15.1

y	x				
	10	30	50	70	n_y
1	7	–	5	16	28
3	–	3	19	3	25
5	1	22	2	2	25
7	–	1	1	–	2
n_x	8	24	27	21	$n = 80$

Якщо ознаки X і Y можуть набувати неперервних значень, то в кореляційній таблиці в першому рядку і першому стовпці вказують інтервали зміни ознак X та Y і значення середин цих інтервалів, а в клітинах на перехресті – відповідну кількість спостережень.

Так, табл. 15.2 відповідає діаграмі розсіяння, що зображена на рис. 15.2. Наприклад, частота 5 означає, що в інтервал зміни X від 4 до 6 і в інтервал зміни Y від 30 до 40 потрапляє 5 спостережень. У клітину

$x = 2 \div 4$, $y = 20 \div 30$ потрапляє 4 точки, а в клітину $x = 6 \div 8$, $y = 40 \div 50$ — 8 точок і под.

За діаграмою розсіяння, що зображена на рис. 15.2, можна висунути гіпотезу про те, що зв'язок між X і Y лінійний.

Кореляційна таблиця дає загальне уявлення про напрям і щільність зв'язку між ознаками X і Y. Якщо, наприклад, числові дані мають тенденцію до розташування вздовж діагоналі таблиці зверху вниз, то це вказує на додатний зв'язок між ознаками. Спостережні значення зосереджені в еліпсі, великою віссю якого є ця діагональ. Цей еліпс називають *кореляційним еліпсом*. Що більше стиснений еліпс, то щільніший кореляційний зв'язок між ознаками.

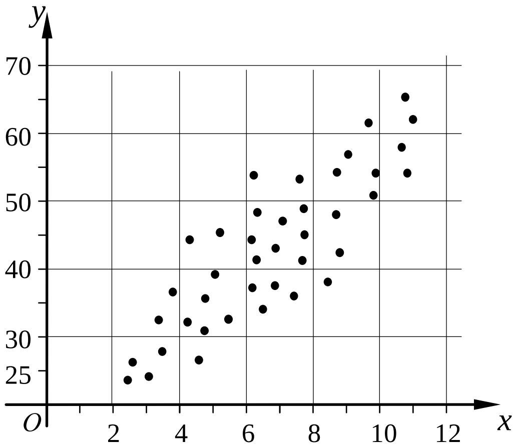


Рис. 15.2

Табл. 15.2

y	x					
	2–4 3	4–6 5	6–8 7	8–10 9	10–12 11	n_{y_j}
20–30 25	4	1				5
30–40 35	2	5	4	1		12
40–50 45		2	8	2		12
50–60 55			2	4	2	8
60–70 65				1	2	3
n_{x_i}	6	8	14	8	4	$n = 40$

15.3. Вибіркові характеристики зв’язку

Основні вибіркові числові характеристики кореляційного зв’язку між випадковими величинами X і Y – це вибірковий кореляційний момент (вибіркова коваріація) і вибірковий коефіцієнт кореляції.

Вибіркова коваріація (*вибірковий кореляційний момент*)

$$K_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \text{ або } K_{xy} = \overline{xy} - \bar{x}\cdot\bar{y} \qquad (1)$$

де $\bar{x}$ та $\bar{y}$ – вибіркові середні, а $\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$.

Якщо деякі пари значень $(x; y)$ спостерігаються n_{xy} разів, то

$$\bar{x} = \frac{1}{n}\sum_i x_i n_{x_i}, \quad \bar{y} = \frac{1}{n}\sum_j y_j n_{y_j} \quad \overline{xy} = \frac{1}{n}\sum n_{xy} xy. \qquad (2)$$

Вибірковий (*емпіричний*) *коефіцієнт кореляції*

$$r_{xy} = \frac{K_{xy}}{S_x \cdot S_y}, \qquad (3)$$

де вибіркове середнє квадратичне відхилення (*стандартна*, або *середня квадратична похибка*) S_x величини X (без урахування кількості ступенів вільності) визначають за однією з формул:

$$S_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad S_x = \sqrt{\frac{1}{n}\sum_{i=1}^{s}(x_i - \bar{x})^2 n_{x_i}}, \quad S_x = \sqrt{\overline{x^2} - \bar{x}^2}; \qquad (4)$$

s – кількість інтервалів змінної X. Вирази для S_y аналогічні.

Квадрати стандартних похибок, тобто величини S_x^2 і S_y^2 – це вибіркові дисперсії змінних X і Y.

Значення r_{XY} може змінюватися від -1 до $+1$. Якщо $\rho_{XY} > 0$, то стохастичний зв’язок між змінними додатний, тобто зі зростанням значень X у середньому збільшується Y. Якщо $r_{XY} < 0$, то стохастичний зв’язок між змінними від’ємний, тобто зі зростанням значень X величина Y в середньому зменшується. Якщо $r_{XY} = \pm 1$, то між змінними існує лінійна функціональна залежність (*повна кореляція*). Якщо $r_{XY} = 0$, то це є свідченням відсутності лише лінійного зв’язку, але не означає, що X і Y статистично незалежні: між ними може існувати інша, відмінна од лінійної, форма залежності.

Робочий варіант формули (3) має вигляд:

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad \text{або} \quad r_{xy} = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{S_x \cdot S_y} \qquad (5)$$

S_x і S_y визначають за формулами (9)–(10).

Якщо деякі пари значень $(x; y)$ спостерігаються у вибірці n_{xy} разів, то формула (5) набирає вигляду:

$$r_{xy} = \frac{\sum n_{xy}xy - n\bar{x}\cdot\bar{y}}{nS_x \cdot S_y}. \qquad (6)$$

У разі двовимірного нормального розподілу коефіцієнт кореляції є мірою лінійної узгодженості між змінними X і Y. Він описує ступінь наближення залежності між випадковими величинами до лінійної функціональної залежності (рис. 15.3).

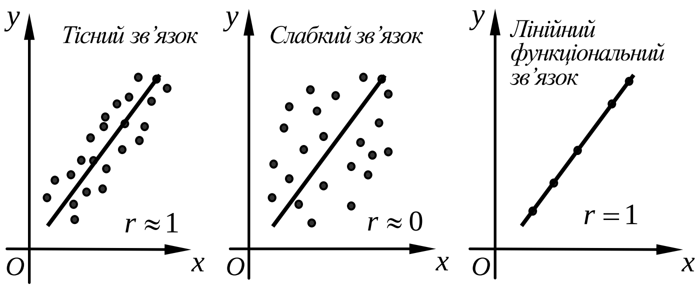


Рис. 15.3

Приклад. Відшукаймо коефіцієнт кореляції r_{xy}, якщо зв'язок між випадковими величинами задано табл. 15.2.

Розв'язання. Скористаємося формулою (6). Нескладно знайти:

$$\sum n_{xy}xy = 4\cdot3\cdot25 + 2\cdot3\cdot35 + 1\cdot5\cdot25 + 5\cdot5\cdot35 + \ldots + 2\cdot11\cdot65 = 12560,$$

$$\bar{x} = 6{,}8,\ \bar{y} = 43,\ S_x = 2{,}358,\ S_y = 11{,}225.$$

Підставивши ці значення у формулу (5) ($n = 40$), отримаємо:

$$r_{xy} = \frac{12560 - 40\cdot6{,}8\cdot43}{40\cdot2{,}358\cdot11{,}225} = 0{,}816.$$

Високе значення коефіцієнта кореляції свідчить про те, що статистичний зв'язок між випадковими величинами X і Y досить близький до лінійного, що видно і з діаграми розсіяння (рис. 15.2). ►

15.4. Парна лінійна регресія

Будемо вважати, що функція регресії лінійна, тобто

$$f(x_i) = \bar{y}(x_i) = a_0 + a_1 x_i \,, \qquad (1)$$

Права частина рівняння регресії залежить від невідомих параметрів a_0, a_1, які необхідно визначити за вибіркою. Основний метод відшукання невідомих параметрів регресії – це метод найменших квадратів (МНК).

Розгляньмо суть МНК. Різниця

$$y_i - f(x_i)$$

– це *відхилення* розрахованого за функцією регресії значення $f(x_i)$ змінної Y і її спостережного значення y_i. Згідно з МНК, невідомі параметри a_0, a_1 функції регресії треба підібрати так, щоб сума квадратів відхилень була мінімальною: $Q = \sum\limits_{i=1}^{n} \left(y_i - f(x_i)\right)^2 \to \min$, тобто

$$\sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2 \to \min \, .$$

Ця сума є функцією від невідомих параметрів a_0 і a_1. Якщо функція диференційовна, то, як відомо, необхідна умова її мінімуму – рівність нулеві перших частинних похідних. Отже, для відшукання невідомих параметрів a_0, a_1 нескладно отримати таку систему двох рівнянь:

$$\begin{cases} \dfrac{\partial Q}{\partial a_0} = 0, \\ \dfrac{\partial Q}{\partial a_1} = 0, \end{cases} \quad \text{або} \quad \begin{cases} -2\sum\limits_{i=1}^{n} (y_i - a_0 - a_1 x_1) = 0, \\ -2\sum\limits_{i=1}^{n} (y_i - a_0 - a_1 x_i) x_i = 0. \end{cases} \qquad (2)$$

Цю систему рівнянь називають *нормальною системою Ґауса*.

Тут a_0, a_1 – шукані величини, x_i, y_i $(i = \overline{1, n})$ вибіркові дані.

Після ділення обох рівнянь системи (2) на n і введення позначень для середніх

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \,, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \,, \quad \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \,, \quad \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i \,, \qquad (3)$$

ця система набирає вигляду:

$$\begin{cases} a_0 + \bar{x}a_1 = \bar{y}, \\ \bar{x}a_0 + \overline{x^2}a_1 = \overline{xy}, \end{cases} \tag{4}$$

звідки отримуємо значення параметрів:

$$a_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{K_{xy}}{S_x^2}, \qquad a_0 = \bar{y} - a_1\bar{x},$$

де K_{xy} – вибіркова коваріація змінних X і Y, яка описує щільність лінійного зв'язку між цими змінними, а S_x^2 – вибіркова дисперсія змінної X.

Середня квадратична похибка рівняння регресії

$$\sigma_R = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n}}. \tag{5}$$

Ця величина є мірою близькості емпіричних даних до теоретичних. Якщо всі фактичні дані збігаються з теоретичними (дані спостережень лежать на прямій), то $\sigma_R = 0$, і маємо ідеальний опис.

Приклад. Обсяг Y виготовленої продукції (в тис. гр. од) в розрахунку на 1 працівника залежно від вартості X основних фондів (в (млн. гр. од) для 10 однорідних підприємств подано в таблиці.

Побудувати рівняння регресії та провести аналіз.

X (млн. гр. од)	1	2	3	4	5	6	7	8	9	10
Y (тис. гр. од)	14	16	15	17	18	20	23	22	24	25

Розв'язання. На підставі форми емпіричної регресійної кривої (на рис. 15.4 – суцільна) можна припустити, що зв'язок між X і Y лінійний.

Нескладно пересвідчитися, що нормальна система рівнянь (4) для заданих показників виглядає так:

$$\begin{cases} 10a_0 + 55a_1 = 195, \\ 55a_0 + 385a_1 = 1176, \end{cases}$$

звідки $a_0 = 12{,}4$, $a_1 = 1{,}27$. Отже, рівняння регресії має вигляд:

$$\hat{y} = 12{,}4 + 1{,}27x.$$

Відповідна пряма зображена на рис. 15.4 пунктиром. Видно, що рівняння регресії добре відтворює фактичну залежність між ознаками.

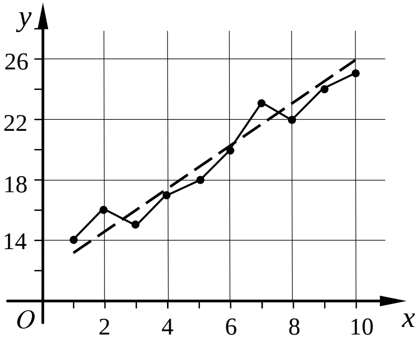


Рис. 15.4

Коефіцієнт $a_1 = 1{,}27$ свідчить, що збільшення основних фондів на 1 млн. гр. од веде до збільшення виробітку одного працівника в середньому на $1{,}27$ тис. гр. од. Параметр $a_0 = 12{,}4$ можна тлумачити як фіксовані витрати, що характеризують середні умови роботи підприємств.

Необхідно розуміти, що значення коефіцієнта a_1, а отже і його тлумачення, залежить від прийнятих одиниць вимірювання.

Вибірковий коефіцієнт кореляції визначаємо за формулою

$$r_{xy} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{S_x \cdot S_y} .$$

Нескладно знайти, що

$$\overline{x} = 5{,}5 \quad \overline{y} = 19{,}4 , \ \overline{xy} = 117{,}2 , \ S_x = 2{,}87 , \ S_y = 3{,}75 .$$

Отже, $r_{xy} \approx 0{,}98$.

Близьке до одиниці значення коефіцієнта кореляції означає, що зв'язок між ознаками близький до лінійного функціонального зв'язку. Це добре ілюструє діаграма розсіяння, зображена на рис. 15.4.

За формулою (5) визначаємо середню квадратичну похибку рівняння регресії: $\sigma_R = 0{,}87$ млн. гр. од. ▶

Додатки. Статистичні таблиці

Додаток 1

Щільність імовірності стандартного

нормального розподілу $f(t) = \dfrac{1}{\sqrt{2\pi}} e^{-t^2/2}$

t	0	1	2	3	4	5	6	7	8	9
0,0	0,3989	0,3989	0,3989	0,3988	0,3986	0,3984	0,3982	0,3980	0,3977	0,3973
0,1	3970	3965	3961	3956	3951	3945	3939	3932	3925	3918
0,2	3910	3902	3894	3885	3876	3867	3856	3847	3836	3825
0,3	3814	3802	3790	3778	3765	3752	3739	3726	3712	3626
0,4	3683	3668	3653	3637	3621	3605	3589	3572	3555	3538
0,5	3521	3503	3485	3467	3448	3429	3410	3391	3372	3352
0,6	3332	3312	3292	3271	3251	3230	3209	3189	3166	3144
0,7	3123	3101	3079	3056	3034	3011	2989	2966	2943	2920
0,8	2897	2874	2850	2827	2803	2780	2756	2732	2709	2685
0,9	2661	2637	2613	2589	2565	2541	2516	2492	2468	2444
1,0	2420	2396	2372	2347	2323	2299	2275	2251	2227	2203
1,1	2179	2155	2131	2107	2085	2059	2036	2012	1989	1965
1,2	1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1,3	1714	1691	1669	1647	0624	1604	0582	1561	1539	1518
1,4	1497	1476	1456	1435	1415	1394	1374	1354	1334	1315
1,5	1295	1276	1257	1238	1219	1200	1182	1163	1145	1127
1,6	1109	1092	1074	1057	1040	1023	1006	0989	0973	0957
1,7	0940	0925	0909	0893	0878	0863	0848	0833	0818	0804
1,8	0790	0775	0761	0748	0734	0721	0707	0694	0681	0669
1,9	0656	0644	0632	0620	0608	0596	0584	0573	0562	0551
2,0	0540	0529	0519	0508	0498	0488	0478	0468	0459	0449
2,1	0440	0431	0422	0413	0404	0396	0387	0379	0371	0363
2,2	0355	0347	0339	0332	0325	0317	0310	0303	0297	0290
2,3	0289	0277	0270	0264	0258	0252	0246	0241	0235	0229
2,4	0224	0219	0213	0208	0203	0198	0194	0189	0184	0180
2,5	0175	0171	0167	0163	0158	0154	0150	0147	0143	0139
2,6	0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2,7	0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2,8	0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2,9	0060	0058	0056	0055	0053	0051	0050	0048	0047	0046
3,0	0044	0043	0042	0040	0039	0038	0037	0036	0035	0034
3,1	0033	0032	0031	0030	0029	0028	0027	0026	0025	0025
3,2	0024	0023	0022	0022	0021	0020	0020	0019	0018	0018
3,3	0017	0017	0016	0016	0015	0015	0014	0014	0013	0013
3,4	0012	0012	0012	0011	0011	0010	0010	0010	0009	0009
3,5	0009	0008	0008	0008	0008	0007	0007	0007	0007	0006
3,6	0006	0006	0006	0005	0005	0005	0005	0005	0005	0004
3,7	0004	0004	0004	0004	0004	0004	0003	0003	0003	0003
3,8	0003	0003	0003	0003	0003	0002	0002	0002	0002	0002
3,9	0002	0002	0002	0002	0002	0002	0002	0002	0001	0001

Додаток 2

Значення функції Лапласа $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_0^z e^{-\frac{t^2}{2}} dt$

z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$
0,00	0,0000	0,40	0,1554	0,80	0,2881	1,20	0,3849
0,01	0,0040	0,41	0,1591	0,81	0,2910	1,21	0,3869
0,02	0,0080	0,42	0,1628	0,82	0,2939	1,22	0,3883
0,03	0,0120	0,43	0,1664	0,83	0,2967	1,23	0,3907
0,04	0,0160	0,44	0,1700	0,84	0,2995	1,24	0,3925
0,05	0,0199	0,45	0,1736	0,85	0,3023	1,25	0,3944
0,06	0,0239	0,46	0,1772	0,86	0,3051	1,26	0,3962
0,07	0,0279	0,47	0,1808	0,87	0,3078	1,27	0,3980
0,08	0,0319	0,48	0,1844	0,88	0,3106	1,28	0,3997
0,09	0,0359	0,49	0,1879	0,89	0,3133	1,29	0,4015
0,10	0,0398	0,50	0,1915	0,90	0,3159	1,30	0,4032
0,11	0,0438	0,51	0,1950	0,91	0,3186	1,31	0,4049
0,12	0,0478	0,52	0,1985	0,92	0,3212	1,32	0,4066
0,13	0,0517	0,53	0,2019	0,93	0,3238	1,33	0,4082
0,14	0,0557	0,54	0,2054	0,94	0,3264	1,34	0,4099
0,15	0,0596	0,55	0,2088	0,95	0,3289	1,35	0,4115
0,16	0,0636	0,56	0,2123	0,96	0,3315	1,36	0,4131
0,17	0,0675	0,57	0,2157	0,97	0,3340	1,37	0,4147
0,18	0,0714	0,58	0,2190	0,98	0,3365	1,38	0,4162
0,19	0,0753	0,59	0,2224	0,99	0,3389	1,39	0,4177
0,20	0,0793	0,60	0,2257	1,00	0,3413	1,40	0,4192
0,21	0,0832	0,61	0,2291	1,01	0,3438	1,41	0,4207
0,22	0,0871	0,62	0,2324	1,02	0,3461	1,42	0,4222
0,23	0,0910	0,63	0,2357	1,03	0,3485	1,43	0,4236
0,24	0,0948	0,64	0,2389	1,04	0,3508	1,44	0,4251
0,25	0,0987	0,65	0,2422	1,05	0,3531	1,45	0,4265
0,26	0,1026	0,66	0,2454	1,06	0,3154	1,46	0,4279
0,27	0,1064	0,67	0,2486	1,07	0,3577	1,47	0,4292
0,28	0,1103	0,68	0,2517	1,08	0,3599	1,48	0,4306
0,29	0,1141	0,69	0,2549	1,09	0,3621	1,49	0,4319
0,30	0,1179	0,70	0,2580	1,10	0,3643	1,50	0,4332
0,31	0,1217	0,71	0,2611	1,11	0,3665	1,51	0,4345
0,32	0,1255	0,72	0,2642	1,12	0,3686	1,52	0,4357
0,33	0,1293	0,73	0,2673	1,13	0,3708	1,53	0,4370
0,34	0,1331	0,74	0,2703	1,14	0,3729	1,54	0,4382
0,35	0,1368	0,75	0,2734	1,15	0,3749	1,55	0,4394
0,36	0,1406	0,76	0,2764	1,16	0,3770	1,56	0,4406
0,37	0,1443	0,77	0,2794	1,17	0,3790	1,57	0,4418
0,38	0,1480	0,78	0,2823	1,18	0,3810	1,58	0,4429
0,39	0,1517	0,79	0,2852	1,19	0,3830	1,59	0,4441

Додаток 2 (закінчення)

z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$
1,60	0,4452	1,85	0,4678	2,20	0,4861	2,70	0,4965
1,61	0,4463	1,86	0,4686	2,22	0,4868	2,72	0,4967
1,62	0,4474	1,87	0,4693	2,24	0,4875	2,74	0,4969
1,63	0,4484	1,88	0,4699	2,26	0,4881	2,76	0,4971
1,64	0,4495	1,89	0,4706	2,28	0,4887	2,78	0,4973
1,65	0,4505	1,90	0,4713	2,30	0,4893	2,80	0,4974
1,66	0,4515	1,91	0,4719	2,32	0,4898	2,82	0,4976
1,67	0,4525	1,92	0,4726	2,34	0,4904	2,84	0,4977
1,68	0,4535	1,93	0,4732	2,36	0,4909	2,86	0,4979
1,69	0,4545	1,94	0,4738	2,38	0,4913	2,88	0,4980
1,70	0,4554	1,95	0,4744	2,40	0,4918	2,90	0,4981
1,71	0,4564	1,96	0,4750	2,42	0,4922	2,92	0,4982
1,72	0,4573	1,97	0,4756	2,44	0,4927	2,94	0,4984
1,73	0,4582	1,98	0,4761	2,46	0,4931	2,96	0,4985
1,74	0,4591	1,99	0,4767	2,48	0,4934	2,98	0,4986
1,75	0,4599	2,00	0,4772	2,50	0,4938	3,00	0,49865
1,76	0,4608	2,02	0,4783	2,52	0,4941	3,20	0,49931
1,77	0,4616	2,04	0,4793	2,54	0,4945	3,40	0,49966
1,78	0,4625	2,06	0,4803	2,56	0,4948	3,60	0,499841
1,79	0,4633	2,08	0,4812	2,58	0,4951	3,80	0,499928
1,80	0,4641	2,10	0,4821	2,60	0,4953	4,00	0,499968
1,81	0,4649	2,12	0,4830	2,62	0,4956	4,50	0,499997
1,82	0,4656	2,14	0,4838	2,64	0,4959	5,00	0,499997
1,83	0,4664	2,16	0,4846	2,66	0,4961		
1,84	0,4671	2,18	0,4854	2,68	0,4963		

Додаток 3

Значення функції $\chi^2_{\alpha;k}$

У таблиці вказано граничні (найбільші) значення $\chi^2_{\alpha;k}$, які випадкова величина, що має χ^2-розподіл із кількістю ступенів вільності k, перевищить зі ймовірністю α.

k \ α	0,99	0,98	0,95	0,90	0,20	0,10	0,05	0,02	0,01	0,001
1	0,000	0,000	0,004	0,016	1,642	2,706	3,841	5,412	6,635	10,827
2	0,020	0,040	0,103	0,211	3,219	4,605	5,991	7,824	9,210	13,815
3	0,115	0,185	0,352	0,584	4,642	6,251	7,815	9,837	11,341	16,268
4	0,297	0,429	0,711	1,064	5,989	7,779	9,488	11,668	13,277	18,465
5	0,554	0,752	1,145	1,610	7,289	9,236	11,070	13,388	15,086	20,517
6	0,872	1,134	1,635	2,204	8,558	10,645	12,592	15,033	16,812	22,457
7	1,239	1,564	2,167	2,833	9,803	12,017	14,067	16,622	18,475	24,322
8	1,645	2,032	2,733	3,490	11,030	13,362	15,507	18,168	20,090	26,125
9	2,088	2,532	3,325	4,168	12,242	14,684	16,919	19,679	21,666	27,877
10	2,558	3,059	3,940	4,865	13,442	15,987	18,307	21,161	23,209	29,588
11	3,053	3,609	4,575	5,578	14,631	17,275	19,675	22,618	24,725	31,264
12	3,571	4,178	5,226	6,304	15,812	18,549	21,026	24,054	26,217	32,909
13	4,107	4,765	5,892	7,042	16,985	19,812	22,362	25,472	27,688	34,528
14	4,660	5,368	6,571	7,790	18,151	21,064	23,685	26,873	29,141	36,123
15	5,229	5,985	7,262	8,547	19,311	22,307	24,996	28,259	30,578	37,697
16	5,812	6,614	7,962	9,312	20,465	23,542	26,296	29,633	32,000	39,252
17	6,408	7,255	8,672	10,085	21,615	24,769	27,587	30,995	33,409	40,790
18	7,015	7,906	9,390	10,865	22,760	25,989	28,869	32,346	34,805	42,312
19	7,633	8,567	10,117	11,651	23,900	27,204	30,144	33,687	36,191	43,820
20	8,260	9,237	10,851	12,443	25,038	28,412	31,410	35,020	37,566	45,315
21	8,897	9,915	11,591	13,240	26,171	29,615	32,671	36,343	38,932	46,797
22	9,542	10,600	12,338	14,041	27,301	30,813	33,924	37,659	40,289	48,268
23	10,196	11,293	13,091	14,848	28,429	32,007	35,172	38,968	41,638	49,728
24	10,856	11,992	13,848	15,659	29,553	33,196	36,415	40,270	42,980	51,179
25	11,524	12,697	14,611	16,473	30,675	34,382	37,652	41,566	44,314	52,620
26	12,198	13,409	15,379	17,292	31,795	35,563	38,885	42,856	45,642	54,052
27	12,879	14,125	16,151	18,114	32,912	36,741	40,113	44,140	46,963	55,476
28	13,565	14,847	16,928	18,939	34,027	37,916	41,337	45,419	48,278	56,893
29	14,256	15,574	17,708	19,768	35,139	39,087	42,557	46,693	49,588	58,302
30	14,953	16,306	18,493	20,599	36,250	40,256	43,773	47,962	50,892	59,703

Додаток 4

t -розподіл Стьюдента

k \ α	0,80	0,40	0,20	0,10	0,05	0,02	0,01	0,001
1	0,325	1,376	3,078	6,314	12,706	31,821	63,657	636,619
2	0,289	1,061	1,866	2,920	4,303	6,965	9,925	31,598
3	0,277	0,978	1,638	2,353	3,182	4,541	5,841	12,941
4	0,271	0,941	1,533	2,132	2,776	3,747	4,604	8,610
5	0,267	0,920	1,476	2,015	2,571	3,365	4,032	6,859
6	0,265	0,906	1,440	1,943	2,447	3,143	3,707	5,959
7	0,263	0,896	1,415	1,895	2,365	2,998	3,499	5,405
8	0,262	0,889	1,397	1,860	2,306	2,896	3,355	5,041
9	0,261	0,883	1,383	1,833	2,262	2,821	3,250	4,781
10	0,260	0,879	1,372	1,812	2,228	2,764	3,169	4,587
11	0,260	0,876	1,363	1,796	2,201	2,718	3,106	4,437
12	0,259	0,873	1,356	1,782	2,179	2,681	3,055	4,318
13	0,259	0,870	1,350	1,771	2,160	2,650	3,012	4,221
14	0,258	0,863	1,345	1,761	2,145	2,624	2,977	4,140
15	0,258	0,866	1,341	1,753	2,131	2,602	2,947	4,073
16	0,258	0,865	1,337	1,746	2,120	2,583	2,921	4,015
17	0,257	0,863	1,333	1,740	2,110	2,567	2,898	3,965
18	0,257	0,862	1,330	1,734	2,101	2,552	2,878	3,922
19	0,257	0,861	1,328	1,729	2,093	2,539	2,861	3,883
20	0,257	0,860	1,325	1,725	2,086	2,528	2,845	3,850
21	0,257	0,859	1,323	1,721	2,080	2,518	2,831	3,819
22	0,256	0,858	1,321	1,717	2,074	2,508	2,819	3,792
23	0,256	0,858	1,319	1,714	2,069	2,500	2,807	3,767
24	0,256	0,857	1,318	1,717	2,064	2,492	2,797	3,745
25	0,256	0,856	1,316	1,708	2,060	2,485	2,787	3,725
26	0,256	0,856	1,315	1,706	2,056	2,479	2,779	3,707
27	0,256	0,855	1,314	1,703	2,052	2,473	2,771	3,690
28	0,256	0,855	1,313	1,701	2,048	2,467	2,763	3,674
29	0,256	0,854	1,311	1,699	2,045	2,462	2,756	3,659
30	0,256	0,854	1,310	1,697	2,042	2,457	2,750	3,646
40	0,255	0,851	1,303	1,684	2,021	2,423	2,704	3,551
60	0,254	0,848	1,296	1,671	2,000	2,390	2,660	3,460
120	0,254	0,845	1,289	1,658	1,980	2,358	2,617	3,373
∞	0,253	0,842	1,282	1,645	1,960	2,326	2,576	3,291

Додаток 5

Розподіл Пуасона $P(X=k)=\dfrac{\lambda^k}{k!}e^{-\lambda}$

k \ λ	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	
0	0,9048	0,8187	0,7408	0,6703	0,6065	0,5488	0,4966	0,4493	0,4066	
1	0,0905	0,1637	0,2222	0,2681	0,3033	0,3293	0,3476	0,3595	0,3659	
2	0,0045	0,0164	0,0333	0,0536	0,0758	0,0988	0,1217	0,1438	0,1647	
3	0,0002	0,0011	0,0033	0,0072	0,0126	0,0198	0,0284	0,0383	0,0494	
4	0,0000	0,0001	0,0003	0,0007	0,0016	0,0030	0,0050	0,0077	0,0111	
5		0,0000	0,0000	0,0001	0,0002	0,0004	0,0007	0,0012	0,0020	
6				0,0000	0,0000	0,0000	0,0001	0,0002	0,0003	
k \ λ	1	2	3	4	5	6	7	8	9	10
0	0,3679	0,1353	0,0498	0,0183	0,0067	0,0025	0,0009	0,0003	0,0001	0,0000
1	0,3679	0,2707	0,1494	0,0733	0,0337	0,0149	0,0064	0,0027	0,0011	0,0005
2	0,1839	0,2707	0,2240	0,1465	0,0842	0,0446	0,0223	0,0107	0,0050	0,0023
3	0,0613	0,1804	0,2240	0,1954	0,1404	0,0892	0,0521	0,0286	0,0150	0,0076
4	0,0153	0,0902	0,1680	0,1954	0,1755	0,1339	0,0912	0,0572	0,0337	0,0189
5	0,0031	0,0361	0,1008	0,1563	0,1755	0,1606	0,1277	0,0916	0,0607	0,0378
6	0,0005	0,0120	0,0504	0,1042	0,1462	0,1606	0,1490	0,1221	0,0911	0,0634
7	0,0001	0,0034	0,0216	0,0595	0,1044	0,1337	0,1490	0,1396	0,1171	0,0901
8	0,0000	0,0009	0,0081	0,0298	0,0653	0,1033	0,1304	0,1396	0,1318	0,1126
9		0,0002	0,0024	0,0132	0,0363	0,0688	0,1014	0,1241	0,1318	0,1251
10		0,0000	0,0008	0,0053	0,0181	0,0413	0,0710	0,0993	0,1186	0,1251
11			0,0002	0,0019	0,0082	0,0213	0,0452	0,0722	0,0970	0,1137
12			0,0001	0,0006	0,0034	0,0126	0,0263	0,0481	0,0728	0,0948
13			0,0000	0,0002	0,0013	0,0052	0,0142	0,0296	0,0504	0,0729
14				0,0001	0,0005	0,0022	0,0071	0,0162	0,0324	0,0521
15				0,0000	0,0002	0,0009	0,0033	0,0090	0,0194	0,0347
16					0,0000	0,0003	0,0014	0,0045	0,0109	0,0217
17						0,0001	0,0006	0,0021	0,0058	0,0128
18						0,0000	0,0002	0,0009	0,0029	0,0071
19							0,0001	0,0004	0,0014	0,0037
20							0,0000	0,0002	0,0006	0,0019
21								0,0001	0,0003	0,0009
22								0,0000	0,0001	0,0004
23									0,0000	0,0002
24										0,0001

Додаток 6.1 Значення функції $F_{\alpha;k_1;k_2}$ (рівень значущості 0,01)

k_2 \ k_1	1	2	3	4	5	6	7	8	9	10	11	12
1	4052	4999	5403	5625	5764	5889	5928	5981	6022	6056	6082	6106
2	98,49	99,01	99,17	99,25	99,30	99,33	99,34	99,36	99,38	99,40	99,41	99,42
3	34,12	30,81	29,48	28,71	28,24	27,91	27,67	27,49	27,34	27,23	27,13	27,05
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,54	14,45	14,37
5	16,26	13,27	12,06	11,39	10,97	10,67	10,45	10,27	10,15	10,05	9,96	9,89
6	13,74	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,79	7,72
7	12,25	9,55	8,45	7,85	7,46	7,19	7,00	6,84	6,71	6,62	6,54	6,47
8	11,26	8,65	7,59	7,01	6,63	6,37	6,19	6,03	5,91	5,81	5,74	5,67
9	10,56	8,02	6,99	6,42	6,06	5,80	5,62	5,47	5,35	5,26	5,18	5,11
10	10,04	7,56	6,55	5,99	5,64	5,39	5,21	5,06	4,95	4,85	4,78	4,71
11	9,85	7,20	6,22	5,67	5,32	5,07	4,88	4,74	4,63	4,54	4,46	4,40
12	9,33	6,93	5,95	5,41	5,06	4,82	4,65	4,50	4,39	4,30	4,22	4,16
14	8,86	6,51	5,56	5,03	4,69	4,46	4,28	4,14	4,03	3,94	3,86	3,80
16	8,53	6,26	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,61	3,55
20	8,10	5,85	4,94	4,43	4,10	3,87	3,71	3,56	3,45	3,37	3,30	3,23
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,25	3,17	3,09	3,03
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,06	2,98	2,90	2,84
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,88	2,80	2,73	2,66
50	7,17	5,06	4,20	3,72	3,41	3,18	3,02	2,88	2,78	2,70	2,62	2,56
70	7,01	4,92	4,08	3,60	3,29	3,07	2,91	2,77	2,67	2,59	2,51	2,45
100	6,90	4,82	3,98	3,51	3,20	2,99	2,82	2,69	2,59	2,51	2,43	2,36
200	6,76	4,71	3,88	3,41	3,11	2,90	2,73	2,60	2,50	2,41	2,34	2,28
400	6,70	4,66	3,83	3,36	3,06	2,85	2,69	2,55	2,46	2,37	2,29	2,23
∞	6,64	4,60	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,24	2,18

Додаток 6.1 (закінчення)

k_2 \ k_1	14	16	20	24	30	40	50	75	100	200	500	∞
1	6142	6169	6208	6234	6258	6286	6302	6323	6334	6352	6361	6366
2	99,43	99,44	99,45	99,46	99,47	99,48	99,48	9,49	99,49	99,49	99,50	99,50
3	26,92	26,83	26,69	26,60	26,50	26,41	26,35	26,27	26,23	2618	26,14	26,12
4	14,24	14,15	14,02	13,93	13,83	13,74	13,69	13,61	13,57	13,52	13,48	13,46
5	9,77	9,68	9,55	9,47	9,38	9,29	9,24	9,17	9,13	9,07	9,04	9,02
6	7,60	7,52	7,39	7,31	7,23	7,14	7,09	7,02	6,99	6,94	6,90	6,88
7	6,35	6,27	6,15	6,07	5,98	5,90	5,85	5,78	5,75	5,70	5,67	5,65
8	5,56	5,48	5,36	5,28	5,20	5,11	5,06	5,00	4,96	4,91	4,88	4,86
9	5,00	4,92	4,80	4,73	4,64	4,56	4,51	4,45	4,41	4,36	4,33	4,31
10	4,60	4,52	4,41	4,33	4,25	4,17	4,12	4,05	4,01	3,96	3,93	3,91
11	4,29	4,21	4,10	4,02	3,94	3,86	3,80	3,74	3,70	3,66	3,62	3,60
12	4,05	3,98	3,86	3,78	3,70	3,61	3,56	3,49	3,46	3,41	3,38	3,36
14	3,70	3,62	3,51	3,43	3,34	3,26	3,21	3,14	3,11	3,06	3,02	3,00
16	3,45	3,37	3,25	3,18	3,10	3,01	2,96	2,89	2,86	2,80	2,77	2,75
20	3,13	3,05	2,94	2,86	2,77	2,69	2,63	2,56	2,53	2,47	2,44	2,42
24	2,93	2,85	2,74	2,66	2,58	2,49	2,44	2,36	2,33	2,27	2,23	2,21
30	2,74	2,66	2,55	2,47	2,38	2,29	2,24	2,16	2,13	2,07	2,03	2,01
40	2,56	2,49	2,37	2,29	2,20	2,11	2,05	1,97	1,94	1,88	1,84	1,81
50	2,46	2,39	2,26	2,18	2,10	2,00	1,94	1,86	1,82	1,76	1,71	1,68
70	2,35	2,28	2,15	2,07	1,98	1,88	1,82	1,74	1,69	1,62	1,56	1,53
100	2,26	2,19	2,06	1,98	1,89	1,79	1,73	1,64	1,59	1,51	1,46	1,43
200	2,17	2,09	1,97	1,88	1,79	1,69	1,62	1,53	1,48	1,39	1,33	1,28
400	2,12	2,04	1,92	1,84	1,74	1,64	1,57	1,47	1,42	1,32	1,24	1,19
∞	2,07	1,99	1,87	1,79	1,69	1,59	1,52	1,41	1,36	1,25	1,15	1,09

Додаток 6.2. Значення функції $F_{\alpha;k_1;k_2}$ (рівень значущості 0,05)

k_2 \ k_1	1	2	3	4	5	6	7	8	9	10	11
1	161	200	216	225	230	234	237	239	241	242	243
2	18,51	19,00	19,16	19,25	19,30	19,33	19,36	19,37	19,38	19,39	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,88	8,84	8,81	8,78	8,76
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,93
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,78	4,74	4,70
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,63	3,60
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,34	3,31
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,13	3,10
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,97	2,94
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,86	2,82
12	4,75	3,88	3,49	3,26	3,11	3,00	2,92	2,85	2,80	2,76	2,72
14	4,60	3,74	3,34	3,11	2,96	2,85	2,77	2,70	2,65	2,60	2,56
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,45
20	4,35	3,49	3,10	2,87	2,71	2,60	2,52	2,45	2,40	2,35	2,31
24	4,26	3,40	3,01	2,78	2,62	2,51	2,43	2,36	2,30	2,26	2,22
30	4,17	3,32	2,92	2,69	2,53	2,42	2,34	2,27	2,21	2,16	2,12
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,07	2,04
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,02	1,98
70	3,98	3,13	2,84	2,50	2,35	2,23	2,14	2,07	2,01	1,97	1,93
100	3,94	3,09	2,70	2,46	2,30	2,19	2,10	2,03	1,97	1,92	1,88
200	3,89	3,04	2,65	2,41	2,26	2,14	2,05	1,98	1,92	1,87	1,83
400	3,86	3,02	2,62	2,39	2,23	2,12	2,03	1,96	1,90	1,85	1,81
∞	3,84	2,99	2,60	2,37	2,21	2,09	2,01	1,94	1,88	1,83	1,79

Додаток 6.2 (закінчення)

k_2 \ k_1	12	13	16	20	30	40	50	75	100	200	500	∞
1	244	245	246	248	250	251	252	253	253	254	254	254
2	19,41	19,42	19,43	19,44	19,46	19,47	19,47	19,48	19,49	19,49	19,50	19,50
3	8,74	8,71	8,69	8,66	8,62	8,60	8,58	8,57	8,56	8,54	8,54	8,53
4	5,91	5,87	5,84	5,80	5,74	5,71	5,70	5,68	5,66	5,65	5,64	5,63
5	4,68	4,64	4,60	4,56	4,50	4,46	4,44	4,42	4,40	4,38	4,38	4,36
6	4,00	3,96	3,92	3,87	3,81	3,77	3,75	3,72	3,71	3,69	3,68	3,76
7	3,57	3,52	3,49	3,44	3,38	3,34	3,32	3,29	3,28	3,25	3,24	3,23
8	3,28	3,23	3,20	3,15	3,08	3,05	3,03	3,00	2,98	2,96	2,94	2,93
9	3,07	3,02	2,98	2,93	2,86	2,82	2,80	2,77	2,76	2,73	2,72	2,71
10	2,91	2,86	2,82	2,77	2,70	2,67	2,64	2,61	2,59	2,56	2,55	2,54
11	2,79	2,74	2,70	2,65	2,57	2,53	2,50	2,47	2,45	2,42	2,41	2,40
12	2,69	2,64	2,60	2,54	2,46	2,42	2,40	2,36	2,35	2,32	2,31	2,30
14	2,53	2,48	2,44	2,39	2,31	2,27	2,24	2,21	2,19	2,16	2,14	2,13
16	2,42	2,37	2,33	2,28	2,20	2,16	2,13	2,09	2,07	2,04	2,02	2,01
20	2,28	2,23	2,18	2,12	2,04	1,99	1,96	1,92	1,90	1,87	1,85	1,84
24	2,18	2,13	2,09	2,02	1,94	1,89	1,86	1,82	1,80	1,76	1,74	1,73
30	2,09	2,04	1.99	1,93	1,84	1,79	1,76	1,72	1,69	1,66	1,64	1,62
40	2,00	1,95	1,90	1,84	1,74	1,69	1,66	1,61	1,59	1,55	1,53	1,51
50	1,95	1,90	1,85	1,78	1,69	1,63	1,60	1,55	1,52	1,48	1,46	1,44
70	1,89	1,84	1,79	1,72	1,62	1,56	1,53	1,47	1,45	1,40	1,37	1,35
100	1,85	1,79	1,75	1,68	1,57	1,51	1,48	1,42	1,39	1,34	1,30	1,28
200	1,80	1,74	1,69	1,62	1,52	1,45	1,42	1,35	1,32	1,26	1,22	1,19
400	1,78	1,72	1,67	1,60	1,49	1,42	1,38	1,32	1,28	1,22	1,16	1,13
∞	1,75	1,69	1,64	1,57	1,46	1,40	1,45	1,28	1,24	1,17	1,11	1,00

Зміст

Підписано до друку 10.023.2020 р. Формат 60x84/$_{16}$.
Папір ксероксний. Гарнітура Times.
Ум. друк. арк. 4,67.
Наклад 50 прим.