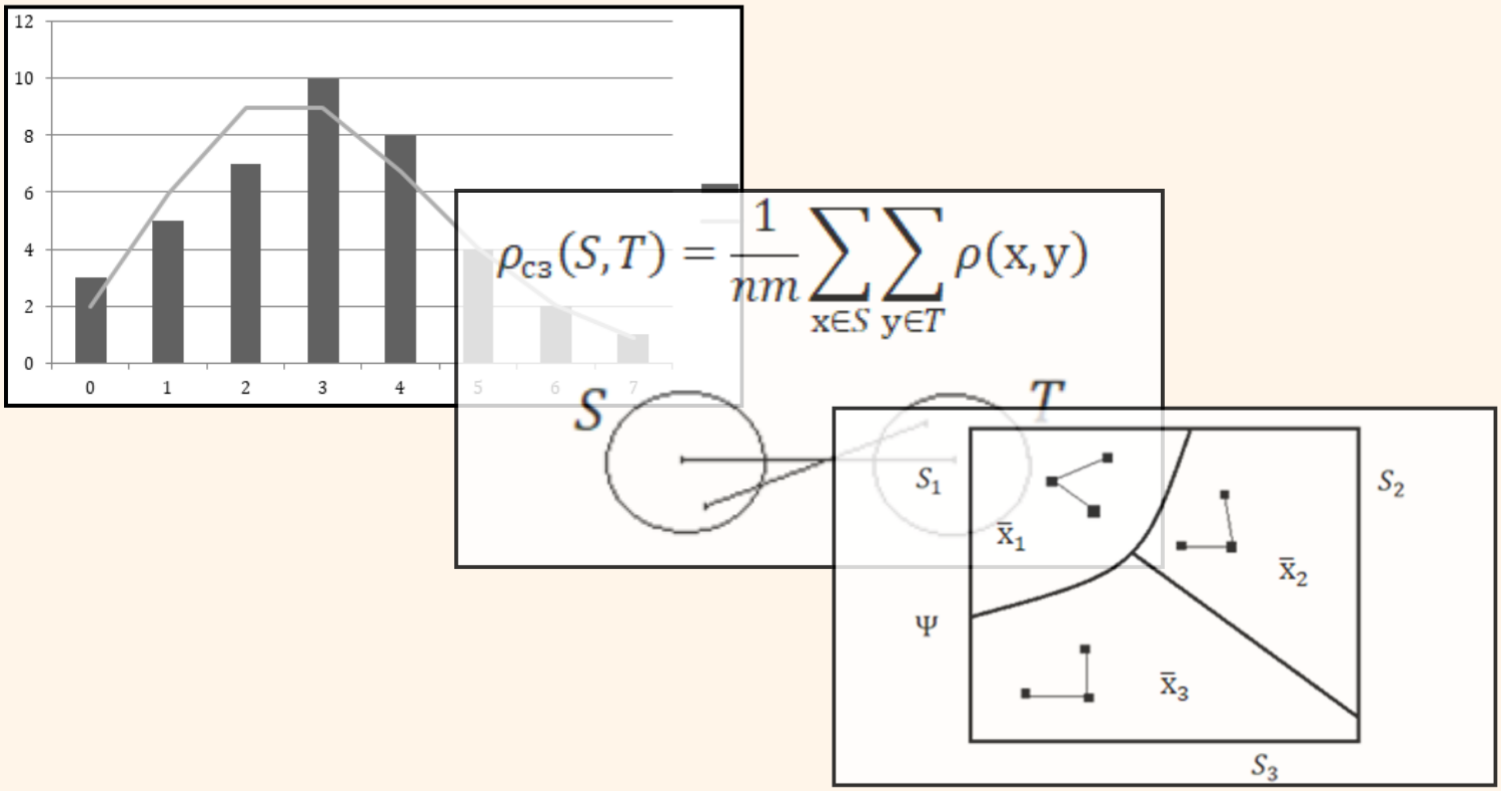
Навчальний посібник

МАТЕМАТИЧНА СТАТИСТИКА



ПЛІЧКО А.М.
АКБАШ К.С.
ЛУНЬОВА М.В.

Статистичні розподіли вибірок та їхні числові характеристики
Статистичні оцінки параметрів генеральної сукупності
Статистичні гіпотези
Дисперсійний аналіз
Кореляційний аналіз
Регресійний аналіз
Непараметричні критерії однорідності статистичних даних
Методи багатовимірної класифікації

2024

Навчальний посібник

МАТЕМАТИЧНА СТАТИСТИКА

Плічко А.М., Акбаш К.С., Луньова М.В.

Кропивницький
Видавець Лисенко С.В.
2024

УДК 519.2
П40

Плічко А.М., Акбаш К.С., Луньова М.В.

П40 Математична статистика: навчальний посібник. – Кропивницький: «КОД», 2024. – 220 с.

ISBN 978-617-8494-04-9

Анотація. Навчальний посібник призначений для студентів спеціальності 112 Статистика, а також дослідників, які займаються міждисциплінарними дослідженнями і використовують методи математичної статистики. Структура посібника охоплює базові теми курсу «Математична статистика» і включає наступні розділи: «Статистичні розподіли вибірок та їхні числові характеристики»; «Статистичні оцінки параметрів генеральної сукупності»; «Статистичні гіпотези»; «Дисперсійний аналіз»; «Кореляційний аналіз»; «Регресійний аналіз»; «Непараметричні критерії однорідності статистичних даних»; «Методи багатовимірної класифікації». Навчальний посібник має на меті виробити вміння та навички застосування теорії математичної статистики до базових задач, що потребують статистичної обробки даних.

Рецензенти:

Мацак І.К., доктор фізико-математичних наук, професор кафедри дослідження операцій факультету комп'ютерних наук та кібернетики Київського національного університету імені Тараса Шевченка;

Чуйков А.С., кандидат фізико-математичних наук, заступник директора з навчально-методичної роботи ВСП «Київський фаховий коледж комп'ютерних технологій та економіки НАУ».

Рекомендовано вченою радою Центральноукраїнського державного університету імені Володимира Винниченка (протокол №9 від 9.03.2024 р.)

ISBN 978-617-8494-04-9

ЗМІСТ

ВСТУП

Математична статистика – це наука, яка вивчає методи обробки результатів спостережень масових випадкових подій, які володіють статистичною стійкістю, тобто закономірністю, з метою виявлення цих закономірностей. Висновки про закономірності, яким підпорядковуються явища, які досліджуються методами математичної статистики, завжди ґрунтуються на обмеженій кількості спостережень. При більшій кількості спостережень ці висновки можуть виявитися іншими. Для отримання точніших висновків щодо закономірностей явищ, математична статистика спирається на теорію ймовірностей.

Хоча математична статистика і спирається на методи і поняття теорії ймовірностей, у деякому розумінні вона розв'язує обернені задачі. Так, у теорії ймовірностей припускають, що ймовірнісна модель явища задана, і на основі цієї моделі обчислюють відповідні ймовірності подій. У математичній статистиці припускають, що ймовірнісна модель явища невідома, і виконують наступне: Припустимо, що у результаті проведених спостережень отримані деякі експериментальні дані. На основі цих даних обирають відповідну їм ймовірнісну модель. А вже далі використовують отриману модель для опису явища чи процесу, який розглядається.

На сьогодні математична статистика – великий розділ математики. У цьому навчальному посібнику розглядаються основні поняття математичної статистики і задачі, які вона розв'язує: оцінювання невідомих параметрів розподілів імовірностей; перевірки параметричних та непараметричних статистичних гіпотез; встановлення форми та ступеня зв'язку між декількома випадковими величинами; методи багатовимірної класифікації. У посібнику також значна увага приділена непараметричним методам математичної статистики, які закладені у сучасних статистичних пакетах і активно використовуються в практичних задачах різних галузей знань.

Навчальний посібник має блочну структуру. У кожному розділі після викладення теоретичного матеріалу подано типові приклади з розв'язаннями, а також питання для самоконтролю. У кінці посібника наведений список використаної літератури та таблиці основних розподілів. Також для кращого розуміння тексту студентам рекомендується повторити основні поняття та твердження з теорії ймовірностей. На багато з них ми наводимо посилання. Посилання в основному тексті на них природні. Наприклад, [6, п.1.11.6] означає посилання на підрозділ 1.11.6 з книжки під номером 6 у списку літератури.

Книга буде корисна як для студентів спеціальності 112 Статистика, так і для науковців, викладачів, аспірантів і студентів природничих і технічних спеціальностей, які використовують у своїх дослідженнях методи математичної статистики.

РОЗДІЛ 1. СТАТИСТИЧНІ РОЗПОДІЛИ ВИБІРОК ТА ЇХНІ ЧИСЛОВІ ХАРАКТЕРИСТИКИ

Основною ***метою*** математичної статистики є систематизація, обробка та використання статистичної інформації. А саме, у статистиці розглядається певна множина однотипних елементів Ω і потрібно знайти певну властивість цих елементів. Це і є метою математичної статистики.

Наприклад, нехай Ω – сукупність яєць виготовлених птахофабрикою протягом місяця. Потрібно оцінити середню вагу одного яйця. Але неможливо перевірити кожне яйце. Тому в статистиці використовують ***вибірковий метод***. Вибирають якусь кількість n яєць, зважують їх, додають результати і суму ділять на n.

Введемо відповідні формальні означення.

Означення. Множина Ω однотипних елементів, яким притаманні певні кількісні ознаки (розмір, вага, тощо) називається ***генеральною сукупністю***. Кількість N елементів множини Ω називають її ***обсягом***. Саме число N здебільшого велике і невідоме.

Означення. Кожна підмножина $A \subset \Omega$ називається ***вибіркою***. Кількість n елементів множини A називають її ***обсягом***. У статистиці це число відоме і значно менше від N $(n \ll N)$.

У математичній статистиці розглядають, зокрема, дві категорії ***задач***:

1. Статистичне оцінювання параметрів генеральної сукупності.
2. Перевірка статистичних гіпотез.

Приклад статистичного оцінювання параметра – оцінка середньої ваги яйця. Звичайно, тут відразу виникає питання точності оцінювання. Це питання детально буде розглянуто далі.

З перевіркою гіпотез справа виглядає складніше, навіть у випадках, подібних до визначення середньої ваги яйця. Наприклад, нехай якась пекарня пече булочки з родзинками. І нехай надійшла скарга, що пекарня не додає родзинок. Скажімо, у булочці має бути 10 родзинок, а покупці скаржаться, що їх менше. Це і є ***статистична гіпотеза***. Далі, починаємо робити так само, як і у випадку з яйцями. Беремо n булочок і підраховуємо середню кількість родзинок. Нехай, наприклад, виявилося 9,5. Чи достатньо цього, щоб зробити висновок про злодійство на пекарні? Може було взяте мале n? Можливо, це трапилася така партія булочок? Детальніше проблеми перевірки статистичних гіпотез також розглянемо далі.

1.1. Варіаційний ряд

а) Дискретний випадок

Отже, нехай задано генеральну сукупність Ω і кількісну ознаку X її елементів, яка набуває скінченну кількість значень. Тобто, задано функцію $X(\omega)$, $\omega \in \Omega$, яка може набувати скінченну кількість значень (як правило невелику). Наприклад, Ω – булочки з родзинками, $X(\omega)$ – кількість родзинок в ω-тій булочці.

Нехай тепер зроблена якась вибірка A з елементами $\omega_1, \ldots, \omega_n$. На цих елементах функція X набуває значень

$$X(\omega_1),\ X(\omega_2), \ldots, X(\omega_n). \qquad (1.1)$$

У прикладі – це кількість родзинок у булочці.

Деякі, а може й багато, чисел в наборі (1.1) можуть бути однаковими. Нехай серед чисел (1.1)

x_1 зустрічається n_1 раз

x_2 зустрічається n_2 раз

... ...

x_k зустрічається n_k раз.

і на множині A інших значень функції $X(\omega)$ не набуває.

Звичайно,

$$n_1 + n_2 + \cdots + n_k = n.$$

Означення. числа x_1,..., x_k, розташовані в порядку зростання, називаються ***варіаційним рядом***, кожен елемент цього ряду називається ***варіантою***, а відношення

$$w_i = \frac{n_i}{n},\ \ i = 1, \ldots k\ \ -$$

відносною частотою варіанти x_i.

Звичайно,

$$w_1 + w_2 + \ldots + w_k = \frac{n_1}{n} + \frac{n_2}{n} + \ldots + \frac{n_k}{n} = 1\ .$$

б) Неперервний випадок

Числова ознака X генеральної сукупності Ω може бути й неперервною. Наприклад – вага яйця.

У цьому випадку поняття варіаційного ряду дещо інше. А саме, для вибірки A множину значень (1.1) (розташовану в порядку зростання) ділять на інтервали (рис.1.1).

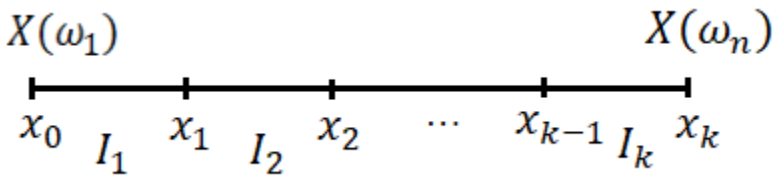


Рис. 1.1

Позначимо через

n_1 – кількість чисел з набору (1.1), які потрапили до проміжку $I_1\ = [x_0, x_1]$;

n_2 – кількість чисел з набору (1.1), які потрапили до проміжку $I_2 = (x_1, x_2]$;

……

n_k – кількість чисел з набору (1.1), які потрапили до проміжку $I_k = (x_{k-1}, x_k]$.

Тут проміжки I_1, I_2,..., I_k називають ***інтервальним варіаційним рядом***, числа

1.2. Дискретний статистичний розподіл вибірки та його числові характеристики

Отже, нехай кількісна ознака X елементів генеральної сукупності Ω для вибірки A набуває значень x_1, x_2,..., x_k відповідно n_1, n_2,..., n_k разів.

Означення. ***Дискретним статистичним розподілом*** вибірки A називається таблиця

x_1	x_2	...	x_k
n_1	n_2	...	n_k
w_1	w_2	...	w_k

Позначення. Для вибірки A і числа $x \in \mathbf{R}$ нехай n_x – кількість значень (1.1) (розташовану в порядку зростання) менших від x (рис. 1.2).

$$\underbrace{X(\omega_1),\ X(\omega_2)\ \ldots}_{n_x\qquad\qquad x}\ .\ \ldots\quad X(\omega_n)$$

Рис. 1.2

Зауваження. Якщо $x_{i-1} < x < x_i$, то $n_x = n_1 + n_2 + n_{i-1}$.

Означення. Функція $F_n(x) = \frac{n_x}{n}$ називається ***емпіричною функцією розподілу вибірки*** A.

Властивості емпіричної функції розподілу:

1) $0 \le F_n(x) \le 1$;

2) $F_n(x_1) = 0$;

3) $F_n(x_k) = 1$;

4) $F_n(x)$ неспадна;

5) $F_n(x)$ неперервна зліва.

У розгорнутому вигляді емпіричну функцію розподілу можна записати так:

$$F_n(x) = \begin{cases} 0, & x \le x_1 \\ w_1, & x_1 < x \le x_2 \\ w_1 + w_2, & x_2 < x \le x_3 \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ w_1 + w_2 + \dots + w_i, & x_i < x \le x_{i+1} \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ w_1 + w_2 + \dots + w_k = 1, & x > x_k. \end{cases}$$

Графік емпіричної функції розподілу зображений на рис.1.3.

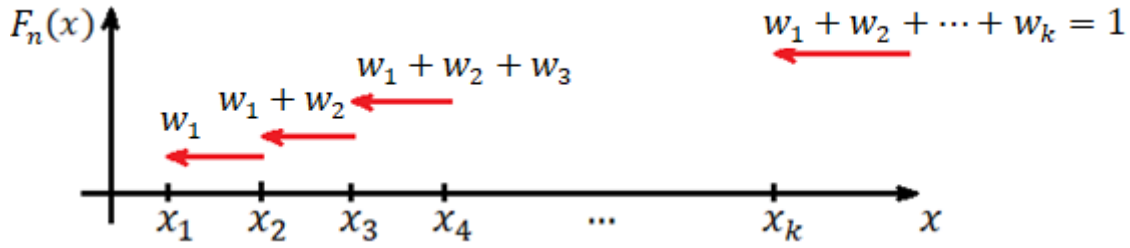


Рис. 1.3

Полігон частот і відносних частот

Дискретний статистичний розподіл вибірки можна зобразити графічно у вигляді ламаних з вершинами (x_i, n_i) або (x_i, w_i).

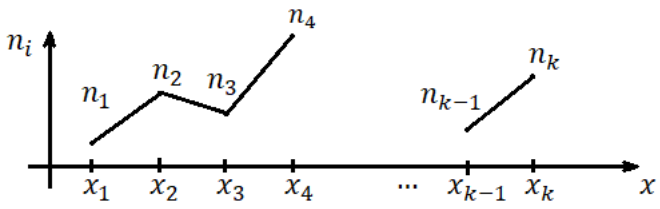


Рис. 1.4

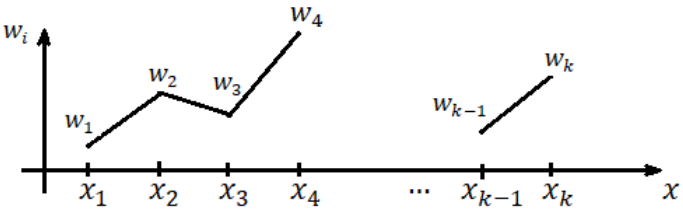


Рис. 1.5

Першу ламану називають ***полігоном частот*** (рис. 1.4), а другу – ***полігоном відносних частот*** (рис. 1.5).

Числові характеристики вибірки

1) ***Вибіркова середня величина*** вибірки A визначається формулою

$$\bar{x} = \sum_{i=1}^{k} x_i w_i = \frac{1}{n} \sum_{i=1}^{k} x_i n_i.$$

2) ***Відхиленням*** варіанти x_i $(i = 1, \dots, k)$ називають число

$$(x_i - \bar{x}) \cdot n_i.$$

Зауваження. $\sum_{i=1}^{k}(x_i - \bar{x}) \cdot n_i = \sum_{i=1}^{k} x_i \cdot n_i - \sum_{i=1}^{k} \bar{x} \cdot n_i = n\bar{x} - n\bar{x} = 0.$

3) ***Модою*** вибірки A називають варіанту, що має найбільшу частоту появи. Її позначають символом Mo. Ілюстрація моди зображена на рис. 1.6.

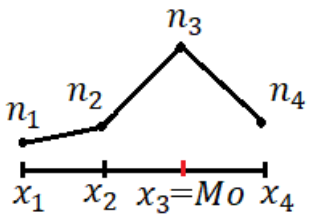


Рис. 1.6

Зауваження. Мода вибірки може бути і не одна. Якщо варіаційний ряд має одну моду, то він називається ***унімодальним***, якщо дві – ***двомодальним*** і т.д.

4) ***Медіаною*** дискретного статистичного розподілу називають число, яке поділяє варіаційний ряд на дві частини, рівні за кількістю частот. Позначають її через Me.

На рис. 1.6 зображена медіана дискретного розподілу.

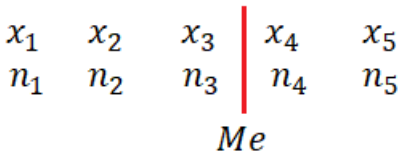


Рис. 1.7

Уточнення. Звичайно, для дискретного розподілу такого числа Me може й не існувати. Тому, спочатку вибирають таке значення x_i, що різниця між кількістю елементів варіаційного ряду менших x_i, і кількістю елементів цього ряду більших від x_i, є мінімальною, а потім за моду вважають середину проміжку (x_i, x_{i+1}).

5) ***Дисперсія вибірки*** відображає розсіяння варіант навколо середньої вибіркової величини і визначається формулою

$$D = \sum_{i=1}^{k}(x_i - \bar{x})^2 w_i.$$

Очевидно,

$$D = \sum_{i=1}^{k} {x_i}^2 w_i - 2\sum_{i=1}^{k} x_i \bar{x} w_i + \sum_{i=1}^{k} \bar{x}^2 w_i = \sum_{i=1}^{k} {x_i}^2 w_i - \bar{x}^2.$$

6) ***Середнє квадратичне відхилення вибірки*** σ – це корінь квадратний з дисперсії

$$\sigma = \sqrt{D}.$$

При обчисленні дисперсії відхилення підноситься до квадрату, тобто змінюється одиниця виміру ознаки X. Для того щоб отримати розсіювання варіант вибірки відносно $\bar{x}$ в тих самих одиницях, в яких вимірюється X, і вводиться величина σ.

1.3. Інтервальний статистичний розподіл вибірки та його числові характеристики

Нагадаємо, що у випадку неперервної числової ознаки X і вибірки A множина значень

$$X(\omega_1),\ X(\omega_2),\dots,X(\omega_n) \tag{1.2}$$

ділилася на інтервали точками

$$x_0 < x_1 < x_2 < \cdots < x_k\,.$$

Інтервальним статистичним розподілом вибірки *A* називається таблиця вигляду:

$[x_0, x_1]$	$(x_1, x_2]$	...	$(x_{k-1}, x_k]$
n_1	n_2	...	n_k
w_1	w_2	...	w_k

де, звичайно,

$$w_i = \frac{n_i}{n},\ \ i = 1,\dots k.$$

Як правило, довжини інтервалів беруться однаковими і тоді таку довжину позначають через h.

Як і у випадку дискретного розподілу, інтервальний статистичний розподіл можна проілюструвати за допомогою гістограм частот, або відносних частот.

Гістограма частот – це фігура, яка складається із прямокутників з основою довжини h і висотою $\frac{n_i}{h}$. На рис. 1.8 зображений приклад гістограми частот інтервального розподілу.

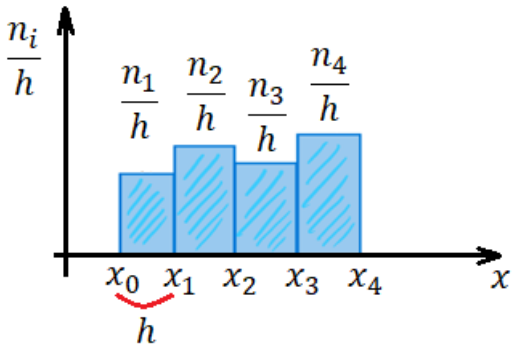


Рис. 1.8

Гістограма відносних частот є фігурою, яка складається з прямокутників, кожен з яких має основу довжиною h і висотою $\frac{w_i}{h}$. На рис. 1.9 зображений приклад гістограми відносних частот інтервального розподілу.

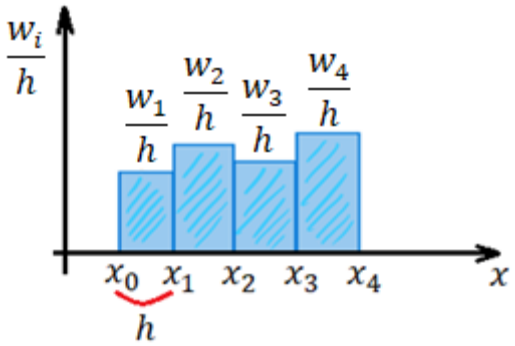


Рис. 1.9

Зауваження.

1. Площа гістограми частот дорівнює

$$S = \sum_{i=1}^{k} h\frac{n_i}{h} = \sum_{i=1}^{k} n_i = n$$

– обсяг вибірки.

2. Площа гістограми відносних частот дорівнює

$$S = \sum_{i=1}^{k} h\frac{w_i}{h} = \sum_{i=1}^{k} w_i = 1.$$

Емпірична функція розподілу $F_n(x)$ для неперервної числової ознаки X і поділу множини її значень на проміжки $I_1, I_2,..., I_k$ задається наступним чином:

$$F_n(x_0) = 0;$$
$$F_n(x_1) = w_1;$$
$$F_n(x_2) = w_1 + w_2;$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$
$$F_n(x_{k-1}) = w_1 + w_2 + \cdots + w_{k-1};$$
$$F_n(x_k) = w_1 + w_2 + \cdots + w_k = 1.$$

Далі покладемо

$$F_n(x) = 0 \text{ на проміжку } (-\infty, x_0)$$
$$F_n(x) = 1 \text{ на проміжку } [x_k, +\infty),$$

а решту точок $(x_i, w_1 + \cdots + w_i)$, $(x_{i+1}, w_1 + \cdots + w_{i+1})$ з'єднаємо ламаною (рис. 1.10).

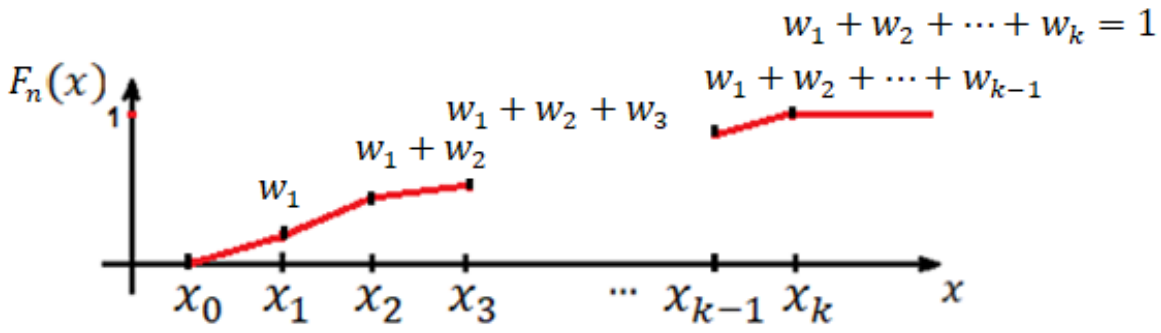


Рис. 1.10

Введемо основні числові характеристики для інтервального розподілу.

Медіана. Проілюструємо спочатку це поняття на графіку емпіричної функції (рис. 1.10).

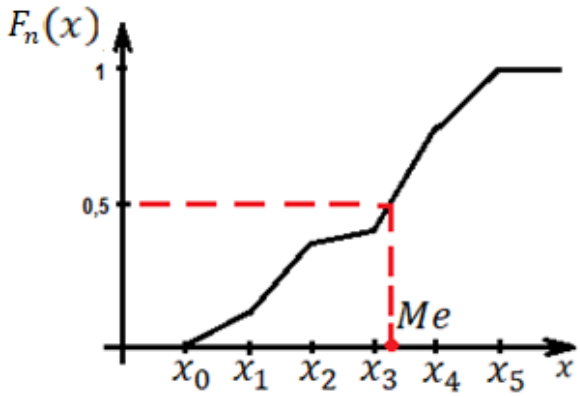


Рис. 1.10

Отже, медіана – це розв'язок рівняння $F_n(x) = 0{,}5$.

Для числового знаходження медіани визначаємо спочатку ***медіанний частинний інтервал***, точніше таку точку x_i, що $F_n(x_{i-1}) \le 0{,}5$, а $F_n(x_i) \ge 0{,}5$. А далі, розв'язуючи лінійне рівняння, знаходимо Me. А саме, з рівняння прямої, що проходить через дві точки, маємо

$$y = F_n(x_{i-1}) + \frac{F_n(x_i) - F_n(x_{i-1})}{h}(x_i - x_{i-1}) = 0{,}5.$$

Звідси

$$Me = x_{i-1} + \frac{0{,}5 - F_n(x_{i-1})}{F_n(x_i) - F_n(x_{i-1})}h.$$

Мода. Для визначення моди потрібно спочатку знайти ***модальний частинний інтервал***, тобто інтервал, до якого потрапляє найбільша кількість значень $X(\omega_j)$. Нехай, це буде інтервал (x_{i-1}, x_i). Тоді, за означенням,

$$Mo = x_{i-1} + \frac{n_i - n_{i-1}}{2n_i - n_{i-1} - n_{i+1}}h.$$

Для визначення середньої величини, дисперсії і середнього квадратичного відхилення для інтервального статистичного розподілу переходять спочатку до відповідного дискретного розподілу. А саме, нехай для $i = 1, \dots, k$

$$x_i^* = \frac{x_{i-1} + x_i}{2}.$$

Тобто x_i^* – це середина відрізка $[x_{i-1},\ x_i]$.

Середня величина інтервального розподілу визначається формулою

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i^* n_i.$$

Дисперсія інтервального розподілу визначається формулою

$$D = D_n = \frac{1}{n}\sum_{i=1}^{k} (x_i^* - \bar{x})^2\, n_i.$$

Середнє квадратичне відхилення інтервального розподілу, це –

$$\sigma = \sigma_n = \sqrt{D}.$$

1.4. Двовимірний дискретний статистичний розподіл вибірки та його числові характеристики

Нехай тепер задано дві (дискретні) числові ознаки X та Y і вибірку $A \subset \Omega$. Наприклад, горішки і родзинки в пакетиках. Таблиця розподілу для них має вигляд:

	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	$\boldsymbol{x_3}$	…	$\boldsymbol{x_j}$	…	$\boldsymbol{x_m}$	$\boldsymbol{n_{y_i}}$
$\boldsymbol{y_1}$	n_{11}	n_{12}	n_{13}	…	n_{1j}	…	n_{1m}	$\boldsymbol{n_{y_1}}$
$\boldsymbol{y_2}$	n_{21}	n_{22}	n_{23}	…	n_{2j}	…	n_{2m}	$\boldsymbol{n_{y_2}}$
$\boldsymbol{y_3}$	n_{31}	n_{32}	n_{33}	…	n_{3j}	…	n_{3m}	$\boldsymbol{n_{y_3}}$
…	…	…	…	…	…	…	…	…
$\boldsymbol{y_i}$	n_{i1}	n_{i2}	n_{i3}	…	n_{ij}	…	n_{im}	$\boldsymbol{n_{y_i}}$
…	…	…	…	…	…	…	…	…
$\boldsymbol{y_k}$	n_{k1}	n_{k2}	n_{k3}	…	n_{kj}	…	n_{km}	$\boldsymbol{n_{y_k}}$
$\boldsymbol{n_{x_j}}$	$\boldsymbol{n_{x_1}}$	$\boldsymbol{n_{x_2}}$	$\boldsymbol{n_{x_3}}$	…	$\boldsymbol{n_{x_j}}$	…	$\boldsymbol{n_{x_m}}$	$\boldsymbol{n}$

і називається ***двовимірним дискретним статистичним розподілом вибірки***.

Тут n_{ij} – частота спільної появи ознак $Y = y_i$, $X = x_j$, а

$$n_{y_i} = \sum_{j=1}^{m} n_{ij}\ , \quad n_{x_j} = \sum_{i=1}^{k} n_{ij}.$$

Зауваження. Звичайно,

$$\sum_{i=1}^{k} n_{y_i} = \sum_{j=1}^{m} n_{x_j} = \sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij} = n,$$

де n – обсяг вибірки A.

Загальні числові характеристики ознаки *X*

Загальне середнє ознаки X має вигляд

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{k} x_j n_{ij} = \frac{1}{n}\sum_{j=1}^{m} x_j n_{x_j}\,.$$

Загальна дисперсія ознаки X має вигляд

$$D_x = \frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{k} x_j^2 n_{ij} - (\bar{x})^2 = \frac{1}{n}\sum_{j=1}^{m} x_j^2 n_{x_j} - (\bar{x})^2.$$

Загальне середнє квадратичне відхилення ознаки X це –

$$\sigma_x = \sqrt{D_x}.$$

Загальні числові характеристики ознаки *Y*

Загальне середнє ознаки Y має вигляд

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{m} y_i n_{ij} = \frac{1}{n}\sum_{i=1}^{k} y_i n_{y_i}.$$

Загальна дисперсія ознаки Y має вигляд

$$D_y = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{m} y_i^2 n_{ij} - (\bar{y})^2 = \frac{1}{n}\sum_{i=1}^{k} y_i^2 n_{y_i} - (y)^2.$$

Загальне середнє квадратичне відхилення ознаки Y це –

$$\sigma_y = \sqrt{D_y}\,.$$

Умовні статистичні розподіли та їхні числові характеристики

Означення. ***Умовним статистичним розподілом ознаки Y при фіксованому значенні $X = x_j$*** називають таблицю варіант ознаки Y, та відповідних їм частот, взятих при фіксованому значенні $X = x_j$

y_1	y_2	y_3	…	y_k
n_{1j}	n_{2j}	n_{3j}	…	n_{kj}

Числові характеристики для такого статистичного розподілу називають ***умовними***. До них належать:

Умовне середнє ознаки Y при $X = x_j$:

$$\bar{y}|_{X=x_j} = \frac{1}{n_{x_j}} \sum_{i=1}^{k} y_i n_{ij}.$$

Умовна дисперсія ознаки Y при $X = x_j$:

$$D\left(Y|_{X=x_j}\right) = \frac{1}{n_{x_j}} \sum_{i=1}^{k} y_i^2 n_{ij} - \left(\bar{y}|_{X=x_j}\right)^2.$$

Умовне середнє квадратичне відхилення Y при $X = x_j$:

$$\sigma\left(Y|_{X=x_j}\right) = \sqrt{D\left(Y|_{X=x_j}\right)}.$$

Означення. ***Умовним статистичним розподілом ознаки X при фіксованому значенні $Y = y_i$*** називають таблицю

x_1	x_2	x_3	…	x_m
n_{i1}	n_{i2}	n_{i3}	…	n_{im}

Умовні числові характеристики для цього розподілу такі:

Умовне середнє ознаки X при $Y = y_i$:

$$\bar{x}|_{Y=y_i} = \frac{1}{n_{y_i}} \sum_{j=1}^{m} x_j n_{ij}.$$

Умовна дисперсія ознаки X при $Y = y_i$:

$$D\left(X|_{Y=y_i}\right)=\frac{1}{n_{y_i}}\sum_{j=1}^{m}x_j^2n_{ij}-\left(\bar{x}|_{Y=y_i}\right)^2.$$

Умовне середнє квадратичне відхилення X при $Y=y_i$:

$$\sigma\left(X|_{Y=y_i}\right)=\sqrt{D\left(X|_{Y=y_i}\right)}\,.$$

Зауваження. При відомих значеннях умовних середніх $\bar{y}|_{X=x_j}$ та $\bar{x}|_{Y=y_i}$ загальні середні ознак X та Y можна обчислити за формулами:

$$\bar{y}=\frac{1}{n}\sum_{j=1}^{m}\left(\bar{y}|_{X=x_j}\right)n_{x_j}$$

та

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{k}\left(X|_{Y=y_i}\right)n_{y_i}.$$

Вибірковий коефіцієнт кореляції

Під час дослідження двовимірного статистичного розподілу вибірки постає потреба з'ясувати наявність зв'язку між ознаками X та Y, який називають ***кореляційним***. Кореляційний зв'язок вимірюється за допомогою **вибіркового коефіцієнта кореляції**

$$r=\frac{K_{xy}}{\sigma_x\cdot\sigma_y}\,.$$

Число K_{xy} у цьому виразі називається **вибірковим коефіцієнтом коваріації** і обчислюється за формулою

$$K_{xy}=\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{m}x_j\cdot y_i\cdot n_{ij}\ -\bar{x}\cdot\bar{y}\,.$$

Коефіцієнт кореляції дає змогу встановити тісноту зв'язку. Він лежить в межах $-1\le r\le 1$. Якщо $r=0$, то зв'язок між ознаками відсутній. Якщо r близький до 1, то зв'язок тісний прямий. Якщо r близький до -1, то зв'язок тісний обернений.

1.5. Емпіричні моменти

Повернемося до однієї ознаки X з таблицею дискретного статистичного розподілу

x_1	x_2	x_3	...	x_k
n_1	n_2	n_3	...	n_k

де

$$\sum_{i=1}^{k} n_i = n.$$

Нехай $m \in N$.

Означення. Початковим емпіричним моментом m-го порядку називається число

$$\nu_m = \frac{1}{n}\sum_{i=1}^{k} x_i^m n_i.$$

Зауваження.

1. При $m = 1$ дістаємо середнє:

$$\nu_1 = \frac{1}{n}\sum_{i=1}^{k} x_i n_i = \bar{x}.$$

2. Дисперсію вибірки можемо записати через початкові емпіричні моменти 1-го і 2-го порядку:

$$D = \frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i - \left(\frac{1}{n}\sum_{i=1}^{k} x_i n_i\right)^2 = \nu_2 - \nu_1^2.$$

Означення. Центральним емпіричним моментом m-го порядку називається число

$$\mu_m = \frac{1}{n}\sum_{i=1}^{k} (x_i - \bar{x})^m \cdot n_i.$$

Зауваження.

1. При $m = 1$ дістаємо

$$\mu_1 = \frac{1}{n}\sum_{i=1}^{k} (x_i - \bar{x}) \cdot n_i = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot n_i - \frac{1}{n} \cdot \bar{x} \cdot \sum_{i=1}^{k} n_i = \bar{x} - \bar{x} = 0.$$

2. При $m = 2$ дістаємо

$$\mu_2 = \frac{1}{n}\sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot n_i = D.$$

Окрім емпіричних моментів 1-го та 2-го порядків на практиці також застосовуються емпіричні моменти 3-го та 4-го порядків. Вирази для центральних емпіричних моментів 3-го та 4-го порядків мають вигляд

$$\mu_3 = \frac{1}{n}\sum_{i=1}^{k} (x_i - \bar{x})^3 \cdot n_i = \cdots = \nu_3 - 3\nu_2\nu_1 + 2(\nu_1)^3;$$

$$\mu_4 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^4 \cdot n_i = \cdots = \nu_4 - 4\nu_3\nu_1 + 6\nu_2(\nu_1)^2 - 3(\nu_1)^4.$$

На практиці використовуються дві важливі числові характеристики вибірки, які обчислюються за допомогою центральних емпіричних моментів 3-го і 4-го порядків.

Коефіцієнт асиметрії

$$k_a = \frac{\mu_3}{\sigma^3}.$$

Коли $k_a < 0$ $(k_a > 0)$ то розподіл X має від'ємну (додатну) асиметрію. Коефіцієнт асиметрії k_a є безрозмірною величиною. Його використовують для оцінки як напрямку, так і сили асиметрії розподілу ознаки X.

Асиметрія називається ***додатною*** $(k_a > 0)$, якщо кількість варіант $x_i < \bar{x}$ статистичного розподілу переважає кількість варіант $x_i > \bar{x}$.

Асиметрія називається ***від'ємною*** $(k_a < 0)$, якщо кількість варіант $x_i > \bar{x}$ статистичного розподілу переважає кількість варіант $x_i < \bar{x}$.

Коефіцієнт ексцесу

$$k_e = \frac{\mu_4}{\sigma^4} - 3.$$

Коефіцієнт ексцесу (або, коротше, *ексцес*) k_e є безрозмірною величиною. Він характеризує "гостровершинність" розподілу ознаки X у порівнянні з нормальним розподілом. k_e зазвичай використовується при дослідженні неперервних ознак генеральних сукупностей, оскільки він оцінює крутизну щільності розподілу неперервної випадкової величини порівняно з нормальним. Для нормального розподілу ексцес $k_e = 0$; при $k_e > 0$ розподіл гостровершинний, при $k_e < 0$ туповершинний.

Відмітимо, що число 3 у формулі ексцесу віднімається ось чому. Для нормального закону розподілу виконується рівність:

$$\frac{\mu_4}{\sigma^4} = 3.$$

Отже, $k_e = 0$.

1.6. Приклади до Розділу 1

Приклад 1. Дискретний розподіл

Виробник мобільних телефонів проводить дослідження, щоб з'ясувати, як часто абоненти змінюють свої телефони. Ознака X – кількість мобільних телефонів, які змінила людина за останні 10 років. Дані згруповані у вигляді дискретного розподілу.

x_i	1	2	3	4	5	6	7
n_i	120	323	781	1220	665	234	11

1. Побудувати полігон частот і відносних частот;
2. Побудувати емпіричну функцію розподілу;
3. Знайти числові характеристики розподілу.

Розв'язання. Тут

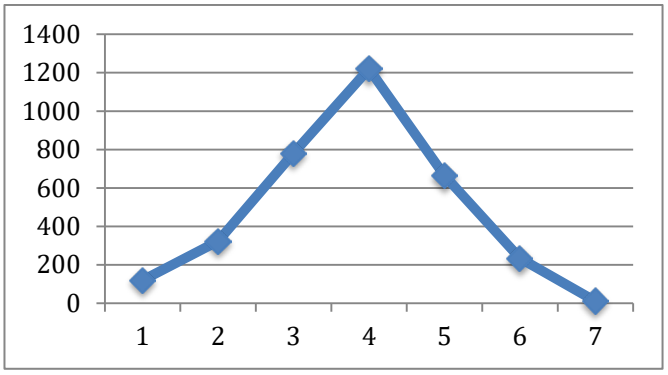
$$n = \sum_{i=1}^{7} n_i = 120 + 323 + 781 + 1220 + 665 + 234 + 11 = 3354.$$

Доповнимо спочатку нашу таблицю значеннями відносних частот, поділивши кожну частоту n_i на $n = 3354$.

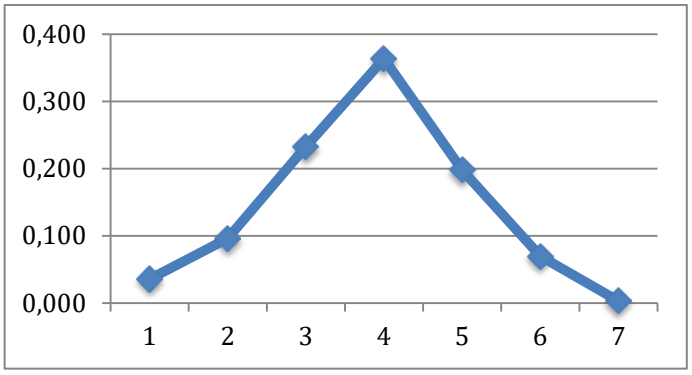
x_i	1	2	3	4	5	6	7
n_i	120	323	781	1220	665	234	11
w_i	0,036	0,096	0,233	0,364	0,198	0,070	0,003

Полігон частот і відносних частот матимуть однаковий вигляд, але відрізнятимуться шкалою. Для полігона частот відкладаємо по осі ординат значення частот n_i, а для полігона відносних частот значення відносних частот w_i.

Полігон частот



Полігон відносних частот

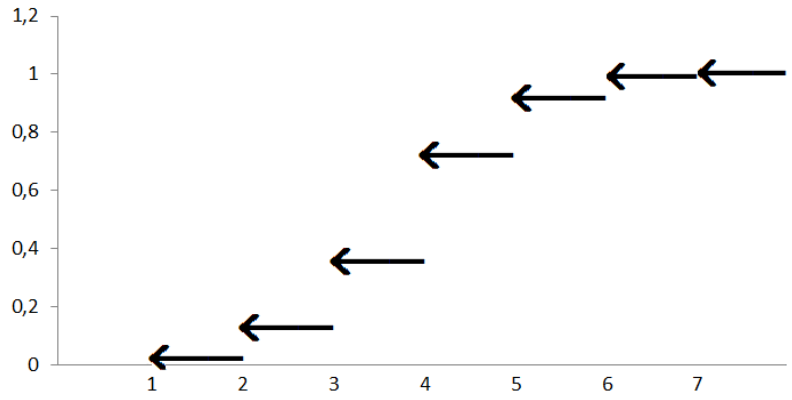


Емпірична функція розподілу.

Побудуємо емпіричну функцію розподілу:

$$F_n(x) = \begin{cases} 0, & x \le 1; \\ 0{,}036, & 1 < x \le 2; \\ 0{,}132, & 2 < x \le 3; \\ 0{,}365, & 3 < x \le 4; \\ 0{,}729, & 4 < x \le 5; \\ 0{,}927, & 5 < x \le 6; \\ 0{,}997, & 6 < x \le 7; \\ 1, & x > 7 \end{cases}$$

Візуалізуємо цю функцію:



Мода.

Мода для даного розподілу дорівнює 4, бо для цього спостереження ми маємо найбільшу частоту появи: 1220.

Медіана.

Для знаходження медіани нам потрібно зрозуміти, яка варіанта перебуває на 1677 місці у впорядкованому ряді (3354/2=1677).

Якщо ми додамо перші три частоти, отримаєм 120+323+781=1224. Далі якщо додамо до цієї суми ще 1220 (цій частоті відповідає варіанта 4), то отримаємо 2444. Відповідно, варіанта з номером 1677 точно знаходиться у діапазоні від 1224 до 2444 і дорівнює 4. Отже, медіана розподілу дорівнює 4.

Середня величина:

$$\bar{x} = \frac{1 \cdot 120 + 2 \cdot 323 + 3 \cdot 781 + 4 \cdot 1220 + 5 \cdot 665 + 6 \cdot 234 + 7 \cdot 11}{3354} = 3{,}8.$$

Дисперсія:

$$D = \frac{1^2 \cdot 120 + 2^2 \cdot 323 + 3^2 \cdot 781 + 4^2 \cdot 1220 + 5^2 \cdot 665 + 6^2 \cdot 234 + 7^2 \cdot 11}{3354} - 3{,}8^2 = 1{,}41.$$

Середнє квадратичне відхилення:

$$\sigma = \sqrt{D} = 1{,}19.$$

Асиметрія.

Підрахуємо емпіричний момент 3-го порядку:

$$\mu_3 = \frac{(1-3{,}8)^3 \cdot 120 + (2-3{,}8)^3 \cdot 323 + (3-3{,}8)^3 \cdot 781 + \cdots + (7-3{,}8)^3 \cdot 11}{3354} = -0{,}33.$$

Тепер можемо знайти саму асиметрію

$$k_a = \frac{-0{,}33}{1{,}19^3} = -0{,}2.$$

Отримали від'ємну лівосторонню асиметрію.

Ексцес.

Підрахуємо емпіричний момент 4-го порядку:

$$\mu_4 = \frac{(1-3{,}8)^4 \cdot 120 + (2-3{,}8)^4 \cdot 323 + (3-3{,}8)^4 \cdot 781 + \cdots + (7-3{,}8)^4 \cdot 11}{3354} = 5{,}71.$$

Далі можемо знайти ексцес

$$k_e = \frac{5{,}71}{1{,}19^4} - 3 = -3{,}17.$$

Ексцес від'ємний, тобто розподіл туповершинний.

Приклад 2. Інтервальний розподіл

За згрупованим рядом середньої місячної температури повітря (січень, м. Одеса):

Інтервали	[-9,4; -7,6]	(-7,6; -5,8]	(-5,8; -4,0]	(-4,0; -2,2]	(-2,2; -0,4]	(-0,4; 1,4]	(-1,4; 3,2]
Частоти	4	2	3	7	6	11	2

1. Побудувати гістограму частот і відносних частот;
2. Побудувати емпіричну функцію розподілу;
3. Полічити числові характеристики розподілу.

Розв'язання. Тут

$$\sum_{i=1}^{7} n_i = 4 + 2 + 3 + 7 + 6 + 11 + 2 = 35.$$

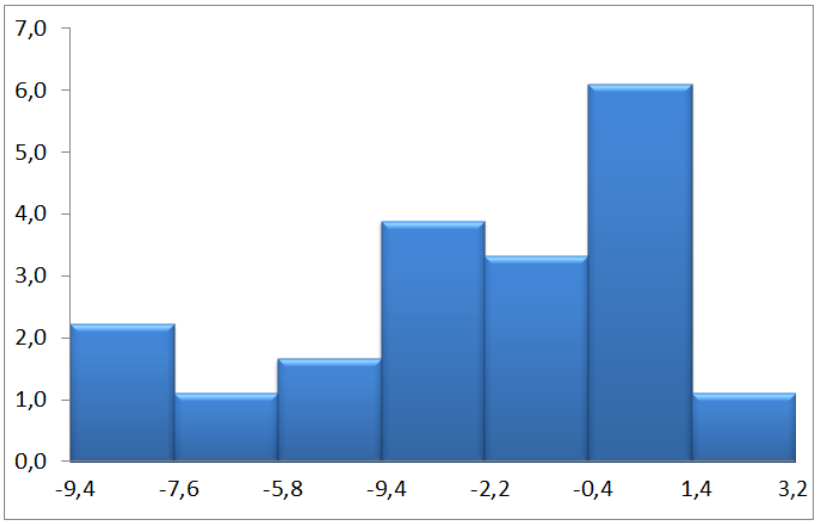
Величина $h = -7{,}6 - (-9{,}4) = 1{,}8.$

Для подальших розрахунків полічимо відносні частоти і середини інтервалів.

x_i	[-9,4; -7,6]	(-7,6; -5,8]	(-5,8; -4,0]	(-4,0; -2,2]	(-2,2; -0,4]	(-0,4; 1,4]	(-1,4; 3,2]
x_i^*	-8,5	-6,7	-4,9	-8,5	-1,3	0,5	3,2
n_i	4	2	3	7	6	11	2
w_i	$\frac{4}{35}$	$\frac{2}{35}$	$\frac{3}{35}$	$\frac{7}{35}$	$\frac{6}{35}$	$\frac{11}{35}$	$\frac{2}{35}$

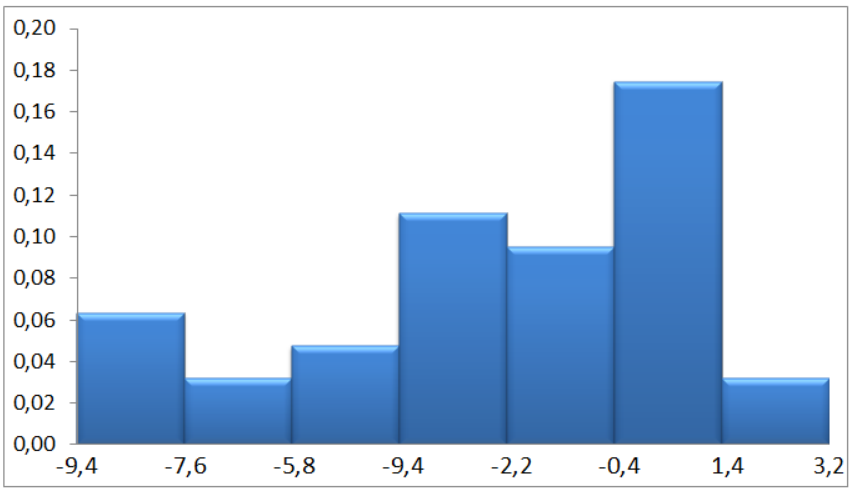
Гістограма частот.

Гістограму частот отримуємо обчисливши значення $\frac{n_i}{h}$ на кожному інтервалі.



Гістограма відносних частот.

Гістограму відносних частот отримуємо знайшовши значення $\frac{w_i}{h}$ на кожному інтервалі.

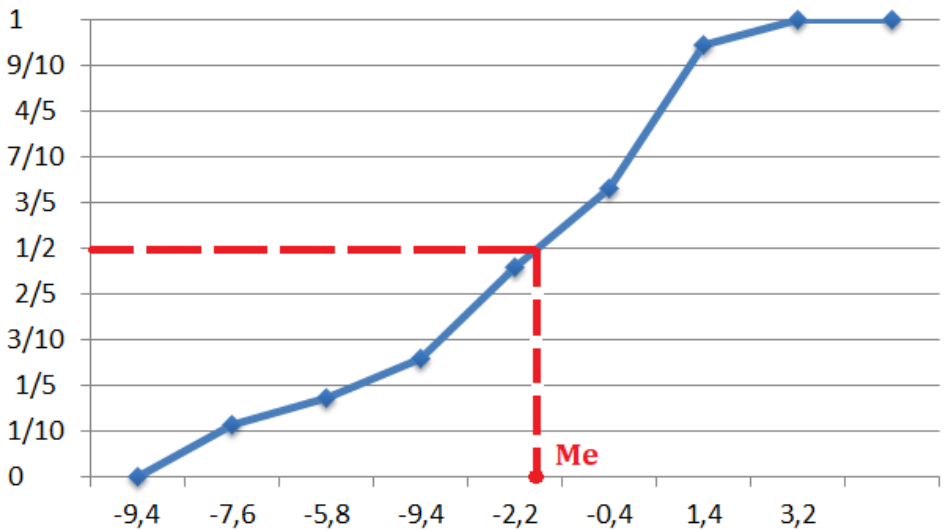


Емпірична функція розподілу.

Побудуємо емпіричну функцію розподілу:

$$F_n(x) = \begin{cases} 0, & x \le -9{,}4; \\ \frac{4}{35}, & -9{,}4 < x \le -7{,}6; \\ \frac{6}{35}, & -7{,}6 < x \le -5{,}8; \\ \frac{9}{35}, & -5{,}8 < x \le -4{,}0; \\ \frac{16}{35}, & -4{,}0 < x \le -2{,}2; \\ \frac{22}{35}, & -2{,}2 < x \le -0{,}4; \\ \frac{33}{35}, & -0{,}4 < x \le 1{,}4; \\ 1, & 1{,}4 < x \le 3{,}2. \end{cases}$$

Візуалізуємо функцію розподілу і відразу проаналізуємо до якого інтервалу потрапляє його медіана:



Медіанний інтервал (-2,2; -0,4].

Тепер знайдемо числові характеристики розподілу:

Середня величина:

$$\bar{x} = \frac{(-8{,}5)\cdot 4 + (-6{,}7)\cdot 2 + (-4{,}9)\cdot 3 + (-8{,}5)\cdot 7 + (-1{,}3)\cdot 6 + 0{,}5\cdot 11 + 3{,}2\cdot 2}{35} = -3{,}41.$$

Дисперсія:

$$D = \frac{(-8{,}5)^2\cdot 4 + (-6{,}7)^2\cdot 2 + (-4{,}9)^2\cdot 3 + (-8{,}5)^2\cdot 7 + (-1{,}3)^2\cdot 6 + 0{,}5^2\cdot 11 + 3{,}2^2\cdot 2}{35} - (-3{,}41)^2 = 16{,}38.$$

Середнє квадратичне відхилення:

$$\sigma = \sqrt{D} = 4{,}05.$$

Медіана.

Ми визначили медіанний інтервал (-2,2; -0,4]. Нагадаємо формулу для обчислення медіани:

$$Me = x_{i-1} + \frac{0{,}5 - F_n(x_{i-1})}{F_n(x_i) - F_n(x_{i-1})} h.$$

Тут $x_{i-1} = -2{,}2; \quad x_i = -0{,}4; \quad F_n(x_{i-1}) = 16/53; \quad F_n(x_i) = 22/35.$
Відповідно

$$Me = -2{,}2 + \frac{0{,}5 - \frac{16}{53}}{\frac{22}{35} - \frac{16}{53}} \cdot 1{,}8 = -1{,}75.$$

Мода.

Спочатку визначимо модальний інтервал. Це інтервал, до якого потрапляє найбільша кількість спостережень. Отже, це – інтервал (-0,4; 1,4], куди потрапляє 11 спостережень. Нагадаємо формулу моди:

$$Mo = x_{i-1} + \frac{n_i - n_{i-1}}{2n_i - n_{i-1} - n_{i+1}} h.$$

Тут $x_{i-1} = -0{,}4; \quad n_i = 11; \quad n_{i-1} = 6$ – частота на домодальному інтервалі; $n_{i+1} = 2$ – частота на післямодальному інтервалі.
Відповідно

$$Mo = -0{,}4 + \frac{11 - 6}{2 \cdot 11 - 6 - 2} \cdot 1{,}8 = 0{,}24.$$

Асиметрія.

Знайдемо центральний емпіричний момент 3-го порядку:

$$\mu_3 = \frac{(-8{,}5 - (-3{,}41))^3 \cdot 4 + (-6{,}7 - (-3{,}41))^3 \cdot 2 + \cdots + (3{,}2 - (-3{,}41))^3 \cdot 2}{35} = 26{,}8.$$

Тепер можемо знайти сам коефіцієнт асиметрії

$$k_a = \frac{26{,}8}{4{,}05^3} = 0{,}4.$$

Отримали додатній коефіцієнт асиметрії.

Ексцес.

Знайдемо центральний емпіричний момент 4-го порядку:

$$\mu_4 = \frac{(-8{,}5 - (-3{,}41))^4 \cdot 4 + (-6{,}7 - (-3{,}41))^4 \cdot 2 + \cdots + (3{,}2 - (-3{,}41))^4 \cdot 2}{35}$$
$$= 409{,}7.$$

Далі можемо підрахувати коефіцієнт ексцесу

$$k_e = \frac{409{,}7}{4{,}05^4} - 3 = -1{,}2.$$

Ексцес від'ємний, тобто розподіл туповершинний.

Приклад 3. Двовимірний розподіл

Залежність урожайності ячменю y_i від кількості внесених добрив на 1 га x_i наведено у вигляді двовимірного статистичного розподілу вибірки:

$Y = y_i$, ц/га	$X = x_i$, га/т				
	0,5	1	1,5	2	2,5
15,5	1	2	—	—	—
16,5	2	4	1	—	—
17,5	—	3	6	1	—
18,5	—	—	4	1	1
19,5	—	—	1	2	1

Обчислити коефіцієнт кореляції r, а також порахувати числові характеристики для умовних розподілів $Y|_{X=1,5}$ та $X|_{Y=16,5}$.

Розв'язання.

Спочатку підрахуємо стовпчикові і рядкові суми частот.

$Y = y_i$, ц/га	$X = x_i$, га/т					
	0,5	1	1,5	2	2,5	$\boldsymbol{n_{y_i}}$
15,5	1	2	0	0	0	**3**
16,5	2	4	1	0	0	**7**
17,5	0	3	6	1	0	**10**
18,5	0	0	4	1	1	**6**
19,5	0	0	1	2	1	**4**
$\boldsymbol{n_{x_j}}$	**3**	**9**	**12**	**4**	**2**	$\boldsymbol{n = 30}$

Числові характеристики ознаки X**.**

Загальне середнє ознаки X:

$$\bar{x} = \frac{0{,}5 \cdot 3 + 1 \cdot 9 + 1{,}5 \cdot 12 + 2 \cdot 4 + 2{,}5 \cdot 2}{30} = 1{,}38.$$

Дисперсія ознаки X:

$$D_X = \frac{0{,}5^2 \cdot 3 + 1^2 \cdot 9 + 1{,}5^2 \cdot 12 + 2^2 \cdot 4 + 2{,}5^2 \cdot 2}{30} - 1{,}38^2 = 0{,}26.$$

Середнє квадратичне відхилення ознаки X:

$$\sigma_X = \sqrt{D_X} = 0{,}51.$$

Числові характеристики ознаки Y.

Загальне середнє ознаки Y:

$$\bar{y} = \frac{15{,}5 \cdot 3 + 16{,}5 \cdot 7 + 17{,}5 \cdot 10 + 18{,}5 \cdot 6 + 19{,}5 \cdot 4}{30} = 17{,}53.$$

Дисперсія ознаки Y:

$$D_Y = \frac{15{,}5^2 \cdot 3 + 16{,}5^2 \cdot 7 + 17{,}5^2 \cdot 10 + 18{,}5^2 \cdot 6 + 19{,}5^2 \cdot 4}{30} - 17{,}53^2 = 1{,}37.$$

Середнє квадратичне відхилення ознаки Y:

$$\sigma_Y = \sqrt{D_Y} = 1{,}17.$$

Коефіцієнт коваріації:

$$K_{xy} = \frac{1}{30}[15{,}5 \cdot 0{,}5 \cdot 1 + 15{,}5 \cdot 1 \cdot 2 + 16{,}5 \cdot 0{,}5 \cdot 2 + 16{,}5 \cdot 1 \cdot 4 + 17{,}5 \cdot 1 \cdot 3 + 17{,}5 \cdot 1{,}5 \cdot 6 + 17{,}5 \cdot 2 \cdot 1 + 18{,}5 \cdot 1{,}5 \cdot 4 + 18{,}5 \cdot 2 \cdot 1 + 18{,}5 \cdot 2{,}5 \cdot 1 + 19{,}5 \cdot 1{,}5 \cdot 1 + 19{,}5 \cdot 2 \cdot 2 + 19{,}5 \cdot 2{,}5 \cdot 1] - 1{,}38 \cdot 17{,}53 = 0{,}45.$$

Коефіцієнт кореляції:

$$r = \frac{K_{xy}}{\sigma_x \cdot \sigma_y} = \frac{17{,}53}{0{,}51 \cdot 1{,}17} = 0{,}76.$$

Отже, існує залежність урожайності ячменю y_i від кількості внесених добрив на 1 га x_i. Зв'язок прямий. Чим більше внесено добрив, тим більша урожайність. Коефіцієнт кореляції дорівнює 0,76.

Умовний розподіл $Y|_{X=1,5}$.

Утворимо дискретний (умовний) розподіл із двовимірної таблиці при $X = 1{,}5$.

y_i	16,5	17,5	18,5	19,5
n_{i3}	1	6	4	1

Умовне середнє ознаки Y при $X = 1{,}5$:

$$\bar{y}|_{X=1,5} = \frac{16{,}5 \cdot 1 + 17{,}5 \cdot 6 + 18{,}5 \cdot 4 + 19{,}5 \cdot 1}{12} = 17{,}92.$$

Умовна дисперсія ознаки Y при $X = 1{,}5$:

$$D\left(Y|_{X=1,5}\right) = \frac{16{,}5^2 \cdot 1 + 17{,}5^2 \cdot 6 + 18{,}5^2 \cdot 4 + 19{,}5^2 \cdot 1}{12} - 17{,}92^2 = 0{,}58.$$

Умовне середнє квадратичне відхилення Y при $X = 1{,}5$:

$$\sigma\left(Y|_{X=1,5}\right) = 0{,}76.$$

Умовний розподіл $X|_{Y=16,5}$.

Утворимо дискретний (умовний) розподіл із двовимірної таблиці при $Y = 16{,}5$.

x_j	0,5	1	1,5
n_{2j}	2	4	1

Умовне середнє ознаки X при $Y = 16{,}5$:

$$\bar{x}|_{Y=16,5} = \frac{0{,}5 \cdot 2 + 1 \cdot 4 + 1{,}5 \cdot 1}{7} = 0{,}93.$$

Умовна дисперсія ознаки X при $Y = 16{,}5$;

$$D\left(X|_{Y=16,5}\right) = \frac{0{,}5^2 \cdot 2 + 1^2 \cdot 4 + 1{,}5^2 \cdot 1}{7} - 0{,}93^2 = 0{,}1.$$

Умовне середнє квадратичне відхилення Y при $Y = 16{,}5$:

$$\sigma\left(X|_{Y=16,5}\right) = 0{,}32.$$

1.7. Питання для самоконтролю до Розділу 1

1. Дати визначення генеральної сукупності.
2. Що називається варіантою, варіаційним рядом?
3. Що таке частота, відносна частота варіант?
4. Дати визначення дискретного статистичного розподілу вибірки.
5. Середнє, дисперсія та середнє квадратичне відхилення для дискретного розподілу вибірки.
6. Що таке медіана, мода дискретного статистичного розподілу?
7. Що називається емпіричною функцією розподілу?
8. Властивості емпіричної функції розподілу.
9. Що називається інтервальним статистичним розподілом вибірки?
10. Середнє, дисперсія та середнє квадратичне відхилення для інтервального розподілу.
11. Як визначається медіана для інтервального статистичного розподілу?
12. Як визначається мода для інтервального статистичного розподілу?
13. Що являє собою полігон частот і відносних частот?
14. Що називається гістограмою частот і відносних частот?
15. Асиметрія і ексцес статистичного розподілу вибірки.

16. Дати означення емпіричної функції розподілу для інтервального статистичного розподілу вибірки.
17. Що називається двовимірним статистичним розподілом вибірки?
18. Формули для обчислення основних числових характеристик ознак X і Y для двовимірного статистичного розподілу вибірки.
19. Вибірковий коефіцієнт кореляції та його властивості.
20. Що називається умовним статистичним розподілом?
21. Умовні числові характеристики для умовного статистичного розподілу.

РОЗДІЛ 2. СТАТИСТИЧНІ ОЦІНКИ ПАРАМЕТРІВ ГЕНЕРАЛЬНОЇ СУКУПНОСТІ

2.1. Загальна інформація

Нагадаємо, що ми розглядаємо генеральну сукупність Ω і певну числову ознаку X цієї сукупності (наприклад вага $X(\omega)$ яйця ω). Значення самої ознаки спочатку невідомі. Можна взяти якесь яйце ω_0 і зважити його. Тоді знатимемо $X(\omega_0)$. Але, яке це яйце? Вони ж не пронумеровані. Нас і не цікавить вага одного конкретного яйця. Нас цікавлять певні параметри генеральної сукупності. Наприклад: середня вага яєць, яка частка серед них важить менше 50 гр., яка більше, і т. п.

Одним із ключових завдань математичної статистики є визначення методів знаходження (точніше, оцінювання) параметрів генеральної сукупності.

Такі методи називаються ***статистичними оцінюваннями***, або ***статистиками***.

Наприклад, для оцінювання середньої ваги $\bar{X}$ партії яєць Ω, яка складається з N яєць, можна запропонувати наступний природний метод. За означенням, ця середня вага дорівнює

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X(\omega_i).$$

Для її оцінювання випадково вибирають яйця $\omega_1, \omega_2, \dots, \omega_n$ $(n \ll N)$, зважують і обчислюють

$$\bar{x}(n) = \frac{1}{n}\sum_{i=1}^{n} X(\omega_i). \qquad (2.1)$$

Число $\bar{x}(n)$ можна вважати наближеним значенням середньої ваги $\bar{X}$.

Кожен метод потребує **обґрунтування**. Чому $\bar{x}(n)$ приблизно дорівнює $\bar{X}$? І чому беручи все більше і більше n ми точніше й точніше оцінюватимемо це середнє значення $\bar{X}$? Такі обґрунтування проводяться в рамках **теорії ймовірностей**.

А саме, розглядаємо числову ознаку $X(\omega)$ як деяку **випадкову величину**. На значення ознаки X для вибірки $A = \{\omega_1, \omega_2, \dots, \omega_n\}$ дивимось як на послідовність незалежних випадкових величин $X_1, \dots, X_n$ з тим же розподілом, що й X. Далі утворюємо нову випадкову величину

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i. \qquad (2.2)$$

Закон великих чисел (ЗВЧ) з теорії ймовірностей стверджує, що при певних природних умовах на X послідовність $\bar{X}_n$ збігається за ймовірністю до математичного сподівання $\mathbf{M}X$ цієї випадкової величини X [6, п.1.21.1].

А в наведеному прикладі математичне сподівання $\mathbf{M}X$ – це і є середнє значення $\bar{X}$.

Отже, ЗВЧ (2.2) є обґрунтуванням для методу (2.1).

Зауваження. Нагадаємо, що в теорії ймовірностей сама випадкова величина X невідома, а відома, як правило, її функція розподілу

$$F(x) = \mathbf{P}\{X(\omega) < x\}, \quad x \in \mathbf{R}.$$

Далі, знаючи функцію розподілу, можна знаходити різні числові параметри випадкової величини X: математичне сподівання, дисперсію, медіану тощо. Це в теорії ймовірностей. А в математичній статистиці функція розподілу $F(x)$ теж невідома. Все що ми можемо: це взяти вибірку $A = \{\omega_1, \omega_2, \dots, \omega_n\}$ і знайти $(X(\omega_1), X(\omega_2), \dots, X(\omega_n))$. Далі, на підставі цих значень, можемо якось оцінити певний параметр ознаки X.

Повернімось до загальної ситуації.

Означення. Нехай задано певний (невідомий) параметр θ ознаки (тобто, випадкової величини) X. Кожен метод $\theta^*(n)$ його оцінки, $n \in \mathbf{N}$, називається ***статистикою*** або ***статистичним оцінюванням цього*** параметр θ.

Зауваження 2.1.

1. Параметр θ – це невідоме стале число.

2. Метод (статистика) – це послідовність нових випадкових величин $\theta^*(n)$, що залежить від n та X.

3. Оцінка параметра θ – це число $\theta(n)$, отримане після застосування методу.

Для статистики природно ввести два наступні поняття.

Означення. Статистична оцінка $\theta^*(n)$ параметра θ називається ***незміщеною***, якщо для кожного n математичне сподівання

$$\mathbf{M}\theta^*(n) = \theta.$$

Незміщеність статистики гарантує, що при її застосуванні не з'являється систематична помилка: оцінка не занижується або не завищується систематично.

Твердження 2.1. Статистика (2.2) незміщена.

Доведення (тут використано властивості математичного сподівання [9, с. 130]).

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

Означення. Статистика $\theta^*(n)$, називається ***слушною*** (або ***конзистентною***), якщо $\theta^*(n) \to \theta$ за ймовірністю при $n \to \infty$.

Твердження 2.2. Статистика (2.2) слушна.

Доведення. Воно випливає із ЗВЧ [6, п.1.21.1].

2.2. Оцінки з мінімальною дисперсією. Нерівність Крамера-Рао

Нехай θ – деякий параметр ознаки X, $\theta^* = \theta^*(n)$ – певний метод його оцінювання (статистика). Нагадаємо, що на ознаку X ми дивимося як на випадкову величину; тоді (невідомий!) параметр θ є деяким параметром цієї випадкової величини, наприклад, математичним сподіванням. Нагадаємо, як будується статистика. Береться n незалежних копій $X_1, \dots, X_n$ випадкової величини X і статистика $\theta^* = \theta_n^*$ – це якась функція від цих копій:

$$\theta^* = h(X_1, \dots, X_n) = h_n(X_1, \dots, X_n).$$

Тобто, статистика будується для кожного n. Наприклад, для оцінювання математичного сподівання $\mathbf{M}X$, використовувалася статистика

$$\theta_n^* = \frac{1}{n}\sum_{i=1}^n X_i = h_n(X_1, \dots, X_n).$$

Статистик може бути багато. Яка з них краща? Та, при застосуванні якої припускається найменша похибка. Як вимірювати похибку?

Однією з найприродніших мір похибки є величина

$$\mathbf{M}(\theta^* - \theta)^2.$$

Як вже було відзначено, природною вимогою до оцінки θ^* є її незміщеність, тобто умова $\mathbf{M}\theta^* = \theta$, бо при $\mathbf{M}\theta^* \neq \theta$ для великих n похибка буде напевно більшою, ніж в оцінок, для яких $\mathbf{M}\theta^* = \theta$. Зауважимо, що, коли оцінка θ^* незміщена, то згадана величина перетворюється в дисперсію:

$$\mathbf{M}(\theta^* - \theta)^2 = \mathbf{M}(\theta^* - \mathbf{M}\theta)^2 = \mathbf{D}\theta^*.$$

Незміщених оцінок параметра θ існує багато. Серед них, для оцінювання θ природно вибирати ті, які мають малу похибку (тобто дисперсію), а в ідеалі – мінімально можливу. Тут відразу постає питання: наскільки малими можуть бути дисперсії оцінок $\theta^*(n)$ параметра θ?

Позначимо через $F(x_1, \dots, x_n, \theta)$ – функцію розподілу випадкового вектора $(X_1, \dots, X_n)$, який залежить від θ:

$$F(x_1, \dots, x_n, \theta) = \mathbf{P}\{X_1 < x_1, \dots, X_n < x_n\}, \quad x_1, \dots, x_n \in \mathbf{R}.$$

Наприклад, якщо $X = (X_1, X_2)$ – двовимірний нормальний розподіл з однаково розподіленими незалежними компонентами з невідомим параметром $\theta = a$ (математичне сподівання) і відомим середнім квадратичним відхиленням $\sigma = 1$, то його функція розподілу має вигляд

$$F(x_1, x_2, \theta) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} \frac{1}{2\pi}\exp\left\{-\frac{(x_1-\theta)^2 + (x_2-\theta)^2}{2}\right\} dx_1 dx_2\,.$$

Так ось, виявляється, що коли розподіл $F(x_1, \dots, x_n, \theta)$ вибірки $(X_1, \dots, X_n)$ досить регулярно залежить від параметра θ, то можна вказати нижню межу дисперсій усіх незміщених оцінок параметра θ, тобто величину, менше від якої дисперсія бути не може: більшою точність оцінювання бути не може.

Пояснимо спочатку, що ми розуміємо під виразом “досить регулярно”, звернувшись до математичного аналізу.

Диференційованість функції за параметром під знаком інтеграла

Щоб уникнути громіздких формул, розглянемо функцію $f(\mathrm{x},\theta)$, де $\mathrm{x}=(x_1,\dots,x_n)$ – вектор, $\theta\in[a,b]$, а частинна похідна цієї функції скрізь братиметься за параметром θ.

Твердження 2.3. (Наслідок теореми Лебега про мажоровану збіжність) [6, I.1.11.8]. Нехай $A\subset\mathbf{R}^n$ – борелівська множина. Якщо для майже всіх $\mathrm{x}\in A$ функція $f(\mathrm{x},\theta)$ диференційована за параметром θ, а похідна $f'(\mathrm{x},\theta)$ мажорується інтегрованою функцією $y(\mathrm{x})$, не залежною від $\theta\in[a,b]$, тобто

$$|f'(\mathrm{x},\theta)|\le y(\mathrm{x}),\ \text{де}\ \int_A y(\mathrm{x})d\mathrm{x}<\infty,$$

то

$$\left(\int_A f(\mathrm{x},\theta)d\mathrm{x}\right)'=\int_A f'(\mathrm{x},\theta)d\mathrm{x}. \qquad (2.3)$$

Доведення. Справді, для всіх $\mathrm{x}\in A$ і $\theta,t\in\mathbf{R}$, за теоремою про середнє значення,

$$\left|\frac{f(\mathrm{x},\theta+t)-f(\mathrm{x},\theta)}{t}\right|=|f'(\mathrm{x},\tau)|_{\tau\epsilon[\theta,\theta+t]}\le y(\mathrm{x}).$$

Тому, внаслідок теореми Лебега про мажоровану збіжність:

$$\left[f(\mathrm{x},\theta)\xrightarrow[\theta\to\theta_0]{}f(\mathrm{x},\theta_0)\Rightarrow\int_A f(\mathrm{x},\theta)dx\longrightarrow\int_A f(\mathrm{x},\theta_0)d\mathrm{x}\right]$$

отримаємо

$$\left|\frac{\int_A f(\mathrm{x},\theta+t)d\mathrm{x}-\int_A f(\mathrm{x},\theta)dx}{t}\right| = \int_A\frac{f(\mathrm{x},\theta+t)-f(\mathrm{x},\theta)}{t}d\mathrm{x}$$

$$\downarrow \qquad (2.3)! \qquad \downarrow$$

$$\left(\int_A f(\mathrm{x},\theta)d\mathrm{x}\right)' = \int_A f'(\mathrm{x},\theta)d\mathrm{x}. \quad \square$$

Означення. Кажуть, що випадковий вектор $\mathrm{X}=(X_1,\dots,X_n)$ ***абсолютно неперервний***, якщо існує така неперервна функція $f(\mathrm{x})$, $\mathrm{x}\in\mathbf{R}^n$, яка називається його ***щільністю***, що для кожної борелівської множини $A\subset\mathbf{R}^n$

$$\mathbf{P}\{\mathrm{X}\in A\}=\int_A f(\mathrm{x})d\mathrm{x}.$$

Теорема 2.1. (про обчислення математичного сподівання) [6, п.1.11.1]. Нехай випадковий вектор X має щільність $f(\mathrm{x})$, $\mathrm{x}\in\mathbf{R}^n$. Тоді для кожної борелівської функції $y(\mathrm{x})$, $\mathrm{x}\in\mathbf{R}^n$,

$$\mathbf{M}y(\mathrm{X})=\int_{\mathbf{R}^n}y(\mathrm{x})f(\mathrm{x})d\mathrm{x}.$$

Умова регулярності. Далі ми припускатимемо, що випадковий вектор X має щільність $f(\mathrm{x},\theta)$, залежну від параметра $\theta\in[a,b]$ і диференційовану за цим параметром.

Означення. Покладемо

$$g(\mathrm{x},\theta)=\frac{f'(\mathrm{x},\theta)}{f(\mathrm{x},\theta)}=[\ln f(\mathrm{x},\theta)]'.$$

Функція $I(\theta)=I(\theta,\mathrm{x})=\mathbf{D}g(\mathrm{x},\theta)$ називається ***кількістю інформації за Фішером*** (про параметр θ на підставі випадкового вектора X).

Далі ми припускатимемо, що похідні f' та g' мажоруються інтегрованими функціями, зокрема, що для $\forall\theta\in[a,b]$

$$\mathbf{M}g(\mathrm{X},\theta)<\infty. \tag{2.4}$$

Лема 2.1. Якщо виконана умова регулярності, то

$$\mathbf{M}g(\mathrm{X},\theta)=0.$$

Доведення.

$$\mathbf{M}g(\mathrm{X},\theta)=\int_{\mathbf{R}^n}\frac{f'(\mathrm{x},\theta)}{f(\mathrm{x},\theta)}f(\mathrm{x},\theta)d\mathrm{x}=[\text{Тв. 2.3}]=\int_{\mathbf{R}^n}f(\mathrm{x},\theta)d\mathrm{x}=0. \quad \square$$

Наслідок 2.1. Якщо

$$\mathbf{M}g^2(\mathrm{X},\theta)<\infty\,, \tag{2.5}$$

то

$$I(\theta,\mathrm{X})=\mathbf{D}g(\mathrm{X},\theta)=\mathbf{M}g^2(\mathrm{X},\theta).$$

Доведення. Справді, за означенням дисперсії,

$$\mathbf{D}g(\mathrm{X},\theta)=\mathbf{M}g^2(\mathrm{X},\theta)-[\mathbf{M}g(\mathrm{X},\theta)]^2=\mathbf{M}g^2(\mathrm{X},\theta). \quad \square$$

Теорема 2.2 (нерівність Крамера-Рао). Нехай $\mathrm{X}=(X_1,\dots,X_n)$ випадковий вектор із щільністю $f(\mathrm{x},\theta)$, $\mathrm{x}\in\mathbf{R}^n$, який задовольняє умову (2.5), а $\theta^*=h(\mathrm{X})$ – така незміщена оцінка параметра θ, що $[h(\mathrm{x})\cdot f(\mathrm{x},\theta)]'$ мажорується інтегрованою функцією. Тоді

$$\mathbf{D}\theta^*\geq\frac{1}{I(\theta)},$$

причому рівність у цій нерівності досягається тоді і тільки тоді, коли для деякої функції $c(\theta)$:

$$g(\mathrm{x},\theta)=c(\theta)(h(\mathrm{x})-\theta),\ \forall\mathrm{x}\in\mathbf{R}^n. \tag{2.6}$$

Доведення. Оскільки оцінка $\theta^*=h(\mathrm{X})$ незміщена, то

$$\theta=\mathbf{M}\theta^*=\mathbf{M}h(\mathrm{X})=(\text{теорема 2.1})=\int_{\mathbf{R}^n}h(\mathrm{x})f(\mathrm{x},\theta)d\mathrm{x}.$$

Диференціюючи цю рівність, отримаємо

$$\begin{aligned}
1 &= \left(\int_{\mathbf{R}^n}h(\mathrm{x})f(\mathrm{x},\theta)d\mathrm{x}\right)' = \int_{\mathrm{R}^n}h(\mathrm{x})f'(\mathrm{x},\theta)d\mathrm{x} = \\
&= \int_{\mathbf{R}^n}h(\mathrm{x})\underbrace{\frac{f'(\mathrm{x},\theta)}{f(\mathrm{x},\theta)}}_{=g(\mathrm{x},\theta)}f(\mathrm{x},\theta)d\mathrm{x} = \int_{\mathbf{R}^n}h(\mathrm{x})g(\mathrm{x},\theta)\cdot f(\mathrm{x},\theta)d\mathrm{x} = \\
&= \mathbf{M}h(\mathrm{X})\cdot g(X,\theta) = \boxed{\mathbf{M}\theta^*\cdot g(X,\theta)}.
\end{aligned} \tag{2.7}$$

Далі

$$\boxed{0} = (\text{Лема } 1) = \theta \cdot \mathbf{M} g(\mathrm{X}, \theta) = \boxed{\mathbf{M}\theta g(\mathrm{X}, \theta)}. \qquad (2.8)$$

Віднімаючи від (2.7) рівність (2.8), дістаємо

$$1 = \mathbf{M}(\theta^* - \theta) g(X, \theta).$$

Піднесемо цю рівність до квадрату і скористаємось нерівністю Коші-Буняковського [6, I.1.13.3]:

$$1 = [\mathbf{M}(\theta^* - \theta) g(\mathrm{X}, \theta)]^2 \le \mathbf{M}(\theta^* - \theta)^2 \cdot \mathbf{M} g^2(\mathrm{X}, \theta) = [\text{наслідок } 2.1] = \qquad (2.9)$$

$$\mathbf{D}\theta^* \cdot \mathbf{D} g(\mathrm{X}, \theta) = \mathbf{D}\theta^* \cdot I(\theta).$$

Це й дає нерівність Крамера-Рао.

Врешті зауважимо, що нерівність Коші-Буняковського перетворюється в рівність тоді й тільки тоді, коли

$$g(\mathrm{X}, \theta) = c(\theta)(\theta^* - \theta) = c(\theta)(h(\mathrm{X}) - \theta). \qquad \square \qquad (2.10)$$

Означення. Незміщену оцінку $\hat{\theta}$ параметра θ називають ***ефективною***, або ***найкращою незміщеною оцінкою***, якщо

$$\mathbf{D}\hat{\theta} = \inf\{\mathbf{D}\theta^* : \mathbf{M}\theta^* = \theta\}.$$

Наслідок 2.2. Якщо для незміщеної оцінки $\hat{\theta}$ параметра θ нерівність Крамера-Рао обертається в рівність, то вона буде ефективною.

Наслідок 2.3. Нехай $f(\mathrm{x}, \theta)$, $\mathrm{x} \in \mathbf{R}^n$ – щільність випадкового вектора X і $h(\mathrm{x})$ – статистика, для якої виконані умови теореми 2.2. Якщо має місце умова (2.6), то $\theta^* = h(\mathrm{X})$ буде ефективною оцінкою параметра θ.

Нагадаємо, що за означенням

$$I(\theta) = I(\theta, \mathrm{X}) = \mathbf{D} g(\mathrm{X}, \theta).$$

Наведемо іншу формулу для інформації $I(\theta)$, яка часом буває зручнішою для використання. Для доведення цієї формули буде потрібна наступна лема.

Лема 2.2. Якщо похідна $f''(x, \theta)$ існує і мажорується інтегрованою функцією, то

$$\mathbf{M} g'(\mathrm{X}, \theta) = -\mathbf{M} g^2(\mathrm{X}, \theta).$$

Доведення.

$$\mathbf{M} g'(\mathrm{X}, \theta) = [\text{Теор. } 2.1] = \int_{\mathbf{R}^n} g'(\mathrm{x}, \theta) \cdot f(\mathrm{x}, \theta) d\mathrm{x} =$$

$$= \int_{\mathbf{R}^n} \left[\frac{f'(\mathrm{x}, \theta)}{f(\mathrm{x}, \theta)}\right] \cdot f(\mathrm{x}, \theta) dx = \int_{\mathbf{R}^n} \frac{f''(\mathrm{x}, \theta) \cdot f(\mathrm{x}, \theta) - [f'(\mathrm{x}, \theta)]^2}{f^2(\mathrm{x}, \theta)} \cdot f(\mathrm{x}, \theta) d\mathrm{x} =$$

$$= \underbrace{\boxed{\int_{\mathbf{R}^n} f''(\mathrm{x}, \theta) d\mathrm{x}}}_{=0} - \int_{\mathbf{R}^n} \left[\frac{f'(\mathrm{x}, \theta)}{f(\mathrm{x}, \theta)}\right]^2 \cdot f(\mathrm{x}, \theta) d\mathrm{x} = -\int_{\mathbf{R}^n} g^2(\mathrm{x}, \theta) d\mathrm{x} = -\mathbf{M} g^2(\mathrm{X}, \theta). \square$$

Наслідок 2.4. В умовах Леми 2.2:

$$I(\theta) = \mathbf{D}g(\mathrm{X}, \theta) = -\mathbf{M}g'(\mathrm{X}, \theta).$$

Доведення. Справді,

$$\mathbf{D}g(\mathrm{X}, \theta) = [\text{Насл. 2.1}] = \mathbf{M}g^2(\mathrm{X}, \theta) = [\text{Лема 2.2}] = -\mathbf{M}g'(\mathrm{X}, \theta). \quad \square$$

А тепер пригадаємо, що при оцінюванні невідомого параметра θ, компоненти $X_1, \dots, X_n$ випадкового вектора є незалежними і однаково розподіленими випадковими величинами. Для абсолютно неперервного випадкового вектора це означає, що його щільність розподілу має вигляд

$$f(\mathrm{x}, \theta) = f(x_1, \dots, x_n, \theta) = \prod_{i=1}^{n} \varphi(x_i, \theta), \tag{2.11}$$

де $\mathrm{x} = (x_1, \dots, x_n) \in \mathbf{R}^n$, а $\varphi(t, \theta)$ – щільність розподілу компонент $X_1, \dots, X_n$ [6, п.1.19.1].

Зауважимо відразу, що тоді

$$g(\mathrm{x}, \theta) = \left[\ln \prod_{i=1}^{n} \varphi(x_i, \theta)\right]' = \left(\sum_{i=1}^{n} \ln \varphi(x_i, \theta)\right)' = \sum_{i=1}^{n} [\ln \varphi(x_i, \theta)]',$$

а

$$\mathbf{D}g(\mathrm{X}, \theta) = \mathbf{D}\sum_{i=1}^{n} [\ln \varphi(x_i, \theta)]' = \sum_{i=1}^{n} \mathbf{D}[\ln \varphi(x_i, \theta)]' = n\mathbf{D}[\ln \varphi(X_1, \theta)]'.$$

Наслідок 2.5. Якщо щільність розподілу випадкового вектора $\mathrm{X} = (X_1, \dots, X_n)$ має вигляд (2.11), то

$$\mathbf{D}\theta^* \geq \frac{1}{nD(\ln \varphi(x_i, \theta))}. \tag{2.12}$$

Зауваження 2.1. З наслідку Леми 2.2 випливає, що цю нерівність можна записати у вигляді

$$\mathbf{D}\theta^* \geq -\frac{1}{n\mathbf{M}[\ln \varphi(x_i, \theta)]'}.$$

Наслідок 2.6. Кожна ефективна оцінка є слушною .

Доведення. Справді, оскільки для ефективної оцінки нерівність (2.12) перетворюється у рівність, то

$$\mathbf{D}\theta^*(n) = \frac{1}{n\mathbf{D}(\ln \varphi(X, \theta))} \underset{n\to\infty}{\longrightarrow} 0.$$

Звідси випливає, що при $n \to \infty$

$$\mathbf{M}(\theta^*(n) - \theta)^2 = \mathbf{M}(\theta^*(n) - \mathbf{M}\theta) = \mathbf{D}\theta^*(n) \to 0.$$

З теорії ймовірностей відомо, що коли для послідовності випадкових величин Z_n, послідовність математичних сподівань $\mathbf{M}{Z_n}^2 \to 0$ при $n \to \infty$, то сама ця послідовність Z_n збігається до нуля за ймовірністю [10, с. 111]. Отже

$$\theta^*(n) - \theta \to 0 \text{ при } n \to \infty \text{ за ймовірністю.}$$

А звідси відразу випливає, що

$$\theta^*(n) \to \theta \text{ при } n \to \infty \text{ за ймовірністю. } \square$$

Теорема-приклад. Нехай $\mathrm{X} = (X_1, \dots, X_n)$ – випадковий вектор, компоненти якого незалежні і мають нормальний розподіл з параметрами $(\theta, 1)$, $\theta \in [a, b]$. Тоді величина

$$\theta^* = \frac{1}{n}\sum_{i=1}^{n} X_i$$

буде ефективною оцінкою математичного сподівання θ.

Доведення. Як ми вже показували, $\theta^* = \bar{X}$ буде незміщеною оцінкою математичного сподівання. Скористаємося означенням ефективності. Потрібно перевірити рівність

$$\mathbf{D}\theta^* = \frac{1}{I(\theta)}.$$

Але спочатку перевіримо виконання умов самої теореми. Пригадаємо, що щільність розподілу нормальної випадкової величини з параметрами $(\theta, 1)$ має вигляд

$$\varphi(t, \theta) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(t-\theta)^2}{2}\right\}. \qquad (2.13)$$

Тоді похідні

$$\varphi'_\theta(t, \theta) = \left[\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(t-\theta)^2}{2}\right\}\right]' = (t-\theta)\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(t-\theta)^2}{2}\right\} =$$

$$= (t-\theta)\varphi(t, \theta). \qquad (2.14)$$

$$[\ln \varphi(t, \theta)]' = \left[\ln \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(t-\theta)^2}{2}\right\}\right]' = t - \theta. \qquad (2.15)$$

Перевіримо існування величин, які фігурують в нерівності Крамера-Рао і обчислимо їх:

1. $g(x, \theta) = [\ln f(x, \theta)]' = [\ln \prod_{i=1}^{n} \varphi(x_i, \theta)]' = \sum_{i=1}^{n} [\ln \varphi(x_i, \theta)]' = \sum_{i=1}^{n} (x_i - \theta)$.
2. $g'(x, \theta) = (\sum_{i=1}^{n} x_i - \theta)' = -n$.
3. $\mathbf{M}\big(g'(\mathrm{X}, \theta)\big) = n$.
4. $\mathbf{M}g^2(\mathrm{X}, \theta) = -\mathbf{M}g'(\mathrm{X}, \theta) = n$.

Врешті, $\forall \mathrm{x} \in \mathbf{R}^n$

$$[h(\mathrm{x}) \cdot f(\mathrm{x}, \theta)]' = \bar{x}\left(\prod_{i=1}^{n} \varphi(x_i, \theta)\right)' = \bar{x}\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x_i-\theta)^2}{2}\right\}\right)' =$$

$$= \bar{x}\frac{1}{(2\pi)^{n/2}}\left[\exp\left\{-\frac{\sum_{i=1}^{n}(x_i-\theta)^2}{2}\right\}\right]' =$$

$$= \bar{x}\frac{1}{(2\pi)^{n/2}}\exp\left\{-\frac{\sum_{i=1}^{n}(x_i-\theta)^2}{2}\right\}\cdot\sum_{i=1}^{n}(x_i-\theta).$$

Чому така функція мажорується інтегровною функцією, незалежною від θ? Тут потрібно скористатися тим, що $\theta \in [a,b]$, а (для $n=1$) функція $t^k e^{-t^2}$ інтегровна [6, п.3.7.4]. Деталі перевірки лишаємо читачеві для самостійного опрацювання.

Обчислимо тепер величини, які фігурують в нерівності Крамера-Рао:

$$\mathbf{D}\theta^* = \mathbf{D}\frac{1}{n}\sum_{i=1}^{n}X_i = \frac{1}{n^2}\sum_{i=1}^{n}\mathbf{D}X_i = \frac{1}{n^2}\cdot n\mathbf{D}X_1 = \frac{1}{n},$$

$$I(\theta) = \mathbf{M}g^2(\mathrm{X},\theta) = (2.14) = n.$$

Отже,

$$\mathbf{D}\theta^* = \frac{1}{I(\theta)}$$

і оцінка

$$\theta^* = \frac{1}{n}\sum_{i=1}^{n}X_i$$

є ефективною. □

Нерівність Крамера–Рао (дискретний розподіл)

Нерівність Крамера-Рао має місце і для дискретного розподілу випадкових векторів. Нагадаємо, що випадковий вектор $\mathrm{X} = (X_1, \dots, X_n)$, заданий на ймовірнісному просторі, має **дискретний розподіл**, якщо $\mathbf{P}(\mathrm{x}) := \mathbf{P}_{\mathrm{X}}(\mathrm{x}) = \mathbf{P}\{\mathrm{X}=\mathrm{x}\} > 0$ лише на елементах певної не більше ніж зліченої множини A з $\mathbf{R}^n$ і, звичайно,

$$\sum\nolimits_{\mathrm{x}\in A}\mathbf{P}(\mathrm{x}) = 1.$$

Ймовірність $\mathbf{P}$ тоді називатимемо ***дискретним розподілом випадкового вектора X***.

Нехай розподіл $\mathbf{P}$ залежить від випадкового параметра θ: $\mathbf{P} = \mathbf{P}(\mathrm{x},\theta)$. На цей розподіл накладаємо наступні умови:

- похідні (за змінною θ) $\mathbf{P}'(\mathrm{x},\theta)$ та $\mathbf{P}''(\mathrm{x},\theta)$ існують, (2.16)
- ряди $\sum_{\mathrm{x}\in A}\boldsymbol{P}'(\mathrm{x},\theta)$ та $\sum_{\mathrm{x}\in A}\mathbf{P}'(\mathrm{x},\theta)$ збігаються абсолютно і рівномірно по θ, (2.17)
- $\mathbf{M}([\ln\mathbf{P}(\mathrm{X},\theta)]')^2 < \infty$; $\mathbf{M}([\ln\mathbf{P}(\mathrm{X},\theta)]'')^2 < \infty$. (2.18)

Теорема 2.3. (нерівність Крамера–Рао для дискретного випадку). Нехай $\mathrm{X} = (X_1, \dots, X_n)$ – дискретний вектор із розподілом $\mathbf{P}(\mathrm{x}, \theta)$, $\mathrm{x} \in \mathbf{R}^n$, для якого виконуються умови (2.16) – (2.18) і нехай $\theta^* = h(\mathrm{X})$ – така незміщена оцінка параметра θ, що ряд

$$\sum_{\mathrm{x} \in A} [h(\mathrm{x})\mathbf{P}(\mathrm{x}, \theta)]' \tag{2.19}$$

збігається абсолютно й рівномірно по θ. Тоді

$$\mathbf{D}\theta^* \geq \frac{1}{I(\theta)},$$

причому рівність у цій нерівності досягається тоді і тільки тоді, коли $\forall \theta \in [a, b]$

$$[\ln \mathbf{P}(\mathrm{X}, \theta)]' = c(\theta) \cdot (h(\mathrm{x}) - \theta).$$

Доведення цієї теореми схоже на доведення для абсолютно неперервного випадку і ми його лишаємо читачеві для самостійного опрацювання.

Наслідок 2.7. Нехай $X_1, \dots, X_n$ – незалежні випадкові величини з одним і тим самим дискретним розподілом $\mathbf{P}(t, \theta)$, $t \in \mathbf{R}$. Нехай виконані умови (2.16) – (2.19). Тоді

$$\mathbf{D}\theta^* \geq \frac{1}{n\mathbf{M}([\ln \mathbf{P}(X_1, \theta)])^2}.$$

2.3. Методи визначення невідомих параметрів

Метод аналогій

Він полягає в тому, що для оцінки невідомих параметрів вибирають ті ж формули, що і для обчислення відповідних параметрів вибірки.

Приклади.

1. Математичне сподівання. Пригадаймо, що маючи генеральну сукупність Ω і її числову ознаку $X(\omega)$, для оцінки середнього значення $\bar{X}$ бралася вибірка $A = \{\omega_1, \omega_2, \dots, \omega_n\}$ і знаходилося вибіркове середнє

$$\bar{x} = \bar{x}(n) = \frac{1}{n}\sum_{i=1}^{n} X(\omega_i).$$

Тому, за методом аналогій, для оцінки невідомого математичного сподівання $\mathbf{M}X$ випадкової величини X розглядаємо n її незалежних копій $X_1, \dots, X_n$ і за оцінку $\mathbf{M}X$ беремо випадкову величину

$$\bar{X}(n) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

2. Дисперсія. Окрім середнього значення $\bar{X}$ числової ознаки $X(\omega)$ часто цікавим є відхилення значень $X(\omega)$ від цього середнього. Наприклад, якщо $\bar{X}$ – це середня вага партії яєць, то може бути цікавим відхилення ваги яєць від середнього значення. Одні яйця більші, інші – менші. Зрозуміло, якщо у партії яйця дуже різної ваги – це погано. Адже їх продають на

штуки. Кількісно таке відхилення можна оцінювати різними способами. Наприклад, якщо задано вибірку $X(\omega_1), \dots, X(\omega_n)$ і $\bar{X} = \bar{X}(n)$ – вибіркове середнє значення, то відхилення можна оцінювати, наприклад, формулами

$$\max_{1\le i\le n} |X(\omega_i) - \bar{X}|; \qquad \frac{1}{n}\sum_{i=1}^{n} |X(\omega_i) - \bar{X}|; \qquad \sqrt{\frac{1}{n}\sum_{i=1}^{n} (X(\omega_i) - \bar{X})^2}.$$

Першу оцінку природно називати ***максимальним відхиленням***, другу – ***середнім відхиленням***, третю – ***середнім квадратичним відхиленням***. Вибір міри відхилення залежить від мети оцінювання та зручності математичного оперування з даною мірою. Ми зупинимося на середньому квадратичному відхиленні. Чому саме на ньому – обґрунтуємо дещо пізніше. Точніше, ми зупинимось не так на середньому квадратичному відхиленні, як на його квадратові

$$\frac{1}{n}\sum_{i=1}^{n} (X(\omega_i) - \bar{X})^2,$$

який називають ***вибірковою дисперсією***. Зокрема, від такої функції легко рахувати похідну. Тоді, за методом аналогій, відповідна статистика для оцінки дисперсії випадкової величини має вигляд

$$D_n = D_n(X) := \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X}_n)^2 ,$$

де $(X_i)_{i=1}^n$ незалежні копії випадкової величини X.

Природно поставити питання: чи буде така статистика незміщеною і слушною? Для відповіді на перше питання, порахуємо математичне сподівання випадкової величини D_n. Для цього покладемо для $\forall i = 1, \dots, n,\ Y_i = X_i - \mathbf{M}X$. Тоді, $\forall i$

$$\mathbf{M}Y_i = \boldsymbol{M}X_i - \mathbf{M}X = 0.$$

$$\mathbf{M}Y_i^2 = \mathbf{M}(X_i - \mathbf{M}X)^2 = \mathbf{D}X.$$

$$\mathbf{D}Y_i = \mathbf{D}(X_i - \mathbf{M}X) = \mathbf{D}X.$$

$$Y_i - \bar{Y}_n = (X_i - \mathbf{M}X) - \frac{1}{n}\sum_{i=1}^{n} (X_i - \mathbf{M}X) = X_i - \mathbf{M}X - \frac{1}{n}\sum_{i=1}^{n} X_i + \mathbf{M}X = X_i - \bar{X}_n.$$

$$\forall i \ne j:\ \mathbf{M}\big(Y_i \cdot Y_j\big) = \mathbf{M}Y_i \cdot \mathbf{M}Y_j = 0.$$

Врешті

$$\mathbf{M}\bar{Y}_n^2 = \frac{1}{n^2}\mathbf{M}\left(\sum_{i=1}^{n} Y_i\right)^2 = \frac{1}{n^2}\mathbf{M}\left(\sum_{i=1}^{n} Y_i^2 + 2\sum_{i<j}^{n} Y_i \cdot Y_j\right) =$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} \mathbf{M}Y_i^2 = \frac{1}{n^2} \cdot n \cdot \mathbf{D}X = \frac{\mathbf{D}X}{n}.$$

Тепер

$$\boxed{\mathbf{M}D_n} = \frac{1}{n}\mathbf{M}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n}\mathbf{M}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}[\mathbf{M}Y_i^2 - 2\mathbf{M}Y_i \cdot \bar{Y}_n] + \frac{1}{n}\mathbf{M}(n\bar{Y}_n^2) =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbf{M}Y_i^2 - 2\mathbf{M}\bar{Y}_n^2 + \mathbf{M}\bar{Y}_n^2 = \frac{1}{n}\cdot n \cdot \mathbf{D}X - \mathbf{M}(\bar{Y}_n)^2 = \mathbf{D}X - \frac{\mathbf{D}X}{n} = \boxed{\frac{n-1}{n}\mathbf{D}X}.$$

Висновок. Статистика D_n зміщена.

Тому, замість D_n часто використовується скоригована статистика

$$\frac{n}{n-1}D_n,$$

яка вже буде незміщеною.

Твердження. Статистики D_n та $\frac{n}{n-1}D_n$ слушні.

Доведення. Оскільки випадкові величини X_i незалежні, то й X_i^2 теж незалежні, $i = 1, \dots, n$. Тому

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i)^2 - 2\frac{1}{n}\sum_{i=1}^{n}X_i \cdot \bar{X}_n + \frac{1}{n}\sum_{i=1}^{n}(\bar{X}_n)^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i)^2 - (\bar{X}_n)^2 \underset{n\to\infty}{\longrightarrow} \mathbf{M}X^2 - (\mathbf{M}X)^2 = \mathbf{D}X.$$

Звідси випливає також збіжність $\frac{n}{n-1}D_n \underset{n\to\infty}{\longrightarrow} \mathbf{D}X$ за ймовірністю. □

Тепер, за оцінку невідомого середнього квадратичного відхилення $\sigma(X)$ природно взяти випадкову величину $\hat{s}_n = \sqrt{\frac{n}{n-1}D_n}$. Звідси, звичайно, випливає, що скоригована статистика $\frac{n}{n-1}D_n$ дорівнює $\hat{s}_n{}^2$. Так ось, надалі цю скориговану статистику для дисперсії надалі позначатимемо через $\hat{s}_n{}^2$. Це виглядить дещо незвично, але така традиція.

Зауваження щодо позначень і термінів. Тут і далі зустрічаються три види величин.

1. Фіксовані (але в математичній статистиці невідомі) характеристики однієї, або кількох, випадкових величин. Як правило, це – числа: наприклад, математичне сподівання $\mathbf{M}X$ випадкової величини X, її дисперсія $\mathbf{D}X$, її середнє квадратичне відхилення σ_X, коваріація $\mathrm{Cov}(X, Y)$ двох випадкових величин X та Y. Але це можуть бути й функції; наприклад, функція розподілу $F(x)$.

2. Статистики, тобто методи оцінювання цих характеристик. Наприклад, для оцінювання невідомого математичного сподівання $\mathbf{M}X$ метод оцінювання полягає в тому, що для кожного n розглядається набір незалежних копій $(X_i)_{i=1}^n$ випадкової величини X і

знаходимо середнє $\bar{X} = \bar{X}(n) = \frac{1}{n}\sum_{i=1}^{n} X_i$. Знайдене середнє і є статистикою для математичного сподівання $\mathbf{M}X$. Отже статистика – це випадкова величина.

3. Реалізації статистик, тобто методів оцінювання. Наприклад, для реалізації описаного методу оцінки невідомого математичного сподівання проводимо n незалежних випробовувань (вимірювань, експериментів) випадкової величини X. Діставши числові результати $(x_i)_{i=1}^{n}$, обчислюємо середнє $\bar{x} = \bar{x}(n) = \frac{1}{n}\sum_{i=1}^{n} x_i$. Воно й буде оцінкою невідомого математичного сподівання. Отже реалізації статистики – це число.

Отже, вводячи позначення, важливо не путати ці три види величин і намагатися позначати їх однотипно. Проте, з одного боку, деякі позначення вже стали традиційними в математичній статистиці, а з іншого – багато авторів до позначення багатьох величин вживають різних позначень (а інколи вживають і різні терміни). Особливо часто у літературі одним і тим самим символом позначають статистику та її реалізацію. Ми намагатимемося дотримуватися означень та позначень книжки [6]; проте не завжди. Переважно, для підкреслень того, що маємо справу зі статистикою, або її реалізацією, часто у індексі ставимо кількість вимірювань n.

Метод максимальної вірогідності

Попередній метод застосовується у випадку, коли про розподіл випадкової величини X (чи ознаки X) нічого не відомо. Проте часто відомим є ***характер розподілу*** випадкової величини X.

Приклад 2.1. Нехай випадкова величина X має двійковий розподіл. Наприклад, розглянемо добре відоме з теорії ймовірностей кидання несиметричної монети. Тут випадкова величина X набуває двох значень 1 – «успіх» з імовірністю p і 0 – «невдача» з імовірністю $1-p$. У математичній статистиці ймовірність p вважається невідомою і її потрібно оцінити. Для цієї оцінки проводять n випробувань (кидають монету). Нехай отримано k успіхів.

Метод максимальної вірогідності полягає в тому, що невідому ймовірність p (чи невідомий параметр p) вибирають так, щоб імовірність k успіхів у n випробуваннях була найімовірнішою. Наприклад, якщо випало два успіхи, то $p = 1$; дві невдачі – $p = 0$; успіх + невдача – $p = ½$. Розглянемо цю оцінку формальніше. Ймовірність k успіхів у n випробуваннях дорівнює

$$C_n^k p^k (1-p)^{n-k}.$$

Для знаходження максимуму цієї функції додатна константа C_n^k значення не має. Тому потрібно знайти таке p, для якого функція $p^k(1-p)^{n-k}$ досягає найбільшого значення. Отримали просту задачу математичного аналізу. Перейдімо до логарифмів

$$k \cdot \ln p + (n-k)\ln(1-p).$$

Знайдемо її похідну й прирівняємо до нуля:

$$\frac{k}{p} - \frac{n-k}{1-p} = 0.$$

Звідси

$$k(1-p) = p(n-k) \Rightarrow k - kp = np - pk \Rightarrow p = \frac{k}{n}.$$

Отриманий результат узгоджується з інтуїцією.

Покажемо, що ця оцінка незміщена, тобто перевіримо, що $\mathbf{M}X = p$. Нехай $(X_i)_1^n$ – незалежні копії випадкової величини X. Тоді $\sum_{i=1}^n X_i$ – це і є кількість успіхів у n випробуваннях, тобто k. Тому

$$\mathbf{M}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \mathbf{P}\{X = 1\} = p.$$

Так само, за ЗВЧ [6, п.1.22.1]

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{} \mathbf{M}X,$$

тобто оцінка слушна.

Приклад 2.2. Розглянемо тепер неперервну випадкову величину, наприклад вагу яйця. У математичній статистиці особливо поширеним є припущення, що випадкова величина X має **нормальний** (або **Гаусівський**) розподіл. Аргументом на користь цього припущення є Центральна гранична теорема (ЦГТ) з курсу теорії ймовірностей [6, п.1.28].

Пригадаймо, що для нормального розподілу щільність ймовірності має вигляд

$$f(x; a, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-a)^2}{2\sigma^2}}. \qquad (2.20)$$

У математичній статистиці значення параметрів a і σ^2 часто вважаються невідомими і їх потрібно знайти. Тут параметр a має сенс математичного сподівання, а σ^2 – дисперсії. Таким чином, у цьому випадку щільність f повністю визначається двома параметрами $\theta_1 = a$ і $\theta_2 = \sigma^2$.

Отже, нехай випадкова величина X має нормальний розподіл з параметрами θ_1 та θ_2. І нехай $X_1, \dots, X_n$ – незалежні копії випадкової величини X.

Тоді для вектора $(X_1, \dots, X_n)$ з теорії ймовірностей відомо, що його щільність ймовірностей має вигляд [6, п.1.9.5]:

$$f(x_1, \dots, x_n; \theta_1, \theta_2) = \frac{1}{\left[\sqrt{2\pi\theta_2}\right]^n} \cdot e^{-\frac{\sum_{i=1}^n (x_i-\theta_1)^2}{2\theta_2}}. \qquad (2.21)$$

Нехай проведено n незалежних випробувань, в результаті яких отримано значення $x_1^0, x_2^0, \dots, x_n^0$. У методі максимальної вірогідності за статистичні оцінки параметрів θ_1, θ_2 вибирають ті їх значення θ_1^*, θ_2^*, за яких задана вибірка буде найімовірнішою, тобто функція

$$f(x_1, \dots, x_n; \theta_1^*, \theta_2^*)$$

досягає максиму в точці $(x_1^0, x_2^0, \dots, x_n^0)$.

На практиці, від функції (2.17) зручно перейти до її логарифма, тобто до функції

$$L(x_1, \dots, x_n, \theta_1, \theta_2) = \ln f(x_1, \dots, x_n, \theta_1, \theta_2) = -\frac{n}{2}(\ln \pi + \ln \theta_2) - \frac{\sum_{i=1}^{n}(x_i - \theta_1)^2}{2\theta_2}.$$

Згідно з необхідною умовою екстремуму:

$$\begin{cases} L'_{\theta_1} = \dfrac{1}{\theta_2}\displaystyle\sum_{i=1}^{n}(x_i^0 - \theta_1) = 0, \\ L'_{\theta_2} = -\dfrac{n}{2\theta_2} + \dfrac{1}{2\theta_2^2} \cdot \displaystyle\sum_{i=1}^{n}\left(x_i^0 - \theta_1\right)^2 = 0 \end{cases}.$$

Із цієї системи рівнянь отримаємо значення θ_1 і θ_2:

$$\theta_1 = \frac{1}{n}\sum_{i=1}^{n} x_i^0 = \bar{x};$$

$$\theta_2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^0 - \bar{x}\right)^2 = D.$$

Отже, параметр θ_1 визначається середнім вибірковим значенням, а параметр θ_2 – вибірковою дисперсією.

2.4. Класичні розподіли математичної статистики

При інтервальних статистичних оцінках параметрів випадкових величин та перевірці статистичних гіпотез, які будуть розглянуті в наступних підрозділах, використовуються розподіли хі-квадрат, Стьюдента і Фішера-Снедекора. Опишемо їх.

Означення. Розподілом хі-квадрат (або **χ^2-розподілом**, або **розподілом Пірсона**) з n степенями свободи називатимемо розподіл випадкової величини

$$\chi_n^2 = \sum_{i=1}^{n} X_i^2,$$

де $X_1, \dots, X_n$ – незалежні стандартні нормальні випадкові величини.

Означення. Розподілом Стьюдента, або **t-розподілом** з n степенями свободи називатимемо розподіл випадкової величини

$$t_n = \frac{X}{\sqrt{\chi_n^2/n}},$$

де випадкові величини X і χ_n^2 незалежні, X має стандартний нормальний розподіл, а χ_n^2 має розподіл хі-квадрат з n степенями свободи.

Означення. Розподілом Фішера або **F-розподілом** з (n, m) степенями свободи, називатимемо розподіл випадкової величини

$$F_{n,m} = \frac{\chi_n^2 /n}{\chi_m^2 /m},$$

де χ_n^2 та χ_m^2 – незалежні випадкові величини, що мають розподіли хі-квадрат з n та m степенями свободи відповідно.

Теорема 2.4. Послідовність $\frac{\chi_n^2 - n}{\sqrt{2n}}$ збігається за розподілом до стандартної нормальної випадкової величини, точніше, якщо $\varphi(x)$ – щільність розподілу стандартної нормальної випадкової величини, то $\forall x \in \mathbf{R}$

$$\mathbf{P}\left\{\frac{\chi_n^2 - n}{\sqrt{2n}} < x\right\} \underset{n\to\infty}{\longrightarrow} \int_{-\infty}^{x} \varphi(t)dt.$$

Доведення. Внаслідок ЦГТ [6, п.1.28], $\forall x \in \mathbf{R}$

$$\mathbf{P}\left\{\frac{\chi_n^2 - \mathbf{M}\chi_n^2}{\sigma(\chi_n^2)} < x\right\} \underset{n\to\infty}{\longrightarrow} \int_{-\infty}^{\infty} \varphi(t)dt.$$

Лишилося згадати, що для стандартної нормальної випадкової величини X $\mathbf{M}X^2 = 1$, а $\mathbf{D}X^2 = 2$ [6, I.1.9.5]. Тоді

$$\mathbf{M}\chi_n^2 = \mathbf{M}\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} \mathbf{M}X_i^2 = n,$$

$$\boldsymbol{\sigma}(\chi_n^2) = \sqrt{\mathbf{D}\chi_n^2} = \sqrt{\mathbf{D}\sum_{i=1}^{n} X_i^2} = \sqrt{\sum_{i=1}^{n} \mathbf{D}X_i^2} = \sqrt{2n}. \ \square$$

Теорема 2.5. Послідовність t_n збігається за ймовірністю до стандартної нормальної випадкової величини.

Доведення. Внаслідок ЗВЧ [6, п.1.21.1],

$$\frac{1}{n}\chi_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \underset{n\to\infty}{\longrightarrow} \mathbf{M}X_1^2 = 1 \quad \text{за ймовірністю.}$$

Отже,

$$t_n = \frac{X}{\sqrt{\chi_n^2/n}} \underset{n\to\infty}{\longrightarrow} X \quad \text{за ймовірністю.}$$

Таким чином, має місце (3.11). □

Щільності й графіки

Розподіл хі-квадрат. Його щільність імовірностей має вигляд

$$f_n(x) = \begin{cases} 0, & x < 0 \\ A_n x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, & x \geq 0, \end{cases}$$

де стала A_n визначається з умови нормування

$$A_n \int_0^{\infty} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} dx = 1.$$

Наприклад, для $n = 4$, графік функції $f_4(x)$ зображений на рис.2.1.

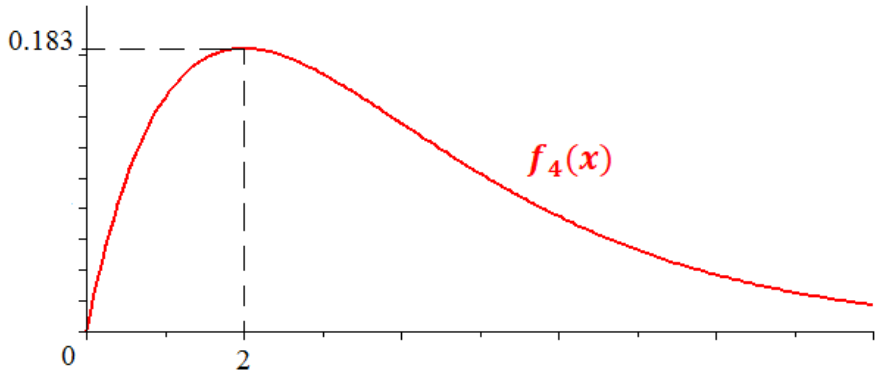


Рис. 2.1

Розподіл Стьюдента. Його щільність імовірностей має вигляд

$$f_n(x) = B_n \left(1 + \frac{x^2}{n}\right)^{\frac{n+1}{2}},$$

де стала B_n визначається з умови нормування

$$B_n \int_0^{\infty} \left(1 + \frac{x^2}{n}\right)^{\frac{n+1}{2}} dx = 1.$$

Наприклад, для $n = 2$, графік функції $f_2(x)$ зображений на рис. 2.2 (намальовано червоним кольором). Синім кольором, для порівняння, зображено графік щільності ймовірностей стандартного нормального розподілу.

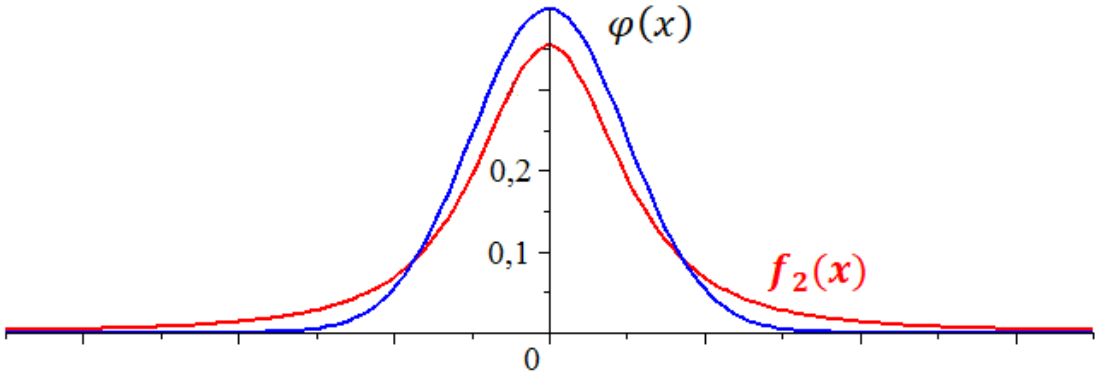


Рис. 2.2

Розподіл Фішера-Снедекора. Його щільність імовірностей має вигляд

$$f_{n,m} = \begin{cases} 0, & x \le 0 \\ C_{n,m} x^{\frac{n}{2}-1} \left(1 + \frac{n}{m} x\right)^{-\frac{n+m}{2}}, & x > 0, \end{cases}$$

де стала $C_{n,m}$ визначається з умови нормування

$$C_{n,m} \int_0^{\infty} x^{\frac{n}{2}-1} \left(1 + \frac{n}{m} x\right)^{-\frac{n+m}{2}} dx = 1.$$

Наприклад, для $n = 4$, $m = 8$ ця щільність має наступний вигляд (рис.2.3):

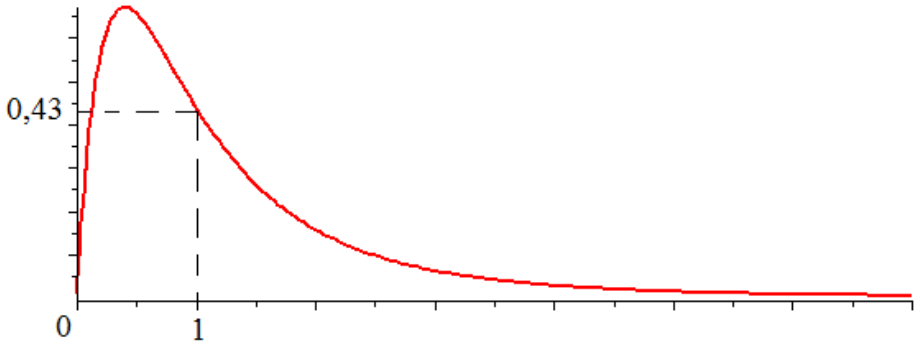


Рис. 2.3

2.5. Інтервальні статистичні оцінки для параметрів генеральної сукупності

У попередньому підрозділі розглядалися методи отримання статистичних оцінок (тобто статистики). Однією з основних вимог до статистики θ^* параметра θ є її ***слушність***, яка означає, що

$$\theta_n^* \to \theta \text{ при } n \to \infty \text{ за ймовірністю,}$$

тобто, що $\forall \delta > 0$

$$\gamma = \gamma(n) = \mathbf{P}\{\theta \in (\theta_n^* - \delta, \theta_n^* + \delta)\} \to 1 \text{ при } n \to \infty. \qquad (2.22)$$

Означення. Кожна ймовірність $\gamma = \gamma(n)$ називається ***надійністю*** статистичного оцінювання статистики θ_n^*, а сам інтервал $(\theta_n^* - \delta, \theta_n^* + \delta)$ - ***довірчим інтервалом***.

Грубо кажучи, слушність статистики означає, що беручи все більше і більше n, ми все точніше й точніше оцінюватимемо невідомий параметр θ, тобто надійність оцінювання стає все вище і вище. Але як точно? Формула (2.22) нічого не стверджує про швидкість збіжності ймовірностей до одиниці.

У цьому підрозділі буде розглянуте наступне питання. Нехай задана випадкова величина X, надійність $\gamma \in (0,1)$, натуральне число n і статистика θ^* невідомого

параметра θ. Як знайти такий інтервал $(\theta_n^* - \delta, \theta_n^* + \delta)$, щоб параметр θ потрапив до цього інтервалу з ймовірністю γ? У загальному випадку відповідь на це питання зовсім не проста. Ми займемося цим питанням для одного конкретного, але дуже важливого випадку: коли відомо, що випадкова величина X має **нормальний розподіл**.

Побудова довірчого інтервалу для математичного сподівання $a = \mathbf{M}X$ випадкової величини X при відомому значенні середнього квадратичного відхилення σ

Прикладом може бути вимірювання певної величини за допомогою якогось приладу, точність вимірювання σ якого відома.

Пригадаємо з теорії ймовірностей, що для нормальної випадкової величини X середнє $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ її незалежних копій $(X_i)_{i=1}^n$ теж є нормально розподіленою випадковою величиною [6]. Окрім того (див. підрозділ 2.3)

$$\mathbf{M}\bar{X}_n = \mathbf{M}X = a,$$

а

$$\mathbf{D}(\bar{X}_n) = \frac{1}{n}\mathbf{D}X,$$

Отже, середнє квадратичне відхилення

$$\sigma(\bar{X}_n) = \sigma/\sqrt{n}.$$

Відповідно, величина

$$Z_n = \frac{\bar{X}_n - a}{\sigma/\sqrt{n}}$$

має стандартний нормальний розподіл з параметрами (0, 1) і не залежить від n, a, та σ.

Позначимо через $\varphi(x)$ щільність стандартного нормального розподілу. Тоді, $\forall x > 0$

$$\mathbf{P}\{|Z_n| < x\} = \int_{-x}^{x} \varphi(u)du = 2\Phi(x).$$

Тут $\Phi(x) = \int_0^x \varphi(u)du$ – це функція Лапласа. Таблицю для неї можна знайти у Додатку 1. Графік функції щільності $\varphi(u)$ має наступний вигляд (рис. 2.4):

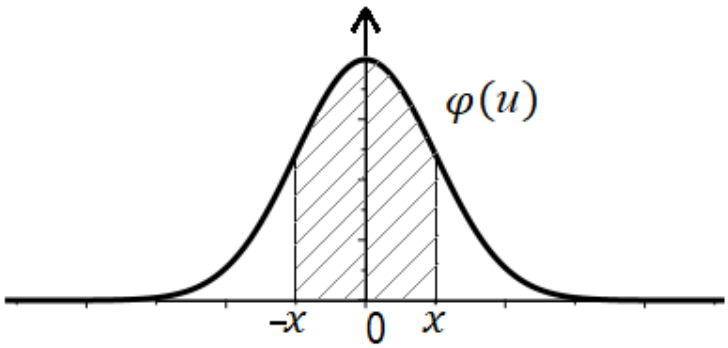


Рис. 2.4

Отже, за статистику θ_n^* для математичного сподівання a беремо випадкову величину $\bar{X}_n$. Тепер рахуємо: $\forall \delta > 0$

$$\mathbf{P}\{a \in (\bar{X}_n - \delta, \bar{X}_n + \delta)\} = \mathbf{P}\{\bar{X}_n - \delta < a < \bar{X}_n + \delta\} = \mathbf{P}\{-\delta < a - \bar{X}_n < \delta\} =$$

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

Тепер, маючи надійність γ з таблиці (Додаток 1) і рівняння

$$2\Phi\left(\frac{\sqrt{n}\delta}{\sigma}\right) = \gamma$$

знаходимо δ і довірчий інтервал $(\bar{X}_n - \delta, \bar{X}_n + \delta)$. Це – випадкова величина. Її реалізацією буде інтервал

$$(\bar{x}_n - \delta, \bar{x}_n + \delta),$$

де $(x_i)_{i=1}^n$ – числові значення випадкових величин $(X_i)_{i=1}^n$, отримані в результаті експерименту, а $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$.

Побудова довірчого інтервалу для математичного сподівання $a = \mathbf{M}X$ при невідомому значенні середнього квадратичного відхилення σ

Нагадаємо, що для довільного n і для незалежних копій $(X_i)_1^n$ випадкової величини X ми позначали $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ і $\hat{s}_n = \sqrt{\frac{n}{n-1}D_n} = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2}$.

Покладемо

$$\tau_{n-1} = \frac{\bar{X}_n - a}{\hat{s}_n/\sqrt{n}}.$$

Відомо, що τ_{n-1} має **розподіл Стьюдента** з $(n-1)$ ступенем свободи [10, с. 537]. Тому власне ми і вживаємо таке позначення.

Ступінь свободи у статистиці - це кількість значень чи спостережень у вибірці, які можуть бути змінені незалежно один від одного без зміни структури вибірки. Тобто це кількість змінних, які залишаються вільними для варіювання після того, як структура вибірки була визначена.

Таблиця ступенів свободи - це таблиця, яка заповнюється у відповідності з типом і кількістю змінних, які використовуються в аналізу статистичних даних. Вона використовується для визначення правильної формули розрахунку критичних значень при проведенні статистичних тестів, таких як t-критерій, F-критерій та xi-квадрат тест.

Тут важливо, що розподіл Стьюдента залежить лише від n, а не від a та σ. Докладніше він був описаний в підрозділі 2.4, а таблиця для цього розподілу міститься у Додатку 3.

Отже, довірчий інтервал для математичного сподівання $a = \mathbf{M}X$ з надійністю γ будуємо так, як і в попередньому випадку. За статистику θ_n^* для a беремо випадкову величину $\bar{X}_n$. Тоді

$$\mathbf{P}\{a \in (\bar{X}_n - \delta, \bar{X}_n + \delta)\} = \mathbf{P}\{|\bar{X}_n - a| < \delta\} = \mathbf{P}\left\{\frac{|\overline{X}_n - a|}{\hat{s}_n/\sqrt{n}} < \frac{\sqrt{n}\delta}{\hat{s}_n}\right\} =$$

$$= \mathbf{P}\left\{|\tau_{n-1}| < \frac{\sqrt{n}\delta}{\hat{s}_n}\right\} = 2\mathrm{T}_{n-1}\left(\frac{\sqrt{n}\delta}{\hat{s}_n}\right),$$

де

$$\mathrm{T}_{n-1}(x) = \int_0^x f_{n-1}(u)du,$$

а $f_{n-1}(u)$ – щільність розподілу випадкової величини t_{n-1}. Отже, з рівняння

$$2\mathrm{T}_{n-1}\left(\frac{\sqrt{n}\delta}{\hat{s}_n}\right) = \gamma$$

за таблицею (Додаток 3) знаходимо δ.

Реалізація. Для вибірки $(x_1, \dots, x_n)$ рахуємо середнє $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ та виправлене середнє квадратичне відхилення $s_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x}_n)^2}$. Тоді довірчий інтервал для математичного сподівання має вигляд

$$(\bar{x}_n - \delta, \bar{x}_n + \delta),$$

де число δ знаходимо із рівнянням

$$2\mathrm{T}_{n-1}\left(\frac{\sqrt{n}\delta}{s_n}\right) = \gamma.$$

Таким чином, цей метод дуже подібний до попереднього, тільки замість нормального розподілу беремо розподіл Стьюдента, а замість відомого середнього квадратичного відхилення – його оцінку s_n.

Побудова довірчих інтервалів із заданою надійністю γ для дисперсії і середнього квадратичного відхилення

Для побудови довірчого інтервалу для дисперсії $\sigma^2 = \mathbf{D}X$ використовують випадкову величину

$$\chi^2_{n-1} = \frac{n-1}{\sigma^2} \cdot (\hat{s}_n)^2. \qquad (2.23)$$

Відомо, що випадкова величина χ^2_{n-1} має розподіл хі-квадрат із $(n-1)$ ступенем свободи (див. [10, с. 536] та підрозділ 2.4); звідси таке позначення. Він залежить лише від n, а значення χ^2_{n-1} можна знайти у таблиці (Додаток 4) для обраного рівня значущості.

Проведемо наступні міркування. Нехай $0 < a < b$. Тоді

$$\mathbf{P}\{a < \chi^2_{n-1} < b\} = \mathbf{P}\left\{\frac{1}{b} < \frac{1}{\chi^2_{n-1}} < \frac{1}{a}\right\}. \qquad (2.24)$$

Підставляючи (2.23) в (2.24) маємо

$$\mathbf{P}\left\{\frac{1}{b} < \frac{\sigma^2}{(n-1)(\hat{s}_n)^2} < \frac{1}{a}\right\} = \mathbf{P}\left\{\frac{(n-1)(\hat{s}_n)^2}{b} < \sigma^2 < \frac{(n-1)(\hat{s}_n)^2}{a}\right\} = \gamma.$$

Отже, довірчий інтервал для σ^2 з надійністю γ має вигляд

$$\left(\frac{(n-1)(\hat{s}_n)^2}{b}; \frac{(n-1)(\hat{s}_n)^2}{a}\right),$$

де числа a, b знаходимо з рівності

$$\mathbf{P}\{a < \chi^2_{n-1} < b\} = \gamma.$$

Але тут рівняння одне, а невідомих – 2.

Пояснимо докладніше, як їх знаходити. У таблицях подано значення функції $\chi^2_{n-1}(u) = \mathbf{P}\{\chi^2_{n-1} > u\},\ u > 0$. Приблизно, графік цієї функції має такий вигляд (рис. 2.2):

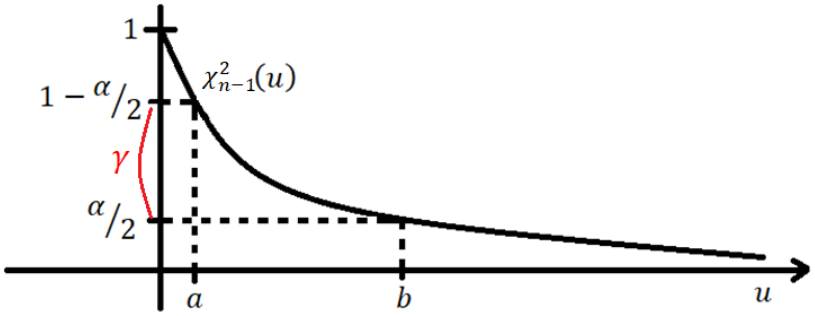


Рис. 2.5

Для знаходження чисел $a,\ b$ покладаємо $\alpha = 1 - \gamma$ і вибираємо довірчий інтервал $(a,\ b)$ так, щоб відповідний проміжок

$$(\mathbf{P}\{\chi^2_{n-1} > b\}, \mathbf{P}\{\chi^2_{n-1} > a\})$$

розташувався посередині інтервалу (a, b), тобто, щоб ймовірності завищення і заниження оцінки були однаковими, тобто числа $a,\ b$ знаходимо з наступних рівнянь

$$a\colon\ \mathbf{P}\{\chi^2_{n-1} > u\} = 1 - \alpha/2; \qquad b\colon\ \mathbf{P}\{\chi^2_{n-1} > u\} = \alpha/2.$$

Врешті, для числової реалізації цього методу, проводимо спостереження $x_1, \dots, x_n$, обчислюємо $s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$ і з наведених вище рівнянь, користуючись таблицями розкладу хі-квадрат, знаходимо значення a та b. Тоді довірчим інтервалом з надійністю γ буде проміжок

$$\left(\frac{(n-1)s_n^2}{b}; \frac{(n-1)s_n^2}{a}\right).$$

Зауваження. Звідси отримаємо довірчий інтервал для середнього квадратичного відхилення σ

$$\left(\frac{\sqrt{n-1}\cdot s_n}{\sqrt{b}}; \frac{\sqrt{n-1}\cdot s_n}{\sqrt{a}}\right).$$

2.6. Приклади до Розділу 2

Приклад 1. Точкові оцінки

200 однотипних деталей були піддані шліфуванню. Результати вимірювання наведені як дискретний статистичний розподіл, поданий у табличній формі:

x_i, мм	3,7	3,8	3,9	4,0	4,1	4,2	4,3	4,4
n_i	1	22	40	79	27	26	4	1

Знайти точкові незміщені статистичні оцінки для середнього значення і дисперсії генеральної сукупності.

Розв'язання.

Оскільки точковою незміщеною оцінкою для середнього значення генеральної сукупності є середня вибіркова величина $\bar{x}$, то обчислимо її:

$$\bar{x} = \frac{3{,}7 \cdot 1 + 3{,}8 \cdot 22 + \cdots + 4{,}4 \cdot 1}{200} = 4{,}004 \text{ (мм)}.$$

Для визначення точкової незміщеної статистичної оцінки для дисперсії генеральної сукупності обчислимо дисперсію вибірки:

$$D = \frac{3{,}7^2 \cdot 1 + 3{,}8^2 \cdot 22 + \cdots + 4{,}4^2 \cdot 1}{200} - (4{,}004)^2 = 0{,}01598.$$

Тоді точкова незміщена статистична оцінка для $\mathbf{D}X$ дорівнюватиме:

$$\frac{n}{n-1} \cdot D = \frac{200}{199} 0{,}01598 = 0{,}01606 \text{ (мм}^2\text{)}.$$

Приклад 2. Побудова довірчого інтервалу для математичного сподівання $\mathbf{M}X$ випадкової величини X при відомому значенні середнього квадратичного відхилення σ

Учні виконують стрибки у висоту. Висота стрибка навмання взятого учня є випадковою величиною з нормальним розподілом імовірностей, причому середнє квадратичне відхилення $\sigma = 5$ см. За даними 100 спостережень здобута статистична оцінка для математичного сподівання випадкової величини $\bar{x} = 135$ см. Знайти довірчий інтервал, в якому з імовірністю 0,9 знаходиться математичне сподівання висоти стрибка учня.

Розв'язання.

Нагадаємо формулу для знаходження довірчого інтервалу:

$$(\bar{x} - \delta, \bar{x} + \delta).$$

Значення δ можемо знайти із наступного співвідношення:

$$2\Phi\left(\frac{\sqrt{n}\delta}{\sigma}\right) = \gamma.$$

У нашому випадку

$$2\Phi\left(\frac{\sqrt{n}\delta}{\sigma}\right) = 0{,}9 \;\Rightarrow\; \Phi\left(\frac{\sqrt{n}\delta}{\sigma}\right) = \frac{0{,}9}{2} = 0{,}45.$$

За таблицею із Додатку 1 знаходимо значення $\frac{\sqrt{n}\delta}{\sigma}$ при $\Phi\left(\frac{\sqrt{n}\delta}{\sigma}\right) = 0{,}45$. Отримаємо, що

$$\frac{\sqrt{n}\delta}{\sigma} = 1{,}66 \quad \Rightarrow \quad \delta = \frac{1{,}66 \cdot \sigma}{\sqrt{n}} = \frac{1{,}66 \cdot 5}{\sqrt{100}} = 0{,}83.$$

Відповідно довірчий інтервал:

$$(135 - 0{,}83;\ 135 - 0{,}83) \quad \Rightarrow \quad (134{,}17; 135{,}83) \quad \Rightarrow$$

$$134{,}17 < \mathbf{M}X < 135{,}83.$$

Приклад 3. Побудова довірчого інтервалу для математичного сподівання $\mathbf{M}X$ при невідомому значенні середнього квадратичного відхилення σ

Соціологічне обстеження освіти 20 співробітників організації дало наступні результати:

Номер анкети	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Кількість років, затрачених на отримання освіти	15	14	15	15	10	13	17	18	15	14	15	15	13	10	18	16	15	14	15	15

У даній організації вважається, що у середньому 15 років (10 років у школі і 5 років у закладі вищої освіти) – стандартний рівень освіти для співробітника. А якщо він навчався менше 15 років, то його рівень освіти не достатній. Потрібно оцінити, чи можна на рівні надійності 0,95 стверджувати, що середній рівень освіти співробітників нижче стандарту.

Розв'язання.

Розрахуємо $\bar{x}$ як звичайне середнє арифметичне, так як ряд у нас не згрупований, а просто наведені показники для кожного співробітника. Отримаємо:

$$\bar{x} = \frac{15 + 14 + 15 + 15 + \cdots + 15}{20} = 14{,}6.$$

Тут для побудови довірчого інтервалу використовується випадкова величина

$$t_n = \frac{\bar{x} - a}{s_n/\sqrt{n}}.$$

Знайдемо s_n:

$$D = \frac{15^2 + 14^2 + 15^2 + 15^2 + \cdots + 15^2}{20} - 14{,}6^2 = 4{,}04;$$

$$s_n = \sqrt{\frac{20}{19} \cdot 4{,}04} = 2{,}062.$$

Величина t_n знаходиться за таблицею розподілу Стьюдента (Додаток 3) з параметрами $\gamma = 0{,}95$ та $k = n - 1$ буде рівна

$$t_n(\gamma = 0{,}95; k = 19) = 2{,}093.$$

Довірчий інтервал лежатиме в межах:

$$\bar{x} - \frac{t_n \cdot s_n}{\sqrt{n}} < a < \bar{x} + \frac{t_n \cdot s_n}{\sqrt{n}}.$$

Підставимо всі отримані значення:

$$14{,}6 - \frac{2{,}093 \cdot 2{,}062}{\sqrt{20}} < \mathbf{M}X < 14{,}6 + \frac{2{,}093 \cdot 2{,}062}{\sqrt{20}} \quad \Rightarrow$$

$$13{,}6 < \mathbf{M}X < 15{,}6.$$

Таким чином з надійністю 95% рівень освіти відповідає стандарту 15 років, який покриває отриманий довірчий інтервал.

Приклад 4. Побудова довірчих інтервалів із заданою надійністю γ для дисперсії $\mathbf{D}X$ і середнього квадратичного відхилення $\sigma = \sigma(X)$

Перевірена партія однотипних телевізорів x_i на чутливість до відеопрограм n_i, дані перевірки наведено як дискретний статистичний розподіл:

n_i, мкв	200	250	300	350	400	450	500	550
x_i	2	5	6	7	5	2	2	1

З надійністю $\gamma = 0{,}99$ побудувати довірчий інтервали для $\mathbf{D}X$ та $\sigma(X)$.

Розв'язання.

Розрахуємо вибіркові числові характеристики:

$$\bar{x} = \frac{200 \cdot 2 + 250 \cdot 5 + \cdots + 550 \cdot 1}{30} = 345 \text{ (мкв)};$$

$$D = \frac{200^2 \cdot 2 + 250^2 \cdot 5 + \cdots + 550^2 \cdot 1}{30} - 345^2 = 7558{,}33 \text{ (мкв}^2\text{)};$$

$$s_n = \frac{30}{29} \cdot 7558{,}33 = 7818{,}97 \quad \Rightarrow \quad s_n = 88{,}42.$$

Знайдемо імовірності:

$$a\colon \mathbf{P}\{\chi_n^2 > u\} = 1 - {}^{\alpha}/_{2}, \qquad b\colon \mathbf{P}\{\chi_n^2 > u\} = {}^{\alpha}/_{2}.$$

де $\alpha = 1 - \gamma = 1 - 0{,}99 = 0{,}01$. Відповідно

$$\mathbf{P}\{\chi_{n-1}^2 > u\} = 1 - \frac{0{,}01}{2} = 0{,}995;$$

$$\mathbf{P}\{\chi_{n-1}^2 > u\} = \frac{0{,}01}{2} = 0{,}005.$$

За таблицею (Додаток 4) знаходимо:

$$\chi_{29}^2\,(0{,}995) = 14{,}3;$$

$$\chi_{29}^2\,(0{,}005) = 52{,}3.$$

Підставляємо всі отримані значення у формули довірчого інтервалу дисперсії:

$$\left(\frac{(n-1)s_n^2}{b}, \frac{(n-1)s_n^2}{a}\right) \quad \Rightarrow \quad \left(\frac{29 \cdot 7818{,}97}{52{,}3}, \frac{29 \cdot 7818{,}97}{14{,}3}\right) \quad \Rightarrow$$

$$4319 \text{ (мкв}^2\text{)} < \mathbf{D}X < 15856{,}5 \text{ (мкв}^2\text{)}.$$

Довірчий інтервал для σ можемо отримати наступним чином:

$$\sqrt{4319} < \sigma < \sqrt{15856{,}5} \quad \Rightarrow$$

$$68{,}3 \text{ (мкв)} < \sigma < 130{,}83 \text{ (мкв)}.$$

2.7. Питання для самоконтролю до Розділу 2

1. Що таке статистичне оцінювання?
2. Яка статистика називається незміщеною?
3. Яка статистика називається слушною?
4. Що таке кількість інформації за Фішером?
5. Запишіть нерівність Крамера-Рао для неперервних розподілів.
6. Яка оцінка невідомого параметра називається ефективною?
7. Як обґрунтувати ефективність оцінки математичного сподівання нормального розподілу?
8. Запишіть нерівність Крамера-Рао для дискретних розподілів.
9. Опишіть метод аналогій для оцінки невідомого математичного сподівання.
10. Опишіть метод аналогій для оцінки невідомої дисперсії.
11. Чому метод аналогій дає зміщену оцінку дисперсії? Як отримати незміщену?
12. В чому полягає метод максимальної вірогідності?

13. Як застосувати метод максимальної вірогідності до оцінки невідомої ймовірності двійкового розподілу?
14. Як застосувати метод максимальної вірогідності до оцінки параметрів нормального розподілу?
15. Дайте означення та опишіть властивості розподілу хі-квадрат.
16. Дайте означення та опишіть властивості розподілу Стьюдента .
17. Дайте означення та опишіть властивості розподілу Фішера-Снедекора.
18. Що таке довірчий інтервал і його надійність?
19. Опишіть алгоритм знаходження довірчого інтервалу для математичного сподівання при відомому значенні середнього квадратичного відхилення.
20. Опишіть алгоритм знаходження довірчого інтервалу для математичного сподівання при невідомому значенні середнього квадратичного відхилення.
21. Опишіть алгоритм знаходження довірчого інтервалу для дисперсії.

РОЗДІЛ 3. СТАТИСТИЧНІ ГІПОТЕЗИ

3.1. Загальні поняття

Означення. Певні судження про генеральну сукупність, сформульовані на підставі вибірки, називають ***статистичними гіпотезами***.

Статистичні гіпотези використовуються для прийняття ***рішень***. Ці рішення називаються ***статистичними***. Вони мають ймовірнісний характер, тобто завжди існує ймовірність помилки.

Означення. Статистичні гіпотези про значення параметрів випадкової величини називаються **параметричними**.

Це можуть бути, наприклад, гіпотези про значення $\mathbf{M}X$, $\mathbf{D}X$ тощо.

Означення. Гіпотезу, що підлягає перевірці, називають ***основною*** і позначають H_0.

Приклад. $H_0: \mathbf{M}X = a,\ a \in \mathbf{R}$.

Кожній гіпотезі протиставляють ***альтернативну*** гіпотезу H_1.

Приклади. $\mathbf{M}X \neq a,\ \mathbf{M}X < a,\ \mathbf{M}X > a$.

Статистичний критерій

Для перевірки правильності статистичної гіпотези вибирають так званий ***статистичний критерій*** Z, керуючись яким, гіпотезу приймають, або відхиляють.

Статистичний критерій є випадковою величиною, розподіл якої відомий.

Приклад. Для перевірки гіпотези $H_0: \mathbf{M}X = a$ для нормальної випадкової величини, коли середнє квадратичне відхилення σ відоме, можна взяти випадкову величину

$$Z = Z(n) = \frac{\bar{X}_n - a}{\sigma/\sqrt{n}},$$

яка має нормальний розподіл з параметрами $(0,1)$.

Алгоритм застосування цього критерію буде описано у даному розділі.

Критична область і критичні точки

Множину M всіх значень статистичного критерію Z можна поділити на дві множини A і $\bar{A}$: $A \cup \bar{A} = M,\ A \cap \bar{A} = \emptyset$.

Означення. Сукупність A значень статистичного критерію Z за яких H_0 приймається, називається ***областю прийняття*** H_0, сукупність $\bar{A}$ значень критерію Z, за яких H_0 відхиляється, називається ***критичною областю***.

Точки (або точка), що поділяють множину M на підмножини A та $\bar{A}$ називають ***критичними***.

Критичну точку позначатимемо z_α. (в природності цього позначення ми скоро переконаємося).

Приклади.

Для гіпотези H_0: $\mathbf{M}X = a$ з альтернативою $\mathbf{M}X < a$ критична область має вигляд (рис. 3.1)

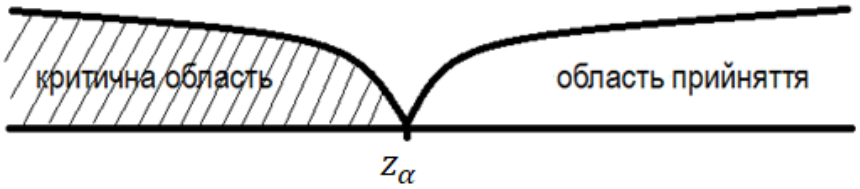


Рис. 3.1

Для гіпотези H_0: $\mathbf{M}X = a$ з альтернативою $\mathbf{M}X > a$ критична область має вигляд (рис. 3.2)

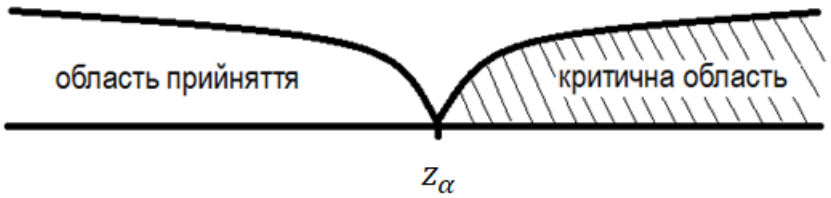


Рис. 3.2

Для гіпотези H_0: $\mathbf{M}X = a$ з альтернативою $\mathbf{M}X \neq a$ критична область має вигляд (рис. 3.3)

Рис. 3.3

Загальний алгоритм перевірки правильності гіпотези

Для перевірки правильності гіпотези H_0 задається ***рівень значущості α***. Найчастіше беруть $\alpha = 0{,}05;\ \ 0{,}01;\ \ 0{,}001.$

В основу перевірки гіпотези H_0 закладений наступний

Принцип. Якщо $\mathbf{P}\{K \in \bar{A}\} \leq \alpha$, то гіпотеза H_0 приймається.

Щоб скористатися цим принципом проводиться спостереження. Якщо результат спостереження $z_o \in A$, (о від observed – спостережуване), то гіпотеза H_0 приймається, а якщо $z_o \in \bar{A}$, то відхиляється.

Опишемо алгоритм перевірки гіпотези докладніше.

1. Формулюється основна гіпотеза H_0 і альтернатива H_1.

2. Вибирається статистичний критерій і рівень значущості.

3. Будуються критична область і критичні точки.

4. Проводиться спостереження (експеримент). У результаті отримаємо вибірку. За результатами цієї вибірки обчислюється спостережуване значення критерію z_o.

5. Якщо $z_o \in A$, то гіпотеза приймається, а якщо $z_o \in \bar{A}$, то відхиляється.

Помилки. Потужність критерію

При перевірці статистичних гіпотез можливі помилки. Розрізняють помилки 1-го роду і 2-го роду.

Помилка 1-го роду. Гіпотеза H_0 є правильною, але її відхилено. Ймовірність цієї помилки позначаємо так: $\mathbf{P}\{K \in \bar{A}|H_0\}$. За умовою, вона повинна дорівнювати рівню значущості α.

Помилка 2-го роду. Гіпотеза H_0 хибна (тобто справедлива гіпотеза H_1), але ми її прийняли. Ймовірність цієї помилки позначаємо так: $\beta = \mathbf{P}\{K \in A|H_1\}$.

Означення. Різницю $1 - \beta = 1 - \mathbf{P}\{K \in A|H_1\}$ називають ***потужністю*** критерію.

3.2. Перевірка гіпотези про числове значення математичного сподівання при відомій (невідомій) дисперсії

Тут ми розглянемо гіпотезу щодо математичного сподівання випадкової величини X, H_0: $\mathbf{M}X = a$, де a – певне число.

Спочатку розглянемо випадок, коли X є нормальною випадковою величиною з відомим середнім квадратичним відхиленням σ. Природнім статистичним критерієм тут є випадкова величина

$$Z = Z_n = \frac{\bar{X}_n - a}{\sigma/\sqrt{n}},$$

де як завжди

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

а X_i – незалежні копії X.

Як ми вже казали, Z є нормальною випадковою величиною з параметрами $(0, 1)$. Врешті, зафіксуємо рівень значущості α.

Перевірка гіпотези $H_0: \mathbf{M}X = a$ з альтернативою $H_1: \mathbf{M}X > a$

Спочатку обчислюємо критичну точку z_α як розв'язок рівняння

$$\mathbf{P}\{Z > z\} = \alpha,$$

де α – рівень значущості.

Покажемо, як це робити. Графік щільності стандартного нормального розподілу має вигляд (рис.3.4)

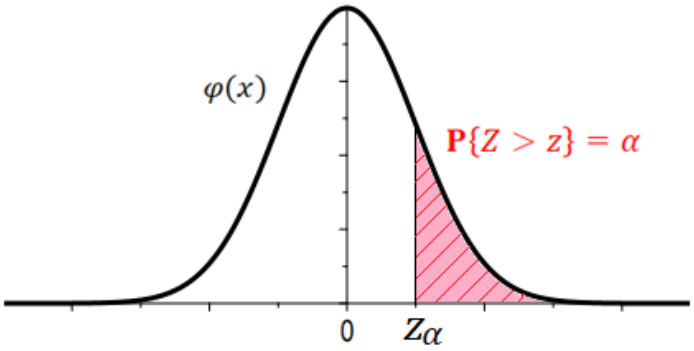


Рис. 3.4

Очевидно,

$$\mathbf{P}\{0 < Z < z_\alpha\} + \mathbf{P}\{Z > z_\alpha\} = \frac{1}{2}.$$

Тому

$$\mathbf{P}\{0 < Z < z_\alpha\} + \alpha = \frac{1}{2}.$$

Звідки

$$\mathbf{P}\{0 < Z < z_\alpha\} = \frac{1 - 2\alpha}{2}$$

та

$$\Phi(z_\alpha) - \Phi(0) = \tfrac{1-2\alpha}{2}.$$

Тут $\Phi(z) = \int_0^z \varphi(x)dx$ – функція Лапласа. Тому

$$\Phi(z_\alpha) = \frac{1 - 2\alpha}{2}.$$

За таблицями значень функції Лапласа (Додаток 1), скориставшись значенням $\frac{1-2\alpha}{2}$ знаходимо аргумент z_α .

Потім проводимо експеримент. Отримуємо значення $x_1, \dots, x_n$. Обчислюємо

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

та спостережуване значення

$$z_{\text{o}} = \frac{\bar{x} - a}{\sigma / \sqrt{n}}$$

випадкової величини Z. Якщо $z_{\text{o}} > z_\alpha$ – гіпотезу H_0 відхиляємо, якщо $z_{\text{o}} \le z_\alpha$ – приймаємо.

Перевірка гіпотези H_0: $\mathbf{M}X = a$ з альтернативою H_1: $\mathbf{M}X < a$

Спочатку обчислюємо критичну точку z_α із рівняння

$$\mathbf{P}\{Z < z\} = \alpha,$$

де α – рівень значущості.

Пригадаємо графік щільності $\varphi(x)$ стандартного нормального розподілу (рис. 3.5). Оскільки α мале, то $z_\alpha < 0$.

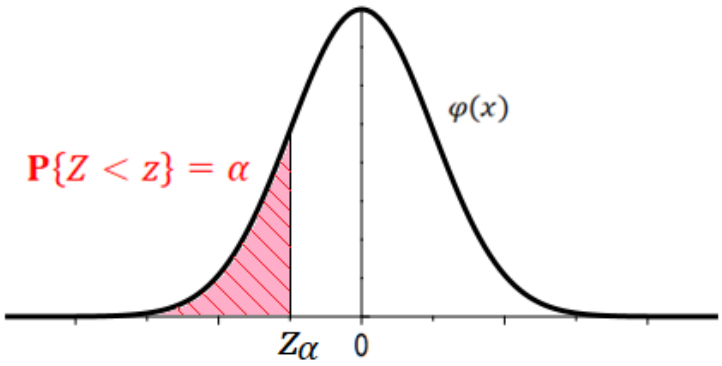


Рис.3.5

Очевидно,

$$\mathbf{P}\{Z < z_\alpha\} + \mathbf{P}\{z_\alpha < Z < 0\} = \frac{1}{2}.$$

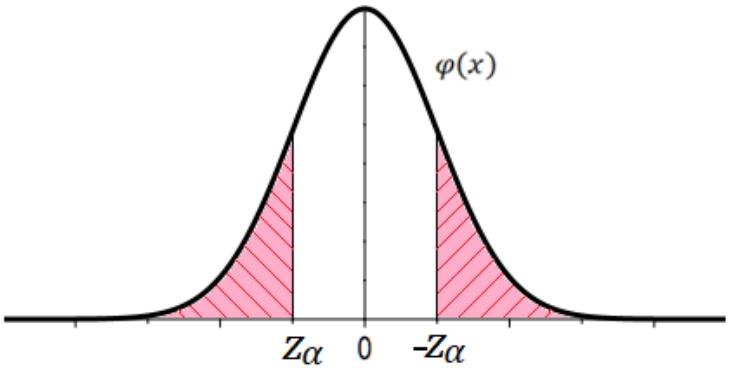


Рис.3.6

Внаслідок симетрії (рис. 3.6),

$$\underset{\alpha}{\mathbf{P}\{Z > -z_\alpha\}} + \underset{\Phi(z_\alpha) - \Phi(0)}{\mathbf{P}\{0 < Z < -z_\alpha\}} = \frac{1}{2}.$$

$$\Phi(0) = 0$$

Звідси

$$\Phi(-z_\alpha) = \frac{1-2\alpha}{2}.$$

З цього рівняння знаходимо $-z_\alpha$, а потім z_α.

Врешті, проводимо експеримент. Отримуємо значення $x_1, \dots, x_n$ випадкової величини X. Обчислюємо $\bar{x}$ та

$$z_{\text{o}} = \frac{\bar{x} - a}{\sigma/\sqrt{n}}.$$

Якщо $z_{\text{o}} < z_\alpha$ – гіпотезу H_0 відхиляємо, якщо $z_{\text{o}} \geq z_\alpha$ – приймаємо.

Перевірка гіпотези $H_0: \mathbf{M}X = a$ з альтернативою $H_1: \mathbf{M}X \neq a$

Тут потрібно знайти дві критичні точки z_α і z'_α з умов

$$\mathbf{P}\{Z > z_\alpha\} = \frac{\alpha}{2} \text{ і } \mathbf{P}\{Z < z'_\alpha\} = \frac{\alpha}{2},$$

де α – рівень значущості.

Звичайно, $z'_\alpha = -z_\alpha$.

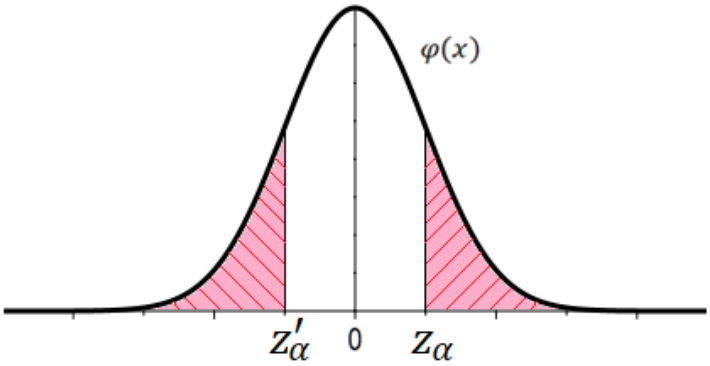


Рис. 3.7

Отже, досить обчислити лише z_α. Для цього скористаємося рівнянням

$$\mathbf{P}\{0 < Z < z_\alpha\} + \mathbf{P}\{Z > z_\alpha\} = \frac{1}{2}.$$

Звідси,

$$\mathbf{P}\{0 < Z < z_\alpha\} + \frac{\alpha}{2} = \frac{1}{2}.$$

Звідси,

$$\Phi(z_\alpha) - \Phi(0) = \frac{1-2\alpha}{2},$$

тобто

$$\Phi(z_\alpha) = \frac{1-2\alpha}{2}.$$

З цього рівняння знаходимо z_α та $z'_\alpha = -z_\alpha$.

Врешті, проводимо експеримент і знаходимо z_o. Якщо $|z_\text{o}| > z_\alpha$ – гіпотезу H_0 відхиляємо, якщо $|z_\text{o}| < z_\alpha$ – приймаємо.

Перевірка гіпотези H_0: $\mathbf{M}X = a$, коли середньоквадратичне відхилення σ невідоме

У цьому випадку за статистичний критерій беруть випадкову величину

$$\tau_{n-1} = \frac{\bar{X}_n - a}{\hat{s}_n/\sqrt{n}},$$

де, пригадаємо, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, $\hat{s}_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2}$.

Випадкова величина t_n має розподіл Стьюдента з $(n-1)$ ступенем свободи (Додаток 3). Наступні кроки перевірки гіпотези H_0 схожі на випадок, коли середньоквадратичне відхилення відоме. Докладніший розгляд цього випадку буде темою для самостійного опрацювання.

3.3. Перевірка правильності гіпотези про рівність математичних сподівань

Зауваження! У даному підрозділі розглядається випадок нормально розподілених випадкових величин. Відповідна теорія для не нормально розподілених випадкових величин представлена у розділі 7.

Незалежність та залежність вибірок

З точки зору теорії ймовірностей ***незалежність*** (наприклад) двох вибірок $x_1, \dots, x_n$ та $y_1, \dots, y_m$ означає, що їх розглядають як реалізації набору незалежних в сукупності випадкових величин $X_1, \dots, X_n, Y_1, \dots, Y_m$. Натомість ***залежність*** (або ***зв'язність***) вибірок однакової довжини $x_1, \dots, x_n$ та $y_1, \dots, y_n$ означає, що їх розглядають як реалізації наборів випадкових величин $X_1, \dots, X_n, Y_1, \dots, Y_n$, величини X_1 та Y_1, X_2 та Y_2, …, X_n та Y_n, як правило, між собою залежні, але їхні різниці $X_i - Y_i$ мають вигляд $\Delta + \varepsilon_i$, $\Delta \in \mathbf{R}$, і випадкові величини (ε_i) незалежні в сукупності (див. напр. [9, с. 366]).

У цьому випадку пари $(X_1, Y_1), \dots, (X_n, Y_n)$ можна інтерпретувати як $2n$ спостережень – по два спостереження на кожному з n об'єктів. Тоді X_i називається ***спостереженням до оброблення***, а Y_j – ***спостереженням після оброблення***, $i = 1, \dots, n$.

Пояснимо різницю між незалежними та зв'язаними вибірками на прикладі. Гіпотезу про дієвість препарату для зниження тиску можна перевіряти різними способами. Можна взяти дві групи кроликів, одній групі давати препарат, іншій – не давати, а потім поміряти середній тиск у кожній групі. Тут вибірки природно вважати незалежними. При цьому, якщо у першій групі тиск не буде істотно меншим, ніж у другій, то робимо висновок про неефективність препарату. А можна перевіряти інакше: взяти лише одну групу кроликів, поміряти і обчислити середній тиск, а потім дати їм препарат і знову поміряти й обчислити середній тиск. Тут, природно, тиски індивідуального кролика до і після прийняття препарату залежні між собою; вони зв'язані із загальним станом здоров'я тваринки, але різниці (Z_i) незалежні між собою, бо стосуються різних кроликів.

Перевірка правильності гіпотези про рівність математичних сподівань для незалежних вибірок

Нехай дано дві нормальні незалежні випадкові величини X, Y і необхідно перевірити гіпотезу

$$H_0: \mathbf{M}X = \mathbf{M}Y.$$

Розглянемо три випадки.

1) Дисперсії $\mathbf{D}X$ і $\mathbf{D}Y$ відомі

Тут задача зводиться до окремого випадку ситуації, яку ми вже розглядали. А саме, потрібно перевірити гіпотезу $\mathbf{M}(X - Y) = 0$. Невелика різниця полягає лише в тому, що для X та Y можуть проводитися різні кількості випробувань (n і m).

Отже, за статистичний критерій тут береться випадкова величина

$$Z_{nm} = \frac{(\bar{X}_n - \mathbf{M}X) - (\bar{Y}_m - \mathbf{M}Y)}{\sigma(\bar{X}_n - \bar{Y}_n)}. \qquad (3.1)$$

Вона має нормальний розподіл з параметрами $(0, 1)$.

Оскільки, внаслідок незалежності,

$$\mathbf{D}(\bar{X}_n - \bar{Y}_n) = \frac{\mathbf{D}X}{n} + \frac{\mathbf{D}Y}{m},$$

то критерій (3.1) набуде вигляду

$$Z_{nm} = \frac{(\bar{X}_n - \mathbf{M}X) - (\bar{Y}_m - \mathbf{M}Y)}{\sqrt{\frac{\mathbf{D}X}{n} + \frac{\mathbf{D}Y}{m}}}.$$

Тепер, перевірку H_0 проводимо за наступною схемою.

1. Знаючи рівень значущості α та вигляд альтернативної гіпотези ($H_1: \mathbf{M}X > \mathbf{M}Y$, $\mathbf{M}X < \mathbf{M}Y$, $\mathbf{M}X \neq \mathbf{M}Y$), користуючись таблицями функції Лапласа (Додаток 1), обчислюємо критичну точку (чи критичні точки) z_α.

2. Проводимо n і m спостережень відповідно $x_1, \dots, x_n$ та $y_1, \dots, y_m$ і обчислюємо спостережуване значення

$$z_{\text{o}} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\mathbf{D}X}{n} + \frac{\mathbf{D}Y}{m}}}.$$

3. Порівнюючи z_α та z_{o}, приймаємо, або відхиляємо гіпотезу.

2) Дисперсії $\mathbf{D}X$ та $\mathbf{D}Y$ невідомі і обсяги вибірок великі ($n, m > 40$)

Як і раніше, позначимо

$$\hat{s}_n(X) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}, \quad \hat{s}_m(Y) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{m}(Y_i - \bar{Y}_m)^2},$$

а ще позначимо

$$\hat{s}_{nm} = \sqrt{\frac{(n-1)\hat{s}_n(X) + (m-1)\hat{s}_m(Y)}{n+m-1}}.$$

З ЦГТ [6, п.1.29] випливає, що при прямуванні $n, m \to \infty$ функція розподілу випадкової величини

$$Z_{nm} = \frac{(\bar{X}_n - \mathbf{M}X) - (\bar{Y}_m - \mathbf{M}Y)}{\hat{s}_{nm}\sqrt{\frac{1}{n} + \frac{1}{m}}} \qquad (3.2)$$

наближається до функції розподілу нормальної випадкової величини з параметрами $(0,\ 1)$. Тому для великих n, m за статистичний критерій перевірки гіпотези H_0: $\mathbf{M}X = \mathbf{M}Y$ можна брати не (3.2), а стандартну нормальну величину. Отже:

1. Для визначення критичної точки z_α (або двох критичних точок) застосовують функцію Лапласа (як і в попередньому випадку).

2. Проводять n і m спостережень випадкової величини X та Y відповідно і обчислюють середні $\bar{x}$ та $\bar{y}$. Далі обчислюють

$$s_n(X) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad s_m(Y) = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(y_i - \bar{y})^2}$$

і, врешті,

$$s_{nm} = \sqrt{\frac{(n-1)s_n^2(X) + (m-1)s_m^2(Y)}{n+m-1}}.$$

Тоді, спостережувана величина критерію матиме вигляд

$$z_{\text{o}} = \frac{\bar{x} - \bar{y}}{s_{nm}\sqrt{\frac{1}{n} + \frac{1}{m}}} \qquad (\mathbf{M}X = \mathbf{M}Y\ !)$$

3. Порівнюють z_{o} з критичною точкою z_α (критичними точками) і приймають або відхиляють гіпотезу.

3) Дисперсії невідомі і обсяги вибірок малі ($n, m < 40$)

Нагадаємо, що за загальний статистичний критерій для перевірки гіпотези рівності математичних сподівань ми брали випадкову величину

$$Z_{nm} = \frac{(\bar{X}_n - \mathbf{M}X) - (\bar{Y}_m - \mathbf{M}Y)}{\hat{s}_{nm}\sqrt{\frac{1}{n} + \frac{1}{m}}}.$$

На підставі ЦГТ [6, п.1.29] для великих m, n ми замінювали її стандартним нормальним розподілом. Але для малих m, n цього робити не можна. Проте відомо, що Z_{nm} має розподіл Стьюдента з $(n + m - 2)$ ступенем свободи [4, с.112]. Тому все робимо як і раніше, замінюючи $\Phi(z)$ розподілом Стьюдента.

Перевірка правильності гіпотези про рівність математичних сподівань для зв'язних вибірок

Розглянемо дві вибірки $x_1, \dots, x_n$ та $y_1, \dots, y_n$. Зауважимо, що коли мова йде про зв'язні вибірки, то мається на увазі, що ці вибірки одного обсягу. Спочатку розглянемо різниці елементів вибірок

$$d_i = x_i - y_i, \quad i = \overline{1,n}.$$

Відповідно спостережувана величина критерію матиме вигляд

$$z_{\text{o}} = \frac{\bar{d}}{s_d/\sqrt{n}},$$

де

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n};$$

$$s_d = \sqrt{\frac{\sum_{i=1}^{n} d_i^2 - \frac{\left(\sum_{i=1}^{n} d_i\right)^2}{n}}{n-1}}.$$

Порівнюючи z_α та z_{o}, приймаємо, або відхиляємо гіпотезу. Для малих вибірок для знаходження z_α використовуємо таблицю Стьюдента для $(n-1)$ ступенів свободи і заданого рівня значущості α (Додаток 3), а для вибірок великого обсягу (на підставі ЦГТ [6, п.1.29]) використовуємо функцію Лапласа (Додаток 1).

3.4. Перевірка правильності гіпотези про рівність дисперсій

Отже, нехай маємо нормальні незалежні випадкові величини X, Y із невідомими дисперсіями $\mathbf{D}X$ і $\mathbf{D}Y$. Хочемо перевірити правильність нульової гіпотези

$$H_0: \mathbf{D}X = \mathbf{D}Y$$

з альтернативною гіпотезою

$$H_1: \mathbf{D}X > \mathbf{D}Y.$$

Гіпотезу H_0 інакше можна записати у вигляді

$$\frac{\mathbf{D}X}{\mathbf{D}Y} = 1.$$

Пригадаємо, що за незміщену статистичну оцінку для $\boldsymbol{DX}$ ми брали випадкову величину

$$(\hat{s}_n)^2(X) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

і так само для Y :

$$(\hat{s}_m)^2(Y) = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y}_m)^2.$$

Тому за статистичний критерій для перевірки гіпотези H_0 природно взяти випадкову величину

$$F = F_{nm} = \frac{(\hat{s}_n)^2(X)}{(\hat{s}_m)^2(Y)}. \qquad (3.3)$$

Ідея застосування цього критерію така. Проводимо n і m вимірювань випадкових величин X і Y відповідно: $x_1, \dots, x_n$ та $y_1, \dots, y_m$.

Рахуємо

$$z_o = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}{\frac{1}{m-1}\sum_{i=1}^{m}(y_i - \bar{y})^2} \quad . \qquad (3.4)$$

Якщо це число близьке до 1, то гіпотезу H_0 приймаємо, якщо далеке – відхиляємо.

Формально, це застосування виглядає так: з теорії ймовірності відомо, що за умови рівності дисперсій: $\mathbf{D}X = \mathbf{D}Y$, випадкова величина (3.3) має розподіл Фішера-Сендекора з $k_1 = (n-1)$ та $k_2 = (m-1)$ ступенями свободи [10, с. 592]. Його значення є в таблицях (див. Додаток 5), точне означення в підрозділі 2.4, а графік щільності ймовірностей $f(x)$ має вигляд (рис. 3.8)

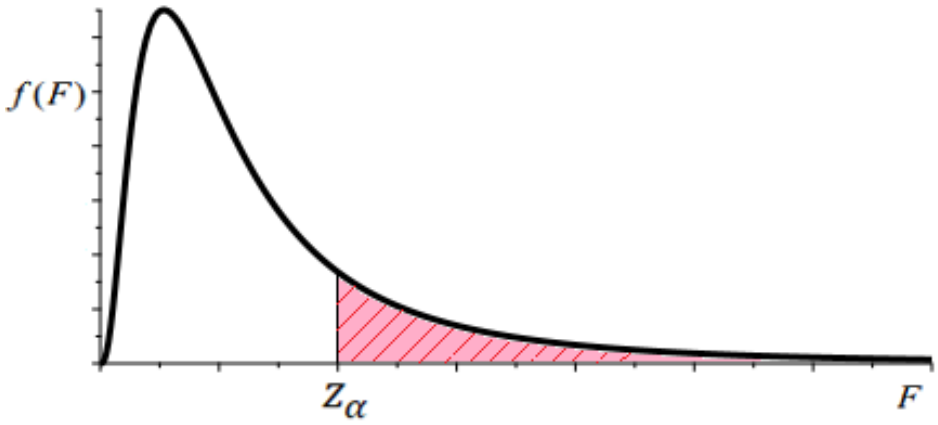


Рис. 3.8

Отже, обираємо рівень значущості α. Далі:

1. З рівняння $\mathbf{P}\{F > z\} = \alpha$, користуючись таблицями (Додаток 5), знаходимо z_α.

Зауваження. У таблицях подано значення для функції

$$\Psi(x) = \mathbf{P}\{F > x\} = \int_x^\infty f(u)du.$$

Тоді

$$\mathbf{P}\{F < x\} = 1 - \mathbf{P}(F > x) = 1 - \Psi(x).$$

2. Проводимо спостереження і за формулою (3.4) знаходимо z_{o}.

3. Якщо $z_{\text{o}} < z_{\alpha}$ – гіпотезу приймають. Якщо ні – відхиляють.

3.5. Перевірка правильності непараметричних статистичних гіпотез

У цьому підрозділі розглядаємо не гіпотези, що стосуються окремих параметрів, а гіпотези, що стосуються цілого розподілу випадкової величини. Спочатку розглянемо окремий випадок.

Дискретний розподіл

Нехай X – дискретна випадкова величина, що набуває лише скінченну кількість значень. Її розподіл повністю характеризується наступною таблицею.

Таблиця 3.1

x_1	...	x_k
p_1	...	p_k

Тобто X набуває значень x_i з ймовірностями p_i.

Означення. Таблиця (3.1) в статистиці називається ***теоретичним розподілом*** випадкової величини X.

Як перевірити, що випадкова величина X справді має цей розподіл? Проводимо n незалежних випробувань. В результаті дістаємо наступну таблицю.

Таблиця 3.2

x_1	...	x_k
n_1	...	n_k
w_1	...	w_k

Тут n_i – кількість результатів випробувань, в яких ми дістали x_i, а $w_i = \frac{n_i}{n}$.

Означення. Таблиця (3.2) називається ***емпіричним розподілом***.

Природно гіпотезу про те, що випадкова величина X має розподіл (3.1) приймати, коли числа p_i та w_i між собою близькі і відхиляти – коли великі. Але, що має бути малим

$$\max_{1 \le i \le k} |p_i - w_i|,$$

чи

$$\sum_{i=1}^{k} |p_i - w_i|,$$

чи може

$$\sum_{i=1}^{k} |p_i - w_i|^2,$$

чи може ще щось? Потрібно обрати критерій.

Загальний випадок

Якщо X – довільна випадкова величина, то можна запропонувати наступну «дискретизацію» випадкової величини X. А саме, ділимо множину значень випадкової величини X на k проміжків $I_1, \dots, I_k$ і нехай

$$p_i = \mathbf{P}\{X \in I_i\}, \qquad i = 1, \dots, k.$$

Дістаємо таблицю.

I_1	...	I_k
p_1	...	p_k

Далі проводимо n незалежних випробувань. У результаті отримуємо таблицю

I_1	...	I_k
n_1	...	n_k
w_1	...	w_k

Тут n_i – кількість результатів випробувань, які потрапили до інтервалу I_i, а $w_i = \frac{n_i}{n}$.

Проте, такий підхід слід уточнити. Яку гіпотезу тут потрібно перевіряти. У загальному випадку розподіл випадкової величини X характеризується її ***функцією розподілу***

$$F(x) = \mathbf{P}\{X < x\}, \qquad x \in \mathbf{R}.$$

Отже, нульовою гіпотезою H_0 тут (природно) є така: випадкова величина X має функцію розподілу $F(x)$. Як перевірити експериментально правильність цієї гіпотези? Проводимо n спостережень. Ділимо множину отриманих значень

$$\{x_1, \dots, x_n\} \qquad (3.5)$$

на інтервали $I_1, \dots, I_k$, а для довільного $x \in I_i$ позначаємо через n_x кількість значень з набору (3.5), які потрапили до $I_1 \cup \dots \cup I_i$. Нагадаємо означення емпіричної функції розподілу.

Означення. Функція

$$F_n(x) = \frac{n_x}{n}$$

називається ***емпіричною функцією розподілу*** випадкової величини X.

Неформально тоді перевірка гіпотези про те, що випадкова величина X має функцію розподілу $F(x)$ виглядає так: порівнюємо $F(x)$ і $F_n(x)$ (рис.3.9).

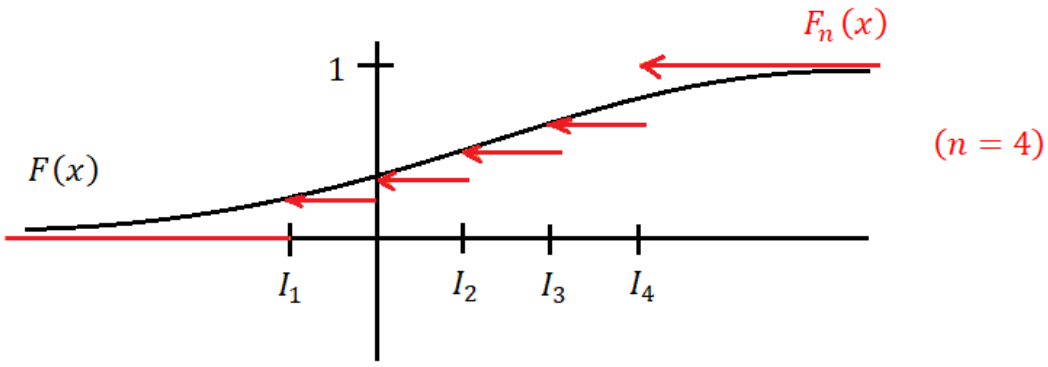


Рис. 3.9

Якщо значення $F(x)$ та $F_n(x)$ «близькі», то гіпотезу H_0 приймаємо, а якщо далекі – відхиляємо. Але, що означає «близькі» чи «далекі»? Потрібен критерій.

Критерій χ^2 (або критерій Пірсона)

Опишемо цей критерій для дискретної випадкової величини X, яка набуває значень $x_1, \dots, x_k$. Гіпотеза H_0 тут має вигляд:

Розподілом випадкової величини X є таблиця

x_1	...	x_k
p_1	...	p_k

1. Для перевірки цієї гіпотези беремо n незалежних копій $X_1, \dots, X_n$ випадкової величини X і вводимо нові випадкові величини.

$$N_i = \text{card}\{X_i : X_j = x_i\}, \qquad i = 1, \dots, k.$$

Звичайно

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

Значення випадкової величини N_i називаються ***емпіричними частотами***. Побудуємо таблицю

x_1	...	x_k
N_1	...	N_k
$\frac{N_1}{n}$	...	$\frac{N_k}{n}$

Якщо гіпотеза H_0 правильна, то згідно ЗВЧ $[6,\ \text{п. } 1.22.1]$, для кожного i

$$\frac{N_i}{n} \underset{n\to\infty}{\longrightarrow} p_i = \mathbf{P}\{X = x_i\}.$$

2. Вибір критерію. Розглянемо випадкову величину

$$Z_n = \sum_{i=1}^{k} \frac{(N_i - n \cdot p_i)^2}{n \cdot p_i}.$$

Можна показати, що коли гіпотеза H_0 правильна, то при $n \to \infty$ $Z_n \to \chi^2$ за ймовірністю, де випадкова величина χ^2 має розподіл хі-квадрат з $(k-1)$ ступенем свободи. [10, с. 623]. Тому, для великих n χ_n^2 можна замінити на χ^2.

3. Задавши рівнем значущості α, з рівняння

$$\mathbf{P}\{\chi^2 > z\} = \alpha$$

знаходимо критичну точку z_α .

4. За результатами спостережень знаходимо спостережуване значення критерію

$$z_\text{o} = \sum_{i=1}^{k} \frac{(n_i - n \cdot p_i)^2}{n \cdot p_i},$$

Тут n_i – значення випадкової величини N_i.

5. Якщо $z_\text{o} < z_\alpha$ – гіпотеза приймається, інакше – відхиляється.

Зауваження. За формою z_o найбільше схоже на відхилення

$$\sum_{i=1}^{k} |p_i - w_i|^2 = \sum_{i=1}^{k} \left(\frac{n_i - n \cdot p_i}{n}\right)^2.$$

Але останнє відхилення, точніше

$$\sum_{i=1}^{k}\left(\frac{N_i - n\cdot p_i}{n}\right)^2 \to 0 \text{ за ймовірністю.}$$

Тому критерієм бути не може.

Критерій Колмогорова

Він ґрунтується на понятті емпіричної функції розподілу. А саме, для випадкової величини X з функцією розподілу $F(x)$ і набором незалежних копій $X_1, \dots, X_n$, її емпірична функція розподілу $\hat{F}_n(x)$ визначається формулою

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i = x\}}, \qquad x \in \mathbf{R},$$

де для довільної випадкової величини Z і $x \in \mathbf{R}$

$$I_{\{Z=x\}}(\omega) := \begin{cases} 1, & Z(\omega) \le x \\ 0, & Z(\omega) > x \end{cases}, \ \omega \in \Omega.$$

Теорема. Нехай $F(x)$ – неперервна функція розподілу випадкової величини X, $\hat{F}_n(x)$ – її емпірична функція розподілу. Позначимо

$D\left(\hat{F}_n, F\right) = D_n = \sup_{x\in R}\left|\hat{F}_n(x) - F(x)\right| \to 0$ за ймовірністю.

Тоді $\forall z \in \mathbf{R}$

$$\lim_{n\to\infty} \mathbf{P}\left\{\sqrt{n}D_n < z\right\} = \begin{cases} 0, & z < 0 \\ 1 - 2\sum_{k=1}^{\infty}(-1)^{n-1}e^{-2n^2z^2}, & z \ge 0 \end{cases} = K(z).$$

Алгоритм застосування теореми. Нехай перевіряється гіпотеза H_0 про те, що випадкова величина X має функцію розподілу $F(x)$.

1. Беремо вибірку $x_1, \dots, x_k$ і знаходимо реалізацію $F_n(x)$ емпіричної функції розподілу $\hat{F}_n(x)$, а потім спостережуване значення критерію

$$z_{\text{o}} = \sqrt{n}\cdot \sup_{1\le i\le n}|F_n(x_i) - F(x_i)|.$$

2. Оскільки для великих n функцію $\mathbf{P}\left\{\sqrt{n}D_n < z\right\}$ можна замінити на $K(z)$, то, задавшись рівнем значущості α з рівняння

$$K(z) = 1 - \alpha$$

знаходимо критичну точку z_α.

3. Якщо $z_{\text{o}} < z_\alpha$ – гіпотезу H_0 приймаємо, а якщо $z_{\text{o}} > z_\alpha$ – відхиляємо.

Перевірка гіпотези про щільність розподілу

Нагадаємо, що коли функція розподілу $F(x)$ випадкова величина X має похідну $f(x) = F'(x)$, то ця похідна називається ***щільністю розподілу*** випадкової величини X. Для перевірки гіпотези про щільність розподілу можна скористатись або критерієм Колмогорова, або критерієм Пірсона. У першому випадку знаючи гіпотетичну щільність $f(x)$ знаходимо функцію розподілу $F(x) = \int_{-\infty}^{x} f(y)dy$ і до $F(x)$ застосовуємо критерій Колмогорова. У другому випадку, застосовуємо метод дискретизації: переходимо до дискретного розподілу і до цього дискретного розподілу застосовуємо критерій Пірсона.

Другий випадок опишемо докладніше.

1. Розбиваємо множину значень випадкової величини X на k частин $I_1, I_2, \dots, I_k$

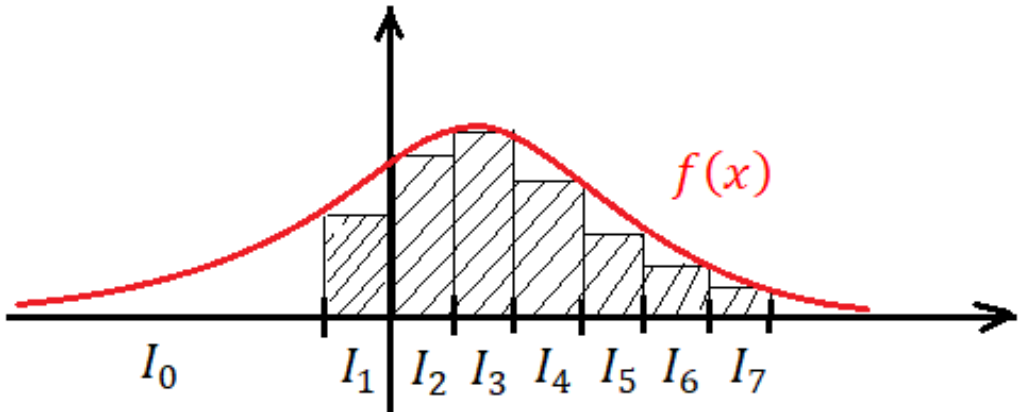


Рис. 3.10

2. Проводимо спостереження. Отримуємо значення $x_1, \dots, x_n$ випадкової величини X. нехай n_i – кількість результатів спостережень, що потрапили до проміжку I_i. Це – емпіричні частоти. Результати спостережень запишемо у вигляді таблиці:

I_1	I_2	I_3	I_4	I_5	I_6	
n_1	n_2	n_3	n_4	n_5	n_6	
$\frac{n_1}{n}$	$\frac{n_2}{n}$	$\frac{n_3}{n}$	$\frac{n_4}{n}$	$\frac{n_5}{n}$	$\frac{n_6}{n}$	←емпіричні ймовірності

3. Обчислюємо теоретичні ймовірності

$$p_i = \int_{I_i} f(x)dx\,, \quad i = 1, \dots, k.$$

4. Знаходимо спостережуване значення

$$z_o = \sum_{i=1}^{n} \frac{(n_i - n \cdot p_i)^2}{n \cdot p_i}\,.$$

5. Задавшись рівнем значущості α і користуючись таблицями розподілу хi-квадрат з $(k-1)$ ступенем свободи з рівняння

$$\mathbf{P}\{\chi^2 > z\} = \alpha$$

Знаходимо критичну точку z_α .

6. Порівнюємо. Якщо $z_{\text{o}} < z_\alpha$ – гіпотезу H_0 приймаємо, в супротивному разі – відхиляємо.

3.6. Гіпотеза про ймовірність події

Окрім гіпотези про розклад ймовірностей, чи про параметри цього розкладу, можна також перевіряти гіпотези про ймовірність того, що дана випадкова величина набуде значень в певній множині $A \subset \mathbf{R}$.

Наприклад. якщо X – вага одного яйця, то це може бути гіпотеза проте, що ймовірність вибрати яйце вагою менше 40 г, дорівнює 0,1.

Отже, перевіряємо гіпотезу

$$H_0\colon \mathbf{P}\{X \in A\} = p_0. \tag{3.6}$$

1. Проводимо n випробувань, тобто беремо n незалежних копій $X_1, \dots, X_n$ випадкової величини X. Позначимо

$$M_n = \operatorname{card}\{X_i\colon X_i \in A\}.$$

Це кількість випадкових величин X_i, значення яких потрапляє до A. M_n є випадковою величиною.

2. Для перевірки гіпотези (3.6) використовують критерій

$$N(0,1) \underset{\text{ЦГТ}}{\leftarrow} U = \frac{\left(\frac{M_n}{n} - p_0\right)\sqrt{n}}{\sqrt{p_0 \cdot q_0}}, \qquad \text{де } q_0 = 1 - p_0. \tag{3.7}$$

Обґрунтування критерію. Звичайно, задаємося рівнем значущості α. Внаслідок ЗВЧ [6, п.1.21.1], коли гіпотеза H_0 правильна, то $\frac{M_n}{n} \to p_0$ при $n \to \infty$ за ймовірністю.

Тобто, якщо гіпотеза H_0 правильна, то

$$\left|\frac{M_n}{n} - p_0\right| \tag{3.8}$$

буде маленьким при великих n. Чому б не провести експеримент, порахувати і сказати: Якщо величина (3.8) менше α – гіпотезу приймаємо, а якщо ні – відхиляємо? Щоб зрозуміти, що так робити не можна розглянемо крайній випадок. Нехай $\alpha = 0{,}05$, $p_0 = 0{,}05$, а $M_n = 0$. Тоді $|0 - 0{,}05| = 0{,}05$.

Отже, ми зробимо висновок, що гіпотеза H_0 правильна при тому, що жоден експеримент її не підтверджує. Тому (3.8) потрібно ділити на середнє квадратичне відхилення (точність вимірювання).

Висновок. За критерій перевірки гіпотези H_0 не можна брати величину $\frac{M_n}{n} - p_0$, потрібно брати

$$\frac{\frac{M_n}{n} - p_0}{\boldsymbol{\sigma}\left(\frac{M_n}{n}\right)}.$$

Порахуємо це середнє квадратичне відхилення. Нехай $Y_i, i = 1, \dots, n$ задана рівністю

$$Y_i = \begin{cases} 1, & X_i \in A \\ 0, & X_i \notin A \end{cases}.$$

При справедливості гіпотези H_0

$$\mathbf{P}\{Y_i = 1\} = p_0$$
$$\mathbf{P}\{Y_i = 0\} = q_0 = 1 - p_0.$$

Окрім того, і це важливо,

$$M_n = \sum_{i=1}^{n} Y_i,$$
$$\mathbf{M}Y_i = 1 \cdot p_0 + 0 \cdot q_0 = p_0,$$
$$\mathbf{D}Y_i = 1^2 \cdot p_0 + 0^2 \cdot q_0 - (MY_i)^2 = p_0 - (p_0)^2 = p_0 \cdot q_0,$$
$$\mathbf{D}\left(\frac{M_n}{n}\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathbf{D}Y_i = \frac{n \cdot p_0 \cdot q_0}{n^2},$$
$$\boldsymbol{\sigma}\left(\frac{M_n}{n}\right) = \sqrt{\frac{p_0 \cdot q_0}{n}}.$$

Тому, за критерій перевірки гіпотези H_0 беремо випадкову величину

$$U_n = \frac{\frac{M_n}{n} - p_0}{\sqrt{\frac{p_0 \cdot q_0}{n}}}.$$

Згідно з ЦГТ, послідовність U_n збігається за розподілом до стандартної нормальної випадкової величини.

Алгоритм. Отже для перевірки гіпотези H_0 використовуємо критерій (3.7). Далі все технічно схоже на перевірку гіпотези про математичне сподівання.

3. А саме, для однієї з альтернативних гіпотез: H_1: $\{p < p_0\}$, $\{p > p_0\}$, $\{p \neq p_0\}$ будується відповідно лівостороння, правостороння чи двостороння критична область (рис. 3.11).

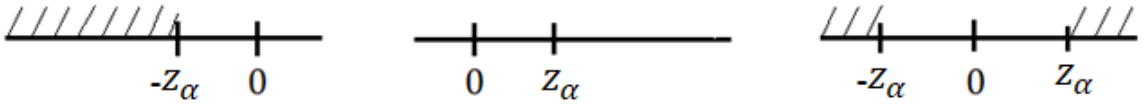


Рис. 3.11

Критична точка вибирається з рівняння

$$\Phi(z) = \frac{1-2\alpha}{2} \qquad \text{або} \qquad \Phi(z) = \frac{1-\alpha}{2}$$

для односторонньої чи двосторонньої критичної області відповідно. У цих рівняннях α – рівень значущості, а Φ – функція Лапласа (значення $\Phi(z)$ знаходяться у Додатку 1).

4. Проводимо n спостережень. Отримуємо значення

$$\{x_1, \dots, x_n\} \tag{3.9}$$

випадкових величин $X_1, \dots, X_n$ відповідно. Знаходимо кількість m тих значень серед набору (3.9), які потрапили до A. Обчислюємо спостережуване значення

$$z_{\text{o}} = \frac{\left(\frac{m}{n} - p_0\right)\sqrt{n}}{\sqrt{p_0 \cdot q_0}}.$$

Якщо число z_o потрапило до критичної області гіпотезу відхиляємо, якщо ні – приймаємо.

3.7. Порівняння двох імовірностей

У попередньому підрозділі ми розглянули питання про перевірку гіпотези про те, що ймовірність певної події дорівнює певному заданому числу. А тепер розглянемо гіпотезу про рівність двох імовірностей. Точніше, нехай маємо дві незалежні випадкові величини X та Y і множину $A \subset \mathbf{R}$. Потрібно перевірити гіпотезу

$$p_1 = \mathbf{P}\{X \in A\} = \mathbf{P}\{Y \in A\} = p_2.$$

Опишемо задачу докладніше. Нехай проводиться n_1 і n_2 незалежних експериментів. Тобто, маємо дві незалежні випадкові величини і ми беремо їхні незалежні копії $X_1, \dots, X_{n_1}$ та $Y_1, \dots, Y_{n_2}$. Нехай $A \subset \mathbf{R}$ – борелівська множина.

Покладемо

$$p_1 = \mathbf{P}\{X \in A\}, \qquad p_2 = \mathbf{P}\{Y \in A\}.$$

Основна гіпотеза тут має вигляд

$$H_0: p_1 = p_2.$$

Для її перевірки природно використовувати наступний критерій

$$\frac{\frac{M_1}{n_1} - \frac{M_2}{n_2}}{\boldsymbol{\sigma}\left(\frac{M_1}{n_1} - \frac{M_2}{n_2}\right)},$$

де $M_1 = \mathrm{card}\{X_i : X_i \in A\},\ M_2 = \mathrm{card}\{Y_i : Y_i \in A\}$.

Знайдемо відповідне середнє квадратичне відхилення. Як ми вже рахували,

$$\mathbf{M}\left(\frac{M_1}{n_1}\right) = p_1, \qquad \mathbf{M}\left(\frac{M_2}{n_2}\right) = p_2,$$

а

$$\mathbf{D}\left(\frac{M_1}{n_1} - \frac{M_2}{n_2}\right) = \frac{\mathbf{D}M_1}{n_1^2} + \frac{\mathbf{D}M_2}{n_2^2} = \frac{n_1 p_1(1-p_1)}{n_1^2} + \frac{n_2 p_2(1-p_2)}{n_2^2} \qquad (3.10)$$

Якщо гіпотеза $H_0: p_1 = p_2 = p$ правильна, то

$$(3.10) = p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right).$$

Отже, якщо гіпотеза H_0 правильна, то можемо вважати, що проведено $n_1 + n_2$ спостережень, в яких подія A настала $M_1 + M_2$ рази. Тому (невідому) ймовірність p можна наблизити оцінкою

$$\frac{M_1 + M_2}{n_1 + n_2}.$$

В результаті, критерій набуде вигляду

$$U_{n_1 n_2} = U = \frac{\frac{M_1}{n_1} - \frac{M_2}{n_2}}{\sqrt{\frac{M_1 + M_2}{n_1 + n_2}\left(1 - \frac{M_1 + M_2}{n_1 + n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad (3.13)$$

Як і раніше, альтернативною гіпотезою H_1 може бути $\{p_1 < p_2\}$, $\{p_1 > p_2\}$, $\{p_1 \neq p_2\}$. Тоді критична область будується точно так само, як і в попередньому підрозділі.

3.8. Приклади до Розділу 3

Приклад 1. Перевірка гіпотези H_0: $\mathbf{M}X = a$ з альтернативою H_1: $\mathbf{M}X > a$

Діаметри кульок є нормальною випадковою величиною з параметрами $(a, 4)$ [мм], де середній діаметр кульок a – невідомий. Було проведено $n = 100$ вимірювань різних кульок і отримано, що $\bar{x} = 225$ мм.

Перевірити гіпотезу H_0: $a = 240$ мм з альтернативною гіпотезою H_1: $a > 240$ мм, якщо рівень значущості $\alpha = 0{,}01$.

Розв'язання.

Оскільки H_1: $a > 240$ мм, то будуємо правобічну критичну область. Записуємо рівняння

$$\Phi(z) = \frac{1 - 2\alpha}{2} = 0{,}49.$$

З нього знаходимо $z_\alpha = 2{,}34$ (за Додатком 1).

Обчислюємо спостережуване значення критерію

$$z_\text{o} = \frac{\bar{x} - a}{\sigma/\sqrt{n}} = \frac{225 - 240}{4/\sqrt{100}} = -37{,}5.$$

Оскільки $z_\text{o} = -37{,}5 < 2{,}34 = z_\alpha$, то гіпотезу H_0 приймаємо.

Приклад 2. Перевірка правильності гіпотези H_0 про рівність двох математичних сподівань

Нехай $n = m = 100$, $\bar{x} = 14{,}465$, $\bar{y} = 17{,}4$, $\alpha = 0{,}01$. Перевірити гіпотезу H_0: $\mathbf{M}X = \mathbf{M}Y$ при альтернативній гіпотезі H_1: $\mathbf{M}X > \mathbf{M}Y$, якщо відомі $\mathbf{D}\text{X} = 10$, $\mathbf{D}\text{Y} = 15$.

Розв'язання.

Критичну точку z_α знаходимо з рівняння

$$\Phi(z) = \frac{1 - 2\alpha}{2} = \frac{1 - 2 \cdot 0{,}01}{2} = 0{,}49.$$

Звідси $z_\alpha = 2{,}34$ (за Додатком 1).

Обчислюємо спостережуване значення критерію

$$z_\text{o} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\mathbf{D}\text{X}}{n} + \frac{\mathbf{D}\text{Y}}{m}}} = \frac{14{,}465 - 17{,}4}{\sqrt{\frac{10}{100} + \frac{15}{100}}} = \frac{-2{,}935}{0{,}5} = -5{,}87.$$

Оскільки $z_\text{o} = -5{,}87 < 2{,}34 = z_\alpha$, то гіпотеза H_0 приймається.

Приклад 3. Перевірка правильності гіпотези H_0 про рівність двох математичних сподівань. Дисперсії $\mathbf{D}X$ та $\mathbf{D}Y$ невідомі і обсяги вибірок великі ($n, m > 40$)

За допомогою двох радіолокаційних приладів вимірювалася по кілька разів відстань до певного об'єкту. Перевірити гіпотезу $H_0: \mathbf{M}X = \mathbf{M}Y$ при альтернативі $\mathbf{M}X > \mathbf{M}Y$ і рівні значущості $\alpha = 0{,}01$. Припускаємо, що помилки вимірювань мають нормальний розподіл з невідомими математичними сподіваннями й дисперсіями. Нехай проведено по $n = m = 100$ вимірювань і обчислено $\bar{x} = 193{,}56$ км та $\bar{y} = 201{,}75$ км, $s(X) = 4{,}4$ та $s(Y) = 4{,}0$.

Розв'язання.

З рівняння

$$\Phi(z) = \frac{1 - 2\alpha}{2}$$

знаходимо $z_\alpha = 2{,}34$ (за Додатком 1).

Знаходимо спостережуване значення критерію:

$$z_\text{o} = \frac{193{,}56 - 201{,}75}{\sqrt{\frac{99}{198}(4{,}4^2 + 4^2)\left(\frac{1}{100} + \frac{1}{100}\right)}} = -28{,}24.$$

Оскільки $z_\text{o} = -27{,}24 < 2{,}34 = z_\alpha$, то гіпотезу H_0 приймаємо.

Приклад 4. Перевірка правильності гіпотези про рівність дисперсій

Під час дослідження стабільності температури в термостаті дістали такі результати: 21,2; 21,8; 21,3; 21,0; 21,4; 21,3. З метою стабілізації температури було використано удосконалений пристрій, після цього заміри температури показали такі результати: 37,7; 37,6; 37,6; 37,4. Чи можна за рівня значущості $\alpha = 0{,}01$ вважати використання удосконаленого пристрою до стабілізатора температури ефективним?

Розв'язання.

Очевидно, що ефективність стабілізаторів без удосконаленого пристрою і з ним залежить від дисперсій вимірюваних ними температур. Отже, задача звелась до порівняння двох дисперсій.

Обчислимо виправлені вибіркові дисперсії

$$\bar{x} = \frac{21{,}2 + 21{,}8 + 21{,}3 + 21{,}0 + 21{,}4 + 21{,}3}{6} = 21{,}33;$$

$$D_X = \frac{21{,}2^2 + 21{,}8^2 + 21{,}3^2 + 21{,}0^2 + 21{,}4^2 + 21{,}3^2}{6} - 21{,}33^2 = 0{,}059;$$

$$s_X^2 = \frac{6}{5} \cdot 0{,}059 = 0{,}071.$$

$$\bar{y} = \frac{37{,}7 + 37{,}6 + 37{,}6 + 37{,}4}{4} = 37{,}58;$$

$$D_Y = \frac{37{,}7^2 + 37{,}6^2 + 37{,}6^2 + 37{,}4^2}{4} - 37{,}58^2 = 0{,}012;$$

$$s_Y^2 = \frac{4}{3} \cdot 0{,}012 = 0{,}016.$$

Знаходимо спостережуване значення критерію (більшу з дисперсій записуємо у чисельник):

$$z_{\text{о}} = \frac{s_X^2}{s_Y^2} = \frac{0{,}071}{0{,}016} = 4{,}46.$$

Критичну точку знаходимо за таблицею (Додаток 8) відповідно до заданого рівня значущості $\alpha = 0{,}01$ і числа ступенів свободи $k_1 = 6 - 1 = 5$, $k_2 = 4 - 1 = 3$, $z_\alpha\ (\alpha = 0{,}01; k_1 = 5;\ k_2 = 3) = 28{,}2$.

Оскільки $z_{\text{о}} = 4{,}46 < 28{,}2 = z_\alpha$, то гіпотезу H_0 приймаємо.

Приклад 5. Критерій χ^2 (критерій Пірсона)

В експериментах з селекцією гороху біолог Георг Мендель спостерігав частоти різного вигляду насіння, що одержані при схрещуванні рослин з круглими жовтими (КЖ), зморшкуватими жовтими (ЗЖ), круглими зеленими (КЗ) і зморшкуватими зеленими (ЗЗ) насінинами. У результаті експерименту були отримані такі результати:

Насіння	Частота n_i	Ймовірність p_i
КЖ	315	9/16
ЗЖ	101	3/16
КЗ	108	3/16
ЗЗ	32	1/16
Всього	$n = 556$	1

Ймовірності визначаються теорією Менделя. При рівні значущості $\alpha = 0{,}05$ перевірити гіпотезу про узгодження експериментальних даних із теоретичними ймовірностями.

Розв'язання.

Знаходимо спостережуване значення критерію

$$z_{\text{о}} = \sum_{i=1}^{k} \frac{(n_i - n \cdot p_i)^2}{n \cdot p_i}$$
$$= \frac{(315 - 556 \cdot 9/16)^2}{556 \cdot 9/16} + \frac{(101 - 556 \cdot 3/16)^2}{556 \cdot 3/16} + \frac{(108 - 556 \cdot 3/16)^2}{556 \cdot 3/16}$$
$$+ \frac{(32 - 556 \cdot 1/16)^2}{556 \cdot 1/16} = 0{,}47.$$

За рівнем значущості $\alpha = 0{,}05$ і числом ступенів свободи $4 - 1 = 3$ за таблицею розподілу χ^2 (Додаток 4) з рівняння

$$\mathbf{P}\{\chi^2 > z\} = 0{,}05$$

знаходимо критичну точку $z_\alpha = 7{,}8$.

Оскільки

$$z_{\text{о}} = 0{,}47 < 7{,}8 = z_\alpha\,,$$

то нульову гіпотезу про узгодження експериментальних даних із теоретичними ймовірностями приймаємо.

Приклад 6. Критерій χ^2 (критерій Пірсона). Дискретний розподіл

Як показали метеорологічні спостереження, в теплий період року (квітень-жовтень) у Херсоні спостерігалися такі кількості днів (к.д.) зі зливами (1976-2015 рр.):

Рік	1976	1977	1978	1979	1980	1981	1982	1983
к.д.	0	0	3	4	5	3	3	1
Рік	1984	1985	1986	1987	1988	1989	1990	1991
к.д.	1	4	7	2	3	1	5	2
Рік	1992	1993	1994	1995	1996	1997	1998	1999
к.д.	2	1	4	3	2	2	2	3
Рік	2000	2001	2002	2003	2004	2005	2006	2007
к.д.	4	5	4	6	6	5	3	1
Рік	2008	2009	2010	2011	2012	2013	2014	2015
к.д.	4	4	2	3	0	3	4	3

Аналіз цих даних свідчить про те, що зливи у південному регіону України є рідким явищем. Тому доцільно емпіричний розподіл апроксимувати розподілом Пуассона.

Перевірити, чи узгоджується даний розподіл із розподілом Пуассона.

Розв'язання.

Позначимо кількість днів зі зливою через x_i, а відповідні частоти через n_i. Ці характеристики, які, по суті, є емпіричним розподілом, а також відповідні розрахунки зведемо у наступну таблицю.

Також у таблиці підрахуємо ймовірності p_i та теоретичні частоти $n \cdot p_i$ за формулою розподілу Пуассона:

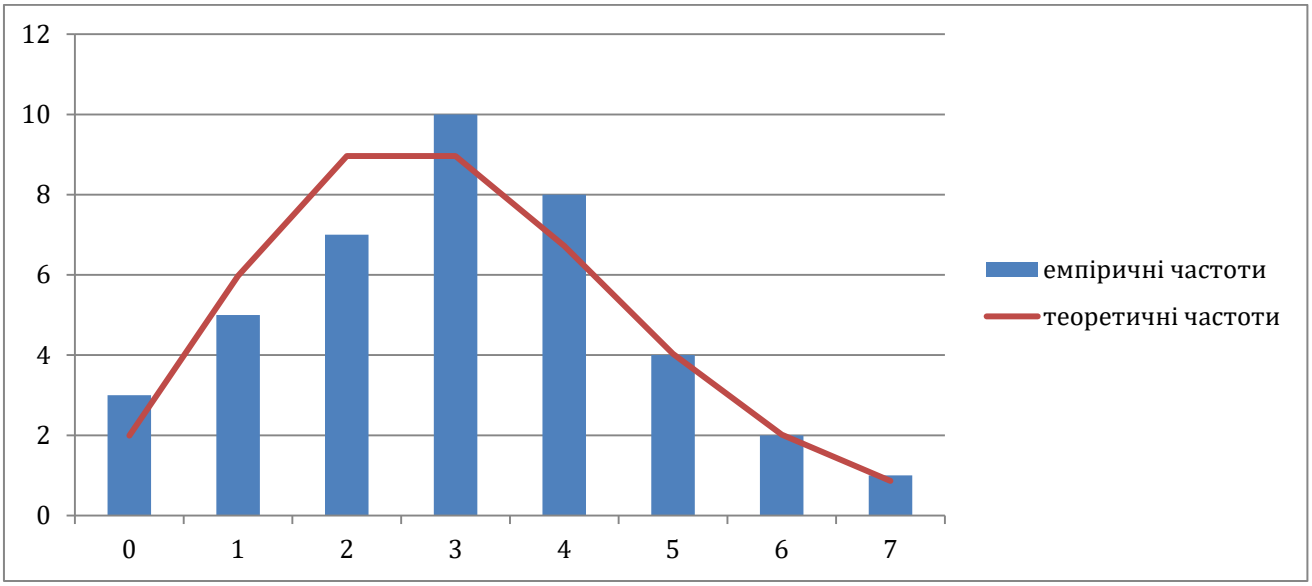
$$P(X = x_i) = \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \quad (x_i = 0,1,2, \dots),$$

де

$$\lambda = \bar{x} = \frac{1}{n} \sum_{i=1}^{8} x_i \cdot n_i = \frac{120}{40} = 3.$$

x_i	n_i	$x_i \cdot n_i$	p_i	$n \cdot p_i$
0	3	0	0,05	2
1	5	5	0,15	6
2	7	14	0,22	9
3	10	30	0,22	9
4	8	32	0,17	7
5	4	20	0,10	4
6	2	12	0,05	2
7	1	7	0,02	1
Сума:	**40**	**120**	**1,00**	**40**

Візуально порівняємо емпіричні n_i та теоретичні $n \cdot p_i$ частоти.



Далі виконаємо перевірку статистичної гіпотези про відповідність емпіричного розподілу

i	n_i	$n \cdot p_i$	$\frac{(n_i - n \cdot p_i)^2}{n \cdot p_i}$
0	3	1,99	0,51
1	5	5,97	0,16
2	7	8,96	0,43
3	10	8,96	0,12
4	8	6,72	0,24
5	4	4,03	0,00
6	2	2,02	0,00
7	1	0,86	0,02
8	0	0,32	0,32
		$z_{\text{o}} = \sum_{k=1}^{7} \frac{(n_i - n \cdot p_i)^2}{n \cdot p_i} =$	1,48

закону Пуассона за допомогою критерію χ^2.

Оскільки закон Пуассона є одно параметричним,то кількість ступенів свободи рівна $k - 2 = 9 - 2 = 7$. Для рівня значущості $\alpha = 0{,}05$ з Додатку 4 маємо: критична точка $\chi^2_{0{,}05}(7) = 14{,}07$. Таким чином, $z_{\text{o}} = 1{,}48 < \chi^2_{0{,}05}(7) = 14{,}07$.

Отже, гіпотеза H_0 про незначущість розбіжностей між емпіричними інтервальними частотами і відповідними частотами закону Пуассона з імовірністю 95% не відхиляється і емпіричний розподіл (кількість днів зі зливами в теплий період року у Херсоні) можна апроксимувати законом Пуассона з оцінкою параметра $\lambda = \bar{x} = 3$.

Приклад 7. Критерій χ^2 (критерій Пірсона). Неперервний розподіл

За заданим інтервальним статистичним розподілом випадкової величини X — маса новонароджених дітей — при рівні значущості $\alpha = 0{,}05$ перевірити правильність H_0 про нормальний закон розподілу ознаки X — маси новонароджених дітей.

Інтервали	I_1	I_2	I_3	I_4	I_5	I_6	I_7
Вага дітей	1—1,5	1,5—2	2—2,5	2,5—3	3—3,5	3,5—4	4—4,5
n_i	10	20	50	35	28	15	12

Розв'язання.

Для розрахунку числових характеристик перейдемо до дискретного розподілу, визначивши середини інтервалів:

Вага дітей	1,25	1,75	2,25	2,75	3,25	3,75	4,25
n_i	10	20	50	35	28	15	12

Підрахуємо числові характеристики отриманого розподілу:

$$\bar{x} = \frac{1{,}25 \cdot 10 + 1{,}75 \cdot 20 + \cdots + 4{,}25 \cdot 12}{170} = 2{,}67;$$

$$D = \frac{1{,}25^2 \cdot 10 + 1{,}75^2 \cdot 20 + \cdots + 4{,}25^2 \cdot 12}{170} - (2{,}67)^2 = 0{,}61;$$

$$\sigma = \sqrt{0{,}61} = 0{,}78.$$

Теоретичні частоти обчислюються за формулою

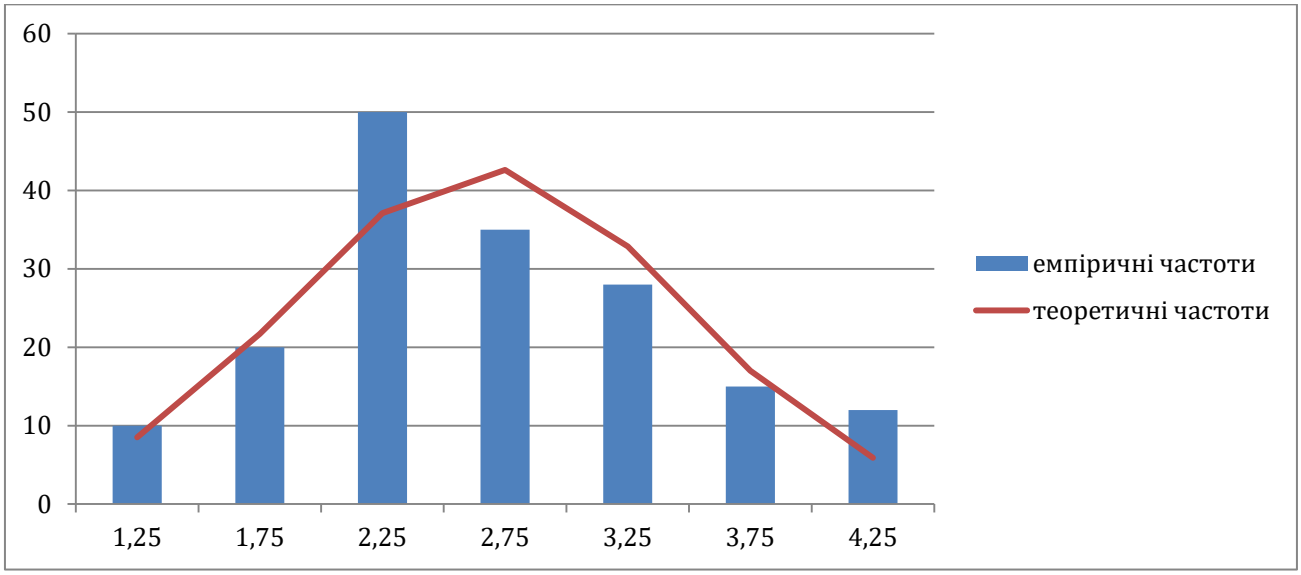
$$n \cdot p_i = n \cdot \left(\Phi\left(\frac{x_{i+1} - \bar{x}}{\sigma}\right) - \Phi\left(\frac{x_i - \bar{x}}{\sigma}\right) \right).$$

Розрахунки для теоретичних частот проведемо у наступній таблиці

x_i	x_{i+1}	n_i	$\frac{x_i - \bar{x}}{\sigma}$	$\frac{x_{i+1} - \bar{x}}{\sigma}$	$\Phi\left(\frac{x_i - \bar{x}}{\sigma}\right)$	$\Phi\left(\frac{x_{i+1} - \bar{x}}{\sigma}\right)$	$n \cdot p_i = n \cdot \left(\Phi\left(\frac{x_{i+1} - \bar{x}}{\sigma}\right) - \Phi\left(\frac{x_i - \bar{x}}{\sigma}\right)\right)$	$\frac{(n_i - n \cdot p_i)^2}{n \cdot p_i}$
1	1,5	10	-2,15	-1,51	-0,4842	-0,4342	8,5	0,26
1,5	2	20	-1,51	-0,87	-0,4342	-0,3066	21,7	0,13
2	2,5	50	-0,87	-0,22	-0,3066	-0,0882	37,1	4,47
2,5	3	35	-0,22	0,42	-0,0882	0,1625	42,6	1,36
3	3,5	28	0,42	1,06	0,1625	0,3558	32,9	0,72
3,5	4	15	1,06	1,70	0,3558	0,4558	17,0	0,24
4	4,5	12	1,70	2,35	0,4558	0,4905	5,9	6,31
							$z_o = \sum_{i=1}^{7} \frac{(n_i - n \cdot p_i)^2}{n \cdot p_i} =$	13,49

Для знаходження значень функції Лапласа використовуємо Додаток 1.

На рисунку зобразимо емпіричні і підраховані теоретичні частоти. Бачимо суттєві відмінності між ними на 3-му і 7-му інтервалах.



Оскільки нормальний закон розподілу має два параметри, то кількість ступенів свободи рівна $k-2-1=7-3=4$ (k – кількість інтервалів). Для рівня значущості $\alpha=0{,}01$ з Додатку 4 маємо критичну точку $\chi^2_{0{,}05}(4)=13{,}28$. Таким чином, $z_{\text{o}}=13{,}49>\chi^2_{0{,}05}(4)=13{,}28$. Тому ряд маси новонароджених дітей не є нормально розподіленим.

Приклад 8. Перевірка гіпотези про щільність розподілу

Середній урожай зерна з 1 гектара є випадковою величиною X. Нехай за даними 100 спостережень ми дістали такий інтервальний розподіл емпіричних імовірностей

Інтервали	I_1	I_2	I_3	I_4	I_5	I_6	I_7
Урожайність, ц/га	13,5-14,5	14,5-15,5	15,5-16,5	16,5-17,5	17,5-18,5	18,5-19,5	19,5-20,5
Емпіричні ймовірності, $\frac{n_i}{n}$	0,06	0,1	0,18	0,28	0,2	0,12	0,06

За допомогою критерію Пірсона перевірити гіпотезу H_0: щільність ймовірності випадкової величини X має вигляд

$$f(x)=\frac{1}{\sqrt{2\pi}\cdot 1{,}5}e^{-\frac{(x-17)^2}{2\cdot(1{,}5)^2}}$$

з рівнем значущості $\alpha=0{,}05$.

Розв'язання.

Розбиття та емпіричні ймовірності вже маємо.

Обчислюємо теоретичні ймовірності $p_i = \int_{I_i} f(x)dx$ через середини інтервалів:

$$p_1 = \frac{1}{\sqrt{2\pi}\cdot 1{,}5} e^{-\frac{(14-17)^2}{2\cdot(1{,}5)^2}} = 0{,}04, \qquad p_2 = \frac{1}{\sqrt{2\pi}\cdot 1{,}5} e^{-\frac{(15-17)^2}{2\cdot(1{,}5)^2}} = 0{,}11, \ldots$$

$$p_7 = \frac{1}{\sqrt{2\pi}\cdot 1{,}5} e^{-\frac{(20-17)^2}{2\cdot(1{,}5)^2}} = 0{,}04.$$

Отримаємо таблицю:

Інтервали	I_1	I_2	I_3	I_4	I_5	I_6	I_7
P_i	0,04	0,11	0,21	0,28	0,21	0,11	0,04

За таблицею розподілу хі-квадрат (Додаток 4) з $k = 7 - 1 = 6$ ступенями свободи з рівняння

$$\mathbf{P}\{\chi^2 > z\} = 0{,}05$$

знаходимо критичну точку $z_\alpha = 12{,}6$.

За формулою знаходимо спостережуване значення критерію

$$z_\text{o} = \sum_{i=1}^{7} \frac{(n_i - n\cdot p_i)^2}{n\cdot p_i}$$

$$= \frac{(100\cdot 0{,}06 - 100\cdot 0{,}04)^2}{100\cdot 0{,}04} + \frac{(100\cdot 0{,}1 - 100\cdot 0{,}11)^2}{100\cdot 0{,}11} + \cdots$$

$$+ \frac{(100\cdot 0{,}06 - 100\cdot 0{,}04)^2}{100\cdot 0{,}04} = 4{,}05.$$

Оскільки $z_\text{o} = 4{,}05 < 12{,}6 = z_\alpha$, то гіпотезу H_0 приймаємо.

Приклад 9. Гіпотеза про ймовірність події

В результаті тривалих спостережень встановлено, що ймовірність повного одужання хворого, який приймав препарат P, дорівнює $p_0 = 0{,}8$. Новий препарат Q призначили $n = 800$ хворим, причому $n = 660$ з них повністю видужали. Чи можна вважати, що препарат Q ефективніший за препарат P з рівнем значущості $\alpha = 0{,}05$.

Розв'язання.

У цьому прикладі, відносна частота

$$p = \frac{m}{n} = \frac{660}{800} = 0{,}825 > 0{,}8 = p_0.$$

Ця нерівність вказує, що новий препарат може бути кращим. Але впровадження нового препарату потребує коштів. Тому, основна гіпотеза $H_0: p = p_0 = 0{,}8$ нам особливо дорога,

якщо так, то ми нових коштів не затрачуємо. Отже рівень значущості α – ймовірність того, що ми введемо новий препарат, а він не кращий від старого, повинна бути малою $\alpha = 0{,}05$. Як ми вже казали, в цьому випадку критична область правостороння,тому $H_1: p > p_0$.

Знаходимо критичну точку з рівняння

$$\Phi(z) = \frac{1-2\alpha}{2} = \frac{1-2\cdot 0{,}05}{2} = 0{,}45.$$

Звідки, $z_\alpha = 1{,}64$ (за Додатком 1).

За критерієм (3.7), знаходимо спостережуване значення

$$z_{\text{o}} = \frac{\left(\frac{m}{n} - p_0\right)\sqrt{n}}{\sqrt{p_0 \cdot q_0}} = \frac{\left(\frac{660}{800} - 0{,}8\right)\sqrt{800}}{\sqrt{0{,}8 \cdot 0{,}2}} \approx 1{,}78.$$

Оскільки $z_{\text{o}} = 1{,}78 > 1{,}64 = z_\alpha$, [потрапляє до критичної області] то гіпотезу H_0 відхиляємо і робимо висновок, що новий препарат кращий.

Приклад 10. Порівняння двох ймовірностей

Від двох постачальників до магазину надійшло $n_1 = 200$ і $n_2 = 300$ однотипних виробів (наприклад яєць). У першій партії виявилося $m_1 = 14$ бракованих, а в другій – $m_2 = 27$. Потрібно на рівні значущості $\alpha = 0{,}05$ оцінити, чи постачальники постачають однаково якісні (чи неякісні) товари?

Розв'язання.

Відносні частоти тут дорівнюють

$$\frac{m_1}{n_1} = \frac{14}{200} = 0{,}07, \quad \frac{m_2}{n_2} = \frac{27}{300} = 0{,}09.$$

Перевіримо гіпотезу $H_0: p_1 = p_2$ (обидві точні ймовірності p_1, p_2 того, що виріб бракований нам невідомі). Нехай альтернативна гіпотеза $H_1: p_1 \neq p_2$.

Для знаходження критичної точки (точніше, 2-х критичних точок) користуємося ЦГТ [3, I.1.28].. Згідно з нею, при великих n_1 та n_2 (3.13) можна замінити на стандартну нормальну величину $N(0,1)$. Отже, критичну точку z_α знаходимо з рівняння

$$\Phi(z) = \frac{1-\alpha}{2} = \frac{1-0{,}05}{2} = 0{,}475 \Longrightarrow z_\alpha = 1{,}96.$$

Далі обчислимо спостережуване значення z_{o}:

$$z_{\text{o}} = \frac{\frac{m_1}{n_1} - \frac{m_2}{n_2}}{\sqrt{\frac{m_1+m_2}{n_1+n_2}\left(1 - \frac{m_1+m_2}{n_1+n_2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} =$$

$$= \frac{\frac{14}{200} - \frac{27}{300}}{\sqrt{\frac{14+27}{200+300}\left(1 - \frac{14+27}{200+300}\right)\left(\frac{1}{200} + \frac{1}{300}\right)}} \approx -0{,}8.$$

Оскільки

$$z'_\alpha = -1{,}96 < z_\text{o} = -0{,}8 < 1{,}96 = z_\alpha\,,$$

то гіпотезу H_0: $p_1 = p_2$ приймаємо. Отже, постачальники постачають однаково якісні товари.

3.9. Питання для самоконтролю до Розділу 3

1. Що таке статистична гіпотеза?
2. Чим параметричні гіпотези відрізняються від непараметричних?
3. Яка гіпотеза називається основною, а яка – альтернативною?
4. Що таке статистичний критерій?
5. Сформулюйте означення критичної області й критичної точки.
6. Опишіть загальний алгоритм перевірки правильності статистичної гіпотези.
7. Як визначаються помилки 1-го та 2-го роду? Що таке потужність критерію?
8. Перевірка гіпотези $\mathbf{M}X = a$ з альтернативою $\mathbf{M}X > a$ і відомим σ.
9. Перевірка гіпотези $\mathbf{M}X = a$ з альтернативою $\mathbf{M}X < a$ і відомим σ.
10. Перевірка гіпотези $\mathbf{M}X = a$ з альтернативою $\mathbf{M}X \neq a$ і відомим σ.
11. Перевірка гіпотез $\mathbf{M}X = a$ з альтернативами $\mathbf{M}X > a$, $\mathbf{M}X < a$, $\mathbf{M}X \neq a$ і невідомим σ.
12. Перевірка правильності гіпотези H_0 про рівність двох математичних сподівань.
13. Перевірка правильності гіпотези про рівність дисперсій.
14. Що таке емпіричний розподіл та емпірична функція розподілу?
15. Опишіть критерій χ^2 (Пірсона) перевірки гіпотези про розподіл.
16. Опишіть критерій Колмогорова.
17. Перевірка гіпотези про ймовірність події.
18. Перевірка гіпотези про рівність двох імовірностей.

РОЗДІЛ 4. ДИСПЕРСІЙНИЙ АНАЛІЗ

Завдання дисперсійного аналізу

Нехай задано певну ознаку генеральної сукупності, тобто (в імовірнісних термінах) – випадкову величину X. У багатьох задачах нас цікавить, якою мірою певний фактор (або комбінація факторів) впливає на цю ознаку (тобто на випадкову величину X).

Наприклад, нехай винайдено новий медичний препарат і нас цікавить, якою мірою він ефективний. Як він впливає на здоров'я пацієнтів? Нас також може цікавити, в яких дозах препарат найефективніший? Звичайно, може бути і кілька препаратів. Тоді нас цікавить ефективність кожного препарату, зокрема і ефективність їхньої комбінації.

Математичне дослідження таких задач і становить предмет дисперсійного аналізу.

4.1. Однофакторний дисперсійний аналіз

Тут розглядається випадок, коли перевіряється вплив лише одного фактора.

У досить загальному вигляді задачу дисперсійного аналізу можна поставити так.

Нехай спостерігаються випадкові величини $X_1, X_2, \dots, X_m$, які незалежні між собою, нормально розподілені і мають одне і те ж середнє квадратичне відхилення σ. При цьому, математичні сподівання $\mathbf{M}X_1,\ \mathbf{M}X_2, \dots,\ \mathbf{M}X_m$ невідомі. Потрібно перевірити гіпотезу

$$H_0: \mathbf{M}X_1 = \mathbf{M}X_2 = \cdots = \mathbf{M}X_m.$$

Якщо $m = 2$, то задача перевірки такої гіпотези вже розглядалася у підрозділі 3.3.

Для перевірки гіпотези проводимо над кожною випадковою величиною $X_1, X_2, \dots, X_m$ n спостережень. Математично це означає, що беремо n незалежних копій кожної випадкової величини $X_i, i = 1, \dots, m$. Отримаємо таблицю, складену з випадкових величин.

X_1	X_{11}	X_{12}	…	X_{1n}	$\bar{X}_1$
X_2	X_{21}	X_{22}	…	X_{2n}	$\bar{X}_2$
…	…	…	…	…	…
X_m	X_{m1}	X_{m2}	…	X_{mn}	$\bar{X}_m$

У результаті фізичних вимірювань дістаємо матрицю відповідних числових значень цих випадкових величин:

	1	2	3	…	n	
1	x_{11}	x_{12}	x_{13}	…	x_{1n}	$\bar{x}_1$
2	x_{21}	x_{22}	x_{23}	…	x_{2n}	$\bar{x}_2$
…	…	…	…	…	…	…
m	x_{m1}	x_{m2}	x_{m3}	…	x_{mn}	$\bar{x}_m$

Тут $\forall i = 1, \dots, m$

$$\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$$

і, відповідно,

$$\bar{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij}$$

– середні. Їх називаємо ***середніми вимірювань***.

Приклади. Нехай ми маємо m мікроскопів. Вони однотипні, точніше мають однакову (і відому) точність, тобто однакові середні квадратичні відхилення. Але, при налаштуванні, мікроскопи можуть давати систематичні помилки. Хочемо виявити їх. Отже, тут перевіряємо один фактор – налаштування. У математичних термінах – перевіряємо рівність математичних сподівань.

Подібним є приклад із годинниками. Нехай вони однотипні, тому їхня точність однакова, А ми хочемо переконатися, чи вони поставлені на один час.

Отже, якщо гіпотеза $H_0: \mathbf{M}X_1 = \mathbf{M}X_2 = \dots = \mathbf{M}X_m$ справедлива, то за ЗВЧ не слід сподіватися на значну різницю між випадковими величинами $(\bar{X}_i)_{i=1}^m$ чи, відповідно, між значеннями $(\bar{x}_i)_{i=1}^m$. Але, як і при перевірці інших гіпотез, цьому поняттю потрібно надати кількісне значення.

Введемо ще одне середнє. А саме, позначимо

$$\bar{X} = \frac{1}{n \cdot m}\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}, \qquad \left(\text{і відповідно} \quad \bar{x} = \frac{1}{n \cdot m}\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij} \right)$$

і назвемо його ***середнім усіх вимірювань***.

Далі будуть потрібні два простих співвідношення

Лема 4.1.

$$\bar{X} = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{n}\sum_{j=1}^{n} X_{ij}\right) = \frac{1}{m}\sum_{i=1}^{m} \bar{X}_i.$$

Лема 4.2. $\forall i = 1, \dots, m$

$$\sum_{j=1}^{n}\left(X_{ij} - \bar{X}_i\right) = \sum_{j=1}^{n} X_{ij} - n\bar{X}_i = 0.$$

Отже, щоб надати поняттю «різниця між $(\bar{X}_i)_{i=1}^m$ мала» кількісне значення, розглянемо наступну випадкову величину

$$Q = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_{ij} - \bar{X}\right)^2,$$

яку називають ***загальною***, або ***повною сумою квадратів відхилень***.

Проаналізуємо цю суму.

$$Q = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\underbrace{X_{ij} - \bar{X}_i}_{} + \underbrace{\bar{X}_i - \bar{X}}_{}\right)^2 = [\text{піднесемо до квадрату}] =$$

$$= \underbrace{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_{ij} - \bar{X}_i\right)^2}_{Q_2} + \underbrace{\sum_{i=1}^{m}\sum_{j=1}^{n}(\bar{X}_i - \bar{X})^2}_{Q_1} + \underbrace{2\sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_{ij} - \bar{X}_i\right)(\bar{X}_i - \bar{X})}_{Q_3}.$$

Лема 4.3.

$$Q_3 = 2\sum_{i=1}^{m}(\bar{X}_i - \bar{X})\underbrace{\sum_{j=1}^{n}\left(X_{ij} - \bar{X}_i\right)}_{=0\ [\text{Л. 4.2.}]} = 0.$$

Висновок.

$$Q = Q_1 + Q_2.$$

Означення. У цій сумі величину Q_1 називають ***сумою квадратів відхилень між серіями***, або інакше, ***розсіюванням за факторами***, а величину Q_2 – ***сумою квадратів відхилень всередині серій***. Q_2 характеризує «залишкове розсіювання», тобто випадкові похибки спостережень.

Означення виправдовує наступна лема.

Лема 4.4.

$$Q_1 = \sum_{i=1}^{m}\sum_{j=1}^{n}(\bar{X}_i - \bar{X})^2 = n\sum_{i=1}^{m}(\bar{X}_i - \bar{X})^2.$$

Зауваження (для прикладів з мікроскопами, чи годинниками). Висновок показує, що «загальне» розсіювання показань приладів Q складається з двох компонент Q_1 і Q_2, які характеризують розсіювання між приладами, тобто відмінність їх систематичних помилок (Q_1) і розсіювання всередині окремих приладів (Q_2), яке за умовою (рівність дисперсій) має однаковий характер для всіх приладів.

Але повернімося до гіпотези H_0. Тепер неформальне твердження: «Якщо гіпотеза H_0 справедлива, то різниця між випадковими величинами $(\bar{X}_i)_{i=1}^m$ мала» може бути сформульована точніше: «Якщо гіпотеза H_0 справедлива, то Q_1 мале». Але, як і для інших гіпотез необхідно оперувати не абсолютними, а відносними числами. Отже, твердження уточнюється: «Якщо гіпотеза H_0 справедлива, то відношення $\frac{Q_1}{Q_2}$ мале». Але $\frac{Q_1}{Q_2}$ не має якогось відомого із самого початку розподілу. Щоб його виявити, розпочнемо з аналізу Q_1.

Твердження 4.1. Випадкова величина Q_1/σ^2 має розподіл χ^2 з $(m-1)$ ступенем свободи.

Доведення. Як ми вже кілька разів казали, середні $\bar{X}_i, i = 1, \dots, m,$ мають нормальний розподіл з математичними сподіваннями

$$\mathbf{M}\bar{X}_i = \mathbf{M}X_i$$

і дисперсіями

$$\mathbf{D}\bar{X}_i = \frac{\mathbf{D}X_i}{n} = \frac{\sigma^2}{n}.$$

Відповідно

$$\frac{\sum_{i=1}^{m}(\bar{X}_i - \bar{X})^2}{\sigma^2/n} \qquad (4.1)$$

має розподіл χ^2 з $(m-1)$ ступенем свободи.

Але $(4.1) = Q_1/\sigma^2$ (Лема 4.4). Це й доводить твердження. □

Лема 4.5 (закон композиції). Якщо випадкові величини X та Y мають розподіли χ^2 з k та l ступенями свободи, то випадкова величина $X + Y$ має розподіл χ^2 із $(k+l)$ ступенями свободи.

Аргументація. За умовою,

$$X = \sum_{i=1}^{k} X_i^2, \qquad Y = \sum_{i=1}^{l} Y_j^2,$$

де X_i та Y_j мають стандартні нормальні розподіли. Тоді

$$X + Y = \sum_{i=1}^{k+1} X_i^2 + \sum_{i=1}^{l+1} Y_j^2.$$

Можна було б сказати, що доведення закінчення, але виникають проблеми з незалежністю, які ми пропускаємо. □

Твердження 4.2. Випадкова величина Q_2/σ^2 має розподіл χ^2 з $m \cdot (n-1)$ ступенями свободи.

Доведення. Зафіксуємо спочатку $1 \le i \le m$. Тоді, як відомо, випадкова величина

$$\frac{1}{\sigma^2}\sum_{j=1}^{n}\left(X_{ij} - \bar{X}_i\right)^2$$

має розподіл χ^2 з $n-1$ ступенем свободи.

Згідно з Лемою 4.5 величина

$$\frac{1}{\sigma^2}\sum_{i=1}^{m}\left[\sum_{j=1}^{n}\left(X_{ij}-\bar{X}_i\right)^2\right]=\frac{Q_2}{\sigma^2}$$

має розподіл χ^2 з $m\cdot(n-1)$ ступенями свободи.

Теорема 4.1. При справедливості гіпотези H_0 випадкова величина

$$F=\frac{\frac{1}{m-1}Q_1}{\frac{1}{m(n-1)}Q_2}$$

має розподіл Фішера-Снедекора з $((m-1);m\cdot(n-1))$ ступенями свободи.

Доведення. Можна показати, що випадкові величини Q_1 та Q_2 незалежні. Тепер застосовуємо Твердження 4.1, 4.2 і означення розподілу Фішера-Снедекора. □

Правило перевірки гіпотези H_0

1. Вибираємо рівень значущості α. З рівняння

$$\mathbf{P}\{F>z\}=\alpha$$

знаходимо критичну точку z_α (Додаток 5).

2. Проводимо спостереження і обчислюємо

$$z_{\text{o}}=\frac{\frac{1}{m-1}n\sum_{i=1}^{m}(\bar{x}_i-\bar{x})^2}{\frac{1}{m(n-1)}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(x_{ij}-\bar{x}_i\right)^2}.$$

3. Якщо $z_{\text{o}}<z_\alpha$ – гіпотеза приймається, якщо ні – відхиляється.

4.2. Двофакторний дисперсійний аналіз

Кількість факторів у розглянутій ситуації може бути більше одного. У цьому підрозділі розглянемо вплив двох різних факторів A і B. Наприклад, у ситуації з мікроскопом, окрім кількох мікроскопів можуть бути ще й кілька різних лаборантів. І потрібно встановити, якою мірою на результати вимірювань впливає різне налаштування мікроскопів, а якою мірою – кваліфікація лаборантів.

Або, наприклад, досліджується вплив двох препаратів, аспірину і парацетамолу на температуру пацієнта.

Основна ідея дисперсійного аналізу тут полягає в розкладі суми квадратів відхилень від середнього на доданки, які відповідають окремим факторам.

Двофакторний дисперсійний аналіз без повторень

Отже, нехай задано два фактори A і B, які впливають на значення деякої випадкової величини X (наприклад, точність мікроскопа і кваліфікація лаборанта). Нашим завданням є з'ясування істотності цих факторів.

Розглянемо випадок, коли за фактором A проведено p груп спостережень $A_1, \dots, A_p$, а за фактором B – q груп спостережень $B^1, \dots, B^q$. Усі індекси, що стосуються фактора A позначатимемо знизу, а індекси, що стосуються фактора B – зверху. Складемо таблицю.

A\B		j (лаборанти)				
		B^1	B^2	…	B^q	
i (мікроскопи)	A_1	X_1^1	X_1^2	…	X_1^q	$\bar{X}_1$
	A_2	X_2^1	X_2^2	…	X_2^q	$\bar{X}_2$
	…	…	…	…	…	…
	A_p	X_p^1	X_p^2	…	X_p^q	$\bar{X}_p$
		$\bar{X}^1$	$\bar{X}^2$	…	$\bar{X}^q$	$\bar{X}$

У прикладі з мікроскопами і лаборантами, це означає, що кожен лаборант сідає за кожен мікроскоп точно один раз.

Припускається, що випадкові величини X_i^j нормально розподілені з невідомими (і можливо різними) математичними сподіваннями $\mathbf{M}X_i^j$ і однією, але невідомою дисперсією σ^2.

Потрібно перевірити тепер вже дві гіпотези про істотність впливу факторів A і B на випадкову величину X. Для цього, подібно до попереднього підрозділу, позначимо

$$\bar{X}_i = \frac{1}{q}\sum_{j=1}^{q} X_i^j,$$

$$\bar{X}^j = \frac{1}{p}\sum_{j=1}^{p} X_i^j,$$

$$\bar{X} = \frac{1}{pq}\sum_{j=1}^{q}\sum_{j=1}^{p} X_i^j,$$

де $i = 1, \dots, p, j = 1, \dots, q$ (вони дописані в таблиці).

Далі, розглянемо випадкову величину

$$Q = \sum_{i=1}^{p}\sum_{j=1}^{q}\left(X_i^j - \bar{X}\right)^2 = \sum_{i=1}^{p}\sum_{j=1}^{q}\left(\underbrace{X_i^j - \bar{X}_i - \bar{X}^j + \bar{X}} + \underbrace{\bar{X}_i - \bar{X}} + \underbrace{\bar{X}^j - \bar{X}}\right)^2 =$$

$$= \underbrace{q\sum_{i=1}^{p}(\bar{X}_i - \bar{X})^2}_{Q_1} + \underbrace{p\sum_{j=1}^{q}\left(\bar{X}^j - \bar{X}\right)^2}_{Q_2} + \underbrace{\sum_{i=1}^{p}\sum_{j=1}^{q}\left(X_i^j - \bar{X}_i - \bar{X}^j + \bar{X}\right)^2}_{Q_3} = Q_1 + Q_2 + Q_3.$$

Вирази Q_1 і Q_2 у цій сумі носять назви ***суми квадратів різниць для факторів A та B*** відповідно, а Q_3 – ***залишкової суми квадратів***. Вона відповідає за їхню спільну дію.

Міркуючи так само, як і в попередньому підрозділі, можна показати, що випадкові величини Q_1/σ^2, Q_2/σ^2 та Q_3/σ^2 мають розподіли χ^2 з $(p-1)$, $(q-1)$ та $(p-1)(q-1)$ ступенями свободи. Тепер для перевірки гіпотез H_A, H^B про вплив факторів A, B, відповідно, користуємося критеріями

$$F_A = \frac{\frac{1}{p-1}Q_1}{\frac{1}{(p-1)(q-1)}Q_3}$$

та

$$F^B = \frac{\frac{1}{q-1}Q_2}{\frac{1}{(p-1)(q-1)}Q_3}.$$

Ці випадкові величини мають розподіли Фішера-Снедекора з $((p-1);(p-1)(q-1))$ та $((q-1);(p-1)(q-1))$ зі ступенями свободи.

Алгоритм перевірки гіпотез H_A та H_B

1. Вибираємо рівень значущості α.

2. Нехай F_A та F^B – випадкові величини, які мають розподіли Фішера-Снедекора з $((p-1);(p-1)(q-1))$ та $((q-1);(p-1)(q-1))$ ступенями свободи відповідно. З рівнянь

$$\mathbf{P}\{F_A > z\} = \alpha \quad \text{та} \quad \mathbf{P}\{F^B > z\} = \alpha$$

знаходимо критичні точки z_A та z^B.

3. Проводимо спостереження і обчислюємо

$$c = \frac{1}{(p-1)(q-1)}\sum_{i=1}^{p}\sum_{j=1}^{q}\left(x_i^j - \frac{1}{q}\sum_{j=1}^{q}x_i^j - \frac{1}{p}\sum_{i=1}^{p}x_i^j + \frac{1}{pq}\sum_{i=1}^{p}\sum_{j=1}^{q}x_i^j\right)^2,$$

а потім спостережувані значення

$$z_o^A = \frac{1}{c(p-1)}\sum_{i=1}^{p}\left(\frac{1}{q}\sum_{j=1}^{q}x_i^j - \frac{1}{pq}\sum_{i=1}^{p}x_i^j\right)^2$$

та

$$z_{\text{o}}^{B} = \frac{1}{c(q-1)} p \sum_{j=1}^{q} \left(\frac{1}{p} \sum_{i=1}^{p} x_i^j - \frac{1}{pq} \sum_{j=1}^{q} x_i^j \right)^2 .$$

4. Якщо $z_{\text{o}}^{A} < z_A$, то гіпотезу H_A про незначущість впливу фактора A приймаємо, якщо $z_{\text{o}}^{A} > z_A$ – відхиляємо.

Те ж саме для фактора B: якщо $z_{\text{o}}^{B} < z^B$ – гіпотезу H^B про незначущість впливу фактора B приймаємо, а якщо $z_{\text{o}}^{B} > z^B$ – відхиляємо.

Двофакторний дисперсійний аналіз з повтореннями

Нехай тепер у кожному блоці проводяться n спостережень. У прикладі з мікроскопами і лаборантами це означає, що кожен лаборант за кожним мікроскопом проводить n вимірювань. У цьому випадку, в кожному блоці стоятиме не випадкова величина, а вектор випадкових величин довжини n.

A\B		j (лаборанти)							
		B^1		B^2		…	B^q		
i (мікроскопи)	A_1	$(X_1^1(k))_1^n$	$\bar{X}_1^1$	$(X_1^2(k))_1^n$	$\bar{X}_1^2$	…	$\left(X_1^q(k)\right)_1^n$	$\bar{X}_1^q$	$\bar{X}_1$
	A_2	$(X_2^1(k))_1^n$	$\bar{X}_2^1$	$(X_2^2(k))_1^n$	$\bar{X}_2^2$	…	$\left(X_2^q(k)\right)_1^n$	$\bar{X}_2^q$	$\bar{X}_2$
	…	…	…	…	…	…	…	…	…
	A_p	$\left(X_p^1(k)\right)_1^n$	$\bar{X}_p^1$	$\left(X_p^2(k)\right)_1^n$	$\bar{X}_p^2$	…	$\left(X_p^q(k)\right)_1^n$	$\bar{X}_p^q$	$\bar{X}_p$
			$\bar{X}^1$		$\bar{X}^2$	…		$\bar{X}^q$	$\bar{X}$

Покладемо

$$\forall i,j: \quad \bar{X}_i^j = \frac{1}{n} \sum_{k=1}^{n} X_i^j(k),$$

$$\forall i: \quad \bar{X}_i = \frac{1}{q} \sum_{j=1}^{q} \bar{X}_i^j,$$

$$\forall j: \quad \bar{X}^j = \frac{1}{p} \sum_{i=1}^{p} \bar{X}_i^j,$$

і, врешті,

$$\bar{X} = \frac{1}{npq}\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n} X_i^j(k).$$

У цьому випадку цілої суми квадратів Q не виписуватимемо (вона дуже довга). Запишемо лише відповідні суми квадратів:

$$Q_1 = nq \cdot \sum_{i=1}^{p} (\bar{X}_i - \bar{X})^2,$$

$$Q_2 = nq \cdot \sum_{j=1}^{q} \left(\bar{X}^j - \bar{X}\right)^2,$$

$$Q_3 = n \cdot \sum_{i=1}^{p}\sum_{j=1}^{q} \left(\bar{X}_i^j - \bar{X}_i - \bar{X}^j + \bar{X}\right)^2,$$

$$Q_4 = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n} \left(X_i^j(k) - \bar{X}_i^j\right)^2.$$

Ці величини зумовлені впливами факторів:

Q_1 – фактора A;

Q_2 – фактора B;

Q_3 – одночасним впливом факторів A та B;

Q_4 – дія випадкової компоненти спостереження одного лаборанта за одним мікроскопом.

Врешті, покладемо

$$S_1^2 = \frac{Q_1}{p-1},$$

$$S_2^2 = \frac{Q_2}{q-1},$$

$$S_3^2 = \frac{Q_3}{(p-1)(q-1)},$$

$$S_4^2 = \frac{Q_4}{pq(n-1)}.$$

Для перевірки гіпотез про значущість факторів, використовуємо наступні критерії:

$$A \quad \rightarrow \quad F_A = \frac{S_1^2}{S_4^2},$$

$$B \quad \rightarrow \quad F^B = \frac{S_2^2}{S_4^2},$$

$$A \cup B \quad \rightarrow \quad F_A^B = \frac{S_3^2}{S_4^2}.$$

Алгоритм перевірки гіпотез про значущість факторів A, B, та одночасного впливу A і B точно такий же, як і в попередньому підрозділі.

4.3. Приклади до Розділу 4

Приклад 1. Однофакторний дисперсійний аналіз

Вивчали вплив якісного фактора на трьох рівнях варіювання (вилив трьох видів добрив) на кількість зерен в колосі ярої пшениці. Результати досліджень наведено в таблиці

Номер досліду	Вид добрив		
	A	B	C
	Кількість зерен у колосі		
1	51	52	42
2	52	54	44
3	56	56	50
4	57	58	52

На основі отриманих результатів оцінити вплив добрив на урожайність ярої пшениці при $\alpha = 0{,}05$.

Розв'язання.

Підрахуємо середні значення за кожною градацією фактора:

$$\bar{x}_1 = \frac{51 + 52 + 56 + 57}{4} = 54; \quad \bar{x}_2 = \frac{52 + 54 + 56 + 58}{4} = 55;$$

$$\bar{x}_3 = \frac{42 + 44 + 50 + 52}{4} = 47.$$

Також знайдемо загальне середнє значення на основі групових середніх:

$$\bar{x} = \frac{54 + 55 + 47}{3} = 52.$$

Далі підрахуємо суми квадратів відхилень:

$$Q_1 = n\sum_{i=1}^{m}(\bar{x}_i - \bar{x})^2 = 4 \cdot [(54-52)^2 + (55-52)^2 + (47-52)^2] = 152.$$

$$Q_2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_i\right)^2$$
$$= [(51-54)^2 + (52-54)^2 + (56-54)^2 + (57-54)^2]$$
$$+ [(52-55)^2 + (54-55)^2 + (56-55)^2 + (58-55)^2]$$
$$+ [(42-47)^2 + (44-47)^2 + (50-47)^2 + (52-47)^2] = 114.$$

Спостережуване значення критерію:

$$F = \frac{\frac{1}{m-1}Q_1}{\frac{1}{m(n-1)}Q_2} = \frac{152/(3-1)}{114/(3\cdot(4-1))} = 6.$$

Критичне значення знаходиться за таблицею розкладу Фішера (Додаток 5):

$$F_\alpha(\alpha = 0{,}05; k_1 = 2; k_2 = 9) = 4{,}3.$$

Оскільки $F = 6 > 4{,}3 = F_\alpha$, то з рівнем надійності $0{,}05$ можна стверджувати, що вплив видів добрив на кількість зерен у колосі ярої пшениці є статистично значущим.

Приклад 2. Двофакторний дисперсійний аналіз (без повторень)

Досліджували вплив різних технологій вирощування (якісний фактор) і різних концентрацій добрив (кількісний фактор) на показники врожаю ярої пшениці. Потрібно оцінити значущість виливу зазначених факторних ознак на результативну ознаку при $\alpha = 0{,}05$. Результати експерименту наведено у таблиці

	Густота стояння ярої пшениці, шт/м²		
Варіант досліду	Технологія 1	Технологія 2	Технологія 3
$N_{30}P_{40}K_{15}$	407	412	395
$N_{30}P_{40}K_{15}$	437	442	405
$N_{90}P_{120}K_{45}$	455	469	425

Розв'язання.

Підрахуємо всі необхідні середні значення:

$$\bar{x}_1 = \frac{407+412+395}{3} = 404{,}7;\ \bar{x}_2 = \frac{437+442+405}{3} = 428;$$

$$\bar{x}_3 = \frac{455+469+425}{3} = 449{,}7.$$

$$\bar{x}^1 = \frac{407+437+455}{3} = 433;\quad \bar{x}^2 = \frac{412+442+469}{3} = 441;$$

$$\bar{x}^3 = \frac{395+405+425}{3} = 408{,}3.$$

$$\bar{x} = \frac{407+412+395+437+442+405+455+469+425}{3\cdot 3} = 427{,}4.$$

Далі знайдемо всі необхідні суми квадратів відхилень:

$$Q_1 = q\sum_{i=1}^{p}(\bar{x}_i - \bar{x})^2 = 3\cdot[(404{,}7-427{,}4)^2 + (428-427{,}4)^2 + (449{,}7-427{,}4)^2]$$
$$= 3038{,}9;$$

$$Q_2 = p\sum_{j=1}^{q}\left(\bar{x}^j - \bar{x}\right)^2 = 3\cdot[(433-427{,}4)^2 + (441-427{,}4)^2 + (408{,}3-427{,}4)^2]$$
$$= 1739{,}6;$$

$$Q_3 = \sum_{i=1}^{p}\sum_{j=1}^{q}\left(x_i^j - \bar{x}_i - \bar{x}^j + \bar{x}\right)^2$$
$$= [(407-404{,}7-433+427{,}4)^2 + (412-404{,}7-441+427{,}4)^2$$
$$+ (395-404{,}7-408{,}3+427{,}4)^2]$$
$$+ [(437-428-433+427{,}4)^2 + (442-428-441+427{,}4)^2$$
$$+ (405-428-408{,}3+427{,}4)^2]$$
$$+ [(455-449{,}7-433+427{,}4)^2 + (469-449{,}7-441+427{,}4)^2$$
$$+ (425-449{,}7-408{,}3+427{,}4)^2] = 229{,}8.$$

Порахуємо спостережувані значення статистик критерію:

$$F_A = \frac{\frac{1}{p-1}Q_1}{\frac{1}{(p-1)(q-1)}Q_3} = \frac{3038{,}9/2}{229{,}8/4} = 26{,}45;$$

$$F^B = \frac{\frac{1}{q-1}Q_2}{\frac{1}{(p-1)(q-1)}Q_3} = \frac{1739{,}6/2}{229{,}8/4} = 15{,}14.$$

За таблицею Фішера (Додаток 5) для рівня $\alpha = 0{,}05$ знайдемо критичні значення:

$$F_{A\alpha}(\alpha = 0{,}05; k_1 = 2; k_2 = 4) = 6{,}9;$$
$$F_{B\alpha}(\alpha = 0{,}05; k_1 = 2; k_2 = 4) = 6{,}9.$$

Оскільки у першому випадку (фактор А) $F_A = 26{,}45 > 6{,}9 = F_{A\alpha}$ то концентрація добрив має статистично значущий вплив на вихідний параметр - густоту стояння ярої пшениці. У другому випадку (фактор В) аналогічно $F^B = 15{,}14 > 6{,}9 = F_{B\alpha}$. Отже, вид технології вирощування також має статистично значущий вплив на вихідний параметр. Причому концентрація добрив ($F_A = 26{,}45$) має більший вплив порівняно з видом технології вирощування ($F^B = 15{,}14$).

Приклад 3. Двофакторний дисперсійний аналіз (з повтореннями)

Чотирьом групам по 4 випробовуваних у різних комбінаціях швидкості пред'явлення і довжини слова було запропоновано завдання з 10 слів для відтворення їх через деякий час. Результати експерименту наведені у таблиці. Необхідно довести значущість припущення про те, що між факторами швидкості пред'явлення слова (А) і довжини слова (В) спостерігається

взаємодія: при великій швидкості пред'явлення краще запам'ятовуються короткі слова, при низькій швидкості – довгі слова.

Рівень значущості обрати рівним $\alpha = 0{,}05$.

Фактор A \ Фактор B	Короткі (B1)	Довгі (B2)
Висока (A1)	7	5
	5	4
	4	3
	7	4
Низька (A2)	4	6
	3	4
	3	7
	4	5

Розв'язання.

Розрахуємо всі необхідні середні значення:

$$\bar{x}_1^1 = \frac{7+5+4+7}{4} = 5{,}8; \quad \bar{x}_1^2 = \frac{5+4+3+4}{4} = 4;$$

$$\bar{x}_2^1 = \frac{4+3+3+4}{4} = 3{,}5; \quad \bar{x}_2^2 = \frac{6+4+7+5}{4} = 5{,}5;$$

$$\bar{x}_1 = \frac{5{,}8+4}{2} = 4{,}9; \quad \bar{x}_2 = \frac{3{,}5+5{,}5}{2} = 4{,}5;$$

$$\bar{x}^1 = \frac{5{,}8+3{,}5}{2} = 4{,}6; \quad \bar{x}^2 = \frac{4+5{,}5}{2} = 4{,}8;$$

$$\bar{x} = \frac{7+5+4+7+5+4+3+4+4+3+3+4+6+4+7+5}{16} = 4{,}7.$$

Далі знайдемо всі необхідні суми квадратів відхилень:

$$Q_1 = nq\sum_{i=1}^{p}(\bar{x}_i - \bar{x})^2 = 4 \cdot 2 \cdot [(4{,}9-4{,}7)^2 + (4{,}5-4{,}7)^2] = 0{,}56;$$

$$Q_2 = nq\sum_{j=1}^{q}\left(\bar{x}^j - \bar{x}\right)^2 = 4 \cdot 2 \cdot [(4{,}6-4{,}7)^2 + (4{,}8-4{,}7)^2] = 0{,}06;$$

$$Q_3 = n\sum_{i=1}^{p}\sum_{j=1}^{q}\left(\bar{x}_i^j - \bar{x}_i - \bar{x}^j + \bar{x}\right)^2$$
$$= 4 \cdot [(5{,}8-4{,}9-4{,}6+4{,}7)^2 + (4-4{,}9-4{,}8+4{,}7)^2 + (3{,}5-4{,}5-4{,}6+4{,}7)^2 + (5{,}5-4{,}5-4{,}8+4{,}7)^2] = 14{,}06;$$

$$Q_4 = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n}\left(x_i^j(k) - \bar{x}_i^j\right)^2$$

$$= [(7-5{,}8)^2 + (5-5{,}8)^2 + (4-5{,}8)^2 + (7-5{,}8)^2] +$$

$$[(5-4)^2 + (4-4)^2 + (3-4)^2 + (4-4)^2]$$

$$+ [(4-3{,}5)^2 + (3-3{,}5)^2 + (3-3{,}5)^2 + (4-3{,}5)^2]$$

$$+ [(6-5{,}5)^2 + (4-5{,}5)^2 + (7-5{,}5)^2 + (5-5{,}5)^2] = 14{,}75.$$

За знайденими сумами обчислимо виправлені дисперсії:

$$S_1^2 = \frac{Q_1}{p-1} = \frac{0{,}56}{1} = 0{,}56;$$

$$S_2^2 = \frac{Q_2}{q-1} = \frac{0{,}06}{1} = 0{,}06;$$

$$S_3^2 = \frac{Q_3}{(p-1)(q-1)} = \frac{14{,}06}{1} = 14{,}06;$$

$$S_4^2 = \frac{Q_4}{pq(n-1)} = \frac{14{,}75}{2\cdot 2\cdot 3} = 1{,}23.$$

Обчислимо спостережувані значення статистик критерію:

$$F_A = \frac{S_1^2}{S_4^2} = \frac{0{,}56}{1{,}23} = 0{,}46;$$

$$F^B = \frac{S_2^2}{S_4^2} = \frac{0{,}06}{1{,}23} = 0{,}05;$$

$$F_A^B = \frac{S_3^2}{S_4^2} = \frac{14{,}06}{1{,}23} = 11{,}44.$$

За таблицею розподілу Фішера (Додаток 5) для рівня $\alpha = 0{,}05$ знайдемо критичні значення:

$$F_{A\alpha}(\alpha = 0{,}05; k_1 = 1; k_2 = 12) = 4{,}8;$$

$$F_{B\alpha}(\alpha = 0{,}05; k_1 = 1; k_2 = 12) = 4{,}8;$$

$$F_{AB\alpha}(\alpha = 0{,}05; k_1 = 1; k_2 = 12) = 4{,}8.$$

Висновки.

Фактор А: $F_A = 0{,}46 < 4{,}8 = F_{A\alpha}$ ⇒ вплив фактора А відсутній.

Фактор В: $F^B = 0{,}05 < 4{,}8 = F_{B\alpha}$ ⇒ вплив фактора В відсутній.

Фактор А та В одночасно: $F_A^B = 11{,}44 > 4{,}8 = F_{AB\alpha}$ ⇒ існує сумісний вплив факторів А та В на ознаку.

Отже, фактори довжини слів і швидкості їхнього пред'явлення окремо не впливають значуще на обсяг відтворення слів. Значущою виявляється взаємодія факторів: короткі слова краще запам'ятовуються при великій швидкості пред'явлення, а довгі – при повільній швидкості пред'явлення.

4.4. Питання для самоконтролю до Розділу 4

1. Сформулюйте суть однофакторного аналізу.
2. Що таке сума квадратів вимірювань для однофакторного аналізу?
3. Який розподіли має сумою квадратів відхилень між серіями?
4. Який розподіли має сума квадратів відхилень всередині серії.
5. Опишіть правило перевірки гіпотези про вплив одного фактора.
6. Сформулюйте суть двофакторного аналізу без повторень.
7. З чого складається сума квадратів вимірювань для двофакторного аналізу без повторень?
8. Які розподіли мають суми квадратів відхилень для двофакторного аналізу без повторень?
9. Опишіть правило перевірки гіпотези про вплив факторів для двофакторного аналізу без повторень.
10. Сформулюйте суть двофакторного аналізу з повтореннями.
11. Опишіть правило перевірки гіпотези про вплив факторів для двофакторного аналізу без повторень.
12. Опишіть правило перевірки гіпотези про вплив факторів для двофакторного аналізу з повтореннями.

РОЗДІЛ 5. КОРЕЛЯЦІЙНИЙ АНАЛІЗ

Кореляційний аналіз передбачає вивчення залежності між випадковими величинами з одночасною кількісною оцінкою ступеня невипадковості їхньої сумісної зміни.

Зміна випадкової величини Y, що відповідає зміні випадкової величини X, розбивається на дві складові – стохастичну, що пов'язана з невипадковою залежністю Y від X, і випадкову (або статистичну), що пов'язана із випадковим характером поведінки самих X та Y.

У цьому розділі буде розглянута стохастична складова зв'язку між Y та X на основі коефіцієнтів кореляції.

5.1. Коефіцієнт кореляції Пірсона

Нагадаємо, що для випадкових величин X та Y їхній ***коефіцієнт коваріації*** визначається формулою

$$\mathrm{Cov}(X,Y) = \mathbf{M}\big((X-\mathbf{M}X)(Y-\mathbf{M}Y)\big),$$

а ***коефіцієнт кореляції*** –

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y},$$

де σ_X та σ_Y – середні квадратичні відхилення X та Y.

Теорема 5.1. Коефіцієнт кореляції $|\rho(X,Y)| \le 1$, причому $|\rho(X,Y)| = 1$ тоді і тільки тоді, коли X та Y зв'язані лінійною залежністю:

$$\text{або } Y = aX + b,$$
$$\text{або } X = a'Y + b'.$$

Доведення достатності.

Нехай, наприклад, $Y = aX + b,\ a \neq 0$. Тоді

$$\mathrm{Cov}(X,Y) = \mathrm{Cov}(X, aX+b) = \mathbf{M}\big[(X-\mathbf{M}X)\big(aX+b-\mathbf{M}(aX+b)\big)\big] =$$
$$= \mathbf{M}[(X-\mathbf{M}X)(aX+b-a\mathbf{M}X-b)] = a\mathbf{M}[(X-\mathbf{M}X)(X-\mathbf{M}X)] = a\mathbf{D}X.$$

З іншого боку,

$$\sigma_X \cdot \sigma_Y = \sqrt{\mathbf{D}X} \cdot \sqrt{\mathbf{D}(aX+b)} = \sqrt{\mathbf{D}X \cdot a^2\mathbf{D}X} = |a| \cdot \mathbf{D}X.$$

Таким чином,

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{a \cdot \mathbf{D}X}{|a| \cdot \mathbf{D}X} = \begin{cases} 1, & a > 0; \\ -1, & a < 0. \end{cases} \quad \square$$

Зауваження. Якщо випадкові величини X та Y незалежні, то їхній коефіцієнт коваріації, а отже й коефіцієнт кореляції, дорівнює 0. Обернене твердження не має місця [6, п.1.15.1].

Врешті, пригадаємо, що випадкові величини X, Y, для яких $\rho(X,Y) = 0$ називаються ***нескорельованими***.

У статистиці для оцінки (невідомого) коефіцієнта кореляції $\rho(X,Y)$ проводять по n випробувань випадкових величин X та Y, отримують їхні числові значення $(x_i)_1^n$ та $(y_i)_1^n$ і обчислюють:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,; \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i\,;$$

$$s_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}\,; \quad s_Y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}\,,$$

$$\operatorname{cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

і, врешті, покладають

$$r = \frac{\operatorname{cov}(X,Y)}{s_X \cdot s_Y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}. \qquad (5.1)$$

Означення. Число r з формули (5.1) називається ***коефіцієнтом кореляції Пірсона***.

Зауваження 1. Таким чином, означення коефіцієнта кореляції Пірсона збігається з означенням вибіркового коефіцієнта кореляції з підрозділу 1.4, а $\operatorname{cov}(X,Y)$ – це дещо підправлений вибірковий коефіцієнт коваріації K_{xy} з цього підрозділу. Різниця полягає в тому, що тут, для того щоб відповідна статистика була незміщеною, сума ділиться на $n-1$.

Зауваження 2. У практичних обчисленнях для малого обсягу вибірки ($n < 100$) часом застосовують ***скорегований коефіцієнт кореляції***

$$r' = r\left[1 + \frac{1-r^2}{2(n-3)}\right]. \qquad (5.2)$$

Перевірка гіпотези про нескорельованість

Розглянемо гіпотезу

$$H_0\colon \rho(X,Y) = 0.$$

Альтернативна гіпотеза у цьому випадку

$$H_1\colon \rho(X,Y) \neq 0.$$

Перевірка гіпотези H_0 здійснюється залежно від обсягу вибірки.

Перед розглядом критеріїв у даному розділі введемо поняття квантилю розподілу.

Означення. ***Квантиль*** – це аргумент z_α функції розподілу $F(z)$, якому відповідає задана ймовірність α, тобто розв'язок рівняння

$$F(z) = \alpha.$$

а) Великий обсяг вибірки ($n \geq 100$)

У цьому випадку використовується критерії Стьюдента. Нехай задано рівень значущості α і нехай обсяг вибірки дорівнює n.

1. За допомогою розкладу Стьюдента з $(n-2)$ ступенями свободи знаходимо критичні точки $-z_\alpha$ та z_α. Для цього розв'язуємо рівняння

$$\mathbf{P}\{|t_{n-2}| > z\} = \alpha.$$

2. На підставі вибірки обчислюємо спостережуване значення

$$z_{\text{o}} = \frac{r}{\sqrt{1-r^2}}\sqrt{n-1},$$

де r знаходимо з формули (5.1).

3. Якщо $z_{\text{o}} \in \left(-z_\alpha(n-2); z_\alpha(n-2)\right)$ – гіпотеза H_0 приймається, якщо ні – відхиляється.

б) Малий обсяг вибірки ($n < 100$)

1. Замість критичного значення z_α, обчисленого за допомогою розподілу Стьюдента, використовується функція розподілу стандартної нормальної випадкової величини $F(x)$. А саме, задавшись рівнем значущості α знаходимо розв'язок z_α рівняння

$$F(z_\alpha) = 1 - \frac{\alpha}{2}.$$

Зокрема, з таблиці квантилів стандартного нормального розподілу (Додаток 2) знаходимо: для $\alpha = 0{,}05 \to z_\alpha = 1{,}96$, а для $\alpha = 0{,}01 \to z_\alpha = 2{,}576$. Врешті, покладаємо

$$u_\alpha = u_\alpha(n) = z_\alpha \cdot \frac{1}{\sqrt{n-3}}. \qquad (5.3)$$

2. Замість вибіркового коефіцієнта кореляції r використовується скорегований коефіцієнт кореляції r'. Далі використовується так зване перетворення Фішера

$$u = \frac{1}{2}\ln\frac{1+r'}{1-r'}.$$

3. Якщо $-u_\alpha < u < u_\alpha$ – гіпотезу H_0 приймаємо, якщо ні – відхиляємо.

Знаходження довірчого інтервалу для великої вибірки ($n > 100$)

Спочатку встановимо рівень надійності $\gamma = 1 - \alpha$ (γ – велике). Далі, користуючись розподілом Стьюдента з $(n-2)$ ступенями свободи знаходимо t_α. Тоді проміжок

$$\left(r - t_\alpha(n-2)\frac{1-r^2}{\sqrt{n-1}};\ r + t_\alpha(n-2)\frac{1-r^2}{\sqrt{n-1}}\right)$$

буде довірчим інтервалом, до якого невідоме значення ρ потрапить з ймовірністю $\gamma = 1 - \alpha$.

Знаходження довірчого інтервалу для малої вибірки ($n < 100$)

Межі r_1 та r_2 довірчого інтервалу для невідомого коефіцієнту кореляції $\rho(X,Y)$ знаходять за наступним алгоритмом.

1. Встановивши рівень надійності $\gamma = 1 - \alpha$ з формули (5.3) знаходимо число $u_\alpha = u_\alpha(n)$.

2. За результатами n випробувань обчислюємо експериментальний коефіцієнт кореляції r. Для цього користуємося формулою

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (5.4)$$

де $(x_i)_1^n$ та $(y_i)_1^n$ – експериментальні значення випадкових величин X та Y.

3. Покладаємо

$$u = \frac{1}{2}\ln\frac{1+r}{1-r},$$

$$u_1 = u - u_\alpha, \qquad u_2 = u + u_\alpha.$$

4. Підставляючи обчислені значення u_1 та u_2 до формули

$$r = \frac{e^{2u} - 1}{e^{2u} + 1}$$

дістаємо межі довірчого інтервалу з надійністю γ:

$$r_1 = \frac{e^{2u_1} - 1}{e^{2u_1} + 1}, \quad r_2 = \frac{e^{2u_2} - 1}{e^{2u_2} + 1}.$$

Перевірка гіпотези рівності двох коефіцієнтів кореляції

Нехай ми маємо два (невідомі) коефіцієнти кореляції ρ_1 та ρ_2 для пар випадкових величин (X_1, Y_1) та (X_2, Y_2) відповідно і хочемо перевірити гіпотезу

$$H_0: \rho_1 = \rho_2$$

з альтернативною гіпотезою

$$H_1: \rho_1 \neq \rho_2.$$

Перевірку цієї гіпотези проводимо за наступним алгоритмом.

1. Задавшись рівним значущості α, з рівняння

$$F(z) = 1 - \frac{\alpha}{2},$$

де $F(z)$ – функція розподілу нормальної стандартної випадкової величини, знаходимо критичну точку z_α (Додаток 2).

2. Провівши n_1 випробувань для пари випадкових величин (X_1, Y_1) та n_2 випробувань для пари випадкових величин (X_2, Y_2), з формули (5.4) знаходимо експериментальні значення r_1 та r_2 коефіцієнтів кореляції для пар (X_1, Y_1) та (X_2, Y_2) відповідно.

3. Обчислюємо

$$u_1 = \frac{1}{2}\ln\frac{1+r_1}{1-r_1},\quad u_2 = \frac{1}{2}\ln\frac{1+r_2}{1-r_2}$$

і, врешті,

$$z_{\text{o}} = \frac{|u_1 - u_2|}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}.$$

4. Якщо $z_{\text{o}} < z_\alpha$ – гіпотеза H_0 приймається, якщо $z_{\text{o}} > z_\alpha$ – відхиляється.

5.2. Коефіцієнт рангової кореляції Спірмена

У статистиці часом виникає проблема аналізу зв'язку між «випадковими величинами», які набувають не числових, або не зовсім числових значень, але які можна впорядкувати за інтенсивністю зростання (або спадання) ознаки. Зв'язок між такими величинами вимірюється за допомогою ***коефіцієнтів рангової кореляції***. Один із них – коефіцієнт рангової кореляції Спірмена, до якого ми зараз перейдемо.

Розглянемо спочатку таку задачу з психології. Психолог хоче дослідити рівень самооцінки пацієнта. Він вибирає (наприклад) 4 риси характеру.

Риси характеру	X (бажані)	Y (наявні)
Сміливість	4	3
Наполегливість	3	4
Замкненість	2	1
Нерішучість	1	2

Пацієнтові спершу пропонується розташувати їх в порядку бажаності, найменш бажаній ознаці приписується число 1 і т. д., найбажанішій – 4. На другому кроці пацієнтові пропонується приписати кожній ознаці число від 1 до 4 розташувавши їх в порядку вираженості цих ознак у нього. Зрозуміло, що коли різниця модулів відповідних чисел для ознак велика, то самооцінка пацієнта низька, а якщо близька до нуля, то висока.

Отже, формально, тут маємо $n = 4$ спостережень двох випадкових величин X та Y

$$\{x_1, x_2, \dots, x_n\} \tag{5.5}$$
$$\{y_1, y_2, \dots, y_n\}, \tag{5.6}$$

які набувають різних натуральних значень від 1 до n.

Означення. ***Коефіцієнтом рангової кореляції Спірмена*** називається число

$$r_s = 1 - \frac{6\cdot\sum_{i=1}^{n}(x_i - y_i)^2}{n^3 - n} = 1 - \frac{6\cdot\sum_{i=1}^{n} {d_i}^2}{n^3 - n}, \tag{5.7}$$

де $d_i = x_i - y_i,\ \ i = 1, \dots, n$. Проте, в наборах (5.5)–(5.6) можуть зустрічатися й однакові значення (наприклад, пацієнт вважає певні риси характеру рівноцінними). У цьому випадку, до формули (5.7) вносять наступні поправки. Позначимо через l_x (відповідно l_y) кількість груп однакових значень в наборі (5.5) (відповідно (5.6)). Нехай $t_k(x)$ (відповідно $t_k(y)$), $k = 1, \dots, l_x$ (відповідно $k = 1, \dots, l_y$) – кількість членів у -й групі, а

$$T_x = \sum_{k=1}^{l_x} (t_k^3(x) - t_k(x)), \qquad T_y = \sum_{k=1}^{l_y} (t_k^3(y) - t_k(y)).$$

Тоді, за означенням,

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} {d_i}^2}{(n^3 - n) - \frac{1}{2}(T_x + T_y)}. \tag{5.8}$$

Зауваження. $|r_s| \le 1$, причому коефіцієнт кореляції Спірмена завжди не перевищує модуля коефіцієнта кореляції Пірсона:

$$r_s \le |r|.$$

Перевірка гіпотези про некорельованість

а) Обсяг вибірки $n > 30$

У цьому випадку, перевірка гіпотези H_0 про некорельованість за Спірменом ($r_s = 0$) проводиться за допомогою розподілу Стьюдента t_{n-2} з $(n-2)$ ступенями свободи. А саме:

1. Встановивши критерій значущості α з рівняння

$$\mathbf{P}\{|t_{n-2}| > t_\alpha\} = \alpha$$

(див таблицю Додаток 3) знаходимо критичне значення t_α.

2. Проводимо експеримент і за отриманими даними (5.5)–(5.6), користуючись формулами (5.7), (5.8) обчислюємо коефіцієнт кореляції Спірмена r_s.

3. Обчислюємо t_s за формулою

$$t_s = \frac{r_s}{\sqrt{1 - {r_s}^2}} \sqrt{n-2}.$$

1. Якщо $t_s < t_\alpha(n-2)$ – гіпотезу H_0 приймаємо, а якщо ні – відхиляємо.

б) Обсяг вибірки $n < 30$

Для перевірки гіпотези H_0, у цьому випадку використовують таблицю критичних значень коефіцієнта рангової кореляції Спірмена (Додаток 6). Отже:

1. З неї, задавшись рівнем значущості α і знаючи обсяг вибірки n, знаходять критичне значення r_α.

2. За отриманими даними експерименту (5.5)–(5.6), користуючись формулами (5.7), (5.8) обчислюємо r_s.

3. Якщо $r_s < r_\alpha$ – гіпотеза H_0 приймається, а якщо ні – відхиляється.

5.3. Коефіцієнт рангової кореляції Кендала

Спочатку введемо поняття ***рангу***. Його ми вже (неявно) використовували в попередньому підрозділі. А тепер опишемо докладніше.

Нехай в результаті експерименту ми отримали значення пар

$$\begin{pmatrix} x_1, x_2, \dots, x_n \\ y_1, y_2, \dots, y_n \end{pmatrix} \tag{5.9}$$

випадкових величин X та Y, які набувають різних натуральних значень від 1 до n. І нехай

$$\begin{pmatrix} x_{i_1}, x_{i_2}, \dots, x_{i_n} \\ y_{i_1}, y_{i_2}, \dots, y_{i_n} \end{pmatrix}$$

– така перестановка набору (5.9), що

$$x_{i_1} \le x_{i_2} \le \dots \le x_{i_n}.$$

Так ось, елементи скінченної послідовності

$$(i_1, i_2, \dots, i_n) \tag{5.10}$$

називаємо ***рангами*** послідовності (5.9).

Опишемо тепер алгоритм обчислення коефіцієнта рангової кореляції Кендала.

1. Послідовність натуральних чисел $(R_i)_{i=1}^{n-1}$ будуємо так:

1) Серед чисел (5.10) є одиниця. Позначимо через R_1 кількість елементів (5.9), що стоять правіше від 1, а потім цю одиницю викидаємо.

2) У новоутвореній послідовності з $(n-1)$ елементів є число 2. Позначимо через R_2 кількість елементів новоутвореної послідовності, що стоять правіше від 2, а потім цю двійку викидаємо.

……

n) Коли ми дійдемо до останнього числа n, то воно лишиться лише одне. Тому правіше від нього нічого не буде і ми R_n не визначаємо.

2. Тепер ***коефіцієнт кореляції Кендала*** визначається за формулою

$$\tau = \frac{4 \cdot \sum_{i=1}^{n-1} R_i}{n(n-1)} - 1. \tag{5.11}$$

Перевірка гіпотези H_0 про некорельованість за Кендалом

Перевірка гіпотези $H_0: \tau = 0$ здійснюється так:

1. Встановлюється рівень значущості α. Потім з рівняння

$$F(z) = 1 - \frac{\alpha}{2},$$

де $F(z)$ – функція розподілу стандартної нормальної випадкової величини, знаходимо точку z_α, тобто розв'язок цього рівняння.

2. За формулою

$$\tau_\alpha = \tau_\alpha(n) = z_\alpha \sqrt{\frac{2(2n+5)}{9n(n+1)}}$$

обчислюємо критичне значення τ_α.

3. На підставі результатів експерименту, з формули (5.11) знаходимо експериментальний коефіцієнт кореляції Кендала τ.

4. Якщо $|\tau| < \tau_\alpha$ – гіпотеза H_0 про некорельованість випадкових величин X та Y приймається. У супротивному випадку, гіпотеза відхиляється.

5.4. Коефіцієнт конкордації (узгодженості) Кендала

Цей коефіцієнт застосовується у випадках, коли хочуть оцінити кореляцію між m випадковими величинами. Тут варто пригадати приклад з попереднього розділу, де розглядалися m лаборантів, що роблять спостереження на m мікроскопах. Але метод конкордації Кендала застосовують тоді, коли мають справу не з числовими, чи не зовсім числовими «випадковими величинами», або якщо дані в групах не нормально розподілені.

Наприклад, нехай 7 школярів виступають на змаганнях з художньої гімнастики, а судять їх 5 експертів (суддів), які присуджують кожному школяреві місце. Хочемо з'ясувати узгодженість їхніх оцінок. Якщо всі судді присуджують кожному окремому учасникові одне і теж саме місце – узгодженість ідеальна. І, навпаки, якщо зовсім різні місця, то узгодженості між суддями зовсім немає.

Формалізуємо тепер цей приклад.

Означення. Нехай є вибірка, скажімо, $\omega_1, \omega_2, \dots, \omega_n$ (у прикладі – це учні) і є m випадкових величин (у прикладі – це судді, або спостерігачі) $X^1, \dots, X^m$. Значення $X^j(\omega_i)$ – це місця, присуджені суддями.

Отже, для фіксованого ω_i $(i = 1, \dots, n)$, кожна випадкова величина X^j $(j = 1, \dots m)$, набуває різних натуральних значень $X^j(\omega_i)$ від 1 до n.

Позначимо $\forall i = 1, \dots, n$

$$d_i = \sum_{i=1}^{m} X^j(\omega_i)\ , \qquad \bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$$

і, врешті,

$$D_i = d_i - \bar{d}.$$

Означення. ***Коефіцієнтом конкордації Кендала*** називається число

$$W = \frac{12 \cdot \sum_{i=1}^{n} {D_i}^2}{m^2(n^3 - n)}. \qquad (5.12)$$

Відомо, що $0 \le W \le 1$.

За наявності однакових рангів у одного спостерігача (у прикладі – це коли один суддя може присуджувати кільком учасникам одне й те саме місце) формула (5.12) набуває вигляду

$$W = \frac{12 \cdot \sum_{i=1}^{n} {D_i}^2}{m^2(n^3 - n) - m\sum_{j=1}^{m} T_j}. \qquad (5.13)$$

У цій формулі, для j-го спостерігача (судді), $j = 1, \dots, m$ через l_j позначається кількість груп з однаковими значеннями, через t_{kj} – кількість однакових значень у k-й групі і, врешті,

$$T_j = \sum_{k=1}^{l_j}\left(t_{kj}^3 - t_{kj}\right).$$

Перевірка гіпотези про некорельованість.

Йдеться про гіпотезу H_0: думки суддів ніяк не узгоджуються між собою. Альтернативна гіпотеза H_1 – думки суддів якось узгоджуються. Для перевірки цієї гіпотези використовують коефіцієнт конкордації Кендала, який обчислюється за формулами (5.12), (5.13).

Неформально цей коефіцієнт використовують так: проводять n випробувань з оцінками m суддів і обчислюють коефіцієнт W. Якщо отримане значення W близьке до нуля – гіпотезу приймають, а якщо ближче до одиниці – відхиляють.

Формально це здійснюється так:

1. Встановлюється рівень значущості α. Для обчислення критичного значення застосовується розклад χ^2_{n-1} з $(n-1)$ ступенем свободи. А саме, з рівняння

$$\mathbf{P}\{\chi^2_{n-1} > z\} = \alpha$$

(див. Додаток 4) знаходимо критичне значення z_α.

2. На підставі вибірки обсягу n, користуючись формулами (5.8), (5.9) знаходимо емпіричне значення

$$z_{\text{o}} = m \cdot (n-1) \cdot W.$$

3. Якщо $z_{\text{o}} < z_\alpha$ – гіпотеза H_0 приймається, а якщо ні – відхиляється.

Зауваження. Гіпотезу H_0 можна перевіряти для $n > 7$.

5.5. Коефіцієнт конкордації (узгодженості) Шукені

Розглянемо приклад, коли на змаганнях з художньої гімнастики виступають n учнів. І Нехай їх судять дві різні суддівські бригади. Наприклад, з Кропивницького і Олександрії. І нехай ми хочемо перевірити узгодженість між оцінками цих бригад. Наприклад, чи вони своїм не підсуджують. Побудуємо таблиці з результатами:

1-а бригада			
	1-й учень	2-й учень	…
1	1	4	
2	1	5	
3	2	5	
4	1	4	
5	2	3	

2-а бригада			
	1-й учень	2-й учень	…
1	3	2	
2	4	1	
3	5	1	
4	3	2	
5	4	3	
6	4	1	

З таблиць виникає підозра неузгодженості. Перша бригада “підсуджує” першому учневі, а друга – другому.

Формалізуємо тепер цей приклад. А саме, розглянемо наступну загальну задачу. Нехай дві групи експертів, чисельністю відповідно m та $\bar{m}$ ставлять перед собою завдання упорядкувати n об'єктів (у нашому прикладі – дві суддівські бригади присуджують місця за виступи школярів). Позначимо через R_i^j $(i = 1, \dots, n, j = 1, \dots, m)$ ранги (тобто місця) запропоновані m експертами в 1-ї групи, а через $\bar{R}_i^j$ – ранги, запропоновані експертами другої групи $(i = 1, \dots, n, j = 1, \dots, \bar{m})$.

Покладемо

$$R_i = \sum_{j=1}^{m} R_i^j, \quad \bar{R}_i = \sum_{j=1}^{m} \bar{R}_i^j, \quad i = 1, \dots, n,$$

і, врешті,

$$L = \sum_{i=1}^{n} R_i \cdot \bar{R}_i.$$

Значення статистики L знаходиться в інтервалі

$$\frac{m \cdot \bar{m} \cdot n \cdot (n+1) \cdot (n+2)}{6} \le L \le \frac{m \cdot \bar{m} \cdot n \cdot (n+1) \cdot (2n+2)}{6}.$$

Можна порахувати, що математичне сподівання

$$\mathbf{M}(L) = \frac{m \cdot \bar{m} \cdot n \cdot (n+1)^2}{4},$$

а дисперсія

$$\mathbf{D}(L) = \frac{m \cdot \bar{m} \cdot (n-1) \cdot n^2 \cdot (n+1)^2}{144}.$$

Означення. ***Коефіцієнтом конкордації Шукені*** називається число

$$\widetilde{W} = \frac{L - \mathbf{M}(L)}{\max(L) - \mathbf{M}(L)}, \qquad (5.10)$$

де $\max(L)$ означає максимум випадкової величини L.

Зауваження. Значення коефіцієнта $\widetilde{W}$ поблизу +1 означають високий степінь узгодженості всередині обох груп експертів та між групами; -1 – високий степінь узгодженості всередині груп і сильну неузгодженість між групами; поблизу 0 – загальну неузгодженість.

5.6. Бісеріальна кореляція

Префікс ***бі*** означає два. Назва коефіцієнту натякає, що йдеться про дві серії випробувань. Формальніше, йдеться про кореляцію між двома випадковими величинами X і Y, де X набуває звичайних числових значень, а Y – набуває лише два значення.

Приклади. Нехай наприклад X – успішність студента (виміряна в середній оцінці), а Y вказує – походить студент з Кропивницького чи з Олександрії, або Y вказує хлопець – це, чи дівчина, або ще щось. І ми хочемо встановити, чи існує зв'язок (тобто кореляція), наприклад між успішністю студента і його походженням.

Точково бісеріальний коефіцієнт кореляції

Отже, нехай X – звичайна (числова) випадкова величина, а Y – двозначна випадкова величина. Значення Y позначатимемо через 0 і 1.

Нехай проведено n випробувань пари (X, Y) і отримано значення

$$\begin{pmatrix} x_1, x_2, \dots, x_n \\ y_1, y_2, \dots, y_n \end{pmatrix} \tag{5.14}$$

де y_i набуває значень 0, 1.

Введемо такі позначення:

n_0 – кількість нулів у другому рядку формули (5.14);

n_1 – кількість одиниць у другому рядку формули (5.14) (звичайно, $n_0 + n_1 = n$);

$\bar{x}_0$ – середнє значення тих елементів першого рядка формули (5.14), для яких відповідні елементи другого рядка дорівнюють нулю:

$$\bar{x}_0 = \frac{1}{n_0} \sum_{i:\, y_i = 0} x_i.$$

$\bar{x}_1$ – середнє значення тих елементів першого рядка формули (5.14), для яких відповідні елементи другого рядка дорівнюють 1:

$$\bar{x}_1 = \frac{1}{n_1} \sum_{i:\, y_i = 1} x_i.$$

s_X – виправлене середнє квадратичне відхилення випадкової величини X:

$$s_X = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2},$$

де

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Означення. ***Точково бісеріальним коефіцієнтом кореляції*** називається число

$$r_{pb} = \frac{\bar{x}_1 - \bar{x}_0}{S_X} \cdot \sqrt{\frac{n_0 \cdot n_1}{n(n-1)}}. \qquad (5.15)$$

Зауваження. $|r_{pb}| \le 1$ і $|r_{pb}| = 0$ тоді і тільки тоді, коли $\bar{x}_0 = \bar{x}_1$.

Перевірка гіпотези про нескорельованість

Йдеться про гіпотезу H_0: між випадковими величинами X та Y немає зв'язку (наприклад, оцінки студентів ніяк не залежать від місця проживання), альтернативна гіпотеза H_1 – залежність існує. Для перевірки гіпотези використовується розподіл Стьюдента t_{n-2} з $(n-2)$ ступенями свободи.

Алгоритм перевірки наступний.

1. Встановивши рівень значущості α, з рівняння

$$\mathbf{P}\{|t_{n-2}| > z\} = \alpha$$

знаходимо критичне значення z_α.

2. На підставі вибірки, знаходимо спостережуване значення

$$z_{\text{o}} = \frac{r_{pb}}{\sqrt{1 - r_{pb}^2}} \cdot \sqrt{n-2}.$$

3. Якщо $|z_{\text{o}}| < z_\alpha$, то гіпотеза H_0 приймається. Якщо ж ні – відхиляється.

Рангово бісеріальний коефіцієнт кореляції

Він використовується тоді, коли X набуває рангових значень (тобто 1,...,n), а Y – дихотомічних (наприклад, 0, 1).

Приклад. Пригадаємо приклад зі змаганнями з художньої гімнастики, де виступають учні Кропивницького і з Олександрії. Нехай X – місце, яке присуджують судді, а $Y = 0$, коли учасник з Кропивницького і $Y = 1$ – коли з Олександрії. Нашим завданням є встановити, чи існує залежність між місцем у змаганнях і походженням учня.

Нехай, як і раніше,

$$\bar{x}_0 = \frac{1}{n_0}\sum_{i:\, y_i = 0} x_i,$$

$$\bar{x}_1 = \frac{1}{n_1} \sum_{i:\, y_i = 1} x_i.$$

За означенням, ***рангово бесеріальний коефіцієнт кореляції***

$$r_{rb} = \frac{2}{n}(\bar{x}_1 - \bar{x}_0). \tag{5.16}$$

Перевірка гіпотези про нескорельованість випадкових величин X та Y здійснюється точно так само, як і у випадку точкового бісеріального коефіцієнта кореляції.

5.7. Спряженість

Спряженість 2×2

Розглянемо дві випадкові величини X, Y кожна з яких набуває лише двох якісних значень, які умовно позначатимемо «+» і «-». Хочемо перевірити, чи існує між X і Y якась залежність.

Приклад. Розглянемо групу пацієнтів $\omega_1, \dots, \omega_n$. Одним із них давали препарат X (напр. парацетамол), а іншим – ні. Стан Y (температура) одних покращився, а інших – ні. Формальніше, випадкові величини X, Y можна записати так

$$\left.\begin{aligned} X(\omega_i) &= \begin{cases} +, & i-\text{му пацієнту давали препарат} \\ -, & i-\text{му пацієнту не давали препарат} \end{cases} \\ Y(\omega_i) &= \begin{cases} +, & \text{стан } i-\text{го пацієнта покращився} \\ -, & \text{стан } i-\text{го пацієнта не покращився} \end{cases} \end{aligned}\right\} \quad i = 1, \dots, n.$$

Таблиця спряженості у цьому випадку має вигляд:

		X	
		«+»	«-»
Y	«+»	a	b
	«-»	c	d

Пояснимо, що в цій таблиці означають літери a, b, c, d. Проведемо спостереження описаних вище випадкових величин X, Y. Отримуємо послідовність пар

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} x_1, \dots, x_i, \dots, x_n \\ y_1, \dots, y_i, \dots, y_n \end{pmatrix}$$

Нагадаємо, кожне x_i і кожне y_i – це знак «+» або «-». Так ось

$$a = \operatorname{card}\{i: x_i = +, y_i = +\};$$
$$b = \operatorname{card}\{i: x_i = -, y_i = +\};$$
$$c = \operatorname{card}\{i: x_i = +, y_i = -\};$$
$$d = \operatorname{card}\{i: x_i = -, y_i = -\}.$$

Нашим завданням є перевірка зв'язку між значеннями X та Y. У прикладі – чи є зв'язок між застосуванням препарату і покращенням стану хворого.

Наведемо кілька величин, які дозволяють оцінювати існування чи відсутність зв'язку.

а) Коефіцієнт асоціації

$$Q = \frac{ad - bc}{ad + bc}. \tag{5.17}$$

Визначається по принципу визначника матриці 2×2. Якщо випадкові величини X та Y незалежні, то $Q = 0$. У випадку повного зв'язку $Q = \pm 1$.

Дисперсія випадкової величини Q дорівнює

$$\mathbf{D}Q = \frac{1}{4}(1 - Q^2)\left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right).$$

Порівняння коефіцієнту Q із отриманим значенням дисперсії (з рахуванням масштабу $\sqrt{\mathbf{D}Q}$) дає змогу отримати оцінку зв'язку. Якщо коефіцієнт Q перевищує значення $\sqrt{\mathbf{D}Q}$ більш як у три рази, то зв'язок між якісними змінними значущий.

б) Коефіцієнт колігації Юла

$$K = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}. \tag{5.18}$$

Термін колігація часто зустрічається в хімії і означає сполучення в групи.

Відомо, що дисперсія цієї випадкової величини

$$\mathbf{D}K = \frac{1}{16}(1 - K^2)\left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right).$$

Аналогічно, якщо коефіцієнт K перевищує значення $\sqrt{\mathbf{D}K}$ більш як у три рази, то зв'язок між якісними змінними значущий.

Між наведеними коефіцієнтами існує співвідношення

$$Q = \frac{2K}{1 + K^2}.$$

в) Коефіцієнт контингенції (подібності), або коефіцієнт Пірсона

$$V = \frac{ad - bc}{\sqrt{(a + d)(a + c)(b + d)(c + d)}}. \tag{5.19}$$

Перевірка гіпотези про значущість зв'язку між випадковими величинами X та Y у цьому випадку проводиться за допомогою критерію χ^2 з одним ступенем свободи. А саме, обчислюється спостережуване значення

$$z_{\text{o}} = V^2 \cdot n\,,$$

де n – обсяг вибірки, і порівнюється з критичним значенням z_α для кількості ступенів свободи 1 та рівня значущості α.

Якщо $z_\text{o} < z_\alpha$, то гіпотеза H_0 про відсутність зв'язку приймається, а якщо $z_\text{o} > z_\alpha$ – відхиляється.

Спряженість $k \times l$

Нехай тепер випадкова величина X набуває не два, а k якісних значень, які ми нумеруємо, а випадкова величина Y набуває l якісних значень, які ми теж нумеруємо. Хочемо встановити існування чи відсутність зв'язку між X та Y.

Приклад. Нехай проводиться опитування населення щодо вирощування канабісу в медичних цілях. Опитують n осіб і групують їх за віком: 18 років – 1, 19 років – 2,... Варіанти відповідей групуємо наступним чином: так – 1, скоріше так ніж ні – 2, не визначився – 3, скоріше ні ніж так – 4, ні – 5. Нехай n_{ij} – кількість осіб віком N_i, що дала *j*-ту відповідь.

Таблиця спряженості набуде вигляду:

		Y				
		1	2		l	Σ
X	1	n_{11}	n_{12}	…	n_{1l}	n_1
	2	n_{21}	n_{22}	…	n_{2l}	n_2
	…	…	…	…	…	…
	k	n_{k1}	n_{k2}	…	n_{kl}	n_k
Σ		n_1^*	n_2^*	…	n_l^*	n

Формальніше: проводимо n спостережень, де

n_{11} – кількість спостережень, у яких $X = 1$ і $Y = 1$;

n_{12} – кількість спостережень, у яких $X = 1$ і $Y = 2$...

Взагалі n_{ij} – кількість спостережень, у яких $X = i$ і $Y = j$.

Позначимо

$$n_i = \sum_{j=1}^{l} n_{ij}\,, \qquad i = 1, \dots, k;$$

$$n_j^* = \sum_{i=1}^{k} n_{ij}.$$

Нагадаємо, що нашим завданням є встановлення зв'язку між випадковими величинами X та Y. За міру цього зв'язку використовуємо статистику

$$\chi^2 = \chi^2(n) = n \cdot \left(\sum_{i=1}^{k} \sum_{j=1}^{l} \frac{n_{ij}^2}{n_i \cdot n_j^*} - 1 \right). \qquad (5.20)$$

яка при незалежності ознак має розподіл хі-квадрат з $(k-1)\cdot(l-1)$ ступенями свободи.

Тому, для рівня значущості α з таблиць розкладу χ^2 (Додаток 4) обчислюємо критичне значення χ_α^2. І якщо $\chi^2 > \chi_\alpha^2$, то з ймовірністю α залежність між випадковими величинами X та Y приймається, а якщо ні – відхиляється.

Проте статистика (5.16) незручна, оскільки, на відміну від коефіцієнта кореляції, який міститься в межах від -1 до 1, її значення $\chi^2(n) \to \infty$ при $n \to \infty$.

Тому, частіше використовують так званий ***коефіцієнт спряженості Пірсона***

$$K_p = \left(\frac{\chi^2}{n + \chi^2} \right)^{1/2}. \qquad (5.21)$$

Значення коефіцієнту лежать у межах від 0 до 1, що дає змогу оцінити тісноту зв'язку. Якщо його значення рівне 0, то величини X та Y не зв'язані між собою, якщо значення близьке до 1, то зв'язок тісний.

Перевірка існування залежності між випадковими величинами X та Y за допомогою коефіцієнта спряженості Пірсона здійснюється так само, як і для статистики (5.20). А саме:

1. Провівши n випробувань, знаходимо χ^2 за допомогою формули (5.20), а за допомогою формули (5.21) знаходимо число K_p.

2. Встановлюємо рівень значущості α. Нехай випадкова величина Z має розподіл χ^2 з $(k-1)(l-1)$ ступенями свободи. Тоді з рівняння

$$P\{Z > z\} = \alpha$$

і таблиць (Додаток 4), знаходимо число χ_α^2.

3. Користуємось тим, що для додатних z функція (від змінної z)

$$\varphi(z) = \left(\frac{z}{n+z} \right)^{1/2}$$

монотонно зростаюча. Тому,

$$z > \chi_\alpha^2 \Leftrightarrow \left(\frac{z}{n+z} \right)^{1/2} > \left(\frac{\chi_\alpha^2}{n+\chi_\alpha^2} \right)^{1/2}.$$

Отже, критичним значенням для випадкової величини $\left(\frac{z}{n+Z} \right)^{1/2}$ з рівнем значущості α буде число

$$K_p^\alpha = \left(\frac{\chi_\alpha^2}{n+\chi_\alpha^2} \right)^{1/2}.$$

4. Якщо отримане з експерименту значення $K_p > K_p^\alpha$, то з ймовірністю α існування залежності між випадковими величинами X та Y приймається, а якщо ні – відхиляється.

5.8. Приклади до Розділу 5

Приклад 1. Коефіцієнт кореляції Пірсона

У результаті спостережень над випадковими величинами X та Y отримана наступна сукупність даних:

x_i	2	4	1	7	3	11	14	15	21	4
y_i	7	6	4	11	2	21	31	23	40	15

Необхідно перевірити гіпотезу про наявність кореляції між випадковими величинами X та Y за допомогою коефіцієнту кореляції Пірсона на рівні значущості $\alpha = 0{,}05$.

Розв'язання.

Знаходимо

$$\bar{x} = \frac{2+4+1+\cdots+4}{10} = 8{,}2; \quad \bar{y} = \frac{7+6+4+\cdots+15}{10} = 16;$$

$$\sum_{i=1}^{10}(x_i-\bar{x})^2 = \frac{(2-8{,}2)^2+(4-8{,}2)^2+\cdots+(4-8{,}2)^2}{10} = 405{,}6;$$

$$\sum_{i=1}^{10}(y_i-\bar{y})^2 = \frac{(7-16)^2+(6-16)^2+\cdots+(15-16)^2}{10} = 1422;$$

$$\sum_{i=1}^{10}(x_i-\bar{x})(y_i-\bar{y}) = \frac{(2-8{,}2)(7-16)+(4-8{,}2)(6-16)+\cdots+(4-8{,}2)(15-16)}{10} = 723;$$

$$r = \frac{\operatorname{cov}(X,Y)}{s_X \cdot s_Y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{723}{\sqrt{405{,}6}\cdot\sqrt{1422}} = 0{,}952.$$

Оскільки вибірка у нас мала ($n < 15$) краще використати скоригований коефіцієнт кореляції

$$r' = r\left[1+\frac{1-r^2}{2(n-3)}\right] = 0{,}952\cdot\left[1+\frac{1-0{,}952^2}{2\cdot(10-3)}\right] = 0{,}958.$$

Перевіримо значущість отриманого коефіцієнту кореляції. Для обраного рівня значущості $\alpha = 0{,}05$ квантиль стандартного нормального розподілу дорівнює $z_{1-\frac{\alpha}{2}} = z_{0{,}975} = 1{,}96$ (за Додатком 2). Далі знаходимо

$$u_\alpha = u_\alpha(n) = z_{1-\frac{\alpha}{2}}\cdot\frac{1}{\sqrt{n-3}} = 1{,}96\cdot\frac{1}{\sqrt{10-3}} = 0{,}74.$$

Використовуючи перетворення Фішера підраховуємо

$$u = \frac{1}{2}\ln\frac{1+r'}{1-r'} = \frac{1}{2}\ln\frac{1+0{,}958}{1-0{,}958} = 1{,}92.$$

Отже, отримане значення $u = 1{,}92$ не потрапляє в діапазон $(-0{,}74;\ 0{,}74)$, тому зв'язок між змінними X та Y є значущим.

Покажемо також, як знайти довірчий інтервал для коефіцієнту r:

$$u_1 = u - u_\alpha = 1{,}92 - 0{,}74 = 1{,}18;$$

$$u_2 = u + u_\alpha = 1{,}92 + 0{,}74 = 2{,}66.$$

Відповідно, дістаємо межі довірчого інтервалу з надійністю $\gamma = 1 - 0{,}05 = 0{,}95$:

$$r_1 = \frac{e^{2u_1} - 1}{e^{2u_1} + 1} = \frac{e^{2\cdot 1{,}18} - 1}{e^{2\cdot 1{,}18} + 1} = 0{,}827;$$

$$r_2 = \frac{e^{2u_2} - 1}{e^{2u_2} + 1} = \frac{e^{2\cdot 2{,}66} - 1}{e^{2\cdot 2{,}66} + 1} = 0{,}99.$$

Приклад 2. Коефіцієнт кореляції Спірмена

У результаті спостережень над випадковими величинами X та Y отримана наступна сукупність даних:

x_i	2	4	7	1	5	9	11	12	17	8
y_i	6	3	5	7	1	2	4	14	18	21

Необхідно перевірити гіпотезу про наявність кореляції між випадковими величинами X та Y за допомогою коефіцієнту кореляції Спірмена на рівні значущості $\alpha = 0{,}05$.

Розв'язання.

Упорядкуємо ряд x_i за зростанням і знайдемо ранг для кожного значення:

x_i	1	2	4	5	7	8	9	11	12	17
R_i	1	2	3	4	5	6	7	8	9	10

Аналогічно, упорядкуємо ряд y_i за зростанням і знайдемо ранг для кожного значення:

y_i	1	2	3	4	5	6	7	14	18	21
R_i^*	1	2	3	4	5	6	7	8	9	10

Для початкового набору x_i та y_i побудуємо таблицю з їх рангами:

R_i	2	3	5	1	4	7	8	9	10	6
R_i^*	6	3	5	7	1	2	4	8	9	10

Далі підрахуємо суму квадратів різниць рангів і сам коефіцієнт кореляції Спірмена:

$$\sum_{i=1}^{n} {d_i}^2 = (2-6)^2 + (3-3)^2 + (5-5)^2 + \cdots + (6-10)^2 = 120;$$

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} {d_i}^2}{n^3 - n} = 1 - \frac{6 \cdot 120}{10^3 - 10} = 0{,}273.$$

Виконаємо перевірку значущості отриманого коефіцієнту. За Додатком 6 знаходимо критичне значення $r_{0,05} = 0{,}636$. Оскільки $r_s = 0{,}273 < 0{,}636 = r_{0,05}$, то зв'язок між змінними не значущий.

Приклад 3. Коефіцієнт кореляції Кендала

В умовах прикладу 2 перевірити гіпотезу про наявність кореляції між випадковими величинами X та Y за допомогою коефіцієнту кореляції Кендала на рівні значущості $\alpha = 0{,}05$.

Розв'язання.

Спочатку упорядковуємо ряд значень x_i за зростанням і запишемо відповідні їм значення y_i:

x_i	1	2	4	5	7	8	9	11	12	17
y_i	7	6	3	1	5	21	2	4	14	18

Виконаємо заміну значень x_i та y_i їх рангами:

R_i	1	2	3	4	5	6	7	8	9	10
R_i^*	7	6	3	1	5	10	2	4	8	9

Для початкової послідовності рангів R_i^*

R_i^* 7; 6; 3; **1**; 5; 10; 2; 4; 8; 9

визначаємо кількість членів, які розташовані праворуч від $R_4^* = 1$. Отримуємо 6 членів.

Викреслюємо $R_4^* = 1$ і отримуємо ряд

R_i^* 7; 6; 3; 5; 10; **2**; 4; 8; 9

Праворуч від $R_6^* = 2$ розташовані 3 члени.

Викреслюємо $R_6^* = 2$ і отримуємо ряд

R_i^* 7; 6; **3**; 5; 10; 4; 8; 9

Праворуч від $R_3^* = 3$ розташовано 5 членів.

Викреслюємо $R_3^* = 3$ і отримуємо ряд

R_i^* 7; 6; 5; 10; **4**; 8; 9

Праворуч від $R_5^* = 4$ розташовані 2 члени.

Викреслюємо $R_5^* = 4$ і отримуємо ряд

R_i^* 7; 6; **5**; 10; 8; 9

Праворуч від $R_3^* = 5$ розташовано 3 члени.

Викреслюємо $R_3^* = 5$ і отримуємо ряд

R_i^* 7; **6**; 10; 8; 9

Праворуч від $R_2^* = 6$ розташовано 3 члени.

Викреслюємо $R_2^* = 6$ і отримуємо ряд

R_i^* **7**; 10; 8; 9

Праворцч від $R_1^* = 7$ розташовано 3 члени.

Викреслюємо $R_1^* = 7$ і отримуємо ряд

R_i^* 10; **8**; 9

Праворуч від $R_2^* = 8$ розташований 1 член.

Викреслюємо $R_2^* = 8$ і отримуємо ряд

R_i^* 10; **9**

Праворуч від $R_2^* = 9$ розташовано 0 членів.

Можемо знайти коефіцієнт кореляції Кендала:

$$\tau = \frac{4 \cdot \sum_{i=1}^{n-1} R_i}{n(n-1)} - 1 = \frac{4 \cdot (6+3+5+2+3+3+3+1+0)}{10 \cdot 9} - 1 = 0{,}155.$$

Перевіримо значущість отриманого коефіцієнту. Для цього розрахуємо критичне значення:

$$\tau_{0,05} = \tau_{0,05}(n) = z_{1-\frac{\alpha}{2}}\sqrt{\frac{2(2n+5)}{9n(n+1)}} = 1{,}96 \cdot \sqrt{\frac{2 \cdot (2 \cdot 10 + 5)}{9 \cdot 10 \cdot (10+1)}} = 0{,}44.$$

Оскільки $|\tau| = 0{,}155 < 0{,}44 = \tau_{0,05}$, то приймається гіпотеза про некорельованість за Кендалом випадкових величин X та Y.

Приклад 4. Коефіцієнт конкордації (узгодженості) Кендала

Нехай $n = 7$ школярів виступають на змаганнях з художньої гімнастики, а судять їх $m = 5$ експертів (суддів), які присуджують кожному школяреві місце. Результати експертних оцінок наведені у таблиці

Школярі	Експерти				
	1	2	3	4	5
1	1	1	2	1	3
2	3	2	1	2	1
3	4	5	7	4	5
4	2	3	5	6	4
5	6	6	6	3	2
6	7	4	4	5	6
7	5	7	3	7	7

Потрібно з'ясувати узгодженість їхніх оцінок. Якщо всі судді присуджують кожному окремому учасникові одне і теж саме місце – узгодженість ідеальна. І, навпаки, якщо зовсім різні місця, то узгодженості між суддями зовсім немає. На основі коефіцієнту конкордації Кендала на рівні значущості $\alpha = 0{,}05$ перевірити гіпотезу про узгодженість думок експертів.

Розв'язання.

Оскільки виставлені бали вже є рангами від 1 до 7, то можемо перейти до знаходження коефіцієнту конкордації. У таблиці наведемо підрахунки рядкових сум рангів та квадрат їх відхилення від середньої суми рангів усіх об'єктів.

Школярі	Експерти					$d_i = \sum_{i=1}^{m} R_i^j$	$D_i = d_i - \bar{d}$	D^2
	1	2	3	4	5			
1	1	1	2	1	3	8	-12	144
2	3	2	1	2	1	9	-11	121
3	4	5	7	4	5	25	5	25
4	2	3	5	6	4	20	0	0
5	6	6	6	3	2	23	3	9
6	7	4	4	5	6	26	6	36
7	5	7	3	7	7	29	9	81
						140		**416**

Середня сума рангів усіх об'єктів дорівнює $\bar{d} = \frac{140}{7} = 20$.

Коефіцієнт конкордації Кендала визначається формулою

$$W = \frac{12 \cdot \sum_{i=1}^{n} {D_i}^2}{m^2(n^3 - n)} \frac{12 \cdot 416}{5^3(7^3 - 7)} = 0{,}119.$$

Перевіримо значущість отриманого коефіцієнту. Знайдемо спостережуване значення

$$\chi^2 = m(n-1)W = 5 \cdot 6 \cdot 0{,}119 = 3{,}57.$$

Далі порахуємо критичне значення χ_α^2 (Додаток 4) для кількості ступенів свободи $(n-1)$:

$$\chi_{0{,}05}^2(6) = 12{,}6.$$

Отримуємо, що $\chi^2 = 3{,}57 < 12{,}6 = \chi_\alpha^2$. Отже, конкордація не значуща, і, відповідно, думки оцінок експертів не узгоджені.

Приклад 5. Коефіцієнт конкордації (узгодженості) Шукені

Дві групи експертів у кількості $m = 6$ і $\overline{m} = 8$ провели ранжування $n = 5$ об'єктів (результати наведені в таблицях).

Номер експерта	Ранжування об'єктів				
Група 1					
1	1	3	4	2	5
2	1	2	3	4	5
3	4	3	2	1	5
4	1	2	3	4	5
5	2	1	3	4	5
6	5	4	3	2	1
R_i	$R_1 = 14$	$R_2 = 15$	$R_3 = 18$	$R_4 = 17$	$R_5 = 26$

Група 2					
1	1	2	3	4	5
2	3	2	1	5	4
3	4	5	1	2	3
4	1	2	3	4	5
5	5	4	2	3	1
6	1	2	3	4	5
7	3	2	4	5	1
8	1	5	4	3	2
$\overline{R}_i$	$\overline{R}_1 = 19$	$\overline{R}_2 = 24$	$\overline{R}_3 = 21$	$\overline{R}_4 = 30$	$\overline{R}_5 = 26$

Потрібно перевірити узгодженість думок експертів за допомогою критерія Шукені.

Розв'язання.

У таблиці підрахуємо стовпчикові суми рангів R_i та $\overline{R}_i$, і на їх основі знайдемо величину:

$$L = \sum_{i=1}^{5} R_i \cdot \overline{R}_i = 14 \cdot 19 + 15 \cdot 24 + 18 \cdot 21 + 17 \cdot 30 + 26 \cdot 26 = 2190.$$

Значення статистики L потрапили до інтервалу

$$\frac{6 \cdot 8 \cdot 5 \cdot (5+1)(5+2)}{6} \le L \le \frac{6 \cdot 8 \cdot 5 \cdot (5+1)(2 \cdot 5+2)}{6};$$

$$1680 \le L \le 5040.$$

Далі підрахуємо математичне сподівання

$$\mathbf{M}(L) = \frac{m \cdot \overline{m} \cdot n \cdot (n+1)^2}{4} = \frac{6 \cdot 8 \cdot 5 \cdot 6^2}{4} = 2160$$

і дисперсію

$$\mathbf{D}(L) = \frac{m \cdot \bar{m} \cdot (n-1) \cdot n^2 \cdot (n+1)^2}{144} = \frac{6 \cdot 8 \cdot 4 \cdot 5^2 \cdot 6^2}{144} = 1200.$$

Тепер можемо знайти сам коефіцієнт конкордації Шукені

$$\widetilde{W} = \frac{L - \mathbf{M}(L)}{\max(L) - \mathbf{M}(L)} = \frac{2190 - 2160}{5040 - 2160} = 0{,}00729.$$

Оскільки $\widetilde{W} = 0{,}00729 \approx 0$, то узгодженість всередині групи експертів відсутня.

Приклад 6. Точково бісеріальна кореляція

За допомогою точково бісеріального коефіцієнту кореляції необхідно перевірити, чи існує зв'язок між статтю людини й самооцінкою. Жіноча стать позначена – 0, чоловіча – 1. Результати дослідження представлені в таблиці

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Стать (X)	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
Самооцінка (Y)	95,7	97	83	98,6	88,6	83	71	90	70	70	70	68,6	70	73	68,5

Розв'язання.

Порахуємо середні значення у групах:

$$\bar{x}_0 = \frac{90 + 70 + 70 + 70 + 68{,}6 + 70 + 73 + 68{,}5}{8} = 72{,}5;$$

$$\bar{x}_1 = \frac{95{,}7 + 97 + 83 + 98{,}6 + 88{,}6 + 83 + 71}{7} = 88{,}1;$$

$$\bar{x} = \frac{95{,}7 + 97 + 83 + \cdots + 68{,}5}{15} = 79{,}8.$$

Далі знайдемо виправлене середнє квадратичне відхилення:

$$s_X = \sqrt{\frac{(95{,}7 - 79{,}8)^2 + (97 - 79{,}8)^2 + \cdots + (68{,}5 - 79{,}8)^2}{14}} = 11{,}54.$$

Підставимо всі отримані величини у формулу точково бісеріального коефіцієнту кореляції:

$$r_{pb} = \frac{\bar{x}_1 - \bar{x}_0}{S_X} \sqrt{\frac{n_0 \cdot n_1}{n(n-1)}} = \frac{88{,}1 - 72{,}5}{11{,}54} \cdot \sqrt{\frac{8 \cdot 7}{15 \cdot 14}} = 0{,}699.$$

Перевіримо значущість отриманого коефіцієнту. Для цього знайдемо спостережуване значення

$$t = \frac{r_{pb}}{\sqrt{1 - r_{pb}^2}}\sqrt{n-2} = \frac{0{,}699}{\sqrt{1 - 0{,}699^2}} \cdot \sqrt{13} = 3{,}53,$$

а також критичне значення t_α для $(n-2)$ ступенів свободи (Додаток 3): $t_{0{,}05}(13) = 2{,}16$.

Оскільки $|t| = 3{,}53 > 2{,}16 = t_\alpha$, то коефіцієнт точково бісеріальної кореляції значущий.

Приклад 7. Рангово бісеріальна кореляція

Необхідно перевірити, чи існує зв'язок між статтю досліджуваного і комунікативними здібностями? Результати дослідження представлені в таблиці. Дівчата – 0; Хлопці – 1.

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Стать (X)	1	0	1	1	0	1	0	0	1	1	1	0	1	1	0
Ранг вербальних здібностей (Y)	1	10	6	9	15	7	8	13	4	3	5	11	12	2	14

Розв'язання.

Підрахуємо середні значення у групах:

$$\bar{x}_0 = \frac{10 + 15 + 8 + 13 + 11 + 14}{6} = 11{,}83;$$

$$\bar{x}_1 = \frac{1 + 6 + 9 + 7 + 4 + 3 + 5 + 12 + 2}{9} = 5{,}44.$$

На основі цих середніх, можемо знайти коефіцієнт рангово бісеріальної кореляції

$$r_{rb} = \frac{2}{n}(\bar{x}_1 - \bar{x}_0) = \frac{2}{15}(5{,}44 - 11{,}83) = -0{,}852.$$

Перевіримо значущість отриманого коефіцієнту. Аналогічно, знайдемо спостережуване значення

$$t = \frac{r_{rb}}{\sqrt{1 - \mathbf{r}_{rb}^2}}\sqrt{n-2} = \frac{-0{,}852}{\sqrt{1 - (-0{,}852)^2}} \cdot \sqrt{13} = -5{,}869,$$

а також критичне значення t_α для $(n-2)$ ступенів свободи (Додаток 3): $t_{0{,}05}(13) = 2{,}16$.

Оскільки $|t| = 5{,}869 > 2{,}16 = t_\alpha$, то коефіцієнт рангово бісеріальної кореляції значущий.

Приклад 8. Спряженість 2×2

Необхідно перевірити, чи існує залежність між статтю та депресивністю особистості? Результати дослідження представлені в таблиці.

Ознаки		Депресія	
		Є ознаки	Ознаки відсутні
Стать	Жінки	9	13
	Чоловіки	4	14

Ознака X (стать): 1 – жінка; 2 – чоловік. Ознака Y (депресія): 1 – є ознаки; 2 – відсутні ознаки. Знайти коефіцієнт асоціації та коефіцієнт колігації Юла. Перевірити їх значущість.

Розв'язання.

В нашому випадку маємо:

$$a = 9; \quad b = 13; \quad c = 4; \quad d = 14.$$

Коефіцієнт асоціації

$$Q = \frac{ad - bc}{ad + bc} = \frac{9 \cdot 14 - 13 \cdot 4}{9 \cdot 14 + 13 \cdot 4} = 0{,}416.$$

Дисперсія випадкової величини Q дорівнює

$$\mathbf{D}Q = \frac{1}{4}(1 - Q^2)\left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right) = \frac{1}{4}(1 - 0{,}416^2)\left(\frac{1}{9} + \frac{1}{13} + \frac{1}{4} + \frac{1}{14}\right) = 0{,}105;$$

$$\sqrt{\mathbf{D}Q} = 0{,}323.$$

Коефіцієнт колігації Юла

$$K = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}} = \frac{\sqrt{9 \cdot 14} - \sqrt{13 \cdot 4}}{\sqrt{9 \cdot 14} + \sqrt{13 \cdot 4}} = 0{,}218.$$

Дисперсія випадкової величини K дорівнює

$$\mathbf{D}K = \frac{1}{16}(1 - K^2)\left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right) = \frac{1}{16}(1 - 0{,}218^2)\left(\frac{1}{9} + \frac{1}{13} + \frac{1}{4} + \frac{1}{14}\right) = 0{,}03;$$

$$\sqrt{\mathbf{D}K} = 0{,}174.$$

Проаналізуємо значущість отриманих коефіцієнтів.

Оскільки $Q = 0{,}416 < 0{,}974 = 3 \cdot \sqrt{\mathbf{D}Q}$, то зв'язок між статтю та депресивністю не значущий.

Аналогічно, $K = 0{,}218 < 0{,}522 = 3 \cdot \sqrt{\mathbf{D}K}$, то зв'язок між статтю та депресивністю не значущий.

Приклад 9. Спряженість $k \times l$

Припустимо, що у результаті перевірки партії електронних ламп трьох типів (по 100 штук кожного типу), виготовлених на п'яти заводах, отримані наступні кількості справних ламп:

Тип лампи	Завод-виробник					n_j^*
	1	2	3	4	5	
1	70	60	20	40	30	220
2	80	90	100	90	70	430
3	30	40	30	20	50	170
n_i	180	190	150	150	150	820

Потрібно перевірити гіпотезу про наявність зв'язку між якістю ламп різного типу і заводом-виробником при рівні значущості $\alpha = 0{,}05$.

Розв'язання.

Знайдемо величину

$$\chi^2 = n\left(\sum_{i=1}^{k}\sum_{j=1}^{l}\frac{n_{ij}^2}{n_i \cdot n_j^*} - 1\right)$$

$$= 820 \cdot \left(\frac{70^2}{180 \cdot 220} + \frac{60^2}{190 \cdot 220} + \frac{20^2}{150 \cdot 220} + \cdots + \frac{50}{150 \cdot 170} - 1\right)$$

$$= 51{,}244.$$

В нашому випадку $k = 3,\ l = 5$. Відповідно матимемо χ^2 з $(k-1)(l-1) = 8$ ступенями свободи. Тому, для рівня значущості $\alpha = 0{,}05$ з таблиць розкладу χ^2 (Додаток 4) обчислюємо критичне значення $\chi^2_{0{,}05}(8) = 15{,}5$.

Маємо, що $\chi^2 = 51{,}244 > 15{,}5 = \chi^2_{0{,}05}(8)$. Відповідно з ймовірністю 0,95 залежність між випадковими величинами X та Y буде значущою.

Можемо також порахувати коефіцієнт K_p, щоб оцінити тісноту зв'язку:

$$K_p = \left(\frac{\chi^2}{n + \chi^2}\right)^{1/2} = \sqrt{\frac{51{,}244}{820 + 51{,}244}} = 0{,}242.$$

Звідси можемо зробити висновок, що зв'язок між змінними не тісний, але значущий.

5.9. Питання для самоконтролю до Розділу 5

1.Що таке коефіцієнт коваріації та коефіцієнт кореляції (Пірсона)?
2. Запишіть формулу для обчислення скорегованого коефіцієнт кореляції Пірсона.
3. Опишіть алгоритм перевірки гіпотези про нескорельованість для великого обсягу вибірки.
4. Опишіть алгоритм перевірки гіпотези про нескорельованість для малого обсягу вибірки.
5. Як знайти довірчий інтервал для коефіцієнт кореляції Пірсона у випадку великої вибірки?
6. Як знайти довірчий інтервал для коефіцієнт кореляції Пірсона у випадку малої вибірки?
7. Опишіть алгоритм перевірки гіпотези про рівність двох коефіцієнтів кореляції Пірсона.
8. Що таке коефіцієнт кореляції Спірмена?
9. Опишіть алгоритм перевірки гіпотези про нескорельованість за Спірменом для великого обсягу вибірки.
10. Опишіть алгоритм перевірки гіпотези про нескорельованість за Спірменом для малого обсягу вибірки.
11. Що таке коефіцієнт кореляції Кендала?
12. Опишіть алгоритм перевірки гіпотези про нескорельованість за Кендадом.
13. Що таке коефіцієнт конкордації (узгодженості) Кендала?
13. Опишіть алгоритм перевірки гіпотези про рівність нулю коефіцієнта конкордації Кендала.
14. Що таке коефіцієнт конкордації (узгодженості) Шукені?
15. Що таке точково бісеріальний коефіцієнт кореляції?
16. Опишіть алгоритм перевірки гіпотези про точково бісеріальну нескорельованість.
17. Що таке рангово бісеріальний коефіцієнт кореляції?
18. Що таке спряжкність 2×2?
19. Дайте означення коефіцієнта асоціації.
20. Дайте означення коефіцієнта колігації Юла.
21. Дайте означення коефіцієнта подібності Пірсона.
22. Що таке спряжкність $k \times l$?
23. Дайте означення коефіцієнта спряженості Пірсона.
24. Як перевірити гіпотезу незалежності випадкових величин за допомогою коефіцієнта спряженості Пірсона?

РОЗДІЛ 6. РЕГРЕСІЙНИЙ АНАЛІЗ

Розглянуті вище методи дисперсійного і кореляційного аналізу дозволяють виявити наявність зв'язку між випадковими величинами і оцінити тісноту цього зв'язку. Наступним кроком є виявлення конкретного функціонального вигляду зв'язку між випадковими величинами, який у статистиці називають регресією.

Кожній величині, яку дістають в результаті експерименту, притаманний елемент випадковості. При спільній появі двох (або більшої кількості) величин у результаті експерименту постає питання залежності однієї випадкової величини від іншої. За вибірковими даними можна знайти лише оцінку істинної регресії, яка містить похибку, що пов'язана з випадковістю вибірки.

6.1. Рівняння парної лінійної регресії

У цьому підрозділі розглянемо приклад лінійної залежності між двома випадковими величинами X та Y. Багато прикладів такої залежності дає фізика. Так, за законом Ома, $U = R \cdot I$ (де U – напруга, I – сила струму, R – опір провідника). Величини U та I можна виміряти вольтметром та амперметром, а опір R залежить від провідника. Це – теоретична лінійна залежність між величинами U та I з невідомим коефіцієнтом R. Як його визначити? Проводимо вимірювання u та i напруги і сили струму й покладаємо $R = \frac{u}{i}$? Але вимірювання неточні. Тому, розглядаємо U та I як випадкові величини. Отже, природно провести багато вимірювань $u_1, \dots, u_n;\ i_1, \dots, i_n$ в одні й ті самі моменти часу і на їх підставі оцінити R. Розглянемо яким чином можна виконати таку оцінку.

Отже, нехай з теорії ми знаємо, що величини x та y зв'язані лінійною залежністю з невідомими коефіцієнтами $\alpha,\ \beta$, які потрібно оцінити:

$$y = \alpha x + \beta.$$

Але точно x і y ми знайти не можемо, тому – це випадкові величини і на практиці матимемо:

$$y = \alpha x + \beta + \varepsilon. \qquad (6.1)$$

Тут ε – випадкова величина (тобто, усі похибки вимірювань x і y віднесені до ε). Вважаємо, що математичне сподівання $\mathbf{M}\varepsilon = 0$ (тобто, немає систематичних помилок), дисперсія $\mathbf{D}\varepsilon = \sigma^2$ (яка не дуже відома і її визначення обговоримо пізніше), а сама випадкова величина ε – нормально розподілена.

Нехай проведено n вимірювань. Отримали значення $(x_i)_1^n$ та $(y_i)_1^n$.

Графічне зображення на площині точок з координатами $(x_i)_1^n$ та $(y_i)_1^n$ називається ***кореляційним полем***.

На рис. 6.1–6.3 наведені основні можливі представлення кореляційних полів у випадку лінійної кореляційної залежності або її відсутності для двох випадкових величин X та Y.

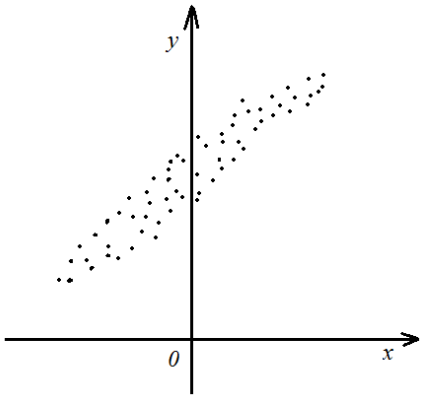


Рис. 6.1

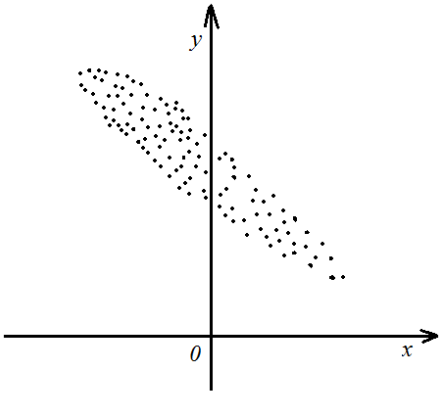


Рис. 6.2

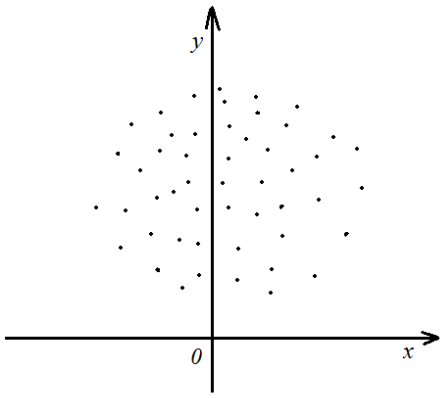


Рис. 6.3

На рис. 6.1 наведений приклад прямого лінійного зв'язку між двома випадковими величинами X та Y. На рис. 6.2 наведений приклад оберненого лінійного зв'язку між X та Y (при збільшенні значень $(x_i)_1^n$ значення $(y_i)_1^n$ зменшуються). На рис. 6.3 наведений приклад, коли кореляційний зв'язок відсутній.

Як знаючи результати вимірювань, оцінити коефіцієнти α, β? Природно, підібрати оцінки

$$\alpha^* = \alpha^*(n)$$

та

$$\beta^* = \beta^*(n)$$

так, щоб y_i та $(\alpha^* x_i + \beta^*)$ відрізнялися якнайменше.

Наведемо ілюстрацію (рис. 6.4).

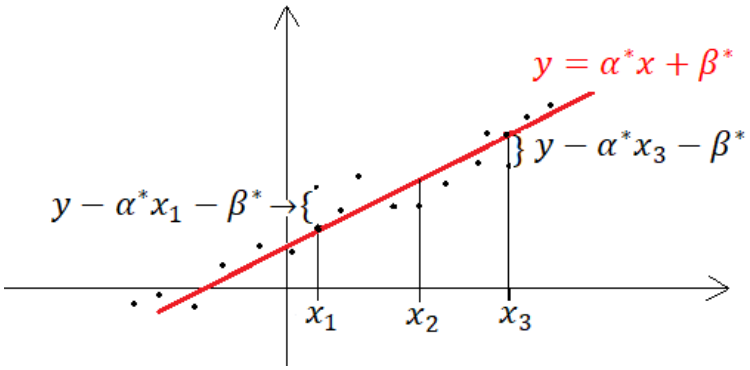


Рис. 6.4

Але, що означає, що величини y_i та $(\alpha x_i + \beta)$ відрізняються якнайменше? Мінімальним має бути $\max_{1 \le i \le n} |y_i - (\alpha x_i + \beta)|$, чи $\sum_{i=1}^n |y_i - (\alpha x_i + \beta)|$? Метод найменших квадратів полягає в тому, що оцінки α^* та β^* потрібно вибирати так, щоб

$$f(\alpha,\beta)=\sum_{i=1}^{n}\left(y_i-(\alpha x_i+\beta)\right)^2\to\min,$$

тобто евклідова відстань між векторами $(y_1,\dots,y_n)$ та $(\alpha x_1+\beta,\dots,\alpha x_n+\beta)$ була мінімальною.

Пригадаємо необхідну умову мінімуму

$$\begin{cases} f'_\beta=0 \\ f'_\alpha=0 \end{cases}.$$

Порахуємо ці похідні

$$\begin{cases} f'_\beta=-2\sum_{i=1}^{n}(y_i-\alpha x_i-\beta)=0 \\ f'_\alpha=-2\sum_{i=1}^{n}(y_i-\alpha x_i-\beta)x_i=0 \end{cases}\Rightarrow$$

$$\begin{cases} n\beta+\left(\sum_{i=1}^{n}x_i\right)\alpha=\sum_{i=1}^{n}y_i \\ \left(\sum_{i=1}^{n}x_i\right)\beta+\left(\sum_{i=1}^{n}x_i^2\right)\alpha=\sum_{i=1}^{n}x_iy_i \end{cases}\xrightarrow[\text{поділимо на } n]{}$$

$$\begin{cases} \beta+\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)\alpha=\frac{1}{n}\sum_{i=1}^{n}y_i \\ \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)\beta+\frac{1}{n}\left(\sum_{i=1}^{n}x_i^2\right)\alpha=\frac{1}{n}\sum_{i=1}^{n}x_iy_i \end{cases}. \qquad (6.2)$$

Позначимо

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i;\quad \bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i;$$

$$\sigma_x^2=\frac{1}{n}\sum_{i=1}^{n}x_i^2-(\bar{x})^2;\quad \sigma_y^2=\frac{1}{n}\sum_{i=1}^{n}y_i^2-(\bar{y})^2;$$

$$K_{xy}=\frac{1}{n}\sum_{i=1}^{n}x_i\cdot y_i-\bar{x}\cdot\bar{y}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y});$$

$$r_{xy}=\frac{K_{xy}}{\sigma_x\sigma_y}.$$

Тоді система (6.2) набуде вигляду:

$$\begin{cases} \beta + \bar{x}\alpha = \bar{y} \\ \bar{x}\beta + \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right)\alpha = \frac{1}{n}\sum_{i=1}^{n} x_i y_i \end{cases}.$$

В отриманій системі віднімемо від 2-го рівняння перше, помножене на $\bar{x}$. В результаті, після елементарних перетворень отримаємо формули для коефіцієнтів рівняння регресії:

$$\alpha^* = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\cdot\bar{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2} = \frac{K_{xy}}{\sigma_x^2}$$

$$\beta^* = \bar{y} - \bar{x}\alpha^*.$$

Помножимо ліву і праву частину першої рівності на $\frac{\sigma_x}{\sigma_y}$, дістанемо

$$\frac{\sigma_x}{\sigma_y}\alpha^* = \frac{K_{xy}}{\sigma_x^2}\cdot\frac{\sigma_x}{\sigma_y} = \frac{K_{xy}}{\sigma_x\sigma_y} = r_{xy} \qquad \Rightarrow \qquad \alpha^* = r_{xy}\frac{\sigma_x}{\sigma_y}.$$

Тоді

$$\beta^* = \bar{y} - \alpha^*\cdot\bar{x} = \bar{y} - r_{xy}\frac{\sigma_x}{\sigma_y}\cdot\bar{x}.$$

Алгоритм застосування методу найменших квадратів

Провівши спостереження, отримуємо результати: $(y_i)_1^n$, $(x_i)_1^n$. Обчислюємо

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,; \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Тоді

$$\alpha^* = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\cdot\bar{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2};$$

$$\beta^* = \bar{y} - \bar{x}\alpha^*.$$

Достатня умова мінімуму

Пригадаймо з математичного аналізу, що рівності нулю частинних похідних – це лише необхідна умова мінімуму. Для того, щоб при знайдених α^*, β^* функція набувала найменшого значення, потрібно, щоб у цій точці матриця других частинних похідних для $f(\alpha,\beta)$ була додатно визначеною.

Порахуємо другі частинні похідні:

$$f''_{\alpha^2} = \left(-2\sum_{i=1}^{n}(y_i - \alpha x_i - \beta)x_i\right)'_{\alpha} = 2\sum_{i=1}^{n} x_i^2,$$

$$f''_{\alpha\beta} = \left(-2\sum_{i=1}^{n}(y_i - \alpha x_i - \beta)x_i\right)'_{\beta} = 2\sum_{i=1}^{n} x_i,$$

$$f''_{\beta^2} = \left(-2\sum_{i=1}^{n}(y_i - \alpha x_i - \beta)\right)'_{\beta} = 2n.$$

Отже, матриця других частинних похідних матиме вигляд

$$M = \begin{bmatrix} 2n & 2\sum_{i=1}^{n} x_i \\ 2\sum_{i=1}^{n} x_i & 2\sum_{i=1}^{n} x_i^2 \end{bmatrix}.$$

Кутовий мінор 1-го порядку $2n > 0$.

Кутовий мінор 2-го порядку

$$\det M = 4\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right].$$

Якщо в нерівності Коші-Буняковського:

$$\sum_{i=1}^{n} a_i b_i \le \sqrt{\sum_{i=1}^{n} a_i^2} \cdot \sqrt{\sum_{i=1}^{n} b_i^2},$$

покладемо $a_i \equiv 1,\ b_i = x_i$, то побачимо, що число $\det M$ невід'ємне. Пригадаємо також, що нерівність Коші-Буняковського обертається в рівність лише коли вектори $(a_1, \dots, a_n)$ і $(b_1, \dots, b_n)$ лінійно залежні. Але а нашому випадку, за побудовою вектори $(1, \dots, 1)$ та $(x_1, \dots, x_n)$ лінійно незалежні. Тому число $\det M$ додатне.

Таким чином, обидва кутові мінори додатні, звідки й випливає, що отримані коефіцієнти дають мінімум.

Статистичний аналіз коефіцієнтів регресії

Статистичні висновки щодо коефіцієнта α регресії $y = \alpha x + \beta$ можна отримати за допомогою статистики

$$t_\alpha = \frac{\alpha^* - \alpha}{S_\alpha},$$

де

$$S_\alpha = \frac{S}{s_X\sqrt{n-1}},$$

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \beta^* - \alpha^* x_i)^2,$$

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,,$$

α – істине значення коефіцієнта регресії; α^* – його вибіркова оцінка.

Статистика t_α при справедливості нульової гіпотези $H_0\colon \alpha = \alpha^*$ має розподіл Стьюдента з $(n-2)$ ступенями свободи.

Отже, за допомогою квантилів розподілу Стьюдента можна перевірити гіпотезу рівності α заданому значенню, гіпотезу про значущість коефіцієнтів регресії (суттєвості його відхилення від нуля), побудувати довірчий інтервал для коефіцієнта α. Значення коефіцієнта α регресії є значущим з рівнем значущості $\gamma = 1 - \alpha$, якщо

$$|\alpha^*| > t_{\frac{1+\gamma}{2}} S_\alpha\,.$$

Гіпотеза про рівність коефіцієнта β заданому значенню приймається, якщо

$$|\alpha - \alpha^*| < t_{\frac{1+\gamma}{2}} S_\alpha\,.$$

І, нарешті, двосторонній $\gamma \cdot 100\%$ довірчий інтервал для α має вигляд

$$\alpha^* - S_\alpha \cdot t_{\frac{1+\gamma}{2}} < \alpha < \alpha^* + S_\alpha \cdot t_{\frac{1+\gamma}{2}}.$$

Статистичні висновки щодо коефіцієнта β можуть бути отримані за допомогою статистики

$$t_\beta = \frac{\beta^* - \beta}{S_\beta}, \qquad S_\beta = S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_X^2}},$$

де, β^* – відповідно вибіркова оцінка, а β істинне значення коефіцієнта; S та s_X визначенні вище для t_α.

При $H_0\colon \beta^* = \beta$ статистика t_β має розподіл Стьюдента з $(n-2)$ ступеннями свободи. Перевірка гіпотез про значення коефіцієнта β і побудова довірчих інтервалів для нього виконуються за аналогією з коефіцієнтом α.

Коефіцієнт парної лінійної кореляції та коефіцієнт детермінації

Пригадаймо з попереднього підрозділу, що

$$r_{xy} = \alpha^* \cdot \frac{\sigma_x}{\sigma_y} = \frac{K_{xy}}{\sigma_x \sigma_y}.$$

Тобто, r_{xy} – це вже добре відомий нам коефіцієнт вибіркової кореляції з підрозділу 1.4 та коефіцієнт кореляції Пірсона (див. (5.1)). У лінійному регресійному аналізі він носить назву ***коефіцієнта парної лінійної кореляції*** і оцінює тісноту лінійного зв'язку між випадковими величинами.

Оцінку адекватності побудованої моделі експериментальним спостереженням дає коефіцієнт детермінації, а також середня похибка апроксимації. Наведемо відповідні означення.

Частку дисперсії, яка пояснюється регресією, у загальній дисперсії результуючої ознаки y характеризує ***коефіцієнт детермінації***:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i^*)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

де $y_i^* = \alpha^* x_i + \beta^*$ – теоретичні (отримані на основі моделі парної лінійної регресії) значення.

Середня похибка апроксимації

$$\bar{A} = \frac{1}{n} \cdot \sum \left| \frac{y_i - y_i^*}{y_i} \right| \cdot 100\%$$

оцінює середнє відхилення підрахованих (теоретичних) значень від фактичних. Допустима границя значень $\bar{A}$ – не більше 8-10%.

6.2. Лінеаризація нелінійної моделі заміною змінних

Якщо лінійне рівняння регресії неадекватно описує наявний зв'язок між ознаками, то потрібно знайти нелінійне рівняння регресії, яке гарно описуватиме істину залежність. Вид функції регресії обирається відштовхуючись від особливостей досліджуваного явища, а також із загального графічного аналізу залежності між ознаками (на основі кореляційного поля).

Суть методу лінеаризації – це перетворення нелінійного рівняння на лінійне переходом від досліджуваних змінних до нових змінних. Найпоширеніші види перетворень наведено у наступній таблиці

Таблиця. Лінеаризуючі функціональні перетворення $(y^* = \beta^* + \alpha^* x^*)$

Вихідна залежність $y = f(x)$	Перетворення змінних		Перетворення коефіцієнтів	
	y^*	x^*	β^*	α^*
$y = \beta^* + \frac{\alpha^*}{x}$	y	$\frac{1}{x}$	β^*	α^*
$y = \frac{\beta^*}{\alpha^* + x}$,	$\frac{1}{y}$	x	$\frac{\beta^*}{\alpha^*}$	$\frac{1}{\beta^*}$
$y = \frac{\beta^* x}{\alpha^* + x}$	$\frac{1}{y}$	$\frac{1}{x}$	$\frac{\beta^*}{\alpha^*}$	$\frac{1}{\beta^*}$
$y = \frac{x}{\beta^* + \alpha^* \cdot x}$	$\frac{x}{y}$	x	β^*	α^*
$y = \beta^* \cdot \alpha^{*x}$	$\lg y$	x	$\lg \beta^*$	$\lg \alpha^*$
$y = \beta^* \cdot x^{\alpha^*}$	$\lg y$	$\lg x$	$\lg \beta^*$	α^*
$y = \beta^* \cdot e^{\alpha^* \cdot x}$	$\ln y$	x	$\ln \beta^*$	α^*
$y = \beta^* \cdot e^{\alpha^* / x}$	$\ln y$	1	$\ln a\beta^*$	α^*
$y = \beta^* + \alpha^* \cdot x^n$	y	x^n	β^*	α^*

Обробка результатів спостережень, обчислення регресії та її статистичний аналіз для лінійно перетвореного рівняння виконується методами лінійного регресійного аналізу, які розглянуто вище.

Для оцінки тісноти нелінійного зв'язку випадкових величин використовується ***індекс кореляції***:

$$\rho_{xy} = \sqrt{1 - \frac{\sum_{i=1}^{n}\left(y_i - y_i^*\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}},$$

$$0 \le \rho_{xy} \le 1.$$

6.3. Множинна лінійна регресія

На практиці часто залежна змінна y пов'язана з впливом не одного, а кількох аргументів. У такому випадку регресію називають ***множинною***.

Якщо аргументи в функції регресії містяться в першому степені, то множинна регресія називається ***множинною лінійною регресією***, у супротивному випадку — ***множинною нелінійною регресією***.

Розглянемо лінійну залежність y від m аргументів $(x_1, x_{2,\dots,} x_m)$:

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_m \cdot x_m \,. \qquad (6.3)$$

Нехай проведено n емпіричних спостережень, для яких отримані наступні набори даних:

$$y_i = \beta_0 + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \cdots + \beta_m \cdot x_{mi}, \quad i = \overline{1,n}. \tag{6.4}$$

Таким чином отримаємо систему лінійних рівнянь

$$\begin{cases} y_1 = \beta_0 + \beta_1 \cdot x_{11} + \beta_2 \cdot x_{21} + \cdots + \beta_m \cdot x_{m1} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 \cdot x_{12} + \beta_2 \cdot x_{22} + \cdots + \beta_m \cdot x_{m2} + \varepsilon_2 \\ y_3 = \beta_0 + \beta_1 \cdot x_{13} + \beta_2 \cdot x_{23} + \cdots + \beta_m \cdot x_{m3} + \varepsilon_3 \\ \cdots \\ y_n = \beta_0 + \beta_1 \cdot x_{1n} + \beta_2 \cdot x_{2n} + \cdots + \beta_m \cdot x_{mn} + \varepsilon_n \end{cases} \tag{6.5}$$

де ε_i – попарно нескорельовані нормально розподілені випадкові величини з $\mathbf{M}(\varepsilon_i) = 0$ і $\mathbf{D}(\varepsilon_i) = \sigma_i^2, \ i = \overline{1,n}.$

Основна задача буде полягати у тому, щоб отримати такі оцінки β_j^* параметрів β_j, де $j = \overline{0,m}$, при яких сума квадратів відхилень (ε_i) фактичних значень y_i ознаки Y від розрахованих y_i^* була б мінімальною:

$$\sum_{i=1}^{n}(y_i - y_i^*)^2 = \sum_{i=1}^{n}(y_i - \beta_0^* - \beta_1^* \cdot x_{1i} - \beta_2^* \cdot x_{2i} - \cdots - \beta_m^* \cdot x_{mi})^2 = \sum_{i=1}^{n}\varepsilon_i^2 \to \min.$$

Запишемо систему (6.5) у векторно-матричній формі:

$$\mathrm{y} = \mathrm{X} \cdot \beta + \varepsilon, \tag{6.6}$$

де

$$\mathrm{y} = \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix}; \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \cdots \\ \beta_m \end{pmatrix}; \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{pmatrix}; \quad \mathrm{X} = \begin{pmatrix} 1\ x_{11} \cdots x_{1m} \\ 1\ x_{21} \cdots x_{2m} \\ \cdots \\ 1\ x_{n1} \cdots x_{nm} \end{pmatrix}$$

Матрицю X розмірністю $(m+1) \times n$ називають ***регресійною матрицею***, а її елементи x_{ij} – ***регресорами***.

Як і у випадку парної лінійної регресії, параметри рівняння (6.3) є величинами сталими, але невідомими. Ці параметри ми також замінимо статистичними точковими оцінками $\beta_0^*, \beta_1^*, \beta_2^*, \ldots, \beta_m^*$, які можна отримати шляхом обробки результатів вибірки. Вони є величинами випадковими. Таким чином, рівності (6.4) відповідатиме статистична оцінка

$$y_i = \beta_0^* + \beta_1^* \cdot x_{1i} + \beta_2^* \cdot x_{2i} + \cdots + \beta_m^* \cdot x_{mi}. \tag{6.7}$$

Статистична оцінка для вектора y буде визначатись вектором

$$\mathrm{y} = \mathrm{X} \cdot \beta^* + \varepsilon, \tag{6.8}$$

де

$$\beta^* = \begin{pmatrix} \beta_0^* \\ \beta_1^* \\ \beta_2^* \\ \cdots \\ \beta_m^* \end{pmatrix}.$$

Із (6.8) можемо отримати вектор похибок:

$$\varepsilon = y - X \cdot \beta^*.$$

Тут, аналогічно, для визначення компонентів вектора β^* застосовується метод найменших квадратів на основі мінімізації суми квадратів усіх похибок, а саме величини $\varepsilon' \cdot \varepsilon$, де ε' – транспонований вектор до ε. В результаті отримаємо, що

$$\boxed{\beta^* = (X' \cdot X)^{-1} \cdot X \cdot y,} \qquad (6.9)$$

де X' — матриця, транспонована до X.

Ранжування факторів

Якби усі пояснюючі змінні x_j $(j = \overline{1,m})$ у рівнянні (6.3) вимірювалися в одних і тих самих одиницях, наприклад, в кг, то безпосередньо співставляючи абсолютні значення отриманих коефіцієнтів регресії β_j^* $(j = \overline{1,m})$, можна було б ранжувати фактори x_j за силою їх дії на y.

Однак у загальному випадку змінні x_j мають різні одиниці виміру і таке ранжування не коректне. У такому разі використовують процедуру нормування коефіцієнтів регресії, тобто обчислюють ***стандартизовані коефіцієнти регресії*** a_j за наступною формулою:

$$a_j = \beta_j^* \cdot \frac{s_{x_j}}{s_Y} \quad (j = \overline{1,m}),$$

де a_j — коефіцієнт регресії після нормування; s_{x_j} – виправлене середнє квадратичне відхилення змінної x_j; s_Y – виправлене середнє квадратичне відхилення ознаки Y.

Маючи значення a_j, можна побудувати рівняння множинної регресії (6.3) у ***стандартизованому масштабі***

$$t_y = a_1 \cdot t_{x_1} + a_2 \cdot t_{x_2} + \cdots + a_m \cdot t_{x_m},$$

де

$$t_y = \frac{y - \bar{y}}{s_Y}; \qquad t_{x_j} = \frac{x_j - \bar{x_j}}{S_{x_j}} \quad (j = \overline{1,m})$$

стандартизовані змінні, для яких середні значення дорівнюють нулю $(\bar{t}_y = \bar{t}_{x_j} = 0)$, а виправлені середні квадратичні відхилення дорівнюють одиниці $(S_{t_y} = S_{t_{x_j}} = 1)$. $\bar{x_j}$ – середнє значення для змінної x_j.

Чим більший модуль $|a_j|$, тим сильніший вплив фактора x_j на ознаку Y, тобто за значеннями $|a_j|$ можна виконати безпосереднє ранжування факторів за силою їхньої дії на y.

Оцінка якості рівняння множинної регресії

За аналогією з парною регресією можна визначити долю результату, яка пояснюється варіацією включених у модель факторів в його загальній дисперсії:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i^*)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}\varepsilon_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Величину R^2 називають ***коефіцієнтом множинної детермінації***. Він служить вимірювачем якості підбору рівняння регресії. Його значення знаходяться у межах від 0 до 1. Чим ближче значення R^2 до одиниці, тим краще рівняння регресії пояснює поведінку y.

Крім коефіцієнта R^2 використовують також інший показник якості моделі – ***коефіцієнт множинної кореляції***

$$R = \sqrt{R^2},$$

який являє собою узагальнення парного коефіцієнту кореляції r_{yx} і характеризує сумісний вплив усіх факторів на результат y. На відміну від парного коефіцієнту кореляції r_{yx} коефіцієнт множинної кореляції R набуває значень у межах від 0 до 1 і не може бути використаний для інтерпретації напрямку зв'язку.

Коефіцієнт R^2 є неспадаючою функцією від кількості пояснюючих змінних. Якщо додати до моделі фактор, який зовсім не впливає на y, то R^2 обов'язково автоматично збільшиться. Цей недолік можна усунути, якщо визначати показник R^2 не через суми квадратів відхилень, а через дисперсії на один степінь свободи. В результаті отримаєм ***нормований коефіцієнт множинної детермінації***:

$$\hat{R}^2 = 1 - \frac{\frac{\sum_{i=1}^{n}\varepsilon_i^2}{(n-m-1)}}{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{(n-1)}} = 1 - \frac{n-1}{n-m-1} \cdot (1 - R^2).$$

Відомо, що $\hat{R}^2$ збільшується при додаванні нового фактору у модель тоді і тільки тоді, коли модуль t-статистики параметра за цією змінною більший від одиниці. Значення $\hat{R}^2$ може навіть зменшуватися при додаванні нового фактора.

Перевірка якості моделі множинної лінійної регресії

Перевірка статистичної якості моделі виконується шляхом перевірки сумісної значущості її коефіцієнтів, тобто виконується перевірка наступної гіпотези:

$$H_0: \ \beta_1 = \beta_2 = \cdots = \beta_m = 0.$$

На практиці замість вказаної гіпотези перевіряють тісно пов'язану з нею гіпотезу про статистичну значущість коефіцієнта детермінації R^2:

$$H_0: \ R^2 = 0.$$

Для перевірки даної гіпотези використовується статистика:

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - m - 1}{m},$$

яка має розподіл Фішера.

Отримане значення статистики F порівнюється з критичним значенням F_α, яке знаходиться з таблиці розподілу Фішера (Додаток 5) за заданим рівнем значущості α і кількістю ступенів свободи $k_1 = m$ та $k_2 = (n - m - 1)$. Якщо $F > F_\alpha$, то гіпотеза H_0 відхиляється, тобто рівняння множинної лінійної регресії статистично значуще.

Як і у випадку парної регресії виконується перевірка статистичної значущості окремих коефіцієнтів рівняння на основі t-статистик:

$$t_{\beta_j^*} = \frac{\beta_j^*}{S_{\beta_j^*}}, \quad j = \overline{0, m},$$

де $S_{\beta_j^*}$ - стандартна похибка параметра β_j^*, яка обчислюється за формулою:

$$\mathrm{S}_{\beta_j^*} = \mathrm{S} \cdot \sqrt{[(\mathrm{X}' \cdot \mathrm{X})^{-1}]_{j+1\ j+1}},$$

де $[(\mathrm{X}' \cdot \mathrm{X})^{-1}]_{j+1\ j+1}$ – діагональний елемент оберненої матриці $(\mathrm{X}' \cdot \mathrm{X})^{-1}$, який стоїть на перетині $(j+1)$-го рядка і $(j+1)$-го стовпчика; S^2 – незміщена оцінка дисперсії σ^2 збурення ε, яка визначається за формулою:

$$S^2 = \frac{\sum_{i=1}^{n} \varepsilon_i^2}{n - m - 1}.$$

Якщо $\left|t_{\beta_j^*}\right| > t_{\text{кр}}$, де $t_{\text{кр}}$ знаходиться із таблиць по значенню $\frac{1+\gamma}{2}$ $(\gamma = 1 - \alpha)$ із кількістю ступенів свободи $k = n - m - 1$, то коефіцієнт β_j^* вважається статистично значущим.

Наведену строгу перевірку значущості коефіцієнтів можна замінити простим порівняльним аналізом:

- ✓ Якщо $\left|t_{\beta_j^*}\right| \leq 1$, то β_j^* статистично незначущий;
- ✓ Якщо $1 \leq \left|t_{\beta_j^*}\right| \leq 2$, то β_j^* відносно значущий, і для уточнення слід скористатися строгою методикою;
- ✓ Якщо $2 \leq \left|t_{\beta_j^*}\right| \leq 3$, то β_j^* статистично значущий;
- ✓ Якщо $\left|t_{\beta_j^*}\right| > 3$, то β_j^* вважається сильно значущим і ймовірність похибки не перевищує 0,001.

Аналогічно як і у випадку парної лінійної регресії для статистично значущих коефіцієнтів моделі можна побудувати довірчі інтервали:

$$\beta_j^* - t_{\frac{1+\gamma}{2}} \cdot S_{\beta_j^*} \leq \beta_j \leq \beta_j^* + t_{\frac{1+\gamma}{2}} \cdot S_{\beta_j^*}.$$

Довірчий інтервал можна побудувати і для індивідуальних прогнозних значень залежної змінної y.

Зафіксуємо значення прогнозних пояснюючих змінних

$$x_{10}, x_{20}, \ldots, x_{p0}$$

і за вектор-стовпчиком

$$x_0 = (1,\ x_{10},\ \ x_{20}, \dots, x_{p0}\ )'$$

знайдемо прогнозні значення залежної змінної y:

$$\hat{y}_0 = \beta_0^* + \beta_1^* \cdot x_{10} + \beta_2^* \cdot x_{20} + \cdots + \beta_m^* \cdot x_{m0}.$$

Відповідно довірчий інтервал для індивідуального прогнозного значення $\hat{y}_0$ по точці x_0 знаходиться за формулою:

$$\hat{y}_0 - t_{\frac{1+\gamma}{2}} \cdot S_{\hat{y}_0} \le y_0 \le \hat{y}_0 + t_{\frac{1+\gamma}{2}} \cdot S_{\hat{y}_0},$$

де стандартна похибка $S_{\hat{y}_0}$ обчислюється за формулою:

$$S_{\hat{y}_0} = S \cdot \sqrt{1 + x_0' \cdot (X' \cdot X)^{-1} \cdot x_0}.$$

6.4. Приклади до Розділу 6

Приклад 1. Парна лінійна регресія

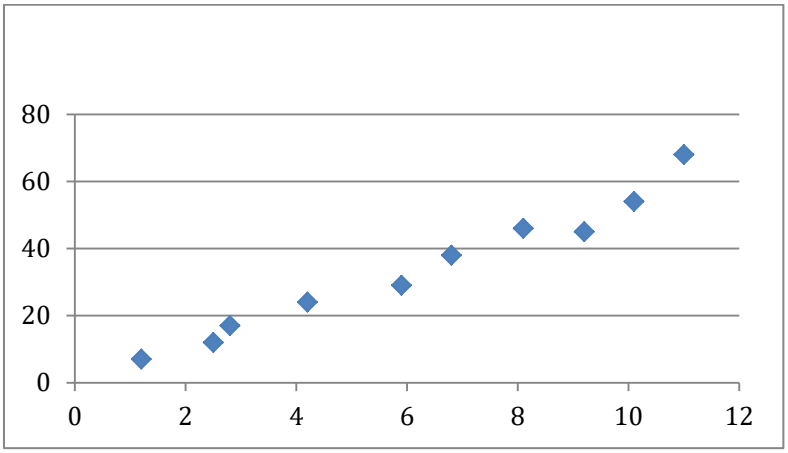
Для сукупності даних знайти оцінки коефіцієнтів α і β регресії $y = \beta + \alpha \cdot x$ та провести їхній статистичний аналіз при рівні значущості $\alpha = 0{,}05$ (не плутати з α – коефіцієнтом регресії!).

x_i	1,2	2,5	2,8	4,2	5,9	6,8	8,1	9,2	10,1	11,0
y_i	7	12	17	24	29	38	46	45	54	68

Оцінити адекватність отриманого рівняння регресії на основі коефіцієнту детермінації.

Розв'язання.

Спочатку побудуємо кореляційне поле і візуально оцінимо вид залежності:



Кореляційне поле чітко вказує на лінійну залежність і демонструє прямий зв'язок між ознаками.

Обчислимо оцінку α^*:

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i = \frac{1{,}2 \cdot 7 + 2{,}5 \cdot 12 + 2{,}8 \cdot 17 + \cdots + 11 \cdot 68}{10} = 269{,}63;$$

$$\bar{x} = \frac{1{,}2 + 2{,}5 + 2{,}8 + \cdots + 11}{10} = 6{,}18; \quad \bar{y} = \frac{7 + 12 + 17 + \cdots + 68}{10} = 34;$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{1{,}2^2 + 2{,}5^2 + 2{,}8^2 + \cdots + 11^2}{10} = 48{,}75;$$

$$\alpha^* = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x} \cdot \bar{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2} = \frac{269{,}63 - 6{,}18 \cdot 34}{48{,}75 - 6{,}18^2} = 5{,}64;$$

$$\beta^* = \bar{y} - \bar{x} \cdot \alpha^* = 34 - 6{,}18 \cdot 5{,}64 = -0{,}84.$$

Далі, перевіримо значущість одержаних коефіцієнтів (суттєвість їх відхилення від нуля). Спочатку обчислимо

$$S_x^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 \frac{(1{,}2 - 6{,}18)^2 + (2{,}5 - 6{,}18)^2 + \cdots + (11 - 6{,}18)^2}{9} = 11{,}73;$$

$$S_x = \sqrt{11{,}73} = 3{,}42.$$

Далі, обчислимо значення дисперсії

$$S^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - y_i^*)^2,$$

де

$$y_i^* = \beta^* + \alpha^* \cdot x_i.$$

У нашому випадку

x_i	y_i	$y_i^* = -0{,}84 + 5{,}64 \cdot x_i$	$(y_i - y_i^*)^2$
1,2	7	5,92	1,16
2,5	12	13,25	1,57
2,8	17	14,94	4,23
4,2	24	22,84	1,35
5,9	29	32,42	11,71
6,8	38	37,50	0,25
8,1	46	44,82	1,38
9,2	45	51,03	36,31
10,1	54	56,10	4,41
11	68	61,17	46,59
		$\sum_{i=1}^{n} (y_i - y_i^*)^2 =$	**108,97**

$$S^2 = \frac{1}{8} \sum_{i=1}^{10} (y_i - y_i^*)^2 = 13{,}62; \quad S = 3{,}69.$$

$$S_\alpha = \frac{S}{S_x \sqrt{n-1}} = \frac{\sqrt{13{,}62}}{3{,}42 \cdot 3} = 0{,}36;$$

$$S_\beta = S \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)S_x^2}} = 3{,}69 \cdot \sqrt{\frac{1}{10} + \frac{6{,}18^2}{9 \cdot 11{,}73}} = 2{,}51.$$

Для рівня надійності $\gamma = 1 - \alpha = 0{,}95$ маємо

$$t_{\frac{1+0{,}95}{2}}(n-2) = t_{0{,}975}(8) = 2{,}896.$$

Перевіримо значимість коефіцієнта α:

$$|\alpha^*| = 5{,}64 > t_{0.975}(8) \cdot S_\alpha = 2{,}896 \cdot 0{,}36 = 1{,}04,$$

відповідно, з рівнем значущості 0,05 робимо висновок про значимість коефіцієнта регресії.

І, нарешті, довірчий інтервал для α дорівнює

$$5{,}64 - 2{,}896 \cdot 0{,}36 \le \alpha \le 5{,}64 + 2{,}896 \cdot 0{,}36;$$

$$4{,}6 \le \alpha \le 6{,}68.$$

Аналогічні задачі розглянемо для коефіцієнта β. Перевіримо гіпотезу $H_0: \beta = 0$. У нашому випадку

$$|\beta^*| = 0{,}84 < t_{0.975} \cdot S_\beta = 2{,}896 \cdot 2{,}51 = 7{,}27.$$

Відповідно, коефіцієнт β з надійністю 0,95 не відрізняються суттєво від нуля, тобто його значення може бути прирівняне до нуля.

Двосторонній довірчий інтервал для β має вигляд

$$-0{,}84 - 7{,}27 \le \beta \le -0{,}84 - 7{,}27;$$

$$-8{,}11 \le \beta \le 6{,}43.$$

Таким чином, рівняння регресії y по x адекватно відображається рівнянням

$$y = 5{,}64 \cdot x.$$

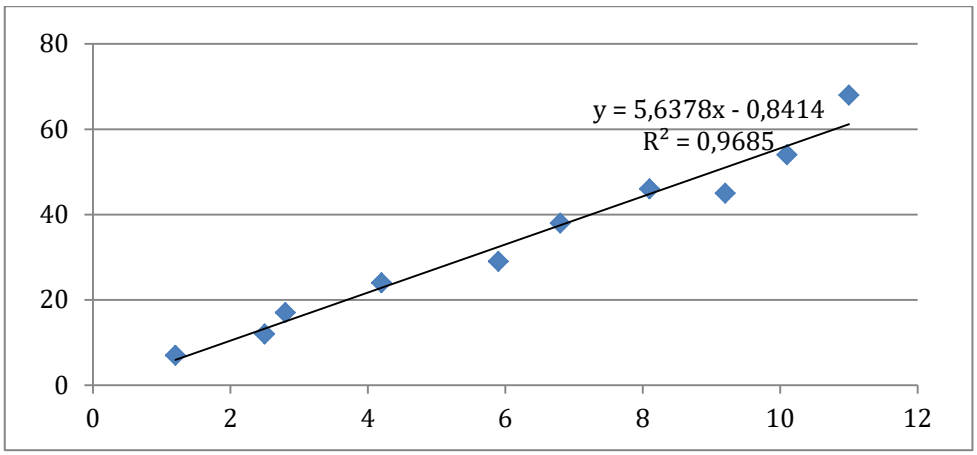
Для оцінки якості отриманої моделі порахуємо також коефіцієнт детермінації:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = (7 - 34)^2 + (12 - 34)^2 + \cdots + (68 - 34)^2 = 3464;$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - y_i^*)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{108{,}97}{3464} = 0{,}969.$$

Як видно, коефіцієнт детермінації дуже високий, що говорить про адекватність отриманої моделі регресії.

Візуалізуємо отримане рівняння регресії на нашому кореляційному полі. Дійсно, розсіювання спостережень навколо прямої регресії мінімальне.



Приклад 2. Нелінійна регресія (лінеаризація)

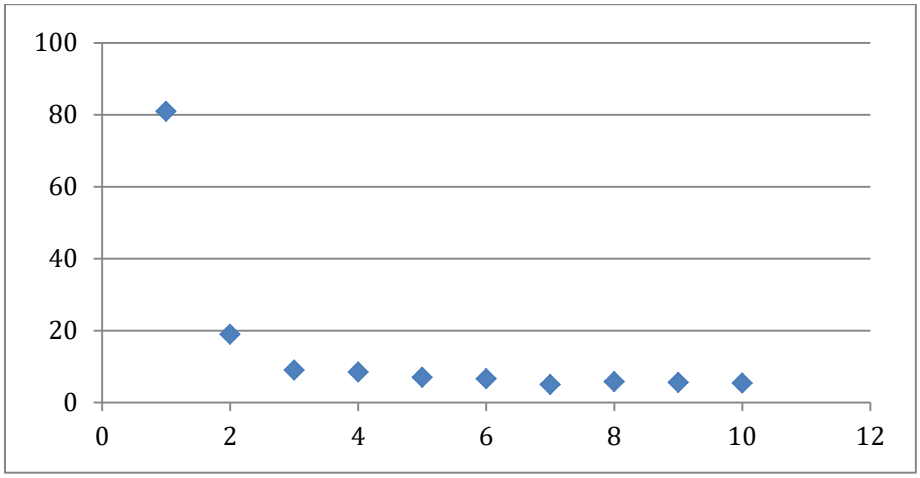
У результаті дослідження залежності між випадковими величинами x та y отримано наступні результати ($n = 10$):

x_i	1	2	3	4	5	6	7	8	9	10
y_i	81	19	9	8,5	7	6,6	5	5,8	5,6	5,4

Необхідно знайти функцію регресії $y = f(x)$ і провести її статистичний аналіз.

Розв'язання.

Спочатку побудуємо кореляційне поле і візуально оцінимо вид залежності:



Видно, що тут можемо використати лінеаризуюче перетворення для функції типу $y = \beta^* \cdot e^{\alpha^*/x}$.

Позначимо

$$y^* = \ln x$$

та

$$x^* = \frac{1}{x}.$$

Тоді знайдемо лінійну регресію $y^* = \beta^* + \alpha^* x^*$.

Для x^* і y^* маємо ряд

$x^* \rightarrow$	1	1/2	1/3	1/4	1/5	1/6	1/7	1/8	1/9	1/10
$y^* \rightarrow$	4,39	2,94	2,20	2,14	1,94	1,89	1,79	1,76	1,72	1,69

Знаходимо

$$\sum_{i=1}^{10} x_i^* = 2{,}93; \quad \left(\sum_{i=1}^{10} x_i^*\right)^2 = 8{,}58;$$

$$\sum_{i=1}^{10} {x_i^*}^2 = 1{,}55; \quad \sum_{i=1}^{10} y_i^* = 22{,}46; \quad \sum_{i=1}^{10} x_i^* \cdot y_i^* = 8{,}67.$$

Тоді

$$\alpha^* = \frac{10 \cdot 8{,}67 - 2{,}93 \cdot 22{,}46}{10 \cdot 1{,}55 - 8{,}58} = 3{,}02; \quad \beta^* = \frac{22{,}46 - 3{,}02 \cdot 2{,}93}{10} = 1{,}361.$$

Відповідно, шукана регресія має вигляд

$$y^* = 1{,}36 + 3{,}02 \cdot x^*$$

або

$$\ln y = 1{,}36 + 3{,}02 \cdot \frac{1}{x} \quad \Rightarrow \quad y = 3{,}9 \cdot e^{\frac{3{,}02}{x}}.$$

Перевіримо значимість коефіцієнтів регресії. Маємо

$$S^2 = \frac{1}{n-2} \cdot \sum_{i=1}^{n} (y_i^* \cdot 1{,}36 - 3{,}02 \cdot x^*)^2 = 0{,}004; \quad (S = 0{,}067);$$

$$S_{x^*}^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} ({x_i}^* - \bar{x}^*)^2 = 0{,}077; \quad (S_{x^*} = 0{,}278);$$

$$S_{\alpha^*} = \frac{S}{S_{x^*} \cdot \sqrt{n-1}} = 0{,}08.$$

Для рівня надійності $\gamma = 1 - \alpha = 0{,}95$ маємо

$$t_{\frac{1+0{,}95}{2}}(f = 8) = t_{0.975}(8) = 2{,}896.$$

Так як $|\alpha^*| = 3{,}02 > t_{0.975}(8) \cdot S_{\alpha^*} = 2{,}896 \cdot 0{,}077 = 0{,}22$, коефіцієнт регресії приймається значущим.

Для дисперсії коефіцієнта β^* маємо

$$S_{\beta^*} = S \cdot \left\{\frac{1}{n} + \frac{{\bar{x}^*}^2}{(n-1) \cdot S_{x^*}^2}\right\}^{\frac{1}{2}} = 0{,}067 \cdot \left\{\frac{1}{10} + \frac{0{,}299^2}{9 \cdot 0{,}077}\right\}^{\frac{1}{2}} = 0{,}032.$$

Тоді $|\beta^*| = 1{,}361 > t_{0.975}(8) \cdot S_{\beta^*} = 2{,}896 \cdot 0{,}032 = 0{,}093$, що також приводить до висновку про значимість коефіцієнта β^* з рівнем значущості $\alpha = 0{,}95$.

Порахуємо теоретичні значення y_i^* для отриманого рівняння регресії та квадрати відхилень емпіричних значень від теоретичних

x_i	y_i	$y_i^* = 3{,}9 \cdot e^{\frac{3{,}02}{x_i}}$	$(y_i - y_i^*)^2$
1	81	79,9	1,175
2	19	17,7	1,811
3	9	10,7	2,796
4	8,5	8,3	0,041
5	7	7,1	0,018
6	6,6	6,5	0,022
7	5	6,0	1,008
8	5,8	5,7	0,012
9	5,6	5,5	0,021
10	5,4	5,3	0,016
		$\sum_{i=1}^{n}(y_i - y_i^*)^2 =$	6,920

Для оцінки тісноти нелінійного зв'язку знайдемо індекс кореляції:

$$\bar{y} = \frac{81 + 19 + 9 + \cdots + 5{,}4}{10} = 15{,}29;$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = (81 - 15{,}29)^2 + (10 - 15{,}29)^2 + \cdots + (5{,}4 - 15{,}29)^2 = 4949{,}13;$$

$$\rho_{xy} = \sqrt{1 - \frac{\sum_{i=1}^{n}(y_i - y_i^*)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \sqrt{1 - \frac{6{,}92}{4949{,}13}} = 0{,}999.$$

Бачимо, що індекс кореляції практично дорівнює одиниці, що говорить про дуже тісний, практично функціональний, нелінійний зв'язок між x та y.

Приклад 3. Множинна лінійна регресія

Ознака Y — лінійно залежна від трьох змінних. Результати спостережень наведено в таблиці:

i	y_i	x_{i1}	x_{i2}	x_{i3}
1	6	1	1	2
2	8	2	2	1
3	14	1	0	0
4	20	3	2	1
5	26	5	2	2

Необхідно:

1) знайти компоненти вектора β^* і побудувати лінійну функцію регресії

$$y_i = \beta_0^* + \beta_1^* \cdot x_{i1} + \beta_2^* \cdot x_{i2} + \beta_3^* \cdot x_{i3}.$$

2) обчислити коефіцієнт детермінації R^2;

3) оцінити ефективність впливу на ознаку Y незалежних змінних x_1, x_2, x_3.

Розв'язання.

1) З умови задачі вектор $\vec{y}$ та регресійна матриця X мають наступний вигляд:

$$\mathrm{y} = \begin{pmatrix} 6 \\ 8 \\ 14 \\ 20 \\ 26 \end{pmatrix}; \qquad \mathrm{X} = \begin{pmatrix} 1 & 1 & 1 & 2 \\ 1 & 2 & 2 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 3 & 2 & 1 \\ 1 & 5 & 2 & 2 \end{pmatrix}; \qquad \mathrm{X}' = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 3 & 5 \\ 1 & 2 & 0 & 2 & 2 \\ 2 & 1 & 0 & 1 & 2 \end{pmatrix}.$$

Розраховуємо

$$\beta^* = (\mathrm{X}' \cdot \mathrm{X})^{-1} \cdot \mathrm{X} \cdot \mathrm{y} = \left[\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 3 & 5 \\ 1 & 2 & 0 & 2 & 2 \\ 2 & 1 & 0 & 1 & 2 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 1 & 2 \\ 1 & 2 & 2 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 3 & 2 & 1 \\ 1 & 5 & 2 & 2 \end{pmatrix} \right]^{-1} \cdot \begin{pmatrix} 1 & 1 & 1 & 2 \\ 1 & 2 & 2 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 3 & 2 & 1 \\ 1 & 5 & 2 & 2 \end{pmatrix} \cdot \begin{pmatrix} 6 \\ 8 \\ 14 \\ 20 \\ 26 \end{pmatrix}$$

$$= \begin{pmatrix} 7{,}98 \\ 6{,}34 \\ -3{,}78 \\ -2{,}58 \end{pmatrix}.$$

Отже, маємо:

$$\beta_0^* = 7{,}9; \quad \beta_1^* = 6{,}34; \quad \beta_2^* = -3{,}78; \quad \beta_3^* = -2{,}58.$$

Рівнянням регресії буде

$$y_i = 7{,}98 + 6{,}34 \cdot x_{i1} - 3{,}78 \cdot x_{i2} - 2{,}58 \cdot x_{i3}.$$

2) Розрахуємо коефіцієнт детермінації R^2. Для цього необхідно порахувати теоретичні значення $\hat{y}_i$

i	y_i	x_{i1}	x_{i2}	x_{i3}	$y_i^* = 7{,}98 + 6{,}34 \cdot x_{i1} - 3{,}78 \cdot x_{i2} - 2{,}58 \cdot x_{i3}$	$(\varepsilon_i^*)^2$
1	6	1	1	2	5,38	0,3844
2	8	2	2	1	10,52	6,3504
3	14	1	0	0	14,32	0,1024
4	20	3	2	1	16,86	9,8596
5	26	5	2	2	26,96	0,9216
					$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 =$	17,618

Також знайдемо суму квадратів відхилень y_i від їх середнього:

$$\bar{y} = \frac{6 + 8 + 14 + 20 + 26}{5} = 14{,}8;$$

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = (6 - 14{,}8)^2 + (8 - 14{,}8)^2 + \cdots + (26 - 14{,}8)^2 = 276{,}8.$$

Підставимо отримані суми у формулу коефіцієнта детермінації:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - y_i^*)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{17{,}618}{276{,}8} = 0{,}936.$$

Відповідно коефіцієнт множинної кореляції дорівнює

$$R = \sqrt{R^2} = 0{,}968.$$

Оскільки коефіцієнт детермінації досить високий, то це говорить про адекватність отриманої моделі.

Підрахуємо стандартизовані коефіцієнти регресії:

$$a_j = \beta_j^* \cdot \frac{S_{x_j}}{S_y} \quad (j = \overline{1, m}).$$

Для цього спочатку знайдемо стандартизовані похибки $S_{\beta_j^*}$:

$$S^2 = \frac{\sum_{i=1}^{n} \varepsilon_i^2}{n - m - 1} = \frac{17{,}618}{5 - 3 - 1} = 17{,}618;$$

$$(\mathrm{X}' \cdot \mathrm{X})^{-1} = \frac{1}{178} \cdot \begin{pmatrix} 173 & -14 & -39 & -41 \\ -14 & \mathbf{32} & -38 & -35 \\ -39 & -38 & \mathbf{123} & 0 \\ -41 & -8 & -35 & \mathbf{91} \end{pmatrix};$$

$$S_{x_1} = 17{,}618 \cdot \sqrt{\frac{32}{178}} = 1{,}78;$$

$$S_{x_2} = 17{,}618 \cdot \sqrt{\frac{123}{178}} = 3{,}489;$$

$$S_{x_3} = 17{,}618 \cdot \sqrt{\frac{91}{178}} = 3{,}001;$$

$$S_y = \sqrt{\frac{\vec{y}' \cdot \mathrm{y}}{n} - (\bar{y})^2} = 7{,}44.$$

Тепер можемо визначити нормовані коефіцієнти регресії:

$$a_1 = 6{,}34 \cdot \frac{1{,}78}{7{,}44} = 1{,}52; \quad a_2 = -3{,}78 \cdot \frac{3{,}489}{7{,}44} = -1{,}77; \quad a_2 = -2{,}58 \cdot \frac{3{,}001}{7{,}44} = -1{,}04.$$

Чим більше модуль $|a_j|$, тим сильніший вплив фактора x_j на ознаку Y. У нашому випадку найсильніше впливає змінна x_2, наступна за впливом змінна x_1, найменший вплив здійснює змінна x_3.

6.5. Питання для самоконтролю до Розділу 6

1. Що таке парна лінійна регресія?
2. Виведіть формули для обчислення коефіцієнтів парної лінійної регресії.
3. Як перевірити гіпотезу про значення коефіцієнта α парної лінійної регресії.
4. Як перевірити гіпотезу про значення коефіцієнта β парної лінійної регресії.
5. Дайте означення коефіцієнта парної лінійної кореляції.
6. Дайте означення коефіцієнта детермінації для парної лінійної регресії.
7. Опишіть метод лінеаризація нелінійної моделі за допомогою заміни змінних.
8. Що таке множинна лінійна регресія?
9. Запишіть формулу для обчислення коефіцієнтів множинної лінійної регресії.
10. Запишіть рівняння множинної лінійної регресії у стандартизованому масштабі.
11. Дайте означення коефіцієнтів множинної детермінації та множинної кореляції.
12. Опишіть метод перевірки гіпотези про сумісну значущості її коефіцієнтів множинної лінійної регресії.

РОЗДІЛ 7. НЕПАРАМЕТРИЧНІ КРИТЕРІЇ ОДНОРІДНОСТІ СТАТИСТИЧНИХ ДАНИХ

У цьому розділі ми розглянемо задачі порівняння вибірок у випадку, коли вони є нормально розподіленими або є порядковими. Для обробки даних, які не підпорядковуються нормальному розподілу, використовують непараметричні методи.

Оскільки у таких критеріях обробляється не самі виміряні значення, а їх ранги, ці критерії нечутливі до викидів. Під викидами розуміють результати проведених експериментів чи випробувань, які суттєво відрізняються від спостережуваних середніх значень. У математичній статистиці існує цілий ряд критеріїв для перевірки того, чи є ці відхилення випадковими, чи їх прояв є наслідком прояву систематичних (чи фіксованих) невипадкових процесів.

Таким чином, ми розглянемо непараметричні критерії однорідності статистичних даних, які є аналогами параметричних методів порівняння математичних сподівань (середніх) та дисперсій, які розглядалися у підрозділах 3.3 та 3.4.

Будь який розподіл можна описати ***параметром положення***, який характеризує центр групування випадкових величин, і ***параметром масштабу***, який характеризує ступінь розсіювання випадкових величин відносно центру групування (наприклад, у випадку нормального розподілу ними є відповідно середнє значення $\mathbf{M}X$ і дисперсія $\mathbf{D}X$).

Коли закон розподілу невідомий, гіпотези про параметри положення і параметри масштабу перевіряються за допомогою спеціальних ***критеріїв зсуву*** і ***масштабу***.

Розглянемо дві випадкові величини X і Y з невідомими функціями розподілу $F(x) = \mathbf{P}\{X < x\}$ і $G(x) = \mathbf{P}\{Y < x\},\ x \in \mathbf{R}$.

Якщо $f(x)$ і $g(x)$ їхні функції щільності, то **гіпотеза зсуву** записується як

$$H_0\colon\ f(x) = g(x) \qquad (7.1)$$

проти альтернативної гіпотези

$$H_1\colon\ f(x) = g(x + \Delta).$$

Або $H_0\colon\ \Delta = 0$ проти альтернативної гіпотези $H_1\colon\ \Delta \neq 0$, де Δ - зсув, який визначається різницею параметрів положення розподілів.

Гіпотези про різницю у дисперсіях (при невідомих розподілах) формулюються як **гіпотези про параметри масштабу**. Наприклад, якщо

$$f(x) = \frac{1}{\tau} \cdot f\left(\frac{x - \mathbf{M}X}{\tau}\right) \quad i \quad g(x) = f(x - \mathbf{M}X)$$

то, відповідно, гіпотеза про параметр масштабу записується як

$$H_0\colon\ \tau = 1 \qquad (7.2)$$

проти альтернативної

$$H_1\colon\ \tau \neq 1.$$

Залежність та незалежність вибірок

Нагадаємо, що дві вибірки $x_1, \dots, x_n$ та $y_1, \dots, y_m$ називаються ***незалежними***, якщо вони є реалізаціям набору незалежних в сукупності випадкових величин $X_1, \dots, X_n$, $Y_1, \dots, Y_m$. Натомість ***залежність*** (або ***зв'язність***) вибірок однакової довжини $x_1, \dots, x_n$ та $y_1, \dots, y_n$ означає, що їх розглядають як реалізації наборів випадкових величин $X_1, \dots, X_n$, $Y_1, \dots, Y_n$, випадкові величини X_1 та Y_1, X_2 та Y_2, …, X_n та Y_n, як правило, між собою залежні, але їхні різниці $X_i - Y_i$ мають вигляд $\Delta + \varepsilon_i$, де випадкові величини (ε_i) незалежні в сукупності.

7.1. Критерії зсуву для порівняння двох незалежних вибірок

Критерій Колмогорова-Смирнова

Для перевірки гіпотези (7.1) використовують **критерії Смирнова**. Опишемо його. Нехай $X_1, \dots, X_n$ – незалежні копії випадкової величини X, а $Y_1, \dots, Y_m$ – незалежні копії випадкової величини Y. Пригадаємо, що для довільної випадкової величини Z і $x \in \mathbf{R}$ позначаємо

$$I_{\{Z=x\}}(\omega) = \begin{cases} 1, Z(\omega) \le x \\ 0, Z(\omega) > x \end{cases}, \qquad \omega \in \Omega.$$

Далі, $\forall x \in \mathbf{R}$ позначаємо

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i=x\}}, \qquad \hat{G}_m(x) = \frac{1}{m}\sum_{j=1}^{m} I_{\{Y_j=x\}}.$$

Ці випадкові величини називаються ***емпіричними функціями*** розподілів випадкових величин X та Y, побудованими за наборами $(X_i)_{i=1}^{n}$ та $(Y_j)_{j=1}^{m}$.

Врешті, покладаємо

$$D_{nm} = \sup_{x \in R} \left| \hat{F}_n(x) - \hat{G}_m(x) \right|.$$

Цю випадкову величину називатимемо «відстанню» між випадковими величинами $\hat{F}_n$ та $\hat{G}_m$. Неформально, можемо казати, що коли відстань D_{nm} маленька, то гіпотезу H_0 приймаємо, якщо велика – відхиляємо і приймаємо гіпотезу H_1.

Теоретичною підставою перевірки гіпотези H_0 є наступна теорема.

Теорема Смирнова. Якщо гіпотеза H_0 справедлива, то $\forall t \in \mathbf{R}$

$$\mathbf{P}\left\{ \sqrt{\frac{nm}{n+m}} D_{nm} \le t \right\} \xrightarrow[n,m\to\infty]{} K(t),$$

де $K(t)$ – функція розподілу Колмогорова:

$$K(t) = \begin{cases} 0, & t \le 0 \\ 1 - 2\sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2 t^2}, & t > 0. \end{cases}$$

Для нас важливо, що для цієї функції існують таблиці (Додаток 7).

Практичне застосування теореми Смирнова відбувається так:

1. Проводяться n спостережень випадкової величини X. Дістаємо числові значення

$$\{x_1, \dots, x_n\}. \qquad (*)$$

Проводяться m спостережень випадкової величини Y. Дістаємо числові значення

$$\{y_1, \dots, y_m\}. \qquad (**)$$

2. За отриманими результатами знаходимо числові значення випадкових величин $\hat{F}_n$ та $\hat{G}_m$. А саме:

$$F^*(x) = \frac{n_x}{n}, \quad G^*(x) = \frac{m_x}{n}, \qquad x \in \mathbf{R},$$

де n_x – кількість значень в наборі $(*)$ менших від x, а m_x – кількість значень в наборі $(**)$ менших від x.

3. Обчислюємо «відстань»

$$D^*_{nm} = \sup_{x \in \mathbf{R}} |F^*_n(x) - G^*_m(x)|.$$

Оскільки функції F^*_n та G^*_m набувають не більше ніж n та m значень відповідно, то з обчисленням такого супремуму немає проблем.

4. Задавшись рівнем значущості α, критичну точку z_α знаходимо з рівняння

$$K(z) = \alpha.$$

5. При великих n та m для розрахунків, замість $\sqrt{\frac{nm}{n+m}} D_{nm}$ використовується гранична функція $K(t)$. Отже

6. Якщо $\sqrt{\frac{nm}{n+m}} D_{nm} < z_\alpha$ – гіпотезу приймаємо, в супротивному разі – відхиляємо.

Критерій Манна-Уітні

Нехай $x_1, x_2, \dots\ x_n$ та $y_1, y_2, \dots\ y_m$ – впорядковані за зростанням вибірки. Для перевірки гіпотези зсуву Манн і Уітні запропонували ранговий критерій, заснований на статистиці (***U-статистиці Манна-Уітні***):

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} h_{ij}, \quad \text{де} \quad h_{ij} = \begin{cases} 1, & x_i < y_j \\ 0, & x_i > y_j \end{cases}.$$

Тут U – точна кількість пар значень x_i та y_j, для яких $x_i < y_j$.

Якщо $U_1(\gamma) \le U \le U_2(\gamma)$, то гіпотеза зсуву відхиляється. Тут $U_1(\gamma)$ та $U_2(\gamma)$ – критичні значення, наведені у таблиці (Додаток 8).

Критерій Вілкоксона

З U-статистикою Манна-Уітні пов'язана ***статистика Вілкоксона***, яка визначається сумою рангів елементів однієї вибірки (припустимо, x_i обсягу n) у загальній впорядкованій послідовності елементів поєднаної вибірки обсягу $(m+n)$:

$$R = mn + \frac{n(n+1)}{2} - U.$$

При $n, m > 20$ застосовується апроксимація

$$W = \frac{R - \frac{n(n+m+1)}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}}.$$

Статистика W апроксимується нормальним розподілом, і гіпотеза зсуву відхиляється з достовірністю α, якщо $|W| \le z_{\frac{1+\gamma}{2}}$, де $z_{\frac{1+\gamma}{2}}$ – $\frac{1+\gamma}{2}$-квантиль нормального розподілу.

Якщо у двох порівнюваних вибірках є співпадаючі значення, то їм рекомендується приписувати середні ранги. При цьому у знаменнику статистики слід використовувати величину

$$\left\{\frac{nm(n+m+1)}{12}\left[1 - \frac{\sum_{i=1}^{k} t_i(t_i^2-1)}{(m+n)(m+n-1)(m+n+1)}\right]\right\}^{\frac{1}{2}},$$

де k – загальна кількість груп співпадаючих величин; t_i – число співпадаючих величин у i-тій групі (слід зауважити, що співпадіння враховуються тільки тоді, коли співпадаючі величини належать різним вибіркам, тобто співпадіння, які цілком складаються із елементів однієї і тієї ж вибірки, на величину W не впливають).

7.2. Критерії зсуву для порівняння двох зв'язних вибірок

Парний критерій Вілкоксона

Одним із варіантів застосування розглянутого критерію Вілкоксона є парний критерій Вілкоксона. Його статистика будується наступним чином. Для двох вибірок x та y однакового обсягу n будується ряд різниць $|x_i - y_i|$, який потім ранжується по зростанню.

У впорядкованому ряду значень $|x_i - y_i|$ знаходиться сума рангів T величин $z_i = x_i - y_i > 0$. Гіпотеза зсуву відхиляється, якщо $T_1(\gamma) \le T \le T_2(\gamma)$, де $T_1(\gamma)$ і $T_2(\gamma)$ – критичні значення, наведені у Додатку 9.

При $n \ge 20$ застосовне наближення

$$T^* = \frac{T - \frac{n(n+1)}{4}}{\sqrt{n(n+1)(2n+1)}}.$$

При $|T^*| < z_{\frac{1+\gamma}{2}}$ гіпотеза зсуву відхиляється (тут $z_\gamma - \gamma$-квантиль стандартного нормального розподілу).

Існує більш точне наближення, у відповідності з ним гіпотеза зсуву відхиляється з надійністю γ, якщо

$$|K| < K(\gamma),$$

де

$$K = \frac{T^*}{2}\left\{1 + \sqrt{\frac{n-1}{n-(T^*)^2}}\right\}; \quad K(\alpha) = \frac{1}{2}z_\gamma + \frac{1}{2}t_\gamma(n-1);$$

z_γ – γ-квантиль стандартного нормального розподілу; $t_\gamma(f)$ – γ-квантиль розподілу Сьтюдента з $n-1$ ступенем свободи.

Критерій знаків

Нехай маємо n пар випадкових величин $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$, стосовно яких відомо, що різниці $Z_i = Y_i - X_i$ можна подати у вигляді

$$Z_i = \Delta + \varepsilon_i, \qquad i = 1, \dots, n,$$

де Δ – невідома константа, а випадкові величини ε_i:

- незалежні (самі випадкові величини $(X_i)_{i=1}^n$ та $(Y_i)_{i=1}^n$ можуть бути залежними);

- симетричні (тобто функції розподілу випадкових величин ε_i та $-\varepsilon_i$ рівні);

- функції розподілу випадкових величин ε_i неперервні, тобто $\forall t \in \mathbf{R}\ \ \mathbf{P}\{\varepsilon_i = t\} = 0$.

Щодо невідомого параметра Δ висувається гіпотеза H_0: $\Delta = 0$.

Альтернативна гіпотеза H_1 може бути як однобічною: $\Delta > 0$, або $\Delta < 0$, так і двобічною $\Delta \neq 0$.

Пари $(X_1, Y_1), \dots, (X_n, Y_n)$ можна інтерпретувати як $2n$ спостережень – по два спостереження на кожному з n об'єктів.

Оскільки кожна випадкова величина ε_i симетрична і неперервна, то $\forall i\ \ \mathbf{P}\{\varepsilon_i = 0\} = 0$, а $\mathbf{P}\{\varepsilon_i > 0\} = \mathbf{P}\{\varepsilon_i < 0\}$. Тому $\mathbf{P}\{\varepsilon_i > 0\} + \mathbf{P}\{\varepsilon_i < 0\} = 1$, отже $\forall i\ \ \mathbf{P}\{\varepsilon_i > 0\} = \mathbf{P}\{\varepsilon_i < 0\} = \frac{1}{2}$.

Покладемо

$$\delta_i = \begin{cases} 1, & \text{коли } \varepsilon_i > 0 \text{ (успіх)} \\ -1, & \text{коли } \varepsilon_i < 0 \text{ (невдача)} \end{cases}, \qquad i = 1, \dots, n.$$

Оскільки $(\varepsilon_i)_1^n$ – незалежні випадкові величини, то й $(\delta_i)_1^n$ будуть незалежними. Отже $(\delta_i)_1^n$ – послідовність незалежних випробувань Бернуллі з ймовірністю успіху ½.

Позначимо через B_n випадкову величину, яка дорівнює кількості успіхів в n незалежних випробуваннях Бернуллі, тобто кількості додатних величин серед $(\varepsilon_i)_1^n$. Ця випадкова

величина B_n добре відома з теорії ймовірностей (див. [9 , с. 57]) і її розподіл імовірностей має приблизно такий вигляд (рис. 7.1)

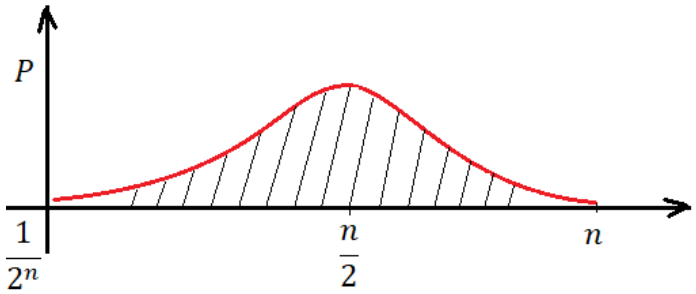


Рис. 7.1

Таким чином (неформально) кількість додатних величин серед $(\varepsilon_i)_1^n$ близька до ½.

Позначимо через D_n кількість додатних різниць серед випадкових величин

$$Z_i = Y_i - X_i = \Delta + \varepsilon_i, \qquad i = 1, \dots, n. \tag{7.3}$$

Якщо гіпотеза H_0 справедлива, тобто $\Delta = 0$, то кількість додатних різниць D_n серед набору (7.3) дорівнює кількості додатних величин серед $(\varepsilon_i)_1^n$, тобто приблизно $\frac{n}{2}$. Якщо ж ця гіпотеза H_0 не справджується, то кількість D_n істотно відрізняється від $\frac{n}{2}$. Тобто вона буде або істотно більшою, або істотно меншою від $\frac{n}{2}$.

Висновок. За статистику для побудови критерію знаків беремо біноміальний розподіл B_n з параметрами $(n, \frac{1}{2})$.

Алгоритм застосування критерію знаків

1. Проводимо випробування випадкових величин $(X_i)_1^n$ та $(Y_i)_1^n$. Дістаємо числові послідовності

$$\{x_1, \dots, x_n\},$$
$$\{y_1, \dots, y_n\}.$$

2. Обчислюємо різниці

$$(y_1 - x_1, y_2 - x_2, \dots, y_n - x_n) \tag{***}$$

3. Знаходимо m_o як кількість додатніх різниць в послідовності $(***)$.

4. Задавшись рівним значущості α з таблиць біноміального розподілу B_n знаходимо критичну точку $m_{\alpha,n}$ як максимальне число, для якого

$$\mathbf{P}\{B_n > m\} \le \alpha$$

(сума ймовірностей менша від α).

2. Якщо $m_o \in [n - m_{\alpha,n}, m_{\alpha,n}]$, то гіпотезу H_0 приймаємо, а якщо ні – відхиляємо. Значення границь $[n - m_{\alpha,n}, m_{\alpha,n}]$ можна знайти у Додатку 12.

Розподіл ймовірностей матиме наступний вигляд (рис.7.2)

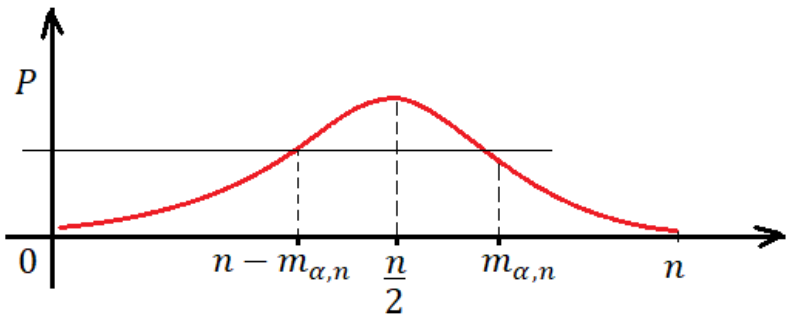


Рис. 7.2

Помилки 1-го та 2-го роду. Коли гіпотеза H_0 справджується, то випадкова величина

$$D_n = \text{card}\{i: Y_i - X_i > 0\}$$

має біноміальний розподіл з параметрами $\left(n, \frac{1}{2}\right)$. Тому D_n майже завжди мало відхиляється від $\frac{n}{2}$. Але D_n може (хоча й зрідка) набувати значень, які істотно відрізняються від $\frac{n}{2}$. При цьому, гіпотезу H_0 відхиляємо і тим самим припускаємося помилки 1-го роду (її ймовірність не перевищує вибраного рівня значущості α).

Гіпотеза H_0 нам особливо дорога. Наприклад, якщо перевіряється новий медичний препарат, то його впровадження вимагає значних коштів. Тому, ми не хотіли б припуститися помилки 1-го роду: відхилити гіпотезу H_0 про те, що новий препарат ніяк не кращий, коли ця гіпотеза правильна.

Перейдемо тепер до помилки 2-го роду. Коли гіпотеза H_0 не справджується (наприклад $\theta > 0$), то, незважаючи на те, що випадкова величина D_n (кількість додатних різниць) майже завжди набуває значень істотно більших ніж $\frac{n}{2}$, вона може, хоча й зрідка, набувати значень поблизу $\frac{n}{2}$. При цьому гіпотезу H_0 приймаємо, отже припускаємося помилки 2-го роду.

Якщо зняти вимогу неперервності розподілів випадкових величин X_i та Y_i, то різниці $Y_i - X_i$ можуть набувати нульових значень з ненульовими ймовірностями. У цьому випадку говорять, що є зв'язки між X_i та Y_i. В цьому випадку теж можна користуватися критерієм знаків для перевірки гіпотези H_0: $\Delta = 0$. Але потрібно відкинути рівні нулеві різниці і застосувати критерій до тих, що лишилися.

Опишемо відповідний алгоритм. Проводимо випробування. Дістаємо числові послідовності

$$\{x_1, \dots, x_n\}, \qquad \{y_1, \dots, y_n\}.$$

обчислюємо різниці $y_i - x_i$ і відкидаємо усі ті, що отримали ненульові різниці

$$(y_1 - x_1, \dots, y_s - x_s), \quad s \le n.$$

А далі проводимо точно таку ж процедуру, як і попереднього разу, тільки скрізь замінюючи n на s.

7.3. Критерії зсуву для порівняння декількох незалежних вибірок

Ці критерії застосовуються для перевірки гіпотези відсутності зсуву для кількох випадкових величин $X_1, \dots, X_k,\ k > 2$.

Критерій Краскела-Уолліса

Цей критерій є узагальненням критерію Манна-Вітні, який ми розглядали раніше, на випадок більший ніж 2 випадкові величини.

Отже, нехай для випадкових величин $X_1, \dots, X_k$ були отримані наступні результати експерименту:

$$X_1: x_1^1, x_1^2, \dots, x_1^{n_1}$$
$$X_2: x_2^1, x_2^2, \dots, x_2^{n_2}$$
$$\dots\dots\dots\dots\dots$$
$$X_k: x_k^1, x_k^2, \dots, x_k^{n_k}$$

Всього $N = \sum_{i=1}^{k} n_i$ значень.

Впорядкуємо усі їх за зростанням і позначимо через R_i^j ранг, тобто номер j-го елемента i-ї вибірки в спільному впорядкованому ряді.

Додаткові припущення:

1. Випадкові величини $X_1, \dots, X_k$ незалежні.

2. Їхні (невідомі) функції розподілу $F_1, \dots, F_k$ неперервні.

Перевіряється нульова гіпотеза

$$H_0: F_1(x) = F_2(x) = \dots = F_k(x)$$

при альтернативі

$$H_1: F_1(x - \Delta_1) = F_2(x - \Delta_2) = \dots = F_k(x - \Delta_k).$$

Для означення статистики критерію Краскела-Уолліса введемо позначення: $\forall i = 1, \dots, k$

$$R_i = \sum_{j=1}^{n_i} R_i^j.$$

Тоді статистика має вигляд

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1).$$

Гіпотеза H_0 відхиляється, якщо $H > H_\alpha$, де H_α – критичне значення, яке відповідає рівневі значущості α і приймається, якщо $H < H_\alpha$.

При малих значеннях: $k \le 5$ та $n_i \le 8$ значення H_α знаходять з таблиць (Додаток 10). При великих значеннях k та n_i застосовують різні апроксимації. Вкажемо одну із них.

Апроксимація Краскела-Уолліса. Введемо наступні позначення:

$$M = \frac{N^3 - \sum_{i=1}^{k} n_i^3}{N(N+1)};$$

$$V = 2(k-1) - \frac{2\{3k^2 - 6k + N(2k^2 - 6k + 1)\}}{5N(N+1)} - \frac{6}{5}\sum_{i=1}^{k}\frac{1}{n_i};$$

$$\nu_1 = (k-1)\frac{(k-1)(M-k+1) - V}{\frac{1}{2}MV};\ \ \nu_2 = \frac{M-k+1}{k-1}\nu_1.$$

Теорема. За відсутності зсуву, тобто при справедливості гіпотези H_0 статистика

$$F = \frac{H(M-k+1)}{(k-1)(M-H)}$$

має F – розподіл Фішера з ν_1 та ν_2 ступенями свободи.

Таким чином, H_0 відхиляється з рівнем значущості α, якщо експериментальне значення $F > F_\alpha(\nu_1, \nu_2)$ і приймаємо супротивному разі. Значення $F_\alpha(\nu_1, \nu_2)$ знаходимо за Додатком 5.

7.4. Критерії зсуву для порівняння декількох зв'язних вибірок

Критерій Фрідмана

Цей критерій являє собою розширення тесту Вілкоксона для наявності більш ніж двох залежних вибірок. Він ґрунтується на рангових послідовностях, які будуються для значень усіх змінних, що беруть участь у критерії.

Отже, нехай маємо k випадкових величин $X_1, \dots, X_k$. І для них були отримані наступні результати експерименту:

$$X_1: x_1^1, x_1^2, \dots, x_1^n$$
$$X_2: x_2^1, x_2^2, \dots, x_2^n$$
$$\dots\dots\dots\dots\dots$$
$$X_k: x_k^1, x_k^2, \dots, x_k^n$$

Виконаємо ранжування, від найбільшого до найменшого, спостережень в середині кожного ряду. Нехай R_i^j – ранг елемента x_i^j у спільному ранжуванні $x_1^j, x_2^j, \dots, x_k^j$. Позначимо

$$R_i = \sum_{j=1}^{n} R_i^j;\ \bar{R}_i = \frac{R_i}{n};\ \bar{\bar{R}} = \frac{k+1}{2}.$$

Для перевірки гіпотези зсуву між параметрами положеннями вибірок рівного об'єму в використовують критерій, який побудовано на статистиці

$$S = \frac{12n}{k(k+1)}\sum_{i=1}^{k}\left(\bar{R}_i - \bar{\bar{R}}\right) = \frac{12}{nk(k+1)}\sum_{i=1}^{k} R_i^2 - 3n(k+1).$$

Гіпотеза зсуву відхиляється, якщо $S < S_{1-\alpha}(n,k)$, де $S_{1-\alpha}(n,k)$ – критичне значення наведене у Додатку 11.

У зв'язку з тим, що таблиці розподілу критерію, що розглядається, побудовані для невеликого діапазону значень, широко застосовуються різноманітні апроксимації. Наведемо приклад найбільш поширених.

При $n \geq 13$ та $k \geq 20$ застосовується апроксимація

$$S_{1-\alpha}(n,k) = \chi_\alpha^2(k-1).$$

Значення $\chi_\alpha^2(k-1)$ знаходимо за Додатком 4.

Для інших значень n та k використовується перетворення

$$F = \frac{(n-1)S}{n(k-1) - S}.$$

Гіпотеза зсуву відхиляється, якщо $F < F_\alpha(f_1, f_2)$, де $F_\alpha(f_1, f_2)$ – α-квантиль розподілу Фішера при f_1 та f_2 ступенях свободи. Значення $F_\alpha(f_1, f_2)$ знаходимо за Додатком 5.

При $n \geq 13$ та $7 \leq k \leq 19$ приймаємо $f_1 = k-1$ та $f_2 = (k-1)(n-1)$.

При $k \geq 8$ та $7 \leq n \leq 12$ приймаємо $f_1 = k-1$ та

$$f_2 = \frac{L^2}{(n-1)\sum_{1\leq i\leq k} V_i^2} - (n-1),$$

де

$$V_i = \frac{1}{n-1}\sum_{j=1}^{n}\left(R_i^j - \bar{R}_i\right);\ \ L = (n-1)\sum_{i=1}^{k} V_i;\ \ i = 1, \ldots, k.$$

Якщо f_2 стає дробовим, то при використанні таблиць необхідно застосувати інтерполяцію.

Критерій Кендала

Для розв'язання цієї ж задачі, а саме дослідження зсуву для декількох зв'язних вибірок, можна використовувати критерій Кендала. Детальний опис критерію наведений у підрозділі 5.4.

7.5. Критерії масштабу

На початку цього розділу було сформульовано гіпотезу щодо рівності параметрів масштабу (див. формулу (7.2)). Згідно з нею, для випадкових величин, які мають нульове математичне сподівання і цільності розподілів $f(x)$ та $g(x)$, гіпотеза H_0 відсутності масштабу має вигляд

$$g(x) = f(x),$$

а гіпотеза H_1 наявності масштабу –

$$g(x) = \frac{1}{\tau} f\left(\frac{x}{\tau}\right), \qquad \tau > 0, \ \ \tau \neq 1\,.$$

Зробимо спочатку два зауваження щодо формули представлення в альтернативній гіпотезі.

1. Нехай, наприклад, $\tau = 2$. Намалюємо графіки функцій $f(x)$ та $f\left(\frac{x}{2}\right)$ (рис.7.3).

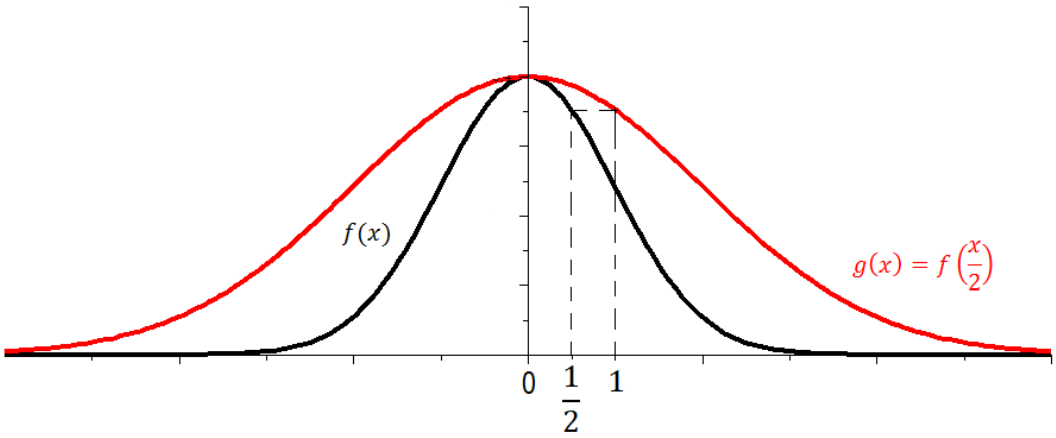


Рис. 7.3

При діленні аргументу x на 2, графік розтягується у 2 рази (змінюється в масштабі у 2 рази). Але, щоб відобразити зміну масштабу для щільності розподілу, функцію $f\left(\frac{x}{\tau}\right)$ брати не можна, бо при $\tau \neq 1$

$$\int_{-\infty}^{\infty} f\left(\frac{x}{\tau}\right) dx = \left[\frac{x}{\tau} = y\right] = \int_{-\infty}^{\infty} f(y)\tau dy = \tau \neq 1.$$

Тому, щоб отримана функція була щільністю розподілу, потрібно взяти $\frac{1}{\tau} f\left(\frac{x}{\tau}\right)$. Тоді інтеграл від неї дорівнюватиме 1.

2. Для випадкових величин X і Y, про які точно відомо, що вони мають щільність і ці щільності можуть відрізнятися лише масштабом і що вони мають однакові математичні сподівання (нехай, для простоти подальших обчислень, нульові), перевірка гіпотези H_0 відсутності масштабу еквівалентна перевірці гіпотези про рівність дисперсій. Знову ж таки в припущенні, що ці дисперсії існують.

Справді, якщо $g(x) = f(x)$, то $\mathbf{D}X = \mathbf{D}Y$. Навпаки, нехай $\mathbf{D}X = \mathbf{D}Y$. Припустимо, що $g(x) = \frac{1}{\tau} f\left(\frac{x}{\tau}\right)$, $\tau \neq 1$. Тоді

$$\mathbf{D}Y = \int_{-\infty}^{\infty} y^2 g(y)dy = \int_{-\infty}^{\infty} \frac{1}{\tau} y^2 f\left(\frac{x}{\tau}\right) d\tau = \begin{bmatrix} \frac{y}{\tau} = x \\ dy = \tau dx \end{bmatrix} = \int_{-\infty}^{\infty} \tau^2 x^2 f(x)dx =$$

$$\tau^2 \int_{-\infty}^{\infty} x^2 f(x)dx = \tau^2 \mathbf{D}X \neq \mathbf{D}X.$$

Протиріччя показує, що $\mathbf{D}X = \mathbf{D}Y \Longrightarrow g(x) = f(x)$. □

Зміни масштабу немає.

Зауваження 7.1. Гіпотезу масштабу можна перевіряти й тоді, коли випадкові величини дисперсій не мають.

Її можна перевіряти і тоді, коли випадкові величини порядкові, тобто коли мова йде не про їхні числові значення, а лише про порядок: яке значення якому передує.

Критерій Ансарі-Бредлі для двох вибірок

Цей критерій є масштабним аналогом критерію Вілкоксона перевірки гіпотези однорідності, який ми розглядали в підрозділі 7.2. Опишемо його. Нехай проведено m випробувань

$$x_1, \dots, x_m$$

випадкової величини X і n випробувань випадкової величини Y:

$$y_1, \dots, y_n.$$

Розмістимо їх в один ряд, впорядковуючи за зростанням

$$z_1 \leq z_2 \leq \dots \leq z_{m+n}. \qquad (****)$$

Для довільного $1 \leq i \leq m$ нехай R_i – ранг елемента x_i у послідовності $(****)$, тобто його номер у цій послідовності.

Статистикою критерію Ансарі-Бредлі є випадкова величина

$$S = \sum_{i=1}^{m} \left\{ \frac{m+n+1}{2} - \left| R_i - \frac{m+n+1}{2} \right| \right\}.$$

Гіпотеза H_0 відсутності масштабу: $\tau = 1$ перевіряється так:

1. Встановлюється рівень значущості γ.
2. З Додатку 13 знаходяться критичні значення
$$S_1(\gamma) = S_1(\gamma, m, n) \text{ та } S_2(\gamma) = S_2(\gamma, m, n).$$
3. За формулою (2) обчислюється спостережуване значення S.
4. Якщо $S_1 < S < S_2$, то гіпотеза H_0 про відсутність масштабу на рівні значущості α приймається. У супротивному випадку, вона відхиляється і приймається альтернативна гіпотеза $H_1: \tau \neq 1$ (зміна масштабу є).

Зауваження. Таблиця в Додатку 13 містить лише значення $S_1(\gamma, m, n)$ та $S_2(\gamma, m, n)$ для обсягів вибірок $m, n \leq 10$.

При $m, n > 10$ можна використовувати асимптотичну нормальність розподілу випадкової величини

$$S^* = \frac{S - \mathbf{M}(S)}{\sqrt{\mathbf{D}(S)}}.$$

У цьому випадку, вибіркове математичне сподівання обчислюється за формулою

$$\mathbf{M}(S) = \begin{cases} \dfrac{m(m+n+2)}{4}, & \text{при } m+n \text{ парному;} \\ \dfrac{m(m+n+1)^2}{4(m+n)}, & \text{при } m+n \text{ непарному.} \end{cases}$$

Натомість, вибіркова дисперсія $D(S)$ – за формулою

$$\mathbf{D}(S) = \begin{cases} \dfrac{mn(m+n-2)(m+n+2)}{48(m+n-1)}, & \text{при } m+n \text{ парному;} \\ \dfrac{mn(m+n+1)[(m+n)^2+3]}{48(m+n)^2}, & \text{при } m+n \text{ непарному.} \end{cases}$$

Тоді гіпотеза H_0 рівності параметрів масштабу для двох випадкових величин приймається з рівнем значущості α, якщо

$$|S^*| < z_{\text{кр}}$$

і відхиляється у супротивному випадку. Число $z_{\text{кр}}$ знаходимо, як і раніше, з рівняння

$$\Phi\left(z_{\text{кр}}\right) = \frac{1+\gamma}{2},$$

де $\Phi(z)$ – функція розподілу стандартної нормальної випадкової величини.

Критерій Бхапкара-Дешпанде для декількох вибірок

Нехай тепер ми маємо $k > 2$ випадкових величин $X_1, X_2, \dots, X_k$ і в результаті експерименту отримано їхні значення:

$$X_1: x_1^1, x_1^2, \dots, x_1^{n_1}$$
$$X_2: x_2^1, x_2^2, \dots, x_2^{n_2}$$
$$\dots\ \dots\ \dots\ \dots\ \dots\ \dots\ \dots\ \dots$$
$$X_k: x_k^1, x_k^2, \dots, x_k^{n_k}$$

Якщо ми будемо всіма можливими способами з кожної вибірки (рядка) брати одне спостереження, то дістанемо $N = \prod_{i=1}^{k} n_i$ підвибірок. Покладемо $\forall i\ \ \bar{n}_i = {}^{n_i}/_{N}$.

Позначимо через v_i^j – кількість таких підвибірок, в яких спостереження з i-ї вибірки (рядка) більше, ніж $j-1$ спостережень. Позначимо

$$u_i^j = \frac{\left(v_i^j\right)^2}{N - \frac{1}{k}}.$$

Теорема (Бхапкара). Якщо $f_1 = f_2 = \cdots = f_k$, то $u_i^j = 0$.

Критерій Бхапкара ґрунтується на статистиці

$$V = N(2k-1)\left[\sum_{i=1}^{k} \bar{n}_i u_i^1 - \left\{\sum_{i=1}^{k} \bar{n}_i u_i^1\right\}^2\right].$$

Для представлення ефективнішої версії цього критерію позначимо $\forall i = 1, \dots, k \;\; d_i = u_i^1 + u_i^k$. Далі, позначимо

$$D = \frac{N(k-1)^2(2k-1)C_{2(k-1)}^{k-1}}{2\left[k^2 + (k^2+4k+2)C_{2(k-1)}\right]}\left[\sum_{i=1}^{k} \bar{n}_i d_i^j - \left\{\sum_{i=1}^{k} \bar{n}_i d_i\right\}^2\right].$$

І, врешті, позначивши $l_i = -u_i^1 + u_i^k \;\; (i = 1, \dots, k)$, розглянемо статистику

$$L = \frac{N(2k-1)(k-1)^2 C_{2(k-1)}^{k-1}}{2k^2\left(C_{2(k-1)}^{k-1} - 1\right)}\left[\sum_{i=1}^{k} \bar{n}_i (l_i)^2 - \left\{\sum_{i=1}^{k} \bar{n}_i l_i\right\}^2\right].$$

Гіпотеза H_0 про відсутність зміни масштабу приймається, якщо відповідно, $V, L, D < \chi_\alpha^2(k-1)$, де $\chi_\alpha^2(k-1)$ – α – квантиль розподілу χi-квадрат з $k-1$ степенем свободи.

Для критеріїв L та D при $k = 3$ критичні значення (L_γ і D_γ) наведені в таблиці (Додаток 14).

7.6. Приклади до Розділу 7

Приклад 1. Критерій Колмогорова-Смирнова

Потрібно визначити, чи однакові функції розподілу листків по розмірам двох ярусів (частот по вибіркам X та Y) крони акації. Дані експерименту наведені в наступній таблиці

Діапазон розмірів		Експеримент	
		$n(X)$	$m(Y)$
8,2	9,8	1	0
9,8	11,4	5	7
11,4	13,0	12	10
13,0	14,6	1	10
14,6	16,2	1	4
16,2	17,8	0	7

Розв'язання.

Знайдемо відносні частоти та значення емпіричних функцій розподілів.

n_i	m_i	$\frac{n_i}{n}$	$\frac{m_i}{m}$	$F_n^*(x)=\frac{n_x}{n}$	$G_m^*(y)=\frac{m_x}{n}$	$\lvert F_n^*(x)-G_m^*(y)\rvert$
1	0	0,050	0,000	0,050	0,000	0,050
5	7	0,250	0,184	0,300	0,184	0,116
12	10	0,600	0,263	0,900	0,447	0,453
1	10	0,050	0,263	0,950	0,711	0,239
1	4	0,050	0,105	1,000	0,816	0,184
0	7	0,000	0,184	1,000	1,000	0,000
$\boldsymbol{n=20}$	$\boldsymbol{m=38}$					$\boldsymbol{\max=0,453}$

$$D_{nm}^* = \sup_{x\in\mathbf{R}}|F_n^*(x)-G_m^*(x)| = \max|F_n^*(x)-G_m^*(x)| = 0{,}453.$$

Задавшись рівнем значущості $\alpha = 0{,}05$, критичну точку $k_{\text{кр}}$ знаходимо з рівняння (Додаток 7)

$$K(k_{\text{кр}}) = \alpha \;\Rightarrow\; k_{\text{кр}} = 1{,}36.$$

$$\sqrt{\frac{nm}{n+m}}D_{nm} = \sqrt{\frac{20\cdot 38}{20+38}}\cdot 0{,}453 = 1{,}64.$$

Маємо, що $\sqrt{\frac{nm}{n+m}}D_{nm} = 1{,}64 > 1{,}36 = k_{\text{кр}}$ – гіпотезу відхиляємо, тобто функції розподілу листків по розмірам у двох ярусів різні.

Приклад 2. Критерій Манна-Уітні та Вілкоксона

Нехай є дві вибірки випадкових величин

$$(n = 8):\ 1{,}2;\ \ 2{,}1;\ \ 3{,}8;\ \ 6{,}4;\ \ 7{,}2;\ \ 9;\ \ 11;\ \ 12{,}4;$$

$$(m = 10):\ 2{,}1;\ \ 2{,}1;\ \ 6{,}1;\ \ 6{,}3;\ \ 9;\ \ 9;\ \ 11{,}2;\ \ 12{,}4;\ \ 13{,}2;\ 13{,}6.$$

Необхідно перевірити гіпотезу зсуву критерієм Манна-Уітні та Вілкоксона на рівні значущості $\alpha = 0{,}05$.

Розв'язання.

Розрахуємо кількість пар, для яких $x_i < y_j$ при всіх $i = 1, \dots, n$ та $j = 1, \dots, m$. Наприклад, для $i = 1$ та різних $j = 1, \dots, 10$ маємо число таких пар, рівне $\sum_{j=1}^{10} h_{1j} = 10$. Далі за аналогією отримаємо:

i	1	2	3	4	5	6	7	8
$\sum h_{ij}$	10	8	8	6	6	4	4	2

Отримаємо:

$$U = \sum_{i=1}^{8}\sum_{j=1}^{10} h_{ij} = 10 + 8 + 8 + 6 + 6 + 4 + 4 + 2 = 48.$$

Для $\gamma = 1 - \alpha = 0{,}05$ із таблиці (Додаток 8) знаходимо

$$U_1(0{,}95) = 17 \text{ та } U_2(0{,}95) = 63.$$

Так як

$$U_1(0{,}95) = 17 \ \le U = 48 \le U_2(0{,}95) = 63,$$

то гіпотеза зсуву відхиляється.

Для статистики Вілкоксона маємо:

$$R = 10 \cdot 8 + \frac{8 \cdot (8+1)}{2} - 48 = 68.$$

При розрахунку статистики W слід мати на увазі, що у нашому прикладі є три групи співпадаючих спостережень: (2,1;2,1;2,1), (9;9;9), (12,4;12,4), тобто $k = 3$, $t_1 = 3$, $t_2 = 3$, $t_3 = 2$.

Розраховуємо

$$\left\{\frac{10 \cdot 8 \cdot (8+10+1)}{12}\left[1 - \frac{3 \cdot (3^2-1) + 3 \cdot (3^2-1) + 2 \cdot (2^2-1)}{(8+10)(8+10-1)(8+10+1)}\right]\right\}^{\frac{1}{2}} = 11{,}202;$$

$$W = \frac{68 - \frac{8 \cdot (8+10+1)}{2}}{\sqrt{\frac{10 \cdot 8 \cdot (8+10+1)}{12}}} = -0{,}714.$$

Квантиль нормального розподілу $z_{\frac{1+0,95}{2}} = 1,96$ (Додаток 2).

Так як $|W| = 0,714 < 1,96 = z_{\frac{1+\gamma}{2}}$, то гіпотеза зсуву відхиляється .

Приклад 3. Парний критерій Вілкоксона

Учням 5 класу вирішили провести тренінги по емоційній розрядці з метою підвищення середнього балу успішності. За допомогою парного критерію Вілкоксона з'ясувати, чи тренінг підвищив середній бал успішності учнів 5-го класу? Результати успішності наведені у наступній таблиці:

№	До	Після
1	3,8	3,9
2	4,2	4,8
3	4,0	4,0
4	3,3	3,4
5	3,9	3,8
6	4,5	4,3
7	4,9	4,9
8	3,8	4,5
9	3,8	4,0
10	4,1	4,2

Розв'язання.

Необхідні розрахунки проведемо у наступній таблиці:

№	До	Після	$x_i - y_i$	$\|x_i - y_i\|$	R_i для $\|x_i - y_i\|$	R_i для $x_i - y_i > 0$
1	3,8	3,9	-0,1	0,1	4	
2	4,2	4,8	-0,6	0,6	9	
3	4,0	4,0	0	0	1,5	
4	3,3	3,4	-0,1	0,1	4	
5	3,9	3,8	0,1	0,1	4	4
6	4,5	4,3	0,2	0,2	7,5	7,5
7	4,9	4,9	0	0	1,5	
8	3,8	4,5	-0,7	0,7	10	
9	3,8	4,0	-0,2	0,2	7,5	
10	4,1	4,2	-0,1	0,1	6	

Тепер можемо знайти значення статистики T як суми рангів величин $z_i = x_i - y_i > 0$:

$$T = 4 + 7,5 = 11,5.$$

За Додатком 9 для $n = 10$ знаходимо $T_1(0,95) = 9$ і $T_2(0,95) = 46$.

Так як $T_1(0{,}95) = 9 \le T = 11{,}5 \le T_2(0{,}95) = 46$, то гіпотеза зсуву відхиляється.

Отже, інтенсивність зсувів у бік збільшення середнього балу успішності у учнів 5 класу не перевищує інтенсивність зсувів у бік її зменшення, то можемо зробити висновок, що тренінг не допоміг у збільшенні середнього балу успішності учнів 5 класу.

Приклад 4. Критерій знаків

Розглянемо експеримент штучного стимулювання дощу. Хмару обприскують певним препаратом. І ми хочемо перевірити ефективність препарату. Для перевірки випадково вибрали 16 пар днів. У кожній парі в один день застосовували препарат, а в інший – ні. У ці дні вимірювався рівень опадів (наприклад, мг/м^2). Отримали наступну таблицю опадів:

Номер пари	Рівень опадів	
	З препаратом, Y_i	Без препаратом, X_i
1	0	1,32
2	2,09	0
3	0,07	0
4	0,30	0,1
5	0	0,44
6	2,55	0
7	1,62	1,05
8	0	0,54
9	0	0
10	1,87	0,62
11	250	0
12	3,15	5,54
13	0,15	0,01
14	2,56	0
15	0	0
16	0	0,75

Питання: Чи свідчать наведені дані про ефективність препарату?

Розв'язання.

Сформулюємо цю задачу в термінах перевірки статистичних гіпотез. Маємо 16 пар спостережень $(X_1, Y_1), (X_2, Y_2), \dots, (X_{16}, Y_{16})$. Тут Y_i рівень опадів при застосуванні препарату X_i – без застосування препарату.

Природно припустити, що випадкові величини X_i та Y_i зв'язані співвідношенням

$$Y_i - X_i = \Theta + \varepsilon_i, \;\; i = 1, \dots, 16,$$

де випадкові величини ε_i незалежні і симетричні. Щодо невідомого параметра Θ висуваємо гіпотезу $H_0: \Theta = 0$. Фізично вона означає неефективність препарату. За альтернативну гіпотезу природно прийняти $H_1: \Theta > 0$ – ефект є.

Скористаємося критерієм знаків. Для цього знайдемо знаки різниць $Y_i - X_i$ у наступній таблиці:

Номер пари	Рівень опадів		$Y_i - X_i$	Знаки $Y_i - X_i$
	З препаратом, Y_i	Без препаратом, X_i		
1	0	1,32	-1,32	–
2	2,09	0	2,09	+
3	0,07	0	0,07	+
4	0,3	0,1	0,2	+
5	0	0,44	-0,44	–
6	2,55	0	2,55	+
7	1,62	1,05	0,57	+
8	0	0,54	-0,54	–
9	0	0	0	
10	1,87	0,62	1,25	+
11	250	0	250	+
12	3,15	5,54	-2,39	–
13	0,15	0,01	0,14	+
14	2,56	0	2,56	+
15	0	0	0	
16	0	0,75	-0,75	–

За критерієм знаків оцінюється наявність зв'язків, тому, нулі відкидаємо. Лишиться 14 знаків. $s = 14$. З них $m_{\text{сп}} = 9$ позитивних. Візьмемо рівень значущості $\alpha = 0{,}025$. З таблиць Додатку 12 знаходимо $m_{\alpha,s} = m_{0{,}025;14} = 11$.

Звичайно,

$$m_{\text{сп}} = 9 < 11 = m_{\alpha,s}.$$

Тому гіпотезу H_0 приймаємо і робимо висновок, що на рівні значущості $\alpha = 0{,}025$ препарат не ефективний. Ми його застосовувати не будемо.

Приклад 5. Критерій Краскела-Уолліса

У результаті спостережень отримано 5 вибірок випадкових величин X_1, X_2, X_3, X_4, X_5. Потрібно перевірити гіпотезу H_0 про відсутність зсуву на рівні значущість $\alpha = 0{,}05$.

X_1: 1; 2; 3; 4; 5; 6;

X_2: 3; 4; 5; 6; 7;

X_3: 7; 8; 9;

X_4: 1; 5; 7; 8; 10; 12;

X_5: 10; 11; 13; 14; 16; 18; 20.

Розв'язання.

Ранжуємо сумісно всі $N = \sum_{i=1}^{5} n_i = 6 + 5 + 5 + 6 + 7 = 27$ вибіркових значень x_i^j. Нагадаємо, що i – це номер вибірки, а R_i^j – ранг j-го спостереження в i-й вибірці.

Розмістимо x_i^j за зростанням і порахуємо ранги у наступній таблиці:

x_i^j	i	R_i^j
1	1	1,5
1	4	1,5
2	1	3
3	1	4,5
3	2	4,5
4	1	6,5
4	2	6,5
5	1	9
5	2	9
5	4	9
6	1	11,5
6	2	11,5
7	2	14
7	3	14
7	4	14
8	3	16,5
8	4	16,5
9	3	18
10	4	19,5
10	5	19,5
11	5	21
12	4	22
13	5	23
14	5	24
16	5	25
18	5	26
20	5	27

Далі, рахуємо суми рангів для кожної вибірки окремо:

$$R_1 = \sum_j R_1^j = 1{,}5 + 3 + 4{,}5 + 6{,}5 + 9 + 11{,}5 = 36; \quad \bar{R}_1 = \frac{36}{6} = 6;$$

$$R_2 = \sum_j R_2^j = 4{,}5 + 16{,}5 + 9 + 11{,}5 + 14 = 45{,}5; \quad \bar{R}_2 = \frac{45{,}5}{5} = 9{,}1;$$

$$R_3 = \sum_j R_3^j = 14 + 16{,}5 + 18 = 48{,}5; \quad \bar{R}_3 = \frac{48{,}5}{3} = 16{,}2;$$

$$R_4 = \sum_j R_4^j = 1{,}5 + 9 + 14 + 16{,}5 + 19{,}5 + 22 = 82{,}5; \quad \bar{R}_4 = \frac{82{,}5}{6} = 13{,}8;$$

$$R_5 = \sum_j R_5^j = 19{,}5 + 21 + 23 + 24 + 25 + 26 + 27 = 165{,}5; \quad \bar{R}_5 = \frac{165{,}5}{7} = 23{,}6.$$

Розрахуємо статистику Краскела-Уолліса:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

$$= \frac{12}{27 \cdot 28} \cdot \left(\frac{36^2}{6} + \frac{45{,}5^2}{5} + \frac{48{,}5^2}{3} + \frac{82{,}5^2}{6} + \frac{165{,}5^2}{7} \right) - 3 \cdot 28 = 18{,}56.$$

Далі використаємо апроксимацію Краскела-Уолліса. Для цього обчислимо

$$M = \frac{N^3 - \sum_{i=1}^{k} n_i^3}{N(N+1)} = \frac{27^3 - (6^3 + 5^3 + 3^3 + 6^3 + 7^3)}{27(28+1)} = 24{,}81;$$

$$V = 2 \cdot 4 - \frac{2 \cdot \{3 \cdot 5^2 - 6 \cdot 5 + 27(2 \cdot 25 - 6 \cdot 5 + 1)\}}{5 \cdot 27 \cdot 28} - \frac{6}{5} \sum_{i=1}^{5} \frac{1}{n_i} = 6{,}46;$$

$$\nu_1 = (5-1) \frac{(5-1)(24{,}81 - 5 + 1) - 6{,}46}{\frac{1}{2} \cdot 24{,}81 \cdot 6{,}46} = 3{,}83;$$

$$\nu_2 = \frac{24{,}81 - 5 + 1}{5 - 1} \cdot 3{,}83 = 19{,}91.$$

Далі знаходимо статистику F:

$$F = \frac{H(M - k + 1)}{(k-1)(M - H)} = \frac{18{,}56 \cdot (24{,}81 - 5 + 1)}{(5-1)(24{,}81 - 18{,}56)} = 15{,}73.$$

Вона має F – розподіл Фішера з ν_1 та ν_2 ступенями свободи.

З таблиць розподілу Фішера (Додаток 5) знаходимо для рівня значущості $\alpha = 0{,}05$ знаходимо критичне значення:

$$F_{kp}(\nu_1 = 3{,}83, \nu_2 = 19{,}91) = 2{,}9.$$

Оскільки

$$F = 15{,}73 > F_{kp} = 2{,}9,$$

то гіпотеза H_0 відхиляється, тобто приймається гіпотеза H_1 про наявність зсуву.

Приклад 6. Критерій Фрідмена

Для трьох вибірок випадкових величин, значення яких наведено у таблиці, перевірити гіпотезу зсуву за критерієм Фрідмена для рівня значущості $\alpha = 0{,}05$.

Номер елемента у вибірці	**Номер вибірки**			$\overline{x_j}$
	1	**2**	**3**	
1	2,1	3,2	4,3	3,20
2	1,8	4,1	2,3	2,73
3	1,7	2,3	3,4	2,47
4	1,8	2,4	3,5	2,57
5	1,9	2,5	3,6	2,67
6	2,4	1,2	3,7	2,43
7	1,7	1,9	3,2	2,27
8	1,6	2,3	2,8	2,23
9	1,5	2,4	2,9	2,26
10	1,7	2,9	3,7	2,83

Розв'язання.

Побудуємо таблицю рангів R_i^j по рядкам

j	***i***		
	1	**2**	**3**
1	1	2	3
2	1	3	2
3	1	2	3
4	1	2	3
5	1	2	3
6	2	1	3
7	1	2	3
8	1	2	3
9	1	2	3
10	1	2	3

Отримаємо

$$R_1 = 11, \qquad R_2 = 20, \qquad R_3 = 29.$$

Порахуємо статистику критерію

$$S = \frac{12}{10 \cdot 3 \cdot 4} \sum_{i=1}^{3} R_i^2 - 3 \cdot 10 \cdot 4 = 0{,}1 \cdot (11^2 + 20^2 + 29^2) - 120 = 16{,}2.$$

Із Додатку 11 знаходимо критичне значення $S_{0,95}(10; 3) = 6{,}1$.

Оскільки $S = 16{,}2 > S_{0,95}(10; 3) = 6{,}1$, то гіпотеза зсуву приймається.

Приклад 7. Критерій Ансарі-Бредлі

Нехай випадкові величини X та Y набувають значень

X: 1,2; 3,4; 6,2; 8,1; 10,2; 11,3; 13,0; 15,9; $(m = 8)$

Y: 0,8; 2,4; 4,2; 5,1; 6,8; 11,4; 13,8; 20,1; 24,2; 26,7. $(n = 10)$

Перевірити гіпотезу H_0 рівності параметрів масштабу за допомогою критерію Ансарі-Бредлі при рівні значущості $\alpha = 0{,}05$.

Розв'язання.

Для цього впорядкуємо їх за зростанням.

ЗН	X/Y	R_i
0,8	Y	1
1,2	X	2
2,4	Y	3
3,4	X	4
4,2	Y	5
5,1	Y	6
6,2	X	7
6,8	Y	8
8,1	Y	9
10,2	Y	10
11,3	Y	11
11,4	X	12
13,0	Y	13
13,8	X	14
15,9	Y	15
20,1	Y	16
24,2	Y	17
26,7	Y	18

У цій таблиці – перший стовпчик - усі 18 значень випадкових величин X та Y впорядкованих за зростанням.

Другий стовпчик містить інформацію з якої вибірки отриманий відповідний елемент.

Третій стовпчик – ранг кожного елемента у загальній вибірці.

Розрахуємо статистику Ансарі-Бредлі, використавши ранги елементів випадкової величини X:

$$S = \sum_{i=1}^{8} \left\{ \frac{8+10+1}{2} - \left| R_i - \frac{8+10+1}{2} \right| \right\} = \sum_{i=1}^{8} \{9{,}5 - |R_i - 9{,}5|\} =$$

$$= 76 + (|2 - 9{,}5| + |4 - 9{,}5| + |7 - 9{,}5| + \cdots + |13 - 9{,}5| + |15 - 9{,}5|) = 49.$$

Тепер, з таблиці (Додаток 13) для $m = 8$, $n = 10$ і $\gamma = 0{,}95$ знаходимо $S_1(\gamma) = 28$; $S_2(\gamma) = 52$.

Оскільки

$$S_1(\gamma) = 28 < S = 49 < S_2(\gamma) = 52,$$

то гіпотеза H_0 про рівність параметрів масштабу для випадкових величин X та Y приймається.

Проілюструємо ще, як у цьому прикладі використати нормальну апроксимацію.

Оскільки число $m + n = 18$ парне, то

$$\mathbf{M}(S) = \frac{8 \cdot 20}{4} = 40; \quad \mathbf{D}(S) = \frac{8 \cdot 10 \cdot 16 \cdot 20}{78 \cdot 17} = 31{,}37; \quad \sqrt{\mathbf{D}(S)} = 5{,}60.$$

Тому

$$S^* = \frac{49 - 40}{5{,}601} = 1{,}61.$$

При $\gamma = 0{,}95$ квантиль стандартного нормального розподілу рівний $Z_{\text{кр}} = 1{,}96$.

Оскільки

$$S^* = 1{,}61 < 1{,}96 = Z_{\text{кр}},$$

то гіпотезу H_0 приймаємо.

Приклад 8. Критерій Бхапкара-Дешпанде

Припустимо, що є $k = 3$ вибірки по 4 спостереження в кожній:

$$X_1: \ 300; \quad 400; \quad 510; \quad 600;$$

$$X_2: \ 250; \quad 440; \quad 510; \quad 900;$$

$$X_3: \ 520; \quad 610; \quad 920; \quad 1070.$$

Перевірити наявність зміни масштабу на рівні значущості $\alpha = 0{,}05$.

Розв'язання.

Тут $\prod_{i=1}^{3} n_i = 4^3 = 64$. Усі такі 64 набори виписувати не будемо. Прямим перебором можна переконатися, що

$$v_1^1 = 35, \quad v_2^1 = 27, \quad v_3^1 = 2,$$
$$v_1^3 = 3, \quad v_2^3 = 11, \quad v_3^3 = 50,$$

Тоді $v_1^1 + v_2^1 + v_3^1 = 35 + 27 + 2 = 64$,

а $v_1^3 + v_2^3 + v_3^3 = 3 + 11 + 50 = 64$.

Далі, знаходимо

$$u_1^1 = \frac{v_1^1}{64} = \frac{35}{64} = 0{,}55; \;\; u_2^1 = \frac{27}{64} = 0{,}42; \;\; u_3^1 = \frac{2}{64} = 0{,}03;$$
$$u_1^3 = \frac{3}{64} = 0{,}05; \;\; u_2^3 = \frac{11}{64} = 0{,}17; \;\; u_3^3 = \frac{50}{64} = 0{,}78;$$
$$d_1 = u_1^1 + u_1^3 = 0{,}55 + 0{,}05 = 0{,}6;$$
$$d_2 = u_2^1 + u_2^3 = 0{,}42 + 0{,}17 = 0{,}59;$$
$$d_3 = u_3^1 + u_3^3 = 0{,}03 + 0{,}78 = 0{,}81.$$

Аналогічно розраховуємо

$$l_1 = -u_1^1 + u_1^3 = -0{,}55 + 0{,}05 = -0{,}5;$$
$$l_2 = -u_2^1 + u_2^3 = -0{,}42 + 0{,}17 = -0{,}25;$$
$$l_3 = -u_3^1 + u_3^3 = -0{,}03 + 0{,}78 = 0{,}75.$$

Обчислюємо статистики критеріїв:

$$V = 64 \cdot (2 \cdot 3 - 1)\left[\sum_{i=1}^{3} \frac{4}{64}\left(u_i^j - \frac{1}{3}\right)^2 - \left\{\sum_{i=1}^{3} \frac{4}{64}\left(u_i^j - \frac{1}{3}\right)\right\}^2\right] =$$

$$= 320 \cdot \left[\frac{\left(0{,}5469 - \frac{1}{3}\right)^2}{16} + \frac{\left(0{,}4219 - \frac{1}{3}\right)^2}{16} + \frac{\left(0{,}03125 - \frac{1}{3}\right)^2}{16} - \right.$$

$$\left. - \left\{\frac{1}{16}[0{,}5469 + 0{,}4219 + 0{,}03125 - 1]\right\}^2\right] = 2{,}89.$$

$$L = \frac{64 \cdot 3 \cdot 4 \cdot C_4^2}{2[9 + (9 + 12 + 2) \cdot C_4^2]}\left[\sum_{i=1}^{3} \frac{1}{16} {d_i}^2 - \left\{\frac{1}{16}\sum_{i=1}^{3} d_i\right\}^2\right] = 1{,}82.$$

З таблиці для $\gamma = 0{,}95$ і $n = 4$ знаходимо критичні значення:

$$L_{0{,}95} = 6{,}15; \; D_{0{,}95} = 4{,}60.$$

Оскільки

$$D = 1{,}82 < 4{,}60 = D_{0{,}95};$$

$$L = 4{,}67 < 6{,}15 = L_{0{,}95},$$

то гіпотеза H_0 приймається. Зміни масштабу немає.

7.7. Питання для самоконтролю до Розділу 7

1. Сформулюйте гіпотезу зсуву і гіпотезу масштабу.
2. Які вибірки називаються залежними, а які незалежними.
3. Сформулюйте теорему Смирнова.
4. Опишіть схему застосування критерію Колмогорова-Смирнова перевірки статистичних гіпотез.
5. У чому полягає критерій Манна-Уітні перевірки статистичних гіпотез?
6. Опишіть схему застосування критерію Вілкоксона для незв'язних вибірок.
7. Опишіть схему застосування парного критерію Вілкоксона для порівняння зв'язних вибірок.
8. Обґрунтуйте критерій знаків перевірки статистичних гіпотез.
9. Якими є помилки 1-го та 2-го роду при застосуванні критерію знаків.
10. У чому полягає критерій зсуву Краскела-Уолліса для порівняння декількох незалежних вибірок?
11. Запишіть формули апроксимації Краскела-Уолліса.
12. Опишіть критерій Фрідмана для порівняння декількох зв'язних вибірок.
13. Опишіть критерій Кендала для порівняння декількох зв'язних вибірок.
14. Якою є суть гіпотези зміни масштабу?
15. Як застосовується критерій Ансарі-Бредлі зміни масштабу для двох вибірок?
16. Як застосовується критерій Бхапкара-Дешпанде для декількох вибірок?

РОЗДІЛ 8. МЕТОД БАГАТОВИМІРНОЇ КЛАСИФІКАЦІЇ

8.1. Кластерний аналіз

У статистиці групування вихідних даних є основним методом розв'язування задачі класифікації.

Традиційно ця задача розв’язується так: із сукупності ознак, які описують об'єкт, вибирається одна, найважливіша для дослідника і проводиться в групування у відповідності зі значеннями цієї ознаки.

Якщо потрібно провести класифікацію за кількома ознаками, то спочатку класифікують їх за першою, найважливішою ознакою, потім кожен із отриманих класів розбивається за другою ознакою (менш важливою) і т. д.

Якщо згадані ознаки не вдається впорядкувати за важливістю, застосовується метод **інтегрального показника**, який залежить від вихідних ознак.

Класифікацію здійснюють за допомогою **методів кластерного аналізу**, які ми й розглянемо у цьому розділі.

Різні схеми класифікації оцінюються за допомогою **цільової функції**, значення якої дозволяють їх порівнювати.

Приклад. Нехай досліджується сукупність n об'єктів (пацієнтів), кожен з яких характеризується k заміряними на ньому ознаками (температура, тиск, тощо). Потрібно розбити цю сукупність на однорідні в певному розумінні групи (класи): здоровий, одужує, хворий,...).

Отримані в результаті розбиття групи називають ***кластерами*** (від *cluster* – група елементів). Методи знаходження кластерів називають ***кластер-аналізом***.

При цьому, розглядають дві задачі класифікації.

1. **Звичайна задача класифікації**: група спостережень розбивається на певні області. Наприклад, коли задано варіаційний ряд

$$x_1 \le x_2 \le \cdots \le x_n$$

Значень випадкової величини X і потрібно поділити його на кілька інтервалів так, щоб елементи одного інтервалу по можливості були близькі між собою. Наприклад, огірки: корнішони, пікулі, нестандарт (рис.8.1).

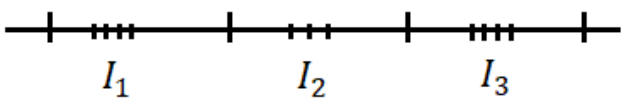


Рис. 8.1

2. **Задача природної класифікації**: знаходження «природного» розшарування спостережень на чітко виражені кластери, які перебувають на певній відстані один від одного. Наприклад, позитивний і негативний тест на коронавірус (рис. 8.2).

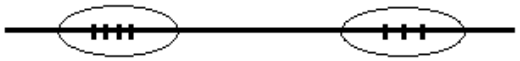


Рис. 8.2

Існують три різні підходи до проблеми кластерного аналізу: евристичний, екстремальний і статистичний.

Евристичний підхід характерний відсутністю формальної моделі і критеріїв для порівняння різних рішень. Наприклад: перебирають огірки: корнішони, пікулі, нестандарт.

Екстремальний підхід передбачає, що модель чітко не формулюється, але встановлюється чіткий критерій розбиття на класи: 0-10 г; 20-40 г; > 40 г.

Статистичний підхід. Його основою є побудова ймовірнісної моделі для досліджуваного процесу. Це підхід особливо зручний для теоретичних проблем, пов'язаних з кластерним аналізом. Цей підхід буде далі об’єктом нашої особливої уваги.

В задачах кластерного аналізу вихідні дані зображуються прямокутною таблицею (матрицею):

$$A = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}.$$

Тут кожен рядок означає результат вимірювання k ознак (параметрів) на одному з n об'єктів.

Наприклад, є n пацієнтів для кожного з них вимірюють k параметрів: температура, тиск, лейкоцити, цукор і т. д.

В конкретних ситуаціях може бути цікаво як групування об'єктів, так і групування ознак.

Числові значення, які входять до матриці A можуть відповідати трьом типам змінних:

1. ***Кількісні*** **–** це ознаки, які задаються кількісними значеннями. Наприклад, температура.

2. ***Рангові*** – це значення, які можна впорядкувати, але з ними не можна проводити арифметичних операцій, як із числами. Прикладом може бути місце учня на змаганнях з гімнастики.

3. ***Якісні*** – це ознаки, які не можна впорядкувати. Наприклад: цей пацієнт уже встає, але сам не їсть; а цей – їсть, але ще не встає.

Для кластерного аналізу використовують **матрицю близькості**

$$R = \left(r_{ij}\right)_{i,j=1}^{n},$$

де r_{ij} визначає міру близькості i-того об'єкта від j-го об'єкта.

Точніше, більшість алгоритмів кластерного аналізу базується на матриці близькості, тому, якщо дані зображені матрицею A, то першим етапом розв'язання задачі пошуку кластерів буде вибір способу обчислення відстані (чи близькості) між об'єктами, або ознаками.

Відстань між об'єктами і міра близькості

Отже, розглядаємо певну множину $\mathbf{X}$ об'єктів. Позначаємо їх прямим шрифтом $\mathrm{x}, \mathrm{y}\ldots$ і трактуємо як вектори

$$\mathrm{x} = (x_1, x_2, \ldots, x_k),$$

компоненти яких трактуємо як результати вимірювань (наприклад, пацієнт і набір виміряних параметрів його стану). Якщо ми хочемо поділити їх на підмножини (кластери), то маємо ввести поняття **міри близькості** чи **відстані**, чи **метрики** між об'єктами: число $\rho(\mathrm{x}, \mathrm{y})$.

Тоді близькі відносно цієї метрики об'єкти вважатимемо належними до однієї множини (кластера).

На міру близькості накладаємо добре відомі з функціонального аналізу умови:

1. $\rho(\mathrm{x}, \mathrm{x}) = 0, \ \rho(\mathrm{x}, \mathrm{y}) \geq 0.$
2. $\rho(\mathrm{x}, \mathrm{y}) = \rho(\mathrm{y}, \mathrm{x}).$
3. $\rho(\mathrm{x}, \mathrm{y}) \leq \rho(\mathrm{x}, \mathrm{z}) + \rho(\mathrm{z}, \mathrm{y}).$

Розглянемо докладніше два природні приклади такої метрики.

Евклідова метрика. Якщо $\mathrm{x} = (x_1, \ldots, x_k), \ \mathrm{y} = (y_1, \ldots, y_k)$, то

$$\rho(\mathrm{x}, \mathrm{y}) = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}.$$

Використання цієї метрики виправдане, коли:

а) Спостереження x, y беруться з генеральних сукупностей, які мають багатовимірний нормальний розподіл, компоненти кожного вектора $\mathrm{x} \in \mathbf{X}$ незалежні і мають одну й ту ж дисперсію

б) Компоненти вектора x однорідні за фізичним значенням і однаково важливі для класифікації.

в) Множина $\mathbf{X}$ «схожа» на евклідів простір $\mathbf{R}^k$.

Звичайно евклідова відстань не має сенсу, якщо ознаки (тобто компоненти x_i вектора x) вимірюються в різних одиницях. Для зведення ознак до однакових одиниць вимірювання, проводять нормування.

$$x_i^{\mathrm{n}} = \frac{x_i - \bar{x}}{s},$$

де $\bar{x} = \frac{1}{k}\sum_{i=1}^{k} x_i$, $s = \sqrt{\frac{1}{n}\sum_{i=1}^{k} (x_i - \bar{x})^2}$.

Зважена евклідова метрика:

$$\rho_{\mathrm{w}}(\mathrm{x},\mathrm{y}) = \sqrt{\sum_{i=1}^{k} w_i (x_i - y_i)^2},$$

де $w_i > 0$.

Вона застосовується, коли кожній компоненті x_i вектора спостережень вдається приписати «вагу» w_i, яка відображає рівень важливості цієї ознаки.

Відстань Хеммінга:

$$\rho_{\mathrm{H}}(\mathrm{x},\mathrm{y}) = \sum_{i=1}^{k} |x_i - y_i|.$$

Вона використовується для визначення відстані між об'єктами, що задані дихотомічними ознаками, тобто такими, що набувають значень $\{0, 1\}$. У цьому випадку відстань Хеммінга дорівнює кількості значень відповідних ознак, які не рівні між собою.

Відстань між кластерами

Нехай тепер S і T – два кластери (підмножини) множини $\mathbf{X}$. Якщо в $\mathbf{X}$ вибрано метрику ρ, то відстань між кластерами можна ввести кількома різними способами. Наведемо кілька природних прикладів такої відстані у випадку, коли кластери S і T скінченні.

1. ***Відстань «найближчого сусіда»***

$$\rho_{\min}(S,T) = \min_{\mathrm{x}\in S,\ \mathrm{y}\in T} \rho(\mathrm{x},\mathrm{y}).$$

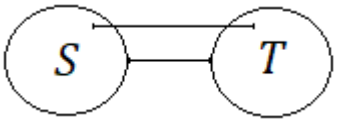


Рис. 8.3

2. ***Відстань «найвіддаленішого сусіда»***

$$\rho_{\max}(S,T) = \max_{\mathrm{x}\in S,\ \mathrm{y}\in T} \rho(\mathrm{x},\mathrm{y}).$$

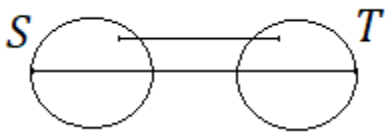


Рис. 8.4

3. ***Відстань між «центрами ваги» кластерів***. Покладемо $\bar{\mathrm{x}} = \frac{1}{n}\sum_{j=1}^{n} \mathrm{x}_j$, де $\left(\mathrm{x}_j\right)_1^n$ усі елементи кластера S, $\bar{\mathrm{y}} = \frac{1}{m}\sum_{j=1}^{m} \mathrm{y}_j$, де $\left(\mathrm{y}_j\right)_1^n$ – усі елементи кластера T.

Тоді

$$\rho_{\mathrm{c}}(S,T) = \rho(\bar{\mathrm{x}}, \bar{y})$$

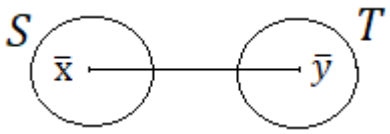


Рис. 8.5

4. ***Відстань, виміряна за принципом середнього зв'язку»***

$$\rho_{\mathrm{a}}(S,T) = \frac{1}{nm}\sum_{\mathrm{x}\in S}\sum_{\mathrm{y}\in T}\rho(\mathrm{x},\mathrm{y})\,.$$

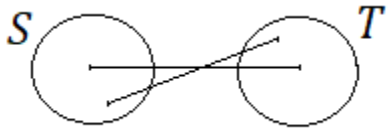


Рис. 8.6

5. ***Відстань, виміряна за принципом «степеневого середнього»***. Нехай $r \in \mathbf{R}$. Тоді

$$\rho_r(S,T) = \left[\frac{1}{nm}\sum_{\mathrm{x}\in S}\sum_{\mathrm{y}\in T}\rho^r(\mathrm{x},\mathrm{y})\right]^{1/r}. \qquad (8.1)$$

Можна показати, що:

1. При $r \to +\infty$ $\rho_r(S,T) \to \rho_{\max}(S,T)$.
2. При $r \to -\infty$ $\rho_r(S,T) \to \rho_{\min}(S,T)$.
3. При $r = 1$ $\rho_r(S,T) \to \rho_{\mathrm{a}}$.

З формули (8.1) випливає, що для трьох кластерів S, T, V

$$\rho_r(V, S\cup T) = \left[\frac{n[\rho_r(V,S)]^r + m[\rho_r(V,T)]^r}{nm}\right]^{1/r}.$$

Функціонали якості розбиття

Існує багато способів розбиття на класи заданої сукупності елементів. Тому, природною є задача порівняння якості цих способів розбиття. З цією метою вводиться поняття ***функціоналу якості*** $Q(S)$, визначеного на сукупності усіх розбиттів.

Найкращим вважається розбиття S^*, на якому функціонал якості $Q(S)$ досягає екстремуму. Далі ми розглянемо найпоширеніші функціонали якості розбиття. Нехай у множині $\mathbf{X}$, яка складається із скінченної кількості елементів $\mathrm{x}_1, \ldots, \mathrm{x}_n$ вибрана метрика ρ і нехай $S = (S_1, \ldots, S_p)$ – фіксована кількість розбиттів множини $\mathbf{X}$. Найпоширенішим функціоналами якості є:

1. ***Сума внутрішньо-класових дисперсій***

$$Q_1(S) = \sum_{l=1}^{p} \sum_{\mathrm{x}_i \in S_l} \rho^2(\mathrm{x}_i, \bar{\mathrm{x}}_l);$$

Нагадаємо $\bar{\mathrm{x}}_l = \frac{1}{\mathrm{card} S_l} \sum_{\mathrm{x} \in S_l} \mathrm{x}_i$, $l = 1, \ldots, p$.

Ілюстрацію цього функціонала зображено на рис. 8.7.

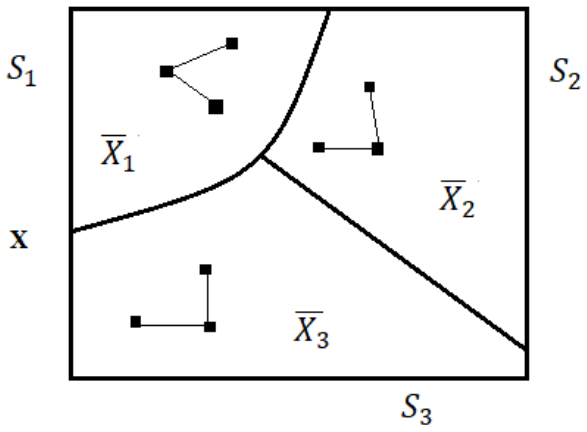


Рис. 8.7

2. ***Сума попарних внутрішньо-класових відстаней між елементами***

$$Q_2(S) = \sum_{l=1}^{p} \sum_{\mathrm{x}_i, \mathrm{x}_j \in S_l} \rho^2(\mathrm{x}_i, \mathrm{x}_j).$$

Ілюстрацію цього функціоналу зображено на рис. 8.8.

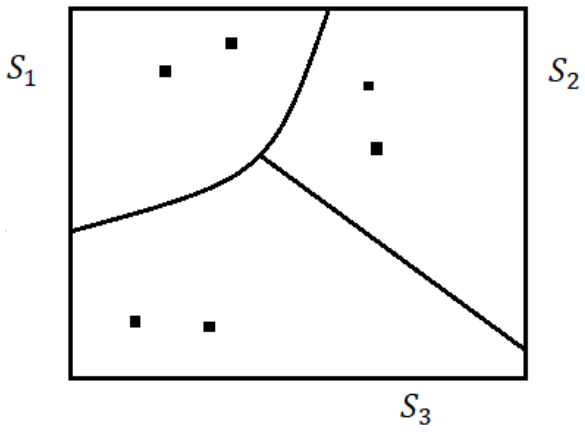


Рис. 8.8

Іноді використовується усереднена версія цього функціоналу

$$Q_2'(S) = \sum_{l=1}^{p} \frac{1}{n_l} \sum_{\mathrm{x}_i, \mathrm{x}_j \in S_l} \rho^2(\mathrm{x}_i, \mathrm{x}_j).$$

Тут n_l – кількість елементів множини S_l.

3. ***Узагальнена внутрішньо-класова дисперсія***

$$Q_3(S) = \det\left(\sum_{l=1}^{p} n_l w_l\right).$$

Тут, для кожного $l = 1, \ldots, p$; w_l означає вибіркову коваріаційну матрицю класу S_l. Її елементи обчислюються за формулою, наведеною нижче. Але спочатку нагадаємо, що кожен об'єкт $\mathrm{x} \in \mathbf{X}$ має вигляд

$$\mathrm{x} = (x_1, \ldots, x_k).$$

Зафіксуємо координату $q \in (1, \ldots, k)$ і клас S_l. Для них позначимо через $\bar{x}_q^l$ середнє значення q-тих компонент (координат) усіх векторів з S_l. Тоді елемент $w_{qm}(l)$ матриці W_l обчислюється за формулою

$$w_{qm}(l) = \frac{1}{n_l} \sum_{X \in S_l} \left(x_q^l - \bar{x}_q^l\right)\left(x_m^l - \bar{x}_m^l\right), \quad q, m, = 1, \ldots, k.$$

Іноді використовується наступна версія цього функціоналу

$$Q'_3(S) = \prod_{l=1}^{p} (\det w_l)^{n_l}.$$

Ієрархічні кластер-процедури

Ієрархічні (або деревоподібні) процедури є найпоширенішими алгоритмами кластерного аналізу при їхній реалізації у комп'ютерних статистичних пакетах. Вони бувають двох типів: **агломеративні** і **дивізимні**. В алгомеративних процедурах початковим є розбиття, що складається з n одноелементних класів, а кінцевим – з одного класу. В дивізимних – навпаки.

Принцип роботи агломеративних (відповідно дивізимних) процедур полягає у послідовному об'єднанні (відповідно розділені) груп елементів. Спочатку найближчих (відповідно найдальших), а потім все дальших (відповідно ближчих).

Приклад агломеративної процедури буде наведено в підрозділі 8.3.

8.2. Дискримінантний аналіз

Метод класифікації з навчанням

У загальному випадку задача розрізнення (дискримінації) з навчанням має такий вигляд. Нехай множина $\mathbf{X}$ векторів $\mathrm{x} = (x_1, \dots, x_k)$ довжини k розбита на підмножини $S_1, \dots, S_p$. Ці множини S_l точно невідомі. Потрібно встановити ***правило***, згідно з яким даний вектор x належить до однієї з множин S_l $(l = 1,2, \dots, p)$. Таке правило формулюють за допомогою "навчання". А саме, проводять експерименти $\mathrm{X}_1, \dots, \mathrm{X}_n$ і встановлюють, які з результатів $\mathrm{x}_1, \dots, \mathrm{x}_n$ цих експериментів до яких множин $S_1, \dots, S_p$ належать. Нагадаємо, що, як це є традиційним в статистиці, до проведення експерименту ми трактуємо $\mathrm{X}_1, \dots, \mathrm{X}_n$ як випадкові вектори, а після проведення, як числові вектори (результати експериментів). Так ось, на підставі отриманої інформації і встановлюється *правило.* За цим правилом ми відносимо новий вектор (щодо якого експерименту не проведено) до якоїсь з множин S_l . А потім, пізніше, проводиться експеримент і виявляється, до якої з множин S_l він насправді належить. Тепер, на підставі нової інформації про вектори $\mathrm{x}_1, \dots, \mathrm{x}_n, \mathrm{x}$ правило уточнюється. Ось на цьому й полягає ***навчання***. Проілюструємо сказане на двох прикладах.

Результатом спостереження над об'єктом є реалізація k-вимірного випадкового вектора $\mathrm{x} = (x_1, \dots, x_k)$. Для побудови правила дискримінації уся множина значень вектора x розбивається на підмножини R_i $(i = 1, .., p)$ так, що при потраплянні вектора x до R_i, об'єкт відноситься до множини S_i.

Приклад (з геології). Нехай ми шукаємо розміри нафтового родовища. Ілюстрація представлена на рис. 8.9.

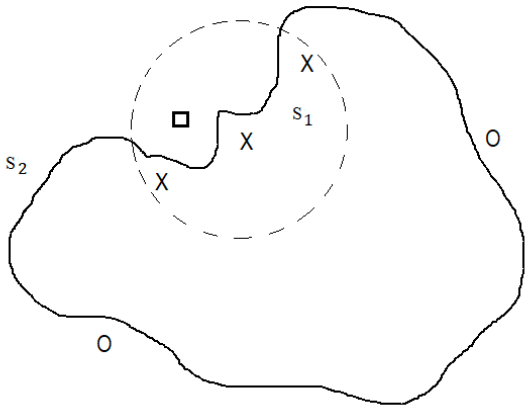


Рис. 8.9

Тут вектор X двовимірний (довгота й широта на земній кулі). S_1 – є нафта; S_2 – немає. Розміри родовища ми не знаємо. Проводимо буріння: $\mathrm{x}_1, \ldots, \mathrm{x}_n$. Хрестики – є нафта; нулики – немає. За якимось правилом, про яке ми говоритимемо далі, окреслюємо розміри родовища. Вони нарисовані пунктиром. Якщо ми хочемо добувати нафту – природно проводити наступне буріння x всередині окресленої пунктиром області. Оце є наше x. Буримо. Намальовано квадратиком. Нафти немає. На підставі нової інформації (про значення x) за нашим правилом інакше окреслюємо розміри родовища і т. д. От у цьому проявляється навчання. З кожним новим бурінням ми отримуємо додаткову інформацію (навчаємося) про розміри родовища.

Приклад (з медичної діагностики). Нехай ми проводимо діагностику раку. Діагноз ставиться на підставі вектора аналізів $\mathrm{x} = (x_1, \ldots, x_k)$: температура, тиск, лейкоцити та інше. Нехай ми мали аналізи пацієнтів $\mathrm{x}_1, \ldots, \mathrm{x}_n$, про яких вже відомо – був у них рак насправді, або ні. На підставі цієї інформації ми за якимось правилом, окреслюємо області S_1 – при цьому наборі даних рак є і S_2 – немає.

А тепер до нас поступає новий пацієнт. Робимо аналізи: вектор x. На підставі нашого правила ставимо діагноз: рак є, або немає. Пізніше виявляється, чи насправді у пацієнта був рак. Тепер вже більше навчившись (маючи додаткову інформацію x) змінюємо наше правило на реалістичніше.

Випадкові вектори

Далі буде описане одне зі згаданих вище правил. Але спочатку нагадаємо необхідні поняття, що стосуються випадкових векторів.

Нехай задано ймовірнісний простір $(\Omega, F, \mathbf{P})$ і лінійний простір $\mathbf{R}^k$. Вимірна функція $\mathrm{X}\colon \Omega \to \mathbf{R}^k$ називається випадковою величиною зі значенням в $\mathbf{R}^k$, або ще – випадковим вектором. Для кожного $\omega \in \Omega$, його значення

$$\mathrm{X}(\omega) = (X_1(\omega), X_2(\omega), \ldots, X_k(\omega)) \tag{8.2}$$

– це вектор довжини k. З випадковим вектором зв'язаний розподіл

$$F_{\mathrm{X}}(A) = P\{\mathrm{X} \in A\},\ A \subset \mathbf{R}^k,$$

де A – вимірна множина. Як і для звичайних (числових) випадкових величин виділяють дискретні та неперервні випадкові розподіли. Якщо випадковий вектор X набуває лише скінчене або злічене число значень $(\mathrm{x}_n)_{n\geq 1}$ з ймовірностями $(p_n)_{n\geq 1}$, то говорять про дискретний розподіл, а якщо існує така невід'ємна інтегрована функція $f(\mathrm{x})$, що для кожної вимірної множини $A \subset \mathbf{R}^k$ має місце рівність

$$\int_A f(\mathrm{x})d\mathrm{x} = F_{\mathrm{X}}(A),$$

то говорять про неперервний розподіл. При цьому функцію $f(\mathrm{x})$ називають щільністю розподілу випадкового вектора X. Як і для числових випадкових величин, точний вигляду вектора X невідомий, та це не дуже й потрібно. Важливим є його розподіл $F_{\mathrm{X}}(A)$. Тому, часто говорять просто про розподіли і знак X пропускають, пишучи просто $\mathbf{P}_F(A)$.

Ймовірнісний дискримінантний аналіз

Отже, нехай у просторі $\mathbf{R}^k$ задано певні класи (множини) $S_1, \dots, S_p$. Нехай точно ці класи не відомі, а відомі лише розподіли $F_1, \dots, F_p$, які певною мірою характеризують ці класи, у тому розумінні, що значення розподілу F_l на S_l більші, а поза S_l – менші. Нашим завданням є отримавши вектор $\mathrm{x}_0 \in \mathbf{R}^k$ і, знаючи розподіли $F_1, \dots, F_p$, вирішити, до якого класу з $S_1, \dots, S_p$ належить цей вектор. Уточнимо згадане завдання для дискретного і неперервного випадків.

а) Дискретні розподіли. Отже, нехай усі розподіли $F_1, \dots, F_p$ дискретні. Точніше, нехай усі множини $S_1, \dots, S_p$ скінченні, або зліченні, і на їхньому об'єднанні $A = \bigcup_{i=1}^{p} S_i$ задані ймовірності $\mathbf{P}_1, \dots, \mathbf{P}_p$, які характеризують ці класи. Отримавши вектор $\mathrm{x}_0 \in A$, обчислюємо значення

$$\mathbf{P}_1(\mathrm{x}_0), \mathbf{P}_2(\mathrm{x}_0) \dots, \mathbf{P}_p(\mathrm{x}_0)$$

і вибираємо серед них найбільше $\mathbf{P}_{l0}(\mathrm{x}_0)$ та робимо висновок, що вектор x_0 належить до класу S_{l0}.

б) Неперервні розподіли. Нехай тепер усі розподіли $F_1, \dots, F_p$ неперервні і мають щільності $f_1, \dots, f_p$, які задано на всьому просторі $\mathbf{R}^k$. Отримавши вектор $\mathrm{x}_0 \in \mathbf{R}^k$, обчислимо

$$f_1(\mathrm{x}_0), f_2(\mathrm{x}_0) \dots, f_p(\mathrm{x}_0);$$

вибираємо серед них найбільше $f_{l0}(\mathrm{x})$ і робимо висновок, що вектор x_0 належить до класу S_{i0}.

Параметри випадкового вектора

У застосовуваннях особливо важливим є випадок, коли $F_1, \dots, F_p$ розподілені нормально. Нормальний вектор визначається своїм математичним сподіванням і певним аналогом дисперсії. Розглянемо їх докладніше.

Поняття математичного сподівання випадкового вектора вводяться просто. За означенням,

$$\mathbf{M}\mathrm{X} = (\mathbf{M}X_1, \mathbf{M}X_2, \dots, \mathbf{M}X_k).$$

Отже, $\mathbf{M}\mathrm{X}$ – це вектор довжини k, складений з математичних сподівань координат. Введення аналога дисперсії для випадкового вектора не таке просте. Нагадаємо спершу, що коваріацією двох випадкових (числових) величин X і Y називається число

$$\mathrm{Cov}(X, Y) = \mathbf{M}\big((X - \mathbf{M}X)(Y - \mathbf{M}Y)\big).$$

Так ось, коваріаційною матрицею випадкових векторів $\mathrm{X} = (X_1, X_2, \dots, X_k)$ та $\mathrm{Y} = (Y_1, Y_2, \dots, Y_k)$ називається наступна числова матриця

$$\mathrm{Cov}(\mathrm{X}, \mathrm{Y}) = \begin{pmatrix} \mathrm{Cov}(X_1, Y_1) & \mathrm{Cov}(X_1, Y_2) & \dots & \mathrm{Cov}(X_1, Y_k) \\ \mathrm{Cov}(X_2, Y_1) & \mathrm{Cov}(X_2, Y_2) & \dots & \mathrm{Cov}(X_2, Y_k) \\ \dots & \dots & \dots & \dots \\ \mathrm{Cov}(X_k, Y_1) & \mathrm{Cov}(X_k, Y_2) & \dots & \mathrm{Cov}(X_k, Y_k) \end{pmatrix}. \qquad (8.3)$$

Отже, коваріація випадкових векторів X та Y, це матриця. Пригадаємо, що для числової випадкової величини X, дисперсія

$$\mathbf{D}X = \mathrm{Cov}(X, X).$$

Тому роль дисперсії для випадкового вектора X відіграє матриця $\mathrm{Cov}(\mathrm{X}, \mathrm{X})$.

Статистика параметрів випадкового вектора

Пригадаємо, що в математичній статистиці самої випадкової величини X (і її параметрів) ми часто не знаємо. Для оцінки невідомого математичного сподівання $\mathbf{M}X$ беруть n незалежних копій $X^1, X^2, \dots, X^n$ випадкової величини X і середнє

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

вважається оцінкою для $\mathbf{M}X$. Тому для оцінки невідомого математичного сподівання $\mathbf{M}\mathrm{X}$ вектора $\mathrm{X} = (X_1, \dots, X_k)$ природно взяти n незалежних копій $\mathrm{X}^i = \big(X_1^i, X_2^i, \dots, X_k^i\big), i = 1, \dots, n$, цього вектора і за оцінку математичного сподівання $\mathbf{M}\mathrm{X}$ вважати вектор

$$\overline{\mathrm{X}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{X}^i = \left(\frac{1}{n}\sum_{i=1}^{n} X_1^i, \frac{1}{n}\sum_{i=1}^{n} X_2^i, \dots, \frac{1}{n}\sum_{i=1}^{n} X_k^i\right).$$

Перейдемо тепер до коваріаційної матриці. Пригадаємо, що для випадкових величин X та Y їхня коваріація $\mathrm{Cov}(X, Y)$ оцінювалася за формулою

$$\overline{\mathrm{cov}}(X, Y) = \frac{1}{n-1}\sum_{i=1}^{n} \big(X^i - \bar{X}\big)\big(Y^i - \bar{Y}\big).$$

Число $n - 1$ замість n ставлять для того, щоб оцінка для $\mathrm{Cov}(X, Y)$ була незміщеною. Тому, для незміщеної оцінки коваріації $\mathrm{Cov}(\mathrm{X}, \mathrm{Y})$ природно брати наступну матрицю. Нехай $\mathrm{X} = (X_1, \dots, X_k)$, $\mathrm{Y} = (Y_1, \dots, Y_k)$. Беремо по n їхніх незалежних копій $\mathrm{X}^i = \big(X_1^i, X_2^i, \dots, X_k^i\big)$, $\mathrm{Y}^i = \big(Y_1^i, Y_2^i, \dots, Y_k^i\big), i = 1, \dots, n$. Тоді оцінка має вигляд

$$\overline{\mathrm{cov}}(\mathrm{X},\mathrm{Y}) = \begin{pmatrix} \overline{\mathrm{cov}}(X_1,Y_1) & \overline{\mathrm{cov}}(X_1,Y_2) & \dots & \overline{\mathrm{cov}}(X_1,Y_k) \\ \overline{\mathrm{cov}}(X_2,Y_1) & \overline{\mathrm{cov}}(X_2,Y_2) & \dots & \overline{\mathrm{cov}}(X_2,Y_k) \\ \dots & \dots & \dots & \dots \\ \overline{\mathrm{cov}}(X_k,Y_1) & \overline{\mathrm{cov}}(X_k,Y_2) & \dots & \overline{\mathrm{cov}}(X_k,Y_k) \end{pmatrix}.$$

Наголосимо, $\overline{\mathrm{cov}}(\mathrm{X},\mathrm{Y})$ – це матриця, складена з випадкових величин.

Нормальний випадковий вектор

Пригадаємо, що випадкова величина X називається нормальною, якщо її щільність розподілу має вигляд

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

Параметр a тут відіграє роль математичного сподівання, а σ – середнього квадратичного відхилення.

Як ввести поняття нормального випадкового вектора? На перший погляд так: це – випадковий вектор

$$\mathrm{X} = (X_1, \dots, X_k), \tag{8.4}$$

де кожна координата $X_1, \dots, X_k$ має нормальний розподіл. Проте таке поняття не дуже зручне. Воно залежить від базису. Якщо

$$\mathrm{e}_1 = (1,0,\dots,0);$$
$$\mathrm{e}_2 = (0,1,\dots,0);$$
$$\dots\dots\dots\dots\dots\dots\dots;$$
$$\mathrm{e}_k = (0,0,\dots,1);$$

то

$$\mathrm{X} = \sum_{i=1}^{n} X_i \cdot \mathrm{e}_i.$$

Коли введемо інший базис в $\mathbf{R}^k$, то випадковий вектор у розумінні (8.4) може перестати бути нормальним. Зручнішим є наступне означення.

Означення. Кажуть, що випадковий вектор (8.4) має нормальний розподіл, якщо $\forall(a_1, \dots, a_k) \in \mathbf{R}^k$ випадкова величина

$$\sum_{i=1}^{n} a_i X_i$$

має нормальний розподіл.

Для нас буде важливим наступне твердження про щільність розподілу нормального випадкового вектора.

Твердження. Нехай $\mathrm{X} = (X_1, \dots, X_k)$ – нормальний випадковий вектор, $\mathrm{a} = (a_1, \dots, a_k) = (\mathbf{M}X_1, \dots, \mathbf{M}X_k)$ його математичне сподівання і нехай його коваріаційна матриця $\Sigma = \mathrm{Cov}(\mathrm{X}, \mathrm{X})$ невироджена, тобто визначник $|\Sigma| \neq 0$. Тоді щільність розподілу випадкового вектора X має вигляд

$$f(\mathrm{x}) = \frac{1}{\sqrt{(2\pi)^2 |\Sigma|}} \exp\left\{-\frac{1}{2}\langle \mathrm{x} - \mathrm{a}, \Sigma^{-1} \cdot (\mathrm{x} - \mathrm{a})\rangle\right\}, \qquad \mathrm{x} \in \mathbf{R}^k.$$

Кутові дужки в цій формулі означають скалярний добуток.

Порівняння значень щільності нормальних векторів

Нагадаємо, що в імовірнісному дискримінантному аналізі з неперервними розподілами ми порівнюємо значення щільності

$$f_1(\mathrm{x}_0), \dots, f_p(\mathrm{x}_0)$$

і відносимо x_0 до того класу, для якого значення щільності найбільше. Звичайно, ці значення можна обчислити і порівняти їх. Але обчислення можуть бути громіздкими. Інколи порівняння вдається зробити без занадто громіздких обчислень. Розглянемо один такий окремий випадок.

Припущення. Нехай випадкові вектори X та Y мають нормальні розподіли з математичними сподіваннями a та b відповідно і однією й тією ж невиродженою коваріаційною матрицею Σ. Нехай f і g – щільності випадкових векторів X та Y відповідно.

Якщо це припущення має місце, то для фіксованого вектора $\mathrm{x} \in \mathbf{R}^k$

$$\frac{1}{\sqrt{(2\pi)^2 |\Sigma|}} \exp\left\{-\frac{1}{2}\langle \mathrm{x} - a, \Sigma^{-1} \cdot (\mathrm{x} - \mathrm{a})\rangle\right\} = f(\mathrm{x}) > g(\mathrm{x}) =$$
$$= \frac{1}{\sqrt{(2\pi)^2 |\Sigma|}} \exp\left\{-\frac{1}{2}\langle \mathrm{x} - \mathrm{b}, \Sigma^{-1} \cdot (\mathrm{x} - \mathrm{b})\rangle\right\}$$

тоді і тільки тоді, коли

$$\langle \mathrm{x} - \mathrm{a}, \Sigma^{-1} \cdot (\mathrm{x} - \mathrm{a})\rangle < \langle \mathrm{x} - \mathrm{b}, \Sigma^{-1} \cdot (\mathrm{x} - \mathrm{b})\rangle$$

Розкриємо дужки

$$\langle \mathrm{x}, \Sigma^{-1} \cdot \mathrm{x}\rangle - \langle \mathrm{x}, \Sigma^{-1} \cdot \mathrm{x}\rangle - \langle \mathrm{a}, \Sigma^{-1} \cdot \mathrm{x}\rangle + \langle \mathrm{a}, \Sigma^{-1} \cdot \mathrm{a}\rangle <$$
$$< \langle \mathrm{x}, \Sigma^{-1} \cdot \mathrm{x}\rangle - \langle \mathrm{x}, \Sigma^{-1} \cdot \mathrm{b}\rangle - \langle \mathrm{b}, \Sigma^{-1} \cdot \mathrm{x}\rangle + \langle \mathrm{b}, \Sigma^{-1} \cdot \mathrm{b}\rangle.$$

Далі скористаємося тим фактом, що коваріаційна матриця симетрична, отже її обернена матриця також симетрична. Поклавши

$$\lambda = \frac{1}{2}\langle \mathrm{a}, \Sigma^{-1} \cdot \mathrm{a}\rangle, \qquad \mu = \frac{1}{2}\langle \mathrm{b}, \Sigma^{-1} \cdot \mathrm{b}\rangle,$$

дістанемо

$$-2\langle \mathrm{x}, \Sigma^{-1} \cdot \mathrm{a}\rangle + 2\lambda < -2\langle \mathrm{x}, \Sigma^{-1} \cdot \mathrm{x}\rangle + 2\mu,$$

і, врешті

$$\langle \mathrm{x}, \Sigma^{-1} \cdot (\mathrm{a} - \mathrm{b})\rangle > \lambda - \mu.$$

Дискримінантний аналіз для двох класів і трьох показників

У цьому пункті ми розглянемо окремий випадок загальної задачі дискримінації, в якому щось можна рахувати вручну і представимо алгоритм проведення підрахунків. А саме, розглянемо випадок дискримінації двох множин S_1 і S_2. Розглянемо випадок, коли $k = 3$, тобто коли маємо лише три показники. Це означає, що значення випадкових векторів, про які йдеться в задачі дискримінації, належать до простору $\mathbf{R}^3$. Вважатимемо також, що ці випадкові вектори мають нормальний розподіл, але його параметри (математичне сподівання і коваріаційна матриця) невідомі. Ці параметри оцінюємо на підставі результатів експерименту. Отже, перший крок алгоритму

1. Знаходження параметрів розподілу.

Нехай було проведено певну кількість експериментів. Результати експериментів потрапили або до класу S_1, або до класу S_2. Нехай до класу S_1 потрапили результати $\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_n$, а до класу S_2 – $\mathrm{y}_1, \mathrm{y}_2, \dots, \mathrm{y}_m$.

Вони мають вигляд:

$$\begin{aligned} \mathrm{x}_1 &= (x_{11}, x_{12}, x_{13}) & \mathrm{y}_1 &= (y_{11}, y_{12}, y_{13}) \\ \mathrm{x}_2 &= (x_{21}, x_{22}, x_{23}) & \mathrm{y}_2 &= (y_{21}, y_{22}, y_{23}) \\ &\dots\dots\dots\dots & &\dots\dots\dots\dots \\ \mathrm{x}_n &= (x_{n1}, x_{n2}, x_{n3}) & \mathrm{y}_m &= (y_{m1}, y_{m2}, y_{m3}). \end{aligned}$$

За означенням, оцінками математичних сподівань є середні

$$\bar{\mathrm{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{x}_i\,, \qquad \bar{\mathrm{y}} = \frac{1}{m}\sum_{i=1}^{m} \mathrm{y}_i.$$

Координати цих середніх мають вигляд:

$$\bar{x}_1 = \frac{1}{n}\sum_{i=1}^{n} x_{i1}\,, \qquad \bar{x}_2 = \frac{1}{n}\sum_{i=1}^{n} x_{i2}\,, \qquad \bar{x}_3 = \frac{1}{n}\sum_{i=1}^{n} x_{i3}\,;$$

$$\bar{y}_1 = \frac{1}{m}\sum_{i=1}^{m} y_{i1}\,, \qquad \bar{y}_2 = \frac{1}{m}\sum_{i=1}^{m} y_{i2}\,, \qquad \bar{y}_3 = \frac{1}{m}\sum_{i=1}^{m} y_{i3}.$$

Перейдімо до знаходження оцінок коваріаційних матриць. Оцінку коваріаційної матриці для векторів $\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_n$ позначимо через S_X, а для векторів $\mathrm{y}_1, \mathrm{y}_2, \dots, \mathrm{y}_m$ – через S_Y. Ці матриці мають вигляд

$$S_X = \begin{pmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{22} & s_{23} \\ s_{31} & s_{32} & s_{33} \end{pmatrix}, \qquad S_Y = \begin{pmatrix} t_{11} & t_{12} & t_{13} \\ t_{21} & t_{22} & t_{23} \\ t_{31} & t_{32} & t_{33} \end{pmatrix}.$$

Застосувавши згадану раніше загальну формулу, елементи цих матриць знаходимо за формулами

$$s_{jk} = \frac{1}{n}\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k), \quad t_{jk} = \frac{1}{m}\sum_{i=1}^{m} (y_{ij} - \bar{y}_j)(y_{ik} - \bar{y}_k), \quad (j, k = 1,2,3).$$

Далі, знаходимо незміщену оцінку сумарної коваріаційної матриці

$$\hat{S} = \frac{1}{n+m-2}(nS_X + mS_Y)$$

і, врешті, знаходимо обернену матрицю S^{-1}. Оскільки ранг матриці S дорівнює 3, то таку операцію можна порахувати вручну.

2. Знаходження параметрів дискримінаційної функції.

Тут варто пригадати останню формулу попереднього пункту. Отже:

а) Знаходимо вектор оцінок коефіцієнтів дискримінантної функції

$$\bar{V} = \hat{S}^{-1}(\bar{\mathrm{x}} - \bar{\mathrm{y}}).$$

б) Нехай

$$X = \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ \dots & \dots & \dots \\ x_{31} & x_{32} & x_{33} \end{pmatrix}, \qquad Y = \begin{pmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \\ \dots & \dots & \dots \\ y_{31} & y_{32} & y_{33} \end{pmatrix}.$$

Для них знаходимо вектори

$$\vec{U}_X = X \cdot \bar{V}, \qquad \vec{U}_Y = Y \cdot \bar{V},$$

де $\vec{U}_X = (U_{X1}, U_{X2}, \dots, U_{Xn})$, $\vec{U}_Y = (U_{Y1}, U_{Y2}, \dots, U_{Ym})$.

в) Знаходимо їхні середні значення:

$$\bar{U}_X = \frac{1}{n}\sum_{i=1}^{n} U_{Xi}\,, \qquad \bar{U}_Y = \frac{1}{m}\sum_{i=1}^{m} U_{Yi}.$$

г) Знаходимо константу

$$\hat{C} = \frac{1}{2}(\bar{U}_X + \bar{U}_Y).$$

3. Розв'язування задачі дискримінації.

Нехай тепер маємо вектор $\mathrm{z} = (z_1, z_2, z_3)$ і потрібно вирішити, до якого з класів S_1, S_2 він належить. Для цього вирішення знаходимо скалярний добуток $\langle \vec{Z}, \bar{V} \rangle$.

Якщо $\langle \vec{Z}, \bar{V} \rangle \geq \hat{C}$, то цей вектор відносимо до множини S_1, а якщо $\langle \vec{Z}, \bar{V} \rangle < \hat{C}$ – до множини S_2.

8.3. Приклади до Розділу 8

Приклад 1. Кластерний аналіз

Провести класифікацію $n = 6$ об'єктів, кожен з яких характеризується двома ознаками:

Номер	1	2	3	4	5	6	
Координати	5	6	5	10	11	10	1-ша ознака
	10	12	13	9	9	7	2-га ознака

Розв'язання.

На рисунку це виглядає так.

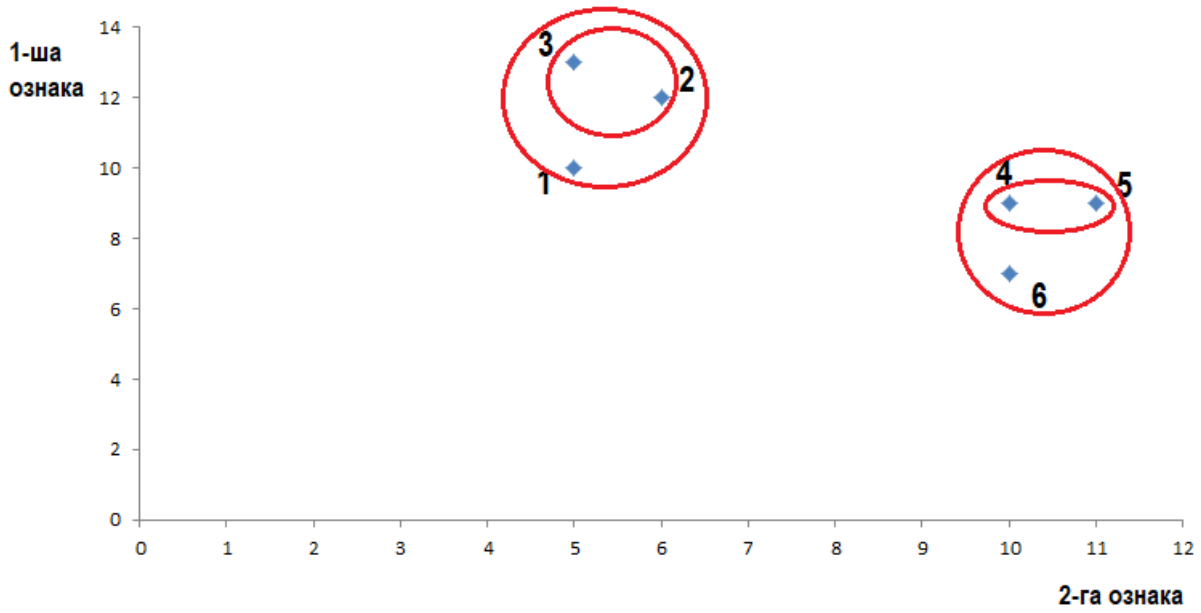


Рис. 8.10

За відстань ρ беремо звичайну евклідову метрику. Тоді, наприклад,

$$\rho_{12} = \sqrt{(5-6)^2 + (10-12)^2} = 2{,}24;$$

$$\rho_{13} = \sqrt{(5-5)^2 + (10-13)^2} = 3.$$

Звичайно

$$\rho_{11} = 0.$$

Аналогічно знаходимо відстані між шістьма об'єктами і отримаємо матрицю відстаней:

$$R_1 = \begin{pmatrix} 0 & 2{,}24 & 3 & 5{,}10 & 6{,}08 & 5{,}83 \\ 2{,}24 & 0 & 1{,}41 & 5 & 5{,}83 & 6{,}4 \\ 3 & 1{,}41 & 0 & 6{,}4 & 7{,}21 & 7{,}81 \\ 5{,}1 & 5 & 6{,}4 & 0 & \mathbf{1} & 2 \\ 6{,}08 & 5{,}83 & 7{,}21 & \mathbf{1} & 0 & 2{,}24 \\ 5{,}83 & 6{,}4 & 7{,}81 & 2 & 2{,}24 & 0 \end{pmatrix}.$$

З матриці видно, що 4-й і 5-й об'єкти найближчі: $\rho_{45} = 1$. Тому їх об'єднують в один кластер. Отримуємо

Номер кластера	1	2	3	4	5
Склад кластера	(1)	(2)	(3)	(4,5)	(6)

Відстань між кластерами визначаємо за метрикою $\rho_{\min}$ «найближчого сусіда». Наприклад

$$\rho_{1,(45)} = \min\{\rho(1,4), \rho(1,5)\} = 5{,}1.$$

Тоді матриця відстаней між кластерами набуде вигляду:

$$R_2 = \begin{pmatrix} 0 & 2{,}24 & 3 & 5{,}10 & 5{,}83 \\ 2{,}24 & 0 & 1{,}41 & 5 & 6{,}4 \\ 3 & 1{,}41 & 0 & 6{,}4 & 7{,}81 \\ 5{,}1 & 5 & 6{,}4 & 0 & 2 \\ 5{,}83 & 6{,}4 & 7{,}81 & 2 & 0 \end{pmatrix}.$$

Найменша відстань – між 2-м і 3-м об'єктами. Об'єднаємо їх в один кластер. Отримаємо 4 кластери

$$S_{(1)}, \quad S_{(2,3)}, \quad S_{(4,5)}, \quad S_{(6)}.$$

Знову порахуємо відстань між кластерами. Дістанемо

$$R_3 = \begin{pmatrix} 0 & 2{,}24 & 5{,}10 & 5{,}83 \\ 2{,}24 & 0 & 5 & 6{,}4 \\ 5{,}1 & 5 & 0 & 2 \\ 5{,}83 & 6{,}4 & 2 & 0 \end{pmatrix}.$$

Найменша відстань – між $S_{(4,5)}$ та $S_{(6)}$. Об'єднаємо їх. В результаті отримаємо 3 кластери:

$$S_{(1)}, \quad S_{(2,3)}, \quad S_{(4,5,6)}.$$

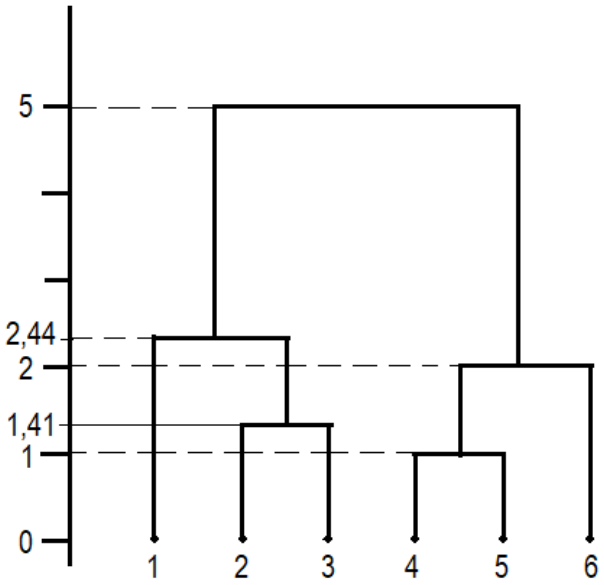
Матриця відстаней матиме вигляд:

$$R_4 = \begin{pmatrix} 0 & 2{,}24 & 5{,}10 \\ 2{,}24 & 0 & 5 \\ 5{,}1 & 5 & 0 \end{pmatrix}.$$

Об'єднаємо тепер кластери $S_{(1)}$ і $S_{(2,3)}$, відстань між якими – 2,24 – найменша.

В результаті отримаємо 2 кластери, відстань між якими, визначена за принципом «найближчого сусіда»: $\rho_{(1,2,3),(4,5,6)} = 5$.

Результати ієрархічної класифікації зображено у вигляді дендрограми



Найприроднішим тут є розбиття на два кластери $S_{(1,2,3)}$ та $S_{(4,5,6)}$.

Приклад 2. Дискримінантний аналіз

Діяльність кожного виробничого об'єднання галузі оцінювалась за такими трьома показниками:

- середньорічна вартість основних виробничих фондів (ОВФ);
- середньооблікова чисельність промислово-виробничого персоналу (ППП);
- балансовий прибуток.

У галузі виділено дві групи: передова, що складається з чотирьох об'єднань, та решта, що включає п'ять об'єднань.

Дані представлені у таблиці.

Група об'єднань/Показники	Вартість ОВФ	Чисельність ППП	Балансовий прибуток
Передова	224228	17115	22981
	151827	14904	21481
	147313	13627	28669
	152253	10545	10199
Інші	46757	4428	11124
	29033	5510	6091
	52134	4214	11842
	37050	5527	11873
	63979	4211	12860

Галузі передано об'єднання Z, у якого за прийнятими трьома показникам отримано такі результати: вартість ОВФ – 55451; чисельність ППП – 9592 тис. осіб; балансовий прибуток – 12840.

Визначити, чи можна віднести нове об'єднання до передової групи підприємств галузі.

Розв'язання.

1. Запишемо вихідні дані у вигляді матриць *X* і *Y* відповідно:

$$X = \begin{pmatrix} x_{11} & x_{13} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ \dots & \dots & \dots \\ x_{n_1 1} & x_{n_1 2} & x_{n_1 3} \end{pmatrix} = \begin{pmatrix} 224228 & 17115 & 22981 \\ 151827 & 14904 & 21481 \\ 147313 & 13627 & 28669 \\ 152253 & 10545 & 10199 \end{pmatrix}$$

та

$$Y = \begin{pmatrix} y_{11} & y_{13} & y_{13} \\ y_{21} & y_{22} & y_{23} \\ \dots & \dots & \dots \\ y_{n_2 1} & y_{n_2 2} & y_{n_2 3} \end{pmatrix} = \begin{pmatrix} 46751 & 4428 & 11124 \\ 29033 & 5510 & 6091 \\ 52134 & 4214 & 11842 \\ 37050 & 5527 & 11873 \\ 63979 & 4211 & 12860 \end{pmatrix}.$$

де $n = 4;\ \ m = 5;$.

Рядок матриці Z:

$$Z^T = (55451; 9592; 12840).$$

2. Знайдемо вектори середніх

$$\bar{X} = \begin{pmatrix} 168905 \\ 14048 \\ 20833 \end{pmatrix};\ \ \bar{Y} = \begin{pmatrix} 45791 \\ 4778 \\ 10758 \end{pmatrix}.$$

3. Визначимо оцінку коваріаційних матриць

$$S_x = \begin{pmatrix} 1023948995 & & \\ 55619765 & 5646869 & \\ 28912429 & 10273637 & 44879664{,}75 \end{pmatrix};$$

$$S_y = \begin{pmatrix} 145841159 & & \\ -6608403 & 371782 & \\ 22784783 & -902483{,}8 & 5750306 \end{pmatrix}.$$

4. Отримаємо незміщену оцінку сумарної коваріаційної матриці

$$\hat{S} = \frac{1}{4+5-2} \cdot \left(4S_x + 5S_y\right) = \begin{pmatrix} 689285968 & & \\ 27062435 & 3492341 & \\ 32796233 & 5226019 & 29752884 \end{pmatrix}.$$

5. Визначимо обернену матрицю до $\hat{S}$

$$\hat{S}^{-1} = \begin{pmatrix} 2{,}10 \cdot 10^{-9} & & \\ -1{,}74 \cdot 10^{-8} & 5{,}32 \cdot 10^{-7} & \\ 7{,}36 \cdot 10^{-10} & -7{,}43 \cdot 10^{-8} & 4{,}59 \cdot 10^{-8} \end{pmatrix}.$$

6. Знайдемо вектор оцінок коефіцієнтів дискримінації

$$a = \hat{S}^{-1} \cdot (\bar{X} - \bar{Y}) = \hat{S}^{-1} \begin{pmatrix} 123115 \\ 9270 \\ 10075 \end{pmatrix} = \begin{pmatrix} 1{,}05 \cdot 10^{-4} \\ 2{,}05 \cdot 10^{-3} \\ -1{,}36 \cdot 10^{-4} \end{pmatrix}.$$

7. Обчислимо оцінки дискримінантної функції

$$\widehat{U}_x = Xa = \begin{pmatrix} 55{,}38 \\ 43{,}48 \\ 39{,}41 \\ 36{,}14 \end{pmatrix}; \quad \widehat{U}_y = Ya = \begin{pmatrix} 12{,}44 \\ 13{,}49 \\ 12{,}47 \\ 13{,}57 \\ 13{,}57 \end{pmatrix}.$$

8. Визначимо середнє значення оцінок дискримінантної функції

$$\bar{u}_x = 43{,}6; \quad \bar{u}_y = 13{,}11.$$

9. Отримаємо константу

$$\hat{C} = \frac{1}{2} \cdot (43{,}6 + 13{,}11) = 28{,}36.$$

10. Визначимо можливість включення об'єднання Z у групу передових об'єднань. Так як матриця Z представлена одним рядком, то $\widehat{U}_y$ позначимо $\widehat{U}_z$.

$$\hat{u}_z = a_1 z_1 + a_2 z_2 + a_3 z_3 = \\ = 1{,}05 \cdot 10^{-4} \cdot 55451 + 2{,}05 \cdot 10^{-3} \cdot 9952 + (-1{,}36) \cdot 10^{-4} \cdot 12840 = \\ = 5{,}81.$$

Середнє значення дискримінантної функції $\hat{u}_z$ менше за константу $\hat{C}$

$$\hat{u}_z = 5{,}81 < \hat{C} = 28{,}34.$$

Отже, об'єднання Z з характеристиками Z^T не може бути віднесеним до групи передових підприємств галузі.

8.4. Питання для самоконтролю до Розділу 8

1. Якою є задача кластерного аналізу?
2. Опишіть задачі типізації.
3. В чому полягають евристичний, екстремальний та статистичний підходи до проблеми кластерного аналізу?
4. Що таке міра близькості та матриця близькості між елементами?
5. Опишіть евклідову та зважену евклідову метрики.
6. Опишіть відстань Хеммінга.
7. Що таке відстань «Найближчого сусіда»?
8. Що таке відстань «Найвіддаленішого сусіда»?
9. Що таке відстань між «центрами ваги» груп?
10. Що таке відстань, виміряна за принципом середнього зв'язку?
11. Що таке відстань, виміряна за принципом «степеневого середнього»?
12. Дайте означення суми внутрішньо-класових дисперсій.
13. Дайте означення суми попарних внутрішньо-класових відстаней між елементами.
14. Дайте означення узагальненої внутрішньо-класової дисперсії.
15. В чому полягає різниця між агломеративними і дивізимними кластер-процедурами?
16. В чому полягає метод класифікації з навчанням? Наведіть приклади.
17. Дайте означення випадкового вектора та його розподілу.
18.Опишіть суть ймовірнісного дискримінантного аналізу для дискретного і неперервного розподілів.
19. Дайте означення математичного сподівання та коваріаційної матриці випадкового вектора.
20. Опишіть статистику параметрів випадкового вектора.
21. Який випадковий вектор називається нормальним?
22. Опишіть алгоритм порівняння значень щільності нормальних векторів.
23. Опишіть процес розв'язування задачі дискримінації для двох класів і трьох показників.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ ТА ЛІТЕРАТУРИ

1. Giraud C. (2021). *Introduction to High-Dimensional Statistics*. Universite Paris Saclay. URL: http://surl.li/tuwmp
2. Montgomery, Douglas C., and Runger, George C. (2018). *Applied Statistics and Probability for Engineers*. John Wiley & Sons, 7th Edition. URL: http://surl.li/tuwnf
3. Боснюк В.Ф. (2020). *Математичні методи в психології: курс лекцій*. Мультимедійне навчальне видання – Харків : НУЦЗУ.
4. Жалдак М.І. , Кузьміна Н.М. , Михалін Г.О. (2009). *Теорія ймовірностей і математична статистика: Підручник для студентів фізико-математичних спеціальностей педагогічних університетів.* Вид. 2, перероб. і доп., Полтава : «Довкілля-К».
5. Жлуктенко В.І. , Наконечний С.І. , Савіна С.С. (2001). *Теорія ймовірностей і математична статистика*. Ч.II, Київ: КНЕУ. URL: http://surl.li/tuwno
6. Карташов М.В. (2008). *Ймовірність, процеси, статистика*, Київ: 'Київський університет'. URL: http://surl.li/tuwnx
7. Лупан І.В. , Авраменко О.В. , Акбаш К.С. (2015). *Комп'ютерні статистичні пакети: навч.-метод. посіб.* - 2-ге вид., Кіровоград : КОД.
8. Руденко В.М. (2012) *Математична статистика*. Навч. посібник, Київ : Центр учбової літератури.
9. Турчин В.М. (2014). *Теорія ймовірностей і математична статистика. Основні поняття, приклади, задачі: підручник для студентів вищих навчальних закладів*. Дніпропетровськ: IMA-прес. URL: http://surl.li/tuwod
10. Турчин В.Н. (2018). *Теория вероятностей и математическая статистика. Учебник для студентов высших учебных заведений.* Днепропетровск : ДЕУ.
11. Школьний Є.П. , Гончарова Л.Д. , Миротворська Н.К. (2020). *Методи обробки та аналізу гідрометеорологічної інформації (збірник задач і вправ): навчальний посібник*. Київ : Міносвіти і науки України.
12. Школьний Є.П. , Лоєва І.Д. , Гончарова Л.Д. (1999). *Обробка та аналіз гідрометеорологічної інформації*. Одеса : Одеський гідрометеорологічний інститут.

ДОДАТКИ. ТАБЛИЦІ СТАТИСТИЧНИХ РОЗПОДІЛІВ

ДОДАТОК 1

ТАБЛИЦЯ ЗНАЧЕНЬ ФУНКЦІЇ ЛАПЛАСА $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{z^2}{2}} dz$

x	*Φ(x)*	*x*	*Φ(x)*	*x*	*Φ(x)*	*x*	*Φ(x)*	*x*	*Φ(x)*
0,00	0,0000	0,52	0,1985	1,04	0,3508	1,56	0,4406	2,16	0,4846
0,01	0,0040	0,53	0,2019	1,05	0,3531	1,57	0,4418	2,18	0,4854
0,02	0,0080	0,54	0,2054	1,06	0,3554	1,58	0,4429	2,20	0,4861
0,03	0,0120	0,55	0,2088	1,07	0,3577	1,59	0,4441	2,22	0,4868
0,04	0,0160	0,56	0,2123	1,08	0,3599	1,60	0,4452	2,24	0,4875
0,05	0,0199	0,57	0,2157	1,09	0,3621	1,61	0,4463	2,26	0,4881
0,06	0,0239	0,58	0,2190	1,10	0,3643	1,62	0,4474	2,28	0,4887
0,07	0,0279	0,59	0,2224	1,11	0,3665	1,63	0,4484	2,30	0,4893
0,08	0,0319	0,60	0,2257	1,12	0,3686	1,64	0,4495	2,32	0,4898
0,09	0,0359	0,61	0,2291	1,13	0,3708	1,65	0,4505	2,34	0,4904
0,10	0,0398	0,62	0,2324	1,14	0,3729	1,66	0,4515	2,36	0,4909
0,11	0,0438	0,63	0,2357	1,15	0,3749	1,67	0,4525	2,38	0,4913
0,12	0,0478	0,64	0,2389	1,16	0,3770	1,68	0,4535	2,40	0,4918
0,13	0,0517	0,65	0,2422	1,17	0,3790	1,69	0,4545	2,42	0,4922
0,14	0,8557	0,66	0,2454	1,18	0,3810	1,70	0,4554	2,44	0,4927
0,15	0,0596	0,67	0,2486	1,19	0,3830	1,71	0,4564	2,46	0,4931
0,16	0,0636	0,68	0,2517	1,20	0,3849	1,72	0,4573	2,48	0,4934
0,17	0,0675	0,69	0,2549	1,21	0,3869	1,73	0,4582	2,50	0,4938
0,18	0,0714	0,70	0,2580	1,22	0,3883	1,74	0,4591	2,52	0,4941
0,19	0,0753	0,71	0,2611	1,23	0,3907	1,75	0,4599	2,54	0,4945
0,20	0,0793	0,72	0,2642	1,24	0,3925	1,76	0,4608	2,56	0,4948
0,21	0,0832	0,73	0,2673	1,25	0,3944	1,77	0,4616	2,58	0,4951
0,22	0,0871	0,74	0,2703	1,26	0,3962	1,78	0,4625	2,60	0,4953
0,23	0,0910	0,75	0,2734	1,27	0,3980	1,79	0,4633	2,62	0,4956
0,24	0,0948	0,76	0,2764	1,28	0,3997	1,80	0,4641	2,64	0,4959
0,25	0,0987	0,77	0,2794	1,29	0,4015	1,81	0,4649	2,66	0,4961
0,26	0,1026	0,78	0,2823	1,30	0,4032	1,82	0,4656	2,68	0,4963
0,27	0,1064	0,79	0,2852	1,31	0,4049	1,83	0,4664	2,70	0,4965
0,28	0,1103	0,80	0,2881	1,32	0,4066	1,84	0,4671	2,72	0,4967
0,29	0,1141	0,81	0,2910	1,33	0,4082	1,85	0,4678	2,74	0,4969
0,30	0,1179	0,82	0,2939	1,34	0,4099	1,86	0,4686	2,76	0,4971
0,31	0,1217	0,83	0,2967	1,35	0,4115	1,87	0,4693	2,78	0,4973
0,32	0,1255	0,84	0,2995	1,36	0,4131	1,88	0,4699	2,80	0,4974
0,33	0,1293	0,85	0,3023	1,37	0,4147	1,89	0,4706	2,82	0,4976
0,34	0,1331	0,86	0,3051	1,38	0,4162	1,90	0,4713	2,84	0,4977
0,35	0,1368	0,87	0,3078	1,39	0,4177	1,91	0,4719	2,86	0,4979
0,36	0,1406	0,88	0,3106	1,40	0,4192	1,92	0,4726	2,90	0,4981
0,37	0,1443	0,89	0,3133	1,41	0,4207	1,93	0,4732	2,92	0,4982
0,38	0,1480	0,90	0,3159	1,42	0,4222	1,94	0,4738	2,94	0,4984
0,39	0,1617	0,91	0,3186	1,43	0,4236	1,95	0,4744	2,96	0,4985
0,40	0,1564	0,92	0,3212	1,44	0,4251	1,96	0,4750	2,98	0,4986
0,41	0,1691	0,93	0,3238	1,45	0,4265	1,97	0,4756	3,00	0,4987
0,42	0,1628	0,94	0,3264	1,46	0,4279	1,98	0,4761	3,20	0,4993
0,43	0,1664	0,95	0,3289	1,47	0,4292	1,99	0,4767	3,40	0,4997
0,44	0,1700	0,96	0,3315	1,48	0,4306	2,00	0,4772	3,60	0,4998
0,45	0,1736	0,97	0,3340	1,49	0,4319	2,02	0,4783	3,80	0,4999
0,46	0,1772	0,98	0,3365	1,50	0,4332	2,04	0,4793	4,00	0,5000
0,47	0,1808	0,99	0,3389	1,51	0,4345	2,06	0,4803	5,00	0,5000
0,48	0,1844	1,00	0,3413	1,52	0,4357	2,08	0,4812	x > 5	0,5000
0,49	0,1879	1,01	0,3438	1,53	0,4370	2,10	0,4821		
0,50	0,1915	1,02	0,3461	1,54	0,4382	2,12	0,4830		
0,51	0,1950	1,03	0,3485	1,55	0,4394	2,14	0,4838		

ДОДАТОК 2

КВАНТИЛІ СТАНДАРТНОГО НОРМАЛЬНОГО РОЗПОДІЛУ

$\gamma = 1 - \alpha$	**0,90**	**0,95**	**0,975**	**0,98**	**0,99**	**0,995**	**0,999**	**0,9995**	**0,9999**
Одностороння	1,282	1,645	1,960	2,054	2,326	2,576	3,090	3,291	3,719
Двостороння	1,645	1,960	2,241	2,326	2,576	2,807	3,291	3,481	3,891

ДОДАТОК 3

ТАБЛИЦЯ ЗНАЧЕНЬ $t(\gamma; k = n - 1)$ РОЗПОДІЛУ СТЬЮДЕНТА, ЩО ЗАДОВОЛЬНЯЮТЬ РІВНІСТЬ $p(t) = 2\int_0^t f(x)dt = \gamma$

$k = n - 1$	γ												
	0,1	**0,2**	**0,3**	**0,4**	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**	**0,95**	**0,98**	**0,99**	**0,999**
1	0,158	0,326	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657	63,662
2	0,142	0,289	0,445	0,617	0,816	1,061	1,336	1,886	2,920	4,303	6,965	9,925	31,598
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	2,638	2,353	3,182	4,541	5,841	12,941
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,694	8,610
5	0,132	0,257	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,859
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,263	0,401	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,405
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,260	0,396	0,540	0,697	0,876	1,086	1,363	1,796	2,201	2,718	3,106	4,487
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,103	2,552	2,872	3,922
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,256	0,390	0,532	0,686	0,859	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,256	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,127	0,256	0,390	0,531	0,684	0,857	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,256	0,389	0,531	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,127	0,256	0,389	0,530	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,646

КРИТИЧНІ ТОЧКИ РОЗПОДІЛУ СТЬЮДЕНТА (t-РОЗПОДІЛУ) ДЛЯ $\alpha = 1 - \gamma$

Число ступенів свободи k	Рівень значущості, α						
	0,2	**0,1**	**0,05**	**0,02**	**0,01**	**0,002**	**0,001**
1	3,08	6,31	12,7	31,82	63,66	127,32	636,62
2	1,89	2,92	4,3	6,97	9,93	14,09	31,6
3	1,64	2,35	3,18	4,54	5,84	7,45	12,94
4	1,53	2,13	2,78	3,75	4,6	5,6	8,61
5	1,48	2,02	2,57	3,37	4,03	4,77	6,86
6	1,44	1,94	2,45	3,14	3,71	4,32	5,96
7	1,42	1,9	2,36	3	3,5	4,03	5,41
8	1,4	1,86	2,31	2,9	3,36	3,83	5,04
9	1,38	1,83	2,26	2,82	3,25	3,69	4,78
10	1,37	1,81	2,23	2,76	3,17	3,58	4,59
11	1,36	1,8	2,2	2,72	3,11	3,5	4,44
12	1,36	1,78	2,18	2,68	3,05	3,43	4,32
13	1,35	1,77	2,16	2,65	3,01	3,37	4,22
14	1,34	1,76	2,14	2,62	2,98	3,33	4,14
15	1,34	1,75	2,13	2,6	2,95	3,29	4,07
16	1,34	1,75	2,12	2,58	2,92	3,25	4,02
17	1,33	1,74	2,11	2,57	2,9	3,22	3,97
18	1,33	1,73	2,1	2,55	2,88	3,2	3,92
19	1,33	1,73	2,09	2,54	2,86	3,17	3,88
20	1,33	1,73	2,09	2,53	2,85	3,15	3,85
21	1,32	1,72	2,08	2,52	2,83	3,14	3,82
22	1,32	1,72	2,07	2,51	2,82	3,12	3,79
23	1,32	1,71	2,07	2,5	2,81	3,1	3,77
24	1,32	1,71	2,06	2,49	2,8	3,09	3,75
25	1,32	1,71	2,06	2,48	2,79	3,08	3,73
26	1,32	1,71	2,06	2,48	2,78	3,07	3,71
27	1,31	1,7	2,05	2,47	2,77	3,06	3,69
28	1,31	1,7	2,05	2,47	2,76	3,05	3,67
29	1,31	1,7	2,04	2,46	2,76	3,04	3,66
30	1,31	1,7	2,04	2,46	2,75	3,03	3,65
40	1,3	1,68	2,02	2,42	2,7	2,97	3,55
60	1,3	1,67	2	2,39	2,66	2,91	3,46
120	1,29	1,66	1,98	2,36	2,62	2,86	3,37
∞	1,28	1,64	1,96	2,33	2,58	2,81	3,29

ДОДАТОК 4

КРИТИЧНІ ТОЧКИ РОЗПОДІЛУ χ^2

Число ступенів свободи, k	Рівень значущості, α							
	0,99	**0,98**	**0,95**	**0,9**	**0,8**	**0,7**	**0,5**	**0,3**
1	0,00016	0,0006	0,0039	0,016	0,064	0,148	0,455	1,07
2	0,02	0,04	0,10	0,21	0,45	0,71	1,39	2,41
3	0,12	0,19	0,35	0,58	1,01	1,42	2,37	3,66
4	0,30	0,43	0,71	1,06	1,65	2,19	3,36	4,90
5	0,55	0,76	1,14	1,61	2,34	3,00	4,35	6,10
6	0,87	1,13	1,63	2,20	3,07	3,83	5,35	7,20
7	1,24	1,56	2,17	2,83	3,82	4,67	6,35	8,40
8	1,65	2,03	2,73	3,49	4,59	5,53	7,34	9,50
9	2,09	2,56	3,32	4,17	5,38	6,39	8,34	10,70
10	2,56	3,06	3,94	4,86	6,18	7,27	9,34	11,80
11	3,10	3,60	4,60	5,60	7,00	8,10	10,30	12,90
12	3,60	4,20	5,20	6,30	7,80	9,00	11,30	14,00
13	4,10	4,80	5,90	7,00	8,60	9,90	12,30	15,10
14	4,70	5,40	6,60	7,80	9,50	10,80	13,30	16,20
15	5,20	6,00	7,30	8,50	10,30	11,70	14,30	17,30
16	5,80	6,60	8,00	9,30	11,20	12,60	15,30	18,40
17	6,40	7,30	8,70	10,10	12,00	13,50	16,30	19,50
18	7,00	7,90	9,40	10,90	12,90	14,40	17,30	20,60
19	7,60	8,60	10,10	11,70	13,70	15,40	18,30	21,70
20	8,30	9,20	10,90	12,40	14,60	16,30	19,30	22,80
21	8,90	9,90	11,60	13,20	15,40	17,20	20,30	23,90
22	9,50	10,60	12,30	14,00	16,30	18,10	21,30	24,90
23	10,20	10,30	13,10	14,80	17,20	19,00	22,30	26,00
24	10,90	12,00	13,80	15,70	18,10	19,90	23,30	27,10
25	11,50	12,70	14,60	16,50	18,90	20,90	24,30	28,10
26	12,20	13,40	15,40	17,30	19,80	21,80	25,30	29,30
27	12,90	14,10	16,20	18,10	20,70	22,70	26,30	30,30
28	13,60	14,80	16,90	18,90	21,60	23,60	27,30	31,40
29	14,30	15,60	17,70	19,80	22,50	24,60	28,30	32,50
30	15,00	16,30	18,50	20,60	23,40	25,50	29,30	33,50

Продовження таблиці

Число ступенів свободи, k	Рівень значущості, α							
	0,2	**0,1**	**0,05**	**0,02**	**0,01**	**0,005**	**0,002**	**0,001**
1	1,64	2,70	3,80	5,40	6,60	7,90	9,50	10,83
2	3,22	4,60	6,00	7,80	9,20	11,60	12,40	13,80
3	4,64	6,30	7,80	9,80	11,30	12,80	14,60	16,30
4	6,00	7,80	9,50	11,70	13,30	14,90	16,90	18,50
5	7,30	9,20	11,10	13,40	15,10	16,30	18,90	20,50
6	8,60	10,60	12,60	15,00	16,80	18,60	20,70	22,50
7	9,80	12,00	14,10	16,60	18,50	20,30	22,60	24,30
8	11,00	13,40	15,50	18,20	20,10	21,90	24,30	26,10
9	12,20	14,70	16,90	19,70	21,70	23,60	26,10	27,90
10	13,40	16,00	18,30	21,20	23,20	25,20	27,70	29,60
11	14,60	17,30	19,70	22,60	24,70	26,80	29,40	31,30
12	15,80	18,50	21,00	24,10	26,20	28,30	31,00	32,90
13	17,00	19,80	22,40	25,50	27,70	29,80	32,50	34,50
14	18,20	21,10	23,70	26,90	29,10	31,00	34,00	36,10
15	19,30	22,30	25,00	28,30	30,60	32,50	35,50	37,70
16	20,50	23,50	26,30	29,60	32,00	34,00	37,00	39,20
17	21,60	24,80	27,60	31,00	33,40	35,50	38,50	40,80
18	22,80	26,00	28,90	32,30	34,80	37,00	40,00	42,30
19	23,90	27,30	30,10	33,70	36,20	38,50	41,50	43,80
20	25,00	28,40	31,40	35,00	37,60	40,00	43,00	45,30
21	26,20	29,60	32,70	36,30	38,90	41,50	44,50	46,80
22	27,30	30,80	33,90	38,70	40,30	42,50	46,00	48,30
23	28,40	32,00	35,20	39,00	41,60	44,00	47,50	49,70
24	29,60	33,20	36,40	40,30	43,00	45,50	48,50	51,20
25	30,70	34,40	37,70	41,60	44,30	47,00	50,00	52,60
26	31,80	35,60	38,90	42,90	45,60	48,00	51,50	54,10
27	32,90	36,70	40,10	44,10	47,00	49,50	53,00	55,50
28	34,00	37,90	41,30	45,40	48,30	51,00	54,50	56,90
29	35,10	39,10	42,60	46,70	49,60	52,50	56,00	58,30
30	36,30	40,30	43,80	48,00	50,90	54,00	57,50	59,70

ДОДАТОК 5

КРИТИЧНІ ТОЧКИ РОЗПОДІЛУ ФІШЕРА (F-РОЗПОДІЛУ)

Рівень значущості 0,05									
$k_2 \backslash k_1$	1	2	3	4	5	6	12	24	∞
1	164,4	199,5	215,7	224,6	230,2	234	244,9	249	254,3
2	18,5	9,2	19,2	19,3	19,3	19,3	19,4	19,5	19,5
3	10,1	9,6	9,3	9,1	9	8,9	8,7	8,6	8,5
4	7,7	6,9	6,6	6,4	6,3	6,2	5,9	5,8	5,6
5	6,6	5,8	5,4	5,2	5,1	5,0	4,7	4,5	4,4
6	6,0	5,1	4,8	4,5	4,4	4,3	4,0	3,8	3,7
7	5,6	4,7	4,4	4,1	4,0	3,9	3,6	3,4	3,2
8	5,3	4,5	4,1	3,8	3,7	3,6	3,3	3,1	2,9
9	5,1	4,3	3,9	3,6	3,5	3,4	3,1	2,9	2,7
10	5,0	4,1	3,7	3,5	3,3	3,2	2,9	2,7	2,5
11	4,8	4,0	3,6	3,4	3,2	3,1	2,8	2,6	2,4
12	4,8	3,9	3,5	3,3	3,1	3,0	2,7	2,5	2,3
13	4,7	3,8	3,4	3,2	3,0	2,9	2,6	2,4	2,2
14	4,6	3,7	3,3	3,1	3,0	2,9	2,5	2,3	2,1
15	4,5	3,7	3,3	3,1	2,9	2,8	2,5	2,3	2,1
16	4,5	3,6	3,2	3,0	2,9	2,7	2,4	2,2	2,0
17	4,5	3,6	3,2	3,0	2,8	2,7	2,4	2,2	2,0
18	4,4	3,6	3,2	2,9	2,8	2,7	2,3	2,1	1,9
19	4,4	3,5	3,1	2,9	2,7	2,6	2,3	2,1	1,8
20	4,4	3,5	3,1	2,9	2,7	2,6	2,3	2,1	1,8
22	4,3	3,4	3,1	2,8	2,7	2,6	2,2	2,0	1,8
24	4,3	3,4	3,0	2,8	2,6	2,5	2,2	2,0	1,7
26	4,2	3,4	3,0	2,7	2,6	2,4	2,1	1,9	1,7
28	4,2	3,3	2,9	2,7	2,6	2,4	2,1	1,9	1,6
30	4,2	3,3	2,9	2,7	2,5	2,4	2,1	1,9	1,6
40	4,1	3,2	2,9	2,6	2,5	2,3	2,0	1,8	1,5
60	4,0	3,2	2,8	2,5	2,4	2,3	1,9	1,7	1,4
120	3,9	3,1	2,7	2,5	2,3	2,2	1,8	1,6	1,3
∞	3,8	3,0	2,6	2,4	2,2	2,1	1,8	1,5	1,0

Продовження таблиці

Рівень значущості 0,01										
$k_2 \backslash k_1$	1	2	3	4	5	6	8	12	24	∞
1	4052	4999	5403	5625	5764	5859	5981	6106	6234	6366
2	98,5	99,0	99,2	99,3	99,3	99,4	99,3	99,4	99,5	99,5
3	34,1	30,8	29,5	28,7	28,2	27,9	27,5	27,1	26,6	26,1
4	21,2	18,0	16,7	16,0	15,5	15,2	14,8	14,4	13,9	13,5
5	16,3	13,3	12,1	11,4	11,0	10,7	10,3	9,9	9,5	9,0
6	13,7	10,9	9,8	9,2	8,8	8,5	8,1	7,7	7,3	6,9
7	12,3	9,6	8,5	7,9	7,5	7,2	6,8	6,5	6,1	5,7
8	11,3	8,7	7,6	7,0	6,6	6,4	6,0	5,7	5,3	4,9
9	10,6	8,0	7,0	6,4	6,1	5,8	5,5	5,1	4,7	4,3
10	10,0	7,6	6,6	6,0	5,6	5,4	5,1	4,7	4,3	3,9
11	9,7	7,2	6,2	5,7	5,3	5,1	4,7	4,4	4,0	3,6
12	9,3	6,9	6,0	5,4	5,1	4,8	4,5	4,2	3,8	3,4
13	9,1	6,7	5,7	5,2	4,9	4,6	4,3	4,0	3,6	3,2
14	8,9	6,5	5,6	5,0	4,7	4,5	4,1	3,8	3,4	3,0
15	8,7	6,4	5,4	4,9	4,6	4,3	4,0	3,7	3,3	2,9
16	8,5	6,2	5,3	4,8	4,4	4,2	3,9	3,6	3,2	2,8
17	8,4	6,1	5,2	4,7	4,3	4,1	3,8	3,5	3,1	2,7
18	8,3	6,0	5,1	4,6	4,3	4,0	3,7	3,4	3,0	2,6
19	8,2	5,9	5,0	4,5	4,2	3,9	3,6	3,3	2,9	2,4
20	8,1	5,9	4,9	4,4	4,1	3,9	3,6	3,2	2,9	2,4
22	7,9	5,7	4,8	4,3	4,0	3,8	3,5	3,1	2,8	2,3
24	7,8	5,6	4,7	4,2	3,9	3,7	3,3	3,0	2,7	2,2
26	7,7	5,5	4,6	4,1	3,8	3,6	3,3	3,0	2,6	2,1
28	7,6	5,5	4,6	4,1	3,8	3,5	3,2	2,9	2,5	2,1
30	7,6	5,4	4,5	4,0	3,7	3,5	3,2	2,8	2,5	2,0
40	7,3	5,2	4,3	3,8	3,5	3,3	3,0	2,7	2,3	1,8
60	7,1	5,0	4,1	3,7	3,3	3,1	2,8	2,5	2,1	1,6
120	6,9	4,8	4,0	3,5	3,2	3,0	2,7	2,3	2,0	1,4
∞	6,6	4,6	3,8	3,3	3,0	2,8	2,5	2,2	1,8	1,0

Продовження таблиці

Рівень значущості 0,001										
$k_2 \backslash k_1$	1	2	3	4	5	6	8	12	24	∞
1	Змінюється від 400 000 до 600 000									
2	998	999	999	999	999	999	999	999	999	999
3	167	148	141	137	135	133	131	128	126	123
4	74,1	61,3	56,2	53,4	51,7	50,5	49	47,4	45,8	44,1
5	47,0	36,6	33,2	31,1	29,8	28,8	27,6	26,4	25,1	23,8
6	35,5	27,0	23,7	21,9	20,8	20,0	19,0	18,0	16,9	15,8
7	29,2	21,7	18,8	17,2	16,2	15,5	14,6	13,7	12,7	11,7
8	25,4	18,5	15,8	14,4	13,5	12,9	12,0	11,2	10,3	9,3
9	22,9	16,4	13,9	12,6	11,7	11,1	10,4	9,6	8,7	7,8
10	21,0	14,9	12,6	11,3	10,5	9,9	9,2	8,5	7,6	6,8
11	19,7	13,8	11 ,6	10,4	9,6	9,1	8,3	7,6	6,9	6,0
12	18,6	13,0	10,8	9,6	8,9	8,4	7,7	7,0	6,3	5,4
13	17,8	12,3	10,2	9,1	8,4	7,9	7,2	6,5	5,8	5,0
14	17,1	11,8	9,7	8,6	7,9	7,4	6,8	6,1	5,4	4,6
15	16,6	11,3	9,3	8,3	7,6	7,1	6,5	5,8	5,1	4,3
16	16,1	11,0	9,0	7,9	7,3	6,8	6,2	5,6	4,9	4,1
17	15,7	10,7	8,7	7,7	7,0	6,6	6,0	5,3	4,6	3,9
18	15,4	10,4	8,5	7,5	6,8	6,4	5,8	5,1	4,5	3,7
19	15,1	10,2	8,3	7,3	6,6	6,2	5,6	5,0	4,3	3,5
20	14,8	10,0	8,1	7,1	6,5	6,0	5,4	4,8	4,2	3,4
22	14,4	9,6	7,8	6,8	6,2	5,8	5,2	4,6	3,9	3,2
24	14,0	9,3	7,6	6,6	6,0	5,6	5,0	4,4	3,7	3,0
26	13,7	9,1	7,4	6,4	5,8	5,4	4,8	4,2	3,6	2,8
28	13,5	8,9	7,2	6,3	5,7	5,2	4,7	4,1	3,5	2,7
30	13,3	8,8	7,1	6,1	5,5	5,1	4,6	4,0	3,4	2,6
40	12,6	8,2	6,6	5,7	5,1	4,7	4,2	3,6	3,0	2,2
60	12,0	7,8	6,2	5,3	4,8	4,4	3,9	3,3	2,7	1,9
120	11,4	7,3	5,8	5,0	4,4	4,0	3,5	3,0	2,4	1,6
∞	10,8	6,9	5,4	4,6	4,1	3,7	3,3	2,7	2,1	1,0

ДОДАТОК 6

КРИТИЧНІ ЗНАЧЕННЯ КОЕФІЦІЄНТУ РАНГОВОЇ КОРЕЛЯЦІЇ СПІРМЕНА

n	α		n	α		n	α	
	0,05	**0,01**		**0,05**	**0,01**		**0,05**	**0,01**
7	0,745	0,893	15	0,518	0,654	23	0,415	0,531
8	0,690	0,857	16	0,500	0,632	24	0,406	0,520
9	0,663	0,817	17	0,485	0,615	25	0,398	0,510
10	0,636	0,782	18	0,472	0,598	26	0,389	0,500
11	0,609	0,754	19	0,458	0,582	27	0,383	0,491
12	0,580	0,727	20	0,445	0,568	28	0,375	0,483
13	0,555	0,698	21	0,435	0,555	29	0,368	0,474
14	0,534	0,675	22	0,424	0,543	30	0,362	0,466

ДОДАТОК 7

КРИТИЧНІ ТОЧКИ СТАТИСТИКИ КОЛМОГОРОВА $\mathbf{D_{nm}}$

Об'єм вибірки, n	Рівень значущості α			
	0,10	**0,05**	**0,02**	**0,01**
1	0,95	0,98	0,99	0,995
2	0,78	0,84	0,90	0,93
3	0,64	0,71	0,78	0,83
4	0,57	0,62	0,69	0,73
5	0,51	0,56	0,62	0,67
6	0,47	0,52	0,58	0,62
7	0,44	0,48	0,54	0,58
8	0,41	0,45	0,51	0,54
9	0,39	0,43	0,48	0,51
10	0,37	0,41	0,46	0,49
11	0,35	0,39	0,44	0,47
12	0,34	0,38	0,42	0,45
13	0,33	0,36	0,40	0,43
14	0,31	0,35	0,39	0,42
15	0,30	0,34	0,38	0,40
16	0,29	0,33	0,37	0,39
17	0,29	0,32	0,36	0,38
18	0,28	0,31	0,34	0,37
19	0,27	0,30	0,34	0,36
20	0,26	0,29	0,33	0,35

КРИТИЧНІ ЗНАЧЕННЯ СТАТИСТИКИ КОЛМОГОРОВА-СМИРНОВА

$$\mathbf{P}\{\boldsymbol{K} > \boldsymbol{k}_{\text{КР}}\} = \boldsymbol{\alpha}$$

$\boldsymbol{\alpha}$	**0,1**	**0,05**	**0,01**
$\boldsymbol{k}_{\text{кр}}$	1,22	1,36	1,63

ДОДАТОК 8

КРИТИЧНІ ЗНАЧЕННЯ $U_1(\gamma)$ ТА $U_1(\gamma)$ КРИТЕРІЯ МАННА УІТНІ

		γ						γ						γ			
		0,9		0,95				0,9		0,95				0,9		0,95	
n	m	U_1	U_2	U_1	U_2	n	m	U_1	U_2	U_1	U_2	n	m	U_1	U_2	U_1	U_2
4	4	1	15	0	16	22	22	171	313	158	326	16	16	83	173	75	181
	5	2	18	1	19		23	179	327	166	340		17	89	183	81	191
	6	3	21	2	22		24	188	340	174	354		18	95	193	86	202
	7	4	24	3	25		25	197	353	182	368		19	101	203	92	212
	8	5	27	4	28		26	205	367	191	381		20	107	213	98	222
	9	6	30	4	32		27	214	380	199	395		21	113	223	103	233
	10	7	33	5	35		28	223	393	207	409		22	119	233	109	243
5	5	4	21	2	23		29	231	407	215	423		23	125	243	115	253
	6	5	25	3	27		30	240	420	223	437		24	131	253	120	263
	7	6	29	5	30	24	24	207	369	192	384	18	18	109	215	99	225
	8	8	32	6	34		25	217	383	201	399		19	116	226	106	236
	9	9	36	7	38		26	226	398	210	414		20	123	237	112	248
	10	11	39	8	42		27	236	412	219	429		21	130	248	119	259
6	6	7	29	5	31		28	245	427	228	444		22	136	260	125	271
	7	8	34	6	36		29	255	441	238	458		23	143	271	132	282
	8	10	38	8	40		30	264	456	247	473		24	150	282	138	294
	9	12	42	10	44		31	274	470	256	488		25	157	293	145	305
	10	14	46	11	49		32	284	484	265	503		26	164	304	151	317
	11	16	50	13	53	26	26	247	429	230	446	20	20	138	262	127	273
	12	17	55	14	58		27	257	445	240	462		21	146	274	134	286
7	7	11	38	8	41		28	268	460	250	478		22	154	286	141	299
	8	13	43	10	46		29	278	476	260	494		23	161	299	149	311
	9	15	48	12	51		30	289	491	270	510	28	32	347	549	315	571
	10	17	53	14	56		31	299	507	280	526		33	359	565	326	598
	11	19	58	16	61		32	310	522	290	542		34	370	582	337	615
	12	21	63	18	66	28	28	291	493	272	512	30	30	338	562	317	583
	13	24	67	20	71		29	302	510	282	530		31	350	580	328	602
	14	26	72	22	76		30	313	527	293	547		32	362	598	340	620
8	8	15	49	13	51		31	325	543	304	564		33	374	616	352	638
	9	18	57	15	54	10	13	37	93	33	97		34	387	633	364	656
	10	20	60	17	63		14	41	99	36	104		35	399	651	375	675
	11	23	65	19	69		15	44	106	39	111		36	411	669	387	693
	12	26	70	22	74		16	48	112	42	118	32	32	388	636	365	659
	13	28	76	24	80		17	51	119	45	125		33	402	654	378	678
	14	31	81	26	86		18	55	125	48	132		34	425	673	391	697
	15	33	87	29	91		19	58	132	52	138		35	428	692	403	717
	16	36	92	31	97		20	62	138	55	145		36	441	711	416	736
9	9	21	60	17	64	12	12	42	102	37	107		37	454	730	428	756
	10	24	66	20	70		13	47	109	41	117		38	467	749	441	775
	11	27	72	23	76		14	51	117	45	123	34	34	443	713	418	738
	12	30	78	26	82		15	55	125	49	131		35	457	733	431	759
	13	33	84	28	89		16	60	132	53	139		36	471	753	445	779
	14	36	90	31	95		17	64	140	57	147		37	485	773	458	800
	15	39	96	34	101		18	68	148	61	155		38	499	793	472	820
	16	42	102	37	107		19	72	156	65	163		39	513	813	485	841
10	10	27	73	23	77		20	77	163	69	171		40	527	833	499	861
	11	31	79	26	84	14	14	61	135	55	141	36	36	471	753	445	779
	12	34	86	29	91		15	66	144	59	151		37	486	774	459	801
20	24	169	311	156	324		16	71	153	64	160		38	500	795	473	822
	25	177	323	163	337		17	77	161	69	169		39	515	815	487	843
	26	185	335	171	349		18	82	170	74	178		40	529	836	501	864
	27	192	348	178	362		19	87	179	78	188	38	38	563	881	533	911
	28	200	360	186	374		20	92	188	83	197		39	578	904	548	934
							21	97	197	88	206		40	594	926	563	957
							22	102	206	93	215	40	40	628	972	596	1004

ДОДАТОК 9

КРИТИЧНІ ЗНАЧЕННЯ СТАТИСТИКИ Т ЗНАКОВОГО РАНГОВОГО КРИТЕРІЮ ВІЛКОКСОНА З НАДІЙНІСТЮ γ ($\gamma = 1 - \alpha$)

n	Рівень надійності γ			
	0,9		0,95	
	T_1	T_2	T_1	T_2
6	2	19	0	21
7	4	24	3	25
8	6	30	4	32
9	9	36	6	39
10	11	44	9	46
11	14	51	11	55
12	18	60	14	64
13	22	69	18	73
14	26	79	22	83
15	31	89	26	94
16	36	100	30	106
17	42	111	35	118
18	48	123	41	130
19	54	136	47	143
20	61	149	53	157
22	76	177	66	187
24	92	208	82	218
26	111	240	99	252
28	131	275	117	289
30	152	313	138	327
32	176	353	160	368

ДОДАТОК 10

КРИТИЧНІ ЗНАЧЕННЯ КРИТЕРІЯ КРАСКАЛА-УОЛЛІСА (α-РІВЕНЬ ЗНАЧУЩОСТІ)

n_1	n_2	n_3	α		n_1	n_2	n_3	α		n_1	n_2	n_3	α	
			0,1	0,05				0,1	0,05				0,1	0,05
$k = 3$														
2	2	2	4,571		5	4	2	4,541	5,273	6	5	2	4,596	5,330
3	2	2	4,500	4,714	5	4	3	4,549	5,656	6	5	3	4,535	3,602
3	3	2	4,556	5,361	5	4	4	4,668	5,657	6	5	4	4,522	5,661
3	3	3	4,622	5,600	5	5	2	4,623	5,338	6	5	5	4,547	5,729
4	2	2	4,458	5,333	5	5	3	4,545	5,705	6	6	2	4,438	5,410
4	3	2	4,511	5,444	5	5	4	4,523	5,666	6	6	3	4,558	5,625
4	3	3	4,709	5,791	5	5	5	4,560	5,780	6	6	4	4,548	5,724
4	4	2	4,555	5,455	6	2	2	4,545	5,345	6	6	5	4,542	5,765
4	4	3	4,545	5,598	6	3	2	4,682	5,348	6	6	6	4,643	5,801
4	4	4	4,654	5,692	6	3	3	4,590	5,615	7	7	7	4,594	5,819
5	2	2	4,373	5,160	6	4	2	4,494	5,340	8	8	8	4,595	5,805
5	3	2	4,651	5,251	6	4	3	4,604	5,610					
5	3	3	4,533	5,648	6	4	4	4,595	5,681					

$k = 4$											
n_1	n_2	n_3	n_4	α		n_1	n_2	n_3	n_4	α	
				0,1	0,05					0,1	0,05
2	2	2	2	5,667	6,167	4	3	2	2	5,750	6,621
3	2	2	2	5,664	6,333	4	3	3	2	5,872	6,795
3	3	2	2	5,745	6,527	4	3	3	3	6,016	6,984
3	3	3	2	5,879	6,727	4	4	2	2	5,808	6,731
3	3	3	3	6,026	7,000	4	4	2	2	5,901	6,874
4	2	2	2	5,755	6,545						

$k = 5$														
n_1	n_2	n_3	n_4	n_5	α		n_1	n_2	n_3	n_4	n_5	α	
					0,1	0,05						0,1	0,05
4	4	4	2		5,914	6,957	3	3	2	2	2	7,026	7,910
4	4	4	3		6,042	7,142	3	3	3	2	2	7,121	8,044
4	4	4	4		6,088	7,235	3	3	3	3	2	7,210	8,200
2	2	2	2	2	6,982	7,418	3	3	3	3	3	7,333	8,333
3	2	2	2	2	6,955	7,682							

ДОДАТОК 11

КРИТИЧНІ ЗНАЧЕННЯ $S_{\alpha}(n, k)$ КРИТЕРІЮ ФРІДМЕНА

n	Надійність $\gamma = 0,95$			Надійність $\gamma = 0,95$		
	k			k		
	3	4	5	3	4	5
3	5,82	7,00	8,30	5,42	6,20	7,47
4	6,31	7,20	8,80	5,10	6,00	7,58
5	6,10	7,32	8,96	5,21	6,12	7,61
6	6,33	7,40		4,83	6,20	
7	6,00	7,63		4,71	6,26	
8	6,25	7,50		5,00	6,30	
9	6,00			4,67		
10	6,10			4,90		
11	6,09			4,91		
12	6,08			4,67		
13	6,00			4,77		

ДОДАТОК 12

ГРАНИЦІ КРИТИЧНИХ ОБЛАСТЕЙ КРИТЕРІЯ ЗНАКІВ

n	Значення α							n	Значення α					
	0,025		0,010		0,005				0,025		0,010		0,005	
5	0	5	0	5	0	5		53	19	34	18	35	17	36
6	1	5	0	6	0	6		54	20	34	19	35	18	36
7	1	6	1	6	0	7		55	20	35	19	36	18	37
8	1	7	1	7	1	7		56	21	35	19	37	18	38
9	2	7	1	8	1	8		57	21	36	20	37	19	38
10	2	8	1	9	1	9		58	22	36	20	38	19	39
11	2	9	2	9	1	10		59	22	37	21	38	20	39
12	3	9	2	10	2	10		60	22	38	21	39	20	40
13	3	10	2	11	2	11		61	23	38	21	40	21	40
14	3	11	3	11	2	12		62	23	39	22	40	21	41
15	4	11	3	12	3	12		63	24	39	22	41	21	42
16	4	12	3	13	3	13		64	24	40	23	41	22	42
17	5	12	4	13	3	14		65	25	40	23	42	22	43
18	5	13	4	14	4	14		66	25	41	24	42	23	43
19	5	14	5	14	4	15		67	26	41	24	43	23	44
20	6	14	5	15	4	16		68	26	42	24	44	23	45
21	6	15	5	16	5	16		69	26	43	25	44	24	45
22	6	16	6	16	5	17		70	27	43	25	45	24	46
23	7	16	6	17	5	18		71	27	44	26	45	25	46
24	7	17	6	18	6	18		72	28	44	26	46	25	47
25	8	17	7	18	6	19		73	28	45	27	46	26	47
26	8	18	7	19	7	19		74	29	45	27	47	26	48
27	8	19	8	19	7	20		75	29	46	27	48	26	49
28	9	19	8	20	7	21		76	29	47	28	48	27	49
29	9	20	8	21	8	21		77	30	47	28	49	27	50
30	10	20	9	21	8	22		78	30	48	29	49	28	50
31	10	21	9	22	8	23		79	31	48	29	50	28	51
32	10	22	9	23	9	23		80	31	49	30	50	29	51
33	11	22	10	23	9	24		81	32	49	30	51	29	52
34	11	23	10	24	10	24		82	32	50	31	51	29	53
35	12	23	11	24	10	25		83	33	50	31	52	30	53
36	12	24	11	25	10	26		84	33	51	31	53	30	54
37	13	24	11	26	11	26		85	33	52	32	53	31	54
38	13	25	12	26	11	27		86	34	52	32	54	31	55
39	13	26	12	27	12	27		87	34	53	33	54	32	55
40	14	26	13	27	12	28		88	35	53	33	55	32	56
41	14	27	13	28	12	29		89	35	54	34	55	32	57
42	15	27	14	28	13	29		90	36	54	34	56	33	57
43	15	28	14	29	13	30		91	36	55	34	57	33	58
44	16	28	14	30	14	30		92	37	55	35	57	34	58
45	16	29	15	30	14	31		93	37	56	35	58	34	59
46	16	30	15	31	14	32		94	38	56	36	58	35	59
47	17	30	16	32	15	32		95	38	57	36	59	35	60
48	17	31	16	32	15	33		96	38	58	37	59	35	61
49	18	31	16	33	16	33		97	39	58	37	60	36	61
50	18	32	17	33	16	34		98	39	59	38	60	36	62
51	19	32	17	34	16	35		99	40	59	38	61	37	62
52	19	33	18	34	17	35		100	40	60	38	62	37	63

ДОДАТОК 13

КРИТИЧНІ ЗНАЧЕННЯ $S_1(\gamma)$ ТА $S_2(\gamma)$ СТАТИСТИКА АНСАРІ-БРЕДЛІ

m	n	γ				m	n	γ			
		0,90		0,95				0,90		0,95	
		S_1	S_2	S_1	S_2			S_1	S_2	S_1	S_2
2	8	2	10	2	10	5	5	10	20	10	20
	9	2	11	2	11		6	1	21	10	23
	10	2	12	2	12		7	11	24	11	24
	11	2	13	2	13		8	12	26	11	26
	12	2	14	2	14		9	13	27	12	28
	13	3	14	2	15		10	14	29	12	30
	14	3	15	2	16		11	14	31	13	32
	15	3	16	2	17		12	15	33	14	34
	16	3	17	2	17		13	16	34	14	36
	17	3	18	2	19		14	16	36	15	38
	18	3	19	2	19		15	17	38	15	40
3	6	4	13	4	13	6	6	15	27	14	28
	7	5	13	4	14		7	16	29	15	30
	8	5	15	4	16		8	17	31	16	32
	9	5	16	4	17		9	18	34	16	35
	10	5	17	5	18		10	18	36	17	37
	11	6	18	5	19		11	19	38	18	40
	12	6	20	5	21		12	20	40	19	41
	13	6	21	5	22		13	21	42	19	41
	14	7	22	6	23		14	22	44	20	46
	15	7	23	6	24	7	7	21	35	19	37
	16	7	24	6	25		8	22	38	20	39
	17	8	25	6	26		9	23	40	21	42
4	5	7	14	6	16		10	24	43	22	44
	6	7	17	7	17		11	25	45	23	47
	7	8	19	7	19	8	8	26	45	26	46
	8	8	20	7	21		9	29	48	27	49
	9	9	21	8	22		10	30	50	28	52
	10	9	23	8	24		11	31	53	29	55
	11	10	24	9	26		12	32	56	30	58
	12	10	26	9	27	9	9	35	55	33	57
	13	11	27	9	29		10	36	58	34	58
	14	11	29	10	30		11	38	61	36	63
	15	12	30	10	32	10	10	43	67	41	69
	16	12	32	11	33						

ДОДАТОК 14

КРИТИЧНІ ЗНАЧЕННЯ $\mathbf{L}_{\gamma}$ І $\mathbf{D}_{\gamma}$ ДЛЯ $\mathbf{k} = \mathbf{3}$

n_1	n_2	n_3	Надійність γ			
			$\mathbf{0,99}$		$\mathbf{0,95}$	
			L	D	L	D
2	2	2			5,33	10,00
2	2	3	6,48	11,43	5,39	8,00
2	2	4	7,33,	13,33	6,10	5,80
2	3	3	7,40	10,00	5,85	5,59
2	3	4	7,14	9,95	5,80	4,54
2	4	4	7,47	8,30	5,55	5,48
3	3	3	7,21	9,38	6,22	4,80
3	4	4	7,82	6,75	6,34	4,18
4	4	4	8,20	7,15	6,15	4,60

Навчальне видання

Плічко Анатолій Миколайович
Акбаш Катерина Сергіївна
Луньова Марія Валентинівна

Математична статистика

навчальний посібник

Редактори: Плічко А.М., Акбаш К.С., Луньова М.В.

Формат 60х84 1/16. Ум. друк. Арк. 12,91. Тираж 98. Зам. 197.

Виготовлювач СПД ФО Лисенко Я.С.
25029, м.Кропивницький, вул.Театральна, 31, кв.4
Свідоцтво суб'єкта видавничої справи ДК № 8096 від 21.03.2024