Уміщено вказівки до виконання 12 практичних робіт із дисципліни «Математична статистика та обробка геологічної інформації». Зокрема, розглянуто задачі побудови варіаційних рядів та гістограм, оцінки основних статистичних характеристик, ідентифікації нормального розподілу, перевірки однорідності двох вибірок, кореляційного та регресійного аналізу. Пояснено базові теоретичні поняття, необхідні для виконання кожної роботи, наведено приклади виконання робіт із коментарями, ілюстраціями, аналізом результатів.

Для студентів ДНУ, які навчаються за спеціальністю «Науки про Землю», а також студентів інших спеціальностей, які вивчають математичну статистику і статистичну обробку даних.

Темплан 2019, поз. 5

Методичні вказівки
до виконання практичних робіт
із дисципліни
«Математична статистика
та обробка геологічної інформації»

Укладачі: канд. техн. наук, доц. О.М. Мацуга
канд. техн. наук, доц. Т.Г. Ємел'яненко

Редактор О.В. Бец
Техредактор Т.І. Севост'янова
Коректор О.В. Бец

Підписано до друку 28.01.2019. Формат 60x84/16. Папір друкарський. Друк плоский. Ум. друк. арк. 3,7. Ум. фарбовідб. 3,7. Обл.- вид. арк. 2,2. Тираж 20 пр. Зам. №

РВВ ДНУ, просп. Гагаріна, 72, м. Дніпро, 49010.
ПП «Ліра ЛТД», вул. Наукова, 5, м. Дніпро, 49107.
Свідоцтво про внесення до Державного реєстру
серія ДК № 6042 від 26.02.2018 р.

Зміст

Вступ 4

Вказівки до виконання практичних робіт 1 – 3 на тему «Первинний статистичний аналіз» 5
Практична робота 1. Побудова варіаційного ряду і гістограми 5
Практична робота 2. Оцінювання статистичних характеристик 10
Практична робота 3. Ідентифікація нормального розподілу 20
Контрольні запитання 24

Вказівки до виконання практичних робіт 4 – 7 на тему «Перевірка однорідності двох вибірок» 25
Практична робота 4. Перевірка рівності середніх та дисперсій у випадку двох залежних вибірок 26
Практична робота 5. Перевірка рівності середніх та дисперсій у випадку двох незалежних вибірок 29
Практична робота 6. Застосування критерію знакових рангів Вілкоксона до двох залежних вибірок 31
Практична робота 7. Застосування критеріїв суми рангів Вілкоксона та Манна – Уїтні до двох незалежних вибірок 33
Контрольні запитання 36

Вказівки до виконання практичних робіт 8 – 12 на тему «Кореляційний та регресійний аналіз» 37
Практична робота 8. Аналіз кореляційного поля. Оцінювання коефіцієнта кореляції Пірсона 37
Практична робота 9. Оцінювання коефіцієнтів рангової кореляції Спірмена і Кендалла 41
Практична робота 10. Оцінювання параметрів лінійної регресії 47
Практична робота 11. Довірче оцінювання регресії та прогнозних значень 52
Практична робота 12. Перевірка значущості та адекватності регресії 56
Контрольні запитання 59

Список рекомендованої літератури 60

Додаток 61

Вступ

Математична статистика – це розділ математики, який набув поширення в усіх сферах діяльності, зокрема і в геології. Методи математичної статистики застосовують для систематизації, обробки й аналізу даних, щоб знайти в них певні закономірності, висунути і перевірити нові гіпотези, прийняти обґрунтовані рішення на їх основі та ін.

Статистичні дані являють собою відомості щодо об'єктів з досить великої сукупності (генеральної сукупності), які мають певні властивості. Статистичні дані ще називають вибіркою об'єктів з генеральної сукупності. Залежно від завдання об'єктами можуть бути мінерали, гірські породи, скам'янілості, свердловини, зразки ґрунту, підземних вод, які вивчають у процесі спостережень або експериментів.

Властивості, що описують указані об'єкти, називають ознаками, змінними або показниками. Їх вимірюють в різних шкалах. Звичайно застосовують такі типи шкал: номінальну, порядкову (ординальну), інтервальну, відносну (шкалу відношень). Номінальні показники визначають належність об'єктів до деяких класів, які не можна упорядкувати, тому такі показники застосовують у випадку здійснення якісної класифікації. Приклад показника в номінальній шкалі – назва шахти на території держави або колір гірської породи. Порядкові показники дозволяють упорядкувати об'єкти, але не дають змоги визначити, наскільки більший або менший один об'єкт від іншого. Прикладами таких показників можуть бути класи забруднюючих речовин за ступенем їх небезпеки (високонебезпечні, помірно небезпечні, малонебезпечні). Інтервальні показники дозволяють не тільки упорядковувати об'єкти, але й виражати і порівнювати відмінності між ними в числовій формі. Особливість показників, вимірюваних у цій шкалі, – наявність умовного нуля. Наприклад, значення температури в градусах за Цельсієм утворюють інтервальну шкалу. Відносні показники подібні до інтервальних, але вони дозволяють відбивати те, у скільки разів один об'єкт більший або менший за інший. Прикладами показників, вимірюваних у такій шкалі, є маса, довжина, вартість.

Одержати уявлення про базові поняття математичної статистики, опанувати методи математичної статистики та навчитися застосовувати їх в процесі обробки геологічних даних – головна мета вивчення дисципліни «Математична статистика та обробка геологічної інформації». Для допомоги студентам у її досягненні розроблено 12 практичних робіт, указівки до виконання яких наведено в пропонованому виданні.

Зокрема, вказівки в стислій, але доступній формі ознайомлюють читача з основами застосування в геології статистичних методів обробки одновимірних наборів даних, що містять інформацію про одну ознаку об'єкта, та двовимірних наборів, які дозволяють вивчати зв'язки між двома ознаками.

Вказівки до виконання практичних робіт 1 – 3 на тему «Первинний статистичний аналіз»

Первинний статистичний аналіз – це базовий складник статистичної обробки вибірки $\left\{x_i; i=\overline{1,N}\right\}$, здійснюваної з метою оцінити основні характеристики розподілу, з якого вилучено вибірку, та ідентифікувати цей розподіл.

Первинний статистичний аналіз включає нижченазвані етапи.

1. Побудова варіаційного ряду.
2. Розбиття варіаційного ряду на класи та побудова гістограми.
3. Побудова графіка емпіричної функції розподілу.
4. Оцінювання статистичних характеристик.
5. Пошук аномальних значень (на основі аналізу гістограми або статистичних характеристик) та їх вилучення з повторним виконанням етапів 1 – 4.
6. Ідентифікація розподілу за даними вибірки.

Практична робота 1
ПОБУДОВА ВАРІАЦІЙНОГО РЯДУ І ГІСТОГРАМИ

Нехай задано вибірку $\left\{x_i; i=\overline{1,N}\right\}$ обсягом N. Деякі елементи можуть повторюватися у вибірці. Позначимо кількість унікальних елементів через r.

Нехай елемент x_1 трапляється у вибірці n_1 разів, елемент x_2 – n_2 разів, …, x_r – n_r разів. Унікальні елементи вибірки $x_1, x_2, ..., x_r$ називають варіантами, а числа $n_1, n_2, ..., n_r$ – їх частотами. Якщо поділити частоту на обсяг вибірки, то можна одержати відносну частоту $p_i = n_i / N$, $i=\overline{1,N}$.

Послідовність варіант, записаних за зростанням, з присвоєними їм частотами та відносними частотами, називають **варіаційним рядом** (табл. 1).

Таблиця 1

Варіаційний ряд

Варіанта x	Частота n	Відносна частота p
x_1	n_1	p_1
x_2	n_2	p_2
…	…	…
x_r	n_r	p_r

Більш інформативний завжди не варіаційний ряд, а ряд, розбитий на класи, на основі якого будують дуже важливий для аналізу графік – гістограму.

Для **розбиття варіаційного ряду на класи** спочатку потрібно оцінити кількість класів M. Загалом величина M досить довільна, проте існує оптимальна кількість класів, яка залежить від обсягу N даних вибірки та типу їх закону розподілу. На практиці M можна оцінити одним із таких способів:

1) обрахунок M залежно від обсягу вибірки:

– за $N<100$ M обчислюють згідно з формулою

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

де $[\cdot]$ – ціла частина;

– у разі $N\geq 100$ застосовують формулу

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

2) визначення M за формулою

$$M\approx 1+3{,}32\lg N \quad \text{або} \quad M\approx 1+1{,}44\ln N.$$

Відтак розбиття ряду на класи виконують шляхом розбиття відрізка $[x_{\min};x_{\max}]$ на M однакових інтервалів, які називають класами. Для цього вдаються до нижченаведеного алгоритму.

1. Знаходять мінімальний $x_{\min}$ та максимальний $x_{\max}$ елементи вибірки.
2. Обчислюють ширину кожного класу:

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

3. Розраховують межі класів:

$$x'_i=x_{\min}+(i-1)h,\ i=\overline{1,M+1},$$

і в такий спосіб одержують класи:

1-й $[x'_1;x'_2)$

2-й $[x'_2;x'_3)$

…

M-й $[x'_M;x'_{M+1}]$

4. Для кожного класу обчислюють:

– частоту n_i – кількість елементів початкової вибірки, що потрапили в i-й клас, тобто в інтервал $[x'_i;x'_{i+1})$;

– відносну частоту $p_i=n_i/N$.

За результатами розбиття ряду на класи будують гістограму.

Гістограма – це стовпчастий графік, у якому за віссю абсцис відкладають межі класів, а за віссю ординат – відносні частоти (рис. 1).

Гістограма дозволяє: 1) побачити, як розподілені елементи вибірки за класами; 2) оцінити типове значення вибірки, встановити, у який спосіб дані згруповані навколо нього; 3) висунути припущення щодо закону розподілу даних, тобто ідентифікувати його (табл. 2); 4) виявити аномальні значення у вибірці: якщо з правого або лівого «хвоста» гістограми є відокремлений клас із малою відносною частотою, то, найімовірніше, елементи, які потрапили до такого класу, є аномальні.

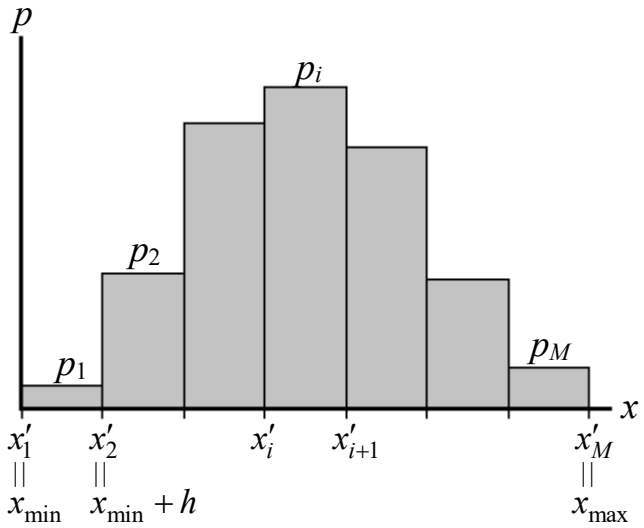
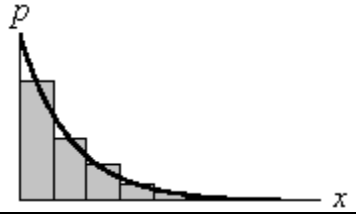
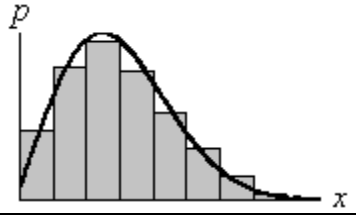
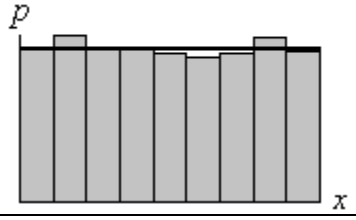
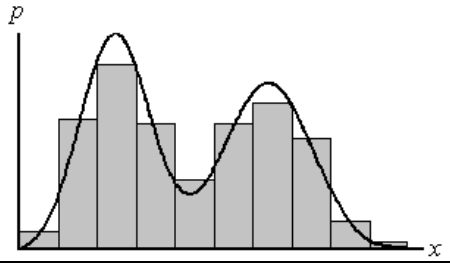


Рис. 1. Гістограма відносних частот

Таблиця 2

Види гістограм та розподіли, які за ними ідентифікують

Тип гістограми	Вигляд гістограми	Розподіли, які ідентифікують за такою гістограмою
Одномодальна симетрична		Нормальний
Унімодальна		Експоненціальний, Вейбулла
Одномодальна асиметрична (з лівою або правою асиметрією)		Логарифмічно нормальний, Вейбулла, Релея, екстремальний найбільшого значення
Рівномірна		Рівномірний
Багатомодальна (у цьому прикладі двомодальна)		Дані неоднорідні, взяті з різних генеральних сукупностей

Слід відзначити, що ідентифікувати розподіл за виглядом гістограми можна, оскільки гістограма – це оцінка функції щільності розподілу показника, зауважимо, що оцінка лише з точністю до константи h. Тому в разі одночасного відображення гістограми та графіка функції щільності потрібно звести їх до одного масштабу шляхом нормування або гістограми відносних частот (поділити відносні частоти на h), або графіка функції щільності (помножити функцію на h).

Приклад 1. Задано вибірку x з 27 елементів (значення кислотно-лужного показника проб води, узятих зі свердловин на території гірничо-збагачувального комбінату):

7 8,4 7,4 7,2 9,2 5 9,5 10,2 7,9 8,8 8,7 8,2 8,8 8,7
3,8 8,2 8,5 8,1 7,1 8,5 8,9 8,8 8,3 7,7 8,4 7,6 7,6

Побудуємо за цією вибіркою варіаційний ряд. З цією метою впорядкуємо елементи вибірки за зростанням і запишемо їх без повторів (стовпець «Варіанта x» у табл. 3). Навпроти кожної варіанти зазначимо її частоту, що вказує, яку кількість разів певний елемент зустрічався у вибірці. У розглядуваному випадку елементи 7,6, 8,2, 8,4, 8,5, 8,7 трапляються двічі, тому їх частота дорівнює 2, елемент 8,8 – тричі, тому його частота дорівнює 3, а для всіх інших елементів частота становить 1. Також для кожної варіанти розрахуємо відносну частоту p, поділивши частоту n на кількість елементів вибірки. У такий спосіб отримаємо варіаційний ряд у вигляді табл. 3.

Таблиця 3

Варіаційний ряд для вибірки з прикладу 1

№ з/п	Варіанта x	Частота n	Відносна частота p ($p = n / 27$)
1	3,8	1	1/27
2	5	1	1/27
3	7	1	1/27
4	7,1	1	1/27
5	7,2	1	1/27
6	7,4	1	1/27
7	7,6	2	2/27
8	7,7	1	1/27
9	7,9	1	1/27
10	8,1	1	1/27
11	8,2	2	2/27
12	8,3	1	1/27
13	8,4	2	2/27
14	8,5	2	2/27
15	8,7	2	2/27
16	8,8	3	3/27
17	8,9	1	1/27
18	9,2	1	1/27
19	9,5	1	1/27
20	10,2	1	1/27

Для контролю підрахуємо суму елементів у стовпці «Частота *n*» табл. 3. Вона має дорівнювати кількості елементів вибірки, у розглядуваному випадку – 27, таким чином, жодної помилки немає. Сума елементів у стовпці «Відносна частота *p*» має дорівнювати 1, ця умова також виконується.

Тепер розіб'ємо варіаційний ряд на класи. Оцінимо кількість класів:

$$M = \left[\sqrt{N}\right] = \left[\sqrt{27}\right] = 5.$$

Кількість класів завжди має бути цілим числом. Якщо $\sqrt{N}$ не ціле число, тоді одержане значення потрібно округлити до цілого.

Розрахуємо ширину кожного класу:

$$h = \frac{x_{\max} - x_{\min}}{M} = \frac{10,2 - 3,8}{5} = 1,28.$$

Слід зауважити, що ширина класу, на відміну від кількості класів, є дійсне число, яке не слід округляти.

Межі класів (їх має бути на одну більше за кількість класів) визначимо так:

$$x'_1 = x_{\min} = 3,8;$$
$$x'_2 = x_{\min} + h = 3,8 + 1,28 = 5,08;$$
$$x'_3 = x_{\min} + 2h = 3,8 + 2 \cdot 1,28 = 6,36;$$
$$x'_4 = x_{\min} + 3h = 3,8 + 3 \cdot 1,28 = 7,64;$$
$$x'_5 = x_{\min} + 4h = 3,8 + 4 \cdot 1,28 = 8,92;$$
$$x'_6 = x_{\min} + 5h = 3,8 + 5 \cdot 1,28 = 10,2 = x_{\max}.$$

Остання межа має дорівнювати максимальному елементу вибірки. Якщо через округлення ширини класів h вона виявилася близькою до $x_{\max}$, але не такою, що дорівнює йому, слід узяти, що вона дорівнює $x_{\max}$.

У результаті матимемо такі класи:

[3,8; 5,08)
[5,08; 6,36)
[6,36; 7,64)
[7,64; 8,92)
[8,92; 10,2].

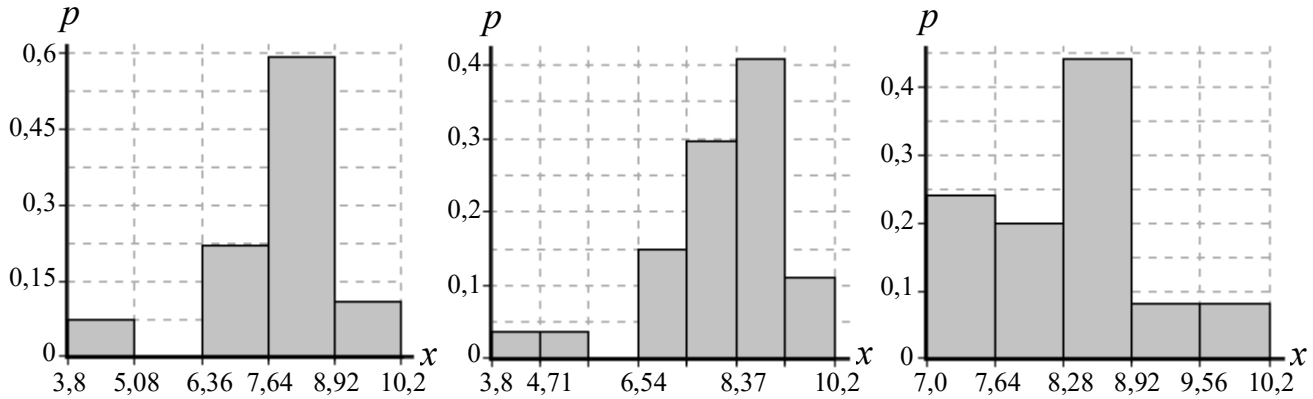
Кожен клас – це інтервал. Підрахуємо, скільки елементів вибірки потрапили до кожного класу, тобто встановимо частоту класів n. Наприклад, в інтервалі [3,8; 5,08) міститься 2 елементи вибірки, тобто частота 1-го класу дорівнює 2. В інтервалі [5,08; 6,36) міститься 0 елементів вибірки, тому частота 2-го класу становить 0, в інтервалі [6,36; 7,64) – 6 елементів вибірки (елемент 7,6 трапляється в цьому інтервалі двічі, тому кількість елементів у класі 6, а не 5). Аналогічно встановимо кількість елементів для інших класів. Також для кожного класу розрахуємо відносну частоту, поділивши частоту на кількість елементів у вибірці, тобто на 27. У результаті матимемо ряд, розбитий на класи (табл. 4).

Таблиця 4

Ряд, розбитий на класи, побудований за вибіркою з прикладу 1

№ з/п	Клас	Частота n	Відносна частота p ($p = n / 27$)
1	[3,8; 5,08)	2	2/27
2	[5,08; 6,36)	0	0
3	[6,36; 7,64)	6	6/27
4	[7,64; 8,92)	16	16/27
5	[8,92; 10,2]	3	1/9

На основі одержаного ряду побудуємо гістограму відносних частот (рис. 2, *а*). Одержана гістограма є одномодальна з правою асиметрією. Крім того, її вигляд дозволяє припустити, що у вибірці є два аномальні значення (3,8 та 5), які утворюють з лівого «хвоста» гістограми відокремлений клас із малою відносною частотою. Щоб у цьому переконатися, розіб'ємо ряд на більшу кількість класів, наприклад на 7 (рис. 2, *б*). І в цьому випадку також елементи 3,8 та 5 утворять окрему групу. Аномальні значення в разі їх виявлення слід вилучати з вибірки і будувати гістограму знову за описаною вище схемою. У даному разі гістограма, побудована за вибіркою, з якої вилучено два аномальні значення (3,8 та 5), більш схожа на одномодальну з лівою асиметрією (рис. 2, *в*). Це дозволяє припустити, що вибіркові дані розподілені за одним із таких законів: Вейбулла, Релея, екстремальним найбільшого значення.



а *б* *в*

Рис. 2. Гістограма відносних частот для даних з прикладу 1:
а – у разі розбиття на 5 класів; *б* – за розбиття на 7 класів;
в – після вилучення аномальних значень

Практична робота 2
ОЦІНЮВАННЯ СТАТИСТИЧНИХ ХАРАКТЕРИСТИК

Вибірка $\left\{x_i; i = \overline{1, N}\right\}$ являє собою підмножину, вибрану випадковим чином з генеральної сукупності. Генеральну сукупність описують певні статистичні характеристики, значення яких можна оцінити за даними вибірки. Оцінки бувають точкові та інтервальні.

Для параметра θ **точкова оцінка** $\overline{\theta}$ дозволяє отримати наближене

значення параметра, **інтервальна оцінка** являє собою інтервал $[\theta_н;\theta_в]$, який із заданою імовірністю містить справжнє значення параметра.

Точкові оцінки бувають зсунені та незсунені. Перевагу в процесі аналізу даних слід віддавати незсуненим оцінкам.

Середнє арифметичне є оцінка справжнього середнього значення (математичного сподівання) генеральної сукупності. Його обчислюють як за початковою вибіркою, так і за варіаційним рядом:

$$\overline{x} = \underbrace{\frac{1}{N}\sum_{i=1}^{N} x_i}_{\text{за вибіркою}} = \underbrace{\frac{1}{N}\sum_{i=1}^{r} x_i n_i}_{\text{за варіаційним рядом}}.$$

Величина середнього арифметичного визначає розташування гістограми (графіка функції щільності) на осі спостережень (рис. 3) та характеризує типове значення показника x.

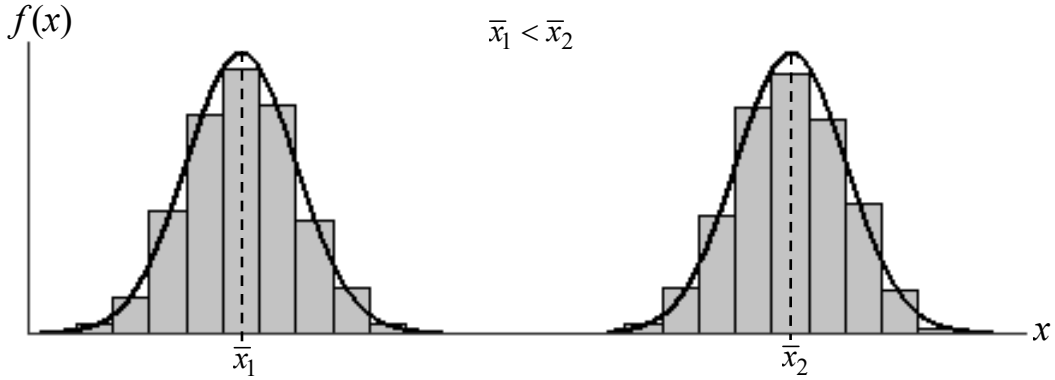


Рис. 3. Нормовані гістограми та графіки функції щільності залежно від середнього арифметичного

Вибіркова медіана – це оцінка справжньої медіани сукупності. Її знаходять за впорядкованою за зростанням вибіркою. Якщо кількість елементів вибірки непарна, то вибіркова медіана – елемент впорядкованої вибірки з номером $(N+1)/2$:

$$\text{Med} = x_{(N+1)/2}.$$

У випадку парної кількості елементів вибіркова медіана – це середнє арифметичне елементів з номерами $N/2$ та $1+N/2$:

$$\text{Med} = \frac{1}{2}\left(x_{N/2} + x_{1+N/2}\right).$$

Медіана, як і середнє, характеризує типове значення показника, але в дещо іншому аспекті. Наприклад, визначено вміст цементу в зразках пісковику. Середнє в такому разі описуватиме середній вміст цементу в досліджуваних зразках, а медіана становитиме таке значення, що в одній половині зразків уміст цементу буде меншим за нього, а в іншій половині – більшим. І середнє, і медіана – типові значення вмісту цементу в зразках пісковику, але їх фізичний зміст різний.

Вибіркова дисперсія є оцінка справжньої дисперсії генеральної

сукупності. Її одержують за даними вибірки у вигляді зсуненої та незсуненої оцінок:

– зсунена:

$$\hat{S}^2 = \underbrace{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2}_{\text{за вибіркою}} = \underbrace{\frac{1}{N}\sum_{i=1}^{r}\left(x_i - \overline{x}\right)^2 n_i}_{\text{за варіаційним рядом}};$$

– незсунена:

$$\overline{S}^2 = \underbrace{\frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2}_{\text{за вибіркою}} = \underbrace{\frac{1}{N-1}\sum_{i=1}^{r}\left(x_i - \overline{x}\right)^2 n_i}_{\text{за варіаційним рядом}}.$$

Вибіркова дисперсія характеризує розсіювання даних вибірки відносно середнього. Чим більша її величина, тим сильніше розкидані дані відносно середнього (рис. 4).

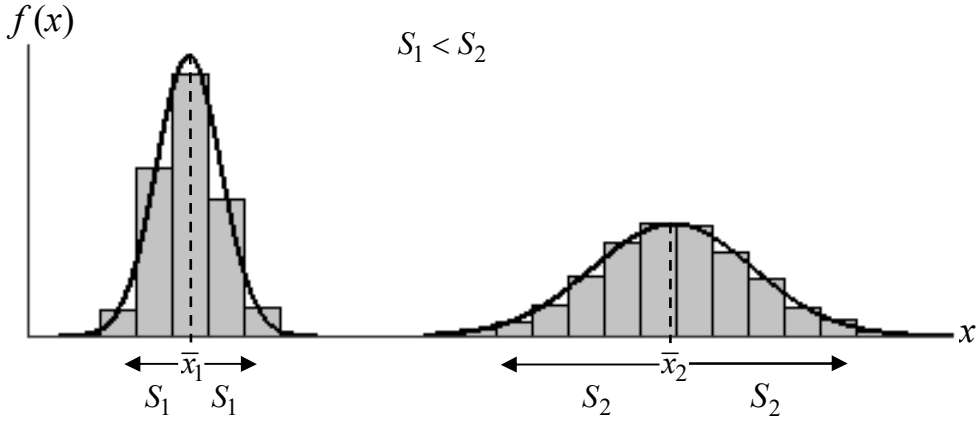


Рис. 4. Нормовані гістограми та графіки функції щільності залежно від дисперсії та середньоквадратичного відхилення

Середньоквадратичне відхилення – це корінь із дисперсії. Відповідно воно описує те саме, що й дисперсія, і його оцінка може бути:

– зсунена:

$$\hat{S} = \sqrt{\hat{S}^2};$$

– незсунена:

$$\overline{S} = \sqrt{\overline{S}^2}.$$

Оцінку **коефіцієнта варіації Пірсона** розраховують за формулою

$$\overline{W} = \frac{\overline{S}}{\overline{x}}.$$

Коефіцієнт відображає відносну варіабельність даних у частках відносно середнього і характеризує якість вибірки. Якщо $\overline{W} < 1$, вибірку вважають якісною.

Оцінку **коефіцієнта асиметрії** розраховують як

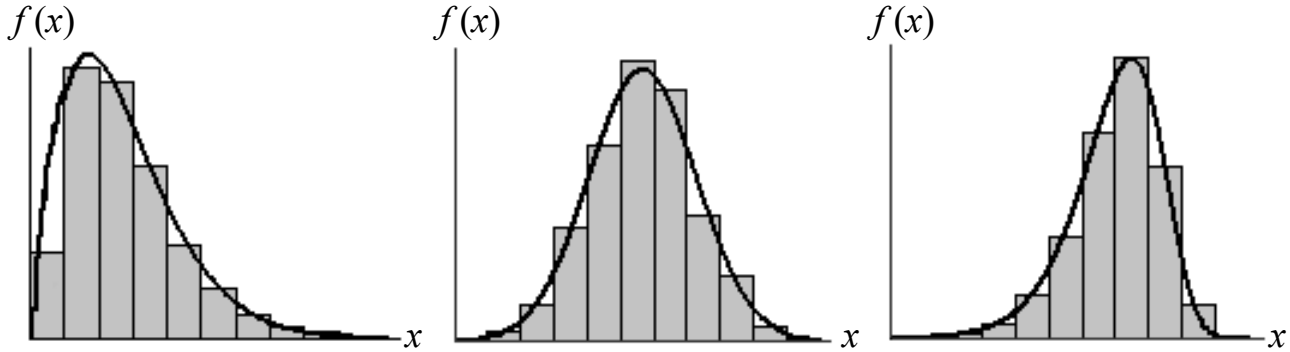
$$\hat{A} = \underbrace{\frac{1}{N\hat{S}^3}\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^3}_{\text{за вибіркою}} = \underbrace{\frac{1}{N\hat{S}^3}\sum_{i=1}^{r}\left(x_i - \overline{x}\right)^3 n_i}_{\text{за варіаційним рядом}}$$

або

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

Обидві оцінки є незсунені у випадку нормально розподілених даних. Якщо розподіл вибіркових даних відмінний від нормального, то оцінки є лише асимптотично незсунені (тобто незсунені у разі великих за обсягом вибірок).

Згаданий коефіцієнт характеризує асиметричність гістограми (графіка функції щільності). У теорії для симетрично розподілених даних $A = 0$; для даних з лівою асиметрією $A > 0$, з правою – $A < 0$ (рис. 5). На практиці точкова оцінка коефіцієнта асиметрії, одержана за даними вибірки, майже ніколи не дорівнює нулю, тому на основі її значення некоректно робити висновок щодо асиметричності розподілу даних. Щоб зробити такий висновок, аналізують довірчий інтервал на коефіцієнт асиметрії.



а *б* *в*

Рис. 5. Нормовані гістограми та графіки функції щільності залежно від коефіцієнта асиметрії:
а – $A > 0$; *б* – $A = 0$; *в* – $A < 0$

Оцінку **коефіцієнта ексцесу** обчислюють як

$$\hat{E} = \underbrace{\frac{1}{N\hat{S}^4}\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^4 - 3}_{\text{за вибіркою}} = \underbrace{\frac{1}{N\hat{S}^4}\sum_{i=1}^{r}\left(x_i - \overline{x}\right)^4 n_i - 3}_{\text{за варіаційним рядом}}$$

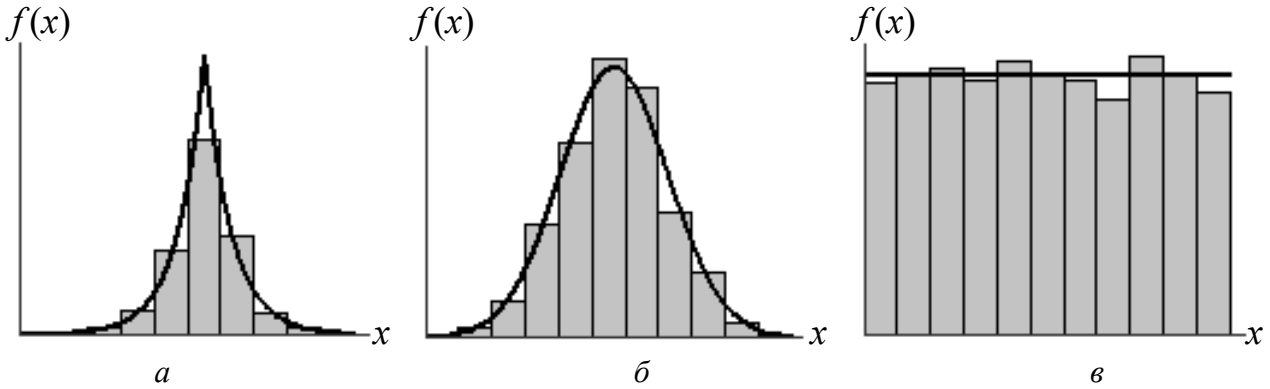
або

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

Загалом обидві оцінки лише асимптотично незсунені. Для нормально розподілених даних оцінка $\overline{E}$ незсунена.

Коефіцієнт ексцесу характеризує гостроту піку гістограми (графіка функції щільності розподілу) відносно нормального розподілу. Теоретично для даних, графік функції щільності розподілу яких відповідає нормальному закону, $E = 0$;

якщо графік функції щільності має гострий пік, то $E > 0$, коли пологий – $E < 0$ (рис. 6). На практиці висновок про гостроту піку графіка функції щільності розподілу роблять, аналізуючи довірчий інтервал на коефіцієнт ексцесу, а не значення точкової оцінки.



а *б* *в*

Рис. 6. Нормовані гістограми та графіки функції щільності залежно від коефіцієнтів ексцесу та контрексцесу:
а – $E > 0$ $(\hat{\chi} < 0{,}515)$; *б* – $E = 0$; *в* – $E < 0$ $(\hat{\chi} > 0{,}63)$

Коефіцієнт контрексцесу оцінюють за формулою

$$\hat{\chi} = \frac{1}{\sqrt{\hat{E} + 3}}.$$

Він, як і коефіцієнт ексцесу, визначає форму розподілу, причому якщо вибіркове значення коефіцієнта $\hat{\chi} < 0{,}515$, у графіка функції щільності гострий пік; за $\hat{\chi} > 0{,}63$ має місце розподіл з пласким графіком функції щільності (рис. 6).

Інтервальна оцінка параметра θ – це $(1-\alpha)\cdot 100\%$ довірчий інтервал $[\theta_н;\theta_в]$. Величина α являє собою імовірність «промаху» параметра повз довірчий інтервал. Найчастіше застосовують $\alpha = 0{,}05$.

Для справжнього середнього $(1-\alpha)\cdot 100\%$ довірчий інтервал визначають як

$$[\bar{x}_н;\bar{x}_в],$$

де

$$\bar{x}_н = \bar{x} - t_{1-\alpha/2,\nu}\sigma\{\bar{x}\}; \qquad \bar{x}_в = \bar{x} + t_{1-\alpha/2,\nu}\sigma\{\bar{x}\};$$

$t_{1-\alpha/2,\nu}$ – квантиль розподілу Стьюдента (табл. Д.2); $\nu = N - 1$;

$\sigma\{\bar{x}\}$ – середньоквадратичне відхилення середнього

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

Для медіани $(1-\alpha)\cdot 100\%$ довірчий інтервал являє собою проміжок

$$[x'_j; x'_k],$$

де x'_j, x'_k – відповідно *j*-й та *k*-й елементи відсортованої вибірки:

$$j \approx \frac{N}{2} - u_{1-\alpha/2}\frac{\sqrt{N}}{2}, \qquad k \approx \frac{N}{2} + 1 + u_{1-\alpha/2}\frac{\sqrt{N}}{2};$$

$u_{1-\alpha/2}$ – квантиль стандартного нормального розподілу ($u_{1-\alpha/2} = 1{,}96$ за $\alpha = 0{,}05$).

Для інших характеристик довірчі інтервали можна визначити лише за припущення, що вибірка вилучена з нормальної генеральної сукупності. У такому разі межі довірчого інтервалу $\left[\theta_н;\theta_в\right]$ параметра θ можна розрахувати як

$$\theta_н = \overline{\theta} - t_{1-\alpha/2,\nu}\sigma\left\{\overline{\theta}\right\}, \qquad \theta_в = \overline{\theta} + t_{1-\alpha/2,\nu}\sigma\left\{\overline{\theta}\right\},$$

де $\sigma\left\{\overline{\theta}\right\}$ – середньоквадратичне відхилення оцінки параметра:

$$\sigma\left\{\overline{S}\right\} = \frac{\overline{S}}{\sqrt{2N}}, \qquad \sigma\left\{\overline{W}\right\} = \overline{W}\sqrt{\frac{1+2\overline{W}^2}{2N}},$$

$$\sigma\left\{\hat{A}\right\} = \sqrt{\frac{6(N-2)}{(N+1)(N+3)}}, \qquad \sigma\left\{\overline{A}\right\} = \sqrt{\frac{6N(N-1)}{(N-2)(N+1)(N+3)}},$$

$$\sigma\left\{\hat{E}\right\} = \sqrt{\frac{24N(N-2)(N-3)}{(N+1)^2(N+3)(N+5)}}, \qquad \sigma\left\{\overline{E}\right\} = \sqrt{\frac{24N(N-1)^2}{(N-3)(N-2)(N+3)(N+5)}}.$$

Слід зауважити, що коли $N > 60$, замість $t_{1-\alpha/2,\nu}$ можна застосовувати квантиль $u_{1-\alpha/2}$.

На основі оцінених статистичних характеристик можна запропонувати простий **спосіб пошуку аномальних значень**. Згідно з ним аномальними вважають елементи вибірки, які не потрапляють до інтервалу

$$\left[\overline{x} - u_{1-\alpha/2}\cdot\overline{S}; \overline{x} + u_{1-\alpha/2}\cdot\overline{S}\right].$$

Строго кажучи, такий спосіб слушний лише для нормально розподілених даних, але з огляду на простоту на практиці до нього часто вдаються і в разі незначного порушення цієї умови.

Приклад 2. Оцінимо основні статистичні характеристики за вибіркою, розглянутою в прикл. 1. Спочатку проведемо розрахунки за початковою вибіркою, без вилучення аномальних значень, виявлених під час аналізу гістограми. Розрахунки виконаємо з точністю 4 знаки після коми.

Сума всіх 27 елементів вибірки дорівнює 216,5. Отже, середнє арифметичне

$$\overline{x} = \frac{216{,}5}{27} = 8{,}018\,5.$$

Для обчислення вибіркової медіани впорядкуємо елементи початкової вибірки за зростанням:

3,8	5	7	7,1	7,2	7,4	7,6	7,6	7,7	7,9	8,1	8,2	8,2	**8,3**
8,4	8,4	8,5	8,5	8,7	8,7	8,8	8,8	8,8	8,9	9,2	9,5	10,2	

З огляду на те що кількість елементів цієї вибірки непарна, вибірковою

медіаною буде елемент під номером $(N+1)/2 = (27+1)/2 = 14$. Отже,

$$\text{Med} = x_{14} = 8{,}3.$$

З метою обчислити оцінки інших статистичних характеристик виконаємо допоміжні розрахунки (табл. 5).

Таблиця 5

Допоміжні розрахунки, необхідні для оцінювання статистичних характеристик

x	$x - \bar{x}$	$(x-\bar{x})^2$	$(x-\bar{x})^3$	$(x-\bar{x})^4$
7,0	–1,018 5	1,037 4	–1,056 6	1,076 2
8,4	0,381 5	0,145 5	0,055 5	0,021 2
7,4	–0,618 5	0,382 6	–0,236 6	0,146 4
7,2	–0,818 5	0,670 0	–0,548 4	0,448 9
9,2	1,181 5	1,395 9	1,649 2	1,948 5
5	–3,018 5	9,111 5	–27,503 1	83,018 6
9,5	1,481 5	2,194 8	3,251 5	4,817 1
10,2	2,181 5	4,758 9	10,381 4	22,646 8
7,9	–0,118 5	0,014 0	–0,001 7	0,000 2
8,8	0,781 5	0,610 7	0,477 3	0,373 0
8,7	0,681 5	0,464 4	0,316 5	0,215 7
8,2	0,181 5	0,032 9	0,006 0	0,001 1
8,8	0,781 5	0,610 7	0,477 3	0,373 0
8,7	0,681 5	0,464 4	0,316 5	0,215 7
3,8	–4,218 5	17,795 9	–75,072 3	316,694 0
8,2	0,181 5	0,032 9	0,006 0	0,001 1
8,5	0,481 5	0,231 8	0,111 6	0,053 7
8,1	0,081 5	0,006 6	0,000 5	0,000 0
7,1	–0,918 5	0,843 7	–0,774 9	0,711 8
8,5	0,481 5	0,231 8	0,111 6	0,053 7
8,9	0,881 5	0,777 0	0,684 9	0,603 7
8,8	0,781 5	0,610 7	0,477 3	0,373 0
8,3	0,281 5	0,079 2	0,022 3	0,006 3
7,7	–0,318 5	0,101 5	–0,032 3	0,010 3
8,4	0,381 5	0,145 5	0,055 5	0,021 2
7,6	–0,418 5	0,175 2	–0,073 3	0,030 7
7,6	–0,418 5	0,175 2	–0,073 3	0,030 7
Разом		$\Sigma_2 = 43{,}100\,7$	$\Sigma_3 = -86{,}971\,7$	$\Sigma_4 = 433{,}892\,4$

Відтак матимемо оцінки:

– для середньоквадратичного відхилення – зсунену та незсунену:

$$\hat{S} = \sqrt{\frac{\Sigma_2}{N}} = \sqrt{\frac{43{,}100\,7}{27}} = 1{,}263\,5,$$

$$\overline{S} = \sqrt{\frac{\Sigma_2}{N-1}} = \sqrt{\frac{43{,}100\,7}{26}} = 1{,}287\,5;$$

– коефіцієнта варіації Пірсона:

$$\overline{W} = \frac{\overline{S}}{\overline{x}} = \frac{1{,}287\,5}{8{,}018\,5} = 0{,}160\,6;$$

– коефіцієнта асиметрії:

$$\hat{A} = \frac{\Sigma_3}{N\hat{S}^3} = \frac{-86{,}971\,7}{27 \cdot 1{,}263\,5^3} = -1{,}596\,9,$$

$$\overline{A} = \frac{\sqrt{N(N-1)}}{N-2}\hat{A} = \frac{\sqrt{27 \cdot 26}}{25}(-1{,}596\,9) = -1{,}692\,4;$$

– коефіцієнта ексцесу:

$$\hat{E} = \frac{\Sigma_4}{N\hat{S}^4} - 3 = \frac{433{,}892\,4}{27 \cdot 1{,}263\,5^4} - 3 = 3{,}305\,5,$$

$$\overline{E} = \frac{N^2-1}{(N-2)(N-3)}\left(\hat{E} + \frac{6}{N+1}\right) = \frac{27^2-1}{25 \cdot 24}\left(3{,}305\,5 + \frac{6}{28}\right) = 4{,}270\,7;$$

– коефіцієнта контрексцесу:

$$\hat{\chi} = \frac{1}{\sqrt{\hat{E}+3}} = \frac{1}{\sqrt{3{,}305\,5+3}} = 0{,}398\,2.$$

Щоб розрахувати межі довірчих інтервалів для характеристик (припускаючи, що дані розподілені нормально), обчислимо спочатку середньоквадратичні відхилення оцінок:

$$\sigma\{\overline{x}\} = \frac{\overline{S}}{\sqrt{N}} = \frac{1{,}287\,5}{\sqrt{27}} = 0{,}247\,8,$$

$$\sigma\{\overline{S}\} = \frac{\overline{S}}{\sqrt{2N}} = \frac{1{,}287\,5}{\sqrt{2 \cdot 27}} = 0{,}175\,2,$$

$$\sigma\{\overline{W}\} = \overline{W}\sqrt{\frac{1+2\overline{W}^2}{2N}} = 0{,}160\,6 \cdot \sqrt{\frac{1+2 \cdot 0{,}160\,6^2}{2 \cdot 27}} = 0{,}022\,4,$$

$$\sigma\{\hat{A}\} = \sqrt{\frac{6(N-2)}{(N+1)(N+3)}} = \sqrt{\frac{6 \cdot 25}{28 \cdot 30}} = 0{,}422\,6,$$

$$\sigma\{\overline{A}\} = 0{,}447\,9,$$

$$\sigma\{\hat{E}\} = \sqrt{\frac{24N(N-2)(N-3)}{(N+1)^2(N+3)(N+5)}} = \sqrt{\frac{24 \cdot 27 \cdot 25 \cdot 24}{28^2 \cdot 30 \cdot 32}} = 0{,}718\,7,$$

$$\sigma\{\overline{E}\} = 0{,}872\,1,$$

Задаючи $\alpha = 0{,}05$ і враховуючи, що в розглядуваному випадку $\nu = 27-1 = 26$, з табл. Д.2 матимемо $t_{1-\alpha/2,\nu} = 2{,}06$. Тоді межі 95% довірчих інтервалів становитимуть

$$\overline{x}_{\text{н}} = \overline{x} - t_{1-\alpha/2,\nu}\sigma\{\overline{x}\} = 8,018\,5 - 2,06 \cdot 0,247\,8 = 7,508,$$
$$\overline{x}_{\text{в}} = \overline{x} + t_{1-\alpha/2,\nu}\sigma\{\overline{x}\} = 8,018\,5 + 2,06 \cdot 0,247\,8 = 8,529;$$

$$S_{\text{н}} = \overline{S} - t_{1-\alpha/2,\nu}\sigma\{\overline{S}\} = 1,287\,5 - 2,06 \cdot 0,175\,2 = 0,926\,6,$$
$$S_{\text{в}} = \overline{S} + t_{1-\alpha/2,\nu}\sigma\{\overline{S}\} = 1,287\,5 + 2,06 \cdot 0,175\,2 = 1,648\,4;$$
$$W_{\text{н}} = \overline{W} - t_{1-\alpha/2,\nu}\sigma\{\overline{W}\} = 0,160\,6 - 2,06 \cdot 0,022\,4 = 0,114\,5,$$
$$W_{\text{в}} = \overline{W} + t_{1-\alpha/2,\nu}\sigma\{\overline{W}\} = 0,160\,6 + 2,06 \cdot 0,022\,4 = 0,206\,7;$$

$$A_{\text{н}} = \hat{A} - t_{1-\alpha/2,\nu}\sigma\{\hat{A}\} = -1,596\,9 - 2,06 \cdot 0,422\,6 = -2,467\,5,$$
$$A_{\text{в}} = \hat{A} + t_{1-\alpha/2,\nu}\sigma\{\hat{A}\} = -1,596\,9 + 2,06 \cdot 0,422\,6 = -0,726\,3,$$

аналогічно на основі оцінки $\overline{A}$:

$$A_{\text{н}} = \overline{A} - t_{1-\alpha/2,\nu}\sigma\{\overline{A}\} = -1,692\,4 - 2,06 \cdot 0,447\,9 = -2,615\,1,$$
$$A_{\text{в}} = \overline{A} + t_{1-\alpha/2,\nu}\sigma\{\overline{A}\} = -1,692\,4 + 2,06 \cdot 0,447\,9 = -0,769\,7,$$

$$E_{\text{н}} = \hat{E} - t_{1-\alpha/2,\nu}\sigma\{\hat{E}\} = 3,305\,5 - 2,06 \cdot 0,718\,7 = 1,825,$$
$$E_{\text{в}} = \hat{E} + t_{1-\alpha/2,\nu}\sigma\{\hat{E}\} = 3,305\,5 + 2,06 \cdot 0,718\,7 = 4,786,$$

так само на основі оцінки $\overline{E}$:

$$E_{\text{н}} = \overline{E} - t_{1-\alpha/2,\nu}\sigma\{\overline{E}\} = 4,270\,7 - 2,06 \cdot 0,872\,1 = 2,474\,2,$$
$$E_{\text{в}} = \overline{E} + t_{1-\alpha/2,\nu}\sigma\{\overline{E}\} = 4,270\,7 + 2,06 \cdot 0,872\,1 = 6,067\,2.$$

Для медіани отримаємо таке:

$$j = \frac{27}{2} - 1,96 \cdot \frac{\sqrt{27}}{2} = 8,408 \approx 8, \qquad k = \frac{27}{2} + 1 + 1,96 \cdot \frac{\sqrt{27}}{2} = 19,592 \approx 20,$$

отже, межами 95% довірчого інтервалу є 8-й та 20-й елементи впорядкованої за зростанням вибірки, тобто 7,6 та 8,7.

Остаточно результати всіх розрахунків, округлені до другого знака після коми, зведемо до табл. 6.

Перевіримо тепер наявність аномальних значень на основі оцінених статистичних характеристик. Визначимо інтервал

$$\left[\overline{x} - u_{1-\alpha/2} \cdot \overline{S}; \overline{x} + u_{1-\alpha/2} \cdot \overline{S}\right] = \left[5,495; 10,542\right].$$

До нього не потрапляють два елементи вибірки – 3,8 і 5, тобто вони є аномальні. Такого ж висновку ми дійшли, аналізуючи гістограму. Хоч на практиці така узгодженість у результатах аналізу не завжди має місце.

Знайдені аномальні значення необхідно вилучити з вибірки. Після цього треба побудувати нову гістограму (див. рис. 2, *в*) та знову обчислити статистичні характеристики (табл. 7). Остаточний аналіз слід проводити саме на їх основі.

Таблиця 6

Статистичні характеристики вибірки з прикладу 1

Характеристика	Оцінка $\hat{\theta}$	Середньоквадратичне відхилення $\sigma\{\hat{\theta}\}$	95% довірчий інтервал	
			$\theta_н$	$\theta_в$
Середнє арифметичне $\overline{x}$	8,02	0,25	7,51	8,53
Медіана Med	8,3	–	7,6	8,7
Середньоквадратичне відхилення $\overline{S}$	1,29	0,18	0,93	1,65
Коефіцієнт варіації Пірсона $\overline{W}$	0,16	0,02	0,11	0,21
Коефіцієнт асиметрії $\hat{A}$	–1,60	0,42	–2,47	–0,73
Коефіцієнт асиметрії $\overline{A}$	–1,69	0,45	–2,62	–0,77
Коефіцієнт ексцесу $\hat{E}$	3,31	0,72	1,83	4,79
Коефіцієнт ексцесу $\overline{E}$	4,27	0,87	2,48	6,07
Коефіцієнт контрексцесу $\hat{\chi}$	0,40	–	–	–

Таблиця 7

Статистичні характеристики вибірки з прикладу 1 (після вилучення аномальних значень)

Характеристика	Оцінка $\hat{\theta}$	Середньоквадратичне відхилення $\sigma\{\hat{\theta}\}$	95% довірчий інтервал	
			$\theta_н$	$\theta_в$
Середнє арифметичне $\overline{x}$	8,31	0,03	8,24	8,37
Медіана Med	8,4	–	7,9	8,7
Середньоквадратичне відхилення $\overline{S}$	0,77	0,11	0,54	0,99
Коефіцієнт варіації Пірсона $\overline{W}$	0,09	0,01	0,07	0,12
Коефіцієнт асиметрії $\hat{A}$	0,26	0,44	–0,64	1,16
Коефіцієнт асиметрії $\overline{A}$	0,28	0,46	–0,68	1,23
Коефіцієнт ексцесу $\hat{E}$	–0,02	0,73	–1,52	1,49
Коефіцієнт ексцесу $\overline{E}$	0,27	0,90	–1,59	2,12
Коефіцієнт контрексцесу $\hat{\chi}$	0,58	–	–	–

Аналіз даних табл. 7 свідчить про таке:

– середнє значення досліджуваного показника для вибіркових даних становить близько 8,31; 95% довірчий інтервал [8,24; 8,37] з імовірністю 0,95 містить справжнє середнє значення, тобто математичне сподівання сукупності;

– медіана вибірки дорівнює 8,4, тобто значення досліджуваного показника для однієї половини об’єктів менше цієї величини, а для іншої половини – більше; довірчий інтервал [7,9; 8,7] з імовірністю 0,95 містить справжню медіану;

– розкид даних відносно середнього дорівнює 0,77, довірчий інтервал [0,54; 0,99] з імовірністю 0,95 містить справжній розкид;

– асиметрія даних близька до 0,28, у межі 95% довірчого інтервалу [–0,68; 1,23] потрапляє нуль, тому можна вважати, що розподіл даних симетричний;

– точкова оцінка коефіцієнта ексцесу дорівнює 0,27, у межі 95% довірчого інтервалу [–1,59; 2,12] потрапляє нуль, отже, справжній коефіцієнт ексцесу можна вважати таким, що дорівнює нулю, гострота піку графіка функції щільності розподілу відповідає нормальній.

Практична робота 3
ІДЕНТИФІКАЦІЯ НОРМАЛЬНОГО РОЗПОДІЛУ

Ідентифікацію здійснюють з метою визначити, з якого розподілу вилучено вибірку або чи вилучено вибірку із заданого розподілу. Ідентифікувати нормальний розподіл за вибіркою $\left\{x_i; i=\overline{1,N}\right\}$ можна одним із перерахованих нижче способів.

1. Візуальний аналіз гістограми

Для даних з нормального розподілу гістограма має бути одномодальною симетричною (рис. 7). Якщо побудована за вибіркою гістограма відповідає цій вимозі, говорять, що ідентифіковано нормальний розподіл, у протилежному випадку – що нормальний розподіл не ідентифіковано.

Слід зауважити, що на практиці гістограма вкрай рідко має ідеально симетричний вигляд. Здебільшого наявне незначне відхилення від ідеальної симетрії. При цьому чим менший обсяг вибірки, тим більшим може бути відхилення.

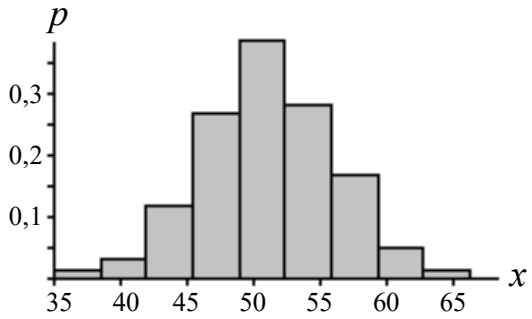


Рис. 7. Приклад гістограми, за якою ідентифіковний нормальний розподіл

2. Ідентифікація на основі коефіцієнтів асиметрії та ексцесу

Для нормального розподілу коефіцієнти асиметрії та ексцесу дорівнюють нулю. Тому на основі вибірки перевіряють, чи можна вважати, що для генеральної сукупності, з якої вилучено вибірку, ці коефіцієнти нульові. З цією метою висувають дві гіпотези:

$$\mathrm{H}_0 : A = 0, \qquad \mathrm{H}_0 : E = 0.$$

Для перевірки гіпотези про рівність нулю коефіцієнта асиметрії

розраховують статистику

$$t_A = \frac{\overline{A}}{\sigma\{\overline{A}\}}.$$

Якщо $|t_A| \le t_{1-\alpha/2,\nu}$, то коефіцієнт асиметрії вважають таким, що дорівнює нулю. Якщо ж $|t_A| > t_{1-\alpha/2,\nu}$, то коефіцієнт асиметрії вважають відмінним від нуля. Тут $t_{1-\alpha/2,\nu}$ – квантиль розподілу Стьюдента з $v = N-1$, значення якого знаходять зі спеціальної таблиці (табл. Д.2).

Щоб перевірити гіпотезу про рівність нулю коефіцієнта ексцесу, обчислюють статистику

$$t_E = \frac{\overline{E}}{\sigma\{\overline{E}\}},$$

яку порівнюють з $t_{1-\alpha/2,\nu}$. Якщо $|t_E| \le t_{1-\alpha/2,\nu}$, то коефіцієнт ексцесу вважають таким, що дорівнює нулю. За умови $|t_E| > t_{1-\alpha/2,\nu}$ коефіцієнт ексцесу вважають відмінним від нуля.

Якщо дійшли висновку, що обидва коефіцієнти дорівнюють нулю, то говорять, що ідентифіковано нормальний розподіл. У протилежному випадку – що нормальний розподіл не ідентифіковано. Цей спосіб застосовний, лише якщо обсяг оброблюваної вибірки не менший 500.

3. Ідентифікація на основі імовірнісного паперу

Цей спосіб ідентифікації найбільш надійний серед наведених. З метою ідентифікувати нормальний розподіл за ймовірнісним папером розраховують значення функції емпіричного розподілу в кожній варіанті варіаційного ряду x_i, $i = \overline{1,r}$, за формулою

$$F_N(x_i) = \begin{cases} p_1, & i = 1, \\ p_1 + p_2, & i = 2, \\ \dots \\ p_1 + p_2 + \dots + p_j, & i = j, \\ \dots \\ p_1 + p_2 + \dots + p_r = 1, & i = r. \end{cases}$$

Відтак для кожної варіанти обчислюють значення квантиля стандартного нормального розподілу порядку $F_N(x_i)$:

$$u_i = u_{F_N(x_i)}.$$

Значення цього квантиля також можна знайти з таблиці квантилів (табл. Д.1).

Значення (x_i, u_i), $i = \overline{1,r}$ відображають на ймовірнісному папері у вигляді точок (рис. 8). Якщо точки на папері утворюють пряму лінію, то говорять, що ідентифіковано нормальний розподіл, в іншому випадку – що нормальний розподіл не ідентифіковано.

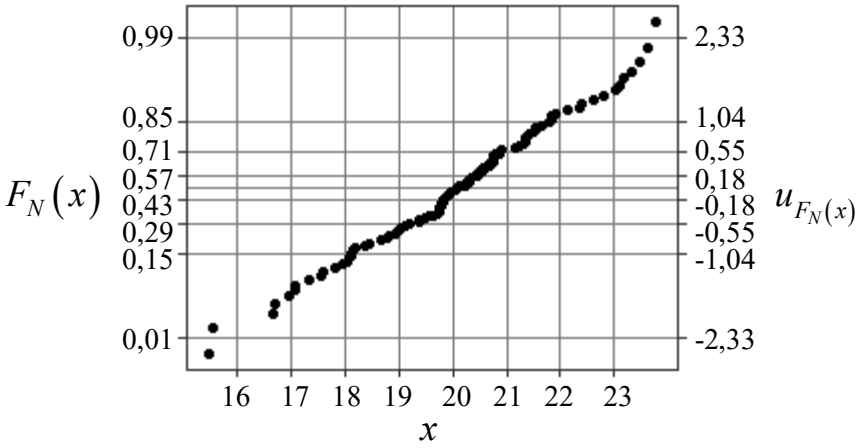


Рис. 8. Імовірнісний папір з нанесеними на нього статистичними даними

Приклад 3. Перевіримо, чи вибірку з прикл. 1 вилучено з нормального розподілу, тобто чи ідентифіковний за нею нормальний розподіл. Для цього застосуємо всі три способи.

1. Візуальний аналіз гістограми

Гістограма, розглянута в практичній роботі 1, за вибіркою без аномальних значень не симетрична, а лівоасиметрична (див. рис. 2, *в*). На основі такої гістограми нормальний розподіл не ідентифіковний. Але обсяг вибірки дуже малий, щоб вважати цей висновок надійним. Тому обов'язково потрібно вдатися й до інших способів.

2. Ідентифікація на основі коефіцієнтів асиметрії та ексцесу

Перевіримо гіпотези про рівність нулю коефіцієнтів асиметрії та ексцесу. Для цього розрахуємо статистики

$$t_A = \frac{\overline{A}}{\sigma\{\overline{A}\}} = \frac{0,28}{0,46} = 0,61,$$

$$t_E = \frac{\overline{E}}{\sigma\{\overline{E}\}} = \frac{0,27}{0,9} = 0,30$$

і порівняємо їх із квантилем Стьюдента.

Задаючи $\alpha = 0,05$ і враховуючи, що в розглядуваному випадку $\nu = 25 - 1 = 24$, з таблиці квантилів розподілу Стьюдента матимемо $t_{1-\alpha/2,\nu} = 2,06$.

У цьому випадку правдиві такі нерівності:

$$|t_A| \le t_{1-\alpha/2,\nu}, \qquad 0,61 \le 2,06;$$

$$|t_E| \le t_{1-\alpha/2,\nu}, \qquad 0,30 \le 2,06.$$

Виконання нерівності $|t_E| \le t_{1-\alpha/2,\nu}$ свідчить, що і коефіцієнт асиметрії, і коефіцієнт ексцесу можна вважати такими, що дорівнюють нулю. Тому за вибіркою без аномальних значень нормальний розподіл ідентифіковано.

3. Ідентифікація на основі ймовірнісного паперу

У прикл. 1 за вибіркою вже було побудовано варіаційний ряд за даними, що містили аномальні значення. Побудуємо тепер варіаційний ряд за даними без аномальних значень 3,8 і 5. Крім того, для кожної варіанти розрахуємо значення емпіричної функції розподілу $F_N(x)$ (табл. 8). Для першої варіанти її значення дорівнює відносній частоті цієї варіанти, для наступних – сумі відносних частот поточної та всіх попередніх варіант. Після цього з табл. Д.1 знайдемо значення $u_{F_N(x)}$ для кожної варіанти (табл. 8).

Таблиця 8

Значення емпіричної функції розподілу $F_N(x)$ та квантилів $u_{F_N(x)}$, обчислені для елементів варіаційного ряду, побудованого за даними з прикладу 1 після вилучення аномальних значень

№ з/п	Варіанта x	Частота n	Відносна частота p	Значення емпіричної функції розподілу $F_N(x)$	x	$u = u_{F_N(x)}$
1	7	1	1/25	1/25 = 0,04	7	–1,75
2	7,1	1	1/25	2/25 = 0,08	7,1	–1,41
3	7,2	1	1/25	3/25 = 0,12	7,2	–1,17
4	7,4	1	1/25	4/25 = 0,16	7,4	–0,99
5	7,6	2	2/25	6/25 = 0,24	7,6	–0,71
6	7,7	1	1/25	7/25 = 0,28	7,7	–0,58
7	7,9	1	1/25	8/25 = 0,32	7,9	–0,47
8	8,1	1	1/25	9/25 = 0,36	8,1	–0,36
9	8,2	2	2/25	11/25 = 0,44	8,2	–0,15
10	8,3	1	1/25	12/25 = 0,48	8,3	–0,05
11	8,4	2	2/25	14/25 = 0,56	8,4	0,15
12	8,5	2	2/25	16/25 = 0,64	8,5	0,36
13	8,7	2	2/25	18/25 = 0,72	8,7	0,58
14	8,8	3	3/25	21/25 = 0,84	8,8	0,99
15	8,9	1	1/25	22/25 = 0,88	8,9	1,17
16	9,2	1	1/25	23/25 = 0,92	9,2	1,41
17	9,5	1	1/25	24/25 = 0,96	9,5	1,75
18	10,2	1	1/25	1	–	–

Значення x та u відобразимо графічно у вигляді точок. Власне кажучи, ці точки потрібно наносити на імовірнісний папір, але на практиці можна відобразити їх і в рівномірній сітці (рис. 9). У цьому випадку дані на одержаному графіку утворюють досить чітку пряму лінію, тому нормальний розподіл ідентифіковано.

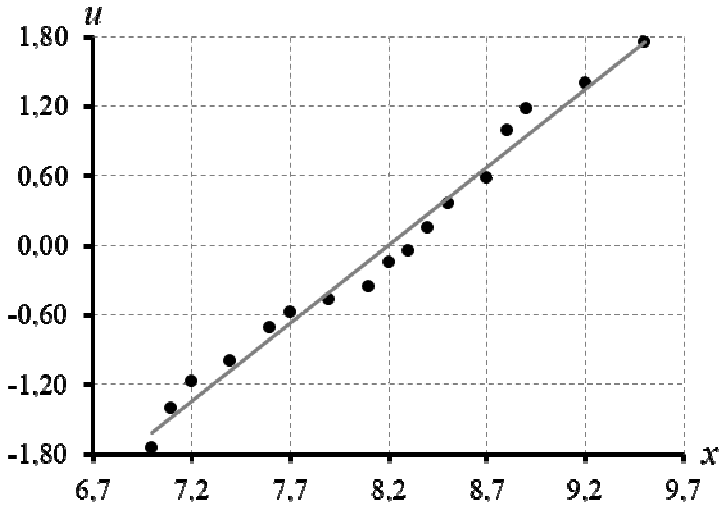


Рис. 9. Графічне відображення масиву $\left\{\left(x_i, u_i\right),\ i=\overline{1,r}\right\}$

Отже, для вибірки з прикл. 1 нормальний розподіл ідентифіковано на основі коефіцієнтів асиметрії та ексцесу, а також ймовірнісного паперу. Вигляд же гістограми не дозволяє припустити нормальність розподілу даних. Але оскільки гістограма – найменш надійний спосіб ідентифікації серед розглянутих, можна вважати, що вибіркові дані вилучено з генеральної сукупності, що має нормальний закон розподілу.

Контрольні запитання

1. У яких шкалах можна вимірювати показники? Які показники вимірюють у кожній з цих шкал?
2. За якою гістограмою ідентифікують нормальний розподіл?
3. Які розподіли ідентифікують за унімодальною гістограмою?
4. Про що свідчить двомодальна гістограма?
5. Якому значенню дорівнює медіана вибірки {4, 2, 7, 8, 2}?
6. Що характеризують вибіркові дисперсія та середньоквадратичне відхилення? Яким співвідношенням вони пов'язані між собою?
7. Якому значенню дорівнює дисперсія вибірки {1, 1, 1, 1, 1}?
8. Які властивості розподілу даних характеризують коефіцієнти асиметрії та ексцесу? Якому значенню вони мають дорівнювати для нормально розподілених даних?
9. Чи можна вважати на рівні значущості 0,05, що дані розподілено симетрично, якщо під час обробки вибірки з 200 елементів одержано $\hat{A}=0{,}12$, $\sigma\left\{\hat{A}\right\}=0{,}05$?
10. Які оцінки називають точковими, а які – інтервальними?
11. У який спосіб можна знайти аномальні значення у вибірці?
12. Які існують способи ідентифікації нормального розподілу?

Вказівки до виконання практичних робіт 4 – 7 на тему «Перевірка однорідності двох вибірок»

Вибірки показників x та y **вважають однорідними**, якщо їх вилучено із генеральних сукупностей з однаковими функціями розподілів. Формально гіпотеза про однорідність двох вибірок має вигляд

$$\mathrm{H}_0 : F(x) = G(x).$$

Під час перевірки однорідності розрізняють випадки залежних та незалежних вибірок. Для кожного випадку розроблено різні критерії.

Залежні є вибірки, одержані в результаті спостереження одних і тих самих об'єктів (процесів, явищ), але в різний час чи різними методами.

Незалежні є вибірки, отримані в результаті спостереження різних об'єктів (процесів, явищ).

Наприклад, визначено вміст певного хімічного елемента в N зразках граніту одним способом, а потім іншим. Таким чином одержано дві залежні вибірки. Якщо заміри було проведено для деякої кількості зразків граніту з родовища A та деякої кількості зразків з родовища B, то мають місце дві незалежні вибірки. Або ж якщо було заміряно вміст хімічного елемента в зразках граніту середнього палеозою та пізнього палеозою, то так само одержано дві незалежні вибірки.

Нехай, наприклад, в N свердловинах виміряно концентрацію певного елемента навесні та восени. У такий спосіб одержано дві залежні вибірки. Якщо заміри проведено в певній кількості свердловин з двох різних місцевостей, то мають місце дві незалежні вибірки.

Слід зауважити, що обсяги залежних вибірок завжди однакові, обсяги незалежних можуть різнитися.

Серед критеріїв для перевірки однорідності вибірок виділяють такі:

– **параметричні** – застосовні, коли розподіл показників x та y нормальний;

– **рангові** – застосовні, якщо розподіл хоча б одного з показників (x чи y) відмінний від нормального.

Неоднорідність вибірок може бути обумовлена відмінністю в параметрі зсуву розподілів (середньому значенні, медіані) або відмінністю в параметрі масштабу (наприклад, у дисперсії). З огляду на це **серед параметричних** виділяють **критерії для перевірки збігу середніх та дисперсій**. **Серед рангових** розрізняють **критерії зсуву** (Вілкоксона, Манна – Уїтні, Ван дер Вардена, медіанний та ін.) і **масштабу** (Клотца, квартильний, Севіджа і т. ін.).

Під час застосування рангових критеріїв доводиться стикатися з поняттям рангу. Ранг – це значення, яке присвоюють елементу вибірки; він дорівнює порядковому номеру цього елемента у впорядкованій за зростанням вибірці. Але якщо у вибірці є однакові елементи, то їх ранг дорівнює середньому арифметичному їх порядкових номерів. Приклад розрахунку рангів для елементів вибірки {4, 1, 6, 9, 4} наведено нижче (табл. 9).

Таблиця 9

Приклад розрахунку рангів

Початкова вибірка	Впорядкована вибірка	Порядковий номер	Ранг	Початкова вибірка	Ранг
4	1	1	1	4	2,5
1	4	2	2,5	1	1
6	4	3	2,5	6	4
9	6	4	4	9	5
4	9	5	5	4	2,5

Практична робота 4
ПЕРЕВІРКА РІВНОСТІ СЕРЕДНІХ ТА ДИСПЕРСІЙ У ВИПАДКУ ДВОХ ЗАЛЕЖНИХ ВИБІРОК

Нехай задано дві залежні вибірки $\left\{x_i; i=\overline{1,N}\right\}$ та $\left\{y_i; i=\overline{1,N}\right\}$. Для перевірки **гіпотези про рівність середніх значень** генеральних сукупностей, з яких вилучено ці вибірки,

$$\mathrm{H}_0 : m_x = m_y$$

за альтернативи

$$\mathrm{H}_1 : m_x \neq m_y$$

застосовують критерій, суть якого полягає в такому.

Обчисливши різниці $z_i = x_i - y_i$, одержують нову вибірку $\left\{z_i; i=\overline{1,N}\right\}$, за якою розраховують середнє значення $\overline{z}$ та середньоквадратичне відхилення S_z:

$$\overline{z} = \frac{1}{N}\sum_{i=1}^{N} z_i, \qquad \overline{S}_z = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(z_i - \overline{z}\right)^2}.$$

На їх основі визначають статистику критерію

$$t = \frac{\overline{z}\sqrt{N}}{\overline{S}_z}.$$

Якщо $|t| \le t_{1-\alpha/2,\nu}$, то немає підстав відхиляти гіпотезу про рівність середніх значень сукупностей, з яких вилучено вибірки, їх вважають рівними. Коли $|t| > t_{1-\alpha/2,\nu}$, вважають, що середні значення істотно відрізняються.

Величина $t_{1-\alpha/2,\nu}$ – це квантиль розподілу Стьюдента з кількістю степенів вільності $\nu = N-1$. Її значення знаходять за спеціальною таблицею (табл. Д.2).

Для перевірки **гіпотези про рівність дисперсій** генеральних сукупностей

$$\mathrm{H}_0 : \sigma_x^2 = \sigma_y^2$$

за альтернативи

$$\mathrm{H}_1 : \sigma_x^2 \neq \sigma_y^2$$

обчислюють статистику

$$f = \begin{cases} \dfrac{\overline{S}_x^2}{\overline{S}_y^2}, & \text{якщо } \overline{S}_x^2 \geq \overline{S}_y^2, \\ \dfrac{\overline{S}_y^2}{\overline{S}_x^2}, & \text{якщо } \overline{S}_x^2 < \overline{S}_y^2, \end{cases}$$

де $\overline{S}_x^2$, $\overline{S}_y^2$ – незсунені дисперсії вибірок *x* та *y* відповідно.

Якщо $f \leq f_{1-\alpha,\nu_1,\nu_2}$, говорять, що дисперсії генеральних сукупностей, з яких вилучено вибірки, збігаються. Якщо $f > f_{1-\alpha,\nu_1,\nu_2}$, вважають, що дисперсії відмінні. Величина $f_{1-\alpha,\nu_1,\nu_2}$ – це квантиль розподілу Фішера з кількістю степенів вільності

$$\nu_1 = \begin{cases} N_1 - 1, \text{якщо } \overline{S}_x^2 \geq \overline{S}_y^2, \\ N_2 - 1, \quad \text{якщо } \overline{S}_x^2 < \overline{S}_y^2 \end{cases} \quad \text{та} \quad \nu_2 = \begin{cases} N_2 - 1, \quad \text{якщо } \overline{S}_x^2 \geq \overline{S}_y^2, \\ N_1 - 1, \text{якщо } \overline{S}_x^2 < \overline{S}_y^2. \end{cases}$$

Її значення беруть зі спеціальних таблиць (табл. Д.3).

Приклад 4. З'ясуємо, чи істотно відрізняються середні концентрації хлору в пробах води, узятих зі свердловин на території гірничо-збагачувального комбінату, станом на вересень та листопад 2018 р. (табл. 10).

Таблиця 10

Дві залежні вибірки

Номер свердловини	Концентрація хлору станом на вересень 2018 р. (вибірка *x*)	Концентрація хлору станом на листопад 2018 р. (вибірка *y*)
1	86	187
2	78	187
3	642	653
4	192	336
5	571	466
6	485	504
7	553	672
8	535	550
9	328	404
10	795	821
11	830	578
12	692	93

З огляду на те що вибірки одержано в результаті спостереження над одними й тими самими об'єктами, але в різний час, вони є залежні. Обсяг кожної $N = 12$. Щоб перевірити, чи відрізняються середні концентрації хлору у вересні та листопаді 2018 р., розрахуємо статистику *t* (із використанням даних табл. 11):

$$t = \frac{\overline{z}\sqrt{N}}{\overline{S}_z} = \frac{28 \cdot \sqrt{12}}{210{,}64} = 0{,}46.$$

Таблиця 11

Допоміжні розрахунки для перевірки рівності середніх у випадку двох залежних вибірок

x	y	$z = x - y$	$z - \overline{z}$	$(z - \overline{z})^2$
86	187	–101	–129	16 641
78	187	–109	–137	18 769
642	653	–11	–39	1 521
192	336	–144	–172	29 584
571	466	105	77	5 929
485	504	–19	– 47	2 209
553	672	–119	–147	21 609
535	550	–15	– 43	1 849
328	404	–76	–104	10 816
795	821	–26	– 54	2 916
830	578	252	224	50 176
692	93	599	571	326 041
Разом		$\sum = 336$	–	$\sum = 488\,060$
		$\overline{z} = \frac{\sum}{N} = \frac{336}{12} = 28$		$\overline{S}_z = \sqrt{\frac{\sum}{N-1}} = \sqrt{\frac{488\,060}{11}} = 210{,}64$

Задаючи $\alpha = 0{,}05$ і враховуючи, що $v = 12 - 1 = 11$, з табл. Д.2 матимемо $t_{1-\alpha/2,\nu} = 2{,}20$. У цьому випадку слушна нерівність $|t| \le t_{1-\alpha/2,\nu}$. Отже, середні значення сукупностей, з яких узято вибірки, можна вважати рівними. У термінах предметної галузі це означає, що середня концентрація хлору в пробах води, узятих зі свердловин у листопаді 2018 р., значно змінилася порівняно з вереснем.

Щоб перевірити, чи має місце статистично значуща різниця в дисперсіях сукупностей, розрахуємо незсунені вибіркові дисперсії:

$$\overline{x} = 482, \qquad \overline{S}_x^2 = 66\ 581{,}84,$$

$$\overline{y} = 454, \qquad \overline{S}_y^2 = 48\ 942{,}93.$$

У розглядуваному випадку $\overline{S}_x^2 > \overline{S}_y^2$. Тому

$$f = \frac{\overline{S}_x^2}{\overline{S}_y^2} = \frac{66\ 581{,}84}{48\ 942{,}93} = 1{,}36.$$

Задаючи $\alpha = 0{,}05$ і враховуючи, що в цьому разі $v_1 = v_2 = 12 - 1 = 11$, з табл. Д.3 отримаємо $f_{1-\alpha,v_1,v_2} = 2{,}82$. У цьому випадку слушна нерівність $f \le f_{1-\alpha,\nu_1,\nu_2}$, отже, дисперсії збігаються. Тобто розкид значень концентрації хлору відносно середнього в листопаді порівняно з вереснем 2018 р. також не змінився.

Практична робота 5
ПЕРЕВІРКА РІВНОСТІ СЕРЕДНІХ ТА ДИСПЕРСІЙ У ВИПАДКУ ДВОХ НЕЗАЛЕЖНИХ ВИБІРОК

Нехай задано дві незалежні вибірки $\left\{x_i; i=\overline{1,N_1}\right\}$ та $\left\{y_i; i=\overline{1,N_2}\right\}$. Для перевірки **гіпотези про рівність середніх значень** двох сукупностей, з яких вилучено ці вибірки,

$$\mathrm{H}_0 : m_x = m_y$$

за альтернативи

$$\mathrm{H}_1 : m_x \neq m_y$$

запропоновано точний критерій лише для випадку, коли слушна гіпотеза $\mathrm{H}_0 : \sigma_x^2 = \sigma_y^2$. Цей критерій полягає в такому.

За кожною з вибірок розраховують середні значення ($\overline{x}$, $\overline{y}$) та незсунені вибіркові дисперсії ($\overline{S}_x^2$, $\overline{S}_y^2$). Далі обчислюють зважене середнє S^2 оцінок S_x^2, S_y^2:

$$S^2 = \frac{(N_1-1)\overline{S}_x^2 + (N_2-1)\overline{S}_y^2}{N_1+N_2-2}$$

і розраховують статистику критерію

$$t = \frac{\overline{x}-\overline{y}}{\sqrt{\dfrac{S^2}{N_1}+\dfrac{S^2}{N_2}}}$$

або з урахуванням вигляду S^2

$$t = \frac{\overline{x}-\overline{y}}{\sqrt{\dfrac{(N_1-1)\overline{S}_x^2 + (N_2-1)\overline{S}_y^2}{N_1+N_2-2}}} \sqrt{\frac{N_1N_2}{N_1+N_2}}.$$

Якщо $|t| \le t_{1-\alpha/2,\nu}$, говорять про рівність середніх значень тих сукупностей, з яких вилучено вибірки. Коли $|t| > t_{1-\alpha/2,\nu}$, вважають, що середні значення відрізняються істотно.

Значення квантиля Стьюдента $t_{1-\alpha/2,\nu}$ з кількістю степенів вільності $\nu = N_1 + N_2 - 2$ знаходять зі спеціальних таблиць (табл. Д.2).

Якщо дисперсії сукупностей відмінні, застосовують наближені критерії, наприклад критерій з правкою Уелча, в основі якого лежить статистика

$$t_1 = \frac{\overline{x}-\overline{y}}{\sqrt{\dfrac{\overline{S}_x^2}{N_1}+\dfrac{\overline{S}_y^2}{N_2}}}.$$

Висновок про рівність середніх значень роблять, якщо $|t_1| \le t_{1-\alpha/2,\nu}$. В

іншому випадку говорять, що має місце істотна різниця в середніх значеннях. При цьому кількість степенів вільності квантиля Стьюдента $t_{1-\alpha/2,v}$ дорівнює

$$v=\left(\frac{\overline{S}_x^2}{N_1}+\frac{\overline{S}_y^2}{N_2}\right)^2\left(\frac{1}{N_1-1}\left(\frac{\overline{S}_x^2}{N_1}\right)^2+\frac{1}{N_2-1}\left(\frac{\overline{S}_y^2}{N_2}\right)^2\right)^{-1}.$$

Перевірку рівності дисперсій у випадку незалежних вибірок здійснюють за тим самим критерієм, що і для залежних вибірок.

Приклад 5. З'ясуємо, чи істотна відмінність між середніми концентраціями хлору в пробах води, узятих із двох різних свердловин на території гірничо-збагачувального комбінату в один і той же час.

Концентрація хлору у свердловині 1 (вибірка *x*):

102,9 142,0 353,1 253,3 169,9 234,4 277,9 175,8

Концентрація хлору у свердловині 2 (вибірка *y*):

424,9 353,1 310,2 422,0 454,2 390,6 372,4.

У цьому разі мають місце незалежні вибірки, оскільки концентрацію хлору було виміряно в різних свердловинах. Обсяг вибірки *x* складає $N_1=8$, а вибірки *y* – $N_2=7$. Розрахуємо середні арифметичні та незсунені дисперсії цих вибірок:

$$\overline{x}=213{,}663, \qquad \overline{S}_x^2=6\ 617{,}574,$$
$$\overline{y}=389{,}629, \qquad \overline{S}_y^2=2\ 400{,}409.$$

Перевіримо спочатку рівність дисперсій сукупностей, з яких вилучено вибірки. Оскільки має місце випадок, коли $\overline{S}_x^2>\overline{S}_y^2$, то

$$f=\frac{\overline{S}_x^2}{\overline{S}_y^2}=\frac{6\ 617{,}574}{2\ 400{,}409}=2{,}757.$$

Задаючи $\alpha=0{,}05$ і враховуючи, що $v_1=N_1-1=7$, $v_2=N_2-1=6$, з табл. Д.3 одержимо $f_{1-\alpha,v_1,v_2}=4{,}21$.

У розглядуваному випадку слушна нерівність $f\le f_{1-\alpha,v_1,v_2}$, отже, дисперсії можна вважати рівними. Тому застосуємо точний критерій для перевірки рівності середніх, для якого слушне таке:

$$S^2=\frac{(8-1)\cdot 6\ 617{,}574+(7-1)\cdot 2\ 400{,}409}{8+7-2}=4\ 671{,}19,$$

$$t=\frac{213{,}663-389{,}629}{\sqrt{\dfrac{4\ 671{,}19}{8}+\dfrac{4\ 671{,}19}{7}}}=-4{,}975, \quad |t|=4{,}975.$$

Задаючи $\alpha=0{,}05$ і враховуючи, що $v=8+7-2=13$, з табл. Д.2 отримаємо $t_{1-\alpha/2,v}=2{,}16$. У цьому разі слушна нерівність $|t|>t_{1-\alpha/2,v}$, отже, середні сукупностей відрізняються, тобто середні концентрації хлору в пробах води, узятих зі свердловин 1 та 2, відмінні одна від одної.

Практична робота 6
ЗАСТОСУВАННЯ КРИТЕРІЮ ЗНАКОВИХ РАНГІВ ВІЛКОКСОНА ДО ДВОХ ЗАЛЕЖНИХ ВИБІРОК

Нехай задано дві залежні вибірки $\left\{x_i; i=\overline{1,N}\right\}$ та $\left\{y_i; i=\overline{1,N}\right\}$. Для перевірки гіпотези про відсутність зсуву у функціях розподілу показників x та y застосовують критерій знакових рангів Вілкоксона.

Формують нову вибірку $\left\{z_i; i=\overline{1,N}\right\}$, де $z_i = x_i - y_i$. Кожному її елементу ставлять у відповідність величину

$$\alpha_i = \begin{cases} 1, & \text{якщо } z_i > 0, \\ 0, & \text{якщо } z_i < 0. \end{cases}$$

Якщо серед елементів вибірки $\left\{z_i; i=\overline{1,N}\right\}$ є нульові, їх вилучають. Нехай N' – кількості ненульових значень z_i.

Далі переходять до вибірки $\left\{|z_i|; i=\overline{1,N'}\right\}$, елементи якої впорядковують за зростанням, присвоюючи їм ранги. Нехай r_i – ранг елемента $|z_i|$.

Статистика критерію дорівнює сумі додатних знакових рангів

$$T = \sum_{i=1}^{N'} \alpha_i r_i .$$

На її основі розраховують стандартизовану статистику

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

де

$$E\{T\} = \frac{1}{4} N'(N'+1), \qquad D\{T\} = \frac{1}{24} N'(N'+1)(2N'+1).$$

Якщо $|u| \le u_{1-\alpha/2}$, вважають, що зсуву у функціях розподілу немає. Якщо ж $|u| > u_{1-\alpha/2}$, говорять, що функції розподілу зсунені одна відносно іншої.

Величина $u_{1-\alpha/2}$ – квантиль стандартного нормального розподілу (табл. Д.1). За $\alpha = 0,05$ має місце рівність $u_{1-\alpha/2} = 1,96$.

Зауваження 1. Перевірку головної гіпотези доцільно проводити на основі стандартизованої статистики, лише коли $N > 15$. Для малих вибірок ($N \le 15$) рішення щодо прийняття чи відхилення головної гіпотези роблять, порівнюючи значення статистики T із табульованими критичними значеннями.

Зауваження 2. Якщо у вибірці $\left\{|z_i|; i=\overline{1,N'}\right\}$ є однакові елементи, то дисперсію статистики T слід обчислювати за уточненою формулою:

$$D\{T\} = \frac{1}{24}\left(N'(N'+1)(2N'+1) - \frac{1}{2}\sum_{j=1}^{g} A_j (A_j - 1)(A_j + 1) \right),$$

де g – кількість в'язок; A_j – кількість елементів у j-й в'язці.

Під час виконання практичної роботи цими зауваженнями можна знехтувати.

Приклад 6. Для вибірок з прикл. 4 перевіримо, чи відсутній зсув у їх функціях розподілу. Для цього розрахуємо статистику T критерію знакових рангів Вілкоксона (табл. 12).

Таблиця 12

Розрахунок статистики T критерію знакових рангів Вілкоксона

x	y	$z = x - y$	s	$\|z\|$	r – ранги для $\|z\|$	$s \cdot r$
86	187	–101	0	101	6	0
78	187	–109	0	109	8	0
642	653	–11	0	11	1	0
192	336	–144	0	144	10	0
571	466	105	1	105	7	7
485	504	–19	0	19	3	0
553	672	–119	0	119	9	0
535	550	–15	0	15	2	0
328	404	–76	0	76	5	0
795	821	–26	0	26	4	0
830	578	252	1	252	11	11
692	93	599	1	599	12	12
Разом						$T = 30$

Вище з метою присвоїти ранги $|z|$ (шостий стовпець табл. 12) вибірку з $|z|$ було впорядковано, її елементам було присвоєно порядкові номери, на основі яких визначено ранги (табл. 13). У даному разі у послідовності $|z|$ немає в'язок, тому ранги збігаються з порядковими номерами.

Таблиця 13

Визначення рангів для |z|

Впорядковані $\|z\|$	11	15	19	26	76	101	105	109	119	144	252	599
Порядкові номери $\|z\|$	1	2	3	4	5	6	7	8	9	10	11	12
Ранги для $\|z\|$	1	2	3	4	5	6	7	8	9	10	11	12

Отже, статистика критерію $T = 30$.

Розрахуємо стандартизовану статистику:

$$E\{T\} = \frac{1}{4} \cdot 12 \cdot (12+1) = 39, \qquad D\{T\} = \frac{1}{24} \cdot 12 \cdot (12+1) \cdot (2 \cdot 12 + 1) = 162{,}5,$$

$$u = \frac{30 - 39}{\sqrt{162{,}5}} = -0{,}71, \qquad |u| = 0{,}71.$$

Узявши $\alpha = 0{,}05$, маємо $u_{1-\alpha/2} = 1{,}96$. У даному разі слушна нерівність $|u| \le u_{1-\alpha/2}$, отже, зсуву у функціях розподілу показників x та y немає.

Практична робота 7
ЗАСТОСУВАННЯ КРИТЕРІЇВ СУМИ РАНГІВ ВІЛКОКСОНА ТА МАННА – УЇТНІ ДО ДВОХ НЕЗАЛЕЖНИХ ВИБІРОК

Нехай задано дві незалежні вибірки $\left\{x_i; i=\overline{1,N_1}\right\}$ та $\left\{y_i; i=\overline{1,N_2}\right\}$. Для перевірки гіпотези про відсутність зсуву у функціях розподілу показників x та y можна застосувати критерій суми рангів Вілкоксона або критерій Манна – Уїтні. Ці критерії є алгебрично еквівалентні, тобто приводять до однакового результату.

Критерій суми рангів Вілкоксона передбачає злиття вибірок в одну:

$$\left\{x_1, x_2, ..., x_{N_1}, y_1, y_2, ..., y_{N_2}\right\}.$$

Елементам спільної вибірки присвоюють ранги:

$$\left\{r\left(x_1\right), r\left(x_2\right), ..., r\left(x_{N_1}\right), r\left(y_1\right), r\left(y_2\right), ..., r\left(y_{N_2}\right)\right\}.$$

Як статистику критерію беруть величину, що дорівнює сумі рангів елементів вибірки x:

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

Далі розраховують стандартизовану статистику

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

де

$$E\{W\} = \frac{1}{2} N_1\left(N_1 + N_2 + 1\right), \qquad D\{W\} = \frac{1}{12} N_1 N_2\left(N_1 + N_2 + 1\right).$$

Якщо $|u| \le u_{1-\alpha/2}$, вважають, що зсуву у функціях розподілу немає. Якщо ж $|u| > u_{1-\alpha/2}$, говорять, що функції розподілу зсунені одна відносно іншої. Величина $u_{1-\alpha/2}$ – це квантиль стандартного нормального розподілу (табл. Д.1). За $\alpha = 0{,}05$ має місце рівність $u_{1-\alpha/2} = 1{,}96$.

Зауваження 3. Перевірка головної гіпотези на основі стандартизованої статистики u слушна лише у випадку, коли $N_1 + N_2 > 25$. Якщо $N_1 + N_2 < 25$, слід звертатися до табульованих критичних значень статистики W.

Приклад 7. Перевіримо відсутність зсуву у функціях розподілу значень концентрації хлору для двох різних свердловин (див. прикл. 5).

Відповідно до табл. 14

$$W = \sum_{i=1}^{8} r\left(x_i\right) = 37{,}5,$$

$$E\{W\} = \frac{1}{2}\cdot 8\cdot\left(8+7+1\right) = 64, \quad D\{W\} = \frac{1}{12}\cdot 8\cdot 7\cdot\left(8+7+1\right) = 74{,}667,$$

$$u = \frac{37{,}5 - 64}{\sqrt{74{,}667}} = -3{,}067, \quad |u| = 3{,}067.$$

Таблиця 14

Розрахунок статистики *W* критерію суми рангів Вілкоксона

Спільна вибірка	Впорядкована спільна вибірка	Порядковий номер	Ранг	Спільна вибірка	Ранг	Сума рангів вибірки *x*
102,9 (*x*)	102,9	1	1	102,9 (*x*)	1	
142,0 (*x*)	142,0	2	2	142,0 (*x*)	2	
353,1 (*x*)	169,9	3	3	351,6 (*x*)	9,5	
253,3 (*x*)	175,8	4	4	253,3 (*x*)	6	
169,9 (*x*)	234,4	5	5	169,9 (*x*)	3	
234,4 (*x*)	253,3	6	6	234,4 (*x*)	5	
277,9 (*x*)	277,9	7	7	277,9 (*x*)	7	
175,8 (*x*)	310,2	8	8	175,8 (*x*)	4	$\sum = 37{,}5$
424,9 (*y*)	353,1	9	9,5	424,9 (*y*)	14	
353,1 (*y*)	353,1	10	9,5	354,6 (*y*)	9,5	
310,2 (*y*)	372,4	11	11	310,2 (*y*)	8	
422,0 (*y*)	390,6	12	12	422,0 (*y*)	13	
454,2 (*y*)	422,0	13	13	454,2 (*y*)	15	
390,6 (*y*)	424,9	14	14	390,6 (*y*)	12	
372,4 (*y*)	454,2	15	15	372,4 (*y*)	11	

Вважаючи, що $\alpha = 0{,}05$, матимемо $u_{1-\alpha/2} = 1{,}96$. Оскільки $|u| > u_{1-\alpha/2}$, то має місце зсув у функціях розподілу значень концентрації хлору для двох різних свердловин.

В основі **критерію Манна – Уїтні** лежить поняття інверсії. Якщо для певного значення x_i у вибірці y є k значень, менших за x_i, то говорять, що для цього x_i має місце k інверсій. Якщо у вибірці y є значення, що дорівнює x_i, це розглядають як півінверсії. Як статистику критерію беруть суму інверсій за всіма x:

$$V = \sum_{i=1}^{N_1} v_i, \qquad v_i = \sum_{j=1}^{N_2} v_{i,j},$$

$$v_{i,j} = \begin{cases} 1, & \text{якщо } x_i > y_j, \\ 0{,}5, & \text{якщо } x_i = y_j, \\ 0, & \text{якщо } x_i < y_j. \end{cases}$$

Слід відзначити, що під час обчислення V величину $v_{i,j}$ можна визначати і за формулою

$$v_{i,j} = \begin{cases} 1, & \text{якщо } x_i < y_j, \\ 0{,}5, & \text{якщо } x_i = y_j, \\ 0, & \text{якщо } x_i > y_j. \end{cases}$$

Це еквівалентне тому, щоб вибірки x та y поміняти місцями. У такому разі статистика V буде пов'язана зі статистикою W критерію Вілкоксона співвідношенням

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

На основі V обчислюють стандартизовану статистику

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

де

$$E\{V\} = \frac{1}{2} N_1 N_2 , \qquad D\{V\} = \frac{1}{12} N_1 N_2 \left(N_1 + N_2 + 1 \right),$$

значення якої порівнюють із квантилем $u_{1-\alpha/2}$ нормального закону розподілу. Якщо $|u| \le u_{1-\alpha/2}$, головну гіпотезу про відсутність зсуву приймають, в іншому випадку відхиляють.

Зауваження 4. Як і під час застосування критерію Вілкоксона, якщо $N_1 + N_2 < 25$, замість розрахунку стандартизованої статистики u слід звертатися до табульованих критичних значень статистики V.

Приклад 8. Застосуємо до даних прикл. 7 критерій Манна – Уїтні.

Для кожного значення x_i розрахуємо кількість інверсій v_i (табл. 15). Наприклад, для $x_1 = 102{,}9$ у вибірці y немає жодного елемента, меншого за x_1, тобто немає жодної інверсії і $v_1 = 0$. Для $x_3 = 353{,}1$ у вибірці y є один елемент, менший за x_3 (це одна інверсія), та один елемент, який дорівнює x_3 (ще півінверсії). Тому $v_3 = 1{,}5$.

Таблиця 15

Розрахунок кількості інверсій v для елементів вибірки x

Порядковий номер i	1	2	3	4	5	6	7	8
Елемент вибірки x_i	102,9	142,0	353,1	253,3	169,9	234,4	277,9	175,8
Кількість інверсій v_i	0	0	1,5	0	0	0	0	0

Сумарна кількість інверсій і буде значенням статистики критерію:

$$V = \sum_{i=1}^{8} v_i = 1{,}5 .$$

Таким чином, матимемо

$$E\{V\} = \frac{1}{2} \cdot 8 \cdot 7 = 28 , \; D\{V\} = \frac{1}{12} \cdot 8 \cdot 7 \left(8 + 7 + 1 \right) = 74{,}667 ,$$

$$u = \frac{1{,}5 - 28}{\sqrt{74{,}667}} = -3{,}067 , \qquad |u| = 3{,}067 .$$

Оскільки $|u| > u_{1-\alpha/2}$, то має місце зсув у функціях розподілу значень концентрації хлору для двох різних свердловин. Як бачимо, результати застосування критеріїв Вілкоксона та Манна – Уїтні повністю збігаються.

Контрольні запитання

1. Які вибірки називають однорідними?
2. Який вигляд має гіпотеза про однорідність двох вибірок?
3. У чому полягає різниця між залежними та незалежними вибірками в задачі перевірки однорідності двох вибірок?
4. За якої умови для перевірки однорідності застосовують параметричні критерії?
5. Який критерій вважають параметричним?
6. Які існують параметричні критерії для перевірки однорідності двох вибірок?
7. За якою формулою розраховують статистику критерію під час перевірки рівності середніх у випадку двох залежних вибірок? А у випадку двох незалежних вибірок?
8. За яким законом розподілені статистики критеріїв для перевірки рівності середніх у випадку залежних та незалежних вибірок?
9. Який вигляд має статистика критерію для перевірки рівності дисперсій? З квантилем якого розподілу її порівнюють?
10. З якого критерію починають перевірку однорідності двох вибірок: рівності середніх чи дисперсій?
11. За допомогою яких рангових критеріїв можна перевірити однорідність двох залежних вибірок? А двох незалежних вибірок?
12. Яку гіпотезу перевіряють за допомогою рангових критеріїв зсуву?
13. Яку величину називають рангом? У який спосіб присвоюють ранги однаковим елементам вибірки?
14. Які ранги відповідають елементам такої вибірки: {3, -1, 9, 2, 0, 5, 3, 6}?
15. Як обчислюють статистику критерію знакових рангів Вілкоксона?
16. У який спосіб знаходять статистику критерію суми рангів Вілкоксона?
17. Що розуміють під інверсією в критерії Манна – Уїтні?
18. Чому збігаються результати застосування критеріїв суми рангів Вілкоксона та Манна – Уїтні?
19. Яким співвідношенням пов’язані статистики критеріїв суми рангів Вілкоксона та Манна – Уїтні?
20. У який спосіб приймають чи відхиляють головну гіпотезу у випадку застосування рангових критеріїв зсуву до малих вибірок?
21. Якому значенню дорівнює квантиль стандартного нормального розподілу $u_{1-\alpha/2}$ за $\alpha = 0,05$?

Вказівки до виконання практичних робіт 8 – 12 на тему «Кореляційний та регресійний аналіз»

У навколишньому світі можна спостерігати багато різних взаємодій, наприклад: між зростом та вагою людини (як правило, чим вища людина за зростом, тим вона більше важить); між обсягом продукції, яку випускає підприємство, та витратами; між часткою зусиль, що їх студент витрачає на підготовку до занять, та результуючою оцінкою і т. ін. Вивчення таких взаємозв'язків у двовимірних даних можна звести до розв'язання двох задач:

1) установлення наявності зв'язку між показниками;

2) за наявності зв'язку – ідентифікація та відновлення форми зв'язку.

Зв'язок між показниками буває двох видів:

1) **функціональний** – коли значення одного показника можна однозначно передбачити на основі значення іншого; наприклад, такий зв'язок існує між віком дерева та кількістю кілець на його зрізі;

2) **стохастичний** – у такому разі однозначного, функціонального зв'язку між показниками немає, можна говорити лише про те, що в разі зміни одного показника середні значення іншого також змінюються; на практиці зазвичай має місце стохастичний зв'язок.

За своєю формою **зв'язок** може бути:

1) **лінійний**;

2) **нелінійний**.

Виявити наявність зв'язку дозволяє **кореляційний аналіз**, що включає аналіз кореляційного поля та коефіцієнтів кореляції.

Якщо зв'язок існує, то постає завдання ідентифікувати та відновити зв'язок у вигляді рівняння, що описує закон зміни середніх значень одного показника залежно від значень іншого. Таке рівняння називають регресією, а його відновлення становить предмет розгляду **регресійного аналізу**.

Практична робота 8
АНАЛІЗ КОРЕЛЯЦІЙНОГО ПОЛЯ. ОЦІНЮВАННЯ КОЕФІЦІЄНТА КОРЕЛЯЦІЇ ПІРСОНА

Кореляційне поле являє собою графічне зображення масиву $\left\{x_i, y_i; i = \overline{1, N}\right\}$, коли за віссю абсцис відкладають значення x_i, а за віссю ординат – відповідні значення y_i (рис. 10). У науковій літературі замість терміна «кореляційне поле» інколи застосовують термін «діаграма розсіювання».

Візуальний аналіз кореляційного поля дозволяє ідентифікувати наявність зв'язку між показниками x та y. Поле у вигляді кола або овалу без нахилу свідчить про те, що зв'язок між x та y відсутній (рис. 10, *а*). Поле у вигляді овалу з нахилом дає змогу говорити про наявність лінійного зв'язку, а нахил

овалу – про додатний (рис 10, *б*) чи від'ємний зв'язок (рис. 10, *в*). Поле складної конфігурації (рис 10, *г*, *д*) свідчить про нелінійний зв'язок між x та y. Якщо в межах кола виділяють декілька сукупностей (рис. 10, *е*), це вказує на неоднорідність даних.

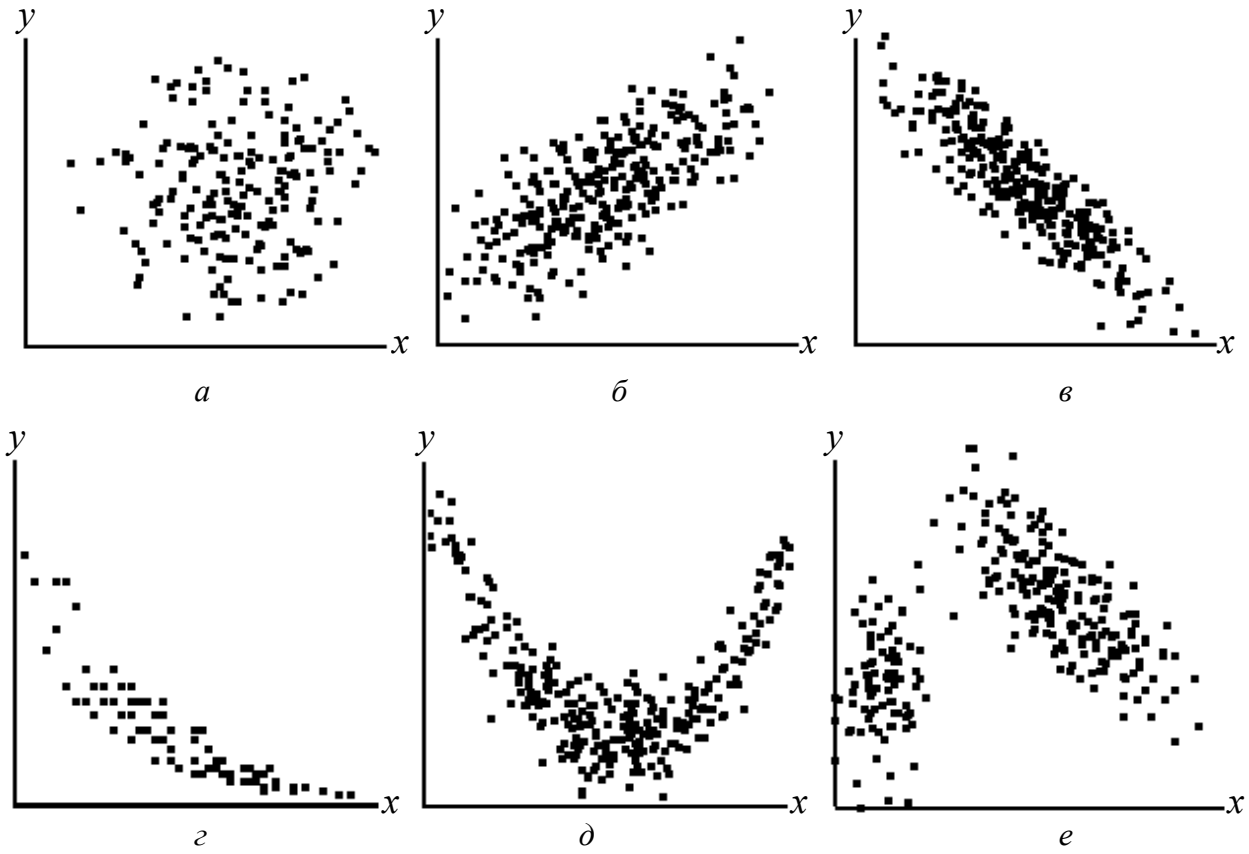


Рис. 10. Кореляційні поля:
а – зв'язок відсутній; *б* – додатний лінійний зв'язок; *в* – від'ємний лінійний зв'язок; *г*, *д* – нелінійний зв'язок; *е* – випадок неоднорідних даних

Кількісно міру залежності між показниками визначають **коефіцієнтом кореляції**. Залежно від закону розподілу показників уводять різні типи коефіцієнтів. Найпростіший є парний коефіцієнт кореляції Пірсона. Він відображає міру лінійної залежності між показниками. У цьому випадку потрібно, щоб їх розподіл був нормальний.

Коефіцієнт кореляції Пірсона має такі властивості:

1) $r \in [-1;1]$;

2) якщо між показниками x та y зв'язок відсутній, то $r = 0$; якщо розподіл показників нормальний, слушне і зворотне твердження: рівність нулю коефіцієнта кореляції свідчить про відсутність зв'язку між показниками *x* та *y*;

3) значення коефіцієнта кореляції, відмінне від нуля, вказує на лінійну залежність між x та y; при цьому якщо $r = \pm 1$, то має місце лінійна функціональна залежність, у противному разі – лінійна стохастична;

4) силу залежності між показниками визначає модуль коефіцієнта $|r|$, знак свідчить лише про те, є зв'язок додатний (рис. 10, *б*) чи від'ємний (рис. 10, *в*).

На практиці коефіцієнт кореляції Пірсона оцінюють за вибіркою $\left\{x_i, y_i; i=\overline{1,N}\right\}$ у такий спосіб:

$$\hat{r} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\hat{S}_x \cdot \hat{S}_y},$$

де

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \qquad \overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i, \qquad \overline{xy} = \frac{1}{N}\sum_{i=1}^{N} x_i y_i,$$

$$\hat{S}_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2}, \qquad \hat{S}_y = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \overline{y}\right)^2}.$$

Статистичне значення $\hat{r}$ завжди відмінне від нуля, тому щоб зробити висновок про відсутність чи наявність лінійної залежності між показниками, висувають гіпотезу $\mathrm{H}_0 : r = 0$. Для її перевірки обраховують статистику

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

Якщо $|t| \le t_{1-\alpha/2,\nu}$, вважають, що коефіцієнт кореляції Пірсона дорівнює нулю (незначущий), лінійний зв'язок між показниками відсутній. Коли $|t| > t_{1-\alpha/2,\nu}$, говорять, що коефіцієнт відмінний від нуля (значущий), між показниками існує лінійний зв'язок (додатний, якщо оцінка коефіцієнта додатна, в іншому випадку – від'ємний). Тут $t_{1-\alpha/2,\nu}$ – квантиль розподілу Стьюдента з кількістю степенів вільності $\nu = N-2$ (табл. Д.2); α – помилка першого роду, зазвичай $\alpha = 0,05$.

Інтервальне оцінювання коефіцієнта здійснюють шляхом призначення $(1-\alpha)\cdot 100\%$ довірчого інтервалу з межами

$$r_{\text{н,в}} = \hat{r} + \frac{\hat{r}\left(1-\hat{r}^2\right)}{2N} \mp u_{1-\alpha/2}\frac{1-\hat{r}^2}{\sqrt{N-1}},$$

де $u_{1-\alpha/2}$ – квантиль стандартного нормального розподілу ($u_{1-\alpha/2} = 1,96$ за $\alpha = 0,05$).

Приклад 9. Необхідно оцінити кореляційний зв'язок між умістом сульфат-іонів (x) та іонів кальцію (y) у підземних водах на території гірничо-збагачувального комбінату. Вибірку значень x та y задано (табл. 16).

Таблиця 16

Вибірка значень *x* та *y*

№ з/п	*x*	*y*	№ з/п	*x*	*y*	№ з/п	*x*	*y*
1	2 675,7	190,4	6	1 964,0	188,5	11	1 618,4	151,8
2	2 437,1	156,4	7	1 911,4	167,0	12	2 361,4	222,6
3	1 938,3	170,3	8	1 888,3	191,6	13	1 983,8	172,0
4	2 149,2	174,5	9	1 637,4	145,3	14	1 917,1	138,2
5	2 254,5	191,3	10	1 666,2	138,2	15	1 758,3	173,6

За даними табл. 16 побудуємо кореляційне поле (рис. 11).

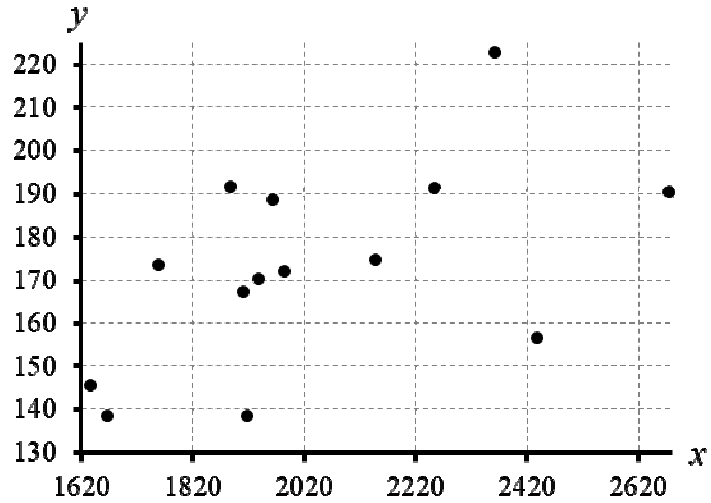


Рис. 11. Кореляційне поле для даних з прикладу 9

За формою поле близьке до овалу з нахилом. Тому можна зробити висновок про наявність лінійного стохастичного зв'язку між умістом сульфат-іонів та іонів кальцію в підземних водах. Нахил поля свідчить про те, що лінійний зв'язок додатний.

Перевіримо тепер наявність зв'язку за допомогою коефіцієнта кореляції Пірсона. Розрахуємо спочатку середні значення показників x, y та їх добутку:

$$\overline{x} = 2\,010{,}74; \qquad \overline{y} = 171{,}447; \qquad \overline{xy} = 348\,609{,}804.$$

Оцінимо середньоквадратичні відхилення показників x та y (у розглядуваному випадку потрібні зсунені середньоквадратичні відхилення):

$$\hat{S}_x = 299{,}226; \qquad \hat{S}_y = 22{,}523.$$

Таким чином, оцінка коефіцієнта Пірсона дорівнює

$$\hat{r} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\hat{S}_x \cdot \hat{S}_y} = \frac{348\,609{,}804 - 2\,010{,}74 \cdot 171{,}447}{299{,}226 \cdot 22{,}523} \approx 0{,}575.$$

Висновок щодо наявності чи відсутності зв'язку зробимо, розрахувавши статистику

$$t = \frac{\hat{r}\sqrt{N-2}}{\sqrt{1-\hat{r}^2}} = \frac{0{,}575 \cdot \sqrt{15-2}}{\sqrt{1-0{,}575^2}} = 2{,}534,$$

$$|t| = 2{,}534.$$

Беручи $\alpha = 0{,}05$ і враховуючи, що $\nu = 15 - 2 = 13$, з табл. Д.2 матимемо $t_{1-\alpha/2,\nu} = 2{,}16$. У даному разі слушна нерівність $|t| > t_{1-\alpha/2,\nu}$. Отже, коефіцієнт кореляції Пірсона відмінний від нуля і між умістом сульфат-іонів та іонів кальцію в підземних водах існує лінійний зв'язок, причому цей зв'язок додатний, оскільки додатна оцінка коефіцієнта кореляції. Одержаний висновок узгоджується з висновком, зробленим на основі аналізу кореляційного поля.

Межі 95% довірчого інтервалу для коефіцієнта є такі:

$$r_{\text{н}} = 0{,}575 + \frac{0{,}575 \cdot \left(1 - 0{,}575^2\right)}{2 \cdot 15} - 1{,}96 \cdot \frac{1 - 0{,}575^2}{\sqrt{15-1}} = 0{,}237,$$

$$r_{\text{в}} = 0{,}575 + \frac{0{,}575 \cdot \left(1 - 0{,}575^2\right)}{2 \cdot 15} + 1{,}96 \cdot \frac{1 - 0{,}575^2}{\sqrt{15-1}} = 0{,}938.$$

Зауваження 5. Модуль коефіцієнта не може перевищувати 1. З огляду на це якщо під час розрахунків верхня межа довірчого інтервалу для коефіцієнта виявляється більшою за 1, то її беруть такою, що дорівнює 1, тобто $r_{\text{в}} = 1$, а коли нижня межа виявляється меншою за –1, то беруть $r_{\text{н}} = -1$.

Отже, у розглядуваному прикладі оцінка коефіцієнта кореляції Пірсона, тобто його точкове наближення, дорівнює 0,575. Інтервал $[0{,}237; 0{,}938]$ з імовірністю 0,95 містить справжнє значення коефіцієнта. Ці значення свідчать про наявність додатного лінійного зв'язку між досліджуваними показниками.

Практична робота 9
ОЦІНЮВАННЯ КОЕФІЦІЄНТІВ РАНГОВОЇ КОРЕЛЯЦІЇ СПІРМЕНА І КЕНДАЛЛА

До коефіцієнтів рангової кореляції Спірмена та Кендалла вдаються в тому випадку, коли розподіл показника x чи y відмінний від нормального. На їх основі виявляють монотонну залежність між показниками. Попередньо масив даних $\left\{x_i, y_i; i = \overline{1, N}\right\}$ переформовують у масив рангів

$$\left\{rx_i, ry_i; i = \overline{1, N}\right\},$$

де rx_i, ry_i – ранги елементів x та y відповідно.

При цьому масив рангів слід упорядкувати за зростанням рангів rx.

Нагадаємо, що ранг – це порядковий номер елемента у впорядкованій за зростанням послідовності. Якщо у впорядкованій послідовності є однакові елементи, то говорять, що вони утворюють в'язку. Елементам в'язки присвоюють однаковий ранг, що дорівнює середньому арифметичному їх порядкових номерів.

Коефіцієнт рангової кореляції Спірмена τ_c має такі властивості:

1) $\tau_c \in [-1; 1]$;

2) за $\tau_c = 0$ монотонна залежність між показниками відсутня;

3) $\tau_c = 1$, якщо $rx_i = ry_i$ для всіх $i = \overline{1, N}$, тобто ранги елементів x та y повністю узгоджені; $\tau_c = -1$, якщо має місце протилежне впорядкування послідовностей рангів.

Коефіцієнт Спірмена дорівнює коефіцієнту кореляції Пірсона, розрахованому за масивом рангів:

$$\tau_c = r\left(rx, ry\right).$$

Оцінку коефіцієнта $\hat{\tau}_c$ можна одержати безпосередньо за цим визначенням або із застосуванням нижчеподаних виразів.

Якщо немає зв'язаних рангів, то слушна формула

$$\hat{\tau}_c = 1 - \frac{6}{N\left(N^2-1\right)} \sum_{i=1}^{N} \left(rx_i - ry_i\right)^2 .$$

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

$$\hat{\tau}_c = \frac{\frac{1}{6}N\left(N^2-1\right) - \sum_{i=1}^{N}\left(rx_i - ry_i\right)^2 - A - B}{\sqrt{\left(\frac{1}{6}N\left(N^2-1\right) - 2A\right)\left(\frac{1}{6}N\left(N^2-1\right) - 2B\right)}},$$

де

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

z – кількість в'язок між рангами rx; A_j – кількість однакових значень x у j-й в'язці;

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

k – кількість в'язок між рангами ry; B_k – кількість однакових значень y у k-й в'язці.

Висновок щодо відсутності/наявності монотонної залежності роблять на основі статистики

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

Якщо $|t| \le t_{1-\alpha/2,\nu}$, вважають, що коефіцієнт кореляції Спірмена дорівнює нулю (незначущий), між показниками відсутня монотонна залежність. Коли $|t| > t_{1-\alpha/2,\nu}$, говорять, що коефіцієнт відмінний від нуля (значущий), між показниками існує монотонний зв'язок, при цьому знак коефіцієнта вказує на те, додатний чи від'ємний зв'язок.

Коефіцієнт рангової кореляції Кендалла τ_k має такі самі властивості, що й коефіцієнт Спірмена. Його оцінку за відсутності в'язок обраховують згідно з виразом

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

де

$$S = \sum_{i=1}^{N-1}\nu_i ; \qquad \nu_i = \sum_{j=i+1}^{N}\nu_{i,j} ;$$

$$\nu_{i,j} = \begin{cases} 1, & \text{якщо } ry_i < ry_j, \\ -1, & \text{якщо } ry_i > ry_j. \end{cases}$$

За наявності зв'язаних рангів оцінку $\hat{\tau}_k$ обчислюють у такий спосіб:

$$\hat{\tau}_k = \frac{S}{\sqrt{\left(\frac{1}{2}N(N-1) - C\right)\left(\frac{1}{2}N(N-1) - D\right)}},$$

де

$$S = \sum_{i=1}^{N-1} \nu_i ; \qquad \nu_i = \sum_{j=i+1}^{N} \nu_{i,j} ;$$

$$\nu_{i,j} = \begin{cases} 1, & \text{якщо } ry_i < ry_j \text{ та } rx_i \neq rx_j, \\ -1, & \text{якщо } ry_i > ry_j \text{ та } rx_i \neq rx_j, \\ 0 & \text{в інших випадках;} \end{cases}$$

$$C = \frac{1}{2}\sum_{j=1}^{z} A_j\left(A_j - 1\right); \qquad D = \frac{1}{2}\sum_{k=1}^{p} B_k\left(B_k - 1\right).$$

Для встановлення факту відсутності/наявності залежності визначають статистику

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

Якщо $|u| \le u_{1-\alpha/2}$, то коефіцієнт кореляції вважають таким, що дорівнює нулю, і стверджують, що монотонної залежності між показниками немає. В іншому випадку говорять, що коефіцієнт відмінний від нуля й існує монотонний зв'язок.

Слід відзначити, що завжди для одних і тих же масивів $|\hat{\tau}_c| > |\hat{\tau}_k|$, а у випадку досить великого N

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

Приклад 10. Задано вибірку з попередньої практичної роботи (див. табл. 16). Оцінимо наявність монотонної залежності між умістом сульфат-іонів (x) та іонів кальцію (y) на основі коефіцієнтів рангової кореляції Спірмена і Кендалла.

Розглянемо спочатку окремо вибірку x і присвоїмо її елементам ранги, потім присвоїмо ранги елементам вибірки y (табл. 17).

Таблиця 17

Ранги елементів вибірок *x* та *y*

Упорядковані значення x	Порядкові номери	Ранги x (rx)	Упорядковані значення y	Порядкові номери	Ранги y (ry)
1 618,4	1	1	138,2	1	1,5
1 637,4	2	2	138,2	2	1,5

Закінчення табл. 17

Упорядковані значення x	Порядкові номери	Ранги x (rx)	Упорядковані значення y	Порядкові номери	Ранги y (ry)
1 666,2	3	3	145,3	3	3
1 758,3	4	4	151,8	4	4
1 888,3	5	5	156,4	5	5
1 911,4	6	6	167	6	6
1 917,1	7	7	170,3	7	7
1 938,3	8	8	172	8	8
1 964	9	9	173,6	9	9
1 983,8	10	10	174,5	10	10
2 149,2	11	11	188,5	11	11
2 254,5	12	12	190,4	12	12
2 361,4	13	13	191,3	13	13
2 437,1	14	14	191,6	14	14
2 675,7	15	15	222,6	15	15

Присвоїмо визначені ранги елементам початкової вибірки, у такий спосіб одержимо масив рангів (табл. 18).

Таблиця 18

Масив рангів

Початкова вибірка		Масив рангів	
x	y	rx	ry
2 675,7	190,4	15	12
2 437,1	156,4	14	5
1 938,3	170,3	8	7
2 149,2	174,5	11	10
2 254,5	191,3	12	13
1 964,0	188,5	9	11
1 911,4	167,0	6	6
1 888,3	191,6	5	14
1 637,4	145,3	2	3
1 666,2	138,2	3	1,5
1 618,4	151,8	1	4
2 361,4	222,6	13	15
1 983,8	172,0	10	8
1 917,1	138,2	7	1,5
1 758,3	173,6	4	9

Масив рангів упорядкуємо за зростанням рангів rx (табл. 19).

Таблиця 19

Масив рангів, упорядкований за зростанням рангів *rx*

rx	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ry	4	3	1,5	9	14	6	1,5	7	11	8	10	13	15	5	12

За цим масивом розрахуємо коефіцієнт рангової кореляції Спірмена (табл. 20).

Таблиця 20

Допоміжні розрахунки для оцінювання коефіцієнта рангової кореляції Спірмена

rx	ry	$rx-ry$	$(rx-ry)^2$
1	4	–3	9
2	3	–1	1
3	1,5	1,5	2,25
4	9	–5	25
5	14	–9	81
6	6	0	0
7	1,5	5,5	30,25
8	7	1	1
9	11	–2	4
10	8	2	4
11	10	1	1
12	13	–1	1
13	15	–2	4
14	5	9	81
15	12	3	9
Разом			$\sum = 253,5$

За масивом y має місце одна в'язка з двох елементів, тому

$$p=1,\ B_1=2,\qquad B=\frac{1}{12}\cdot\left(2^3-2\right)=\frac{1}{2}.$$

Тоді оцінку коефіцієнта Спірмена обчислимо за уточненою формулою:

$$\hat{\tau}_c=\frac{\frac{1}{6}\cdot 15\cdot 224-\sum_{i=1}^{15}\left(rx_i-ry_i\right)^2-\frac{1}{2}}{\sqrt{\left(\frac{1}{6}\cdot 15\cdot 224\right)\left(\frac{1}{6}\cdot 15\cdot 224-1\right)}}=\frac{560-253,5-0,5}{559,5}=0,547.$$

Щоб зробити висновок стосовно відсутності/наявності зв'язку, обчислимо

$$t=\frac{\hat{\tau}_c\sqrt{N-2}}{\sqrt{1-\hat{\tau}_c^2}}=\frac{0,547\cdot\sqrt{15-2}}{\sqrt{1-0,547^2}}=2,356,\quad |t|=2,356.$$

Задавши $\alpha=0,05$ і враховуючи, що $\nu=15-2=13$, з табл. Д.2 матимемо $t_{1-\alpha/2,\nu}=2,16$. Слушність нерівності $|t|>t_{1-\alpha/2,\nu}$ свідчить, що коефіцієнт кореляції Спірмена відмінний від нуля і між показниками x та y (вмістом сульфат-іонів та іонів кальцію) існує монотонний зв'язок, причому він додатний, оскільки додатна оцінка коефіцієнта.

Розрахуємо тепер оцінку коефіцієнта рангової кореляції Кендалла. Для цього обчислимо для кожного значення *ry* величину *v*. Для її визначення потрібно порівняти поточне значення *ry* з усіма значеннями, яким відповідає вищий ранг *rx*. Якщо поточне значення менше, додамо 1 до *v*, якщо більше, додамо –1, якщо вони рівні, то 0. При цьому якщо під час порівняння поточного значення *ry* з тим, якому відповідає вищий ранг *rx*, виявимо рівність відповідних *rx*, то незважаючи на те як співвідносяться ранги *ry*, до *v* додамо 0 (табл. 21).

Таблиця 21

Допоміжні розрахунки для оцінювання коефіцієнта рангової кореляції Кендалла

rx	*ry*	Розрахунок *v*
1	4	$v_1 = -1-1+1+1+1-1+1+1+1+1+1+1+1+1=8$
2	3	$v_2 = -1+1+1+1-1+1+1+1+1+1+1+1+1=9$
3	1,5	$v_3 = 1+1+1+0+1+1+1+1+1+1+1+1+1=11$
4	9	$v_4 = 1-1-1-1+1-1+1+1+1-1+1=1$
5	14	$v_5 = -1-1-1-1-1-1-1+1-1-1=-8$
6	6	$v_6 = -1+1+1+1+1+1+1-1+1=5$
7	1,5	$v_7 = 1+1+1+1+1+1+1+1=8$
8	7	$v_8 = 1+1+1+1+1-1+1=5$
9	11	$v_9 = -1-1+1+1-1+1=0$
10	8	$v_{10} = 1+1+1-1+1=3$
11	10	$v_{11} = 1+1-1+1=2$
12	13	$v_{12} = 1-1-1=-1$
13	15	$v_{13} = -1-1=-2$
14	5	$v_{14} = 1$
15	12	
Разом		$\sum = 42$

У даному разі також скористаємося уточненою формулою, оскільки мають місце в’язки. При цьому $D = \frac{1}{2} \cdot \left(2^2 - 2\right) = 1.$

Оцінка коефіцієнта Кендалла становить

$$\hat{\tau}_k = \frac{42}{\sqrt{\left(0{,}5 \cdot 15 \cdot 14\right)\left(0{,}5 \cdot 15 \cdot 14 - 1\right)}} = \frac{42}{104{,}5} = 0{,}402 .$$

Це значення менше за оцінку коефіцієнта кореляції Спірмена, як і має бути згідно з теорією. Щоб зробити висновок стосовно відсутності/наявності зв’язку, обчислимо статистику

$$u = \frac{3\hat{\tau}_k\sqrt{N(N-1)}}{\sqrt{2(2N+5)}} = \frac{3\cdot 0{,}402\cdot\sqrt{15\cdot(15-1)}}{\sqrt{2\cdot(2\cdot 15+5)}} = 2{,}088, \quad |u| = 2{,}088.$$

Узявши $\alpha = 0{,}05$, матимемо $u_{1-\alpha/2} = 1{,}96$. У цьому випадку слушна нерівність $|u| > u_{1-\alpha/2}$. Відтак робимо висновок, що коефіцієнт кореляції Кендалла відмінний від нуля і між показниками *x* та *y* (вмістом сульфат-іонів та іонів кальцію) існує монотонний зв'язок, причому цей зв'язок додатний, оскільки додатна оцінка коефіцієнта.

Отже, обидва коефіцієнти рангової кореляції, Спірмена і Кендалла, свідчать про наявність додатної монотонної залежності між показниками. Вигляд кореляційного поля вказував, що залежність лінійна. З огляду на те що лінійна залежність монотонна, одержані результати коректні.

Практична робота 10
ОЦІНЮВАННЯ ПАРАМЕТРІВ ЛІНІЙНОЇ РЕГРЕСІЇ

Якщо під час аналізу кореляційного поля та коефіцієнтів кореляції встановлено, що між показниками *x* та *y* існує лінійний зв'язок, подальше завдання можна звести до відновлення лінійної регресії.

Регресією називають залежність середніх значень показника *y* від значень показника *x*, виражену рівнянням

$$\bar{y}(x) = f(x). \qquad (1)$$

Функцію *f*, що описує цю залежність, називають **функцією регресії**. Якщо вона лінійна відносно параметрів, то говорять про лінійну регресію:

$$\bar{y}(x) = a + bx.$$

Графічне відображення рівняння (1) називають **лінією регресії** (рис. 12).

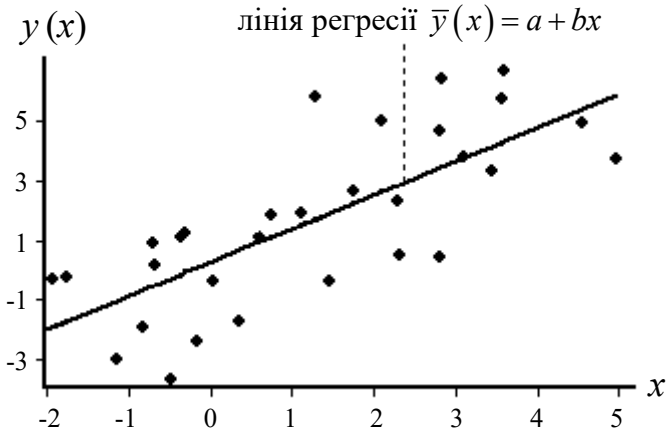


Рис. 12. Кореляційне поле та лінія регресії

Залежність фактичних значень показника *y* від значень показника *x* можна виразити рівнянням

$$y(x) = f(x) + \varepsilon,$$

а у випадку лінійної регресії – рівнянням

$$y(x) = a + bx + \varepsilon,$$

де ε – це залишок (випадкова помилка), стосовно якого припускають, що для кожного значення x він розподілений за нормальним законом з нульовим математичним сподіванням та однаковою дисперсією.

Відновлення лінійної регресії проводять за такою схемою:

1) знаходження точкових та інтервальних оцінок параметрів регресії;
2) перевірка параметрів регресії на рівність нулю (незначущість);
3) накладання довірчих інтервалів на регресію та прогнозні значення;
4) перевірка значущості та адекватності відновленої регресії.

Перші два кроки схеми становлять предмет розгляду цієї практичної роботи, третій та четвертий буде розглянуто в наступних роботах.

Для **знаходження оцінок параметрів** регресії традиційно застосовують **метод найменших квадратів**, відповідно до якого шукають такі оцінки параметрів, за яких розсіювання точок кореляційного поля відносно відновленої регресії було б мінімальним. Формально оцінки параметрів знаходять з умови мінімуму **дисперсії залишків**

$$S_\varepsilon^2 = \frac{1}{N-2}\sum_{i=1}^{N}\varepsilon_i^2 = \frac{1}{N-2}\sum_{i=1}^{N}\left(y_i - \hat{\bar{y}}(x_i)\right)^2 = \frac{1}{N-2}\sum_{i=1}^{N}\left(y_i - \hat{a} - \hat{b}x_i\right)^2.$$

Залишки ε_i дорівнюють різниці між фактичними значеннями залежної змінної y та значеннями, обчисленими на основі відновленої регресії (рис. 13).

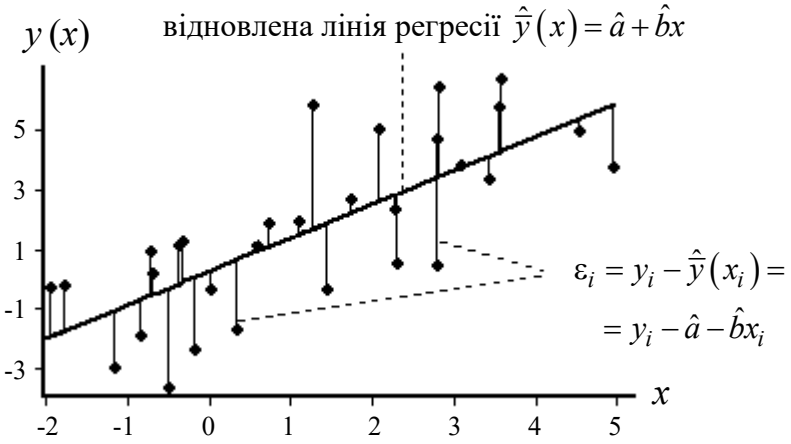


Рис. 13. Графічне зображення залишків

Оцінки параметрів лінійної регресії, за яких дисперсія залишків мінімальна, обчислюють за формулами

$$\hat{b} = \hat{r}\frac{\hat{S}_y}{\hat{S}_x}, \qquad \hat{a} = \bar{y} - \hat{b}\bar{x}.$$

Інтервальні оцінки параметрів a та b знаходять як інтервали $[a_н; a_в]$ і $[b_н; b_в]$ відповідно, що з імовірністю $1-\alpha$ містять справжні значення параметрів. Межі цих інтервалів розраховують таким чином:

$$a_{\text{н}} = \hat{a} - t_{1-\alpha/2,\nu} \cdot \sigma\{\hat{a}\}, \qquad a_{\text{в}} = \hat{a} + t_{1-\alpha/2,\nu} \cdot \sigma\{\hat{a}\};$$
$$b_{\text{н}} = \hat{b} - t_{1-\alpha/2,\nu} \cdot \sigma\{\hat{b}\}, \qquad b_{\text{в}} = \hat{b} + t_{1-\alpha/2,\nu} \cdot \sigma\{\hat{b}\},$$

де $t_{1-\alpha/2,\nu}$ – квантиль розподілу Стьюдента за $\nu = N-2$ (табл. Д.2); $\sigma\{\hat{a}\}$, $\sigma\{\hat{b}\}$ – середньоквадратичні відхилення оцінок $\hat{a}$, $\hat{b}$ відповідно, які знаходять за формулами

$$\sigma\{\hat{a}\} = \sqrt{D\{\hat{a}\}} = \sqrt{\frac{S_\varepsilon^2}{N}\left(1 + \frac{\overline{x}^2}{\hat{S}_x^2}\right)};$$
$$\sigma\{\hat{b}\} = \sqrt{D\{\hat{b}\}} = \sqrt{\frac{S_\varepsilon^2}{N\hat{S}_x^2}}.$$

Щоб перевірити, **чи параметри регресії незначущі**, висувають гіпотези про їх рівність нулю:

$$\text{H}_0 : a = 0, \qquad \text{H}_0 : b = 0.$$

Для перевірки гіпотези щодо параметра *a* визначають статистику

$$t_a = \frac{\hat{a}}{\sigma\{\hat{a}\}},$$

яку порівнюють з квантилем розподілу Стьюдента $t_{1-\alpha/2,\nu}$ за $\nu = N-2$. Якщо слушна нерівність $|t_a| \le t_{1-\alpha/2,\nu}$, то параметр *a* вважають незначущим (таким, що дорівнює нулю). У такому разі можна стверджувати, що справжня лінія регресії проходить через початок координат. Коли $|t_a| > t_{1-\alpha/2,\nu}$, параметр *a* значущий (відмінний від нуля).

Перевірку параметра *b* на рівність нулю проводять на основі статистики

$$t_b = \frac{\hat{b}}{\sigma\{\hat{b}\}}.$$

Якщо $|t_b| \le t_{1-\alpha/2,\nu}$, говорять, що параметр *b* незначущий (дорівнює нулю). Це свідчить про незначущість відновленої регресії. У разі, коли $|t_b| > t_{1-\alpha/2,\nu}$, параметр *b* регресії вважають значущим (відмінним від нуля) і стверджують, що регресія значуща.

Приклад 11. За даними з практичної роботи 8 відновимо лінійну регресію з метою спрогнозувати вміст іонів кальцію (y) на основі вмісту сульфат-іонів (x) у підземних водах на території гірничо-збагачувального комбінату. При цьому будемо використовувати обчислені в роботі 8 оцінки $\overline{x}$, $\overline{y}$, $\hat{S}_x$, $\hat{S}_y$, $\hat{r}$.

Обрахуємо оцінки параметрів регресії:

$$\hat{b} = \hat{r}\frac{\hat{S}_y}{\hat{S}_x} = 0{,}575 \cdot \frac{22{,}523}{299{,}226} = 0{,}043\ 28;$$

$$\hat{a} = \overline{y} - \hat{b}\overline{x} = 171,447 - 0,043\ 28 \cdot 2\ 010,74 = 84,422 .$$

Використовуючи знайдені оцінки, побудуємо графік відновленої регресії $\hat{\hat{y}}(x) = 84,422 + 0,043\ 28x$ і відобразимо його разом із кореляційним полем (рис. 14).

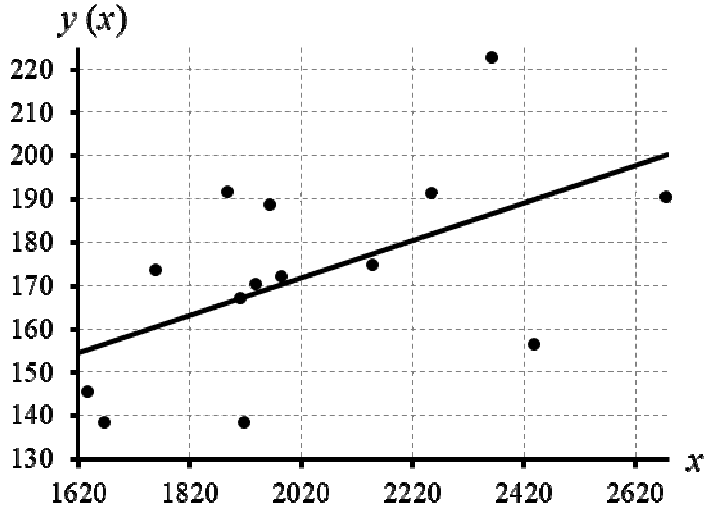


Рис. 14. Кореляційне поле та відновлена лінія регресії

Тепер визначимо інтервальні оцінки параметрів регресії та перевіримо гіпотезу про рівність параметрів нулю.

Для початку обчислимо в кожній точці x значення відновленої функції регресії, величину залишку і на їх основі розрахуємо дисперсію залишків (табл. 22).

Таблиця 22

Допоміжні розрахунки для визначення дисперсії залишків

x	y	$\hat{\hat{y}}(x) = \hat{a} + \hat{b}x$	$\varepsilon = y - \hat{\hat{y}}(x)$	ε^2
2 675,7	190,4	200,226 3	–9,826 3	96,556 1
2 437,1	156,4	189,899 7	–33,499 7	1 122,229 1
1 938,3	170,3	168,311 6	1,988 4	3,953 6
2 149,2	174,5	177,439 4	–2,939 4	8,639 9
2 254,5	191,3	181,996 8	9,303 2	86,550 3
1 964	188,5	169,423 9	19,076 1	363,896 8
1 911,4	167	167,147 4	–0,147 4	0,021 7
1 888,3	191,6	166,147 6	25,452 4	647,823 4
1 637,4	145,3	155,288 7	–9,988 7	99,773 6
1 666,2	138,2	156,535 1	–18,335 1	336,177 2
1 618,4	151,8	154,466 4	–2,666 4	7,109 4
2 361,4	222,6	186,623 4	35,976 6	1 294,316 3
1 983,8	172	170,280 9	1,719 1	2,955 4
1 917,1	138,2	167,394 1	–29,194 1	852,294 8
1 758,3	173,6	160,521 2	13,078 8	171,054 4
Разом				$\sum = 5\ 093,352$

Дисперсія залишків дорівнює

$$S_\varepsilon^2 = \frac{1}{N-2}\sum_{i=1}^{N}\varepsilon_i^2 = \frac{5\ 093{,}352}{15-2} = 391{,}796.$$

Розрахуємо тепер середньоквадратичні відхилення для оцінок:

$$\sigma\{\hat{a}\} = \sqrt{D\{\hat{a}\}} = \sqrt{\frac{S_\varepsilon^2}{N}\left(1+\frac{\overline{x}^2}{\hat{S}_x^2}\right)} = \sqrt{\frac{391{,}796}{15}\cdot\left(1+\frac{2\ 010{,}74^2}{299{,}226^2}\right)} = 34{,}721,$$

$$\sigma\left\{\hat{b}\right\} = \sqrt{D\left\{\hat{b}\right\}} = \sqrt{\frac{S_\varepsilon^2}{N\hat{S}_x^2}} = \sqrt{\frac{391{,}796}{15\cdot 299{,}226^2}} = 0{,}017\ 08.$$

На їх основі визначимо межі довірчих інтервалів $[a_н; a_в]$, $[b_н; b_в]$ для параметрів a та b відповідно (як і в попередніх роботах, $\alpha = 0{,}05$, $t_{1-\alpha/2,13} = 2{,}16$):

$$a_н = \hat{a} - t_{1-\alpha/2,\nu}\cdot\sigma\{\hat{a}\} = 84{,}422 - 2{,}16\cdot 34{,}721 = 9{,}425;$$

$$a_в = \hat{a} + t_{1-\alpha/2,\nu}\cdot\sigma\{\hat{a}\} = 84{,}422 + 2{,}16\cdot 34{,}721 = 159{,}419;$$

$$b_н = \hat{b} - t_{1-\alpha/2,\nu}\cdot\sigma\left\{\hat{b}\right\} = 0{,}043\ 28 - 2{,}16\cdot 0{,}017\ 08 = 0{,}006\ 38;$$

$$b_в = \hat{b} + t_{1-\alpha/2,\nu}\cdot\sigma\left\{\hat{b}\right\} = 0{,}043\ 28 + 2{,}16\cdot 0{,}017\ 08 = 0{,}080\ 17.$$

Перевіримо параметри регресії на рівність нулю. Для параметра a матимемо таке:

$$t_a = \frac{\hat{a}}{\sigma\{\hat{a}\}} = \frac{84{,}422}{34{,}721} = 2{,}431;$$

оскільки $|t_a| \le 2{,}16$, параметр a відмінний від нуля.

Для параметра b слушний такий вираз:

$$t_b = \frac{\hat{b}}{\sigma\left\{\hat{b}\right\}} = \frac{0{,}04328}{0{,}01708} = 2{,}534;$$

як бачимо, $|t_b| > 2{,}16$, справжнє значення b на рівні значущості 0,05 відмінне від нуля.

Остаточні результати зведемо в табл. 23.

Таблиця 23

Оцінки параметрів регресії з результатами перевірки їх значущості

Параметр регресії	Оцінка	Середньо-квадратичне відхилення оцінки	95% довірчий інтервал	Статистика t	Квантиль	Значущість параметра
a	84,422	34,721	[9,425; 159,419]	2,431	2,16	$\neq 0$
b	0,043 28	0,0170 8	[0,006 38; 0,080 17]	2,534	2,16	$\neq 0$

Точкова оцінка параметра a дорівнює 84,422, хоча справжнє значення параметра може лежати в досить широких межах [9,425; 159,419]. Параметр a

на рівні значущості 0,05 відмінний від нуля, отже, справжня лінія регресії не проходить через початок координат.

Наближене значення параметра b дорівнює 0,043 28, а справжнє може лежати в межах [0,006 38; 0,080 17]. Параметр b на рівні значущості 0,05 відмінний від нуля, тобто регресія значуща.

Практична робота 11
ДОВІРЧЕ ОЦІНЮВАННЯ РЕГРЕСІЇ ТА ПРОГНОЗНИХ ЗНАЧЕНЬ

Нехай знайдено оцінки параметрів регресії і, отже, визначено оцінку регресійного рівняння, а саме

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

На її основі прогнозують, яким буде середнє значення показника y за відомого значення показника x. У такий спосіб одержують наближене значення середнього, справжнє середнє залишається невідомим. Але для нього можна встановити довірчі межі. З цією метою будують **довірчий інтервал на регресію** $\left[\bar{y}_{\text{н}}(x); \bar{y}_{\text{в}}(x)\right]$, який з імовірністю $1-\alpha$ містить справжнє середнє значення показника y. Його межі розраховують за формулами

$$\begin{aligned} \bar{y}_{\text{н}}(x) &= \hat{\bar{y}}(x) - t_{1-\alpha/2,\nu} \cdot \sigma\left\{\hat{\bar{y}}(x)\right\}, \\ \bar{y}_{\text{в}}(x) &= \hat{\bar{y}}(x) + t_{1-\alpha/2,\nu} \cdot \sigma\left\{\hat{\bar{y}}(x)\right\}, \end{aligned} \tag{2}$$

де

$$\sigma\left\{\hat{\bar{y}}(x)\right\} = \sqrt{\frac{S_\varepsilon^2}{N} + \left(\sigma\left\{\hat{b}\right\} \cdot (x - \bar{x})\right)^2};$$

$\nu = N - 2$, $t_{1-\alpha/2,\nu}$ – квантиль розподілу Стьюдента (табл. Д.2).

Межі довірчого інтервалу $\left[\bar{y}_{\text{н}}(x); \bar{y}_{\text{в}}(x)\right]$ можна розрахувати для всіх точок вибірки і тоді графічно зобразити його разом з лінією регресії (рис. 15, *а*). Слід наголосити, що цей довірчий інтервал не є паралельний лінії регресії, його межі завжди розходяться за віддалення x від $\bar{x}$.

Якщо хочуть спрогнозувати, не яким за конкретного x буде середнє значення y, а яке взагалі можна спостерігати значення y, то будують **довірчий інтервал на прогнозне значення** $\left[y_{\text{н}}(x); y_{\text{в}}(x)\right]$. Він з імовірністю $1-\alpha$ містить справжнє значення показника y. Його межі визначають у такий спосіб:

$$\begin{aligned} y_{\text{н}}(x) &= \hat{\bar{y}}(x) - t_{1-\alpha/2,\nu} \cdot \sigma\left\{\hat{y}(x)\right\}, \\ y_{\text{в}}(x) &= \hat{\bar{y}}(x) + t_{1-\alpha/2,\nu} \cdot \sigma\left\{\hat{y}(x)\right\}, \end{aligned} \tag{3}$$

де

$$\sigma\left\{\hat{y}(x)\right\} = \sqrt{\left(\sigma\left\{\hat{\bar{y}}(x)\right\}\right)^2 + S_\varepsilon^2}.$$

Довірчий інтервал на прогнозне значення можна подати графічно (рис. 15, *б*). Він завжди ширший за інтервал на регресію і містить приблизно $(1-\alpha)\cdot 100\%$ точок вибірки. Довірчий інтервал на прогнозне значення також називають **толерантними межами**.

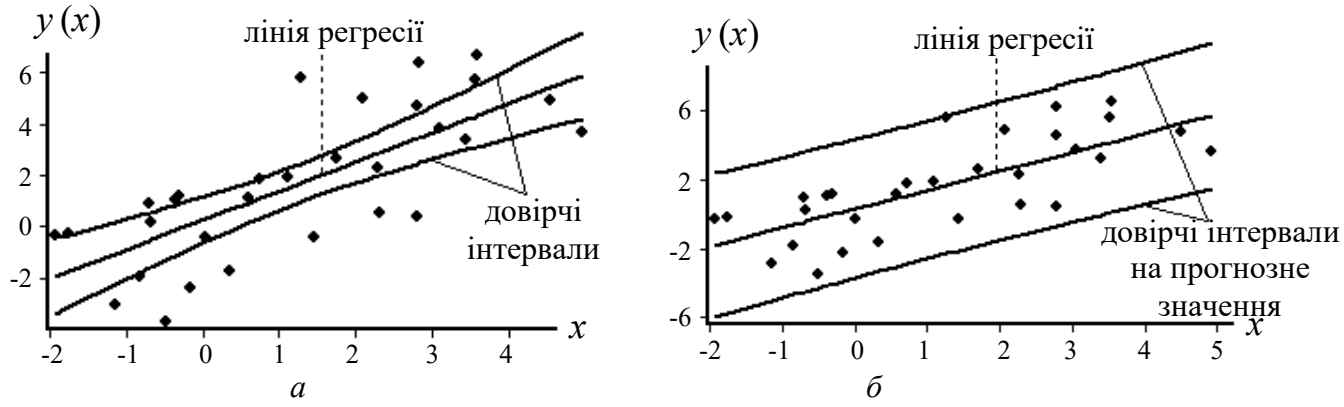


Рис. 15. Графічне зображення довірчого інтервалу:
а – на регресію; *б* – на прогнозне значення

Слід відзначити, що за досить великого обсягу вибірки значення $\sigma\{\hat{y}(x)\}$ у виразі (3) можна розрахувати за спрощеною формулою:

$$\sigma\{\hat{y}(x)\} = \sqrt{S_\varepsilon^2}\ .$$

Але в практичній роботі обсяг вибірки замалий для її застосування, тому потрібно послуговуватися точною формулою.

Для наочності нижче довірчі інтервали на регресію та прогнозне значення наведено на одному графіку (рис. 16).

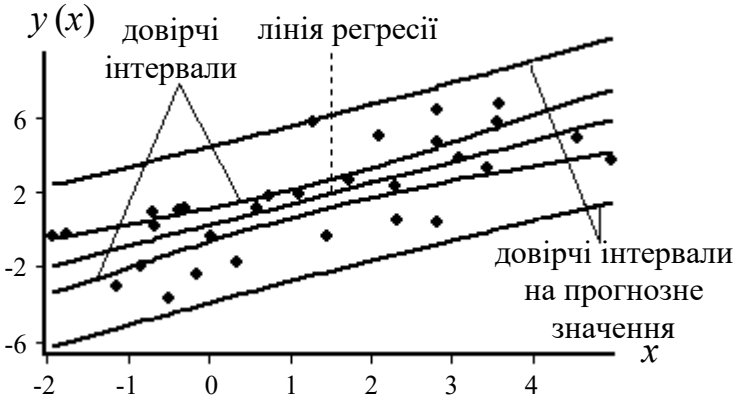


Рис. 16. Графічне зображення довірчих інтервалів на регресію та прогнозне значення

Приклад 12. Побудуємо довірчі інтервали на регресію та прогнозне значення. Візьмемо $\alpha = 0{,}05$, тоді $t_{1-\alpha/2,\nu} = 2{,}16$ ($\nu = N-2 = 15-2 = 13$).

У кожній точці x обчислимо межі 95% довірчого інтервалу $\left[\overline{y}_{\text{н}}(x); \overline{y}_{\text{в}}(x)\right]$, послуговуючись формулами (2), а також межі 95% довірчого інтервалу $\left[y_{\text{н}}(x); y_{\text{в}}(x)\right]$, застосовуючи формули (3) (табл. 24).

Таблиця 24

Розрахунок 95% довірчих інтервалів на регресію та прогнозне значення

Початкові дані		Значення регресії $\hat{\bar{y}}(x) = \hat{a} + \hat{b}x$	$\sigma\{\hat{\bar{y}}(x)\}$	95% довірчий інтервал на регресію		$\sigma\{\hat{y}(x)\}$	95% довірчий інтервал на прогнозне значення	
x	y			$\bar{y}_н(x)$	$\bar{y}_в(x)$		$y_н(x)$	$y_в(x)$
2 675,7	190,4	200,226	12,454	173,325	227,128	23,386	149,712	250,740
2 437,1	156,4	189,900	8,897	170,683	209,116	21,701	143,025	236,775
1 938,3	170,3	168,312	5,258	156,954	179,670	20,480	124,074	212,549
2 149,2	174,5	177,439	5,631	165,276	189,603	20,579	132,988	221,891
2 254,5	191,3	181,997	6,592	167,758	196,235	20,863	136,933	227,060
1 964	188,5	169,424	5,173	158,251	180,597	20,459	125,233	213,614
1 911,4	167	167,147	5,385	155,516	178,779	20,513	122,839	211,456
1 888,3	191,6	166,148	5,522	154,220	178,075	20,550	121,760	210,535
1 637,4	145,3	155,289	8,172	137,637	172,940	21,414	109,034	201,544
1 666,2	138,2	156,535	7,794	139,700	173,371	21,273	110,585	202,485
1 618,4	151,8	154,466	8,428	136,263	172,670	21,513	107,998	200,935
2 361,4	222,6	186,623	7,873	169,617	203,630	21,302	140,610	232,636
1 983,8	172	170,281	5,131	159,197	181,365	20,448	126,113	214,449
1 917,1	138,2	167,394	5,355	155,827	178,961	20,505	123,102	211,686
1 758,3	173,6	160,521	6,687	146,078	174,964	20,893	115,393	205,650

Відобразимо ці інтервали на графіку разом з лінією регресії та кореляційним полем (рис. 17).

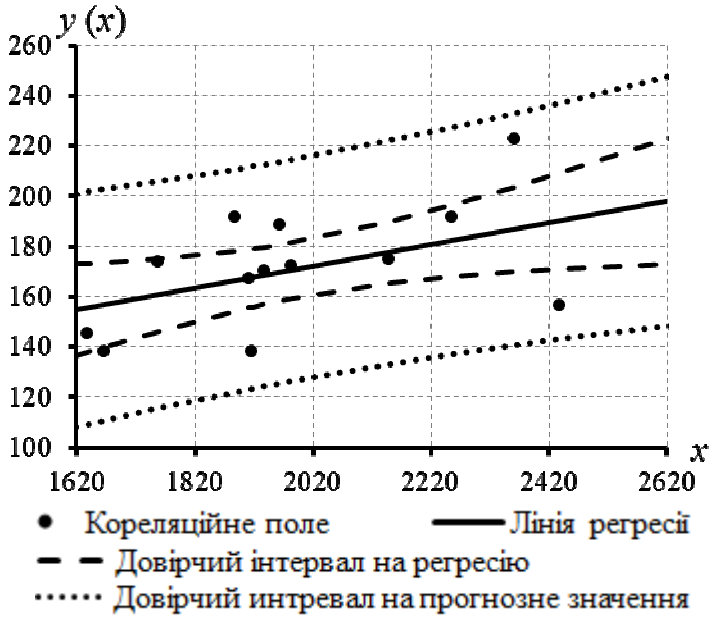


Рис. 17. Кореляційне поле та лінія регресії разом з 95% довірчими інтервалами

Зауваження 6. У розглядуваному випадку майже всі точки кореляційного поля потрапили в межі обох інтервалів. Це обумовлено дуже малим обсягом

вибірки. На практиці коли обсяг вибірки більший, інтервал на регресію значно вужчий і до нього потрапляє мала частина точок, а в інтервал на прогнозні значення потрапляє близько 95% точок.

Розглянуті довірчі інтервали застосовують на практиці таким чином. Нехай під час чергових спостережень на території гірничо-збагачувального комбінату було виявлено вміст сульфат-іонів у підземних водах, що складав 2 000 мг/л. Тоді, застосовуючи рівняння регресії

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

можна сказати, що наближене середнє значення показника y (вмісту іонів кальцію) становить

$$\hat{\bar{y}}(2\ 000) = 84,422 + 0,043\ 28 \cdot 2\ 000 = 170,982.$$

У цій точці можна накласти 95% довірчий інтервал на регресійне значення:

$$\sigma\left\{\hat{\bar{y}}(2\ 000)\right\} = \sqrt{\frac{S_\varepsilon^2}{N} + \left(\sigma\left\{\hat{b}\right\} \cdot (2\ 000 - \bar{x})\right)^2} =$$

$$= \sqrt{\frac{391,796}{15} + \left(0,017\ 08 \cdot (2\ 000 - 2\ 010,74)\right)^2} = 5,114;$$

$$\bar{y}_{\text{н}}(2\ 000) = \hat{\bar{y}}(2\ 000) - t_{1-\alpha/2,\nu} \cdot \sigma\left\{\hat{\bar{y}}(2\ 000)\right\} = 170,982 - 2,16 \cdot 5,114 = 159,936,$$

$$\bar{y}_{\text{в}}(2\ 000) = \hat{\bar{y}}(2\ 000) + t_{1-\alpha/2,\nu} \cdot \sigma\left\{\hat{\bar{y}}(2\ 000)\right\} = 170,982 + 2,16 \cdot 5,114 = 182,028.$$

Отже, в точці $x = 2\ 000$ довірчий інтервал на регресію є $[159,936;182,028]$.

Також можна визначити 95% довірчий інтервал на прогнозне значення:

$$\sigma\left\{\hat{y}(2\ 000)\right\} = \sqrt{\left(\sigma\left\{\hat{\bar{y}}(2\ 000)\right\}\right)^2 + S_\varepsilon^2} = \sqrt{5,114^2 + 391,796} = 20,444;$$

$$y_{\text{н}}(2\ 000) = \hat{\bar{y}}(2\ 000) - t_{1-\alpha/2,\nu} \cdot \sigma\left\{\hat{y}(2\ 000)\right\} = 170,982 - 2,16 \cdot 20,444 = 126,823,$$

$$y_{\text{в}}(2\ 000) = \hat{\bar{y}}(2\ 000) + t_{1-\alpha/2,\nu} \cdot \sigma\left\{\hat{y}(2\ 000)\right\} = 170,982 + 2,16 \cdot 20,444 = 215,141.$$

Довірчий інтервал на прогнозне значення за $x = 2\ 000$ є $[126,823;215,141]$.

Таким чином, коли вміст сульфат-іонів у підземних водах дорівнює 2 000, наближене середнє значення вмісту іонів кальцію становить 170,982. З імовірністю 0,95 довірчий інтервал $[159,936;182,028]$ містить справжнє середнє значення вмісту іонів калію, а довірчий інтервал $[126,823;215,141]$ – справжнє прогнозне значення.

Практична робота 12
ПЕРЕВІРКА ЗНАЧУЩОСТІ ТА АДЕКВАТНОСТІ РЕГРЕСІЇ

У випадку коли маємо справу з відновленою регресією, постають такі питання:

1) чи є регресія значуща (чи пояснює вона хоч би деякою мірою зміни показника *y*);

2) чи є регресія адекватна (наскільки добре вона пояснює зміни показника *y*).

Щоб дати відповідь на ці питання, застосовують *F*-тест, обчислюють коефіцієнт детермінації та аналізують діагностичну діаграму.

За допомогою ***F*-тесту** перевіряють, чи є регресія значуща. Для цього розраховують статистику

$$F = \frac{\sum_{i=1}^{N}\left(\hat{\bar{y}}(x_i) - \bar{y}\right)^2}{S_\varepsilon^2},$$

яку порівнюють з квантилем розподілу Фішера $f_{1-\alpha,\nu_1,\nu_2}$, беручи $\nu_1 = 1$, $\nu_2 = N - s$. Величина s дорівнює кількості параметрів, від яких залежить регресія. Для лінійної регресії $s = 2$. Значення квантиля знаходять за спеціальною таблицею (табл. Д.3).

Якщо $F \le f_{1-\alpha,\nu_1,\nu_2}$, регресію вважають незначущою. Якщо ж $F > f_{1-\alpha,\nu_1,\nu_2}$, роблять висновок про значущість регресії.

Слід відзначити, що у випадку лінійної регресії *F*-тест еквівалентний перевірці гіпотези про рівність нулю параметра *b* регресії і слушні такі співвідношення: $F = t_b^2$, $f_{1-\alpha,\nu_1,\nu_2} = t_{1-\alpha,\nu_2}^2$.

Якщо за допомогою *F*-тесту встановлено значущість регресії, це ще не свідчить про її адекватність, оскільки вона може пояснювати досить малу частину змін показника *y*. Зробити висновок про адекватність регресії можна на основі коефіцієнта детермінації.

Коефіцієнт детермінації R^2 – показник, що визначає, якою мірою регресія пояснює варіабельність показника *y*. Значення коефіцієнта детермінації обчислюють шляхом піднесення до квадрата оцінки коефіцієнта кореляції Пірсона:

$$R^2 = \hat{r}^2 \cdot 100\%.$$

У такий спосіб коефіцієнт детермінації розраховують, якщо відновлюють лінійну регресію. Загалом його можна визначити за формулою

$$R^2 = \left(1 - \frac{(N-s)S_\varepsilon^2}{N\hat{S}_y^2}\right) \cdot 100\%.$$

Зрозуміло, що $R^2 \in [0;100]$. Чим більше значення R^2, тим більш адекватна регресія.

Про адекватність регресії також можна зробити висновок,

проаналізувавши діагностичну діаграму.

Діагностична діаграма – це точковий графік, на якому за віссю абсцис відкладають значення регресії $\hat{\bar{y}}(x_i)$, а за віссю ординат – залишки, тобто величини $\varepsilon_i = y_i - \hat{\bar{y}}(x_i)$. Залишки являють собою «непояснені» похибки оцінювання y. З аналізу діагностичної діаграми можуть випливати такі висновки. Якщо поле залишків колоподібне чи овалоподібне без кутового нахилу (рис. 18, *а*, *б*), то обрано адекватну модель регресії. Якщо ж виявлено окремі кластери або форма поля діаграми відмінна від зазначеної (рис. 18, *в, г*), дослідник має зосередитися на пошуку причин неадекватності відновленої регресії.

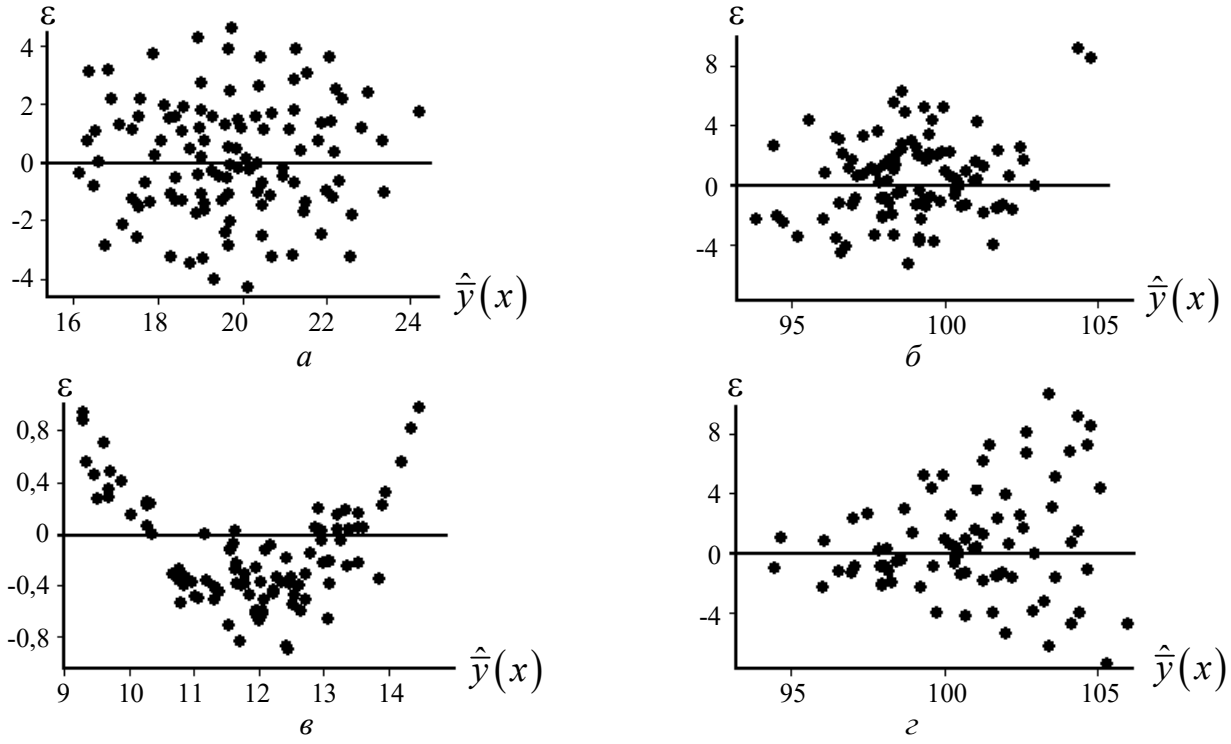


Рис. 18. Діагностичні діаграми:
а – регресія адекватна; *б* – регресія адекватна, але в даних є 2 аномальні значення;
в – регресія неадекватна, доцільно відновити іншу модель;
г – дисперсія залишків не є постійна

Приклад 13. Перевіримо, чи є відновлена у практичній роботі 10 лінійна регресія значуща та адекватна.

Розрахуємо статистику *F*-тесту, скориставшись даними табл. 22 та розрахованим у роботі 8 значенням $\bar{y} = 171{,}447$. Значення статистики дорівнює

$$F = \frac{\sum_{i=1}^{N}\left(\hat{\bar{y}}(x_i) - \bar{y}\right)^2}{S_\varepsilon^2} = \frac{2\ 515{,}728}{391{,}796} = 6{,}421.$$

Узявши $\alpha = 0{,}05$ і врахувавши, що $\nu_1 = 1$, $\nu_2 = 13$, з табл. Д.3 матимемо $f_{1-\alpha,\nu_1,\nu_2} = 4{,}67$. Отже, $F > f_{1-\alpha,\nu_1,\nu_2}$ і лінійна регресія значуща.

Щоб перевірити правильність зроблених розрахунків, повернемося до результатів практичної роботи 10, під час виконання якої було знайдено $t_b = 2{,}534$. Має бути правдива рівність $F = t_b^2$. У розглядуваному випадку це так: $6{,}421 = 2{,}534^2$.

Щоб встановити, чи є регресія адекватна, обчислимо коефіцієнт детермінації:

$$R^2 = \hat{r}^2 \cdot 100\% = 0{,}575^2 \cdot 100\% = 33{,}1\%.$$

Для контролю розрахуємо R^2 і за іншою формулою:

$$R^2 = \left(1 - \frac{S_\varepsilon^2 (N - s)}{\hat{S}_y^2 N}\right) \cdot 100\% = \left(1 - \frac{391{,}796 \cdot (15 - 2)}{22{,}523^2 \cdot 15}\right) \cdot 100\% = 33{,}1\%,$$

звідси бачимо, що результати збігаються. Зауважимо, що коефіцієнт детермінації на рівні 33,1% не можна назвати високим.

Щоб переконатися в адекватності регресії, побудуємо також діагностичну діаграму (рис. 19). Значення $\hat{\hat{y}}(x)$ та ε, за якими треба її побудувати, вже розраховано в практичній роботі 10 (див. табл. 22).

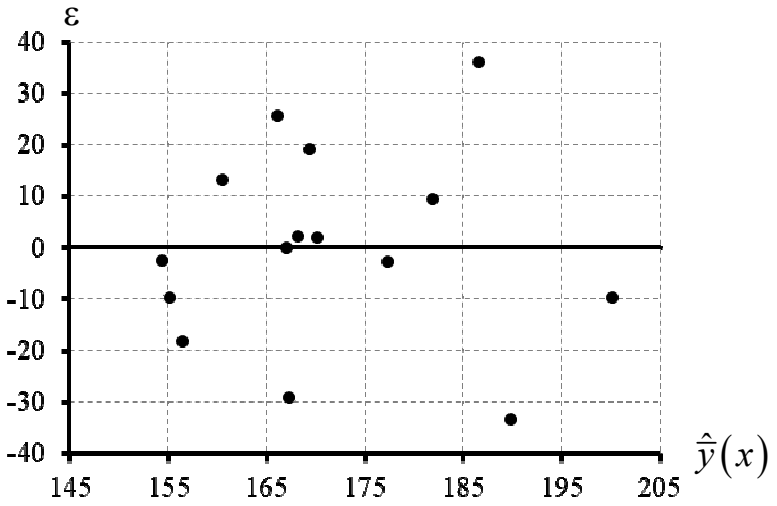


Рис. 19. Діагностична діаграма

Точки на діагностичній діаграмі утворюють овал без нахилу, що свідчить про адекватність лінійної регресії. Враховуючи вигляд кореляційного поля та діагностичної діаграми, можемо вважати, що відновлена лінійна регресія адекватна, а дещо відносна адекватність відповідно до коефіцієнта детермінації пов'язана з досить великими залишками ε.

Контрольні запитання

1. Які задачі розв'зують за допомогою кореляційного та регресійного аналізу?

2. У який спосіб будують кореляційне поле? Які висновки роблять на його основі?

3. Яку залежність між показниками x та y ідентифікують у разі кореляційного поля, зображеного на рис. 20?

Рис. 20. Приклад кореляційного поля

4. Наявність яких залежностей оцінюють за допомогою коефіцієнта кореляції Пірсона?

5. Які властивості має коефіцієнт кореляції Пірсона?

6. За якою формулою обчислюють оцінку коефіцієнта кореляції Пірсона?

7. Які залежності оцінюють рангові коефіцієнти кореляції Спірмена і Кендалла?

8. Які властивості мають рангові коефіцієнти кореляції Спірмена і Кендалла?

9. Яким співвідношенням пов'язані між собою коефіцієнти кореляції Спірмена і Кендалла за великого обсягу вибірки?

10. Яким чином пов'язані між собою коефіцієнти кореляції Пірсона і Спірмена?

11. Яку залежність називають регресією?

12. Який метод застосовують для знаходження оцінок параметрів лінійної регресії?

13. Що являють собою залишки лінійної регресії?

14. За якою формулою обраховують дисперсію залишків?

15. У який спосіб перевіряють значущість параметрів регресії? Про що свідчить значущість параметра b лінійної регресії $\bar{y}(x) = a + bx$?

16. Які відмінності довірчого інтервалу на регресію та довірчого інтервалу на прогнозне значення?

17. З якою метою застосовують F-тест?

18. Як обчислюють коефіцієнт детермінації і що він характеризує?

19. У який спосіб будують діагностичну діаграму? Які висновки роблять на її основі?

Список рекомендованої літератури

Айвазян, С.А. Прикладная статистика. Основы моделирования и первичная обработка данных [Текст]: справ. изд. / С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин. – М.: Финансы и статистика, 1983. – 471 с.

Большев, Л.Н. Таблицы математической статистики [Текст] / Л.Н. Большев, Н.В. Смирнов. – М.: Наука, 1965. – 464 с.

Дэвис, Дж. С. Статистический анализ данных в геологии [Текст]: пер. с англ.: в 2 кн. Кн. 1. / Дж. С. Дэвис. – М.: Недра, 1990. – 319 с.

Ивченко, Г.И. Математическая статистика [Текст] / Г.И. Ивченко, Ю.И. Медведев. – М.: Высш. шк., 1984. – 248 с.

Кендалл, М. Стастические выводы и связи [Текст] / М. Кендалл, А. Стьюарт. – М.: Наука, 1973. – 900 с.

Крамер, Г. Математические методы статистики [Текст] / Г. Крамер. – М.: Мир, 1975. – 648 с.

Лагутин, М.Б. Наглядная математическая статистика [Текст]: учеб. пособие / М.Б. Лагутин. – М.: БИНОМ: Лаб. знаний, 2013. – 472 с.

Павловский, З. Введение в математическую статистику [Текст] / З. Павловский. – М.: Статистика, 1967. – 284 с.

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

Сигел, Э. Практическая бизнес-статистика [Текст]: пер. с англ. / Э. Сигел. – 4-е изд. – М.: Издат. дом «Вильямс», 2002. – 1 056 с.

Статистична обробка даних [Текст] / В.П. Бабак, А.Я. Білецький, О.П. Приставка, П.О. Приставка. – К.: МІВВЦ, 2001. – 388 с.

Хан, Г. Статистические модели в инженерных задачах [Текст] / Г. Хан, С. Шапиро. – М.: Мир, 1969. – 395 с.

Чини, Р.Ф. Статистические методы в геологии [Текст]: пер. с англ. / Р.Ф. Чини. – М.: Мир, 1986. – 189 с.

Додаток
ТАБЛИЦІ КВАНТИЛІВ

Таблиця 1

Квантилі стандартного нормального розподілу u_p

$a \setminus b$	–0,09	–0,08	–0,07	–0,06	–0,05	–0,04	–0,03	–0,02	–0,01	0,00
–3,6	0,000 1	0,000 1	0,000 1	0,000 1	0,000 1	0,000 1	0,000 1	0,000 1	0,000 2	0,000 2
–3,5	0,000 2	0,000 2	0,000 2	0,000 2	0,000 2	0,000 2	0,000 2	0,000 2	0,000 2	0,000 2
–3,4	0,000 2	0,000 3	0,000 3	0,000 3	0,000 3	0,000 3	0,000 3	0,000 3	0,000 3	0,000 3
–3,3	0,000 3	0,000 4	0,000 4	0,000 4	0,000 4	0,000 4	0,000 4	0,000 5	0,000 5	0,000 5
–3,2	0,000 5	0,000 5	0,000 5	0,000 6	0,000 6	0,000 6	0,000 6	0,000 6	0,000 7	0,000 7
–3,1	0,000 7	0,000 7	0,000 8	0,000 8	0,000 8	0,000 8	0,000 9	0,000 9	0,000 9	0,001 0
–3,0	0,001 0	0,001 0	0,001 1	0,001 1	0,001 1	0,001 2	0,001 2	0,001 3	0,001 3	0,001 3
–2,9	0,001 4	0,001 4	0,001 5	0,001 5	0,001 6	0,001 6	0,001 7	0,001 8	0,001 8	0,001 9
–2,8	0,001 9	0,002 0	0,002 1	0,002 1	0,002 2	0,002 3	0,002 3	0,002 4	0,002 5	0,002 6
–2,7	0,002 6	0,002 7	0,002 8	0,002 9	0,003 0	0,003 1	0,003 2	0,003 3	0,003 4	0,003 5
–2,6	0,003 6	0,003 7	0,003 8	0,003 9	0,004 0	0,004 1	0,004 3	0,004 4	0,004 5	0,004 7
–2,5	0,004 8	0,004 9	0,005 1	0,005 2	0,005 4	0,005 5	0,005 7	0,005 9	0,006 0	0,006 2
–2,4	0,006 4	0,006 6	0,006 8	0,006 9	0,007 1	0,007 3	0,007 5	0,007 8	0,008 0	0,008 2
–2,3	0,008 4	0,008 7	0,008 9	0,009 1	0,009 4	0,009 6	0,009 9	0,010 2	0,010 4	0,010 7
–2,2	0,011 0	0,011 3	0,011 6	0,011 9	0,012 2	0,012 5	0,012 9	0,013 2	0,013 6	0,013 9
–2,1	0,014 3	0,014 6	0,015 0	0,015 4	0,015 8	0,016 2	0,016 6	0,017 0	0,017 4	0,017 9
–2,0	0,018 3	0,018 8	0,019 2	0,019 7	0,020 2	0,020 7	0,021 2	0,021 7	0,022 2	0,022 8
–1,9	0,023 3	0,023 9	0,024 4	0,025 0	0,025 6	0,026 2	0,026 8	0,027 4	0,028 1	0,028 7
–1,8	0,029 4	0,030 1	0,030 7	0,031 4	0,032 2	0,032 9	0,033 6	0,034 4	0,035 1	0,035 9
–1,7	0,036 7	0,037 5	0,038 4	0,039 2	0,040 1	0,040 9	0,041 8	0,042 7	0,043 6	0,044 6
–1,6	0,045 5	0,046 5	0,047 5	0,048 5	0,049 5	0,050 5	0,051 6	0,052 6	0,053 7	0,054 8
–1,5	0,055 9	0,057 1	0,058 2	0,059 4	0,060 6	0,061 8	0,063 0	0,064 3	0,065 5	0,066 8
–1,4	0,068 1	0,069 4	0,070 8	0,072 1	0,073 5	0,074 9	0,076 4	0,077 8	0,079 3	0,080 8
–1,3	0,082 3	0,083 8	0,085 3	0,086 9	0,088 5	0,090 1	0,091 8	0,093 4	0,095 1	0,096 8
–1,2	0,098 5	0,100 3	0,102 0	0,103 8	0,105 6	0,107 5	0,109 3	0,111 2	0,113 1	0,115 1
–1,1	0,117 0	0,119 0	0,121 0	0,123 0	0,125 1	0,127 1	0,129 2	0,131 4	0,133 5	0,135 7
–1,0	0,137 9	0,140 1	0,142 3	0,144 6	0,146 9	0,149 2	0,151 5	0,153 9	0,156 2	0,158 7
–0,9	0,161 1	0,163 5	0,166 0	0,168 5	0,171 1	0,173 6	0,176 2	0,178 8	0,181 4	0,184 1
–0,8	0,186 7	0,189 4	0,192 2	0,194 9	0,197 7	0,200 5	0,203 3	0,206 1	0,209 0	0,211 9
–0,7	0,214 8	0,217 7	0,220 6	0,223 6	0,226 6	0,229 6	0,232 7	0,235 8	0,238 9	0,242 0
–0,6	0,245 1	0,248 3	0,251 4	0,254 6	0,257 8	0,261 1	0,264 3	0,267 6	0,270 9	0,274 3
–0,5	0,277 6	0,281 0	0,284 3	0,287 7	0,291 2	0,294 6	0,298 1	0,301 5	0,305 0	0,308 5
–0,4	0,312 1	0,315 6	0,319 2	0,322 8	0,326 4	0,330 0	0,333 6	0,337 2	0,340 9	0,344 6
–0,3	0,348 3	0,352 0	0,355 7	0,359 4	0,363 2	0,366 9	0,370 7	0,374 5	0,378 3	0,382 1
–0,2	0,385 9	0,389 7	0,393 6	0,397 4	0,401 3	0,405 2	0,409 0	0,412 9	0,416 8	0,420 7
–0,1	0,424 7	0,428 6	0,432 5	0,436 4	0,440 4	0,444 3	0,448 3	0,452 2	0,456 2	0,460 2
0,0	0,464 1	0,468 1	0,472 1	0,476 1	0,480 1	0,484 0	0,488 0	0,492 0	0,496 0	0,500 0

Закінчення табл. 1

$a \setminus b$	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,500 0	0,504 0	0,508 0	0,512 0	0,516 0	0,519 9	0,523 9	0,527 9	0,531 9	0,535 9
0,1	0,539 8	0,543 8	0,547 8	0,551 7	0,555 7	0,559 6	0,563 6	0,567 5	0,571 4	0,575 3
0,2	0,579 3	0,583 2	0,587 1	0,591 0	0,594 8	0,598 7	0,602 6	0,606 4	0,610 3	0,614 1
0,3	0,617 9	0,621 7	0,625 5	0,629 3	0,633 1	0,636 8	0,640 6	0,644 3	0,648 0	0,651 7
0,4	0,655 4	0,659 1	0,662 8	0,666 4	0,670 0	0,673 6	0,677 2	0,680 8	0,684 4	0,687 9
0,5	0,691 5	0,695 0	0,698 5	0,701 9	0,705 4	0,708 8	0,712 3	0,715 7	0,719 0	0,722 4
0,6	0,725 7	0,729 1	0,732 4	0,735 7	0,738 9	0,742 2	0,745 4	0,748 6	0,751 7	0,754 9
0,7	0,758 0	0,761 1	0,764 2	0,767 3	0,770 4	0,773 4	0,776 4	0,779 4	0,782 3	0,785 2
0,8	0,788 1	0,791 0	0,793 9	0,796 7	0,799 5	0,802 3	0,805 1	0,807 8	0,810 6	0,813 3
0,9	0,815 9	0,818 6	0,821 2	0,823 8	0,826 4	0,828 9	0,831 5	0,834 0	0,836 5	0,838 9
1,0	0,841 3	0,843 8	0,846 1	0,848 5	0,850 8	0,853 1	0,855 4	0,857 7	0,859 9	0,862 1
1,1	0,864 3	0,866 5	0,868 6	0,870 8	0,872 9	0,874 9	0,877 0	0,879 0	0,881 0	0,883 0
1,2	0,884 9	0,886 9	0,888 8	0,890 7	0,892 5	0,894 4	0,896 2	0,898 0	0,899 7	0,901 5
1,3	0,903 2	0,904 9	0,906 6	0,908 2	0,909 9	0,911 5	0,913 1	0,914 7	0,916 2	0,917 7
1,4	0,919 2	0,920 7	0,922 2	0,923 6	0,925 1	0,926 5	0,927 9	0,929 2	0,930 6	0,931 9
1,5	0,933 2	0,934 5	0,935 7	0,937 0	0,938 2	0,939 4	0,940 6	0,941 8	0,942 9	0,944 1
1,6	0,945 2	0,946 3	0,947 4	0,948 4	0,949 5	0,950 5	0,951 5	0,952 5	0,953 5	0,954 5
1,7	0,955 4	0,956 4	0,957 3	0,958 2	0,959 1	0,959 9	0,960 8	0,961 6	0,962 5	0,963 3
1,8	0,964 1	0,964 9	0,965 6	0,966 4	0,967 1	0,967 8	0,968 6	0,969 3	0,969 9	0,970 6
1,9	0,971 3	0,971 9	0,972 6	0,973 2	0,973 8	0,974 4	0,975 0	0,975 6	0,976 1	0,976 7
2,0	0,977 2	0,977 8	0,978 3	0,978 8	0,979 3	0,979 8	0,980 3	0,980 8	0,981 2	0,981 7
2,1	0,982 1	0,982 6	0,983 0	0,983 4	0,983 8	0,984 2	0,984 6	0,985 0	0,985 4	0,985 7
2,2	0,986 1	0,986 4	0,986 8	0,987 1	0,987 5	0,987 8	0,988 1	0,988 4	0,988 7	0,989 0
2,3	0,989 3	0,989 6	0,989 8	0,990 1	0,990 4	0,990 6	0,990 9	0,991 1	0,991 3	0,991 6
2,4	0,991 8	0,992 0	0,992 2	0,992 5	0,992 7	0,992 9	0,993 1	0,993 2	0,993 4	0,993 6
2,5	0,993 8	0,994 0	0,994 1	0,994 3	0,994 5	0,994 6	0,994 8	0,994 9	0,995 1	0,995 2
2,6	0,995 3	0,995 5	0,995 6	0,995 7	0,995 9	0,996 0	0,996 1	0,996 2	0,996 3	0,996 4
2,7	0,996 5	0,996 6	0,996 7	0,996 8	0,996 9	0,997 0	0,997 1	0,997 2	0,997 3	0,997 4
2,8	0,997 4	0,997 5	0,997 6	0,997 7	0,997 7	0,997 8	0,997 9	0,997 9	0,998 0	0,998 1
2,9	0,998 1	0,998 2	0,998 2	0,998 3	0,998 4	0,998 4	0,998 5	0,998 5	0,998 6	0,998 6
3,0	0,998 7	0,998 7	0,998 7	0,998 8	0,998 8	0,998 9	0,998 9	0,998 9	0,999 0	0,999 0
3,1	0,999 0	0,999 1	0,999 1	0,999 1	0,999 2	0,999 2	0,999 2	0,999 2	0,999 3	0,999 3
3,2	0,999 3	0,999 3	0,999 4	0,999 4	0,999 4	0,999 4	0,999 4	0,999 5	0,999 5	0,999 5
3,3	0,999 5	0,999 5	0,999 5	0,999 6	0,999 6	0,999 6	0,999 6	0,999 6	0,999 6	0,999 7
3,4	0,999 7	0,999 7	0,999 7	0,999 7	0,999 7	0,999 7	0,999 7	0,999 7	0,999 7	0,999 8
3,5	0,999 8	0,999 8	0,999 8	0,999 8	0,999 8	0,999 8	0,999 8	0,999 8	0,999 8	0,999 8
3,6	0,999 8	0,999 8	0,999 9	0,999 9	0,999 9	0,999 9	0,999 9	0,999 9	0,999 9	0,999 9

Примітка. У табл. 1 потрібно відшукати значення величини p, тоді квантиль u_p буде дорівнювати сумі значень з першого стовпця (a) та першого рядка (b): $u_p = a + b$.

Таблиця 2

Квантилі розподілу Стьюдента $t_{1-\alpha/2,\nu}$ ($\alpha = 0{,}05$)

ν	$t_{1-\alpha/2,\nu}$	ν	$t_{1-\alpha/2,\nu}$
		35	2,03
1	12,7	36	2,03
2	4,30	37	2,03
3	3,18	38	2,02
4	2,78	39	2,02
5	2,57	40	2,02
6	2,45	41	2,02
7	2,36	42	2,02
8	2,31	43	2,02
9	2,26	44	2,02
10	2,23	45	2,01
11	2,20	46	2,01
12	2,18	47	2,01
13	2,16	48	2,01
14	2,14	49	2,01
15	2,13	50	2,01
16	2,12	51	2,01
17	2,11	52	2,01
18	2,10	53	2,01
19	2,09	54	2,00
20	2,09	55	2,00
21	2,08	56	2,00
22	2,07	57	2,00
23	2,07	58	2,00
24	2,06	59	2,00
25	2,06	60	2,00
26	2,06	65	2,00
27	2,05	70	1,99
28	2,05	75	1,99
29	2,05	80	1,99
30	2,04	90	1,99
31	2,04	100	1,98
32	2,04	110	1,98
33	2,03	120	1,98
34	2,03	∞	1,96

Таблиця 3

Квантилі розподілу Фішера $f_{1-\alpha,\nu_1,\nu_2}$ ($\alpha = 0{,}05$)

$\nu_2 \setminus \nu_1$	1	5	7	10	11	13	30	60	120	∞
1	161,4	230,2	236,8	241,9	243,0	244,7	250,1	252,2	253,3	254,3
2	18,51	19,30	19,35	19,40	19,40	19,42	19,46	19,48	19,49	19,50
3	10,13	9,01	8,89	8,79	8,76	8,73	8,62	8,57	8,55	8,53
4	7,71	6,26	6,09	5,96	5,94	5,89	5,75	5,69	5,66	5,63
5	6,61	5,05	4,88	4,74	4,70	4,66	4,50	4,43	4,40	4,37
6	5,99	4,39	4,21	4,06	4,03	3,98	3,81	3,74	3,70	3,67
7	5,59	3,97	3,79	3,64	3,60	3,55	3,38	3,30	3,27	3,23
8	5,32	3,69	3,50	3,35	3,31	3,26	3,08	3,01	2,97	2,93
9	5,12	3,48	3,29	3,14	3,10	3,05	2,86	2,79	2,75	2,71
10	4,96	3,33	3,14	2,98	2,94	2,89	2,70	2,62	2,58	2,54
11	4,84	3,20	3,01	2,85	2,82	2,76	2,57	2,49	2,45	2,40
12	4,75	3,11	2,91	2,75	2,72	2,66	2,47	2,38	2,34	2,30
13	4,67	3,03	2,83	2,67	2,63	2,58	2,38	2,30	2,25	2,21
14	4,60	2,96	2,76	2,60	2,57	2,51	2,31	2,22	2,18	2,13
15	4,54	2,90	2,71	2,54	2,51	2,45	2,25	2,16	2,11	2,07
16	4,49	2,85	2,66	2,49	2,46	2,40	2,19	2,11	2,06	2,01
17	4,45	2,81	2,61	2,45	2,41	2,35	2,15	2,06	2,01	1,96
18	4,41	2,77	2,58	2,41	2,37	2,31	2,11	2,02	1,97	1,92
19	4,38	2,74	2,54	2,38	2,34	2,28	2,07	1,98	1,93	1,88
20	4,35	2,71	2,51	2,35	2,31	2,25	2,04	1,95	1,90	1,84
21	4,32	2,68	2,49	2,32	2,28	2,22	2,01	1,92	1,87	1,81
22	4,30	2,66	2,46	2,30	2,26	2,20	1,98	1,89	1,84	1,78
23	4,28	2,64	2,44	2,27	2,24	2,18	1,96	1,86	1,81	1,76
24	4,26	2,62	2,42	2,25	2,22	2,15	1,94	1,84	1,79	1,73
25	4,24	2,60	2,40	2,24	2,20	2,14	1,92	1,82	1,77	1,71
26	4,23	2,59	2,39	2,22	2,18	2,12	1,90	1,80	1,75	1,69
27	4,21	2,57	2,37	2,20	2,17	2,10	1,88	1,79	1,73	1,67
28	4,20	2,56	2,36	2,19	2,15	2,09	1,87	1,77	1,71	1,65
29	4,18	2,55	2,35	2,18	2,14	2,08	1,85	1,75	1,70	1,64
30	4,17	2,53	2,33	2,16	2,13	2,06	1,84	1,74	1,68	1,62
40	4,08	2,45	2,25	2,08	2,04	1,97	1,74	1,64	1,58	1,51
60	4,00	2,37	2,17	1,99	1,95	1,89	1,65	1,53	1,47	1,39
120	3,92	2,29	2,09	1,91	1,87	1,80	1,55	1,43	1,35	1,25
∞	3,84	2,21	2,01	1,83	1,79	1,72	1,46	1,32	1,22	1,00