МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ОДЕСЬКИЙ ДЕРЖАВНИЙ ЕКОЛОГІЧНИЙ УНІВЕРСИТЕТ

ЛОБОДА Н. С., КУЗА А. М.

МЕТОДИ МАТЕМАТИЧНОЇ СТАТИСТИКИ У ГІДРОЕКОЛОГІЧНИХ ДОСЛІДЖЕННЯХ

Конспект лекцій

Одеса
Одеський державний екологічний університет
2020

УДК 504.4

Л68

Рекомендовано методичною радою Одеського державного екологічного університету Міністерства освіти і науки України як конспект лекцій (протокол № 9 від 29.06.2017 р.)

Лобода Н.С., Куза А.М.
Методи математичної статистики у гідроекологічних дослідженнях: конспект лекцій. Одеса, ОДЕКУ, 2017. 94 с.

У конспекті лекцій розглянуті математичні моделі, їх класифікації, ціль побудови та основні завдання, які можуть бути вирішені за допомогою математичного моделювання при вирішенні гідроекологічних задач; викладені теоретичні та практичні основи стохастичного моделювання водного середовища у природних та порушених водогосподарською діяльністю умовах; нейромережеве моделювання гідроекологічних систем; детерміністичні математичні моделі.

ISBN 978-966-186-013-0

ЗМІСТ

ПЕРЕДМОВА

Конспект лекцій складений відповідно до програми курсу «Методи математичної статистики у гідроекологічних дослідженнях», який входить до складу дисциплін з підготовки спеціалістів за спеціальністю «Екологія, охорона навколишнього середовища та збалансоване природокористування», спеціалізація – «Гідроекологія» - фаховий шифр 6.040106.

Загальна мета дисципліни «Методи математичної статистики у гідроекологічних дослідженнях» полягає в забезпеченні студентів об'ємом теоретичних знань і практичних навичок, необхідних для статистичної обробки та аналізу даних і їх практичного застосування до вирішення практичних задач гідроекології.

Загальний обсяг навчального часу визначається робочим навчальним планом та становить 30 годин лекцій, 15 годин практичних занять, 60 годин самостійної роботи студентів.

В результаті вивчення дисципліни «Методи математичної статистики у гідроекологічних дослідженнях» студенти повинні знати методи аналізу однорідності та обробки рядів даних на основі параметричних та непараметричних критеріїв, методи визначення статистичних параметрів, основи теорії випадкових функцій, поняття про функціональні та статистичні залежності, основи теорії кореляції, особливості перевірки статистичних гіпотез у різних задачах.

Після вивчення дисципліни студенти повинні вміти - установлювати закономірності динаміки водності, мінералізації та забруднення річок на основі коефіцієнта галинності, виявляти статистичні «викиди», перевіряти гіпотези про існування тенденцій, надавати прогнози щодо показників якості води за допомогою статистичних методів.

Вивчення дисципліни «Методи математичної статистики у гідроекологічних дослідженнях» ґрунтується на знаннях, одержаних студентами за такими дисциплінами учбового плану – «Вища математика», «Гідрохімія», «Гідрометрія», «Гідравліка», «Гідрологія». Вивчення дисципліни потребує розуміння процесів виникнення та формування гідрологічних явищ, гідрохімічного складу вод знання й впливу на них антропогенних чинників. Знання, одержані в результаті вивчення дисципліни «Методи математичної статистики у гідроекологічних дослідженнях», будуть використовуватися у курсовому проектуванні, при написанні кваліфікаційних робіт.

1 ПОНЯТТЯ МАТЕМАТИЧНОЇ СТАТИСТИКИ

1.1 Задачі математичної статистики

Математична статистика є прикладною математичною дисципліною, спорідненою теорії ймовірностей. Задача математичної статистики полягає у тому, щоб на основі властивостей деякої підмножини (вибірки) зробити висновки про властивості усієї множини в цілому. Уся множина в цілому має назву генеральної сукупності.

Часові ряди однорідних за своїм походженням величин можна розглядати як вибірку з генеральної сукупності (статистичний ряд). При цьому до статистичного ряду ставиться вимога щодо відсутності зв'язків між її членами. Це означає, що статистична залежність елементами одного й того статистичного ряду має бути відсутньою. Ця вимога не завжди виконується через що при оцінки статистичних параметрів можуть визначатися з похибками.

Теоретичним обгрунтуванням можливості застосування статистичних методів до опису закономірностей деякого фізичного процессу на основі спостережених даних можуть слугувати ***закон великих чисел*** та ***центральна гранична теорема***, які ще мають назву граничних теорем теорії ймовірностей.

У вузькому розумінні під законом великих чисел розуміють комплекс теорем, для кожної з яких установлюється факт наближення середніх характеристик до деяких сталих. Фізичний зміст закону великих чисел зводиться до такого: індівідуальні особливості кожного випадкового явища практично не впливають на середній результат, а випадкові відхилення нівелюються. Закон великих чисел є законом сталості середніх величин. Іншими словами, закон великих чисел дозволяє зробити такі висновки: *при достатньо великій кількості спостережень отримані за вибірками статистичні характеристики достатні для опису генеральної сукупності у цілому. Відомі теореми, які відносяться до закону великих чисел, - це теореми Чебишева, Бернуллі та інш.*

Під центральною граничною теоремою розуміють групу теорем, зміст яких міститься у такому: подія, що відбувається у результаті підсумовування або визначення добутку великої кількості незалежних або малозалежних подій є випадковою і підкорюється нормальному закону розподілу. Якщо фізичне явище розглядати як результат додавання випадкових незалежних подій, то легко прийти до ствердження, що воно може представлятись випадковою величиною, яка підкорюється нормальному закону розподілу. Але аналіз вихідної інформації показує, що статистичний розподіл величин у більшості випадків не може описуватись

нормальним законом розподілу. Справа у тому, що події, які приводять до формування досліджуваних величин залежні одна від одної. В результаті виникає також зв'язок між по передніми та наступними членами рядів стоку. Структура таких часових рядів стоку може бути описаною за допомогою так званих автокореляційних функцій.

Задачі статистичних методів є ретроспективно-описувальними та прогностичними. На початку необхідно виявити закономірності формування досліджуваного явища, побудувати його математичну модель і вже за математичною моделлю виконувати прогноз, який надається в ймовірнісній формі.

Математичні моделі можна класифікувати за ступенем їх невизначеності – детерміністичні і стохастичні. Детерміністична модель не містить у собі випадкових компонент, а стохастична – містить. У більшості стохастичних моделей ряди спостережених величин розглядаються або як ***послідовность незалежних випадкових величин***, або використовується та чи інша модіфікація ***простого ланцюга Маркова.*** Ступінь залежності між членами ряду характеризується коефіцієнтом автокореляції.

1.2 Випадкова подія

Випадкова подія є подією, яка в результаті експерименту (випробування) може відбутись, а може не відбутись. Наприклад, вміст забруднювальної речовини у воді може перевищити ГДК в результаті скиду комунальних вод, а може й не перевищити. Розглядаючи випадкові події, приходимо до висновку, що кожна з них має ту або іншу можливість своєї появи: одна – більшу, друга – меншу. Для того, щоб кількісно порівняти між собою події за ступенем можливості їх появи, необхідно з кожній події поставити у відповідність число, яке тим більше, чим більш імовірна подія. Таке число має назву імовірності події. ***Імовірність події є чисельною мірою ступеня об'єктивної можливості її здійснення.***

Для співвідношення різних подій за ступенем можливості їхнього здійснення встановлюють одиницю вимірювання. Імовірність достовірної події (***достовірна подія є подією яка в результаті експерименту обов'язково має відбутися***) приймається рівною одиниці, імовірність недостовірної події (***недостовірна подія*** **є** подією, яка в даному випробуванні ніяк не може відбутися) дорівнює 0. Значення імовірностей інших подій знаходяться у інтервалі від 0 до 1. Сума ймовірностей розглядуваної випадкової події завжди дорівнює 1.

Якщо в результаті випробувань розглядається не сама подія, а її кількісна характеристика, то в теорії ймовірностей використовується *поняття* випадкової величини. ***Випадковою називається величина, яка***

внаслідок випробування набуває того чи іншого значення, наперед невідомо, якого саме.

Для статистичного опису випадкової величини використовуються закони розподілу та їх параметри. ***Законом розподілу випадкової величини називається співвідношення, яке установлює зв'язок випадковою величиною та характеристиками ймовірності її появи. Закон розподілу, представлений у виді математичного рівняння має назву теоретичного закону розподілу. Параметри цих рівнянь є статистичними параметрам.***

1.3 Дискретна випадкова величина та закон її розподілу

Якщо можливі значення випадкової величини можна перерахувати, то випадкова величина відноситься до *дискретної*, тобто вона описується кінцевим числом значень. Дискретна випадкова величина X, можливі значення якої x_1, x_2, x_3, …,x_n з імовірністної точки зору буде повністю описана, якщо кожному значенню випадкової величини поставити у відповідність значення імовірності його появи у випробуванні. Наприклад, у результататі n іспитів випадкова величина приймає n значень x_1, x_2, x_3, …,x_n з відповідними ймовірностями p_1, p_2, p_3, …, p_n . ***Законом розподілу випадкової дискретної величини*** називається співвідношення, яке установлює зв'язок між можливими значеннями випадкової величини та її ймовірностями. Такий закон розподілу називається таблицею розподілу

x_i	x_1	x_2	x_3	…	x_n
p_i	p_1	p_2	p_3	…	p_n

Сума імовірностей $\sum_{i=1}^{n} p_i = 1.$

Якщо таблицю розподілу представити у графічному вигляді (рис.1.1), то ми отримаємо багатокутник розподілу.

Закон розподілу, представлений у вигляді таблиці, або багатокутника є вичерпною імовірнісною характеристикою випадкової величини.

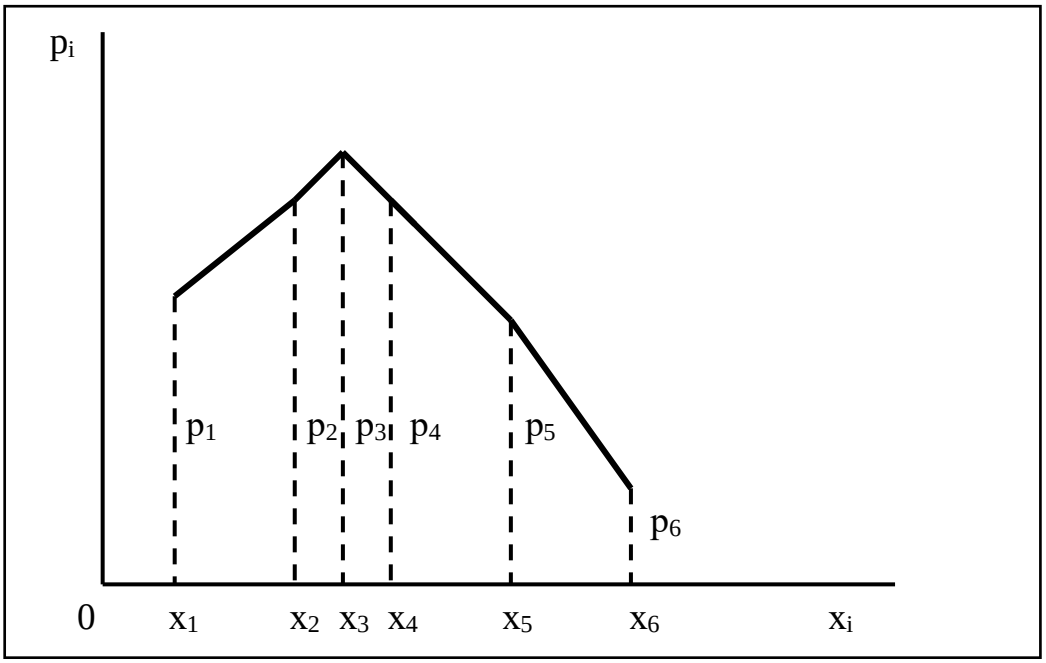


Рисунок 1.1 – Багатокутник розподілу

Визначити імовірність появи кожного значення випадкової дискретної величини у результаті n випробувань можливо, якщо представити її як відношення числа випадків, коли розглядувана величина спостерігалася до загальної кількості випробувань

$$p_i = \frac{m}{N}, \tag{1.1}$$

де p_i - імовірність появи випадкової величини;

i - порядковий номер випробування (експерименту);

m - число випадків, коли значення $X = x$ спостерігалося у результаті проведення випробування;

N - загальна кількість проведених випробувань, в результаті яких отримані різні значення випадкової величини $X = x$, які описують дискретну випадкову величину X.

1.4 Неперервна випадкова величина та закон її розподілу

Поряд з дискретними величинами існують неперервні. Випадкові величини, які заповнюють деякий проміжок на вісі x, називають *неперервними*. Кожне окреме значення неперервної випадкової величини не має ніякої відмінної від нуля ймовірності. Для того, щоб задати закон розподілу неперервної випадкової величини використовується поняття

функції розподілу $F(x) = p(X < x)$, де $F(x)$ не імовірність події X=x, а імовірність того, що X<x, де x – деяка поточна змінна. Функція розподілу F(x) має такі властивості:

- функція розподілу F(x) є монотонно неспадною, тобто при $x_2 > x_1$, $F(x_2) > F(x_1)$ (рис.1.2).
- у мінус нескінченності, коли $x \to -\infty$, $F(-\infty)=0$
- у плюс нескінченності, коли $x \to +\infty$, $F(+\infty)=1$

У гідрологічних розрахунках здебільшого використовується не інтегральна функція розподілу F(x), а функція забезпеченості $P(x)$, яка пов'язана з інтегральною функцією співвідношенням

$$P(x)=1-F(x) . \qquad (1.2)$$

Забезпеченість випадкової величини $P(x)$ x є імовірністю того, що випадкова величина X прийме значення більше деякого заданого значення x, тобто P(x)=p(X>x).

Для дискретних випадкових величин функція забезпеченості може представлятись у вигляді

$$F(x)= P(x<X) = \sum_{x_i<x} p_i (X = x_i) , \qquad (1.3)$$

де нерівність $x_i < x$ вказує, що підсумовування відбувалося не для усіх x_i, а лише для тих, що меньші ніж x.

Таким чином, функція розподілу дискретної випадкової величини x може представлятись у вигляді ступінчастої функції. Інтегральна функція $F(x) = p(X < x)$, представлена у графічному вигляді, має назву інтегральної кривої розподілу (рис.1.2, рис.1.3).

Розглянемо неперервну випадкову величину X, що має властивість монотонності. Імовірність попадання безперервної випадкової величини у ділянку на осі x, яка змінюється від x до $x+\Delta x$, розраховується таким чином

$$p(x < X < x + \Delta x) = F(x + \Delta x) - F(x) , \qquad (1.4)$$

де права частина є приростом інтегрального функції розподілу на ділянці довжиною Δx.

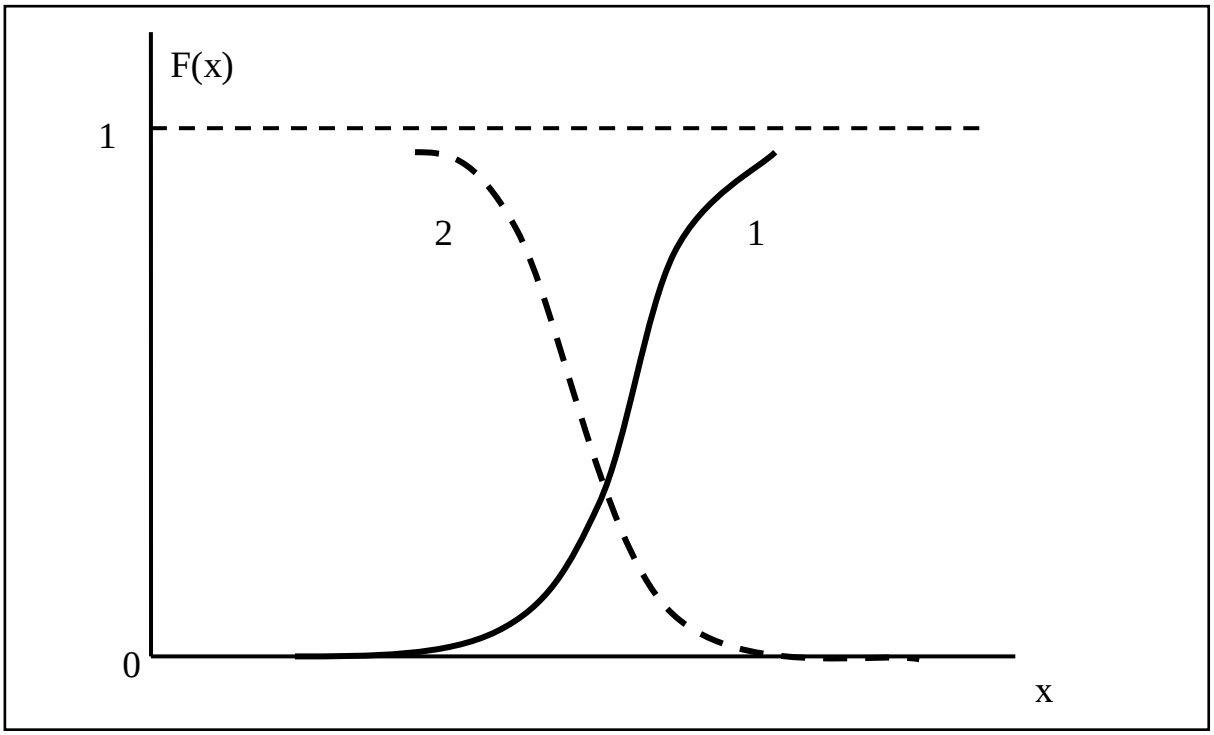


Рисунок 1.2 – Інтегральна функція розподілу F(x) випадкової величини x (1) та функція забезпеченості (2)

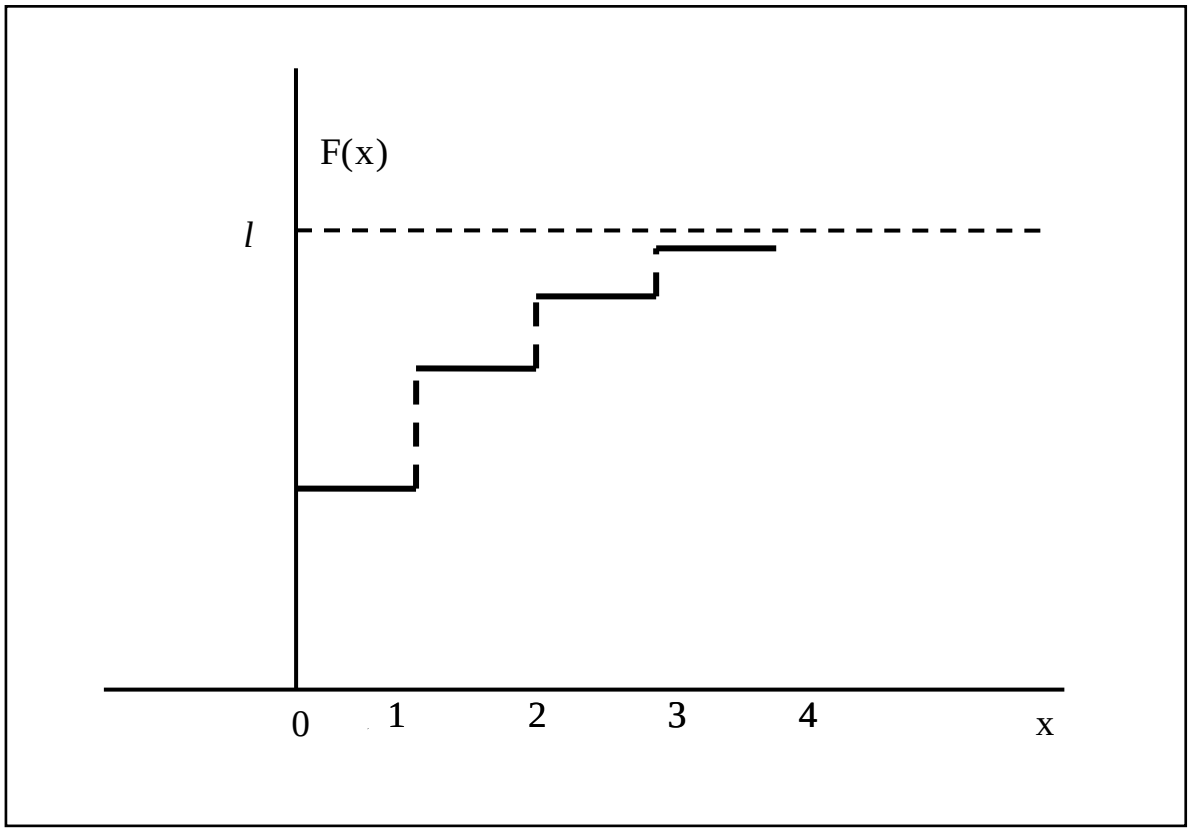


Рисунок 1.3 - Інтегральна функція розподілу F(x) випадкової величини x, визначена за дискретними даними

Якщо цей приріст поділити на Δx, то отримаємо середню імовірність. Вона ж є похідною від функції $F(x)$ в точці

$$\lim_{\Delta x \to 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} = F'(x), \qquad (1.5)$$

звідки $F'(x) = f(x)$

Похідна $F'(x)$ має назву щільності розподілу безперервної випадкової величини f(x) і називається диференційним законом розподілу випадкової величини. ***Похідна*** $F'(x) = f(x)$ ***характеризує щільність ймовірності неперервної випадкової величини*** X ***у точці*** x ***і носить назву щільності розподілу безперервної випадкової величини або диференціального закону розподілу випадкової безперервної величини.*** Ця форма закону розподілу існує тільки для неперервних випадкових величин.

Добуток $f(x)dx$ називають елементом ймовірності, який характеризує ймовірність попадання випадкової величини X на елементарну ділянку Δx. Ймовірність попадання на відрізок від α до β має дорівнювати сумі елементів ймовірності, тобто інтегралу

$$p(\alpha < X < \beta) = \int_{\alpha}^{\beta} f(x)dx. \qquad (1.6)$$

Згідно з виразом (1.6) вирішується протилежна задача визначення функції $F(x)$ через щільність $f(x)$. Якщо

$$F(x) = p(X < x) = p(-\infty < X < x), \qquad (1.7)$$

то $$F(x) = \int_{-\infty}^{+x} f(x)dx. \qquad (1.8)$$

Геометрично $F(x)$ є площею під кривою розподілу, яка розташовується зліва від точки x (рис.1.4).

Площа під кривою, що лежить лівіше перетину *x*, є функцією розподілу $F(x)$. Загальна площа під кривою розподілу дорівнює одиниці.

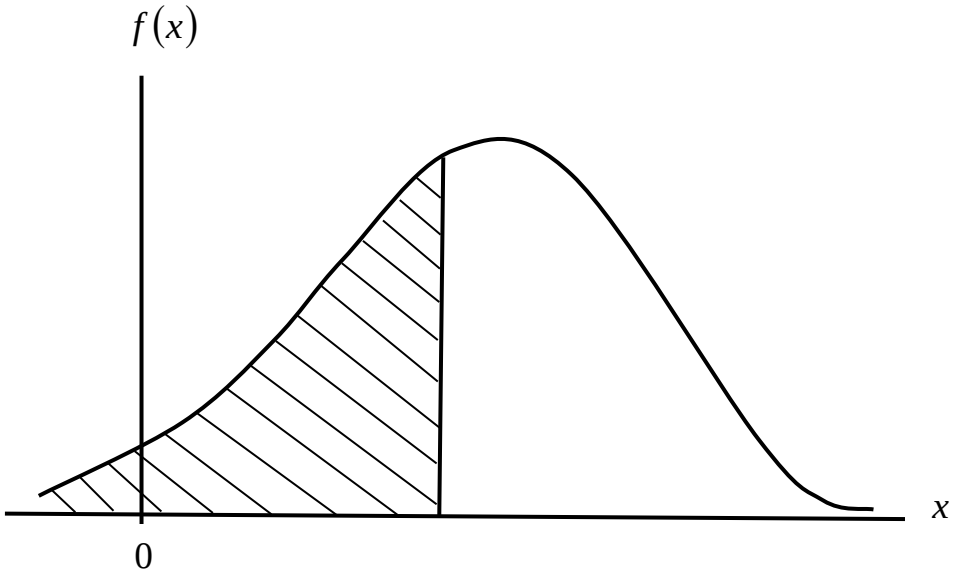


Рисунок 1.4 – Крива щільності $f(x)$ випадкової величини X

Диференціальна функція розподілу безперервної випадкової величини *x* має такі властивості:

1. Щільність розподілу є завжди невід'ємною , тобто

$$f(x) \geq 0. \tag{1.8}$$

2. Інтеграл від щільності розподілу у межах від $-\infty$ до ∞ дорівнює одиниці

$$\int_{-\infty}^{+\infty} f(x)dx = 1. \tag{1.9}$$

3. Щільність розподілу є розмірною величиною і її розмірність є зворотною розмірності величини *x*.

З геометричної точки зору перераховані властивості указують на те, що крива щільності розподілу лежить не нижче осі абсцис та повна площа, обмежена кривою $f(x)$ і віссю абсцис, дорівнює 1.

Між інтегральною, диференціальною функціями та функцією забезпеченості існують певні співвідношення, які записуються таким чином

$$p(\alpha < X < \beta) = F(\beta) - F(\alpha), \tag{1.10}$$

$$p(\alpha < X < \beta) = \int_{\alpha}^{\beta} f(x)dx, \qquad (1.11)$$

$$P(x) = 1 - \int_{-\infty}^{x} f(x)dx = 1 - F(x) = \int_{x}^{\infty} f(x)dx \ . \qquad (1.12)$$

З (1.12) витікає, що через одну форму закону розподілу можна перейти до іншої.

1.5 Статистичні параметри законів розподілу

Однією з основних задач математичної статистики є установлення ймовірнісних властивостей випадкової величини на основі визначених законів розподілу. Кожен закон розподілу випадкової величини являє собою математичну функцію, яка повністю описує випадкову величину з ймовірнісної точки зору. На практиці для такого опису використовуються окремі числові характеристики, які відображають головні риси закону розподілу.

Числові характеристики випадкової величини, які виражають у стислій формі особливості закону розподілу, називаються статистичними параметрами.

Статистичні параметри, визначені по рядах спостережень (вибірках), відрізняються від відповідних параметрів генеральної сукупності і називаються ***оцінками статистичних параметрів.***

Для оцінок статистичних параметрів на основі вибірок розроблені спеціальні статистичні методи. Найбільш універсальним є метод статистичних моментів, якій не залежить ні від якого теоретичного закону розподілу.

Кількість статистичних параметрів, які використовуються в теоретичному законі розподілу, не повинна бути великою. Світовий досвід показує, що при розрахунках стоку найбільш оптимальними для практичного застосування є такі теоретичні закони розподілу, для описування яких достатньо двох або трьох статистичних параметрів – математичного сподівання, дисперсії та залежного від неї коефіцієнта варіації, характеристики асиметрії розподілу. ***Математичне сподівання*** показує центр статистичного розподілу випадкової величини. ***Дисперсія*** характеризує розсіювання випадкової величини відносно її центра статистичного розподілу. ***Коефіцієнт варіації*** є характеристикою мінливості випадкової величини. На відміну від середнього квадратичного

відхилення коефіцієнт варіації є відносною величиною. ***Коефіцієнт асиметрії*** характеризує несиметричність (асиметричність) розподілу випадкової величини відносно математичного сподівання і може бути як від'ємним, так і додатним. Ексцес характеризує сплюснутість або витягнутість кривої розподілу випадкової величини у порівнянні з кривою нормального розподілу.

Кожен із законів розподілу випадкової величини являє собою певний математичний вираз, за яким розраховуються випадкові величини заданої забезпеченості при заданих статистичних параметрах розподілу. Найбільш відомі закони розподілу випадкових величин: нормальний, логарифмічно нормальний, Пірсона III, трьохпараметричний гама-розподіл Крицького-Менкеля. З метою прискорення розрахунків використовують не теоретичні вирази, а спеціально розроблені таблиці ординат законів розподілу. Серед методів визначення статистичних параметрів найбільше поширення має метод моментів, графо-аналітичний метод, метод найбільшої правдоподібності.

1.6 Поняття про математичне сподівання, моду та медіану

Математичне сподівання характеризує положення випадкової величини на числовій осі з урахуванням того, що всі значення мають різні ймовірності появи.

Математичне сподівання випадкової величини m_x є центром статистичного розподілу випадкової величини X, відносно якого групуються члени сукупності та являє собою середнє зважене по ймовірності значення випадкової величини X, тобто

$$m_x = \frac{x_1 p_1 + x_2 p_2 + ... + x_N p_N}{p_1 + p_2 + ... + p_N} = \frac{\sum_{i=1}^{N} x_i p_i}{\sum_{i=1}^{N} p_i}. \qquad (1.13)$$

У (1.13) кожне значення x_i має ваговий коефіцієнт, пропорційний ймовірності цього значення.

Враховуючи, що для дискретних випадкових величин $\sum_{i=1}^{N} p_i = 1$, з (1.13) отримаємо

$$m_x = \sum_{i=1}^{N} x_i p_i . \qquad (1.14)$$

Математичне сподівання може також позначатися як $M[X]$.

Таким чином, ***математичне сподівання випадкової величини є сумою добутків усіх можливих значень випадкової величини на ймовірності цих значень.***

Геометрична інтерпретація цього математичного поняття є наступною.

Нехай на числовій осі розташовані точки із абсцисами $x_1, x_2, ..., x_N$, які мають маси $p_1, p_2, ..., p_N$, і $\sum_{i=1}^{N} p_i = 1$. Тоді математичне сподівання, яке визначається за (1.13), є нічим іншим як абсцисою центра тяжіння даної системи матеріальних точок.

Математичне сподівання для безперервної випадкової величини записується через інтеграл

$$m_x = \int_{-\infty}^{+\infty} x f(x) dx, \qquad (1.15)$$

де $f(x)dx$ - елемент ймовірності, який несе теж саме смислове навантаження, що й p_i для дискретних випадкових величин.

Як вже зазначалося, гідрологічні ряди представляють собою вибірки, для яких $n << N$. Таким чином, замість ймовірності p використовують частоту $p_i^* = \frac{m_i}{n}$. ***Оцінкою або емпіричним еквівалентом математичного сподівання, розрахованого за вибіркою, є середнє арифметичне значення*** $\bar{x}$

$$m_x^* = \bar{x} = \sum_{i=1}^{n} x_i p_i^* = \sum_{i=1}^{n} x_i \frac{m_i}{n}, \qquad (1.16)$$

де n - загальне число незалежних випробувань (спостережень);

$\sum_{i=1}^{n} m_i = n$;

p_i^* - частота або емпірична ймовірность.

Якщо число появи кожного значення випадкової величини дорівнює одиниці ($m_1, m_2, ..., m_n=1$), тобто кожне значення випадкової величини повторюється в досліді один раз, то $p_i^* = \frac{1}{n}$. Відповідно,

$$m_x^* = \hat{m}_x = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \,. \qquad (1.17)$$

Середнє арифметичне значення має ту саму розмірність, що й величина, по якій воно розраховано.

Модою *Mo* (m_0) ***випадкової величини*** називають найбільш ймовірне її значення. На графіку мода є точкою з найбільшою ординатою кривої розподілу. Розрізняють одно- та багатомодальні розподілення. Якщо багатокутник розподілу або крива розподілу мають більше одного максимума, то розподіл називається полімодальним (рис. 1.5) (Е.С. Вентцель, 1969)

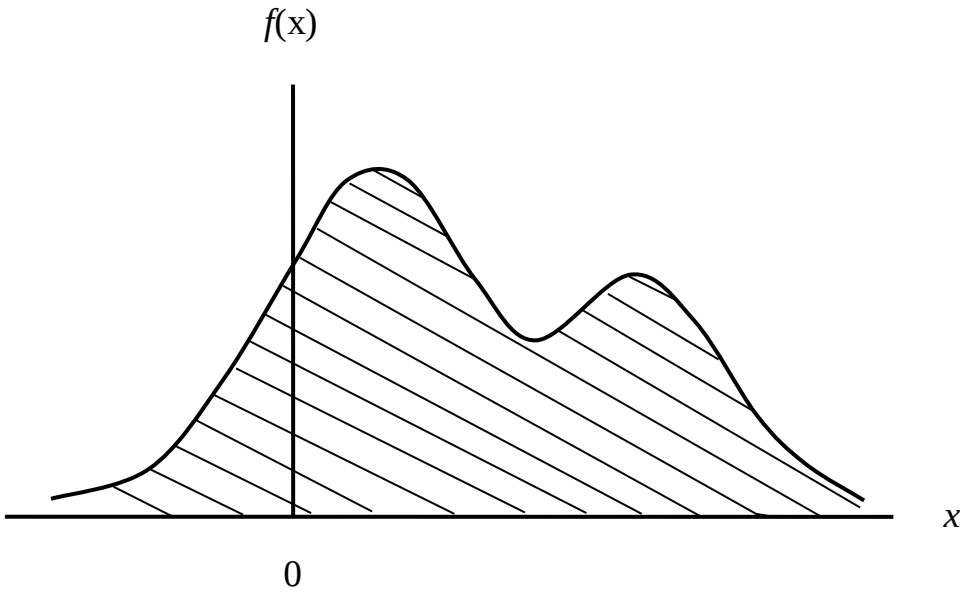


Рисунок 1.5 - Багатокутник розподілу з двома максимумами

Для неперервної випадкової величини мода є значенням з найбільшою щільністю ймовірності, для дискретної - з найбільшою ймовірністю.

Медіаною випадкової величини називається таке її значення *Me* або m_e, для якого виконується умова

$$p(X < Me) = p(X > Me), \qquad (1.18)$$

тобто однаково ймовірним є те, буде випадкова величина більша чи менша за значення Me. Геометрично медіана представляється як абсциса точки, в якій площа обмежена кривою розподілу ділиться навпіл (рис.1.6) (Е.С. Вентцель, 1969)

$$\int_{-\infty}^{Me} f(x)dx = \int_{Me}^{+\infty} f(x)dx. \qquad (1.19)$$

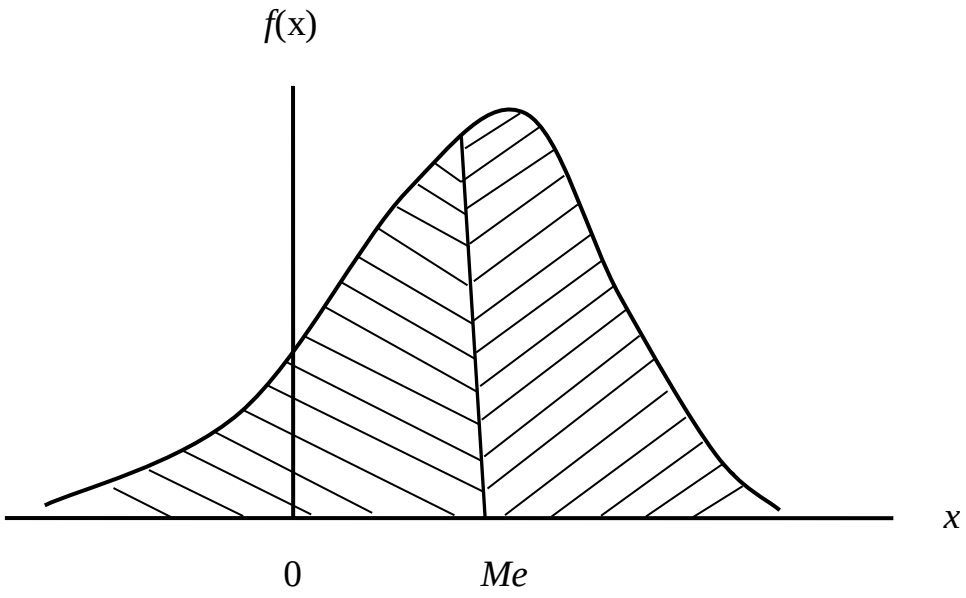


Рисунок 1.6 – Положення медіани на осі абсцис під кривою розподілу

Для дискретних випадкових величин береться таке правило, що, коли число членів ряду x_i непарне та дорівнює $2m+1$, то медіана ряду буде дорівнювати значенню x_{m+1} , яке є середнім членом ранжованого ряду, тобто $Me = x_{m+1}$. Якщо число членів ряду парне, тобто дорівнює $2m$, то за медіану береться середнє значення між центральними величинами ранжованого ряду (В.А. Шелутко, 1991).

Питання для самоперевірки

1. У чому головна задача математичної статистики?
2. Що являє собою центральна гранична теорема?
3. На які класи поділяються математичні моделі?
4. Дайте визначення «випадкової події» та «випадкової величини».

5. Яка подія є «достовірною» і яка «недостовірною»?
6. У чому полягає закон розподілу випадкової величини?
7. Коли випадкова величина називається «дискретною»?
8. Коли випадкова величина називається «неперервною»?
9. Як визначити забезпеченість випадкової величини $P(x)$?
10. Які властивості має диференціальна функція розподілу безперервної випадкової величини x ?
11. Назвіть основні статистичні параметри законів розподілу?
12. Що являє собою «математичне сподівання», «мода», «медіана» випадкової величини?

2 ВИЗНАЧЕННЯ СТАТИСТИЧНИХ ПАРАМЕТРІВ ЗА МЕТОДОМ МОМЕНТІВ

2.1 Поняття про статистичні моменти та їх зв'язок із статистичними параметрами

Поняття моментів перенесено в математичну статистику із розділу фізики "Механіка", де момент являє собою добуток сили на плече. Плече – відстань від точки, у якій прикладена сила, до точки опори. Значення дискретної випадкової величини розглядається як матеріальна точка на числовій осі з масою, пропорційною ймовірності появи цієї випадкової величини. Якщо плечем є відстань від нуля числової осі до матеріальної точки, то такі статистичні моменти називаються початковими. Коли ж для визначення статистичного моменту береться відстань від математичного сподівання до розглядуваної матеріальної точки, то статистичний момент отримує назву центрального.

Для описування властивостей кривих розподілу широко використовують початкові та центральні статистичні моменти.

Початкові моменти s-го порядку дискретної величини X являють собою суму добутків

$$\alpha_s = \sum_{i=1}^{N} x_i^s p_i, \tag{2.1}$$

де α_s - початковий момент s-го порядку.

Для неперервної випадкової величини сума записується через інтеграл

$$\alpha_s = \int_{-\infty}^{\infty} x^s f(x) dx. \tag{2.2}$$

Таким чином, для дискретних випадкових величин оцінка початкових моментів s-того порядку розраховується за формулою

$$\hat{\alpha}_s = \sum_{i=1}^{n} x_i^s p_i^* \text{ , де } p_i^* = \frac{1}{n}. \tag{2.3}$$

Перший початковий момент (s=1) є нічим іншим як математичним сподіванням

$$\alpha_1 = m_x = \sum_{i=1}^{N} x_i p_i, \tag{2.4}$$

а його оцінка є середнім арифметичним значенням $\bar{x}$

$$\hat{\alpha}_1 = \hat{m}_x = \bar{x} = \sum_{i=1}^{n} x_i p_i^* = \frac{\sum_{i=1}^{n} x_i}{n}. \qquad (2.5)$$

Центральні моменти s-го порядку β_s дискретних випадкових величин описуються таким виразом

$$\beta_s = \sum_{i=1}^{N} (x_i - m_x)^s p_i ; \qquad (2.6)$$

відповідно для неперервних випадкових величин -

$$\beta_s = \int_{-\infty}^{\infty} (x - m_x)^s f(x) dx. \qquad (2.7)$$

Для будь-якої випадкової величини центральний момент першого порядку дорівнює нулю.

Оцінка центрального моменту s-го порядку на основі вибірки довжиною n виконується в такий спосіб

$$\hat{\beta}_s = \sum_{i=1}^{n} (x_i - \bar{x})^s p_i = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^s . \qquad (2.8)$$

При s=1

$$\beta_{s=1} = \sum_{i=1}^{n} (x_i - m_x) p_i = \sum_{i=1}^{n} x_i p_i - m_x \sum_{i=1}^{n} p_i = m_x - m_x = 0, \qquad (2.9)$$

або

$$\hat{\beta}_{s=1} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) = \frac{1}{n} \sum_{i=1}^{n} x_i - \frac{1}{n} \sum_{i=1}^{n} \bar{x} = 0. \qquad (2.10)$$

Таким чином, перший центральний момент дорівнює нулю.

При $s = 2$

$$\beta_2 = \sum_{i=1}^{N} (x_i - m_x)^2 p_i \qquad (2.11)$$

або

$$\hat{\beta}_2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 . \qquad (2.12)$$

Другий центральний момент характеризує розсіювання випадкової величини відносно її центра розподілу і має назву дисперсії D_x

$$\beta_2 = D_x = \sigma_x^2 . \qquad (2.13)$$

Квадратний корінь із дисперсії називається середнім квадратичним відхиленням σ_x.Оцінка середнього квадратичного відхилення розраховується таким чином

$$\hat{\sigma}_x = \sqrt{D_x} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}} \ . \qquad (2.14)$$

Оцінки другого центрального моменту мають від'ємні зміщення (заниження). Для їх усунення в (2.14) вводиться поправочний коефіцієнт $\sqrt{n/(n-1)}$, з урахуванням якого (2.14) набуває вигляду

$$\hat{\sigma}_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} . \qquad (2.15)$$

З метою порівняння мінливості різномасштабних випадкових величин використовують безрозмірну характеристику $C_V = \sigma_x / m_x$, яка має назву ***коефіцієнта варіації*** і за даними спостережень оцінюється таким чином

$$\hat{C}_v = \frac{\hat{\sigma}_x}{\bar{x}} . \qquad (2.16)$$

У більш детальному вигляді оцінка коефіцієнта варіації знаходиться за виразом

$$\hat{C}_V = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\bar{x}^2(n-1)}} = \sqrt{\frac{\sum_{i=1}^{n}(k_i - 1)^2}{n-1}}\ , \qquad (2.17)$$

де $k_i = x_i / \bar{x}$ - модульний коефіцієнт.

При $s = 3$ з використанням (2.16) та (2.17) отримаємо вирази для розрахунків третього центрального моменту для генеральної сукупності

$$\beta_3 = \sum_{i=1}^{N}(x_i - m_x)^3 p_i\ , \qquad (2.18)$$

та вибірки

$$\hat{\beta}_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3\ . \qquad (2.19)$$

Третій центральний момент характеризує несиметричність (асиметричність) розподілу випадкової величини відносно математичного сподівання і може бути як від'ємним, так і додатнім.

Нормування β_S по σ_x^3 дозволяє отримати безрозмірний параметр статистичного розподілу, названий коефіцієнтом асиметрії C_S

$$C_S = \frac{\beta_3}{\sigma_x^3}\ , \qquad (2.20)$$

який при розрахунках за вибірками представляється у вигляді

$$\hat{C}_S = \hat{\beta}_3 / \hat{\sigma}_x^3 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n\hat{\sigma}_x^3}\ . \qquad (2.21)$$

Вираз (2.21) при використанні модульних коефіцієнтів $k_i = x_i / \bar{x}$ набуває вигляду

$$\hat{C}_S = \frac{\sum_{i=1}^{n}(k_i - 1)^3}{n\hat{C}_V^3}\ . \qquad (2.22)$$

Як і оцінка параметра C_V, оцінка коефіцієнта асиметрії за виразами (2.21) і (2.22) є зміщеною відносно відповідного параметра генеральної сукупності. Від'ємна зміщенність може бути усунена шляхом введення поправочного коефіцієнта $n^2/(n-1)(n-2)$.

Таким чином, кінцевий вигляд формули для розрахунків коефіцієнта асиметрії має такий вигляд

$$\hat{C}_S = \frac{n}{(n-1)(n-2)} \frac{\sum_{i=1}^{n}(k_i - 1)^3}{\hat{C}_V^3}. \qquad (2.23)$$

Для випадкової величини, яка підлягає нормальному закону розподілу, $C_S = 0$, тобто крива щільності ймовірностей є симетричною відносно математичного сподівання.

Четвертий центральний момент ($s = 4$), покладений в основу характеристики гостровершинності кривої розподілу випадкової величини, має назву ексцесу. Формула для визначення четвертого центрального моменту за вибіркою набуває вигляду

$$\hat{\beta}_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4. \qquad (2.24)$$

Четвертий центральний момент характеризує сплюснутість або витягнутість кривої розподілу випадкової величини у порівнянні з кривою нормального розподілу. Для випадкової величини з нормальним законом розподілу співвідношення β_4 / σ_x^4 завжди дорівнює 3. Таким чином, нормування β_4 по σ_x^4 дозволяє отримати безрозмірний статистичний параметр, названий ексцесом

$$E = \frac{\beta_4}{\sigma_x^4} - 3 \qquad (2.25)$$

або

$$\hat{E} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n\hat{\sigma}_x^4} - 3. \qquad (2.26)$$

Якщо $E > 0$, то крива розподілу витягнута відносно нормального закону розподілу, для якого $E = 0$. Коли ж $E < 0$, крива розподілу приплюснута по відношенню до кривої нормального розподілу (рис.2.1).

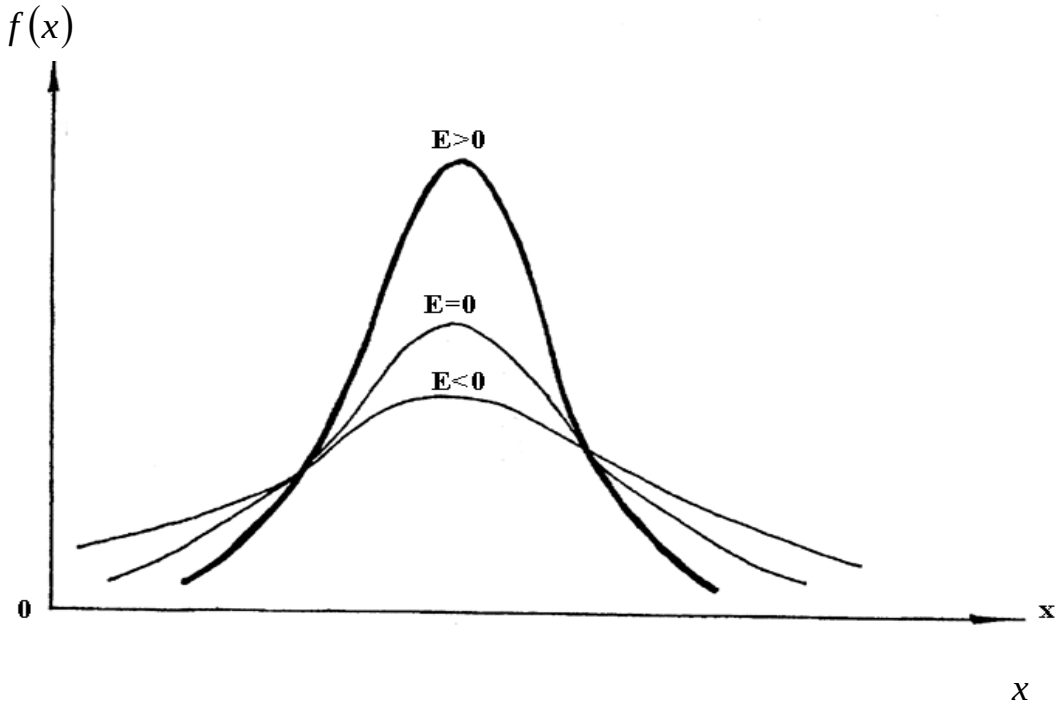


Рисунок 2.1 - Криві розподілу з різними значеннями ексцесу Е.

2.2 Вимоги до оцінок статистичних параметрів

Статистичні параметри, розраховані за вибірками, називаються вибірковими оцінками статистичних параметрів. Значення вибіркових оцінок відрізняються від значень відповідних параметрів генеральної сукупності.

Для того, щоб по оцінках статистичних параметрів достатньо вірогідно характеризувати параметри генеральної сукупності, ці оцінки повинні задовольняти вимогам незміщеності, ефективності та умотивованості.

Припустимо, що з генеральної сукупності взяті m вибірок, за якими розраховані m оцінок параметрів.

Оцінка статистичного параметра називається ***незміщеною***, якщо її математичне сподівання дорівнює параметру генеральної сукупності

$$M\left[\widehat{\theta}_m\right]=\theta, \tag{2.27}$$

де M - позначка математичного сподівання;

$\widehat{\theta}_m$ - оцінка статистичного параметра, яка є випадковою величиною, що змінюється від вибірки до вибірки;

θ - значення параметра генеральної сукупності.

Якщо оцінка є незміщеною, то відсутність систематичних похибок п гарантується.

Незміщена оцінка, яка має найменшу дисперсію серед усіх можливих незміщених оцінок параметра, розрахованих по вибірках одного й того ж об'єму, називається ***ефективною***, тобто повинна виконуватись умова

$$D\left[\widehat{\theta}_m\right]\to 0 \text{ при } m\to\infty, \tag{2.28}$$

де $D\left[\widehat{\theta}_m\right]$- дисперсія, тобто розсіювання випадкової величини $\widehat{\theta}_m$ відносно математичного сподівання θ.

Оцінка параметра $\widehat{\theta}_m$ називається ***умотивованою***, якщо вона по ймовірності збігається до параметра генеральної сукупності θ

$$\lim_{n\to\infty}\left(p\left|\widehat{\theta}_m-\theta\right|<\varepsilon\right)=1, \tag{2.29}$$

де ε - дуже мале додатне число;

$\widehat{\theta}_m$ - вибіркова оцінка параметра;

θ - значення параметра генеральної сукупності.

У методі моментів тільки середнє арифметичне значення є незміщеною, ефективною та умотивованою оцінкою.

Як було зазначено вище, для усунення систематичних похибок при розрахунках параметрів σ_x, C_V, C_S за вибірками вводилися відповідні поправочні коефіцієнти: $\sqrt{n/(n-1)}$ - при обчисленні середньоквадратичного відхилення та коефіцієнта варіації, а також $n^2/[(n-1)(n-2)]$ - при обчисленні коефіцієнта асиметрії C_S.

Особливістю методу моментів є той факт, що при розрахунках за цим методом збільшується внесок значних відхилень від центра розподілу.

2.3 Поняття про систематичні та випадкові похибки визначення статистичних параметрів

Похибки оцінок статистичних параметрів за вибірками можна поділити на систематичні та випадкові (А.В. Рождественський, А.В. Єжов, А.В.Сахарюк, 1990). ***Систематичні похибки*** виникають за рахунок чинників, які однаково впливають на результат при багатократному повторенні вимірювань. ***Випадкові похибки*** виникають внаслідок впливу комплексу чинників, кожне з яких урахувати неможливо, що приводить при кожному вимірюванні до різних похибок як за числовим значенням, так і за знаком. Систематичні похибки не можуть бути вилучені або зменшені при багатократному вимірюванні. У той же час випадкові похибки зменшуються при збільшенні числа вимірювань однієї і тієї ж величини. Як відомо, середнє арифметичне значення при збільшенні числа вимірювань наближається до значення математичного сподівання генеральної сукупності. Систематичні похибки, коли вони відомі, можуть бути усунені або шляхом зміни методу вимірювань (розрахунків), або шляхом введення у розрахункові формули відповідних поправочних коефіцієнтів. Випадкові ж похибки можна лише оцінити. При цьому постає питання про те, як саме виконується оцінка випадкових похибок.

2.4 Похибки визначення оцінок статистичних параметрів при розрахунках за методом моментів

Формула середньоквадратичного відхилення середніх арифметичних значень від математичного сподівання отримана аналітичним шляхом. При цьому береться припущення, що нормальний закон розподілу вибіркових середніх зберігається і для вибірок, які відхиляються від нормального закону розподілу:

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}, \qquad (2.30)$$

де $\sigma_{\bar{x}}$ - середнє квадратичне відхилення вибіркової середньої арифметичної величини від математичного сподівання генеральної сукупності;

σ_x - середнє квадратичне відхилення випадкової величини для ряду спостережень довжиною n, яке визначається за формулою (2.23).

Середнє квадратичне відхилення коефіцієнтів варіації σ_{C_V} рекомендується обчислювати за такою формулою

$$\sigma_{C_V} = C_V \sqrt{(1 + C_V^2)/2n} \qquad (2.31)$$

або за формулою Є. Г. Блохінова

$$\sigma_{C_V} = \frac{C_V}{n + 4C_V^2} \sqrt{\frac{n}{2}(1 + C_V^2)}\,. \qquad (2.32)$$

Середня квадратична похибка коефіцієнта асиметрії σ_{C_S} визначається за формулою С.М.Крицького та М.Ф.Менкеля

$$\sigma_{C_S} = \sqrt{\frac{6}{n}(1 + 6C_V^2 + 5C_V^4)} \qquad (2.33)$$

Середня квадратична похибка відношення C_S/C_V визначається за формулою А.Ш. Резніковського

$$\sigma_{C_S/C_V} = \frac{1}{C_V}\sqrt{\frac{6}{n}}\quad . \qquad (2.34)$$

Для визначення якості розрахунків статистичних параметрів за даними спостережень необхідно ввести критерій якості, а саме – допустиму похибку розрахунків. Допустиму похибку краще представляти у частках або відсотках від значення параметра, точність визначення якого оцінюється. Наприклад, відносне середнє квадратичне відхилення середньої арифметичної величини $\bar{x}$, визначеною за вибіркою довжиною n, від математичного сподівання m_x генеральної сукупності розраховується таким чином

$$\varepsilon_{\bar{x}} = \frac{\sigma_{\bar{x}}}{\bar{x}} = \frac{\sigma_x}{\bar{x}\sqrt{n}} = \frac{C_V}{\sqrt{n}} \cdot 100\%\,. \qquad (2.35)$$

Із виразу (2.35) видно, що відносна випадкова похибка визначення середнього арифметичного значення пропорційна коефіцієнту C_V та зворотна довжині вибірки n. Із збільшенням довжини вибірки відносна похибка $\varepsilon_{\bar{x}}$ зменшується.

Відносна випадкова похибка визначення коефіцієнта варіації за вибірковими даними розраховується таким чином

$$\varepsilon_{C_V} = \frac{\sigma_{C_V}}{C_V} \cdot 100\%. \qquad (2.36)$$

Відносні випадкові похибки визначення коефіцієнта асиметрії або відношення C_S / C_V по ряду спостережень розраховуються за виразами

$$\varepsilon_{C_S} = \frac{\sigma_{C_S}}{C_S} \cdot 100\% \ ; \quad \varepsilon_{C_S / C_V} = \frac{\sigma_{C_S / C_V}}{\sigma_{C_S / C_V}} \cdot 100\% \ . \qquad (2.37)$$

Використовуючи формули (2.35) – (2.36), можна знайти довжину рядів, які забезпечують необхідну точність визначення статистичних параметрів. Так, при C_V=0,6 та заданій допустимій похибці визначення середнього арифметичного $\varepsilon_{\bar{x}} = 10\%$, згідно із формулою (2.35), необхідною довжиною ряду є n=30, а при C_V=1,5 необхідною кількість спостережень обчислюється величиною 144. Зазначимо, що ряди найчастіше мають довжину меншу ніж 50 років і лише окремі з них досягають довжини в 100 років і більше.

У зв'язку з великою похибкою визначення коефіцієнта асиметрії за даними спостережень його нормують за співвідношенням C_S / C_V у межах окремих районів. Що стосується такого статистичного параметра як ексцес E, то ця характеристика практично не використовується через низьку точність її визначення по рядах спостережень.

Питання для самоперевірки

1. Що являє собою перший початковий момент?
2. Чому дорівнює центральний момент першого порядку для випадкових величин?
3. Що характеризують другий, третій і четвертий центральні моменти?
4. Як розрахувати коефіцієнти варіації та асиметрії?
5. Що представляють собою оцінки статистичних параметрів?
6. Яка оцінка статистичних параметрів є одночасно незміщеною, ефективною та умотивованою?
7. Дайте визначення «систематичних» і «випадкових» похибок.
8. Як розраховується середня квадратична похибка основних статистичних параметрів?
9. Який критерій є показником якості розрахунків статистичних параметрів?

3 ВИЗНАЧЕННЯ СТАТИСТИЧНИХ ПАРАМЕТРІВ ГРАФО-АНАЛІТИЧНИМ МЕТОДОМ

Графо-аналітичний метод або метод квантілей запропонований в 1960 році Г.А. Алексєєвим і являє собою спрощений спосіб розрахунків статистичних параметрів. В ньому використовується згладжена емпірична крива забезпеченостей і аналітичний (теоретичний) закон розподілу. При його застосуванні приймається умова збіжності теоретичної кривої розподілу з емпіричною хоча б у трьох точках. Цей метод дозволяє оцінити статистичні параметри по рядах стоку безпосередньо для того теоретичного розподілу ймовірностей, який в більшій мірі відповідає емпіричному.

3.1 Емпірична функція забезпеченості

У гідрологічних розрахунках закон розподілу випадкової величини задається у вигляді функції забезпеченості. Емпірична забезпеченість заданого значення $X = x$ являє собою ймовірність того, це значення буде перевищене. Для її обчислення використовується формула (1.3), записана через емпіричну частоту, тобто

$$P^*(x) = p^*(X \geq x). \qquad (3.1),$$

Для розрахунків емпіричної функції забезпеченості необхідно для кожного x_i з вибірки довжиною n $(i = 1.2,...,n)$ визначити кількість випадків, коли випадкова величина X набула значення більшого або рівного x_i і поділити знайдене m на загальну кількість випробувань.

З цією метою гідрологічний ряд розташовується в убутному порядку (ранжирується). Приймається, що кожне значення ряду спостерігається один раз, тобто

$$p^*(X = x_i) = \frac{1}{n}. \qquad (3.2)$$

Емпірична забезпеченість визначається шляхом послідовного підсумовування $p^*(X = x_i)$ от найбільшого члена вибірки до відповідного m-того значення ранжируваного ряда

$$P^* = \sum_{i=1}^{m} p^* = \sum_{i=1}^{m} \frac{1}{n} = \frac{m}{n}, \qquad (3.3),$$

Формула (3.3) є справедливою лише для випадку, коли всі значення X представлені в одній вибірці.

У протилежному випадку отримуємо, що забезпеченість першого члена ранжируваної вибірки дорівнює $P_1^* = \frac{1}{n}$, а останнього $P_n^* = \frac{n}{n} = 1$. Таким чином, виходить, що значення випадкової величини, більші або менші тих, що увійшли до однієї виборки, стають неможливими. Це суперечить досвіду практики, з якого витікає, що які б значення не увійшли до вибірки, завжди можливі значення більші або менші спостережених. Цей недолік виключається шляхом застосування таких формул:

- формула Хазена

$$P^* = \frac{m - 0.5}{n}; \tag{3.4}$$

- формула С.М. Крицького та М.Ф. Менкеля

$$P^* = \frac{m}{n+1}; \tag{3.5}$$

- формула М.М. Чегодаєва

$$P^* = \frac{m - 0.3}{n + 0.4}. \tag{3.6}$$

3.2 Застосування графо-аналітичного методу до теоретичного закону розподілу Пірсона III

Розглянемо випадок, коли як теоретичний використовується закон Пірсона III. За графіком емпіричної кривої забезпеченості визначаються величини стоку в характерних точках з забезпеченістю 5, 50 та 95 відсотків (%). Виходячи з припущення, що в цих точках емпірична крива забезпеченості співпадає з теоретичною, звернемося до закону розподілу Пірсона Ш. Цей теоретичний закон розподілу випадкової величини надається у виді таблиці нормованих відхилень $\Phi_P(C_S)$, які залежать від забезпеченості P і коефіцієнта асиметрії C_S:

$$\Phi_P(C_S) = \frac{x_P - \bar{x}}{\hat{\sigma}_x}. \qquad (3.7)$$

Для трьох характерних точок (P=5%, 50%, 95%) з формули (3.7) визначаються випадкові величини x_P:

$$x_5 = \bar{x} + \hat{\sigma}_x \Phi_5 \ ; \qquad (3.8)$$

$$x_{50} = \bar{x} + \hat{\sigma}_x \Phi_{50} \ ; \qquad (3.9)$$

$$x_{95} = \bar{x} + \hat{\sigma}_x \Phi_{95} \qquad (3.10)$$

з трьома невизначеними параметрами: $\bar{x}$, $\hat{\sigma}_x$ та $\hat{C}_S$. Параметр $\hat{C}_S$ входить у рівняння (3.7) в силу того, що $\Phi_P(C_S)$ є функцією C_S.

Для визначення коефіцієнта асиметрії використовується коефіцієнт скісності S, який функціонально пов'язаний з C_S і представлений у табл.(3.1) .

Таблиця 3.1 – Значення коефіцієнта асиметрії C_S та скісності S для біноміальної кривої розподілу Персона III

C_S	$\frac{x_p - \bar{x}}{\sigma_x} = \frac{k_p - 1}{C_V} = \hat{O}(P, \tilde{N}_S)$						$\hat{O}_5 - \hat{O}_{95}$	$\hat{S}$
	Φ_1	Φ_2	Φ_5	$\hat{O}_{10}$	Φ_{50}	$\hat{O}_{95}$		
0.0	2.33	2.02	1.64	1.28	0.00	-1.64	3.28	0.00
0.1	2.40	2.11	1.67	1.29	-0.02	-1.61	3.28	0.03
0.2	2.47	2.16	1.70	1.30	-0.03	-1.58	3.28	0.06
0.3	2.54	2.21	1.72	1.31	-0.05	-1.52	3.27	0.09
0.4	2.61	2.26	1.75	1.32	-0.07	-1.52	3.27	0.11
0.5	2.68	2.31	1.77	1.32	-0.08	-1.49	3.26	0.16
0.6	2.75	2.35	1.80	1.33	-0.10	-1.45	3.25	0.17
0.7	2.82	2.40	1.82	1.33	-0.12	-1.42	3.24	0.20
0.8	2.89	2.45	1.84	1.34	-0.13	-1.38	3.22	0.22
0.9	2.96	2.50	1.86	1.34	-0.15	-1.35	3.21	0.25
1.0	3.02	2.54	1.88	1.34	-0.16	-1.32	3.20	0.28
1.1	3.09	2.58	1.89	1.34	-0.18	-1.28	3.17	0.31
1.2	3.15	2.62	1.92	1.34	-0.19	-1.24	3.16	0.34
1.3	3.21	2.57	1.94	1.34	-0.21	-1.20	3.14	0.37

Продовження таблиці 3.1

C_S	$\frac{x_p - \bar{x}}{\sigma_x} = \frac{k_p - 1}{C_V} = \hat{O}(P, \tilde{N}_S)$						$\hat{O}_5 - \hat{O}_{95}$	$\hat{S}$
	Φ_1	$\hat{O}_2$	Φ_5	Φ_{10}	$\hat{O}_{50}$	$\hat{O}_{95}$		
1.4	3.27	2.71	1.95	1.34	-0.22	-1.17	3.12	0.39
1.5	3.33	2.74	1.96	1.33	-0.24	-1.13	3.09	0.42
1.6	3.39	2.78	1.97	1.33	-0.25	-1.10	3.07	0.45
1.7	3.44	2.82	1.98	1.32	-0.27	-1.06	3.04	0.49
1.8	3.50	2.85	1.99	1.32	-0.28	-1.02	3.01	0.51
1.9	3.55	2.88	2.00	1.31	-0.29	-0.98	2.98	0.54
2.0	3.60	2.91	2.00	1.30	-0.31	-0.95	2.95	0.57
2.1	3.65	2.94	2.01	1.29	-0.32	-0.91	2.92	0.59
2.2	3.68	2.95	2.02	1.27	-0.33	-0.88	2.90	0.63
2.3	3.73	2.98	2.01	1.26	-0.34	-0.85	2.86	0.64
2.4	3.78	3.02	2.00	1.25	-0.35	-0.82	2.82	0.68
2.5	3.82	3.05	2.00	1.23	-0.36	-0.79	2.79	0.69
2.6	3.85	3.05	2.00	1.21	-0.37	-0.76	2.76	0.72
2.7	3.92	3.10	2.00	1.19	-0.38	-0.74	2.74	0.74
2.8	3.96	3.12	2.00	1.18	-0.39	-0.71	2.71	0.76
2.9	4.01	3.12	1.99	1.15	-0.39	-0.69	2.68	0.78
3.0	4.05	3.14	1.97	1.13	-0.40	-0.66	2.63	0.80
3.1	4.09	3.14	1.97	1.11	-0.40	-0.64	2.62	0.81
3.2	4.11	3.14	1.96	1.09	-0.41	-0.62	2.59	0.83
3.3	4.15	3.14	1.95	1.08	-0.41	-0.60	2.56	0.85
3.4	4.18	3.15	1.94	1.08	-0.41	-0.59	2.53	0.86
3.5	4.21	3.16	1.93	1.04	-0.41	-0.57	2.50	0.87
3.6	4.24	3.17	1.93	1.03	-0.42	-0.56	2.48	0.89
3.7	4.26	3.18	1.91	1.01	-0.42	-0.54	2.45	0.90
3.8	4.29	3.18	1.90	1.00	-0.42	-0.53	2.43	0.91
3.9	4.32	3.20	1.90	0.98	-0.41	-0.51	2.41	0.92
4.0	4.34	3.20	1.90	0.96	-0.41	-0.50	2.40	0.92
4.1	4.36	3.22	1.89	0.95	-0.41	-0.49	2.38	0.93
4.2	4.39	3.21	1.88	0.93	-0.41	-0.48	2.36	0.94
4.6	4.46	3.27	1.84	0.87	-0.40	-0.44	2.28	0.97
4.7	4.49	3.28	1.83	0.85	-0.40	-0.43	2.26	0.97
4.8	4.50	3.29	1.81	0.82	-0.39	-0.42	2.23	0.98
4.9	4.51	3.30	1.80	0.80	-0.39	-0.41	2.21	0.98
5.0	4.54	3.32	1.78	0.78	-0.38	-0.40	2.18	0.98
5.1	4.57	3.32	1.76	0.76	-0.38	-0.39	2.15	0.98
5.2	4.59	3.33	1.74	0.73	-0.37	-0.38	2.15	0.98

Коефіцієнт скісності розраховується за такою формулою

$$\hat{S} = \frac{x_5 + x_{95} - 2x_{50}}{x_5 - x_{95}} = \frac{\Phi_5 + \Phi_{95} - 2\Phi_{50}}{\Phi_5 - \Phi_{95}}. \qquad (3.11)$$

За табл.3.1 відповідно визначеному за гідрологічним рядом $\hat{S}$ встановлюється коефіцієнт $\hat{C}_S$ та нормовані ординати Φ_5, Φ_{50}, Φ_{95}.

Для отримання математичного виразу, який буде визначати середньоквадратичне відхилення, віднімаємо з лівої та правої частин рівняння (3.8) відповідні частини рівняння (3.10):

$$x_5 - x_{95} = \hat{\sigma}_x(\Phi_5 - \Phi_{95}), \qquad (3.12)$$

звідки

$$\hat{\sigma}_x = \frac{x_5 - x_{95}}{\Phi_5 - \Phi_{95}}. \qquad (3.13)$$

Середнє арифметичне значення знаходять з рівняння (3.9)

$$\bar{x} = x_{50} - \hat{\sigma}_x \Phi_{50}. \qquad (3.14)$$

Коефіцієнт варіації розраховується за відношенням

$$\hat{C}_V = \frac{\hat{\sigma}_x}{\bar{x}}. \qquad (3.15)$$

Якщо коефіцієнт скісності S від'ємний, то це свідчить про від'ємну асиметрію $(C_S < 0)$ розподілу. У таких випадках наведені в таблиці ординат кривої забезпеченості Пірсона Ш величини беруться з протилежним знаком:

$$\frac{x_p - \bar{x}}{\hat{\sigma}_x} = -\hat{O}_{P'}, \qquad (3.16)$$

для значень забезпеченості $P' = 100 - P$ (у відсотках) і при додатному значенні коефіцієнта асиметрії $C_S^* = |C_S|$. Ординати кривої розраховуються за формулою

$$x_P = \bar{x} - \hat{\sigma}_x \hat{O}_{P'} \ ; \tag{3.17}$$

Для трьох характерних точок отримуємо значення

$$x_5 = \bar{x} - \hat{\sigma}_x \hat{O}_5 \ ; \tag{3.18}$$

$$x_{50} = \bar{x} - \hat{\sigma}_x \hat{O}_{50} ; \tag{3.19}$$

$$x_{95} = \bar{x} - \hat{\sigma}_x \hat{O}_{95} . \tag{3.20}$$

Тоді

$$S' = -S = \frac{2x_{50} - x_5 - x_{95}}{x_5 - x_{95}} = \frac{\hat{O}_5 + \hat{O}_{95} - 2\hat{O}_{50}}{\hat{O}_5 - \hat{O}_{95}} ; \tag{3.21}$$

$$\hat{\sigma}_x = \frac{x_5 - x_{95}}{\hat{O}_5 - \hat{O}_{95}} ; \tag{3.22}$$

$$\bar{x} = x_{50} + \hat{\sigma}_x \hat{O}_{50} . \tag{3.23}$$

Зрозуміло, що хоч у графо-аналітичному методі розрахунків статистичних параметрів і використовується теоретичний закон розподілу, отримані статистичні характеристики є все ж таки тільки оцінками статистичних параметрів генеральної сукупності, бо вони спираються на емпіричну криву забезпеченості, побудовану за даними вибірки.

Графо-аналітичний метод також застосовується для такого теоретичного закону розподілу випадкової величини як логарифмічно-нормальний закон. Похибки визначення статистичних параметрів за вибірками установлюються по тим же формулам, що й в методі моментів.

Питання для самоперевірки

1. На чому заснований графо-аналітичний метод?
2. У чому полягає емпірична функція розподілу? Як її розрахувати?
3. Що являє собою теоретична крива розподілу?
4. У чому відмінність емпіричної і теоретичної кривих розподілу?
5. У чому полягає закон розподілу Пірсона III типу?
6. З якою характеристикою функціонально пов’язаний коефіцієнт скісності *S*?
7. Як розраховуються похибки визначення статистичних параметрів за графо-аналітичним методом?

4 ТЕОРЕТИЧНІ ЗАКОНИ РОЗПОДІЛУ, ЇХ ОСОБЛИВОСТІ ТА МЕЖІ ЗАСТОСУВАННЯ

Теоретичні закони розподілу ймовірностей базуються або на визначених теоретичних схемах, або є узагальненням емпіричних розподілів (Richard H. McCuen, 2002).

Вимоги до теоретичних кривих розподілу величин:

1. У рівнянні кривої повинно бути якнайменше параметрів, які чисельно визначаються за вибірковими даними.
2. Через те, що значення стоку та концентрацій забруднюючих речовин завжди додатні, крива розподілу не повинна знаходитися в області від'ємних значень.
3. Верхня межа кривої розподілу необмежена.
4. Теоретичні криві розподілу повинні бути одномодальними, що витікає з умови однорідності і незалежності досліджуваних величин.

4.1 Диференціальне рівняння кривих щільності ймовірностей К.Пірсона

Розподіл ймовірностей $y = f(z)$ повинен задовольняти таким умовам: на початку та на кінці графіка щільності ймовірностей $y = 0$ та між початком і кінцем досягає максимального значення.

Отже, перша похідна $\frac{dy}{dz}$ повинна дорівнювати нулю у трьох точках: на початку, в кінці та у точці, яка відповідає моді.

Величина z представляє собою центровану і нормовану вихідну величину x, тобто

$$z = \frac{x - m_x}{m_x} = k - 1, \tag{4.1}$$

де k - модульний коефіцієнт.

Сукупність теоретичних кривих розподілу випадкових величин Пірсона можна отримати в результаті розв'язання диференціального рівняння (В.А. Шелутко, 1991)

$$\frac{dy}{dz} = \frac{y(z + d)}{\varphi(z)}, \tag{4.2}$$

де d – відстань між модою та математичним сподіванням на графіку кривої розподілу (рис.4.1);

$y = f(x))$ - щільність ймовірності;

$\varphi(z))$ - ряд Маклорена, який має вигляд

$$\varphi(z) = b_0 + b_1 z + b_2 z^2 + \ldots + b_n z^n . \qquad (4.3)$$

Рівняння (4.3) відповідає усім висунутим раніше вимогам. Дійсно, на кінцях розподілу при $y = 0$, $\frac{dy}{dz} = 0$; при $z = -d$, тобто в точці максимуму (m_o) значення $\frac{dy}{dz}$ також дорівнює нулю.

Для побудови кривих розподілу Пірсон використовував тільки три перших члена ряду Маклорена, тобто

$$\varphi(z) = b_0 + b_1 z + b_2 z^2 , \qquad (4.4)$$

звідки

$$\frac{dy}{dz} = \frac{y(z+d)}{b_0 + b_1 z + b_2 z^2} . \qquad (4.5)$$

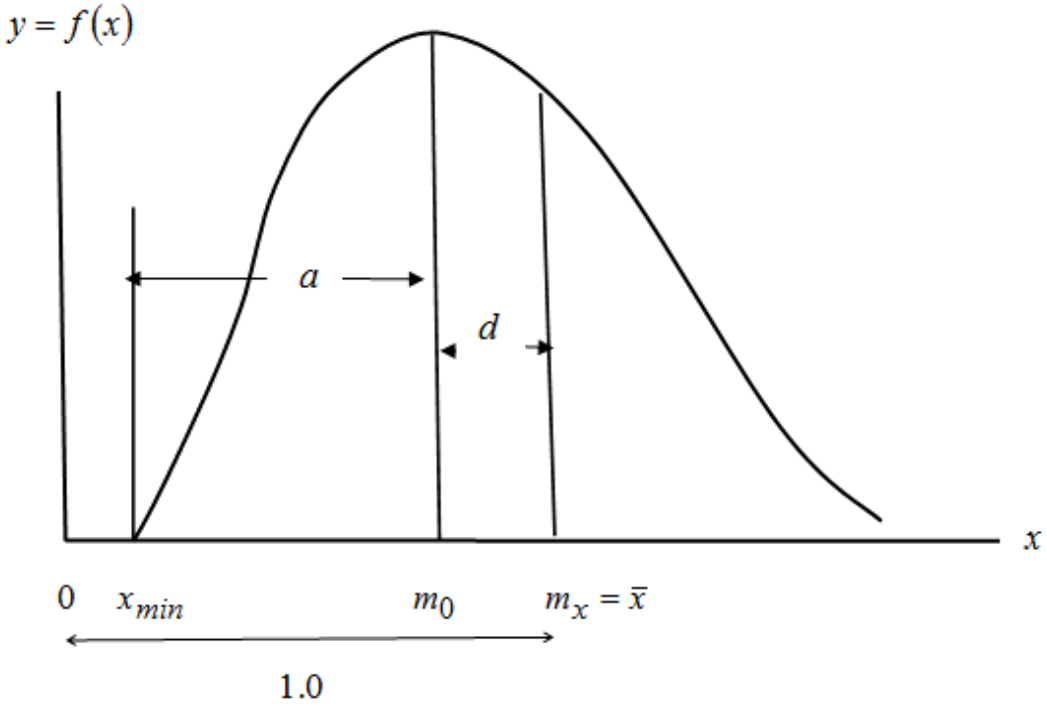


Рисунок 4.1 - Крива Пірсона III (C_S >0)

Значення z та d вважаються відомими. Для розрахунків щільності ймовірності необхідно знайти значення b. Після інтегрування (4.5) приходимо до рівняння

$$-nb_0\beta_{n-1} - b_1(n+1)\beta_n - b_2(n+2)\beta_{n+1} = \beta_{n+1} + d\beta_n, \qquad (4.6)$$

де β_n- центральні моменти порядку n.

Розглянувши цей вираз при різних n $(n = 0,1,2)$ та враховуючи, що $\beta_0 = 1$ і $\beta_1 = 0$, отримуємо систему рівнянь

$$\begin{cases} -b_1 = d \\ -b_0 - 3b_2\beta_2 = \beta_2 \\ -3b_1\beta_2 - 4b_2\beta_3 = \beta_3 + d\beta_2 \end{cases}, \qquad (4.7)$$

яка дозволяє виразити b_0 , b_1, b_2 через центральні моменти другого та третього порядків.

З системи рівнянь (4.7) витікає, що для розрахунків перших трьох параметрів ряду Маклорена, необхідно використовувати третій центральний момент, який визначається за експериментальними даними з великими похибками. Отже подальше збільшення числа членів Маклорена немає сенсу.

В залежності від числових значень коефіцієнтів b_0, b_1, b_2 з вихідного рівняння можна отримати сім різних типів кривих забезпеченостей. В гідрологічних розрахунках використовуються два з них: нормальний та біноміальний асиметричний (Пірсона III).

4.2 Нормальний закон розподілу випадкової величини

Нормальний закон розподілу є частковим випадком розподілу Пірсона. Для знаходження відповідної функції щільності ймовірностей Пірсон використав лише перший член ряду Маклорена b_0, вважаючи, що інші його члени дорівнюють нулю, тобто

$$\frac{dy}{dz} = \frac{y(z+d)}{b_0}. \qquad (4.8)$$

Звернемося до загального розв’язку системи рівнянь (4.7).

1.Оскільки b_1=0, математичне сподівання і мода співпадають, тобто $m_x = m_0$.

2. Параметр $-b_0$ дорівнює дисперсії випадкової величини.

3. Оскільки β_3=0, то C_S=0, тобто розподіл симетричний.

Розв'язання диференціального рівняння (4.8) зводиться до такого

$$\frac{dy}{dz} = \frac{yz}{b_0} \; ; \tag{4.9}$$

$$\int \frac{dy}{y} = \int -\frac{z}{\sigma_z^2} dz ; \tag{4.10}$$

$$ln|y| = -\frac{1}{\sigma_z^2}\frac{z^2}{2} + c \tag{4.11}$$

Але ж $y = f(z) \geq 0$, тоді $|y| = y$ та $ln|y| = ln\, y$, тобто

$$ln\, y = -\frac{1}{\sigma_z^2}\frac{z^2}{2} + c . \tag{4.12}$$

У точці $z = 0$ величина $c = ln\, y$. Позначимо її як $\ln y_0$. Це значення буде максимальним через те, що для всіх інших z виконується нерівність $ln\, y < ln\, y_0$. Крім того, якщо d=0, то значення k, відносно якого ведуться розрахунки, буде співпадати з модою, тобто мати найбільшу щільність ймовірності. З урахуванням цього прийдемо до виразу

$$y = y_0 e^{-\frac{z^2}{2\sigma_z^2}} . \tag{4.13}$$

З властивостей щільності ймовірності витікає, що

$$\int_{-\infty}^{\infty} f(z)dz = 1. \tag{4.14}$$

Отже,

$$y_0 \int_{-\infty}^{\infty} e^{-\frac{z^2}{2\sigma_z^2}} dz = 1. \tag{4.15}$$

Після деяких перетворень ($t = \frac{z}{\sqrt{2\sigma_z^2}} = \frac{z}{\sigma_z\sqrt{2}}$; $dz = \sigma_z\sqrt{2}dt$) отримаємо, що

$$y_0 = \frac{1}{\sigma_z\sqrt{2\pi}}. \tag{4.16}$$

Загальний вигляд рівняння кривої нормального розподілу такий

$$f(z) = \frac{1}{\sigma_z\sqrt{2\pi}} e^{-\frac{z^2}{2\sigma_z^2}}. \tag{4.17}$$

Враховуючи, що $m_z = 0$, а $\sigma_z = C_V$, можна записати, що

$$y = \frac{1}{\sigma_z\sqrt{2\pi}} e^{-\frac{z^2}{2\sigma_z^2}} = \frac{1}{C_V\sqrt{2\pi}} e^{-\frac{(k-1)^2}{2C_V^2}}. \tag{4.18а}$$

Для вихідного ряду x

$$f(x) = \frac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{(x-m_x)^2}{2\sigma_x^2}}. \tag{4.18б}$$

Якщо замість значень x використати центровані та нормовані значення випадкової величини X , представлені у вигляді

$$t = \frac{x - m_x}{\sigma_x} = \frac{k-1}{C_V}, \tag{4.19}$$

і , ураховуючи, що $\sigma_t = 1$, отримаємо

$$f(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}. \tag{4.20}$$

Вираз (4.20) задається у вигляді таблиці для практичного використання (табл.1.2).

Таблиця 4.1 – Значення функції $y = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$

(t)	y	(t)	y	(t)	y	(t)	y
0,0	0,399	1,0	0,242	2,0	0,054	3,0	0,004
0,1	0,397	1,1	0,218	2,1	0,044	3,2	0,0024
0,2	0,391	1,2	0,194	2,2	0,036	3,4	0,0012
0,3	0,381	1,3	0,171	2,3	0,028	3,6	0,0009
0,4	0,368	1,4	0,150	2,4	0,022	3,7	0,0004
0,5	0,352	1,5	0,130	2,5	0,018	3,8	0,0003
0,6	0,333	1,6	0,111	2,6	0,014	3,9	0,0002
0,7	0,312	1,7	0,094	2,7	0,010	4,0	0,0001
0,8	0,290	1,8	0,079	2,8	0,008		
0,9	0,266	1,9	0,066	2,9	0,006		

Нормальний розподіл має декілька особливостей.

1. У зв'язку з тим, що функція (4.20) має дійсні значення при будь-яких значеннях незалежної змінної X, область її визначення така: $-\infty < X < +\infty$.

2. Функція *f(x)* є парною, тобто *f(-X)=f(+X),* а нормальна крива розподілу симетрична відносно осі ординат.

3. Нормальна крива розподілу не перетинає осі *x*.

4. Крива щільності розподілу симетрична відносно моди.

5. Нормальний закон розподілу випадкової величини є двох параметричним, тобто в ньому використовуються два параметри – математичне сподівання m_X та дисперсія σ_X^2.

6. Параметр σ_x є характеристикою форми кривої розподілу: чим більше σ_x, тим максимальна ордината менше, а крива сплющується (рис.4.2)

7. Якщо змінювати m_X, то крива щільності розподілу буде переміщуватися уздовж осі x, зберігаючи свою форму.

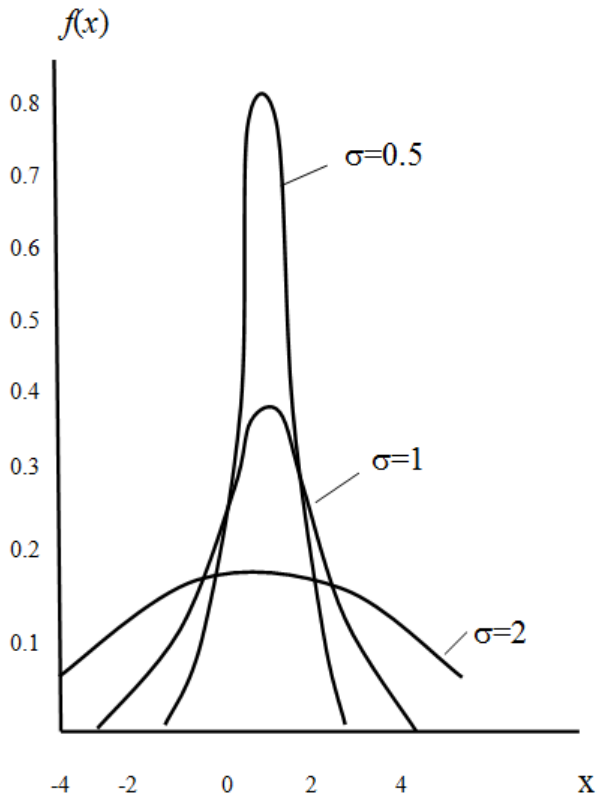


Рисунок 4.2 – Графік щільності ймовірностей нормального закону розподілу при різних середніх квадратичних відхиленнях σ_x

Вперше нормальний закон розподілу випадкових величин був розроблений для аналізу похибок вимірювань. На цій основі він і отримав розповсюдження у багатьох галузях науки і техніки, в тому числі і в природничих науках, де широко використовується для оцінки точності розрахунків, визначення довірчих інтервалів і таке інше. У практичних дослідженнях іноді допускається використовувати його до деяких рядів з асиметричним розподілом. Однак, похибка прийнятого припущення про нормальність розподілу може привести до неправильних вирішень поставлених задач.

4.3 Закон розподілу Пірсона III

Диференціальне рівняння Пірсона III (4.2) може бути розписаним для перших двох членів ряду Маклорена. Для нормального закону розподілу була прийнята жорстка умова про те, що розподіл симетричний, тобто мода розподілу співпадає з математичним сподіванням і радіус асиметрії *d=0.*

Припустимо, що випадкова величина X додатна, отже $x_{min} \geq 0$. Позначимо через a відстань від m_x до моди (див. рис. 4.1). Тоді сума $a + d$ є відстанню від математичного сподівання m_x до x_{min}, звідки

$$a + d = m_x - x_{min}. \tag{4.21}$$

Якщо замість *x* використовувати модуль стоку, то отримаємо

$$a + d = 1 - k_{min}, \tag{4.22}$$

оскільки $k_{min} \geq 0$, то $a + d \leq 1$.

Загальний вигляд кривої Пірсона III при $C_S > 0$ (додатній асиметрії) показаний на рис.4.1.

Рівняння розподілу Пірсона III типу можна вивести, якщо розглянути два перших члени ряду Маклорена (b_0 та b_1), тобто

$$\frac{dy}{dz} = \frac{y(z + d)}{b_0 + b_1 z}. \tag{4.23}$$

За умови $b_2 = b_3 ... = 0$, система рівнянь (4.7) приймає вигляд

$$\begin{cases} -b_1 = d \\ -b_0 = \beta_2 \\ -3b_1\beta_2 = \beta_3 + d\beta_2 \end{cases}. \tag{4.24}$$

З системи рівнянь (4.24) витікає, що

$$d = \frac{\beta_3}{2\beta_2}. \tag{4.25}$$

Після інтегрування (4.23) отримаємо наступний вираз

$$f(z) = ce^{\frac{z}{b_1}}(b_0 + b_1 z)^{\frac{1}{b_1}\left(d - \frac{b_0}{b_1}\right)}. \tag{4.26}$$

При мінімальному значенні z маємо $f(z) = 0$ і оскільки $b_1 \neq 0$ (асиметричний розподіл), то

$$b_0 + b_1 z_{\min} = 0, \tag{4.27}$$

звідки

$$z_{min} = -\frac{b_0}{b_1}, \tag{4.28}$$

Оскільки відстань від $z_{\min}$ до центру розподілу m_k дорівнює $a + d$ можна установити зв'язок між параметрами рівняння (4.26) та графічними характеристиками розподілу, що знайшло своє відображення у рівнянні виду (В.А. Шелутко, 1984, 1991)

$$f(z) = y = y_0 e^{-z/d}(1 + z/a)^{a/d}, \tag{4.29}$$

де z відраховується від моди (m_o);

a - відстань від мінімального значення до моди ;

y_0 - модальна ордината.

Параметри рівняння b_0, b_1 знаходять за допомогою методу моментів.

Установлено, що параметри рівняння (4.29) зв`язані з центральними моментами розподілу таким чином

$$a + d = 2\beta_2^2 / \beta_3. \tag{4.30}$$

Для практичного застосування в (4.30) β_2 та β_3 замінюють на статистичні параметри m_x, C_V, C_S. Знаючи, що $\beta_2 = \sigma_x^2$, а $C_V = \sigma_x / m_x$, запишемо

$$\beta_2 = C_V^2 m_x^2. \tag{4.31}$$

Третій центральний момент, як було показано, зв`язаний з коефіцієнтом асиметрії C_S та середнім квадратичним відхиленням в такому вигляді

$$\beta_3 = C_S \sigma_x^3 = C_S C_V^3 m_x^3. \tag{4.32}$$

Підставляючи (4.31) та (4.32) в (4.30), отримуємо

$$a + d = 2C_V^4 m_x^4 / (C_S C_V^3 m_x^3) = 2C_V m_x / C_S. \tag{4.33}$$

В загальному вигляді співвідношення між x_{min} та іншими статистичними параметрами розподілу записується у вигляді

$$m_x - x_{min} = 2C_V m_x / C_S, \tag{4.34}$$

або при використанні модульних коефіцієнтів

$$1 - k_{min} = 2C_V / C_S. \tag{4.35}$$

де $k_{min} = x_{min}/m_x$ - модульний коефіцієнт мінімального значення ряду.

Згідно із (4.35) розглянемо три можливих варіанти для k_{min}:

$$1)\ \text{якщо}\ k_{min} = 0,\ \text{то}\ a + d = 1 - k_{min} = \frac{2C_V}{C_S} = 1,\ \text{тобто}\ C_S = 2C_V; \tag{4.36}$$

$$2)\ \text{якщо}\ k_{min} > 0,\ \text{то}\ C_S = \frac{2C_V}{1 - k_{min}},\ \text{звідки}\ C_S > 2C_V; \tag{4.37}$$

$$3)\ \text{якщо}\ k_{min} < 0,\ C_S < 2C_V. \tag{4.38}$$

Третій варіант суперечить природі рядів стоку та концентрацій хімічних речовин, які завжди додатні, отже крива розподілу Пірсона III може використовуватись тільки тоді, коли $C_S \geq 2C_V$.

Головні властивості розподілу Пірсона III наступні: крива розподілу обмежена нижньою *(x*=0*)* і не обмежена верхньою границями. При $x \to \infty$ крива наближається до осі абсцис та описується трьома статистичними параметрами: m_x, C_V, C_S, а при $x_{min} = 0$, *коли* $C_S = 2C_V$, є двопараметричним розподілом.

4.4 Логарифмічно-нормальний закон розподілу

Нормальний закон розподілу використовується для випадкових величин, які змінюються в межах від -∞ до +∞. З метою застосування нормального закону розподілу до рядів спостережених величин, область визначення яких змінюється від 0 до +∞, необхідно перетворити вихідну змінну (А.В. Рождественський, А.І. Чеботарьов, 1974).

У нашому розпорядженні є суттєво додатня випадкова величина X з областю визначення $0 \le x < \infty$, несиметрично розподілена відносно центру розподілу m_x. Перетворимо її у нову, яка підлягає нормальному закону розподілу. З цією метою введемо нову змінну $u = ln\, x$. Якщо $0 \le x < +\infty$, то $-\infty < ln\, x < +\infty$. Тоді рівняння нормального закону розподілу набере вигляду

$$y = f(u) = \frac{1}{\sigma_u \sqrt{2\pi}} e^{-\frac{(u-m_u)^2}{2\sigma_u^2}}. \qquad (4.39)$$

Статистичні параметри випадкової величини u оцінюються за даними ряду стоку таким чином

$$\hat{m}_u = \frac{\sum_{i=1}^{n} ln\, x_i}{n} = \hat{m}_{ln\, x}; \qquad (4.40)$$

$$\hat{\sigma}_{\ln u} = \sqrt{\frac{\left(\sum_{i=1}^{n} \left(\ln x_i - \hat{m}_{\ln x}\right)^2\right)}{n-1}}. \qquad (4.41)$$

Для визначення ординат кривої розподілу використовують таблицю нормального закону розподілу (табл.4.1).

Можливій інший варіант розрахунків, за яким на базі нормального розподілу отримується новий закон, який можна застосовувати до величин стоку.

Нехай нам відомий закон розподілу деякої величини u. Необхідно отримати закон розподілу величини x, розподіл якої невідомий, але x функціонально зв'язана з u, тобто

$$x = \varphi(u), \text{ а } u = \psi(x). \qquad (4.42)$$

Тоді новий закон розподілу представляється у вигляді

$$f(x) = f(u)\frac{du}{dx}. \qquad (4.43)$$

Якщо розподіл $u=lnx$ описується нормальним законом, то розподіл випадкової величини x можна представити таким чином

$$\frac{du}{dx} = \frac{1}{x}, \qquad (4.44)$$

$$f(x) = f(u)\frac{1}{x}. \qquad (4.45)$$

Рівняння логарифмічно-нормального закону розподілу запишеться у вигляді

$$f(x) = \frac{1}{\sigma_{\ln x}\sqrt{2\pi}} e^{-\frac{(\ln x - m_{\ln x})^2}{2\sigma^2_{\ln x}}} \frac{1}{x}. \qquad (4.46)$$

Для визначення ординат кривої розподілу використовують таблицю логарифмічно-нормального закону розподілу.

Логарифмічно-нормальний закон розподілу використовується для рядів з високим коефіцієнтом асиметрії, а саме

$$C_S = C_V(3 + C_V). \qquad (4.47)$$

4.5 Розрахунки характеристик стоку заданої забезпеченості за теоретичними законами розподілу

Емпіричні криві забезпеченості не дають можливості визначити величини стоку за межами вихідної інформації. Безпосередньо графічна екстраполяція емпіричної кривої забезпеченості у область малих (p < 5%) та великих (p > 95%) значень забезпеченості має суб'єктивний характер і може привести до значних похибок у розрахунках. Теоретичні закони розподілу у даному випадку "вирівнюють" емпіричний розподіл стокової величини та екстраполюють його за межі спостережених даних.

Найчастіше застосовуються закон Пірсона Ш та логарифмічно-нормальний. Ці теоретичні закони розподілу представлені у вигляді таблиць. Наприклад, ординати теоретичної кривої забезпеченості Пірсона Ш надаються у вигляді нормованих відхилень від середньої величини стоку при заданому C_V

$$\Phi\left(P, C_s\right) = \frac{x_P - \bar{x}}{\sigma_x} = \frac{k_P - 1}{C_v}. \qquad (4.48)$$

Щоб скористатися таблицями теоретичних розподілів, спочатку необхідно визначити оцінки статистичних параметрів за матеріалами спостережень.

Розрахунки характеристик стоку за теоретичним розподілом Пірсона Ш виконуються за виразом

$$x_P = \left(\Phi_P C_v + 1\right)\bar{x} \ . \qquad (4.49)$$

Питання для самоперевірки

1. Які існують вимоги до теоретичних кривих розподілу величин?
2. Які особливості має нормальний закон розподілу?
3. За яких умов може використовуватися крива розподілу Пірсона ІІІ?
4. У чому полягають головні властивості розподілу Пірсона ІІІ?
5. Який закон розподілу застосовується до рядів спостережених величин, область визначення яких змінюється від 0 до +∞?
6. Як знайти ординати теоретичної кривої забезпеченості Пірсона Ш?
7. Яку функцію виконує теоретичний закон розподілу?
8. Через що виникають похибки у розрахунках при використанні емпіричної кривої?

5 ОСНОВИ ТЕОРІЇ КОРЕЛЯЦІЇ

5.1 Залежні та незалежні події. Теорема множення ймовірностей

Подія A називається незалежною від події B, якщо ймовірність події A не залежить від того, відбулася подія B чи ні (О.С. Вентцель, 1969).

Подія A називається залежною від події B, якщо ймовірність події A змінюється в залежності від того, відбулася подія B чи ні. ***Ймовірність події A, обчисленої за умови, що мала місце інша подія B, називається умовною ймовірністю події A та позначається як $p(A/B)$.***

Ймовірність добутку двох подій дорівнює добутку ймовірності однієї з них на умовну ймовірність другої, визначеної за умови, що перша мала місце

$$p(AB) = p(A)\,p(B/A); \tag{5.1}$$

або

$$p(AB) = p(B)\,p(A/B)\ . \tag{5.2}$$

Доведемо теорему множення для схеми випадків. Нехай можливі результати випробувань становлять n випадків. Припустимо, що події A сприяють m випадків, а події B - k випадків. Існують випадки, які сприяють події A та події B одночасно, кількість таких випадків становить L (рис. 1.15).

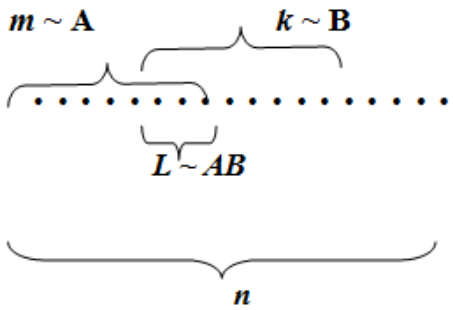


Рисунок 5.1 – Схема випадків для доведення теореми множення

Ймовірність сумісної появи подій A та B становить

$$p(AB)=\frac{L}{n}; \tag{5.3}$$

а ймовірності появи кожної з подій дорівнюють

$$p(A)=\frac{m}{n}; p(B)=\frac{k}{n}. \tag{5.4}$$

Обчислимо $p(B/A)$, тобто умовну ймовірність події B, припускаючи, що подія A мала місце. Якщо відомо, що подія A відбулася, то з усіх раніше можливих n випадків залишаються можливими тільки ті m, які сприяли події A. З них L випадків сприяють події B. Отже,

$$p(B/A)=\frac{L}{m}. \tag{5.5}$$

Підставляючи вираз $p(AB)$, $p(A)$ та $p(B/A)$ в формулу (1.183), отримаємо тотожність

$$\frac{L}{n}\equiv\frac{m}{n}\cdot\frac{L}{m}. \tag{5.6}$$

Наслідки теореми множення такі.

Декілька подій називаються незалежними, коли будь-яка з них не залежить від будь-якої сукупності решти подій.

Ймовірність добутку двох незалежних подій дорівнює добутку ймовірності цих подій.

5.2 Закони розподілу системи випадкових величин

Якщо результат випробування описується не однією випадковою величиною, а декількома, то говорять, що вони утворюють систему випадкових величин.

В процесі узагальнення результатів спостережень при вирішенні інженерно-гідрологічних задач часто доводиться мати справу з декількома (двома чи більше) випадковими величинами, що утворюють систему. Наприклад, зливові опади описуються не однією, а як мінімум двома випадковими величинами: тривалістю дощу та його інтенсивністю. Течія води характеризується швидкістю і напрямком. Гідрологохімічний режим

водної екосистеми може описуватися показниками якості води декількох річок та ін.

При вивченні системи випадкових величин вже не можна обмежуватися дослідженням властивостей окремих складових: необхідно враховувати залежності між цими складовими.

Системи випадкових величин, як і просто випадкова величина, описуються законами розподілу: інтегральним, диференціальним.

Як і для незалежних випадкових величин, можливий повний опис ймовірнісних властивостей системи шляхом побудови її закону розподілу і частковий, заснований на обчисленні окремих числових характеристик.

Інтегральна функція розподілу системи двох випадкових величин X і Y є ймовірністю сумісного виконання двох нерівностей: $X < x, Y < y$; тобто

$$F(x,y) = p\big((X < x),(Y < y)\big). \tag{5.7}$$

Основні властивості функції розподілу системи двох випадкових величин $F(x,y)$ такі:

$F(x,y)$ є неспадною функцією обох аргументів, тобто

- при $x_2 > x_1 \quad F(x_2, y) \geq F(x_1, y);$ (5.8)

- при $y_2 > y_1 \quad F(x, y_2) \geq F(x, y_1).$ (5.9)

При $x \to -\infty$ або $y \to -\infty$ інтегральна функція розподілу дорівнює нулю

$$F(x,-\infty) = F(-\infty, y) = F(-\infty,-\infty) = 0. \tag{5.10}$$

При одному з аргументів, який дорівнює $+\infty$, функція розподілу системи перетворюється на функцію розподілу випадкової величини, що відповідає іншому аргументу

$$F(x,+\infty) = F_1(x), \quad F(+\infty, y) = F_2(y), \tag{5.11}$$

де $F_1(x)$ і $F_2(y)$ - відповідно функції розподілу випадкових величин X та Y.

Якщо обидва аргументи відповідають $+\infty$, то функція розподілу системи дорівнює одиниці

$$F(+\infty,+\infty)=1. \tag{5.12}$$

Розподіл системи безперервних випадкових величин можна охарактеризувати не тільки інтегральною функцією розподілу, але й *щільністю розподілу*, яка розглядається як приріст інтегральної функції на відрізках Δx та Δy

$$\lim_{\substack{\Delta x\to 0\\ \Delta y\to 0}} \frac{F(x+\Delta x,y+\Delta y)-F(x+\Delta x,y)-F(x,y+\Delta y)+F(x,y)}{\Delta x\,\Delta y}\ . \tag{5.13}$$

Якщо $F(x,y)$ не тільки безперервна, але й диференціюється, то вираз (5.13) являє собою другу змішану частинну похідну функції $F(x,y)$ по x та по y. Позначимо цю похідну як $f(x,y)$

$$f(x,y)=\frac{\partial^2 F(x,y)}{\partial x\partial y} \tag{5.14}$$

і назвемо щільністю розподілу системи.

Для $f(x,y)$ можлива геометрична інтерпретація у вигляді деякої поверхні, яка називається *поверхнею розподілу* (рис. 5.1).

Між інтегральною функцією системи двох випадкових величин та диференціальною функцією існує зв'язок

$$F(x,y)=\int_{-\infty}^{x}\int_{-\infty}^{y} f(x,y)dxdy. \tag{5.15}$$

Звернемо увагу на дві властивості щільності розподілу системи випадкових величин:

1. Щільність розподілу системи є функцією невід'ємною

$$f(x,y)\geq 0\ . \tag{5.16}$$

2. Подвійний інтеграл у нескінченних границях від щільності розподілу системи дорівнює одиниці

$$\int_{-\infty}^{+\infty}\int f(x,y)dxdy=1\ . \tag{5.17}$$

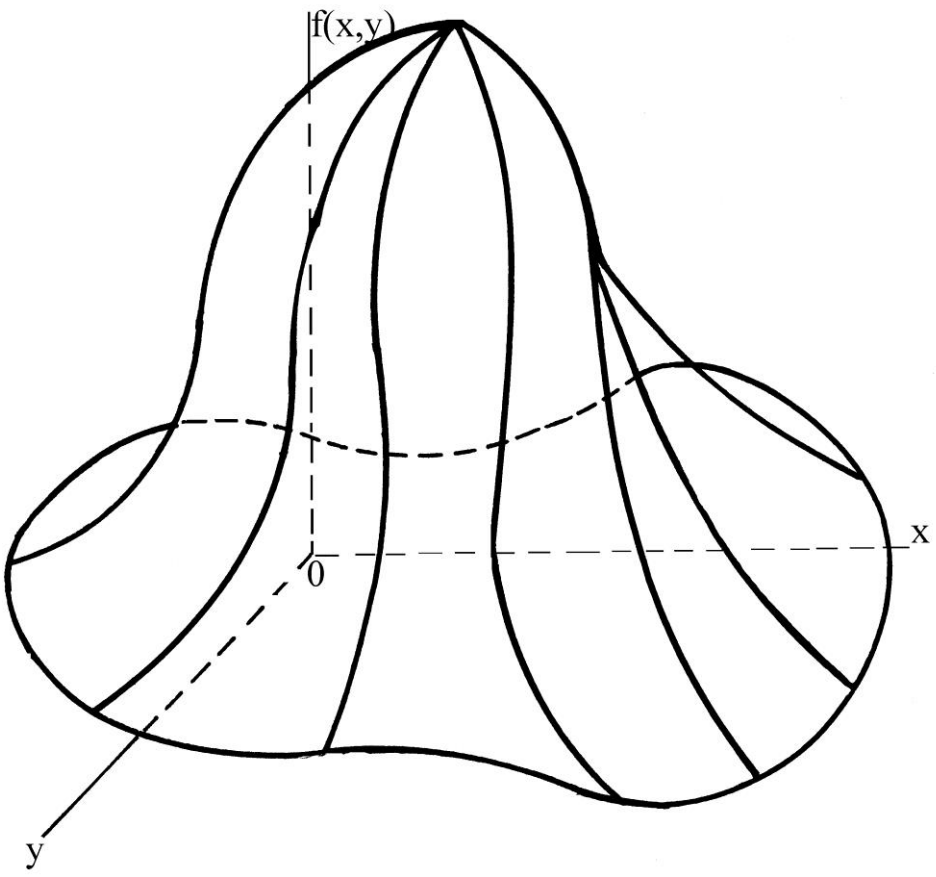


Рисунок 5.1 – Поверхня розподілу двох випадкових величин

5.3 Умовні закони розподілу випадкових величин

Як відзначалося в попередньому розділі, за законом розподілу системи двох випадкових величин можна визначити розподіл окремих величин, що входять до системи

$$F_1(x) = F(x,\infty);\ F_2(y) = F(\infty,y) \tag{5.18}$$

чи інакше

$$F_1(x) = F(x,\infty) = \int_{-\infty}^{x}\int_{-\infty}^{\infty} f(x,y)dxdy. \tag{5.19}$$

Після диференціювання (5.19) по x, одержимо вираз для щільності розподілу величини X:

$$f_1(x) = F_1'(x) = \int_{-\infty}^{\infty} f(x,y)dy. \qquad (5.20)$$

Щільність розподілу величини Y запишемо аналогічно:

$$f_2(y) = F_2'(y) = \int_{-\infty}^{\infty} f(x,y)dx. \qquad (5.21)$$

Формули (5.20) та (5.21) показують, що для того, щоб одержати розподіл окремої величини, яка входить до системи, необхідно проінтегрувати щільність розподілу системи по всій області визначення аргументу, який відповідає іншій випадковій величині.

Обернену задачу в загальному випадку вирішити неможна: знаючи розподіл окремих величин, що входять до системи, не завжди можна знайти закон розподілу системи.

При цьому недостатньо знати розподіл кожної з величин, потрібно знати залежність між ними, яку можна охарактеризувати за допомогою умовних законів розподілу.

Умовним законом розподілу величини X, яка входить до системи (X,Y), називається закон її розподілу, обчислений за умови, що інша випадкова величина Y набула визначеного значення y.

Умовним законом розподілу величини Y, яка входить до системи (X,Y), називається закон її розподілу, обчислений за умови, що інша випадкова величина X набула визначеного значення x.

Умовний закон розподілу, який можна задати інтегральною функцією розподілу або щільністю, позначається в такий спосіб:

$$F(x/y) \text{ чи } f(x/y). \qquad (5.22)$$

Теорема множення для системи двох випадкових величин записується таким чином

$$f(x,y) = f_1(x) \cdot f(y/x) \qquad (5.23)$$

або

$$f(x,y) = f_2(y) \cdot f(x/y). \qquad (5.24)$$

Умовний закон розподілу кожної з випадкових величин представляється у вигляді

$$f(y/x)=\frac{f(x,y)}{f_1(x)}, \tag{5.25}$$

$$f(x/y)=\frac{f(x,y)}{f_2(y)}. \tag{5.26}$$

Зміст формул (5.25) та (5.26) описується таким чином: **щільність розподілу системи двох випадкових величин дорівнює щільності розподілу однієї з них, помноженої на умовну щільність іншої, обчисленої за умови, що перша величина набула конкретного значення.**

5.4 Залежні і незалежні випадкові величини, кореляційний момент, коефіцієнт кореляції

Характер залежності випадкових величин, які утворюють систему, може бути різним. У деяких випадках залежність може бути настільки тісною, що, знаючи значення однієї випадкової величини, можна точно вказати значення іншої. Така надійна залежність близька до функціональної. ***Якщо ж, знаючи значення випадкової величини X, не можна точно вказати значення випадкової величини Y, а тільки закон її розподілу в залежності від значення, якого набула величина X, то така залежність називається ймовірнісною чи стохастичною.*** Іншими словами, у стохастичній залежності зміна випадкової величини X зумовлює зміну розподілу випадкової величини Y. Цей розподіл можна описати рівнянням умовного закону розподілу вигляду (5.25). Функціональна залежність розглядається при цьому як крайній, граничний випадок ймовірнісної залежності. В іншому крайньому випадку залежність між випадковими величинами може бути слабкою настільки, що розглядувані величини можна вважати незалежними.

Сформулюємо найважливіше для подальшого викладення поняття про незалежні випадкові величини. ***Випадкова величина Y називається незалежною від випадкової величини X, якщо закон розподілу Y не залежить від того, якого значення набула величина X***

$$f(y/x)= f_2(y). \tag{5.27}$$

Якщо Y залежить від X, то

$$f(y/x) \neq f_2(y). \tag{5.28}$$

Для незалежних випадкових величин теорема множення законів розподілу набуває вигляду:

$$f(x,y) = f_1(x) \cdot f_2(y). \tag{5.29}$$

Ймовірнісна залежність між випадковими величинами може бути більш-менш тісною.

Числовою характеристикою тісноти зв'язку в системі двох випадкових величин є коваріаційний момент, який описується співвідношенням

$$K_{xy} = \int_{-\infty}^{\infty}\int (x - m_x)(y - m_y) f(x,y)dxdy. \tag{5.30}$$

Оскільки для незалежних випадкових величин щільність ймовірності системи випадкових величин дорівнює добутку безумовних ймовірностей кожної з цих величин (5.29), інтеграл (5.30) можна представити як добуток двох інтегралів

$$K_{xy} = \int_{-\infty}^{\infty} (x - m_x) f_1(x)dx \int_{-\infty}^{\infty} (y - m_y) f_2(y)dy \ , \tag{5.31}$$

де інтеграл

$$\int_{-\infty}^{\infty} (x - m_x) f_1(x)dx \tag{5.32}$$

являє собою ніщо інше, як перший центральний момент, який дорівнює нулю. З тієї же причини дорівнює нулю і другий співмножник, звідки можна зробити висновок, що ***для незалежних випадкових величин*** $K_{xy} = 0$. Таким чином, якщо кореляційний момент двох випадкових величин відмінний від нуля, то це є ознакою наявності залежності між ними. На практиці зазвичай використовується не сам кореляційний момент, для якого також вживають термін ***коваріація*** $cov(x,y)$, а його нормоване (безрозмірне) значення, яке називають ***коефіцієнтом кореляції***

$$r_{xy} = \frac{K_{xy}}{\sigma_x \sigma_y} \ , \tag{5.33}$$

причому

$$-1 \le r_{xy} \le 1 \ , \tag{5.34}$$

де σ_x, σ_y - середні квадратичні відхилення величин X, Y.

При $K_{xy} = 0$ величина r_{xy} також дорівнює 0, тобто ***для незалежних випадкових величин коефіцієнт кореляції дорівнює нулю***. У той же час, для функціонально зв'язаних величин $|r_{xy}| = 1.0$. Випадкові величини, для яких коваріація і коєффіцієнт кореляції дорівнюють нулю, називаються *некорельованими*. Однак поняття некорельованості та незалежності випадкових величин не є еквівалентними. З незалежності випадкових величин випливає їхня некорельованість, але з некорельованості не випливає їхня незалежність. Умова незалежності випадкових величин більш жорстка, ніж умова некорельованості, оскільки коефіцієнт кореляції цілком характеризує не будь-яку залежність, а тільки так звану лінійну. Лінійна стохастична залежність випадкових величин полягає в тім, що при зростанні однієї випадкової величини інша теж має тенденцію до зростання чи убування за лінійним законом

$$M[Y / X = x] = \varphi(x) \ , \tag{5.35}$$

де $\varphi(x)$ - лінійна функція x.

Коефіцієнт кореляції характеризує ступінь наближення кореляційного зв'язку між випадковими величинами X та Y до функціональної залежності. Зв'язок між випадковими величинами тим тісніший, чим більший за абсолютною величиною коефіцієнт кореляції. Наявність додатного значення коефіцієнта кореляції між двома випадковими величинами означає, що при зростанні однієї з них інша має тенденцію в середньому зростати $(0 < r_{xy} < 1)$; від'ємного - що при зростанні однієї з випадкових величин інша має тенденцію в середньому убувати $(-1 < r_{xy} < 0)$. Кореляційний зв'язок між двома випадковими величинами X та Y при $0 < r_{xy} < 1$ називають прямим, а при $-1 < r_{xy} < 0$ - оберненим.

5.5 Нормальний закон розподілу для системи двох випадкових величин, рівняння лінійної регресії

Найбільше поширення в практичному використанні має так званий класичний нормальний закон розподілу системи двох нормально-розподілених випадкових величин. У загальному випадку щільність нормального розподілу двох випадкових величин виражається формулою:

$$f(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-r_l^2}}e^{-\frac{1}{2(1-r_l^2)}\left[\frac{(x-m_x)^2}{\sigma_x^2}-\frac{2r_l(x-m_x)(y-m_y)}{\sigma_x\sigma_y}+\frac{(y-m_y)^2}{\sigma_y^2}\right]},\quad(5.36)$$

де $r_н$ - коефіцієнт кореляції між двома випадковими, нормально розподіленими величинами.

Цей закон залежить від п’яти параметрів: $m_x, m_y, \sigma_x, \sigma_y, r_н$.

Припустимо, що випадкові величини X і Y підлягають нормальному закону і некорельовані, тобто $r_н$=0. Тоді вираз (2.217) набере вигляду:

$$f(x,y)=\frac{1}{2\pi\sigma_x\sigma_y}e^{-\frac{(x-m_x)^2}{2\sigma_x^2}-\frac{(y-m_y)^2}{2\sigma_y^2}}. \quad (5.37)$$

Легко переконатися в тім, що випадкові величини X і Y, підпорядковані закону розподілу зі щільністю (5.37), не тільки некорельовані, але й незалежні. Дійсно, (5.37) можна представити як добуток щільностей розподілу окремих величин, що входять до системи

$$f(x,y)=\frac{1}{\sigma_x\sqrt{2\pi}}e^{-\frac{(x-m_x)^2}{2\sigma_x^2}}\cdot\frac{1}{\sigma_y\sqrt{2\pi}}e^{-\frac{(y-m_y)^2}{2\sigma_y^2}}, \quad (5.38)$$

яка відповідає теоремі множення законів розподілу (5.29), що розглядається як необхідна і достатня умова незалежності випадкових величин.

Таким чином, для системи двох випадкових величин, підпорядкованих нормальному закону, з некорельованості величин випливає їхня незалежність.

Визначаючи умовний закон розподілу величини Y при $r_н \neq 0$ відповідно до виразу (5.25), одержимо:

$$f(y/x)=\frac{1}{\sigma_y\sqrt{1-r_н^2}\sqrt{2\pi}}e^{-\frac{1}{2(1-r_н^2)}\left[\frac{y-m_y}{\sigma_y}-r_н\frac{x-m_x}{\sigma_x}\right]^2}. \qquad (5.39)$$

Приведемо вираз (5.39) до вигляду:

$$f(y/x)=\frac{1}{\sigma_y\sqrt{1-r_н^2}\sqrt{2\pi}}e^{-\frac{1}{2\sigma_y^2(1-r_н^2)}\left[y-m_y-r_н\frac{\sigma_y}{\sigma_x}(x-m_x)\right]^2}. \qquad (5.40)$$

Очевидно, що це є щільність розподілу нормального закону з центром розсіювання, який визначається таким чином

$$m_{y/x}=m_y+r_н\frac{\sigma_y}{\sigma_x}(x-m_x) \qquad (5.41)$$

Середнє квадратичне відхилення установлюється за формулою

$$\sigma_{y/x}=\sigma_y\sqrt{1-r_н^2}\,. \qquad (5.42)$$

З формул (5.41) і (5.42) випливає, що умовний закон розподілу величини Y характеризується тільки зміною математичного сподівання при зміні x, а її умовна дисперсія від x не залежить.

Величина $m_{y/x}$ називається умовним математичним сподіванням величини Y при даному x. Залежність (5.41) можна зобразити на площині xOy, відкладаючи умовне математичне сподівання по осі ординат. У відповідності з (5.41), на рисунку отримаємо пряму лінію, яка називається лінією регресії Y по X (рис. 5.2).

Рівняння умовного математичного сподівання (5.41) називають рівнянням регресії Y по X.

Рівняння регресії X по Y можна представити у вигляді

$$m_{x/y}=m_x+r_н\frac{\sigma_x}{\sigma_y}(y-m_y). \qquad (5.43)$$

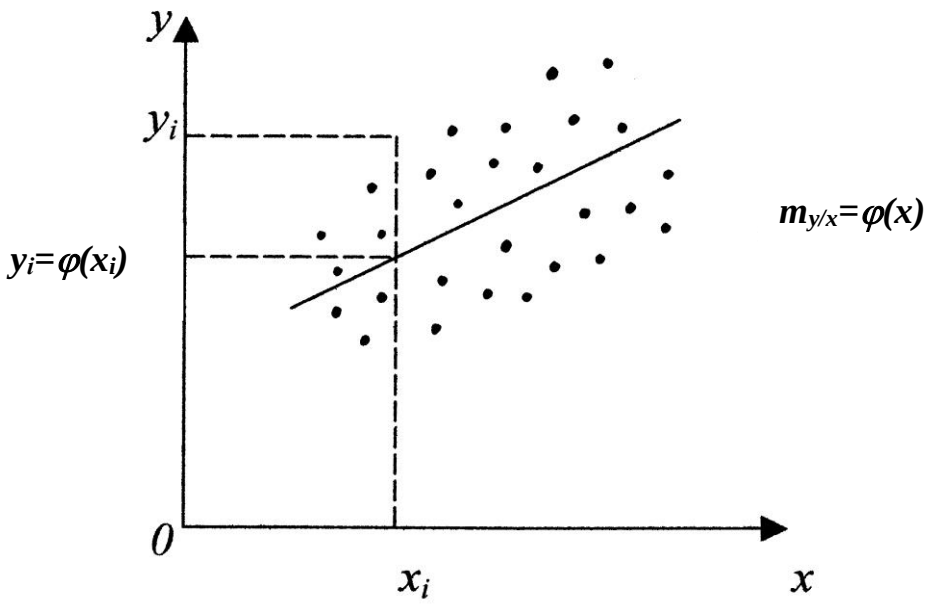


Рисунок 5.2 – Графічний вигляд рівняння лінійної регресії $m_{y/x} = \varphi(x)$

Рівняння (5.41) та (5.43) можна привести до вигляду рівнянь, які описують прямі

$$m_{y/x} = \alpha_1 x + \beta_1 , \tag{5.44}$$

$$m_{x/y} = \alpha_2 y + \beta_2 , \tag{5.45}$$

де $\alpha_1, \beta_1, \alpha_2, \beta_2$ - коефіцієнти рівнянь прямих, причому

$$\alpha_1 = r_н \frac{\sigma_y}{\sigma_x}; \quad \beta = m_y - \alpha_1 m_x ; \tag{5.46}$$

$$\alpha_2 = r_н \frac{\sigma_x}{\sigma_y}; \quad \beta_1 = m_x - \alpha_2 m_y , \tag{5.47}$$

де величини m_x, m_y - це безумовні математичні сподівання випадкових величин X і Y, тобто центри їх розподілу без урахування залежності між X та Y.

Лінії регресії (5.46) та (5.47) співпадають тільки у випадку існування лінійної функціональної залежності між величинами Y та X. Для незалежних X і Y лінії регресії паралельні осям координат. Якби величини Y та X були пов'язані одна з одною функціональною залежністю, то кожному фіксованому $X = x$ відповідало б тільки одне значення $Y = y$. Але у випадку нефункціонального зв'язку кожному значенню $X = x$ відповідають значення y, що з'являються сумісно для заданого x і групуються навколо центра m_y. Тобто кожна точка на лінії (рис. 5.2) є центром розподілу залежної змінної y при заданому x. Коли змінюється x, то змінюються за умовним законом розподілу і значення y, але ці зміни лише частково відображають зміни x. Теж саме відбувається і під час розгляду регресії X по Y. Кожному $Y = y$ відповідає змінна X, значення якої групуються навколо величини m_x.

Якщо ми будемо змінювати y, то зміниться й x, та ця зміна знов таки буде лише частково реагувати на відхилення y. Отже, реакція y на зміну x та реакція x на зміну y не однакова, тому і лінії регресії (5.44) та (5.45) не співпадають.

Стохастичну залежність можна представити графічно у вигляді еліпса розсіювання корелятивно зв'язаних випадкових величин Y та X (рис.5.3).

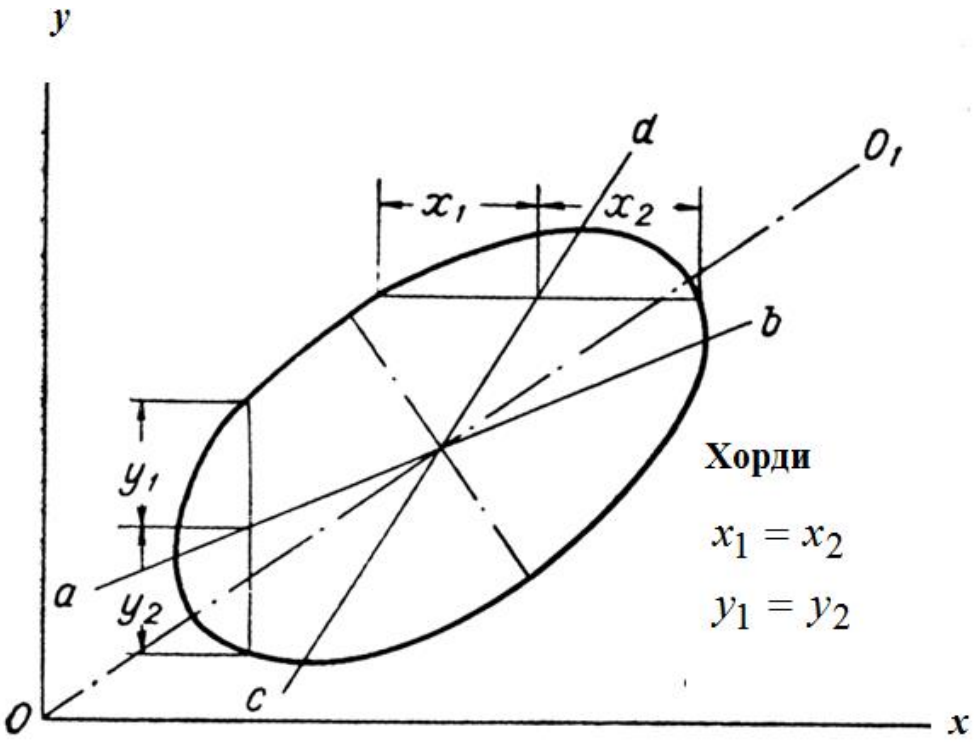


Рисунок 5.3 – Еліпс розсіювання та лінії регресії

Пряма ab є лінія регресії Y по X. Лінія ab ділить навпіл вертикальні хорди еліпса, який відображає розсіювання змінної y при заданому x. Лінія cd ділить навпіл хорди, паралельні осі x.

Проміжна лінія – вісь еліпса $\varnothing\varnothing_1$, має кутовий коефіцієнт σ_y / σ_x, який дорівнює відношенню середньоквадратичних відхилень випадкових величин Y та X. Коли б величини Y та X були пов'язані між собою функціональною залежністю, то лінія регресії Y по X, також як і X по Y, проходила б під кутом до осі x з тангенсом σ_y / σ_x. У випадку не функціонального зв'язку зміни $X = x$ лише частково проявляються у коливаннях змінної y. Тому лінія регресії Y по X пройде під меншим кутом до осі x ніж вісь еліпса $\varnothing\varnothing_1$, тобто й кутовий коефіцієнт буде менший за співвідношення σ_y / σ_x. Коли ж за аргумент береться y, то залежна від $Y = y$ змінна x теж лише частково реагує на відхилення аргументу. Відповідно лінія регресії X по Y пройде під меншим кутом до осі y (або під більшим кутом до осі x), ніж вісь $\varnothing\varnothing_1$. Ступінь невідповідності коливань, тобто ступінь розходження ліній регресії вимірюється коефіцієнтом кореляції.

5.6 Рівняння парної лінійної регресії, оцінка їх параметрів за вибірковими даними

Отримані рівняння парної лінійної регресії (5.44), (5.45), представлені у вигляді рівнянь умовного математичного сподівання для системи випадкових величин, широко застосовуються як математична модель, що описує стохастичну залежність між двома випадковими величинами.

Пошук рівнянь парної лінійної регресії застосовується для опису стохастичних залежностей між випадковими величинами. Взята до розрахунку математична модель (лінійна парна регресія) описує зв'язок генеральних сукупностей залежних випадкових величин X та Y. Задача користувача полягає в тому, щоб за обмеженими у часі даними спостережень (вибірками), зробити висновки про характер зв'язку та оцінити статистичні параметри рівняння регресії.

Для вибіркових даних рівняння умовного математичного сподівання представляється у вигляді

$$\tilde{y}_i = \tilde{y}(x_i) = \hat{m}_{y/x} = ax_i + b, \qquad (5.48)$$

де x_i - дискретні значення випадкової величини X;

y - дискретні значення випадкової величини Y;

$\tilde{y}_i$ - значення випадкової величини Y, розраховані за рівнянням регресії;

a,b - шукані параметри рівняння.

5.6.1. Оцінка параметрів рівняння лінійної регресії за даними спостережень

Для отримання розрахункових параметрів a і b використовується метод найменших квадратів. Оскільки йдеться про рівняння умовного математичного сподівання, то можна сказати, що функція регресії $\tilde{y}(x) = m_{y/x}$ є функцією, яка мінімізує середню квадратичну похибку визначення $\tilde{y}(x)$ при обчисленні $\tilde{y}_i$ по величині x_i. Це означає, що для довільної функції $U(x)$ справедлива нерівність

$$M[Y - U(X)]^2 \geq M[Y - \tilde{y}(x)]. \qquad (5.49)$$

Для того, щоб виконувалася нерівність (5.49), оцінки параметрів, які входять до рівняння регресії, обчислюються за ***методом найменших квадратів***. Це метод обробки емпіричного числового матеріалу, вимога якого полягає в тому, щоб сума квадратів відхилень даних спостережень y_i від лінії регресії $\tilde{y}(x)$була найменшою, тобто

$$\Delta = \sum_{i=1}^{n} [y_i - \tilde{y}(x_i)]^2 = min\ , \qquad (5.50)$$

де y_i - спостережені значення випадкової величини Y;

$\tilde{y}(x_i)$ - значення випадкової величини Y, розраховані за рівняннями регресії для заданих x_i;

n - число спільно спостережених значень y_i та x_i.

Припустимо, що за математичну модель, яка описує стохастичний кореляційний зв'язок між випадковими величинами Y і X, обране рівняння лінійної регресії вигляду (5.48). На основі вихідних даних за спільний період спостережень необхідно оцінити коефіцієнти α й b та одержати тим самим можливість розрахунку («передбачення») значень випадкової величини Y при заданих x_i. Запис рівняння набуває вигляду

$$\tilde{y}(x_i) = ax_i + b\ , \qquad (5.51)$$

де $a=\hat{\alpha}_1; b=\hat{\beta}_1;$ тобто a і b є оцінками параметрів моделі (5.48).

Відповідно до методу найменших квадратів a і b повинні бути такими, щоб сума Δ досягала свого мінімуму. Вимога екстремуму означає, що частинні похідні Δ, узяті по a і b, дорівнюють нулю

$$\frac{\partial\Delta(a,b)}{\partial a}=\frac{\partial\left[\left(\sum_{i=1}^{n}y_i-ax_i-b\right)^2\right]}{\partial a}=0\ ; \tag{5.52}$$

$$\frac{\partial\Delta(a,b)}{\partial b}=\frac{\partial\left[\left(\sum_{i=1}^{n}y_i-ax_i-b\right)^2\right]}{\partial b}=0. \tag{5.53}$$

Розв’язуючи рівняння (5.52) і (5.53) відносно a і b, одержуємо

$$a=\frac{\overline{xy}-\bar{x}\cdot\bar{y}}{\overline{x^2}-\bar{x}^2}; \tag{5.54}$$

$$b=\bar{y}-a\bar{x}\ . \tag{5.56}$$

Чисельник дробу, який знаходиться в правій частині рівняння (5.54), є оцінкою коваріаційного моменту $\hat{K}_{xy}$, розрахованого за дискретною вибіркою завдовжки n

$$\hat{K}_{x,y}=\hat{\mathrm{cov}}(x,y)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}), \tag{5.57}$$

а знаменник – оцінкою дисперсії випадкової величини X

$$\hat{\sigma}_x^2=S_x^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=\overline{x^2}-\bar{x}^2. \tag{5.58}$$

Оцінка коефіцієнта кореляції, яка відображає тісноту лінійного зв'язку між розглядуваними рядами спостережень за випадковими величинами X та Y, записується у вигляді

$$\hat{r}_{x,y} = r = \frac{\hat{K}_{x,y}}{S_x S_y} = \frac{\hat{cov}(x,y)}{S_x S_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 (y_i - \bar{y})^2}}, \quad (5.59)$$

де S_y - оцінка середнього квадратичного відхилення випадкової величини Y.

Оцінка дисперсії S_y^2 випадкової величини Y виконується за формулою

$$\hat{\sigma}_y^2 = S_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \overline{y^2} - \bar{y}^2. \quad (5.60)$$

Оцінка параметра a рівняння лінійної парної регресії виражається через коефіцієнт кореляції $\hat{r}_{x,y} = \hat{r}$ та середнє квадратичне відхилення випадкових величин Y та X, розрахованих за даними спостережень й позначеними як S_y та S_x

$$a = \hat{r}_{x,y}\frac{S_y}{S_x} = r\frac{S_y}{S_x}. \quad (5.61)$$

З урахуванням отриманих результатів можна зазначити, що вони відповідають складовим рівняння умовного математичного сподівання для системи нормально розподілених величин.

5.6.2 Оцінка вірогідності моделі лінійної регресії

Оцінка вірогідності моделі передбачає установлення ступеня відповідності між реальним стохастичним зв'язком випадкових величин X та Y й математичною моделлю, яка описує цей зв'язок. Основою такої оцінки може служити величина Δ (5.50), яка є сумою квадратів нев'язок ε_i між спостереженими і розрахованими за рівнянням лінійної регресії значеннями випадкової величини Y

$$\Delta = \sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{n}\left[y_i - (ax_i + b)\right]^2. \quad (5.6\ 2)$$

Критерієм якості розрахунку є середнє квадратичне відхилення спостережених y_i і розрахованих $\tilde{y}_i = ax_i + b$ значень

$$S = \sqrt{\frac{\sum_{i=1}^{n} \varepsilon_i^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} [y_i - (ax_i + b)]^2}{n}}. \qquad (5.63)$$

Величина S носить також назву *похибки апроксимації*.

Рівняння (5.6 3) легко перетворюється до вигляду

$$S^2 = \overline{y^2} - \bar{y}^2 - 2a \cdot \hat{cov}(x,y) + a^2 \left[\overline{x^2} - \bar{x}^2 \right]. \qquad (5.64)$$

З урахуванням (5.59)-(5.61) похибка апроксимації S записується у вигляді

$$S^2 = S_y^2 \left(1 - r^2\right) \qquad (5.65)$$

або

$$S = S_y \sqrt{1 - r^2}. \qquad (5.66)$$

Однак середня квадратична похибка апроксимації S не є досить інформативною оцінкою, оскільки нев'язка розрахунку визначається в першу чергу вірогідністю визначення коефіцієнтів регресії. Щоб переконатися у вірогідності побудованої моделі, необхідно оцінити вірогідність коефіцієнтів a та b, для чого будуються довірчі інтервали.

5.6.3 Визначення довірчих інтервалів для коефіцієнтів рівняння лінійної регресії

Обчислюючи на основі наявних вибіркових даних оцінку $\hat{\theta}$ деякого параметра θ, ми усвідомлюємо, що насправді величина $\hat{\theta}$ є лише наближеним значенням параметра генеральної сукупності θ, навіть у тому випадку, коли ця оцінка обґрунтована, незміщена й умотивована. Виникає питання, як сильно може відхилятися це наближене вибіркове значення (оцінка) від відповідного значення генеральної сукупності?

Вирішення цього питання досягається шляхом установлення довірчих інтервалів.

Щоб отримати уявлення про точність оцінки $\hat{\theta}$, треба визначити можливу похибку. Призначимо досить велику ймовірність β (наприклад, β=0,95), щоб подію з ймовірністю β можна було б вважати достовірною, і знайдемо таке значення ε, для якого виконується така умова (О.С. Вентцель, 1969)

$$p\left(\left|\hat{\theta}-\theta\right|<\varepsilon\right)=\beta \ . \tag{5.67}$$

Тоді діапазон можливої похибки при заміні θ на $\hat{\theta}$ буде дорівнювати $\pm\varepsilon$; великі за абсолютною величиною похибки можливі з малою ймовірністю $1-\beta$, яка має назву рівня значущості.

Перепишемо (1.247) у вигляді

$$p\left(\hat{\theta}-\varepsilon<\theta<\hat{\theta}+\varepsilon\right)=\beta \ . \tag{5.68}$$

Рівність (5.68) означає, що із ймовірністю β значення параметра θ потрапляє в інтервал

$$I_\beta=\left(\hat{\theta}-\varepsilon;\hat{\theta}+\varepsilon\right), \tag{5.69}$$

який має назву довірчого.

Чим менше для вибраної ймовірності є довірчий інтервал, тим точнішу оцінку θ отримаємо.

Величина θ не є випадковою, проте випадковим є сам інтервал та його положення на числовій осі відносно центра $\hat{\theta}$. Через це величина β розглядається не як ймовірність попадання точки θ в інтервал, а ймовірність того, що визначений інтервал “накриє” точку θ.

Якщо розподіл оцінки параметра близький до нормального, довірчий інтервал має вигляд

$$\left[\hat{\theta}-t_{\nu,\alpha}\sigma_{\hat{\theta}};\hat{\theta}+t_{\nu,q}\sigma_{\hat{\theta}}\right], \tag{5.70}$$

де $t_{\nu,q}$ - статистика Стьюдента при рівні значущості q й числі степенів вільності ν (додаток Б);

$\sigma_{\hat{\theta}}$ - середнє квадратичне відхилення вибіркової оцінки $\hat{\theta}$.

Так, наприклад, довірчий інтервал для математичного сподівання визначається співвідношенням

$$\bar{x} - t_{\nu,q}\sigma_{\bar{x}} < m_x < \bar{x} + t_{\nu,q}\sigma_{\bar{x}} \ , \tag{5.71}$$

де

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} . \tag{5.72}$$

Якщо випадкова величина Y розподілена нормально, то не тільки оцінки математичного сподівання цієї величини, а й оцінки коефіцієнтів регресії розподілені нормально (Є.П. Школьний, І.Д. Лоєва, Л.Д. Гончарова, 1999; Є.П. Школьний, Л.Д. Гончарова, Н.К. Миротворська, 2000).

Довірчі інтервали коефіцієнтів регресії можна побудувати в такий спосіб

$$a - t_{\nu,q}\sigma_a < m_a < a + t_{\nu,q}\sigma_a, \text{ або } \quad \left\lfloor a - t_{\nu,q}\sigma_a ; a + t_{\nu,q}\sigma_a \right\rfloor; \tag{5.74}$$

$$b - t_{\nu,q}\sigma_b < m_b < b + t_{\nu,q}\sigma_b, \quad \text{або } \left\lfloor b - t_{\nu,q}\sigma_a ; b + t_{\nu,q}\sigma_a \right\rfloor, \tag{5.75}$$

де m_a та m_b - математичні сподівання параметрів a та b. Якщо звернутися до рівняння (5.44), яке є моделлю, то $\hat{m}_a = \hat{\alpha}_1 = a$,а $\hat{m}_b = \hat{\beta}_1 = b$.

Оцінки a та b є незміщеними, ефективними та умотивованими оцінками параметрів рівняння лінійної регресії (Є.П. Школьний, І.Д. Лоєва, Л.Д. Гончарова, 1999; Є.П. Школьний, Л.Д. Гончарова, Н.К. Миротворська, 2000)

$$\sigma_a = \frac{S_y}{nS_x} \ , \tag{5.76}$$

$$\sigma_b = \frac{S_y}{\sqrt{n}}\left[1 + \frac{1}{C_V^2}\right]^{\frac{1}{2}} = \frac{S_y}{\sqrt{n}}\left[1 + \frac{\bar{x}^2}{S_x^2}\right] = \frac{S_y}{\sqrt{n}}\frac{(S^2{}_x + \bar{x}^2)}{S^2{}_x} . \tag{5.77}$$

Довірчі інтервали для математичних сподівань α_1 та β_1, які входять у модель лінійної регресії (5.44), визначаються таким чином

$$a + t_{\nu,q}\sigma_a \frac{S_y}{nS_x} \le \alpha_1 \le a + t_{\nu,q}\sigma_a \frac{S_y}{nS_x} \ ; \qquad (5.78)$$

$$b + t_{\nu,q}\sigma_b \frac{S_y}{\sqrt{n}}\left[1 + \frac{1}{C_V^2}\right]^{\frac{1}{2}} \le \beta_1 \le b + t_{\nu,q}\sigma_b \frac{S_y}{\sqrt{n}}\left[1 + \frac{1}{C_V^2}\right]^{\frac{1}{2}}. \qquad (5.79)$$

5.6.4 Визначення довірчого інтервалу для коефіцієнта кореляції в рівнянні лінійної регресії

Якщо величини X та Y мають розподіл, близький до нормального, а обсяг вибірки великий ($n > 30$), то розподіл коефіцієнта кореляції буде близький до нормального. Середня квадратична похибка визначення $\hat{r}_{xy}$ за спостереженими даними у цьому випадку визначається за формулою

$$\hat{\sigma}_r = S_r = \frac{1 - \hat{r}_{x,y}^2}{\sqrt{n-1}}. \qquad (5.80)$$

Довірчий інтервал коефіцієнта кореляції записується у вигляді

$$\hat{r}_{x,y} - t_{\nu,q}\frac{1 - r^2}{\sqrt{n-1}} < r_н < \hat{r}_{x,y} + t_{\nu,q}\frac{1 - \hat{r}_{xy}^2}{\sqrt{n-1}}. \qquad (5.81)$$

Якщо ж довжина ряду невелика ($n < 30...40$), а коефіцієнт кореляції значний ($r > 0{,}3$), то розподіл вибіркових значень $\hat{r}_{xy}$ істотно відрізняється від нормального. При значеннях $\hat{r}_{xy}$, що наближаються до 1, крива розподілу вибіркових оцінок коефіцієнта кореляції стає усе більш асиметричною. Тому, для оцінки точності розрахунку значення коефіцієнта кореляції використовують z - перетворення Фішера

$$z = \frac{1}{2}\ln\left[\frac{1 + r_{x,y}}{1 - r_{x,y}}\right]. \qquad (5.82)$$

Закон розподілу оцінок z близький до нормального із параметрами

$$\bar{z} = \ln\left[\frac{1+r_{x,y}}{2\left(1-r_{x,y}\right)} + \frac{r_{x,y}}{2(n-1)}\right] \tag{5.83}$$

та

$$\sigma_z^2 = 1/(n-3). \tag{5.84}$$

Розрахунки довірчого інтервалу для коефіцієнта кореляції виконуються таким чином:

- надаються оцінки параметрів

$$\hat{z} = \frac{1}{2} ln \frac{1+\hat{r}_{x,y}}{1-\hat{r}_{x,y}}, \tag{5.85}$$

$$\hat{\sigma}_z^2 = \frac{1}{\sqrt{(n-3)}}; \tag{5.86}$$

- будується довірчий інтервал для z

$$\hat{z} - t_{v,q}\sigma_z < z < \hat{z} + t_{v,q}\sigma_z \ , \tag{5.87}$$

тобто

$$z_1 < z < z_2 \ ;$$

- відбувається зворотний перехід від z_1 та z_2 до $\hat{r}_{1xy}$ та $\hat{r}_{2xy}$ за допомогою оберненого перетворення:

$$r_{x,y} = \frac{e^{2z}-1}{e^{2z}+1} \ , \tag{5.88}$$

$$\hat{z} = \frac{1}{2}\ln\frac{1+\hat{r}_{x,y}}{1-\hat{r}_{x,y}}; \frac{1+\hat{r}_{x,y}}{1-\hat{r}_{x,y}} = \exp\{2\hat{z}\} \ , \tag{5.89}$$

$$1+\hat{r}_{xy} = exp\, 2\hat{z} - \hat{r}_{xy} \cdot exp\, 2\hat{z} \ , \tag{5.90}$$

$$\hat{r}_{x,y}\left(1+exp\, 2\hat{z}\right) = exp\, 2\hat{z} - 1, \tag{5.91}$$

звідки

$$\hat{r}_{x,y} = \frac{exp\,2\hat{z} - 1}{exp\,2\hat{z} + 1}. \tag{5.92}$$

Використовуючи границі довірчих інтервалів для статистики z, за допомогою рівняння (5.92), одержуємо границі довірчого інтервалу коефіцієнта кореляції:

$$r_1 = \frac{exp\,2\hat{z}_1 - 1}{exp\,2\hat{z}_1 + 1} \tag{5.93}$$

та

$$r_2 = \frac{exp\,2\hat{z}_2 - 1}{exp\,2\hat{z}_2 + 1}\ . \tag{5.94}$$

Тоді

$$r_1 < m_{r_{xy}} < r_2\,. \tag{5.95}$$

Питання для самоперевірки

1. Що називається умовною ймовірністю події?
2. Чому дорівнює ймовірність добутку двох подій?
3. Що являє собою щільність розподілу безперервної випадкової величини?
4. Що називається умовним законом розподілу величини X, яка входить до системи (X,Y)?
5. Що називається умовним законом розподілу величини Y, яка входить до системи (X,Y)?
6. Чому дорівнює щільність розподілу системи двох випадкових величин?
7. Яка залежність називається ймовірнісною чи стохастичною?
8. Коли випадкова величина Y називається незалежною від випадкової величини X?
9. Що оцінює коефіцієнт кореляції
10. Чому дорівнює коефіцієнт кореляції, який характеризує функціональну залежність між двома випадковими величинами, які входять до системи?
11. У яких випадках коефіцієнт кореляції дорівнює нулю і коли одиниці?

6 ПЕРЕВІРКА СТАТИСТИЧНИХ ГІПОТЕЗ

6.1 Постановка задачі про перевірку статистичних гіпотез

Під *статистичними гіпотезами* розуміють такі гіпотези, котрі відносяться або до виду, або до окремих параметрів розподілу випадкової величини.

Сформулюємо задачу статистичної перевірки гіпотези у загальному виді. Нехай $^0f(x,\Theta)$ - закон розподілу випадкової величини x, який залежить від одного параметра Θ. Припустимо, що необхідно перевірити гіпотезу про те, що $\Theta = \Theta_0$, і цю гіпотезу назвемо *нульовою* або *основною* H_0. Гіпотезу про те, що $\Theta = \Theta_1$ назвемо *конкуруючою* або *альтернативною* H_1. Перевірка гіпотези H_0 відносно конкуруючої гіпотези H_1 виконується на основі вибірки, що складається з n незалежних спостережень $x_1, x_2, x_3, ..., x_n$ над випадковою величиною X (Школьний Є.П., Лоєва І.Д., Гончарова Л.Д., 1999).

Всю можливу множину N вибірок об'ємом n можна розділити на дві неперетинних множини: u_1 і u_2, таких, що гіпотеза H_0 повинна бути відкинутою, якщо вибірка, яка розглядається, потрапляє до підмножини u_1, і прийнятою, якщо вибірка належить до підмножини u_2.

Для зручності під час перевірки гіпотези використовують статистичні параметри K, одержаними на основі вибірок за визначеним правилом. Оскільки K параметри є числами, які зображуються точками на числовій осі, то підмножини u_1 і u_2 вибірок зводяться до двох одномірних областей W_1 і W_2. Область W_1 параметрів k називають *критичною областю*, а область W_2 - *областю припустимих значень*. Оскільки область W_2 складається з точок, які не увійшли до області W_1, то область W_1 однозначно визначає область W_2, і навпаки (Школьний Є.П., Лоєва І.Д., Гончарова Л.Д., 1999).

Приймаючи чи відкидаючи гіпотезу H_0 можна припустити помилку двох видів. *Помилка першого роду* полягає у тому, що нульова гіпотеза H_0 відкидається, коли вона є правильною. *Помилка другого роду* допускається тоді, коли приймається гіпотеза H_0, у той час, коли правильною є гіпотеза H_1.

Імовірність помилки першого роду позначається через α і називається *рівнем значущості.* α завжди задається дослідником і найчастіше дорівнює 0,05 або 5%.

Імовірності помилок першого і другого роду однозначно визначаються вибором критичної області W_1. Критичну область W_1 відокремлює від області прийняття гіпотези H_0 критична точка $k_{кр.}$. Можуть розглядатися

правостороння, лівостороння і двостороння критичні області. Вони позначені на рис. 6.1.

1. Для правосторонньої критичної області (рис.6.1 а), припускається що . гіпотеза H_0 є вірною. Тоді імовірність того, ідо гіпотеза H_0 відкидається тобто, що робиться помилка I роду, дорівнює

$$P(k > k_{кр}) = \alpha \tag{6.1}$$

2. Для лівосторонньої критичної області (рис.6.1 б)

$$P(k < k_{кр}) = \alpha \tag{6.2}$$

3. Для двосторонньої критичної області (рис.6.1 в)

$$P(k < k'_{êð}) + P(k > k''_{êð}) = \alpha \tag{6.3}$$

Рівень значущості впливає на те, з якою ймовірністю основна гіпотеза будет прийнята, цю ймовірність називають довірчою і позначають через

$$P = 1 - \alpha \tag{6.4}$$

Критичний параметр k можна знайти за допомогою таких **теорем** (Школьний Є.П., Лоєва І.Д., Гончарова Л.Д., 1999)**:**

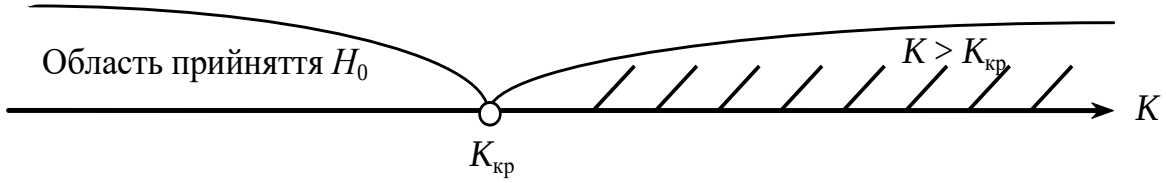
Теорема 1. За умови, що незалежні величини $z : z_1, z_2, ..., z_n$ підпорядковуються нормальному закону розподілу, і сума їх квадратів

$$\chi^2 = \sum_{i=1}^{n} Z_i^2 \tag{6.5}$$

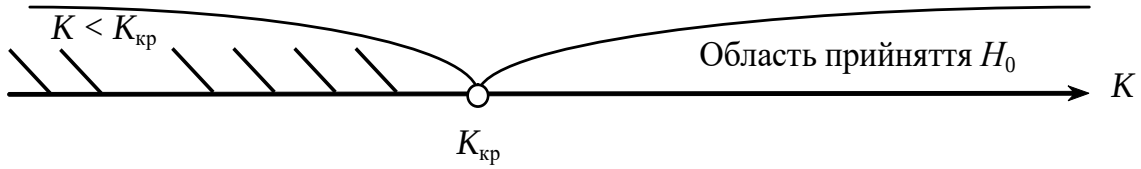
підпорядковується закону розподілу, який має назву χ^2, з ν - числом ступенів свободи. Ступені свободи є парметром розподілу.

Теорема 2. Якщо дві незалежні випадкові величини u і υ, підпорядковуються χ^2 - розподілу з числами ступенів волі ν_1 і ν_2 відповідно, то випадкова величина підлягає розподілу, який називається *законом Фішера-Снедекора.* Розподіл Фішера-Снедекора двопараметричний з параметрами ν_1 і ν_2.

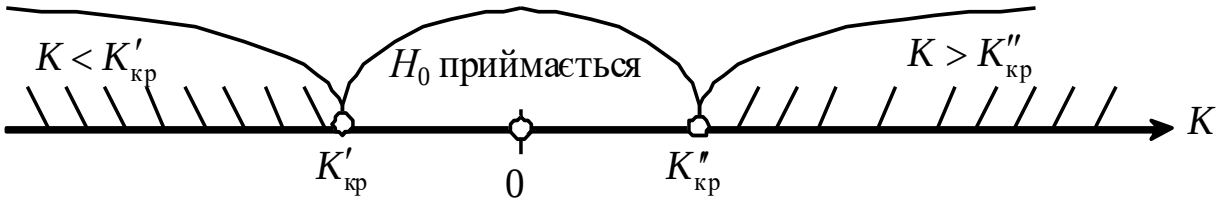
$$F = \frac{u/\nu_1}{\upsilon/\nu_2} \tag{6.6}$$



а) правостороння



б) лівостороння



в) двостороння

Рисунок 6.1 – Види критичних областей

Теорема 3. Якщо ми маємо дві незалежні випадкові величини u і υ, такі, що $\grave{e}$ - підпорядковується нормальному закону, а $\upsilon - \chi^2$ розподілу з числом ступенів волі ν, то випадкова величина t підлягає розподілу Стьюдента, цей закон є однопараметричним, з параметром ν.

$$t = \frac{u}{\sqrt{\frac{\upsilon}{\nu}}}, \tag{6.7}$$

6.2 Перевірка статистичної гіпотези про однорідність членів статистичної сукупності

Якісний стан річки залежить від багатьох факторів. Аналізуючи дані спостережень можна виявити певні закономірності у змінах стану водних об'єктів. Наприклад, відомо, що досліджувана річка забруднена фенолами внаслідок діяльності нафтопереробного заводу. Середня річна концетрація фенолів у воді за період з 1980 по 2010 роки змінювалася у границях 0,0004–0,0001 мг/дм^3, ГДК фенолів становить 0,001 мг/дм^3 (для питних потреб). В 2011 році середньорічна концетрація фенолів у воді склала 0,005 мг/дм^3 (перевисила ГДК у 5 разів), тобто різко відрізнялася від звичних середніх річних концентрацій. Такі значення величин прийнято називати "***випадіннями***" або "***викидами***". *Випадіння чинять вплив на середнє значення, але особливо на значення центральних моментів*, оскільки при їх оцінці в суму входять доданки, які являють собою великі різниці між випадіннями і середніми значеннями, котрі підносяться до другого, третього чи четвертого степеня у залежності від того, оцінку якого моменту треба знайти (Шитиков В.К., Розенберг Г.С.,Зинченко Т.Д., 2003).

Для того, щоб вирішити чи треба залишати «випадіння» у вихідній сукупності чи вилучити, перевіряється статистична гіпотеза H_0 про однорідність членів статистичної сукупності.

Середнє значення, як випадкова величина, підлягає нормальному закону розподілу. Вибиремо із вибірки $x_{\min}$ і $x_{\max}$. Випадкові величини $\bar{x} - x_{\min}$ та $x_{\max} - \bar{x}$ також будуть підлягати нормальному закону, тому можна позначити

$$u = \left| x_{екстр} - \bar{x} \right| , \tag{6.8}$$

де $x_{åêſòð}$ об'єднує мінімальну і максимальну величини. Випадкова величина $è$ відповідає одній з умов теореми 3.

Якщо вважати, що випадкова величина X має нормальний розподіл, тоді на основі теореми 1 слід визнати, що величина

$$\upsilon = \sum_{i=1}^{n} \left(x_i - \bar{x} \right) = \chi^2 \tag{6.9}$$

Підлягає розподілу χ^2 з числом степенів вільності $\nu = n - 1$. Отже, ця величина задовольняє другу умову теореми 3, на її основі отримаємо випадкову величину

$$t = \frac{\left|x_{екстр} - \bar{x}\right|}{\sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}}} = \frac{\left|x_{екстр} - \bar{x}\right|}{S_x}, \qquad (6.10)$$

що підлягає розподілу Стьюдента.

$t_{кр}(\alpha, \nu)$ з умови (для правосторонньої критичної області) відповідає формулі (6.1) і на рівні значущості α, який установлюється дослідником, знаходять по таблиці 6.1.

Таблиця 6.1 – Значення критерію Стьюдента при рівні значущості α = 0.05

Число ступенів свободи	Рівень значущості 0.05	Число ступенів свободи	Рівень значущості 0.05
1	12.70	18	2.10
2	4.30	19	2.09
3	3.18	20	2.09
4	2.78	21	2.08
5	2.57	22	2.07
6	2.45	23	2.07
7	2.36	24	2.06
8	2.31	25	2.06
9	2.26	26	2.06
10	2.23	27	2.05
11	2.20	28	2.05
12	2.18	29	2.05
13	2.16	30	2.04
14	2.14	40	2.02
15	2.13	60	2.00
16	2.12	120	1.98
17	2.11		

Таким чином, якщо $t < t_{кр}$, то гіпотеза H_0 про однорідність членів вибірки не відхиляється. Коли $t > t_{кр}$, гіпотеза H_0 відхиляється і приймається альтернативна гіпотеза H_1 про те, що $x_{екстр}$ «випадіння» не належать до тієї ж генеральної сукупності, що і решта членів цієї сукупності. Необхідно неоднорідні члени вилучити з вибірки, і знову провести оцінку відповідних параметрів.

6.3 Перевірка статистичної гіпотези про однорідність двох нормально розподілених рядів

Однорідними називають такі ряди, які підпорядковуються одному і тому ж закону розподілу, тобто належать до однієї і тієї ж генеральної сукупності, і для характеристики особливостей цього розподілу достатньо мати два параметри: середнє арифметичне значення та дисперсію σ_x^2.

Перевірка на однорідність полягає у перевірці двох гіпотез: про незначущість різниці в дисперсіях і про незначущість різниці у середніх. У випадку коли хоча б одна гіпотеза не приймається, то відкидається гіпотеза про однорідінсть двох рядів X і Y.

Якщо випадкові величини X і Y описуються нормальним законом розподілу, то оцінки дисперсій σ_x^2 і σ_y^2 підлягають розподілу з параметрами $\nu_1 = m-1$ та $\nu_2 = n-1$, де m - довжина вибірки X, а n - довжина вибірки Y. У такому випадку розподіл випадкової величини описується законом Фішера-Снедокора

$$F = \frac{\sigma_x^2}{\sigma_y^2}. \qquad (6.11)$$

Статистична характеристика F має назву критерію Фішера-Снедокора. При розрахунках у чисельник ставлять більшу з дисперсій. Перевірка нульової гіпотези здійснюється шляхом порівняння розрахованого значення F з критичним $F_{кр}(\alpha, \nu_1, \nu_2)$ (табл.6.2).

Нульова гіпотеза не відкидається, коли $F < F_{êð}$. У протилежному випадку приймається альтернативна гіпотеза, тобто різниця між дисперсіями рядів X і Y значуща і не може пояснюватися тільки впливом випадкових флуктуацій у вибірках X і Y.

Якщо розподіл випадкових величин X і Y описується нормальним законом, то середні арифметичні значення $\bar{x}$ та $\bar{y}$ також нормально розподілені. Отже, випадкова величина $u = \bar{x} - \bar{y}$ теж нормально розподілена. Дисперсії σ_x^2 і σ_y^2 підлягають χ^2 - розподілу, тоді випадкова величина також підялягає закону χ^2

$$U = (m-1)\sigma_x^2 + (n-1)\sigma_y^2, \qquad (6.12)$$

де $\nu = m + n - 2$ - кількість степенів вільності.

Таблиця 6.2 - Значення критерію Фішера F для рівня значущості 0.05

ν_1 - число степенів вільності для більшої дисперсії

ν_2 - число степенів вільності для меншої дисперсії

ν_2	ν_1								
	12	14	16	20	24	30	40	50	75
10	2.9	2.9	2.8	2.8	2.7	2.7	2.7	2.6	2.6
11	2.8	2.7	2.7	2.7	2.6	2.6	2.5	2.5	2.5
12	2.7	2.6	2.6	2.5	2.5	2.5	2.4	2.4	2.4
13	2.6	2.6	2.5	2.5	2.4	2.4	2.3	2.3	2.3
14	2.5	2.5	2.4	2.4	2.4	2.3	2.3	2.2	2.2
15	2.5	2.4	2.3	2.3	2.3	2.3	2.2	2.2	2.2
16	2.4	2.4	2.3	2.3	2.2	2.2	2.2	2.1	2.1
17	2.4	2.3	2.3	2.3	2.2	2.2	2.1	2.1	2.0
18	2.3	2.3	2.3	2.3	2.2	2.1	2.1	2.0	2.0
19	2.3	2.3	2.2	2.2	2.1	2.1	2.0	2.0	2.0
20	2.3	2.2	2.2	2.1	2.1	2.0	2.0	2.0	1.9
21	2.3	2.2	2.2	2.1	2.1	2.0	2.0	1.9	1.9
22	2.2	2.2	2.1	2.1	2.0	2.0	1.9	1.9	1.9
23	2.2	2.1	2.1	2.0	2.0	2.0	1.9	1.9	1.8
24	2.2	2.1	2.1	2.0	2.0	1.9	1.9	1.9	1.8
25	2.2	2.1	2.1	2.0	2.0	1.9	1.9	1.8	1.8
26	2.2	2.1	2.1	2.0	2.0	1.9	1.9	1.8	1.8
27	2.1	2.1	2.0	2.0	1.9	1.9	1.8	1.8	1.8
28	2.1	2.1	2.0	2.0	1.9	1.9	1.8	1.8	1.8
29	2.1	2.1	2.0	1.9	1.9	1.9	1.8	1.8	1.7
30	2.1	2.0	2.0	1.9	1.9	1.8	1.8	1.8	1.7
32	2.1	2.0	2.0	1.9	1.9	1.8	1.8	1.7	1.7
34	2.1	2.0	2.0	1.9	1.8	1.8	1.7	1.7	1.7
36	2.0	2.0	1.9	1.9	1.8	1.8	1.7	1.7	1.7
38	2.0	2.0	1.9	1.9	1.8	1.8	1.7	1.7	1.6
40	2.0	2.0	1.9	1.8	1.8	1.7	1.7	1.7	1.6
42	2.0	1.9	1.9	1.8	1.8	1.7	1.7	1.6	1.6
44	2.0	1.9	1.9	1.8	1.8	1.7	1.7	1.6	1.6
46	2.0	1.9	1.9	1.8	1.8	1.7	1.7	1.6	1.6
48	2.0	1.9	1.9	1.8	1.7	1.7	1.6	1.6	1.6
50	2.0	1.9	1.9	1.8	1.7	1.7	1.6	1.6	1.6
100	1.9	1.8	1.8	1.7	1.6	1.6	1.5	1.5	1.4
200	1.8	1.7	1.7	1.6	1.6	1.5	1.7	1.4	1.4
∞	1.8	1.6	1.6	1.6	1.5	1.5	1.4	1.4	1.3

Виходячи з вище сказаного, випадкова величина має розподіл, який може описуватись розподілом Стьюдента

$$t = \frac{\left|\bar{x} - \bar{y}\right|}{\sqrt{\frac{(m-1)\sigma_x^2 + (n-1)\sigma_y^2}{m+n-2}\left(\frac{1}{m} + \frac{1}{n}\right)}}. \qquad (6.13)$$

Гіпотеза H_0 про незначущість різниці між середніми арифметичними значеннями перевіряється шляхом порівняння критерію t з $t_{êð.}(\alpha,\nu)$ за допомогою таблиці 6.1.

H_0 приймається, коли $t < t_{êð.}(\alpha,\nu)$.

Статистична гіпотеза про однорідність рядів X і Y приймається тільки тоді, коли справедливі обидві гіпотези H_0 і $H_0^{/}$.

6.4 Перевірка статистичної гіпотези про однорідність членів статистичної сукупності на основі непараметричних критеріїв

Якщо випадкові величини не підлягають нормальному розподілу або якщо невідомо до якого закону розподілу відноситься випадкова величина, вживаються непараметричні критерії. Одним з таких критеріїв є критерій Вілкоксона, буває двох видів.

Інверсійний критерій Вілкоксона полягає у тому, що випадкові величини, які належать до вибірок X і Y розташовують у загальній послідовності у порядку збільшення (або зменшення) їх значень, наприклад:

$$X : x_1, x_2, x_3, x_4, x_5$$
$$Y : y_1, y_2, y_3, y_4, y_5 y_6$$
$$y_1 x_1 y_2 y_3 x_2 x_3 y_4 y_5 y_6 x_4 x_5$$

Якщо якому-небудь значенню x передує деяке значення y, то кажуть, що ця пара утворює інверсію. В загальній послідовності число інверсій u дорівнює

$$u = 1 + 3 + 3 + 6 + 6 = 19$$

В однорідних рядах, кожне з котрих має не менше 10 членів, число інверсій розподіляється приблизно за нормальним законом з математичним сподіванням

$$\mathrm{m_u} = \frac{m \cdot n}{2} \tag{6.14}$$

і дисперсію

$$\sigma_{\mathrm{u}}^2 = \frac{m \cdot n}{12}(m + n + 1), \tag{6.15}$$

де n і m - число членів у першій та другій вибірках.

Нульова гіпотеза H_0 полягає у належності вибірок X і Y до однієї генеральною сукупності. Шляхом установлення рівня значущості α або довірчої ймовірності $p = 1 - \alpha$, знаходять границі допустимих значень u, що відділяють область прийняття гіпотези від критичної області. Якщо значення критерію, яке отримане за даними спостережень, попаде до критичної області, то нульова гіпотеза відхиляється і з імовірністю ∂ приймається альтернативна гіпотеза.

Область прийняття гіпотези H_0 визначається нерівністю

$$\mathrm{m_u} - t_{\hat{e}\partial}(\alpha, \nu)\sigma_u \le u \le \mathrm{m_u} + t_{\hat{e}\partial}(\alpha, \nu)\sigma_u, \tag{6.16}$$

а критична область – нерівностями

$$u < m_u - t_{кр}(\alpha, \nu)\sigma_u, \tag{6.17}$$

$$u > m_u + t_{кр}(\alpha, \nu)\sigma_u, \tag{6.18}$$

де $\sigma_u = \sqrt{\sigma_u^2}$ - середній квадратичний відхил числа інверсій;

$t_{кр}(\alpha, \nu)$ - критерій Стьюдента для рівня значущості α і числа степенів вільності $\nu = m + n - 2$.

Критерій однорідності Вілкоксона відповідає задачі порівняння тільки двох вибірок. Але він може вживатися для попарного порівняння вибірок в S пунктах спостережень деякого регіону, який вважається однорідним.

Ранговий критерій Вілкоксона полягає у тому, що для кожного значення випадкової величини обчислюється величина зміни ознаки. Всі зміни впорядковують за абсолютною величиною (без урахування знака). Потім рангам приписують знак зміни і підсумовують ці "знакові ранги" - в результаті виходить значення критерію Вілкоксона.

6.5 Перевірка статистичної гіпотези про відповідність емпіричного розподілу теоретичному

Емпіричним законом є згрупований ряд випадкової величини

$$\begin{matrix} x_1; x_2; ...; x_k \\ m_1; m_2; ...; m_k . \end{matrix} \qquad (6.19)$$

Закон розподілу можна представити у *інтервальних теоретичних частотах*

$$\begin{matrix} \tilde{x}_1; \tilde{x}_2; ...; \tilde{x}_k \\ \tilde{m}_1; \tilde{m}_2; ...; \tilde{m}_k . \end{matrix} \qquad (6.20)$$

Допустима розбіжність між емпіричними m_i і теоретичними $\tilde{m}_i$ інтервальними частотами встановлюється через *перевірку статистичної гіпотези про відповідність емпіричного розподілу теоретичному закону* на заданому рівні значущості.

Емпіричні частоти, як випадкові величини, мають нормальний розподіл, сума їх квадратів підлягає χ^2 *розподілу.* Оскільки теоретичні інтервальні частоти $\tilde{m}_i$ - величини невипадкові, то розподілу χ^2 підлягає і така сума квадратів

$$\sum_{i=1}^{k} \frac{\left(m_i - \tilde{m}_i\right)^2}{\tilde{m}_i} = \chi^2 . \qquad (6.21)$$

Випадкова величина χ^2 використовується у якості критерію Пірсона, за допомогою якого відбувається перевірка гіпотези H_0 відносно альтернативної гіпотези H_1 (перевірка узгодженості між емпіричними та теоретичними частотами).

Гіпотеза H_0 полягає у тому, що розбіжності між емпіричними і теоретичними частотами є незначущими на рівні значущості α (визначає правосторонню критичну область рис. 6.3)

$$P\left[\ \chi^2 > \chi^2_{кр}(\nu,\alpha)\ \right] = \alpha, \tag{6.22}$$

Область прийняття гіпотези H_0 визначається нерівністю $\chi^2 < \chi^2_{кр}(\nu,\alpha)$, а критична область – нерівністю $\chi^2 > \chi^2_{кр}(\nu,\alpha)$.

Якщо $\chi^2 < \chi^2_{кр}(\nu,\alpha)$, то гіпотеза H_0 не відхиляється.

Якщо $\chi^2 > \chi^2_{кр}(\nu,\alpha)$, то приймається альтернативна гіпотеза H_1.

Рівень значущості α визначається дослідником. Величина ν є числом ступенів волі, котре відіграє роль параметра χ^2 розподілу. χ^2 залежить від виду теоретичного розподілу, яким апроксимується емпіричний розподіл (Єріна А.М., Головач А.В., Козирєв О.В., 1993), і визначається за таким правилом

$$\nu = k - s, \tag{6.23}$$

де k - кількість часткових інтервалів.

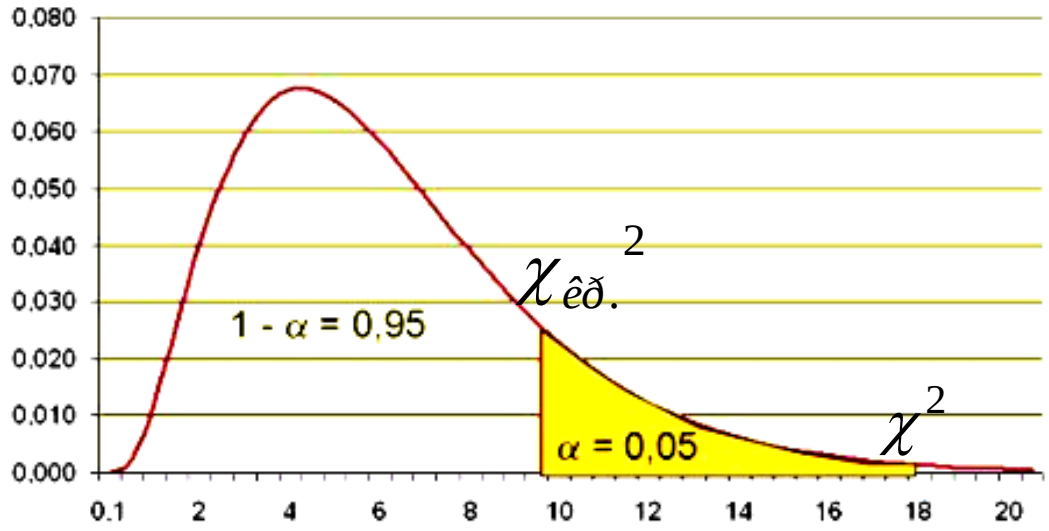


Рисунок 6.3 – Зв’язок між емпіричним і теоретичним розподілами (при рівні значущості $\alpha = 0{,}05$)

При використанні критерію χ^2 *необхідно, щоб частота (емпірична та теоретична) у кожній градації була не менше 5*. Для цього градації, які мають частоти менші за 5, необхідно об'єднати з сусідніми. S - кількість лінійних зв’язків, відносно частот m_{i}, що реалізується при статистичному оцінюванні моментів, які беруть участь при розрахунках параметрів теоретичного розподілу. Лінійний зв'язок описується рівністю

$$\sum_{i=1}^{k} m_i = n, \tag{6.24}$$

де n - загальний об’єм статистичної сукупності, реалізується завжди.

Приклади визначення ν для деяких розподілів.

а) Припустимо, що *емпіричний розподіл апроксимується нормальним розподілом.* Апроксимація («наближення») – процесс підбору теоретичного закону, який буде відповідати емпіричному.

Нормальний розподіл має два параметри- m_x і σ_x^2, оцінками яких є

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} m_i \tilde{x}_i \tag{6.25}$$

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{k} m_i (\tilde{x}_i - \bar{x})^2. \tag{6.26}$$

Рівності (6.25) і (6.26) є лінійними формами відносно інтервальних частот m_{i}. Таким чином, з урахуванням лінійного зв’язку (6.24) для нормального розподілу маємо $S = 3$, а $\nu = k - 3$.

б) При апроксимації емпіричного розподілу *розподілом Пірсона I* типу, для визначення параметрів цього розподілу, окрім статистичних характеристик $\bar{x}$ і S_x^2, необхідно використовувати оцінки основних моментів $\hat{r}_3$ і $\hat{r}_4$, які безпосередньо пов’язані з оцінками третього

$$\hat{\mu}_3 = \frac{1}{n}\sum_{i=1}^{k} m_i \left(\tilde{x}_i - \bar{x}\right)^3 \tag{6.27}$$

і четвертого центральних моментів

$$\hat{\mu}_4 = \frac{1}{n}\sum_{i=1}^{k} m_i \left(\tilde{x}_i - \bar{x}\right)^4 . \qquad (6.28)$$

Рівняння (6.27)-(6.28) є лінійними залежностями відносно інтервальних частот m_{i}. Отже, при визначенні параметрів цього розподілу треба використовувати лінійні залежності (6.25)-(6.28), що з урахуванням рівності (6.24) дає $S = 5$, а $\nu = k - 5$.

в) При апроксимації емпіричного розподілу *розподілом Пірсона III* типу параметри розраховуються за допомогою $\bar{x}$, S_x^2 і $\hat{r}_3$. Тому для III типу розподілів Пірсона маємо $S = 4$, а $\nu = k - 4$.

Для перевірки зазначеної гіпотези також можна використовувати критерій Колмогорова, критерій Романовського.

6.6 Перевірка статистичної гіпотези про однорідність двох рядів за критерієм Гнеденко-Королюка

Цей метод застосовується у випадку коли

$$N = 2n < 60. \qquad (6.29)$$

Вихідний ряд довжиною N розбивається на дві вибірки однакової довжини n. Якщо число елементів непарне, то перший або останній елемент відкидається. Розбивка ряду відбувається посередині.

Для кожної вибірки довжиною n підраховується інтегральна функція розподілу $F(x)$. Для розрахунків кожна з вибірок розміщується у порядку зростання. Емпіричне значення ординат інтегральної функції визначається за формулою

$$d_m = \frac{m'}{n}, \qquad (6.30)$$

де m' - порядковий номер члену вибірки, розташованого у порядку зростання.

Аналізуючи графіки інтегральних функцій розподілу, знаходимо для будь-якого x_i найбільше відхилення між значеннями $F(x_i)$, виражене в частках від одиниці і позначимо його d_m. Величина d_m повинна бути додатною

$$0 \le d_m \le 1 . \qquad (6.31)$$

Величина d_m випадкова і показує міру відхилення емпіричних розподілів двох вибірок. Якщо помножити випадкову величину d_m на n, отримаємо *закон розподілу Гнеденко-Королюка*

$$C_m = d_m \cdot n \ . \tag{6.32}$$

Використовуючи закон (6.32), можна розрахувати ймовірність потрапляння C_m у критичну область

$$p(C_m > C_{\hat{e}\eth.}) = \alpha_{\tilde{n}} \tag{6.33}$$

Величину $\alpha_{\tilde{n}}$ визначають за таблицею 6.3.

Якщо $\alpha_c > \alpha$ (α - рівень значущості дорівнює 5%), то нульова гіпотеза приймається.

6.7 Перевірка статистичної гіпотези про однорідність двох рядів за критерієм Колмогорова-Смірнова

Критерій згоди Колмогорова-Смірнова є більш результативним, ніж критерій χ^2 і може бути використаний для перевірки гіпотези про відповідність емпіричного розподілу будь-якому теоретичному безперервному розподілу $F(x)$ з заздалегідь відомими параметрами (Степнов М.Н., 1985). На відміну від критерію Гнеденко-Королюка критерій Колмогорова-Смірнова не потребує парної кількості елементів, а вихідний ряд не обов'язково розбивати на дві вибірки посередині. *Критерій Колмогорова-Смірнова рекомендовано застосовувати при для статистичних рядів довжиною* $N = 2n \geq 60$.

Умова, яка ставиться при використанні критерію Колмогорова-Смірнова

$$kl/(k+l) = n \geq 15 \tag{6.34}$$

де k та l - довжина вибірок, на які розбито вихідний ряд (розбивають ряд у місті найімовірніших змін характеру величини).

Вихідний ряд розбивається на дві вибірки довжиною k і l, щоб виконувалась умова (6.34).

Таблиця 6.3 – Розподіл функції Гнеденко-Королюка

n	C																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
5	100	87.3	35.71	7.94	0.79													
6	100	93.07	47.4	14.29	2.60	0.22												
7	100	96.27	57.52	5.30	0.82	0.06												
8	100	98.01	66.01	28.27	8.70	1.87	0.25	0.02										
9	100	98.95	73.01	35.17	12.59	3.36	0.63	0.07	0.00									
10	100	99.45	78.69	41.75	16.78	5.24	1.23	0.21	0.02	0.00								
11	100	99.71	83.26	47.92	21.15	5.77	1.70	0.38	0.06	0.01	0.00							
12	100	99.85	86.9	53.61	25.58	9.95	3.14	0.79	0.15	0.02	0.00							
13	100	99.92	89.78	58.82	29.99	12.65	4.43	1.26	0.29	0.05	0.01	0.00						
14	100	99.96	92.06	63.55	34.33	15.49	5.90	1.88	0.49	0.10	0.02	0.00						
15	100	99.98	93.83	67.81	38.55	18.44	7.55	2.62	0.77	0.18	0.04	0.01	0.00					
16	100	99.99	95.23	71.64	42.63	21.45	9.33	3.50	1.12	0.30	0.07	0.01	0.00					
17	100	99.99	96.31	75.06	46.54	24.50	11.24	4.50	1.56	0.46	0.12	0.02	0.00					
18	100	100	97.15	78.10	50.26	27.54	13.24	5.60	2.07	0.67	0.18	0.04	0.01	0.00				
19	100	100	97.81	80.81	53.79	30.57	15.32	6.81	2.67	0.92	0.28	0.07	0.02	0.00				
20	100	100	98.31	83.20	57.13	33.56	17.45	8.11	3.35	1.23	0.40	0.10	0.03	0.01	0.00			
21	100	100	98.70	85.31	60.28	36.50	19.63	9.48	4.11	1.59	0.55	0.17	0.04	0.01	0.00			
22	100	100	99.01	87.17	63.24	39.37	21.84	10.93	4.93	2.00	0.73	0.24	0.07	0.02	0.00			
23	100	100	99.24	88.80	66.01	42.18	24.06	12.43	5.83	2.47	0.95	0.32	0.10	0.03	0.01	0.00		
24	100	100	99.42	90.24	68.60	44.90	26.28	13.98	6.78	2.99	1.20	0.43	0.14	0.04	0.01	0.00		
25	100	100	99.50	91.50	71.02	47.55	28.50	15.58	7.79	3.56	1.48	0.56	0.19	0.06	0.02	0.00		
26	100	100	99.66	92.60	73.27	50.10	30.71	17.20	8.85	4.18	1.81	0.71	0.26	0.08	0.02	0.01	0.00	
27	100	100	99.74	93.57	75.37	54.56	32.90	18.86	9.96	4.84	2.17	0.98	0.33	0.11	0.03	0.01	0.01	
28	100	100	99.80	94.41	77.32	54.94	35.06	20.53	11.10	5.55	2.56	1.09	0.42	0.15	0.05	0.01	0.01	
29	100	100	99.85	95.14	79.12	57.22	37.20	22.21	12.29	6.30	2.99	1.31	0.53	0.20	0.07	0.02	0.01	0.00
30	100	100	99.88	95.78	80.80	59.41	39.29	23.91	13.50	7.09	3.46	1.56	0.65	0.25	0.09	0.03	0.01	0.00

Для кожної вибірки розраховуються ординати емпіричної інтегральної функції розподілу за формулою (6.30). Обидві інтегральні криві зіставляються і для x_i знаходять найбільше відхилення d_{kl} в частках від одиниці. Величина d_{kl} є *мірою розходження між емпіричним розподілом двох вибірок*.

Якщо d_{kl} є функцією від випадкової величини $F(x)$, то вона теж є випадковою. Далі визначається статистична величина

$$z = d_{kl}\sqrt{n}, \qquad (0 \le d \le 1), \tag{6.35}$$

z - величина, яка підлягає розподілу Колмогорова. За таблицею 6.4 перевіряється умова $z \le z_{\hat{e}\eth.}$

$$p(z < z_{\hat{e}\eth.}) = L(z). \tag{6.36}$$

Якщо ймовірність p перевищує рівень значущості $\alpha = 5\%$, тобто

$$p(z < z_{\hat{e}\eth.}) > 0.05), \tag{6.37}$$

тоді приймається нульова гіпотеза.

Можна графічно порівняти емпіричний і теоретичний розподіли за критерієм Колгоморова-Смірнова (рис.6.4 а) і за критерієм Пірсона (рис.6.4 б).

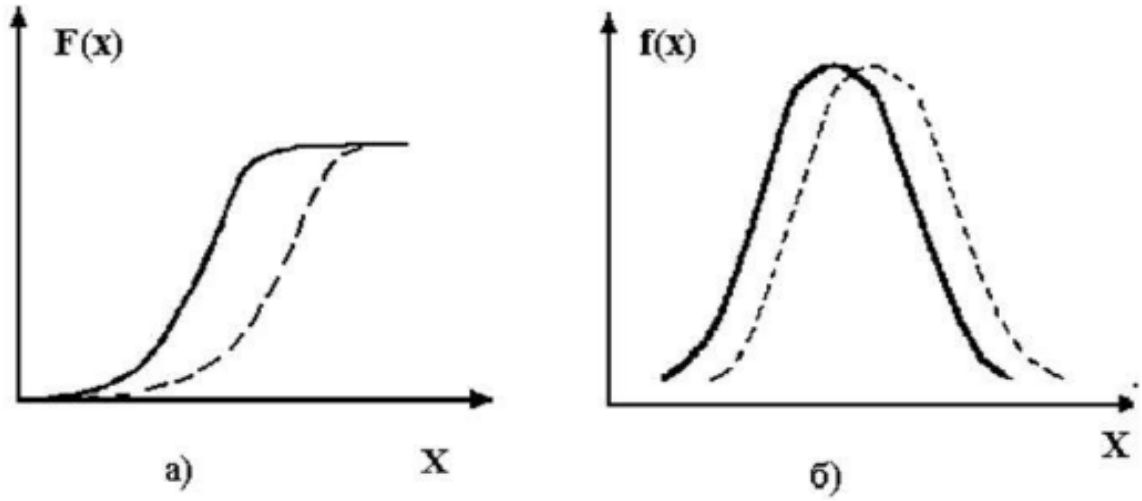


Рисунок 6.4 – Емпіричний (пунктирна лінія) і теоретичний (суцільна лінія) розподіли за критеріями Колгоморова-Смірнова (а) і за критерієм Пірсона (б) (Степнов М.Н., 1985)

Таблиця 6.4 – Функція розподілу Колгоморова $L(z)$

z	0	1	2	3	4	5	6	7	8	9
0,2	0,000000	000000	000000	000000	000000	000000	000000	000000	000001	000004
0,3	0,000009	000021	000046	000091	000171	000303	000511	000826	001285	001929
0,4	0,002808	003972	005476	007377	009730	012589	016005	020022	024682	030017
0,5	0,036055	042814	050306	058534	067497	077183	087577	098656	110394	122760
0,6	0,135718	149229	163255	177752	192677	207987	223637	239582	255780	272188
0,7	0,288765	305471	322265	339114	355981	372833	389640	406372	423002	439505
0,8	0,455858	472039	488028	503809	519365	534682	549745	564545	579071	593315
0,9	0,607269	620928	634285	647337	660081	672515	684836	696445	707941	719126
1,0	0,730000	740566	750825	760781	770436	779794	788860	797637	806130	814343
1,1	0,822282	829951	837356	844502	851395	858040	864443	870610	876546	882258
1,2	0,887750	893030	898102	903973	907648	912134	916435	920557	924506	928288
1,3	0,931908	935371	938682	941847	944871	947758	950514	953144	955651	958041
1,4	0,960318	962487	964551	966515	968383	970159	971846	973448	974969	976413
1,5	0,977782	979080	980310	981475	982579	983623	984610	985544	986427	987261
1,6	0,988048	988791	989492	990154	990777	991364	991917	992438	992928	993389
1,7	0,993823	994230	994612	994972	995309	995625	995922	996200	996460	996704
1,8	0,996932	997146	997346	997533	997707	997870	998023	998165	998297	998421
1,9	0,998536	998644	998744	998837	998924	999004	999079	999149	999213	999273
2,0	0,999329	999381	999429	999473	999514	999553	999588	999620	999651	999679
2,1	0,999705	999728	999750	999771	999790	999807	999823	999837	999851	999863
2,2	0,999874	999886	999895	999904	999912	999920	999927	999933	999939	999944
2,3	0,999949	999954	999958	999961	999965	999968	999971	999974	999976	999978
2,4	0,999980	999982	999984	999985	999987	999988	999989	999990	999991	999992

6.8 Перевірка гіпотези про існування тренда у часовому ряді, за допомогою критерію Аббе

Коли статистична неоднорідність ряду даних установлена і у фондових матеріалах є вказівки на наслідки інтенсивних водогосподарських перетворень, є сенс виявити тренд у хронологічній послідовності щорічних концентрацій забруднюючих речовин.

Тренд – загальна тенденція зміни досліджуваних величин в бік зростання або зменшення. *Динамічний тренд* – це функція, яка характеризує тенденцію зміни явища у часі (Тарасова В.В., 2008).

Для розв'язання цієї задачі можна використати *критерій Аббе*. В його основі лежить порівняння дисперсії значень випадкової величини X з сумою квадратів їх послідовних різниць S^2, яка менш чутлива до систематичної зміни математичного сподівання. Величина S^2 розраховується за формулою

$$S^2 = \frac{1}{2(N-1)} \cdot \sum_{i=1}^{N-1} \left(x_{i+1} - x_i\right)^2, \qquad (6.38)$$

де N - довжина вибірки (кількість років спостережень);

x_{i+1} та x_i - наступне та попереднє значення хронологічного ряду.

Висувається нульова гіпотеза, яка передбачає, що тренд існує. Для первірки гіпотези про відсутність систематичних змін в упорядкованій послідовності розраховується відношення

$$Z = \frac{S^2}{\sigma_X^2}, \qquad (6.39)$$

де σ_X^2 - дисперсія вихідного ряду.

Якщо $Z \geq Z_{кр}$, то можна зробити висновок, що ряд спостережень не має систематичного зсуву математичного сподівання (тренд відсутній), але коли $Z < Z_{кр}$, то тренд існує.

Критичні значення $Z_{кр}$ для $\alpha = 0{,}05$ при N від 4 до 300 наведені у таблиці 6.5.

Недоліком критерію Аббе є те, що генеральна сукупність, з якої вилучається ряд спостережень, припускається нормальною, тому функція Z може «зреагувати» на циклічні коливання стоку, що безумовно впливає на загальну мінералізацію річкової води.

Таблиця 6.5 – Критичні значення розподілу величин

Число даних	5%-ий рівень значущості	Число даних	5%-ий рівень значущості
4	0,390	35	0,729
5	0,410	36	0,733
6	0,446	37	0,736
7	0,468	38	0,740
8	0,491	39	0,743
9	0,512	40	0,746
10	0,531	41	0,749
11	0,548	42	0,752
12	0,564	43	0,755
13	0,578	44	0,758
14	0,591	45	0,760
15	0,603	46	0,763
16	0,614	47	0,765
17	0,624	48	0,768
18	0,633	49	0,770
19	0,642	50	0,772
20	0,560	51	0,774
21	0,657	52	0,776
22	0,665	53	0,778
23	0,671	54	0,780
24	0,678	55	0,782
25	0,684	56	0,784
26	0,689	57	0,785
27	0,695	58	0,878
28	0,700	59	0,789
29	0,705	60	0,791
30	0,709	100	0,837
31	0,714	150	0,867
32	0,718	200	0,885
33	0,722	300	0,906
34	0,726		

6.9 Перевірка гіпотези про статистичну значущість коефіцієнта кореляції і коефіцієнтів рівняння регресії

Часто у дослідженнях постає питання про реальність установлених на основі спостережених даних зв'язків, тому що можливо, що встановлені значення $\hat{r}_{x,y}, a$ й b зумовлені випадковістю вибірок. Вирішення такого питання називають *оцінкою статистичної значущості параметрів* і воно зв'язане з перевіркою статистичних гіпотез.

Висунемо нульову гіпотезу щодо тісноти розглянутого зв'язку

$$H_0 : r_{xy} = 0, \tag{6.40}$$

тобто коефіцієнт кореляції є статистично незначущим.

Альтернативна гіпотеза є такою

$$H_1 : r_{xy} \neq 0, \tag{6.41}$$

тобто коефіцієнт кореляції є статистично значущим.

Якщо розподіл вибіркових оцінок r_{xy} відповідає нормальному закону розподілу (що справедливо при великих n), то для перевірки нульової гіпотези як критерій можна використовувати статистику

$$t = \frac{r_{xy}}{\sqrt{\sigma^2_{r_{xy}}}}, \tag{6.42}$$

розподіл якої підлягає розподілу Стьюдента.

Значення t визначається за вибірковими оцінками коефіцієнта кореляції $\hat{r}_{x,y}$ і його стандарту $\sqrt{\sigma_r^2} = S_r$

$$\hat{t} = \frac{|\hat{r}_{xy}|}{S_r} \tag{6.43}$$

і порівнюється з критичним $t_{кр}$, котре залежить від числа степенів вільності $\nu = n - 1$ й рівня значущості q.

При $t < t_{кр}$ нульова гіпотеза приймається, а при $t > t_{кр}$ - відхиляється, тобто значення коефіцієнта кореляції визнається статистично значущим.

При невеликих n і високих значеннях $\hat{r}_{x,y}$ оцінка статистичної значущості $\hat{r}_{x,y}$ виконується за допомогою z-перетворення Фішера, тобто оцінюється не величина $\hat{r}_{x,y}$ безпосередньо, а статистика $\hat{z}$. Якщо z значуще, то і коефіцієнт кореляції є статистично значущою величиною.

Оцінка значущості коефіцієнтів рівняння регресії a і b виконується аналогічним чином.

Як вже відзначалося, розподіл вибіркових оцінок a і b вважається нормальним. Тоді у відповідності до теорем математичної статистики для перевірки гіпотези про a і b можна використовувати статистику t, яка підлягає розподілу Стьюдента

$$\hat{t}_a = \frac{|a|}{S_a} \quad \text{и} \quad \hat{t}_b = \frac{|b|}{S_b} \tag{6.44}$$

Якщо $\hat{t} > t_{кр}(\nu, q)$, то коефіцієнти регресії вважаються значущими.

Питання для самоперевірки

1.Дайте визначення «статистичних гіпотез».

2.Яка область називається «критичною» і яка областю «припустимих значень»?

3.Що називають помилкою першого роду?

4.Як визначається рівень значущості α?

5.Які значення величин прийнято називати "випадіннями" або "викидами"?

6.Коли відхиляється нульова гіпотеза?

7.За яким принципом формується «нульова» гіпотеза і «альтернативна»?

8.Які ряди називаються однорідними?

9.Який критерій використовується для перевірка гіпотези про статистичну значущість коефіцієнта кореляції та коефіцієнтів регресії ?

10.За яких умов використовуються критерії Фішера і Стьюдента?

11.Який критерій використовуюь для перевірки статистичної однорідності двох вибірок, які не підлягають нормальному закону розподілу?

12.Дайте визначення довірчого інтервалу.

13. Який принцип використання критерію Гнеденко-Королюка?

14. Як перевірити гіпотезу про наявність тренда?

15. За допомогою яких критеріев виконується перевірка статистичної гіпотези про однорідність двох рядів?

ЛІТЕРАТУРА

1. Лобода Н.С., Гопченко Є.Д. Стохастичні моделі у гідрологічних розрахунках: навч. посібник для студ. вищих навч. закл. / Одес. держ. еколог. ун-т. Одеса: Екология, 2006. 200 с.

2. Лобода Н.С. Методи статистичного аналізу у гідрологічних розрахунках і прогнозах: навч.посіб. / Одес. держ. еколог. ун-т. Одеса: Екологія. 2010. 184 с.

3. Сніжко С. І. Оцінка та прогнозування якості природних вод. Підручник. - К.: Ніка-Центр, 2001. 264 с.

4. Школьний Є.П., Гончарова Л.Д.,Миротворська Н.К. Методи обробки та аналізу гідрометеорологічної інформації (збірник задач і вправ): навчальний посібник. К.: Міносвіти і науки України, 2000. 419с.

5. Школьний Є.П., Лоєва І.Д., Гончарова Л.Д. Обробка та аналіз гідрометеорологічної інформації / Одес. держ. еколог. ун-т. Одеса: 1999. 600с.

6. www.library-odeku.16mb.com.

7. Рождественский А.В. Статистические методы в гидрологии / А.В. Рождественський, А.И. Чеботарев. Л.: Гидрометеоиздат, 1974. 424с.

8. Дронов С.В. Многомерный статистический анализ: учебное пособие. / Дронов С.В. Барнаул, изд. Алт. гос. ун-та, 2003. 213 с.

9. Афифи А. Статистический анализ: Подход с использованием ЭВМ. / А. Афифи, С. Эйзен.; пер. с англ. И.С. Енюкова, И.Д. Новикова, под ред. Г.П. Башарина. М.: Мир, 1982. 488 с.

10. Третьяков А.С. Статистические методы в прикладных географических исследованиях: Учебно-методическое пособие / Третьяков А.С. Харьков.: Шрифт, 2004. 96 с.

11. Андерсон Т. Введение в многомерный статистический анализ / пер. с англ. Ю.Ф. Кичатова и др. Москва.: Физико-иптематической литературы, 1963. 500 с.

12. Richard H. McCuen Modeling Hydrologic Change: Statistical Methods / Department of Civil and Environmental Engineering. University of Maryland. Lewis Publishers CRC Press: New York, Washington, D.C., 2002. 448 p.

13. Шитиков В.К., Розенберг Г.С., Зинченко Т.Д. Количественная гидроэкология: методы системной идентификации / Ин-т экологии Волжс.бассейна РАН. Тольятти, 2003. 463 с.

14. Степнов М.Н. Статистические методы обработки результатов механических испытаний: Справочник. М.: Машиностроение, 1985. 232 с.

15. Головач А.В., Єріна А.М., Козирєв О.В. Статистика: підручник / за наук.ред.д.екон.н. С.С. Герасименка. 2-ге вид., перероб. і доп. К.: Вища школа, 1993. 468 с.

16. Тарасова В.В. Екологічна статистика (з блочно-модульною формою контролю знань): підручник / рецен. д.екон.н. Парфенцева Н.О., д.екон.н. Малюга Н.М., д.с.-г.н. Смаглій А.Ф. К.: Центр учбової літератури, 2008. 392 с.

Навчальне електронне видання

ЛОБОДА НАТАЛІЯ СТЕПАНІВНА,
КУЗА АНТОНІНА МИКОЛАЇВНА

МЕТОДИ МАТЕМАТИЧНОЇ СТАТИСТИКИ У ГІДРОЕКОЛОГІЧНИХ ДОСЛІДЖЕННЯХ

Конспект лекцій

Видавець і виготовлювач
Одеський державний екологічний університет
вул. Львівська, 15, м. Одеса, 65016
тел./факс: (0482) 32-67-35
E-mail: info@odeku.edu.ua
Свідоцтво суб'єкта видавничої справи
ДК № 5242 від 08.11.2016