

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ  
ОДЕСЬКИЙ ДЕРЖАВНИЙ ЕКОЛОГІЧНИЙ УНІВЕРСИТЕТ

Н.С. ЛОБОДА, І.В. КАТИНСЬКА

**ПРИКЛАДНІ АСПЕКТИ ЗАСТОСУВАННЯ МЕТОДІВ  
МАТЕМАТИЧНОЇ СТАТИСТИКИ У ГІДРОЕКОЛОГІЧНИХ  
ДОСЛІДЖЕННЯХ**

Навчальний посібник

Одеса  
Одеський державний екологічний університет  
2023

УДК 504.4.054: 519

*Л68*

**Лобода Н. С., Катинська І. В.**

**Л68** Прикладні аспекти застосування методів математичної статистики у гідроекологічних дослідженнях: навчальний посібник. Одеса : Одеський державний екологічний університет, 2023. 99 с.

ISBN 978-966-186-272-1

У навчальному посібнику розглянуті найбільш вживані у практиці методи математичної статистики, які використовуються підчас обробки та аналізу даних гідрологічних, гідрохімічних та гідробіологічних спостережень в умовах значного антропогенного навантаження на водне середовище. Підписання угоди про асоціацію між Україною та ЄС установило нові правила та стандарти природоохоронної діяльності, серед яких виділяється сектор «якість води та управління водними ресурсами». Нові методики розрахунків передбачають застосування різного роду статистичних оцінок, законів розподілу, статистичних гіпотез та статистичних критеріїв. У посібнику наведені статистичні методи оцінки випадковості коливань досліджуваної величини, визначення її підпорядкованості нормальному закону розподілу, описані критерії для встановлення статистичної однорідності рядів, виявлення «викидів» у рядах даних, пошуку тенденцій та статистично значущих трендів у хронологічному ході. Значна увага приділена питанням оцінки екологічних ризиків недосягнення доброго стану водних об'єктів згідно із положеннями Водної Рамкової Директиви (ВРД). Методи багатовимірного статистичного аналізу наведені у виді прикладів їх практичного застосування (розрахунків та інтерпретації) для вирішення гідроекологічних задач.

Запропонований навчальний посібник призначений для студентів різного рівня вищої освіти: бакалаврів, магістрів, аспірантів, які працюють у сфері охорони природного навколишнього середовища, зокрема водного, а також може бути успішно використаний спеціалістами гідрологами та гідробіологами.

УДК 504.4.054: 519

**Рецензенти:**

*Доцент кафедри екологічної безпеки та гідравліки Національного університету  
«Одеська політехніка», доцент **Мельник С.В.***

*Декан природоохоронного факультету Одеського державного екологічного  
університету, доктор технічних наук, професор **Чугай Ангеліна Володимирівна***

*Затверджено вченою радою Одеського державного екологічного університету  
Міністерства освіти і науки України як навчальний посібник для здобувачів вищої  
освіти за спеціальністю «Екологія» (протокол №5 від 29. 06. 2023 р.)*

ISBN 978-966-186-272-1

© Лобода Н.С., Катинська І.В. 2023,

© Одеський державний екологічний університет, 2023

## ЗМІСТ

ВСТУП.....	5
РОЗДІЛ 1 СТАТИСТИЧНІ ГІПОТЕЗИ ТА ЇХ ПЕРЕВІРКА .....	7
1.1 Загальна постановка задачі про перевірку статистичних гіпотез .	7
1.2 Критерії значущості та закони їх розподілу .....	11
1.3 Перевірка статистичної гіпотези про однорідність членів статистичної сукупності (перевірка належності «викидів» до генеральної сукупності) .....	14
1.4 Перевірка статистичної гіпотези про незначущість різниць між оцінками дисперсій .....	18
1.5. Перевірка статистичної гіпотези про незначущість різниць між середніми значеннями.....	19
1.6 Перевірка статистичної гіпотези про однорідність членів статистичної сукупності на основі непараметричних критеріїв (критерій Вілкоксона).....	22
РОЗДІЛ 2 КРИТЕРІЇ ВИПАДКОВОСТІ.....	25
2.1 Ранговий критерій випадковості із застосуванням коефіцієнта Кендалла .....	25
2.2 Критерій Аббе .....	28
2.3 Критерій Дарбіна-Уотсона.....	30
РОЗДІЛ 3 ПЕРЕВІРКА НАЛЕЖНОСТІ ВИБІРКИ ДО НОРМАЛЬНО РОЗПОДІЛЕНОЇ СУКУПНОСТІ.....	33
3.1 Нормальний закон розподілу випадкової величини.....	33
3.2. Оцінка ступеня відхилення емпіричного закону розподілу випадкової величини від нормального .....	39
3.3 Критерій Гаусса .....	41
3.4 Критерій Діксона-Томпсона .....	43
РОЗДІЛ 4 ДОВІРЧІ ІНТЕРВАЛИ ТА ОЦІНКА ВІРОГІДНОСТІ ПАРАМЕТРІВ РІВНЯННЯ ЛІНІЙНОЇ ПАРНОЇ РЕГРЕСІЇ .....	46
4.1 Визначення довірчого інтервалу для математичного сподівання .....	47
4.2 Визначення довірчих інтервалів для коефіцієнтів рівняння лінійної парної регресії.....	49
4.3 Визначення довірчих інтервалів для коефіцієнта кореляції .....	51

4.4	Перевірка гіпотези про статистичну значущість коефіцієнта кореляції і коефіцієнтів рівняння лінійної парної регресії .....	53
	<b>РОЗДІЛ 5 ЗАСТОСУВАННЯ МЕТОДІВ МАТЕМАТИЧНОЇ СТАТИСТИКИ ПРИ ОЦІНКАХ ЕКОЛОГІЧНИХ РИЗИКІВ .....</b>	<b>56</b>
5.1	Поняття про забезпеченість випадкової величини та процентиль .....	56
5.2	Оцінка ризиків недосягнення екологічних цілей .....	58
5.3	Визначення екологічних ризиків за стохастичною моделлю ....	62
	<b>РОЗДІЛ 6 ЗАСТОСУВАННЯ ДИСКРИМІНАНТНОГО АНАЛІЗУ ДО ПОБУДОВИ ШКАЛ КЛАСИФІКАЦІЇ ПОКАЗНИКІВ ЯКОСТІ ВОДИ.....</b>	<b>69</b>
6.1	Схема побудови розв'язувального правила .....	69
6.2	Побудова розв'язувального правила. Квадратична, лінійна та спрощена дискримінантна функція.....	76
	<b>РОЗДІЛ 7 ВИКОРИСТАННЯ ВЗАЄМНОЇ КОРЕЛЯЦІЙНОЇ ФУНКЦІЇ ТА ЇЇ ЗАСТОСУВАННЯ ДО РОЗРАХУНКІВ САМООЧИЩЕННЯ ВОДИ НА ДІЛЯНЦІ РУСЛА .....</b>	<b>84</b>
	<b>ВИКОРИСТАННІ ДЖЕРЕЛА .....</b>	<b>91</b>
	<b>ПРЕДМЕТНИЙ ПОКАЖЧИК .....</b>	<b>95</b>

## ВСТУП

Після підписання угоди про асоціацію з Європейським Союзом та його державами-членами в сфері охорони довкілля [1] перед Україною постали задачі впровадження законодавства ЄС в галузі довкілля. Робота українських організацій відбувається в межах восьми секторів, серед яких важливе місце займає сектор – “Якість води та управління водними ресурсами”. Європейська Комісія розробила загальний методологічний підхід до впровадження Водної Рамкової директиви ЄС, яка має забезпечити законодавчу базу для досягнення “доброго статусу водних об’єктів” [2]. Вирішення поставленої задачі передбачає аналіз факторів навантаження на водне середовище та їх впливу, хімічний моніторинг поверхневих вод, ідентифікацію та визначення значно змінених водних об’єктів, оцінку екологічних ризиків та інше. Одним із документів Угоди є також Директива № 91/676/ЄС (зі змінами і доповненнями, внесеними Регламентом (ЄС) № 1882/2003 отримала назву Нітратної директиви) про захист вод від забруднення, спричиненого надходженням сполук азоту з сільськогосподарських земель та тваринницьких ферм [3]. Нітратна Директива передбачає визначення поверхневих вод чутливих до забруднення та виділення уразливих зон з подальшим розробленням кодексу кращих методів ведення сільського господарства.

У сучасних умовах неможливо розглядати водні ресурси територій без вивчення наслідків антропогенного впливу на водні об’єкти, особливо з точки зору зміни їх екологічного стану, що знайшло своє відображення у роботах професіоналів-викладачів кафедри гідроекології та водних досліджень та інших кафедр ОДЕКУ, які приймали участь у такого роду дослідженнях [4, 5, 6].

Забезпеченість курсів, у яких розглядаються проблеми, пов’язані із охороною, збереженням та відновленням водних екосистем, не можна визнати задовільною. Більша частина видань, у яких розглядаються методи математичної статистики, не урахує особливостей підготовки студентів-екологів, які спеціалізуються на розгляді сучасних негативних змін водних ресурсів та водних екосистем, що відбуваються під впливом глобального потепління та посилюються значним антропогенним навантаженням на водні об’єкти [7].

Зміни клімату викликають порушення умов формування стоку та призводять до змін його кількісних характеристиках як за рік у цілому, так і по гідрологічних сезонах (зима, весна, літо-осінь). Антропогенні навантаження у вигляді зрошення, осушування, вилучання та перекиду стоку, а також скидів забруднених вод у поверхневі водотоки неминуче впливають на якість вод та їх гідроекологічний стан. Установлено, що антропогенний вплив посилює наслідки негативних змін у водних екосистемах за рахунок глобального потепління, особливо у зоні недостатнього зволоження [8].

Призначення даного посібника полягає у тому, щоб надати спеціалістам-гідроекологам можливості засвоїти основні прийоми обробки та аналізу інформації, яка стосується як суто гідрологічних, так і гідроекологічних показників, допомогти навчитися застосувати, аналізувати та інтерпретувати результати розрахунків за певними методами, які широко використовуються у інших галузях, але не в гідроекології.

Основу посібника складає зміст дисциплін із застосування методів математичної статистики («Методи математичної статистики у гідроекологічних дослідженнях», курс для бакалаврів, спеціальність 101), у тому числі окремі методи багатовимірного статистичного аналізу («Методи багатовимірного аналізу при вирішенні гідроекологічних задач», курс для магістрів, спеціальність 101), але в ньому розглядаються питання використання цих методів в гідроекології.

Посібник складається з семи розділів, кожен з яких може вивчатися окремо.

## РОЗДІЛ 1

### СТАТИСТИЧНІ ГІПОТЕЗИ ТА ЇХ ПЕРЕВІРКА

#### 1.1 Загальна постановка задачі про перевірку статистичних гіпотез

*Гіпотезою є наукове припущення, яке висувається щодо пояснення якогось явища і потребує перевірки та доведення.* У природознавстві, різних галузях техніки і економіки часто для з'ясування того чи іншого випадкового факту звертаються до висловлювання гіпотез, які можна перевірити статистично, тобто спираючись на результати спостережень, представлених у випадкових вибірках. *Під статистичними гіпотезами розуміють такі гіпотези, які відносяться або до виду, або до окремих параметрів статистичного розподілу випадкової величини.* Наприклад, статистичною є гіпотеза про те, що середні річні концентрації іонів натрію описуються нормальним законом розподілу. Статистичною буде також гіпотеза про те, що середні за місяць значення мінералізації на близько розташованих створах гідрохімічних спостережень (вище та нижче міста) суттєво не розрізняються. *Якщо гіпотеза висувається відносно параметру закону розподілу, то її називають параметричною. Статистичні гіпотези, які висуваються відносно закону розподілу, називаються непараметричними.*

*Гіпотеза, яка має особливо важливе значення, називається нульовою або основною гіпотезою.* Гіпотеза позначається через  $H$ , нульова гіпотеза – через  $H_0$ . Гіпотеза та її вміст записуються у такій формі

$$H_0 : a^* = a. \quad (1.1)$$

При такому записі нульова гіпотеза передбачає, що вибіркова характеристика  $a^*$  і є характеристикою генеральної сукупності  $a$ .

Сформулюємо задачу статистичної перевірки гіпотези у загальному вигляді. Нехай  $f(x, \Theta)$  – закон розподілу випадкової величини  $x$ , який залежить від параметра  $\Theta$ . Припустимо, що необхідно перевірити гіпотезу про те, що  $\Theta = \Theta_0$ . Будемо називати *гіпотезу нульовою* і позначимо її

через  $H_0$ . Гіпотезу про те, що  $\Theta = \Theta_1$  назвемо *конкуруючою* і позначимо її через  $H_1$ . Гіпотезу  $H_0$  іноді називають основною, а гіпотезу  $H_1$  – *альтернативною*. Альтернативна гіпотеза є протилежною нульовій гіпотезі.

Отже, стоїть задача перевірки гіпотези  $H_0$  відносно конкуруючої гіпотези  $H_1$  на основі вибірки, що складається з  $n$  незалежних спостережень  $x_1, x_2, x_3, \dots, x_n$  за випадковою величиною  $X$ .

Статистична сукупність розглядається як випадкова вибірка з генеральної сукупності. Отже, всю можливу множину  $N$  вибірок об'ємом  $n$  можна розділити на дві неперетинні підмножини (позначимо їх через  $u_1$  та  $u_2$ ), таких, що висунута гіпотеза повинна бути відкинута, якщо вибірка, яка розглядається, потрапляє до підмножини  $u_1$ , і прийнятою, якщо вибірка належить до підмножини  $u_2$ .

Більш зручно, проте, мати діло не безпосередньо з вибірками, а з деякими статистичними параметрами  $k$ , одержаними на основі вибірок за визначаючим правилом. Ці статистичні параметри називаються критеріями. Критерій – це правило для перевірки виконання або невиконання даного твердження (гіпотези). ***Статистичний критерій – це правило, яке дозволяє прийняти або усунути статистичну гіпотезу на основі даної вибірки.***

Оскільки статистичні параметри  $k$  є числами, а останні зображуються точками на числовій осі, підмножини  $u_1$  та  $u_2$  вибірок зводяться до двох підмножин точок числової осі або до двох одновірних областей  $W_1$  та  $W_2$ . Область  $W_1$  параметрів  $k$  називають **критичною областю**, а область  $W_2$  – **областю допустимих 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.

Виникає питання про те, якими принципами треба керуватися при будівництві критичної області  $W_1$ . Ці принципи полягають у тому, що приймаючи чи відкидаючи гіпотезу  $H_0$  можна припустити помилку двох видів.

**Помилка першого роду** полягає у тому, що нульова гіпотеза  $H_0$  відкидається, тобто приймається гіпотеза  $H_1$  тоді, коли в дійсності все ж таки вірною є гіпотеза  $H_0$ .

**Помилка другого роду** допускається тоді, коли приймається гіпотеза  $H_0$ , у той час, коли вірною є гіпотеза  $H_1$ .

Сутність помилок першого і другого роду ілюструє таблиця 1.1



Імовірності помилок першого і другого роду однозначно визначаються вибором критичної області  $W_1$ . Умовимося для будь-якої критичної області  $W_1$  позначити через  $\alpha$  імовірність помилки першого роду.

Таблиця 1.1 – Помилки першого і другого роду [9]

Гіпотеза $H_0$	є вірною	є невірною
Відкидається	Помилка I роду	Правильне рішення
Приймається	Правильне рішення	Помилка II роду

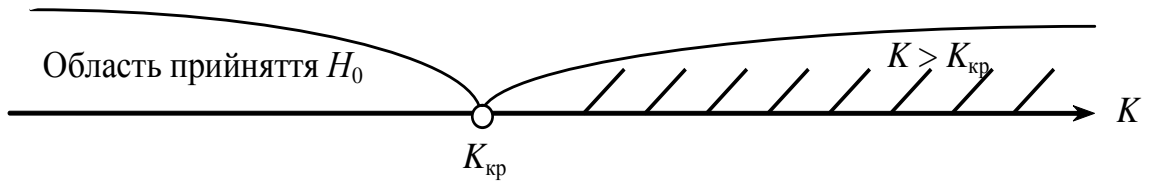
Ймовірність *помилки першого роду називають рівнем значущості*. Критичну область  $W_1$  відокремлює від області прийняття гіпотези  $H_0$  критична точка  $k_{кр}$ , де  $k_{кр}$  – параметр, за допомогою якого задається статистичний критерій (правило). Можуть розглядатися правостороння, лівостороння, двостороння і симетрична двостороння критичні області. Наприклад, при порівнянні двох випадкових величин необхідно оцінити величину їх розбіжності і необхідно врахувати як додатні, так і від’ємні розбіжності. У такому випадку використовують двосторонній критерій. Односторонній критерій використовують у тих випадках, коли необхідно упевнитися, що одна величина більше (менше) іншої.

Правило, яке дозволяє прийняти або усунути статистичну гіпотезу на основі даної вибірки будується таким чином. Якщо значення певного параметру (статистики)  $k$  попадає у критичну область, то нульова гіпотеза відхиляється. При цьому критична область на числовій осі виділяється за допомогою  $k_{кр}$ . Критичні області показані на рис. 1.1.

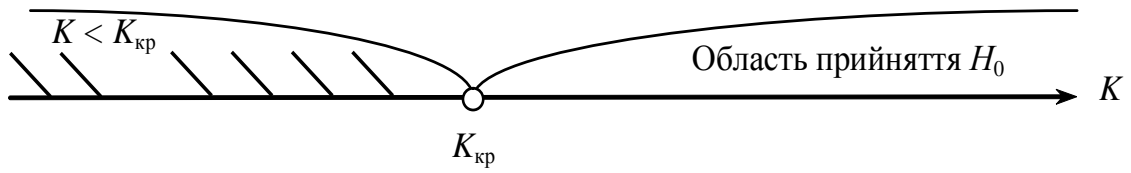
У свою чергу  $k_{кр}$  залежить від рівня значущості  $\alpha$ .

Припустимо, йдеться про правосторонню критичну область (рис. 1.1а). Нехай гіпотеза  $H_0$  є вірною. Тоді імовірність того, що гіпотеза  $H_0$  відкидається тобто, що робиться помилка I роду, дорівнюватиме

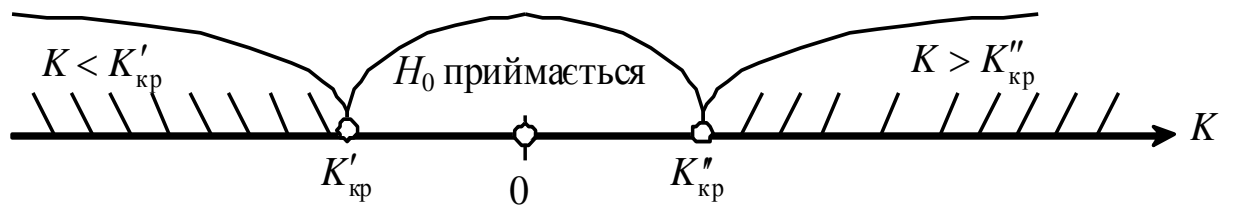
$$p(k > k_{кр}) = \alpha. \quad (1.2)$$



а) правостороння



б) лівостороння



в) двостороння

Рисунок 1.1 – Види критичних областей

Для лівосторонньої критичної області (рис. 1.1б)

$$p(k < k_{кр}) = \alpha, \quad (1.3)$$

для двосторонньої критичної області (рис. 1.1в)

$$p(k < k_{кр2}) + p(k > k_{кр1}) = \alpha . \quad (1.4)$$

Якщо двостороння критична область є симетричною, то

$$p(k < -k_{кр}) + P(k > k_{кр}) = \alpha \quad (1.5)$$

або

$$p(k > k_{кр}) = \frac{\alpha}{2}. \quad (1.6)$$

Рівень значущості впливає на те, з якою ймовірністю основна гіпотеза буде прийнята, цю ймовірність називають довірчою і позначають таким чином

$$p = 1 - \alpha \quad (1.7)$$

## 1.2 Критерії значущості та закони їх розподілу

Відповідь на запитання мала чи велика ймовірність здійснення нульової гіпотези надається за допомогою критеріїв значущості. Критерії значущості застосовуються з метою визначення чи будуть деякі статистики, знайдені за двома або більше сукупностями сильно розрізнятися між собою. Якщо різниця перебільшує випадкові варіації, то вона визнається значущою. Межа між суттєвою і несуттєвою різницею задається за допомогою рівня значущості  $\alpha$ . Критерії значущості базуються на законах розподілу різних статистичних характеристик (статистик), які у першому розділі були позначені як параметр  $k$ . Критерії значущості встановлюються за допомогою закону розподілу параметру (статистики)  $k$ . На базі закону розподілу та заданого рівня значущості визначаються критичні та довірча області. Рівень значущості  $\alpha$  допомагає знайти межу між суттєвою та несуттєвою різницею між характеристиками, які порівнюються. **Рівень значущості є ймовірністю невиконання нульової гіпотези.** Ця ймовірність має бути малою. Вибір рівня значущості робиться вільно, на основі досвіду. Як правило, він приймається рівним 10%, 5% або 1%.

Значення статистики  $k$ , ймовірність появи якої менше заданого рівня значущості, утворюють критичну область (рис.1.1). Якщо досліджувана статистика попадає в критичну область, то нульова гіпотеза відхиляється.

Якщо відомий закон розподілу  $k$ , то за рівнем значущості встановлюють критичну та довірчу області (рис.1.2).

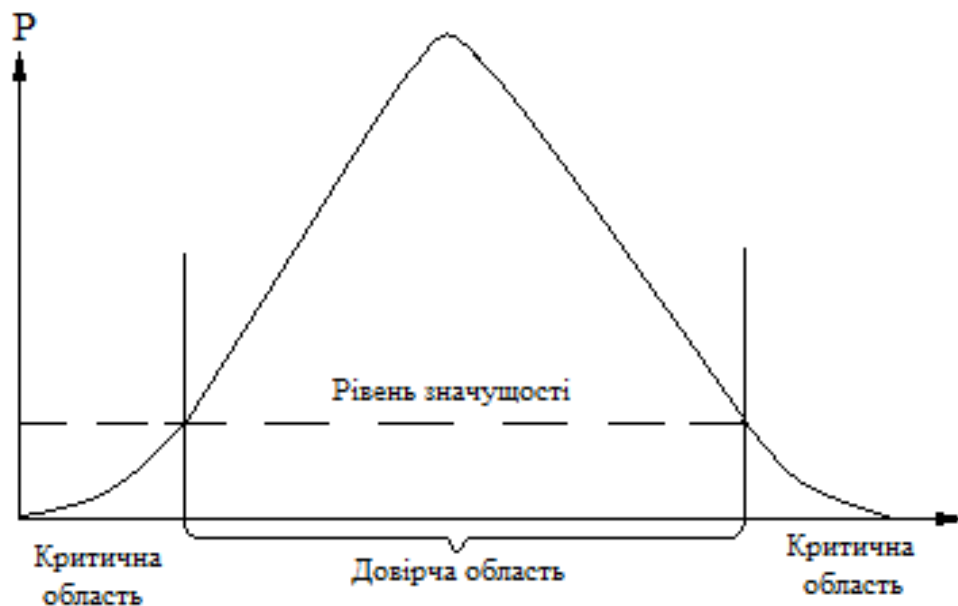


Рисунок 1.2 – Закон розподілу статистики  $k$  і устанавлення довірчої і критичної областей за рівнем значущості  $\alpha$

Для знаходження параметра  $k$  будемо використовувати такі теореми:

*Теорема 1.* Нехай незалежні величини  $z : z_1, z_2, z_3, \dots, z_u$  підпорядковуються нормальному закону розподілу. Тоді сума квадратів цих величин

$$\chi^2 = \sum_{i=1}^n z_i^2 \quad (1.8)$$

підпорядковується закону розподілу, який описується такою щільністю ймовірності

$$f(\chi^2) = \begin{cases} \frac{1}{2^{\frac{\nu}{2}} \Gamma(\frac{\nu}{2})} (\chi^2)^{\frac{\nu}{2}-1} e^{-\frac{\chi^2}{2}}, \text{ при } \chi^2 \geq 0 \\ 0, \text{ при } \chi^2 < 0 \end{cases}, \quad (1.9)$$

де  $\nu$  – число ступенів свободи. Число ступенів свободи є параметром розподілу. Закон розподілу виду (1.9) називають  $\chi^2$ -розподілом.

*Теорема 2.* Нехай маємо дві незалежні випадкові величини  $u$  та  $v$ , які підпорядковуються  $\chi^2$ -розподілу з числами ступенів волі  $\nu_1$  та  $\nu_2$  відповідно. Тоді випадкова величина виду

$$F = \frac{u/\nu_1}{v/\nu_2} \quad (1.10)$$

підпорядковується розподілу

$$f(F) = \begin{cases} \frac{1}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})} \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} \frac{F^{\frac{\nu_1-2}{2}}}{\left(1 + \frac{\nu_1}{\nu_2} F\right)^{\frac{\nu_1+\nu_2}{2}}}, \text{ при } F \geq 0 \\ 0, \text{ при } F < 0 \end{cases}, \quad (1.11)$$

який має назву **закону Фішера-Снедекора**.

Закон розподілу Фішера-Снедекора є двопараметричним, з параметрами  $\nu_1$  та  $\nu_2$ .

*Теорема 3.* Нехай маємо дві незалежні випадкові величини  $u$  та  $v$  такі, що  $u$  – підпорядковується нормальному закону, а  $v$  –  $\chi^2$ -розподілу з числом ступенів свободи  $\nu$ .

Тоді випадкова величина  $t$

$$t = \frac{u}{\sqrt{\frac{v}{\nu}}} \quad (1.12)$$

підпорядковується розподілу

$$f(t) = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}, -\infty < t < \infty, \quad (1.13)$$

який має назву **розподілу Стюдента**. Як видно, цей закон є однопараметричним, з параметром  $\nu$ .

### **1.3 Перевірка статистичної гіпотези про однорідність членів статистичної сукупності (перевірка належності «викидів» до генеральної сукупності)**

Аналізуючи дані спостережень можна виявити певні закономірності у змінах якісного стану водних об'єктів. Наприклад, відомо, що досліджувана річка забруднена фенолами внаслідок діяльності нафтопереробного заводу. Середня річна концентрація фенолів у воді за період з 1980 по 2010 роки змінювалася у границях 0,0004-0,0001 мг/дм<sup>3</sup>, ГДК фенолів становить 0,001 мг/дм<sup>3</sup> (для питних потреб). В 2011 році середньорічна концентрація фенолів у воді склала 0,005 мг/дм<sup>3</sup> (перевищила ГДК у 5 разів), тобто отримане значення різко відрізняється від інших значень. Такі значення величин прийнято називати "**випадіннями**" або "**викидами**". "*Випадіння*" чинять вплив на статистичні параметри, які визначаються за даними спостережень (*середнє арифметичне значення, дисперсія, асиметрія, ексцес*).

Відомо, що середні значення випадкових величин також є випадковими і вони підпорядковуються нормальному закону розподілу. Виберемо максимальне  $x_{\max}$  та мінімальне  $x_{\min}$  значення. Випадкові величини  $\bar{x} - x_{\max}$  та  $\bar{x} - x_{\min}$  також підпорядковуються нормальному закону, тому можна позначити

$$u = |x_{\text{екстр}} - \bar{x}|, \quad (1.14)$$

де  $x_{\text{екстр}}$  – параметр, який являє собою мінімальне, або максимальне значення випадкової величини;

$\bar{x}$  – середня арифметична величина, розрахована за даними спостережень.

Якщо вважати, що випадкова величина  $X$  описується нормальним законом розподілу, то на основі теореми 1 випадкова величина

$$v = \sum_{i=1}^n (x_i - \bar{x})^2 = \chi^2, \quad (1.15)$$

підпорядковується розподілу  $\chi^2$  з числом ступенів свободи  $v = n - 1$ .

Таким чином, величина  $v$  задовольняє другій умові *теореми 3* і на основі цієї теореми отримаємо випадкову величину  $t$ , яка підпорядковується розподілу Стьюдента

$$t = \frac{u}{\sigma_x} = \frac{|x_{\text{екстр}} - \bar{x}|}{\sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}}, \quad (1.16)$$

де  $u$  – значення, розраховане за формулою (1.14);

$n$  – довжина вибірки;

$\sigma_x$  – середнє квадратичне відхилення випадкової величини.

Для перевірки членів ряду на однорідність членів статистичної сукупності формулюють основну гіпотезу  $H_0$  про те, що “викид” належить до тієї ж генеральної сукупності, що й інші члени вибірки. Для її

перевірки розраховують емпіричний критерій Стьюдента для рівня значущості  $\alpha = 0,05$  і числа ступенів свободи  $\nu$  (табл.1.2). Далі перевіряється умова  $t > t_{кр}(\alpha, \nu)$  або  $t < t_{кр}(\alpha, \nu)$  (тобто чи входить «викид» до критичної області). В залежності від результатів приймають або відхиляють гіпотезу  $H_0$ .

Якщо приймається гіпотеза  $H_0$ , то це означає, що «викид» досліджуваного ряду величин належить до тієї ж генеральної сукупності, що й інші члени вибірки. **Якщо відхиляється гіпотеза  $H_0$ , то це означає, що «випадіння» не належать до тієї ж генеральної сукупності,** що і решта членів розглядуваної сукупності. У такому разі неоднорідні члени вилучаються з вибірки, і знову проводиться оцінка відповідних параметрів. Але «випадіння» не відкидаються зовсім. Їх треба пильно вивчати, оскільки вони відбивають ті чи інші аномальні особливості процесів, наслідком яких вони є.

Наведений приклад показує як можна задавати статистику  $k$  для розроблення правила (критерію), за яким приймається чи відхиляється певна гіпотеза. Екстремальні значення випадкових величин мають такі стохастичні властивості, що можуть бути охарактеризовані статистикою  $t$ , яка підпорядковується закону розподілу Стьюдента.

### Приклад розрахунків

Для р. Лопань – м. Харків спостерігається максимальне забруднення по важких металах та по біогенним речовинам за період спостереження 1990-2015 рр. При визначенні показників екологічного ризику у випадку забруднення важкими металами виявлено, що в 1991 році значення ризику дуже сильно відрізняється від всього ряду отриманих цих значень. Висуваємо нульову гіпотезу про те, що значення екологічного ризику  $R$  за 1991 р. належить до тієї ж генеральної сукупності, що і інші члени ряду.

В даному випадку за  $x_{екстр}$  береться значення, яке й перевіряється на викид, тобто  $R=5,15$ . За період спостереження з 1990 по 2015 роки середнє арифметичне значення дорівнює 1,18. Згідно із формулою (1.16) значення статистики Стьюдента розраховувалося таким чином



$$t = \frac{|5,15 - 1,18|}{\sqrt{\frac{18,83}{26-1}}} = 4,56$$

Щоб зробити висновок, чи належить дане значення до тієї ж генеральної сукупності, його необхідно порівняти з  $t_{кр}(\alpha, \nu)$  (на рівні значущості  $\alpha$ , який установлюється дослідником).

При заданому  $\alpha=0,05$   $t_{кр} = 2,06$ . Порівнюючи  $t$  з  $t_{кр}$ , встановлено, що  $t > t_{кр}$ . Отже, нульова гіпотеза про однорідність ряду ризиків відкидається.

Таблиця 1.2 – Значення критерію Стьюдента при рівні значущості  $\alpha = 0.05$

<b>Число ступенів свободи</b>	<b>Рівень значущості 0.05</b>	<b>Число ступенів свободи</b>	<b>Рівень значущості 0.05</b>
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		

## 1.4 Перевірка статистичної гіпотези про незначущість різниць між оцінками дисперсій

Розглянемо, як утворюється критерій перевірки значущості або незначущості різниць між оцінками дисперсій двох вибірок з випадкових величин  $X$  і  $Y$ . Нульова гіпотеза  $H'_0$  полягає у тому, що різниця між оцінками дисперсій приймається незначущою.

Відомо, що ряди випадкових величин  $X$  і  $Y$  підпорядковуються нормальному закону розподілу. На основі *теорему 1* сума їх квадратів підпорядковується  $\chi^2$  розподілу. Отже, випадкові величини

$$u = \sum_{i=1}^n (x_i - \bar{x})^2, \quad (1.17)$$

$$v = \sum_{i=1}^m (y_i - \bar{y})^2, \quad (1.18)$$

підпорядковуються розподілу  $\chi^2$  з числами ступенів свободи  $\nu_1 = n - 1$  та  $\nu_2 = m - 1$ , де  $m$  – довжина вибірки  $X$ , а  $n$  – довжина вибірки  $Y$ . Згідно із *теоремою 2* запишемо

$$\frac{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}{\frac{\sum_{i=1}^m (y_i - \bar{y})^2}{m-1}} = F, \quad (1.19)$$

де  $\bar{x}, \bar{y}$  – середні арифметичні значення вибірок з випадкових величин  $X$  і  $Y$ , довжиною  $n$  та  $m$  відповідно.

$F$  – випадкова величина, яка підпорядковується закону Фішера-Снедекора.

Рівняння (1.19) зазвичай записується таким чином

$$F = \frac{\sigma_x^2}{\sigma_y^2} = \frac{S_x^2}{S_y^2}, \quad (1.20)$$

де  $\sigma_y^2, \sigma_y^2$  або  $S_y^2, S_y^2$  – дисперсії вибірок.

При розрахунках  $F$  у чисельник ставлять більшу з дисперсій. **Перевірка нульової гіпотези здійснюється шляхом порівняння розрахованого значення критерію Фішера  $F$  з критичним  $F_{кр}(\alpha, \nu_1, \nu_2)$ .**

Число ступенів свободи  $\nu_1 = n - 1, \nu_2 = m - 1$

Якщо

$$F < F_{кр}(\alpha, \nu_1, \nu_2) \quad (1.21)$$

то гіпотеза  $H'_0$  (**нульова гіпотеза**) не відхиляється, і навпаки.

У протилежному випадку приймається **альтернативна гіпотеза**, тобто визнається, що різниця між дисперсіями рядів  $X$  і  $Y$  є значущою і не пояснюється лише впливом випадкових флуктуацій у вибірках  $X$  і  $Y$ .

### **1.5. Перевірка статистичної гіпотези про незначущість різниць між середніми значеннями**

Нульовою гіпотезою є гіпотеза про незначущість середніх арифметичних значень, визначених за двома вибірками випадкових величин  $X$  і  $Y$ . Середні арифметичні значення  $\bar{x}$  та  $\bar{y}$  є випадковими величинами, які підкорюються нормальному закону розподілу. Отже, випадкова величина  $u = \bar{x} - \bar{y}$  теж нормально розподілена. Дисперсії  $\sigma_x^2$  і  $\sigma_y^2$  підлягають  $\chi^2$  – розподілу, тоді  $(n-1)\sigma_x^2$  та  $(m-1)\sigma_y^2$  теж підпорядковуються цьому закону. У зв'язку із цим розподіл випадкової величини  $U$

$$U = (n-1)\sigma_x^2 + (m-1)\sigma_y^2, \quad (1.22)$$

також підлягає закону  $\chi^2$ , де  $\nu = m + n - 2$  – кількість ступенів свободи.

Величини  $u$  та  $U$  задовольняють умовам *теорема 3*. Отже, можна використати статистику (1.11)

$$t = \frac{|\bar{x} - \bar{y}|}{\sqrt{\frac{(n-1)\sigma_x^2 + (m-1)\sigma_y^2}{m+n-2} \left( \frac{1}{m} + \frac{1}{n} \right)}}, \quad (1.23)$$

яка підпорядковується розподілу Стьюдента.

*Гіпотеза  $H_0$  про незначущість різниці між середніми арифметичними значеннями перевіряється шляхом порівняння критерію Стьюдента  $t$  з його критичним значенням  $t_{кр.}(\alpha, \nu)$  за допомогою спеціально розробленої таблиці (табл.1.3). Нульова гіпотеза  $H_0$  приймається, коли  $t < t_{кр.}(\alpha, \nu)$ .*

Критерії Стьюдента та Фішера у описаній вище формі застосовуються для оцінки статистичної однорідності рядів. *Однорідними називають такі ряди даних, які підпорядковуються одному і тому ж закону розподілу, тобто належать до однієї і тієї ж генеральної сукупності.* Оскільки у нормальних законах розподілу статистичними параметрами є оцінки математичного сподівання та дисперсії, то оцінку однорідності рядів часто виконують перевіряючи належність середніх арифметичних та дисперсій до однієї й тієї ж генеральної сукупності.

*Статистична гіпотеза про однорідність рядів  $X$  і  $Y$  приймається тільки тоді, коли справедливі обидві гіпотези: про незначущість різниць між дисперсіями та незначущість різниць між середніми.*

Таблиця 1.3 – Значення критерію Фішера  $F$  для рівня значущості 0.05

$\nu_1$  – число ступенів свободи для більшої дисперсії

$\nu_2$  – число ступенів свободи для меншої дисперсії

$\nu_2$	$\nu_1$								
	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
$\infty$	1.8	1.6	1.6	1.6	1.5	1.5	1.4	1.4	1.3

## 1.6 Перевірка статистичної гіпотези про однорідність членів статистичної сукупності на основі непараметричних критеріїв (критерій Вілкоксона)

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:

$$\begin{aligned} X &: x_1, x_2, x_3, x_4, x_5 \\ Y &: y_1, y_2, y_3, y_4, y_5, y_6 \\ u &: x_1 y_2 y_3 x_2 x_3 y_4 x_4 y_5 y_6 x_5 \end{aligned}$$

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

$$u = 1 + 3 + 3 + 4 + 6 = 17.$$

В однорідних рядах, кожне з котрих має не менше 10 членів, число інверсій розподіляється приблизно за нормальним законом з математичним сподіванням [9]

$$m_u = \frac{m \cdot n}{2} \quad (1.24)$$

і дисперсію

$$\sigma_u^2 = \frac{m \cdot n}{12} (m + n + 1), \quad (1.25)$$

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  $\alpha$  або довірчої ймовірності  $p = 1 - \alpha$ , знаходять границі допустимих значень  $u$ , які 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  $p$  приймається альтернативна гіпотеза.

Область прийняття гіпотези  $H_0$  визначається нерівністю

$$m_u - t_{кр}(\alpha, \nu)\sigma_u \leq u \leq m_u + t_{кр}(\alpha, \nu)\sigma_u, \quad (1.26)$$

а критична область – нерівностями

$$u < m_u - t_{кр}(\alpha, \nu)\sigma_u, \quad (1.27)$$

$$u > m_u + t_{кр}(\alpha, \nu)\sigma_u, \quad (1.28)$$

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

$t_{кр}(\alpha, \nu)$  – критерій Стьюдента для рівня значущості  $\alpha$  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.

### Контрольні запитання:

1. Дати визначення гіпотези.
2. Дати визначення нульової гіпотези.
3. Дати визначення альтернативної гіпотези.
4. Які статистичні гіпотези називаються параметричними?
5. Які статистичні гіпотези називаються непараметричними?
6. Дати визначення рівню значущості.
7. У чому полягає помилка першого роду?
8. У чому полягає помилка другого роду?
9. У яких випадках потрібен двосторонній критерій?
10. У яких випадках потрібен односторонній критерій?
11. За допомогою якого критерію значущості перевіряється статистична гіпотеза про наявність «викидів» у даних спостережень?
12. Які ряди називають однорідними?
13. За допомогою якого критерію перевіряється гіпотеза  $H_0$  про незначущість різниці між дисперсіями?
14. За допомогою якого критерію перевіряється Гіпотеза  $H_0$  про незначущість різниці між середніми арифметичними значеннями?
15. У яких випадках приймається гіпотеза про однорідність рядів  $X$  і  $Y$ ?



## РОЗДІЛ 2

### КРИТЕРІЇ ВИПАДКОВОСТІ

У гідрометеорологічних та гідроекологічних дослідженнях часто висувається припущення, що коливання досліджуваних величин випадкові. У випадкових рядах результати спостережень незалежні та можуть спостерігатися у будь-якому порядку. Для установлення випадковості коливань досліджуваної величини розроблені спеціальні критерії.

#### 2.1 Ранговий критерій випадковості із застосуванням коефіцієнта Кендалла

Моріс Кендалл (1907-1983 рр.) – англійський вчений, автор численних праць зі статистики і теорії ймовірностей. Його ім'я носить один з коефіцієнтів рангової кореляції.

Нульова гіпотеза полягає у тому, що приймається пропозиція про випадковість коливань досліджуваної характеристики. Якщо у коливаннях простежується певна тенденція, яка є статистично значущою, то її називають трендом. Таким чином, нульова гіпотеза надає пропозицію, що тренду немає. Альтернативна гіпотеза полягає у тому, що існування тренду підтверджується [10].

Критерій рангової кореляції (коефіцієнт Кендалла) використовується для визначення значущості лінійного тренду досліджуваної характеристики. Спочатку необхідно визначити кількість інверсій ( $P$ ). Про інверсію говориться, що коли в упорядкованому (убутному) ряді членів будь-якому значенню  $u_i$  передують значення  $u_j$ , де  $u_j > u_i$ , то говорять, що ця пара утворює інверсію. Отже, треба установити, скільки чисел у досліджуваній вибірці перевищує кожне число вибірки. Для зручності необхідно досліджуваний ряд розташувати в убутному порядку (табл. 2.1).

Кількість пар дорівнює  $\frac{1}{2}n(n-1)$ , а математичне сподівання числа  $P$  для випадкового ряду визначається за формулою

$$m_p = \frac{1}{4}n(n-1), \quad (2.1)$$

де  $m_p$  – математичне сподівання числа  $P$ ;  
 $n$  – довжина ряду.

Таблиця 2.1 – Значення показника якості води, розрахованого за  $I_1$  [11] та інверсій ( $P$ ) у досліджуваному створі

<i>№ з/п</i>	<i>Роки</i>	$I_1$	$I_1$ (убуттє)	$P$
1	1996	3,33	3,33	16
2	2000	2,34	2,91	15
3	1997	1,68	2,9	14
4	1998	2,4	2,7	13
5	1999	2,5	2,64	12
6	2001	2,1	2,61	11
7	2002	1,67	2,53	10
8	2003	2,9	2,5	9
9	2004	2,3	2,4	8
10	1995	2,2	2,37	7
11	2005	2,7	2,34	6
12	2006	2,33	2,33	5
13	2009	2,53	2,3	4
14	2010	2,61	2,2	3
15	2007	2,64	2,1	2
16	2008	2,91	1,68	1
17	1994	2,37	1,67	0
				<b><math>\Sigma P=136</math></b>

Якщо визначене  $P$  перевищує математичне сподівання (2.1), то це вказує на наявність додатного тренду, якщо число  $P$  менше  $m_p$ , то це вказує на наявність від’ємного тренду.

Число  $P$  пов'язане простим співвідношенням з коефіцієнтом рангової кореляції  $\tau$  (коефіцієнт Кендала), який може змінюватися від  $-1$  до  $+1$  ( $-1 < \tau < 1$ ) і визначається за формулою

$$\tau = \frac{4P}{n(n-1)} - 1, \quad (2.2)$$

де  $\tau$  – коефіцієнт Кендалла;

$P$  – число випадків, коли  $u_j > u_i$  при  $j > i$ ;

$n$  – довжина досліджуваного ряду.

Математичне сподівання  $\tau$  для випадкового ряду дорівнює нулю ( $m_\tau = 0$ ), а дисперсія

$$\sigma_\tau^2 = \frac{2(2n+5)}{9n(n-1)}, \quad (2.3)$$

Середнє квадратичне відхилення розраховується за формулою

$$\sigma_\tau = \sqrt{\frac{2(2n+5)}{9n(n-1)}}, \quad (2.4)$$

Значущість тренду встановлюється за виконанням умови

$$|\tau| > 2\sigma_\tau. \quad (2.5)$$

### Приклад розрахунків

Кількість інверсій  $P$  дорівнює 136.

Математичне сподівання числа  $P$  для досліджуваного ряду за (2.1)

$$m_p = \frac{1}{4}n(n-1) = \frac{1}{4}17(17-1) = 68$$

Оскільки  $P$  перевищує математичне сподівання  $m_p$  ( $136 > 68$ ), то це вказує на наявність додатного тренду.

Розраховуємо коефіцієнт Кендалла за (2.2)

$$\tau = \frac{4P}{n(n-1)} - 1 = \frac{4 * 136}{17(17-1)} - 1 = 1$$

Дисперсію для досліджуваного ряду визначаємо за (2.3)

$$\sigma_\tau^2 = \frac{2(2n+5)}{9n(n-1)} = \frac{2(2*17+5)}{9*17(17-1)} = 0,0319,$$

Середнє квадратичне відхилення визначаємо за (2.4)

$$\sigma_\tau = \sqrt{\frac{2(2n+5)}{9n(n-1)}} = \sqrt{\frac{2*(2*17+5)}{9*17(17-1)}} = 0,179$$

Значущість тренду встановлюється за виконанням умови (2.5)

Оскільки  $\tau = 1$ , а  $2\sigma_\tau = 2*0,179 = 0,357$ , то виконується умова  $|\tau| > 2\sigma_\tau$ , що свідчить про значущість додатного тренду.

## 2.2 Критерій Аббе

Нульова гіпотеза полягає у тому, що коливання досліджуваної характеристики приймаються випадковими (тренду немає), тобто перевіряється гіпотеза про відсутність систематичних змін у спостереженнях. 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

$$S^2 = \frac{1}{2(N-1)} \cdot \sum_{i=1}^{N-1} (x_{i+1} - x_i)^2, \quad (2.6)$$

де  $N$  – довжина вибірки (кількість спостережень);

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

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

$$Z = \frac{S^2}{\sigma_X^2}, \quad (2.7)$$

де  $\sigma_X^2$  – 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 2.2.

Таблиця 2.2 – Критичні значення розподілу величини  $Z$

<b>Число даних</b>	<b>5%-ий рівень</b>	<b>Число даних</b>	<b>5%-ий рівень</b>
<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>
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

Таблиця 2.2 – Продовження

Число даних	5%-ий рівень	Число даних	5%-ий рівень
17	0,624	48	0,768
18	0,633	49	0,770
<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>
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		

Недоліком критерію Аббе є те, що генеральна сукупність, з якої вилучається ряд спостережень, приймається нормальною, тому функція  $Z$  може «зреагувати» на циклічні коливання, якщо вони властиві досліджуваній характеристиці (наприклад, рядам стоку).

### 2.3 Критерій Дарбіна-Уотсона

Нульова гіпотеза полягає у прийнятті припущенні про відсутність у ряді автокореляції [12]. Нульова гіпотеза полягає у Критерій Дарбіна-Уотсона  $D$  використовується при вивченні залишків, які отримані в результаті віднімання з вихідного ряду систематичних елементів, коли треба встановити, чи не залишилось в ньому ще будь-якої систематизації. Залишки можуть розглядатися як різниця між фактичним  $y_i$  і розрахунковим  $\tilde{y}_i$  значенням випадкової величини:  $y_i - \tilde{y}_i$ . Якщо якість використаної математичної моделі (наприклад, моделі лінійної парної

регресії) задовільна, то між залишками не має бути кореляційного зв'язку. Якщо між залишками буде додатна кореляція, то це означає, що застосована математична модель не є придатною до вирішення поставленої задачі.

У цьому випадку статистика Дарбіна-Уотсона  $D$  записується у виді

$$D = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2}, \quad (2.8)$$

де  $e_i = y_i - \tilde{y}_i$ .

Рівняння (2.8) може бути представленим наступним чином

$$D = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2} = \frac{\sum_{i=2}^n e_i^2}{\sum_{i=1}^n e_i^2} - 2 \frac{\sum_{i=2}^n e_i e_{i-1}}{\sum_{i=1}^n e_i^2} + \frac{\sum_{i=2}^n e_{i-1}^2}{\sum_{i=1}^n e_i^2}. \quad (2.9)$$

↓  
«1»

↓  
«1»

$$D = 1 - 2 \frac{\sum_{i=2}^n e_i e_{i-1}}{\sum_{i=1}^n e_i^2} + 1 = 2 - 2 \frac{\sum_{i=2}^n e_i e_{i-1}}{\sum_{i=1}^n e_i^2} = 2 \left( 1 - \frac{\sum_{i=2}^n e_i e_{i-1}}{\sum_{i=1}^n e_i^2} \right) = 2[1 - r(1)], \quad (2.10)$$

$$D = 1 - 2r(1) + 1 = 2 - 2r = 2[1 - r(1)], \quad (2.11)$$

де  $r(1)$  – коефіцієнт кореляції між суміжними значеннями залишків.

Якщо зв'язок позитивний і добре виражений, то коефіцієнт кореляції наближається до 1, тоді  $D \rightarrow 0$ . Якщо зв'язок негативний і добре виражений, то коефіцієнт кореляції наближається до  $-1$ , тоді  $D \rightarrow 4$ .

Якщо зв'язок слабкий і коефіцієнт кореляції наближається до нуля, то  $D \rightarrow 2$ .

На практиці застосування критерію Дарбіна-Уотсона базується на порівнянні  $D$  з теоретичними значеннями  $d_L$  та  $d_U$  для заданої довжини ряду  $n$ , кількості незалежних змінних  $k'$  у рівнянні регресії та рівня значущості  $\alpha$  (табл.2.2).

Якщо  $D < d_L$ , то гіпотеза про незалежність випадкових відхилень відхиляється (має місце позитивна автокореляція).

Якщо  $D > d_U$ , то гіпотеза про незалежність випадкових відхилень не відхиляється.

Якщо  $d_L < D < d_U$ , то немає достатніх підстав для прийняття рішення.

Якщо  $D > 2$ , то з  $d_L$  та  $d_U$  порівнюється не сам коефіцієнт  $D$ , а вираз  $4 - D$ .

### **Контрольні запитання:**

1. Які ряди вважаються випадковими?
2. Дати визначення тренду.
3. Дати визначення поняттю залишки.
4. Для чого використовується критерій Дарбіна-Уотсона?
5. Для чого використовується критерій Аббе?



## РОЗДІЛ 3

### ПЕРЕВІРКА НАЛЕЖНОСТІ ВИБІРКИ ДО НОРМАЛЬНО РОЗПОДІЛЕНОЇ СУКУПНОСТІ

#### 3.1 Нормальний закон розподілу випадкової величини

Серед законів розподілу безперервних випадкових величин нормальний закон розподілу або закон Гаусса займає особливе місце. Цей закон добре розроблений і доступний, до нього наближаються інші закони розподілу при певних умовах. 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.

Нормальний закон розподілу є частковим випадком розподілу Пірсона.

Сукупність теоретичних кривих розподілу випадкових величин Пірсона можна отримати в результаті вирішення диференціального рівняння виду

$$\frac{dy}{dx} = \frac{y(x+d)}{\varphi(x)}, \quad (3.1)$$

де  $d$  – відстань між модою  $m_0$  та математичним сподіванням  $m_x$  на графіку кривої розподілу (рис.3.1);

$y = f(x)$  – щільність ймовірності;

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

$$\varphi(x) = b_0 + b_1x + b_2x^2 + \dots + b_nx^n. \quad (3.2)$$

Якщо використовувати лише перший член ряду Маклорена  $b_0$ , приймаючи, що інші його члени дорівнюють нулю, то отримаємо диференціальне рівняння нормального закону розподілу

$$\frac{dy}{dx} = \frac{y(x+d)}{b_0} . \quad (3.3)$$

Нормальний закон розподілу має такі особливості.

1. Математичне сподівання і мода співпадають, тобто  $m_x = m_0$ .
2. Параметр  $-b_0$  дорівнює дисперсії випадкової величини  $\sigma_x^2$ .
3. Коефіцієнт асиметрії дорівнює нулю ( $C_S=0$ ), тобто розподіл симетричний.

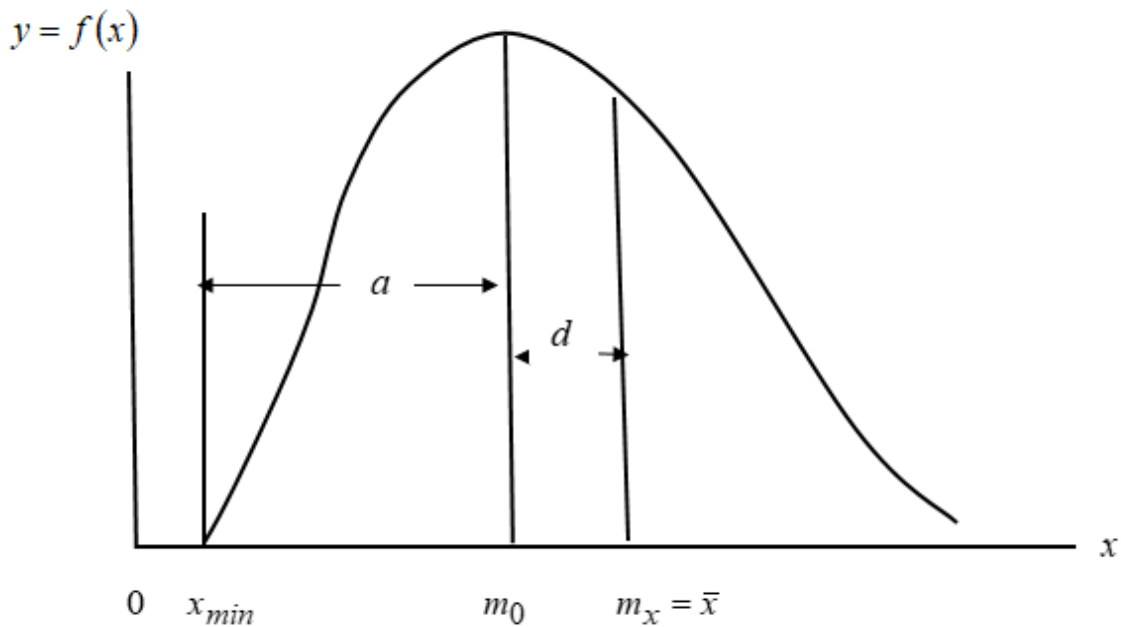


Рисунок 3.1 – Крива Пірсона III ( $C_S > 0$ )

Рівняння щільності розподілу випадкової величини, яка підкорюється нормальному закону розподілу має такий вигляд

$$f(x) = \frac{1}{\sigma_x \sqrt{2\pi}} e^{-\frac{(x-m_x)^2}{2\sigma_x^2}} . \quad (3.4)$$

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

$$t = \frac{x - m_x}{\sigma_x} = \frac{k - 1}{C_V}, \quad (3.5)$$

де  $C_V$  – коефіцієнт варіації;

$k$  – модульний коефіцієнт,

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

$$f(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}. \quad (3.6)$$

Вираз (3.6) задається у вигляді таблиці для практичного використання (табл.3.1).

Таблиця 3.1 – 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 випадкової величини має декілька особливостей.

1. У зв'язку з тим, що функція (3.6) має дійсні значення при будь-яких значеннях незалежної змінної  $X$ , область її визначення така:  $-\infty < X < +\infty$ .

2. Функція  $f(x)$  є парною, тобто  $f(-X) = f(+X)$ .

3. Крива розподілу не перетинає осі  $x$ .

4. Крива щільності розподілу симетрична відносно моди.

5. Нормальний закон розподілу випадкової величини є двопараметричним, тобто в ньому використовуються два параметри – математичне сподівання  $m_x$  та дисперсія  $\sigma_x^2$ .

6. Параметр  $\sigma_x$  є характеристикою форми кривої розподілу: чим більше  $\sigma_x$ , тим максимальна ордината менше, а крива сплющується (рис.3.2)

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

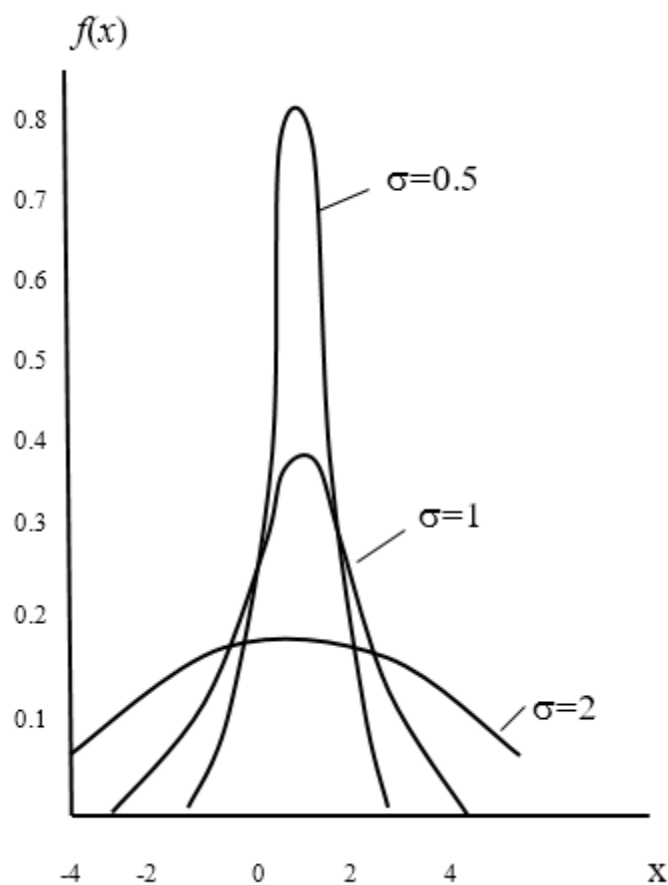


Рисунок 3.2 – Графік щільності ймовірностей нормального закону розподілу при різних середніх квадратичних відхиленнях  $\sigma_x$

Інтегральна функція випадкової величини, яка підлягає нормальному закону розподілу має такий вигляд

$$F(x) = \frac{1}{\sigma_x \sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(x-m_x)^2}{2\sigma_x^2}} dx, \quad (3.7)$$

або

$$F(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt. \quad (3.8)$$

При розрахунках використовується спеціальна функція інтеграла ймовірностей або функція Лапласа.

Функція Лапласа визначається таким чином

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{z^2}{2}} dz, \quad (3.9)$$

де  $\frac{t^2}{2} = z^2$ ,  $t = z\sqrt{2}$ ,  $dt = \sqrt{2}dx$ .

Тоді

$$F(x) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^x e^{-z^2} dz. \quad (3.10)$$

Інтегральна функція нормально розподіленої величини може бути розрахованою через функцію Лапласа (табл.3.2)

$$F(x) = \frac{1}{2} [\Phi(z) + 1] \quad (3.11)$$

Формула (3.13) може бути використана для розрахунку ймовірностей перевищень заданої величини.

Так, наприклад, розрахуємо ймовірність перевищення  $t=1,64$ . В цьому випадку  $z = t/\sqrt{2} = 1,16$  і  $\Phi(z) = 0,900$  (табл. 3.2). Звідси

$$F(x) = \frac{1}{2} \cdot (0,900 + 1) = 0,95.$$

Через інтегральну функцію  $F(x)$  можна визначати функцію забезпеченості (ймовірності перевищення заданого  $x$ ). Функція забезпеченості розраховується таким чином

$$P(x) = 1 - F(x) = 0,05.$$

Таблиця 3.2 – Значення функції Лапласа  $\Phi(z)$

$z$	$\Phi(z)$	$z$	$\Phi(z)$	$z$	$\Phi(z)$
0,00	0,000	0,50	0,529	1,20	0,900
0,05	0,056	0,60	0,601	1,50	0,966
0,10	0,125	0,70	0,678	1,80	0,989
0,20	0,223	0,80	0,742	2,00	0,995
0,30	0,329	0,90	0,797	2,50	0,9996
0,40	0,428	1,00	0,843	3,00	1,0000

Аналогічно можуть бути розраховані координати функції розподілу і функції забезпеченості нормованого ряду для різних значень ймовірностей.

Так, при  $F(x) = 0,01$  за формулою (3.11)

$$0,01 = \frac{1}{2} [\Phi(z) + 1].$$

Звідки

$$\Phi(z) = -0,98.$$

Використовуючи табл.3.2, знаходимо за  $\Phi(z)$  величину  $z$

$$z = -1,65$$

Враховуючи, що  $z = t/\sqrt{2}$ , отримуємо  $t = -1,65 \cdot 1,41 = -2,33$ .

Точно так для нормованих значень ряду  $X$  розраховувались координати інших точок функції розподілу Гауса.

По координатам функцій розподілу за формулою  $P(x) = 1 - F(x)$  можуть бути отримані координати функцій забезпеченості (табл. 3.3) та побудовані криві забезпеченості.

Таблиця 3.3 – Ординати кривої забезпеченості нормованих значень ряду  $X$

$$(t_p = \frac{x_p - m_x}{\sigma_x})$$

$P\%$	0,01	0,10	1,00	3	5	10	25	50
$t_p$	3,72	3,09	2,33	1,88	1,64	1,28	0,67	0,00

Таблиця 3.3 – Продовження

$P\%$	75	90	95	97	99	99,9	99,99
$t_p$	-0,67	-1,28	-1,65	-1,88	-2,23	-3,09	-3,72

### 3.2 Оцінка ступеня відхилення емпіричного закону розподілу випадкової величини від нормального

Рішення про можливу апроксимацію (наближення) емпіричного закону розподілу випадкової величини до нормального закону розподілу можна прийняти аналізуючи вибіркові значення коефіцієнтів асиметрії та ексцесу.

Коефіцієнт асиметрії розраховується за формулою

$$C_S = \frac{n}{(n-1)(n-2)} \frac{\sum_{i=1}^n (k_i - 1)^3}{C_V^3}, \quad (3.12)$$

де  $k_i = x_i / \bar{x}$  – модульний коефіцієнт;

$\bar{x}$  – середнє арифметичне значення;

$n$  – довжина вибірки (довжина ряду спостережень);

$C_V$  – коефіцієнт варіації, який розраховується таким чином

$$C_V = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\bar{x}^2 (n-1)}} = \sqrt{\frac{\sum_{i=1}^n (k_i - 1)^2}{n-1}} . \quad (3.13)$$

Ексцес характеризує сплющеність або витягнутість кривої розподілу випадкової величини у порівнянні з кривою нормального розподілу. Безрозмірний статистичний параметр, названий ексцесом, розраховується за формулою

$$E = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{n\sigma_x^4} - 3. \quad (3.14)$$

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.3.3).

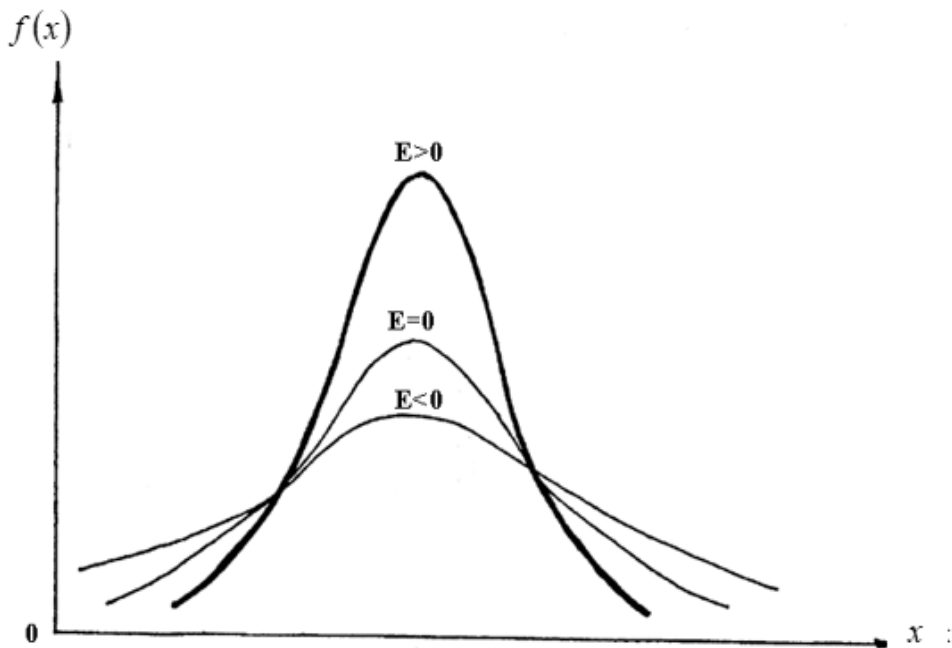


Рисунок 3.3 – Криві розподілу з різними значеннями ексцесу  $E$

Якщо модель нормального закону розподілу відповідає даним спостережень, то відношення коефіцієнтів асиметрії та ексцесу до їх середніх квадратичних відхилень також розподілені нормально із



математичним сподіванням, яке дорівнює нулю і дисперсією, що дорівнює одиниці.

Нульова гіпотеза полягає у тому, що нормальний закон розподілу задовільно описує вибірку. Вона приймається, якщо

$$\frac{C_S}{\sigma_{C_S}} \leq 3 \quad (3.15)$$

та

$$\frac{E}{\sigma_E} \leq 3, \quad (3.16)$$

де

$$\sigma_{C_S} = \sqrt{\frac{6}{n}(1 + 6C_V^2 + 5C_V^4)} \quad (3.17)$$

та

$$\sigma_E = \sqrt{\frac{24}{n}}. \quad (3.18)$$

Нульова гіпотеза відхиляється, коли умови (3.15) та (3.16) не виконуються.

### 3.3 Критерій Гаусса

У цій методиці аксіоматичним є допущення, ще більшість результатів господарської діяльності, в тому числі й ті, які зумовлюють забруднення навколишнього середовища (зокрема, водного середовища), являють собою випадкові величини і підкоряються закону, близькому до нормального (закону Гаусса). Нормальний закон розподілу випадкової величини дуже часто використовується при вивченні показників ризику.

Перевірка відповідності досліджуваної випадкової величини  $R$  на підпорядкованість її нормальному закону розподілу виконується за критерієм (правилом) Гаусса, яке записується таким чином

$$\frac{\sigma_R}{\rho_R} = \sqrt{\frac{\pi}{2}} \approx 1,25, \quad (3.19)$$

де  $\sigma_R$  – середнє квадратичне відхилення показника ризику  $R_i$  від його середнього арифметичного значення ( $R_{сep}$ )

$$\sigma_R = \sqrt{\frac{\sum_{i=1}^k (R_i - R_{сep})^2}{k-1}}, \quad (3.20)$$

$\rho_R$  – середнє арифметичне відхилення виду

$$\rho_R = \frac{1}{k} \sum_{i=1}^k |R_i - R_{сep}|, \quad (3.21)$$

де  $k$  – довжина ряду значень  $R$ , які осереднювались

$$R_{сep} = \frac{1}{k} \sum_{i=1}^k R_i. \quad (3.22)$$

Приклад результатів розрахунків представляється у вигляді такої таблиці:

Таблиця 3.4 – Статистичні параметри ряду показника ризиків  $R$  (за даними ДАВР про хімічний склад води у створі Дністер – Біляївка) [13]

Середнє арифметичне значення $R_{сp}$	Середнє квадратичне відхилення $\sigma_R$	Середнє арифметичне відхилення $\rho_R$	Критерій Гаусса
87,2	30,1	24,1	1,24

Нульова гіпотеза про те, що нормальний закон розподілу задовільно описує вибірку, приймається, оскільки відношення  $\frac{\sigma_R}{\rho_R}$  набуло значення близького до 1,25.

### 3.4 Критерій Діксона-Томпсона

Нульова гіпотеза містить у собі твердження, що усі значення вибірки належать до однієї сукупності, яка підкорюється нормальному закону розподілу. Альтернативна гіпотеза передбачає, що екстремальне значення вибірки не належить до нормально розподіленої сукупності. Для проведення тесту Діксона-Томпсона дані ранжуються від найменшого до найбільшого, причому найменше позначається як  $X_1$ , а найбільше – як  $X_n$ . Нижній індекс вказує на ранг значення від найменшого до найбільшого. Статистика  $R$  і її критичне значення  $R_c$  залежать від розміру вибірки (табл.3.5). Нульова гіпотеза відхиляється, якщо  $R$  більше  $R_c$ .

Наприклад, треба перевірити найбільше значення на викид з такої вибірки ранжованої вибірки 25, 38, 40, 54, 58, 60, 92, 94, 118, 145, 150, 159, 171, 208, 212, 171, 208, 212, 234, 262, 340, 500, 538, 6690. Довжина ряду дорівнює 24. Відповідне рівняння з таблиці 3.5 дозволяє визначити статистику Діксона-Томпсона:

$$R = \frac{X_n - X_{n-2}}{X_n - X_3} = \frac{6690 - 500}{6690 - 40} = \frac{6190}{6650} = 0,931.$$

Критичні значення, отримані з таблиці 3.5, становлять 0,405, 0,443 і 0,485 для рівнів значущості 5%, 2,5% і 1% відповідно. У всіх випадках  $R > R_c$ , отже нульова гіпотеза відхиляється. Найбільше значення є викидом, якщо припустити належність вибірки до нормального закону розподілу.

Перевірка найменшого значення з вибірки (11, 18, 21, 37, 41, 57, довжина ряду дорівнює 6 ) на викид виконується за допомогою рівняння виду

$$R = \frac{X_2 - X_1}{X_n - X_1} = \frac{18 - 11}{57 - 11} = \frac{7}{46} = 0.152.$$

Таблиця 3.5 – Рівняння для визначення статистики Діксона-Томпсона та критичні значення [14]

Розмір вибірки	Статистика $R$ для перевірки на викид найменшого значення	Статистика $R$ для перевірки на викид найбільшого значення	$m^*$	Критичні значення для різних рівнів значущості		
				5%	2,5%	1%
Від 3 до 7	$R = \frac{X_2 - X_1}{X_n - X_1}$	$R = \frac{X_n - X_{n-1}}{X_n - X_1}$	3	0,943	0,970	0,968
			4	0,765	0,829	0,889
			5	0,641	0,707	0,777
			6	0,560	0,626	0,693
Від 8 до 10	$R = \frac{X_2 - X_1}{X_{n-1} - X_1}$	$R = \frac{X_n - X_{n-1}}{X_n - X_2}$	7	0,503	0,562	0,630
			8	0,549	0,610	0,678
			9	0,506	0,565	0,630
			10	0,472	0,528	0,590
Від 11 до 13	$R = \frac{X_3 - X_1}{X_{n-1} - X_1}$	$R = \frac{X_n - X_{n-2}}{X_n - X_2}$	11	0,570	0,617	0,670
			12	0,540	0,586	0,637
			13	0,515	0,560	0,610
			14	0,538	0,583	0,632
Від 14 до 25	$R = \frac{X_3 - X_1}{X_{n-2} - X_1}$	$R = \frac{X_n - X_{n-2}}{X_n - X_3}$	15	0,518	0,562	0,611
			16	0,499	0,542	0,590
			17	0,482	0,525	0,574
			18	0,467	0,509	0,556
			19	0,455	0,495	0,541
			20	0,444	0,482	0,528
			21	0,431	0,470	0,516
			22	0,422	0,461	0,506
			23	0,414	0,452	0,494
			24	0,405	0,443	0,485
			25	0,397	0,435	0,480

\* порядковий номер елемента, у ряду значень, ранжованих від найменшого значення до більшого

Критичні значення, отримані з таблиці 3.5, становлять 0.560, 0.626 і 0.693 для рівнів значущості 5%, 2.5% і 1% відповідно. У всіх випадках  $R < R_C$ , отже нульова гіпотеза приймається. Результати показують, що найменше значення суттєво не відрізняється від показників, які можна очікувати за припущенням нормальної популяції. Тому найнижче значення не є викидом, якщо припускається нормальний розподіл випадкової величини.

### Контрольні запитання:

1. Указати область визначення нормально розподіленої випадкової величини.
2. Як виглядає крива нормального розподілу відносно моди (симетричне чи несиметричне її положення)?
3. Скільки параметрів у математичному виразі для нормального закону розподілу?
4. Для чого використовується функція Лапласа?
5. Чому дорівнює коефіцієнт асиметрії нормально розподіленої випадкової величини?
6. Чому дорівнює ексцес нормально розподіленої величини?
7. Для чого використовується критерій Діксона-Томпсона?

## РОЗДІЛ 4

### ДОВІРЧІ ІНТЕРВАЛИ ТА ОЦІНКА ВІРОГІДНОСТІ ПАРАМЕТРІВ РІВНЯННЯ ЛІНІЙНОЇ ПАРНОЇ РЕГРЕСІЇ

Обчислюючи на основі наявних вибірових даних оцінку  $\hat{\theta}$  деякого параметра  $\theta$ , ми усвідомлюємо, що насправді величина  $\hat{\theta}$  є лише наближеним значенням параметра генеральної сукупності  $\theta$ , навіть у тому випадку, коли ця оцінка є обґрунтованою, незміщеною й умотивованою. Виникає питання, як сильно може відхилитися це наближене вибірове значення (оцінка) від відповідного значення генеральної сукупності? Вирішення цього питання досягається шляхом установалення довірчих інтервалів.

Щоб отримати уявлення про точність оцінки  $\hat{\theta}$ , треба визначити можливу похибку. Призначимо досить велику (близьку до 1) ймовірність  $\beta$  (наприклад,  $\beta=0,95$ ), щоб подію з ймовірністю  $\beta$  можна було б вважати достовірною, та знайдемо таке значення  $\varepsilon$ , для якого виконується наступна умова

$$p(|\hat{\theta} - \theta| < \varepsilon) = \beta. \quad (4.1)$$

Тоді діапазон можливої похибки при заміні  $\theta$  на  $\hat{\theta}$  буде дорівнювати  $\pm \varepsilon$ . Великі за абсолютною величиною похибки будуть з'являтися тільки з малою ймовірністю  $\alpha = 1 - \beta$ , яка має назву рівня значущості.

Перепишемо (4.1) у вигляді

$$p(\hat{\theta} - \varepsilon < \theta < \hat{\theta} + \varepsilon) = \beta. \quad (4.2)$$

Рівність (4.2) означає, що із ймовірністю  $\beta$  значення параметра  $\theta$  потрапляє в інтервал

$$I_{\beta} = (\hat{\theta} - \varepsilon; \hat{\theta} + \varepsilon), \quad (4.3)$$

який має назву довірчого. Ймовірність  $\beta$  називають довірчою. Чим меншим для обраної ймовірності  $\beta$  є довірчий інтервал, тим точнішу оцінку  $\theta$  отримаємо. **Якщо оцінка параметру виходить за межі довірчого інтервалу, то вона визнається несумісною із даними спостережень.**

Щоб знайти довірчий інтервал для будь-якого статистичного параметру потрібно знати закон розподілу оцінки  $\hat{\theta}$ , яка залежить від закону розподілу випадкової величини  $X$ . Найчастіше приймається, що випадкова величина підкорюється нормальному закону розподілу.

#### 4.1 Визначення довірчого інтервалу для математичного сподівання

Нехай випадкова величина  $X$  підкорюється нормальному закону розподілу. Треба знайти довірчий інтервал для математичного сподівання  $m_x$ . Якщо ряд спостережень підкорюється нормальному закону розподілу, то його середні арифметичні значення також підлягають нормальному закону розподілу, а дисперсії –  $\chi^2$ -розподілу [9]. З цією метою скористаємося однією з теорем представлених у розділі 1.

*Теорема 3. Нехай маємо дві незалежні випадкові величини  $u$  та  $v$  такі, що  $u$  – підпорядковується нормальному закону, а  $v$  –  $\chi^2$ -розподілу з числом ступенів свободи  $\nu$ .*

*Тоді випадкова величина  $t$*

$$t = \frac{u}{\sqrt{\frac{v}{\nu}}} \quad (4.4)$$

*підпорядковується закону розподілу Стьюдента.*

*Якщо  $u = m_x - \bar{x}$ ,  $v = \sigma_x^2$ ,  $\nu = n - 1$ , то статистику Стьюдента  $t$  можна представити у виді*

$$t = \frac{m_x - \bar{x}}{\sqrt{\frac{\sigma_x^2}{\nu}}}, \quad (4.5)$$

Ймовірність попадання статистики Стьюдента  $t$  у довірчий інтервал (довірчу область, показану на рис.1.2) дорівнює

$$p[|t| < t_{KP}] = \beta = 1 - \alpha, \quad (4.6)$$

де  $\alpha$  – рівень значущості.

Можна записати (4.6) у вигляді

$$p\left[\frac{|x - m_x|}{\sigma_x} \sqrt{n} < t_{KP}\right] = 1 - \alpha. \quad (4.7)$$

Тоді з (4.7) випливає

$$p\left[|x - m_x| < t_{KP} \frac{\sigma_x}{\sqrt{n}}\right] = 1 - \alpha, \quad (4.8)$$

або

$$p\left[\bar{x} - t_{KP} \frac{\sigma_x}{\sqrt{n}} < m_x < \bar{x} + t_{KP} \frac{\sigma_x}{\sqrt{n}}\right] = 1 - \alpha. \quad (4.9)$$

Довірчий інтервал для математичного сподівання може визначатися співвідношенням

$$\bar{x} - t_{KP} \sigma_{\bar{x}} < m_x < \bar{x} + t_{KP} \sigma_{\bar{x}}, \quad (4.10)$$

де

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}. \quad (4.11)$$

Значення  $\sigma_{\bar{x}}$  є середньою квадратичною похибкою визначення середнього арифметичного за даними спостережень.



## 4.2 Визначення довірчих інтервалів для коефіцієнтів рівняння лінійної парної регресії

У гідроекологічних та гідрометеорологічних розрахунках часто використовується рівняння лінійної парної регресії [15]. Побудоване рівняння часто використовується для виявлення лінійного тренду у хронологічних коливаннях досліджуваних характеристик. Коефіцієнти лінійної парної кореляції у таких випадках можуть бути значно нижчими за 1. Для установлення статистичної значущості виділеного лінійного тренду використовують довірчі інтервали та показники значущості для коефіцієнтів рівняння та його коефіцієнту кореляції [16].

Лінійна парна регресія являє рівняння умовного математичного сподівання системи залежних величин  $(X, Y)$ . Для випадку, коли розглядається залежність випадкової величини  $Y$  від  $X$  умовне математичне сподівання  $m_{y/x}$  записується у виді

$$m_{y/x} = m_y + r_{xy} \frac{\sigma_y}{\sigma_x} (x - m_x), \quad (4.12)$$

де  $\sigma_x, \sigma_y$  – середні квадратичні відхилення випадкових величин  $X$  та  $Y$ ;

$r_{xy}$  – коефіцієнт кореляції між випадковими величинами  $X$  та  $Y$ ;

$m_y, m_x$  – безумовні математичні сподівання випадкової величини  $X$  та випадкової величини  $Y$ .

Для вибірових даних рівняння умовного математичного сподівання (4.12) представляється у вигляді

$$\tilde{y}_i = \tilde{y}(x_i) = \hat{m}_{y/x} = ax_i + b, \quad (4.13)$$

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 (4.13);

$a, b$  – параметри рівняння регресії або коефіцієнти регресії.

Параметри  $a$  та  $b$  розраховуються за даними спостережень таким чином

$$a = r_{x,y} \frac{\sigma_y}{\sigma_x}. \quad (4.14)$$

$$b = \bar{y} - a\bar{x}, \quad (4.15)$$

де  $\sigma_x, \sigma_y$  – середні квадратичні відхилення випадкових величин  $X$  та  $Y$ ;

$r_{xy}$  – коефіцієнт кореляції;

$\bar{x}, \bar{y}$  – середні арифметичні значення випадкових величини  $X$  та  $Y$ .

Якщо випадкова величина  $Y$  розподілена нормально, то не тільки оцінки математичного сподівання цієї величини, а й оцінки коефіцієнтів регресії розподілені нормально [17].

Довірчі інтервали коефіцієнтів регресії будуються в такий спосіб

$$a - t_{v,q}\sigma_a < m_a < a + t_{v,q}\sigma_a, \quad \text{або} \quad [a - t_{v,q}\sigma_a; a + t_{v,q}\sigma_a]; \quad (4.16)$$

$$b - t_{v,q}\sigma_b < m_b < b + t_{v,q}\sigma_b, \quad \text{або} \quad [b - t_{v,q}\sigma_b; b + t_{v,q}\sigma_b], \quad (4.17)$$

де  $m_a$  та  $m_b$  – математичні сподівання параметрів  $a$  та  $b$ .

де  $t_{v,q}$  – статистика Стюдента при рівні значущості  $q$  й числі ступенів свободи  $v$  (див. табл. 1.2);

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

$$\sigma_a = \frac{\sigma_y}{n\sigma_x}, \quad (4.18)$$

$$\sigma_b = \frac{\sigma_y}{\sqrt{n}} \left[ 1 + \frac{1}{C_V^2} \right]^{\frac{1}{2}} = \frac{\sigma_y}{\sqrt{n}} \left[ 1 + \frac{\bar{x}^2}{\sigma_x^2} \right] = \frac{\sigma_y}{\sqrt{n}} \frac{(\sigma_x^2 + \bar{x}^2)}{\sigma_x^2}. \quad (4.19)$$

### 4.3 Визначення довірчих інтервалів для коефіцієнта кореляції

Оцінка коефіцієнта кореляції, який відображає тісноту лінійного зв'язку між рядами спостережень, що представляють собою спостережені сукупності випадкових величин  $Y$  та  $X$ , записується у вигляді

$$r_{xy} = r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 (y_i - \bar{y})^2}}, \quad (4.20)$$

або

$$r_{x,y} = r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y (n-1)}. \quad (4.21)$$

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

$$\sigma_r = S_r = \frac{1 - r_{x,y}^2}{\sqrt{n-1}}. \quad (4.22)$$

Довірчий інтервал математичного сподівання коефіцієнта кореляції записується у вигляді

$$r_{x,y} - t_{v,q} \frac{1 - r_{xy}^2}{\sqrt{n-1}} < m_r < r_{x,y} + t_{v,q} \frac{1 - r_{xy}^2}{\sqrt{n-1}}. \quad (4.23)$$

Якщо ж довжина ряду невелика ( $n < 30 \dots 40$ ), то розподіл вибірових значень  $r_{xy}$  істотно відрізняється від нормального. Нелінійне логарифмічне перетворення Фішера при невеликих  $n$  має розподіл близький до нормального. Перетворення Фішера виконують за формулою

$$\hat{z} = \frac{1}{2} \ln \left[ \frac{1+r_{x,y}}{1-r_{x,y}} \right]. \quad (4.24)$$

Закон розподілу оцінок  $z$  близький до нормального із такими параметрами

$$z = \ln \left[ \frac{1+r_{x,y}}{2(1-r_{x,y})} + \frac{r_{x,y}}{2(n-1)} \right] \quad (4.25)$$

та дисперсією

$$\sigma_z^2 = 1/(n-3). \quad (4.26)$$

Розрахунки довірчого інтервалу для коефіцієнта кореляції виконуються таким чином:

- проводиться перетворення за (4.24);
- визначається середнє квадратичне відхилення

$$\sigma_z = \frac{1}{\sqrt{n-3}}$$

- будується довірчий інтервал для  $z$

$$\hat{z} - t_{v,q} \sigma_z < z < \hat{z} + t_{v,q} \sigma_z, \quad (4.27)$$

тобто

$$z_1 < z < z_2 ;$$

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

$$r_1 = r_{x,y1} = \frac{e^{2z_1} - 1}{e^{2z_1} + 1}, \quad (4.28)$$

$$r_2 = r_{x,y2} = \frac{e^{2z_2} - 1}{e^{2z_2} + 1}, \quad (4.29)$$

границі довірчого інтервалу коефіцієнта кореляції:

$$r_1 < r_{xy} < r_2. \quad (4.30)$$

#### 4.4 Перевірка гіпотези про статистичну значущість коефіцієнта кореляції і коефіцієнтів рівняння лінійної парної регресії

Після отримання за даними спостережень точкових оцінок моделі лінійної парної регресії та визначення довірчих інтервалів постає питання про вірогідність статистичного зв'язку між випадковими величинами. Вирішення такого питання за допомогою статистичних методів називають оцінкою статистичної значущості параметрів і це рішення зв'язане з перевіркою статистичних гіпотез.

Висунемо нульову гіпотезу щодо тісноти розглянутого зв'язку

$$H_0 : r_{xy} = 0, \quad (4.31)$$

тобто коефіцієнт кореляції є статистично незначущим (не відрізняється від нуля).

Альтернативна гіпотеза є такою

$$H_1 : r_{xy} \neq 0, \quad (4.32)$$

тобто коефіцієнт кореляції є статистично значущим (відрізняється від нуля).

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

$$t = \frac{r_{xy}}{\sqrt{\sigma_{r_{xy}}^2}}, \quad (4.33)$$

щільність ймовірності якої підлягає розподілу Стьюдента (див. Теорему 3 з розділу 1).

Значення статистики Стьюдента  $t$  визначається за вибірковими оцінками коефіцієнта кореляції  $r_{x,y}$  і його стандарту  $\sqrt{\sigma_r^2} = \sigma_r$

$$t = \frac{|r_{xy}|}{\sigma_r} \quad (4.34)$$

і порівнюється з критичним  $t_{кр}$ , котре залежить від числа ступенів свободи  $\nu = n - 1$  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  $n$  оцінка статистичної значущості  $r_{x,y}$  виконується за допомогою  $z$ -перетворення Фішера, тобто оцінюється не величина  $r_{x,y}$  безпосередньо, а статистика  $\hat{z}$ .

$$t = \frac{|\hat{z}|}{\sigma_z}. \quad (4.35)$$

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  $b$  приймається нормальним. Тоді у відповідності до теорем математичної статистики для перевірки гіпотези про статистичну значущість параметрів  $a$  і  $b$  можна використовувати статистику  $t$ , яка підлягає розподілу Стьюдента

$$t_a = \frac{|a|}{\sigma_a} \quad \text{і} \quad t_b = \frac{|b|}{\sigma_b}. \quad (4.36)$$

Якщо  $t > t_{кр}(v, q)$ , то коефіцієнти регресії  $a$  і  $b$  вважаються статистично значущими.

Для наближеної оцінки значущості коефіцієнтів рівняння лінійної парної регресії приймається правило, що оцінка статистичного параметра буде статистично значущою, коли вона перевищує вдвічі її середнє квадратичне відхилення

$$r \geq 2\sigma_r \quad a \geq 2\sigma_a \quad b \geq 2\sigma_b. \quad (4.37)$$

### Контрольні запитання:

1. Як співвідносяться рівень значущості і довірча ймовірність?
2. Записати довірчий інтервал для математичного сподівання випадкової величини.
3. Записати довірчий інтервал для математичного сподівання  $m_a$  коефіцієнта  $a$  рівняння регресії  $y = ax + b$
4. Записати довірчий інтервал регресії для математичного сподівання  $m_b$  коефіцієнта  $b$  в рівнянні  $y = ax + b$
5. Записати довірчий інтервал для коефіцієнта кореляції, який використовується для рядів спостережень довжиною більше 30.
6. Записати довірчий інтервал коефіцієнта кореляції, який використовується для рядів спостережень довжиною менше 30.
7. Якому закону розподілу підлягає статистика Стюдента?
8. Коли нульова гіпотеза про те, що коефіцієнт регресії  $a$  не відрізняється від нуля, відкидається?
9. Коли нульова гіпотеза про те, що коефіцієнт кореляції  $r$  не відрізняється від нуля, відкидається?
10. Коли нульова гіпотеза про те, що коефіцієнт регресії  $b$  не відрізняється від нуля, відкидається?

## РОЗДІЛ 5

### ЗАСТОСУВАННЯ МЕТОДІВ МАТЕМАТИЧНОЇ СТАТИСТИКИ ПРИ ОЦІНКАХ ЕКОЛОГІЧНИХ РИЗИКІВ

#### 5.1 Поняття про забезпеченість випадкової величини та процентиль

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 параметрами.*

Визначити ймовірність 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

$$p_i = \frac{m}{N}, \quad (5.1)$$

де  $p_i$  – ймовірність появи випадкової величини;

$i$  – порядковий номер випробування (експерименту);

$m$  – число випадків, коли значення  $X = x$  спостерігалася у результаті проведення випробування;

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

У природничих науках закон розподілу випадкової величини часто задається у вигляді функції забезпеченості. Емпірична забезпеченість заданого значення  $X = x$  являє собою ймовірність того, це значення буде перевищене



$$P^*(x) = p^*(X \geq x), \quad (5.2)$$

де  $P^*(x)$  – емпірична забезпеченість;

$p^*(X \geq x)$  – емпірична ймовірність того, що випадкова величина  $X$  буде приймати значення більші або рівні заданому  $x$ .

Для розрахунків емпіричної функції забезпеченості необхідно для кожного значення  $x_i$  з вибірки довжиною  $n$  ( $i = 1, 2, \dots, n$ ) визначити кількість випадків, коли випадкова величина  $X$  прийняла значення більше або рівне  $x_i$  і поділити знайдене  $m$  на загальну кількість випробувань  $n$ . З цією метою ряд спостережень розташовується в убутному порядку (ранжується). Емпірична забезпеченість кожного значення розраховується як

$$P^*(x_i) = \frac{m}{n} \quad (5.3)$$

де  $m$  – порядковий номер значення  $x_i$  у ранжованому ряді, який і є кількістю випадків, коли  $X \geq x_i$ .

Формула (5.3) 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 члена вибірки –  $P_n^* = \frac{n}{n} = 1$ . Таким чином, виходить, що поява

значення випадкової величини, яке буде більше або менше тих, що увійшли до однієї вибірки, унеможлиблюється. 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 формули виду

$$P^* = \frac{m}{n+1}, \quad (5.4)$$

де  $m$  – порядковий номер члена ряду ранжованого в убутному порядку.

Поняття процентиль є протилежним поняттю забезпеченості. У теорії ймовірностей процентиль ототожнюється із значенням інтегральної функції розподілу випадкової величини. Процентиль являє собою ймовірність того, що випадкова величина  $X$  буде менше значення  $x$ , тобто

$$P^*(x) = p^*(X < x), \quad (5.5)$$

де  $P^*(x)$  – емпірична процентиль;

$p^*(X < x)$  – емпірична ймовірність того, що випадкова величина  $X$  буде приймати значення менші заданого  $x$ .

Для того, щоб визначити процентиль, треба розташовувати ряд не в убутному, а зростаючому порядку, користуючись формулою (5.4), де  $m$  – порядковий номер члена ряду ранжованого у порядку зростання. 90% процентиль відповідає 10% забезпеченості і навпаки, 60% процентиль відповідає 40% забезпеченості і навпаки.

## 5.2 Оцінка ризиків недосягнення екологічних цілей

Згідно Водної Рамкової Директиви (ВРД) ЄС (ВРД ЄС 2000/60/ЄС, 2006; Directive 2000/60/EC, 2000) для кожного з основних річкових басейнів України має бути розроблений план управління, метою якого є досягнення у встановлені строки екологічних цілей – “доброго” екологічного стану масивів поверхневих та підземних вод, а також “доброго” екологічного потенціалу штучних або істотно змінених масивів поверхневих вод.

Оцінка ризиків недосягнення екологічних цілей (доброго екологічного стану водного об’єкту) пов’язана, насамперед, із оцінкою антропогенних навантажень. З метою визначення основних антропогенних впливів на стан поверхневих вод в Україні була затверджена методика, яка розглядає «процес аналізу антропогенних навантажень як «процес оцінки ризику недосягнення екологічних цілей» у відповідності із задачами, поставленими Водною Рамковою Директивою [18]. В цій методиці [19] пропонуються спеціально розроблені критерії для оцінки виникнення ризику недосягнення екологічних цілей для певного водного об’єкту на основі даних про гідроморфологічні зміни, об’єми вилучення вод з річок та водойм, а

також даних про скиди забруднених вод. Методика передбачає надання інформації про точкове та дифузне забруднення поверхневих водотоків водами, які надходять з тваринницьких ферм та з сільськогосподарських полів. Окрім показників антропогенних навантажень у цій методиці рекомендується визначати екологічний ризик по фізико-хімічних показниках, серед яких значна увага приділяється біогенним речовинам. Детальний опис цієї методики та приклад її застосування надається в роботі авторів [20].

В залежності від якісних або кількісних показників антропогенних навантажень для кожного виду розрахунків виділяється 3 категорії наслідків антропогенного впливу: «без ризику»; «можливо під ризиком»; «під ризиком».

Результати оцінки основних антропогенних навантажень та їхніх впливів є основою для розроблення та виконання програми заходів для досягнення екологічних цілей [21].

Критерії оцінки ризику при використанні хімічних та фізико-хімічних показників визначаються за двома категоріями ризику: «під ризиком» та «без ризику». Порогові значення хімічних та фізико-хімічних показників наведені в табл. 5.1. Перевищення порогових значень показників/індикаторів антропогенного навантаження стає підставою до висновку, що розглядуваний масив поверхневих вод підпадає під ризик недосягнення екологічних цілей (для розчиненого кисню – навпаки).

Таблиця 5.1 – Критерії ризику для хімічних та фізико-хімічних показників

Річки	Оксиген* (% насичення)	<i>BCK5</i> ** мг/дм <sup>3</sup>	<i>NH<sub>4</sub></i> ** мг/дм <sup>3</sup>	<i>NH<sub>4</sub></i> *** мг/дм <sup>3</sup>	<i>PO<sub>4</sub></i> *** мг/дм <sup>3</sup>	<i>pH</i> мг/дм <sup>3</sup>
Малі F<2000 км <sup>2</sup>	75	5	0,4	0,15	0,2	6,5-8,5
Середні 2000<F<50000	70	6	0,6	0,2	0,3	
Великі F<50000км <sup>2</sup>	60	7	0,8	0,3	0,4	

**Примітка:** \*10% процентиль – всі сезони, порівняльні умови вимірювання, щонайменше 12 вимірювань; \*\*90% процентиль – всі сезони, репрезентативні умови, щонайменше 12 вимірювань; \*\*\* – середньорічне значення.

На рисунку 5.1 представлена інтегральна крива розподілу концентрацій азоту амонійного. Для малих річок критерієм ризику недосягнення екологічних цілей є критичне значення  $0,4 \text{ мг/дм}^3$ . Це значення показане червоною лінією. Фактичне значення, показане зірочкою, відповідає 90% процентилю і розташоване набагато вище критичного. Якщо  $C_{\Phi} > C_{KP}$ , то це означає існування ризику недосягнення екологічних цілей. Значення концентрацій аміаку, які попадають у зону недосягнення екологічних цілей відокремлені на графіку червоною лінією.

Зазначена методика оцінювання антропогенних навантажень є призначена для первинного аналізу екологічного стану річок України з метою забезпечення досягнення всіма поверхневими водними об'єктами доброго екологічного та хімічного статусу згідно із угодою про асоціацію між Україною та Європейським Союзом. Її недолік полягає у тому, що вона розглядає обмежену кількість хімічних та фізико-хімічних речовин і не включає до себе, наприклад, такі забруднювальні речовини як важкі метали.

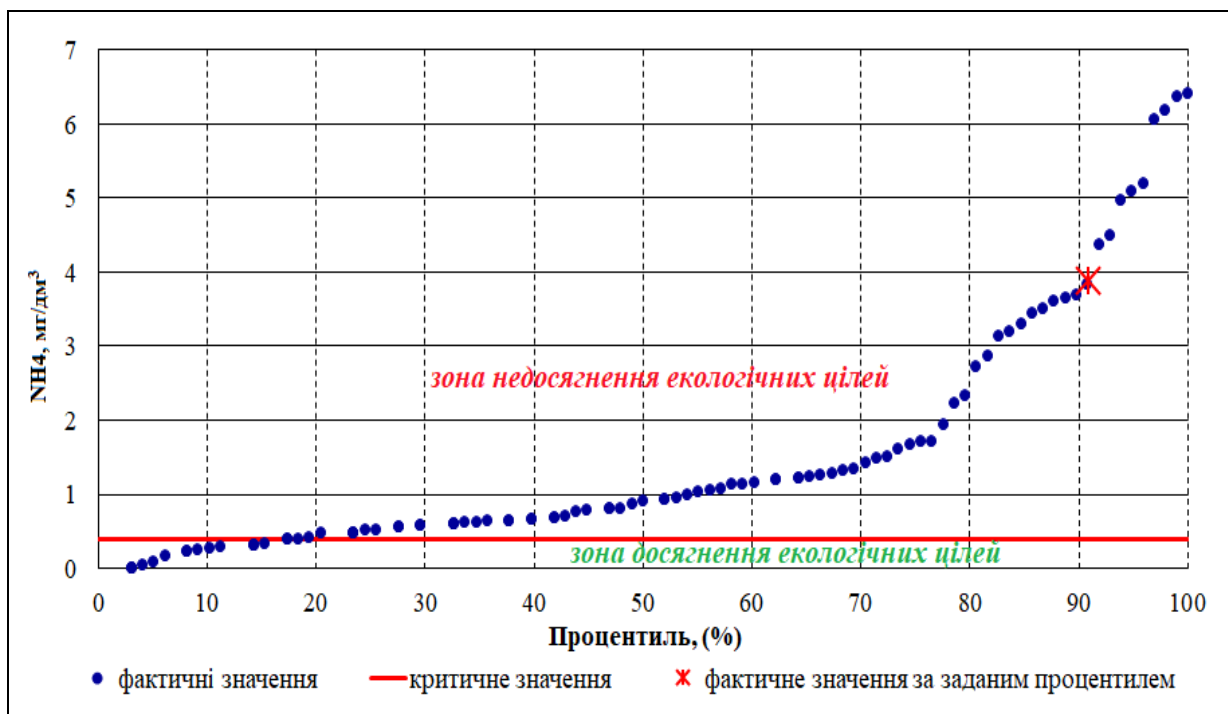


Рисунок 5.1 – Інтегральна крива розподілу показника  $NH_4$  та виділення зони ризику у створі р. Кривий Торець – м. Дружківка за період з 1990 по 2001 роки (червона лінія – критерій ризику  $NH_4$ -  $0,4 \text{ мг/дм}^3$  )

На рисунку 5.2 показане зірочкою фактичне насичення киснем, яке нижче за критичне дорівнює 75% для 10% процентилю (малі річки). Та частина інтегральної кривої, що знаходиться вище критичного (75%) значення, характеризує зону досягнення екологічних цілей, а та що нижче – зону недосягнення. У зоні недосягнення екологічних цілей насиченість киснем нижче допустимої.

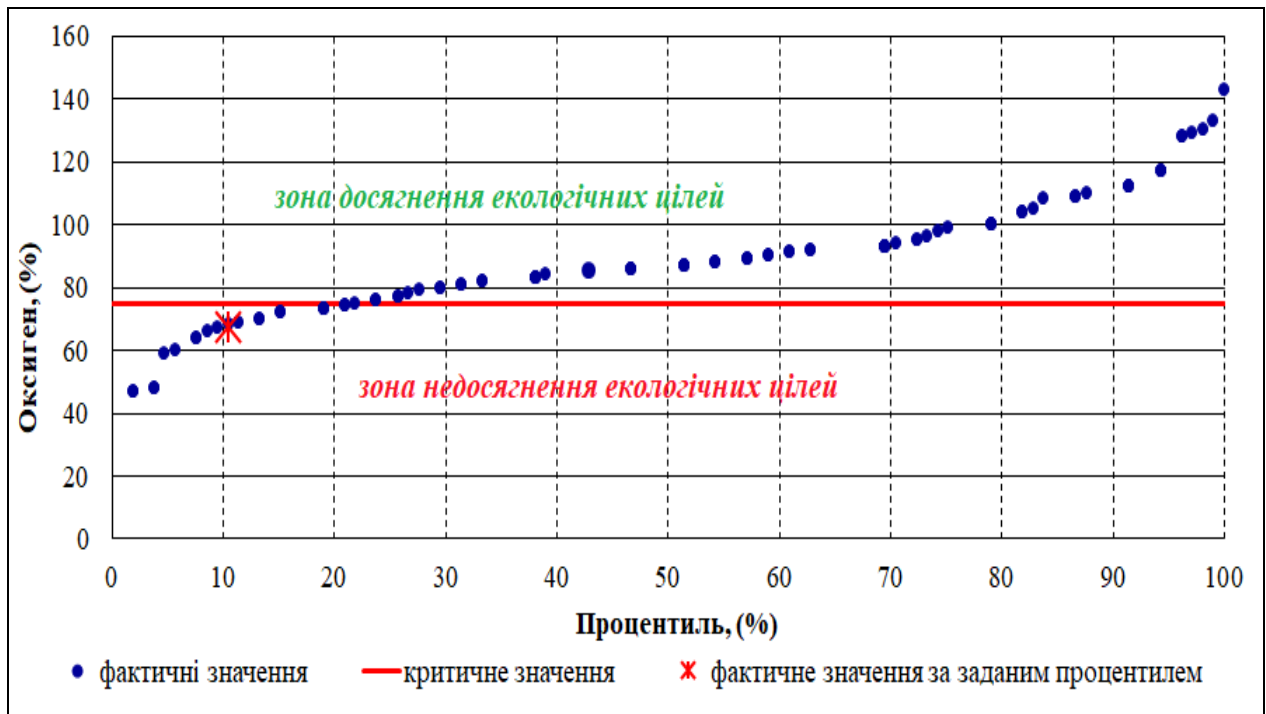


Рисунок 5.2 – Інтегральна крива розподілу показника кисню та виділення зони ризику у створі р. Кривий Торець – м. Дружківка за період з 1990 по 2001 роки (червона лінія відповідає критерію ризику, який дорівнює 75% насичення киснем)

Результати оцінки наслідків забруднення хімічними та фізико-хімічними речовинами представляються, зазвичай, у вигляді таблиці 5.2.

Таблиця 5.2 – Оцінка ризику щодо антропогенного навантаження для хімічних та фізико-хімічних показників за даними моніторингу у створі р. Кривий Торець – м. Дружківка ( $F=1590 \text{ км}^2$ ) за 1990-2001 рр.

Показник	Фактичні значення	Критичні значення	Оцінка ризику
Оксиген, (%насичення) – 10% центиль	67,6	75	«під ризиком»
БСК5, (мг/дм <sup>3</sup> ) – 90% центиль	6,44	5,00	«під ризиком»
$NH_4$ , (мг/дм <sup>3</sup> ) – 90% центиль	3,88	0,40	«під ризиком»
$NH_4$ , (мг/дм <sup>3</sup> ) – середньорічне значення	1,52	0,15	«під ризиком»
$PO_4$ , (мг/дм <sup>3</sup> ) – середньорічне значення	0,35	0,20	«під ризиком»
$pH$ , (мг/дм <sup>3</sup> ) – середньорічне значення	8,13	6,5-8,5	«без ризику»

### 5.3 Визначення екологічних ризиків за стохастичною моделлю

*Ризик* можна розглядати як якісну або кількісну характеристику ситуації, що має невизначеність результату, при обов'язковій наявності несприятливих наслідків [22]. У загальному випадку кількісна оцінка екологічного ризику на базі стохастичної моделі може визначатися як добуток ймовірності виникнення небезпечної екологічної події помноженої на наслідки цієї події. Показником екологічних наслідків забруднення річок може слугувати перевищення фактичної концентрації забруднювальної речовини  $C$  над її граничною допустимою концентрацією СГДК:  $(C/СГДК)$  [23]. Це відношення або сума таких відношень під час забруднення декількома речовинами лежить в основі визначення більшості комплексних індексів забруднення. Таким чином, комплексні індекси забруднення можуть відігравати роль показника екологічного збитку і використовуватись як основа для розрахунків ризику. Перевагою такого підходу є те, що для кожного індексу забруднення існує вже розроблена

градація (шкала оцінки ступеня забруднення), де кожному інтервалу забруднення відповідає своя якісна характеристика забруднення.

При вирішенні задач оцінки ризиків розраховуються показники  $R'$ , які базуються на визначенні співвідношення концентрацій забруднювальної речовини та її ГДК [24]:

$$R' \cong C_i > C_{ГДКi} \quad (5.6)$$

$$R' = C_i / C_{ГДКi} > 1 \quad (5.7)$$

$$R' = C_{ГДКi} / C_i > 1 \quad (5.8)$$

де  $R'$  – кількісний показник ризику;

$C_i$  – рівень концентрації  $i$ -ї забруднювальної речовини;

$C_{ГДКi}$  – гранично допустима концентрація для  $i$ -ї забруднювальної речовини.  $C_{ГДКi}$  призначається в залежності від виду водокористувача.

З урахуванням ймовірності настання ризикової події показник ризику  $R'$  набуває вигляду

$$R' = \sum_{i=1}^n \frac{C_i}{C_{ГДКi}} \frac{N_{as}}{N_i} > 1, \quad (5.9)$$

де  $C_i$  – концентрація  $i$ -тої забруднювальної речовини;

$C_{ГДКi}$  – гранично допустима концентрація  $i$ -тої забруднювальної речовини;

$N_{as}$  – кількість проб з хімічним показником, коли ГДК було перевищене;

$N_i$  – загальне число відібраних проб.

Підчас розрахунків екологічних ризиків аксіоматичним є допущення, що більшість наслідків господарської діяльності, в тому числі й ті, які зумовлюють забруднення навколишнього середовища (зокрема, водного середовища), являють собою випадкові величини і підкоряються статистичного закону розподілу, близькому до нормального. Перевірка

рядів  $R'$  за критерієм Гаусса показує, що у більшості випадків вони можуть розглядатися як такі, що підпорядковуються нормальному закону.

Одним з методів вирішення завдань експертного визначення ризиків є метод побудови шкали кількісного і якісного показника ризику. Для цих цілей використовується метод семантичного диференціала, що дозволяє словами, які відображають різні ситуації щодо якості водного середовища, оцінити стан ризикових подій в даний момент [25]. Шкала градацій складається відповідно до числа Міллера, рівного  $7 \pm 2$  [26].

Градація шкали ризиків може виглядати так: надзвичайно низький, дуже низький, низький, середній, високий, дуже високий, надзвичайно високий. Градація шкали комплексних показників якості води, наприклад, ІЗВ [25] також утримує у собі 7 класів і виглядає таким чином: дуже чиста, чиста, помірно забруднена, забруднена, брудна, дуже брудна, надзвичайно брудна. Таким чином, градація показників ризику та ступеня забруднення води можуть бути узгоджені між собою, що знайшло своє відображення у роботах під керівництвом проф. Лободи Н.С. [13, 27, 28, 29].

Першим етапом у процесі узгодження класифікацій показників якості води та показників ризику є побудова графіку зв'язку між значеннями ІЗВ та значеннями ризиків  $R'$ .

Другий етап полягає у визначенні границь градацій  $R'$ , які відповідають границям ІЗВ за допомогою отриманої залежності  $R' = f(IЗВ)$ .

Третій етап розрахунків передбачає побудову кривої забезпеченості показника  $R'$ . Забезпеченість є ймовірністю перевищення заданого значення випадкової величини. Криві забезпеченості являють собою один із видів математичного опису статистичного закону розподілу випадкової величини.

На четвертому етапі за установленими граничними значеннями  $R$  визначаються відповідні значення їх забезпеченості на основі використання емпіричної кривої ймовірності перевищення заданої випадкової величини.

В залежності від класу якості води установлюються зони ризику, надається якісна оцінка рівня збитку та ймовірність попадання у кожну зону.

Клас якості води з метою наступного семантичного узгодження показників якості та показників ризику установлювався за індексом ІЗВ (табл.5.3).



Таблиця 5.3 – Критерії оцінки якості вод за індексом забруднення води (ІЗВ)

Клас якості води	Характеристика класу	Величина ІЗВ
I	Дуже чиста	$\leq 0,30$
II	Чиста	0,31 – 1,00
III	Помірно забруднена	1,01 – 2,50
IV	Забруднена	2,51 – 4,00
V	Брудна	4,01 – 6,00
VI	Дуже брудна	6,01 – 10,0
VII	Надзвичайно брудна	$> 10,0$

Гідрохімічний індекс забрудненості ІЗВ, введений в дію Держкомгідрометом колишнього СРСР, відноситься до категорії показників, що найчастіше використовуються для оцінки якості водних об'єктів [30]. Він визначається як середнє арифметичне значення перевищення концентрації певних речовин (азот амонійний, азот нітритний, нафтопродукти, феноли, розчинений кисень, БСК):

$$ІЗВ = \frac{1}{6} \sum_{i=1}^n \frac{C_i}{ГДК_i}, \quad (5.10)$$

де  $C_i$  – середня концентрація одного з шести показників якості води;

$ГДК_i$  – гранично допустима концентрація показників якості води (господарське – питне водопостачання), у відповідності із галуззю водопостачання.

За необхідністю розраховується так званий модифікований індекс ІЗВ, в якому використовується також шість показників, а інші беруться за найбільшим відношенням до ГДК.

Приклад графіка зв'язку показника ризику та індексу забруднення води показаний на рис. 5.3. Видно, що показник ризику  $R'$  зростає із збільшенням показника якості вод (ІЗВ). Тобто, чим більше забруднення річки, тим більший показник ризику. Наведена залежність описується через експоненту. Після логарифмування експоненціальна функція

перетворюються на рівняння прямої лінії, тіснота лінійного зв'язку яких оцінюється через коефіцієнт лінійної кореляції  $R$ .

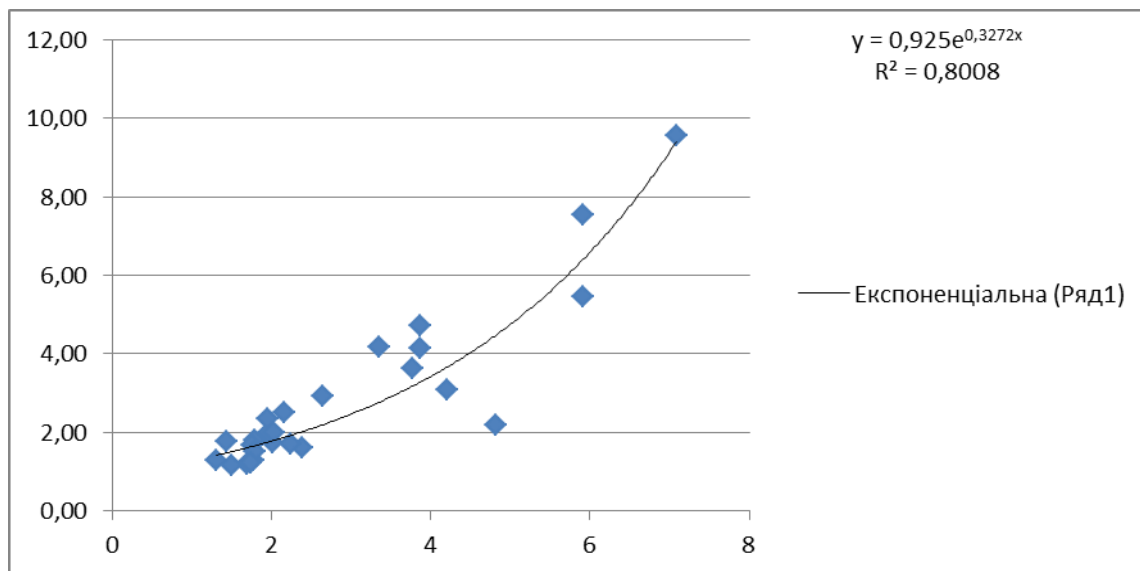


Рисунок 5.3 – Графік зв'язку між ризиком  $R'$  та показником ІЗВ

За залежністю  $R' = f(I3B)$  виконується перехід від границь класів ІЗВ (табл. 5.4) до відповідних границь  $R'$ . Таким чином здійснюється перший та другий етапи семантичного узгодження класифікацій індексів якості води та показників ризику.

На основі емпіричної кривої забезпеченості показників ризику  $R'$  (рис.5.4) установлювався діапазон забезпеченостей, який відповідає кожному інтервалу (градації) ІЗВ. Ймовірність попадання у кожну градацію розраховувалася як  $P_2 - P_1$ , де  $P_2$  – є забезпеченістю величини  $x_2$ , а  $P_1$  – забезпеченість величини  $x_1$ , при чому  $x_2 > x_1$ . Кожній градації ІЗВ відповідає певна градація екологічного ризику. Так, наприклад, градації «помірно забруднені води» для ІЗВ відповідає градація «зона прийняттого ризику» для показника ризику  $R'$ , градації «дуже брудна» – «зона катастрофічного ризику» і так далі.

Таблиця 5.4 – Якісна та кількісна шкала екологічних ризиків забруднення

Клас якості води	Характеристика класу	ІЗВ	Ризик	Діапазон забезпеченостей	Ймовірність попадання у діапазон	Зона ризику
I	Дуже чиста	$\leq 0,30$	$\leq 0,351$	$\leq 94$	6,0	Зона відсутності ризику
II	Чиста	0,31-1,00	0,352-0,472	93,9-89,3	4,6	Зона відсутності ризику
III	Помірно забруднена	1,01-2,50	0,473-0,894	89,2-75,0	14,2	Зона прийнятнього ризику
IV	Забруднена	2,51-4,00	0,895-1,680	74,9-53,9	21,0	Зона допустимого ризику
V	Брудна	4,01-6,00	1,69-3,94	53,8-21,2	32,6	Зона критичного ризику
VI	Дуже брудна	6,01-10,0	3,95-21,5	21,1-0,02	21,1	Зона катастрофічного ризику
VII	Надзвичайно брудна	$> 10,0$	$> 21,5$	0,02	0,02	Зона незворотної втрати якості об'єкта

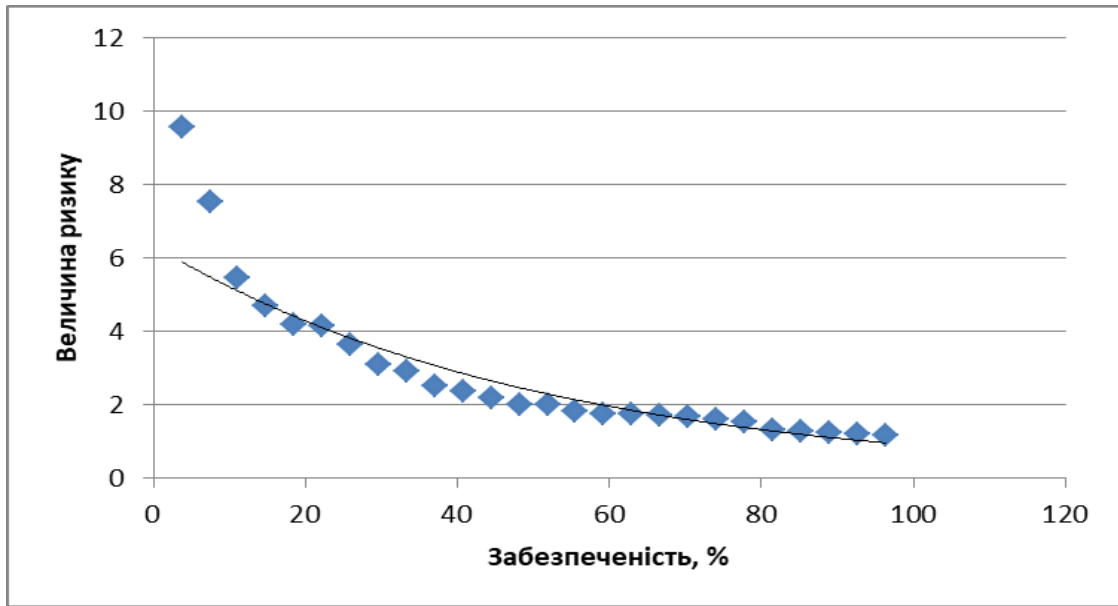


Рисунок 5.4 – Графік емпіричної кривої забезпеченості ризику забруднення для річки Харків

### Контрольні запитання:

1. Що таке екологічний ризик?
2. Дати визначення емпіричної забезпеченості заданого значення  $X = x_i$
3. Дати визначення процентиль заданого значення  $X = x_i$
4. Яким чином розраховується емпірична забезпеченість?
5. Яким чином розраховується емпірична процентиль?
6. У якому випадку устанавлюється висновок про існування ризику недосягнення екологічних цілей?
7. Сутність методу семантичного узгодження.

## РОЗДІЛ 6

### ЗАСТОСУВАННЯ ДИСКРИМІНАНТНОГО АНАЛІЗУ ДО ПОБУДОВИ ШКАЛ КЛАСИФІКАЦІЇ ПОКАЗНИКІВ ЯКОСТІ ВОДИ

#### 6.1 Схема побудови розв'язувального правила

*Дискримінантний аналіз є статистичним методом багатовимірною аналізу, який дозволяє визначати статистичну значущість різниці між двома або більше групами об'єктів по декількох змінних одночасно .*

У гідроекології часто виникає задача щодо віднесення того чи іншого об'єкта або спостереження до одного з відомих класів. Такими класами можуть бути класи якості вод, до яких має бути віднесений той чи інший об'єкт. В практиці гідроекологічного прогнозування часто виникає потреба складати прогноз здійснення або нездійснення того чи іншого явища. Наприклад, можна прогнозувати ступінь ризику порушення екологічного стану водного об'єкту в результаті надходження комунальних стічних вод у поверхневий водоток, або прогнозувати реакцію водного об'єкту на забруднення сполуками азоту за допомогою коефіцієнта вразливості [29]

**Прогноз, який надається за принципом: відбудеться явище чи ні, називають альтернативним.** Поставлені перед дискримінантним аналізом задачі називаються задачами прийняття альтернативних рішень, або задачами класифікації. Питання про віднесення об'єкту до тієї чи іншої сукупності (групи) за принципом «так або ні» вирішується шляхом порівнювання ознак, характерних для кожної із розглядуваних сукупностей, із ознаками самого розглядуваного об'єкту [31].

У теорії розпізнавання образів задача класифікації формулюється таким чином: на основі відомостей про окремих представників різних класів навчаючої системи із характерними для них ознаками (предикторами) необхідно знайти вирішальне правило, за яким той чи інший об'єкт може бути віднесеним до одного з класів.

Сформульована задача є типовою задачею розпізнавання образів. Суть її полягає у тому, що, по-перше, необхідно розділити весь простір

образів на два підпростори, у першому з яких явище відбувається, а в другому – ні. По-друге, треба побудувати правило, за допомогою якого можна віднести образ, який підлягає розпізнаванню, до того чи іншого з підпросторів.

Нехай ми маємо множину  $V$  векторів-предикторів (образів), що складають простір зображень  $R_V$ . Припустимо, що цей простір розділяється на два підпростори  $R_{V1}$  і  $R_{V2}$ . У першому з них розташовується множина  $V_1$  образів  $X$ , при яких явище відбувається, а у другому – множина  $V_2$  образів  $X$ , коли явище не відбувається.

Наперед за все треба побудувати поверхню, яка б розділяла підпростори  $R_{V1}$  і  $R_{V2}$ . Наведемо для пояснення прості приклади.

Нехай простір буде двовимірним  $R_V = R_V(x_1, x_2)$  (рис.6.1). На площині  $(x_1, x_2)$  показана крива  $x_1 = f(x_2)$ , яка відділяє підпростір  $R_{V1}$  від підпростору  $R_{V2}$ . У тривимірному просторі  $R_V = R_V(x_1, x_2, x_3)$  (рис.6.2) підпростори  $R_{V1}$  і  $R_{V2}$  розділяє деяка поверхня, рівняння якої має вид  $x_3 = \varphi(x_1, x_2)$ .

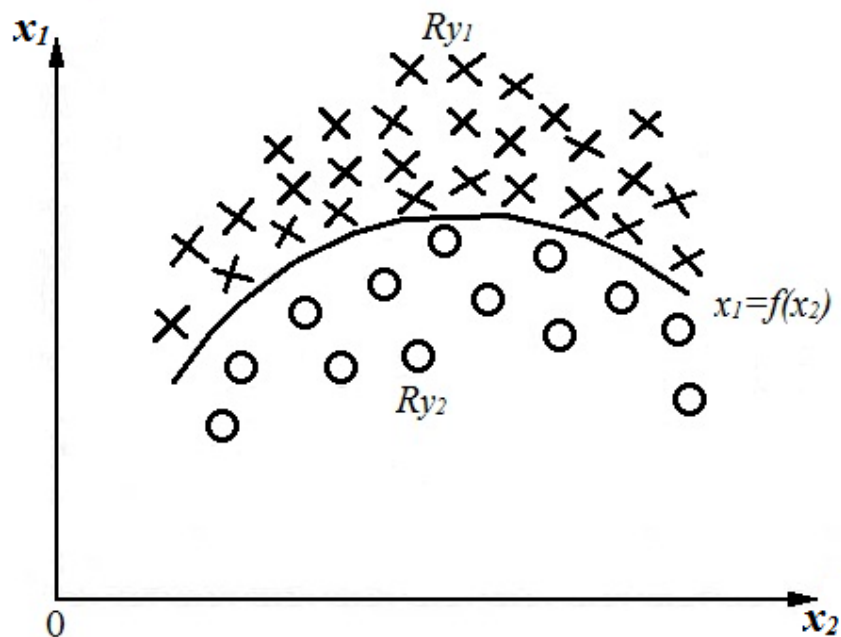


Рисунок 6.1 – Образи і розділяюча функція в двовимірному просторі

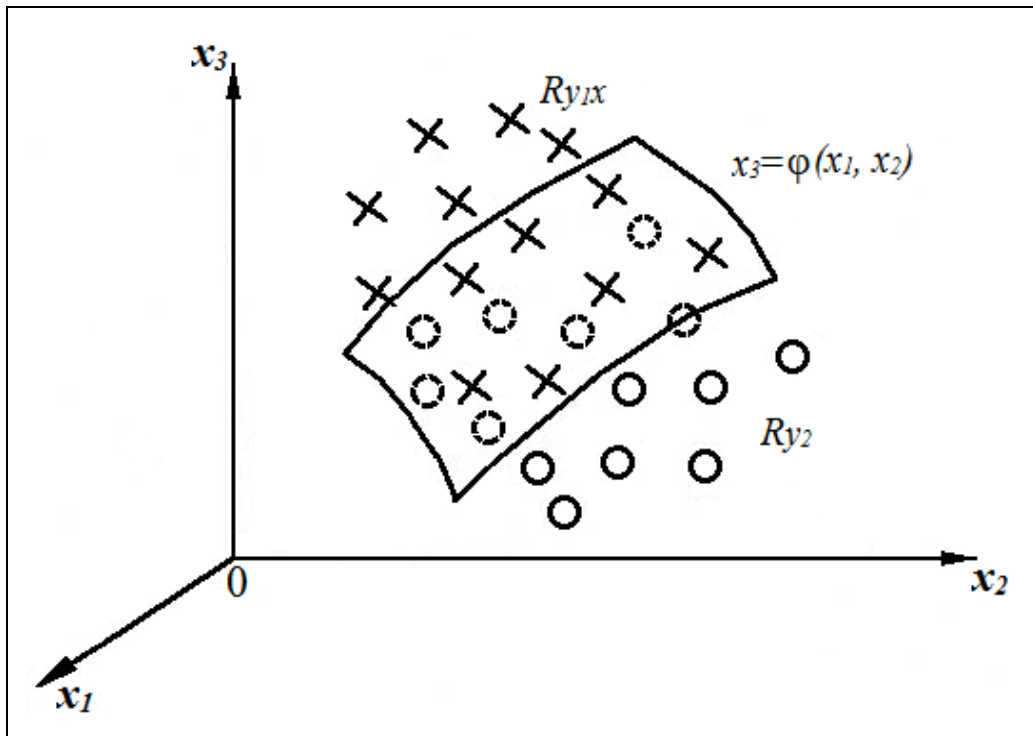


Рисунок 6.2 – Образи та розділяюча поверхня в трьохвимірному просторі

Більш складні умови виникають, коли розглядаються образи із багатовимірному простору  $R_V = R_V(x_1, x_2, \dots, x_n)$ . Поверхня, що розділяє цей простір на підпростори  $R_{V1}$  і  $R_{V2}$  називається розділяючою гіперповерхнею, а її рівняння має вид:

$$F(x_1, x_2, \dots, x_n) = 0. \quad (6.1)$$

Задачею дискримінантного аналізу є отримання правила, за допомогою якого є підстава віднести вектор  $X$ , що підлягає розпізнаванню, до підпростору  $R_{V1}$ , або підпростору  $R_{V2}$ . Це правило називають розв'язувальним правилом. Якщо за розв'язувальним правилом  $X \in R_{V1}$ , то явище прогнозується («так»). Якщо  $X \in R_{V2}$ , то явище не прогнозується («ні»). **Етап, який складається з побудови розділяючої гіперповерхні та розв'язувального правила, носить назву етапу навчання.**

**Прийняття рішення про належність вектора  $X$  до підпросторів  $R_{V1}$  чи  $R_{V2}$  називають етапом розпізнавання.** Множина  $V$  векторів-предикторів, на основі якої реалізуються перелічені етапи, називається

навчаючою сукупністю. Крім неї, створюється ще й *перевірною сукупністю*, яка використовується для перевірки адекватності моделі альтернативного прогнозу. Саме розв'язувальне правило може бути представлене у вигляді деякої математичної функції, яку назовемо дискримінантною.

*Функції, які забезпечують можливість віднесення об'єкта, який підлягає класифікації, до однієї з виділених груп, називають дискримінантними.*

Застосування дискримінантного аналізу, з однієї сторони, передбачає принципову роздільність класів, а з іншої сторони, допускає їхню часткову "перекриваність". Таким чином, рішення про віднесення явища або об'єкту до того чи іншого класу приймається з тією чи іншою долею похибки. **"Перекриваність" класів виражається у тому, що інтервал числових значень ознак у об'єктів, які належать до різних класів, перекривається.** Це утруднює віднесення об'єкта або явища до визначеного класу, якщо їх кількісна ознака лежить у області перекриття.

На рисунку 6.3 показана область перекриття функцій розподілу щільності ймовірностей величини  $X$ . Коли класифікація виконується по одній із ознак, то водний об'єкт за екологічним станом буде віднесений до першого класу, якщо ознака  $x$  буде меншою за  $x_0$  ( $x < x_0$ ), і до другого класу, якщо  $x > x_0$ . Неправильні рішення позначимо через  $\delta_A$  та  $\delta_B$ . У якості  $x_0$  можна вибрати середину інтервалу перекриття  $x_0'$  або абсцису точки перетину кривих щільності розподілу  $x_0''$ . Але більш доцільно вибрати таку точку  $x_0^*$ , для якої буде виконуватись  $\alpha_A = \alpha_B$ . За цієї умови максимальна похибка (найбільша з двох) буде приймати найменше значення.

Таким чином, коли розглядається одна ознака, то задача зводиться до побудови точки, яка розділяє дві сукупності. Результат класифікації визначається відносно цієї точки. Як уже відмічалось, при використанні двох ознак розділ сукупностей відбувається відносно прямої, трьох – відносно поверхні, і т.д.



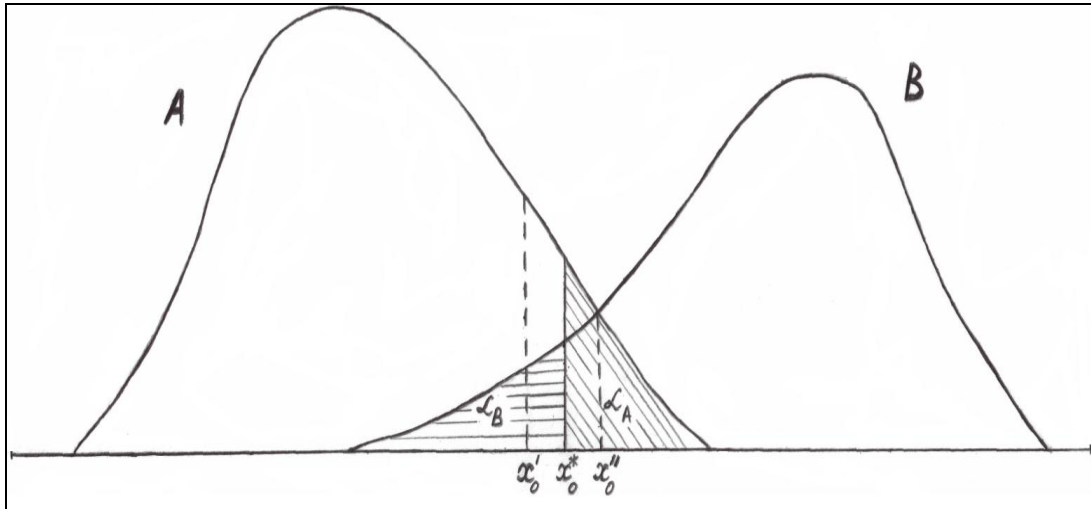


Рисунок 6.3 – Вирішальні правила при розмежуванні  
одновимірних сукупностей

Розглянемо основні ідеї теорії розпізнавання образів [9]. Позначимо через  $H_1$  гіпотезу, що образ  $X \in V_1$ . Альтернативною буде гіпотеза  $H_2$ , про те, що  $X \in V_2$ . Задача розпізнавання полягає у тому, що треба знайти правило, яке дозволяє обґрунтовано прийняти гіпотезу  $H_1$  або  $H_2$ . Всіляка процедура перевірки гіпотез передбачає, що приймаючи те чи інше рішення, ми можемо припустити помилку 1-го чи 2-го роду. Нагадаємо, що помилку 1-го роду ми припускаємо, коли відкидаємо правильну гіпотезу. Помилка 2-го роду пов'язана з прийняттям невірної гіпотези. Помилку другого роду ще називають “похибкою хибної тривоги”.

Будемо вважати, що відомими є умовні ймовірності класів  $V_1$  і  $V_2$ :

$$P(x_1, x_2, \dots, x_n / V_1) \quad (6.2)$$

та

$$P(x_1, x_2, \dots, x_n / V_2) \quad (6.3)$$

Позначимо ймовірність помилки 1-го роду через  $P_a$ , а 2-го роду через  $P_b$ . Знаючи ймовірності (6.2) та (6.3), а також апріорні ймовірності  $P(V_1)$  і  $P(V_2)$  класів  $V_1$  і  $V_2$ , можна розрахувати ймовірності помилок 1-го й 2-го роду.

Розв'язувальне правило або дискримінантна функція будується на основі функції подібності

$$\lambda(x_1, x_2, \dots, x_n) = \frac{P(x_1, x_2, \dots, x_n / V_1)}{P(x_1, x_2, \dots, x_n / V_2)} \quad (6.4)$$

а величину

$$\frac{\delta_b P(V_2)}{\delta_a P(V_1)} = \theta \quad (6.5)$$

називають порогом.

Розв'язувальне правило записується таким чином

$$\text{вектор } X \in V_1, \text{ якщо } \lambda(x_1, x_2, \dots, x_n) > \theta \quad (6.6)$$

$$\text{вектор } X \in V_2, \text{ якщо } \lambda(x_1, x_2, \dots, x_n) < \theta. \quad (6.7)$$

Якщо є підстави вважати, що  $\delta_a = \delta_b$  і  $P(V_1) = P(V_2)$ , то  $\theta = 1$  й розв'язувальне правило приймає вигляд:

$$\text{вектор } X \in V_1, \text{ якщо функція подібності } \lambda(x_1, x_2, \dots, x_n) > 1, \quad (6.8)$$

$$\text{вектор } X \in V_2, \text{ якщо функція подібності } \lambda(x_1, x_2, \dots, x_n) < 1.$$

Ураховуючи (6.5), можна зазначити, що розв'язувальне правило базується на нерівності

$$\frac{P(x_1, x_2, \dots, x_n / V_1)}{P(x_1, x_2, \dots, x_n / V_2)} > \frac{\delta_b P(V_2)}{\delta_a P(V_1)}, \quad (6.9)$$

яка може бути представленою у логарифмічному вигляді

$$\ln \frac{P(x_1, x_2, \dots, x_n / V_1)}{P(x_1, x_2, \dots, x_n / V_2)} > \ln \frac{\delta_b P(V_2)}{\delta_a P(V_1)} \quad (6.10)$$

Функцію виду

$$F(x_1, x_2, \dots, x_n) = \ln P(x_1, x_2, \dots, x_n / V_1) - \ln P(x_1, x_2, \dots, x_n / V_2) + \frac{\delta_a P(V_1)}{\delta_b P(V_2)} \quad (6.11)$$

називають дискримінантною функцією.

При перенесенні порогу у ліву частину (6.11) розв'язувальне правило приймає вид:

$$X \in V_1, \text{ якщо } F(x_1, x_2, \dots, x_n) > 0; \quad (6.12)$$

$$X \in V_2, \text{ якщо } F(x_1, x_2, \dots, x_n) < 0; \quad (6.13)$$

Ясно, що рівняння  $F(x_1, x_2, \dots, x_n) = 0$  є рівнянням розділюючої поверхні для підпросторів  $R_{V_1}$  і  $R_{V_2}$ .

Методи, що основані на теорії статистичних рішень, мають такі обмеження: для їх реалізації необхідно знати щільності умовних розподілів образів у класах  $V_1$  і  $V_2$ . Ці закони розподілів є багатовимірними, і на основі множин векторів-предикторів класів  $V_1$  і  $V_2$ , отримання їх аналітичною виду є дуже складною задачею. Тому вважають, що вид законів розподілу є відомим. У такому разі задача зводиться до необхідності на основі вибірок векторів-предикторів отримати оцінки параметрів цих законів. Ця процедура носить назву відновлення закону розподілу. При практичних реалізаціях цих методів найбільш часто приймають, що класи векторів-предикторів підпорядковуються умовним нормальним законам розподілу. У дійсності ці припущення строго не виконуються. Дуже добре відомо, що для багатьох метеорологічних величин, які виступають у ролі предикторів, нормальний закон розподілу не виконується. Але, як показує досвід, це не вносить суттєвих похибок, якщо більшість предикторів має одномодальний розподіл. Ця умова у більшості випадків виконується [9].

## 6.2 Побудова розв'язувального правила. Квадратична, лінійна та спрощена дискримінантна функція

Будемо вважати, що вектори-предиктори

$$X_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \dots \\ x_{nj} \end{pmatrix} \quad (6.14)$$

підпорядковуються багатовимірному нормальному закону розподілу. Параметрами його, як відомо, є вектори математичних сподівань

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_n \end{pmatrix} \quad (6.15)$$

і матриці коваріацій розміром  $n \times n$ . Тому щільності нормальних умовних розподілів для класів  $V_1$  і  $V_2$  мають вигляд:

$$P(x_1, x_2, \dots, x_n / V_1) = \frac{1}{(2\pi)^{n/2} |K_1|^{1/2}} \exp \left[ -\frac{1}{2} (X - \mu_1)' K_1^{-1} (X - \mu_1) \right] \quad (6.16)$$

$$P(x_1, x_2, \dots, x_n / V_2) = \frac{1}{(2\pi)^{n/2} |K_2|^{1/2}} \exp \left[ -\frac{1}{2} (X - \mu_2)' K_2^{-1} (X - \mu_2) \right] \quad (6.17)$$

де  $\mu_1, \mu_2, K_1, K_2$  – вектори математичних сподівань і матриці коваріацій для першого та другого класів.

Після підстановки (6.16) і (6.17) до дискримінантної функції виду (6.11) прийдемо до такого рівняння:

$$\begin{aligned} F(x) = & -\frac{n}{2} \ln 2\pi - \frac{1}{2} \ln |K_1| - \frac{1}{2} (X - \mu_1)' K_1^{-1} (X - \mu_1) + \\ & + \frac{n}{2} \ln 2\pi + \frac{1}{2} \ln |K_2| + \frac{1}{2} (X - \mu_2)' K_2^{-1} (X - \mu_2) + \ln \frac{P(V_1)}{P(V_2)} \end{aligned} \quad (6.18)$$

Після скорочень дискримінантна функція набуде виду

$$F(x) = \frac{1}{2} \left[ (X - \mu_2)' K_2^{-1} (X - \mu_2) - (X - \mu_1)' K_1^{-1} (X - \mu_1) + \ln \frac{|K_2|}{|K_1|} \right] + \ln \frac{P(V_1)}{P(V_2)} \quad (6.19)$$

Вважається, що ціни помилок першого і другого роду однакові  $\delta_a = \delta_b$ . Дискримінантна функція  $F(x)$ , яка визначається формулою (6.19), носить назву квадратичної дискримінантної функції. Така назва пов'язується з тим, що перші два члени у квадратних дужках являють собою квадратичні форми, тобто многочлени ступеня не більше другого.

Дискримінантна функція (6.20) містить у собі операції обернення коваріаційних матриць  $K_1$  і  $K_2$ . Ця операція може привести до негативних наслідків, коли матриці коваріацій є погано обумовленими. У такому разі похибки, що містяться в коваріаціях, можуть привести до великих похибок розрахунків коефіцієнтів квадратичних форм і, таким чином, до помилок на етапі розпізнавання. У ряді випадків ці похибки можуть бути більшими ніж ті похибки, які ми робимо, приймаючи умову:

$$K_1 = K_2 = K, \quad (6.20)$$

тобто вважаючи, що матриці коваріацій двох класів однакові. Частіше за все приймається умова:

$$K = \frac{K_1 + K_2}{2} \quad (6.21)$$

Якщо прийняти умову (6.21) у дискримінантній функції виду (6.19) і вважати, що  $P(V_1) = P(V_2)$ , то отримаємо:

$$\begin{aligned} F(x) &= \frac{1}{2} \left[ (X - \mu_2)' K^{-1} (X - \mu_2) - (X - \mu_1)' K^{-1} (X - \mu_1) \right] = \\ &= (\mu_1 - \mu_2)' K^{-1} X + \frac{1}{2} \left[ \mu_2' K^{-1} \mu_2 - \mu_1' K^{-1} \mu_1 \right] \end{aligned} \quad (6.22)$$

Дискримінантна функція виду (6.22) є лінійною дискримінантною функцією.

Використання дискримінантного аналізу значно спрощується, якщо є підстави вважати коваріації рівними нулю. Це можна зробити, якщо всі недіагональні елементи матриць коваріацій класів  $V_1$  і  $V_2$  значно менші від діагональних (дисперсій), і ними можна знехтувати.

Тоді

$$K = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix}, \quad (6.23)$$

а її обернена матриця виглядає таким чином

$$K^{-1} = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \dots \frac{1}{\sigma_n^2} \end{pmatrix}, \quad (6.24)$$

тобто операція обернення матриць коваріацій значно спрощується. При цьому спрощується й процедура розрахунків коефіцієнтів квадратичних форм у квадратичній дискримінантній функції (6.11).

Коли виконується умова (6.20) і, крім того, можна вважати коваріаційну матрицю діагональною, значно спрощується вид лінійної дискримінантної функції (6.22). У цьому випадку вона дорівнює:

$$F(x) = \sum_{i=1}^n \frac{\mu_{1i} - \mu_{2i}}{\sigma_i^2} x_i + \sum_{i=1}^n \frac{\mu_{2i}^2 - \mu_{1i}^2}{\sigma_i^2}. \quad (6.25)$$

Рівняння (6.25) є спрощеною лінійною дискримінантною функцією.

Звичайно, при практичному використанні розглянутих дискримінантних функцій замість складових векторів математичних сподівань і дисперсій предикторів використовують їх статистичні оцінки, тобто середні значення і вибірккові дисперсії предикторів.

**Критерієм якості проведення роздільної поверхні є число Махаланобіса, яке характеризує відстань між центрами класів.** При побудові лінійної дискримінантної функції число Махаланобіса визначається за матричним рівнянням виду

$$\Delta = (\mu_1 - \mu_2)' K^{-1} (\mu_1 - \mu_2), \quad (6.26)$$

а при використанні спрощеної лінійної дискримінантної функції (6.25) за таким рівнянням

$$\Delta = \sum_{i=1}^n (\mu_{1i} - \mu_{2i})^2 / \sigma_i^2. \quad (6.27)$$

Чим більше число Махаланобіса, тим менше ймовірність похибки класифікації. При  $\Delta=11$  ймовірність похибки класифікації досягає 5%.

**Прогноз за дискримінантною функцією випускається таким чином:**

*якщо  $F(x_1, x_2, \dots, x_n) \geq 0$  – прогнозується поява досліджуваного явища;*

*якщо  $F(x_1, x_2, \dots, x_n) < 0$  – досліджуване явище не прогнозується.*

Прогнозною датою появи досліджуваного явища  $D'$  є дата, коли:

$$D' = D_{F \geq 0} \quad (6.28)$$

де  $D'$  – прогнозна дата появи досліджуваного явища;

$D_{F \geq 0}$  – дата, на яку дискримінантна функція  $F$  за конкретних умов стає рівною або більшою за нуль.

Оцінка точності прогнозу виконується шляхом розрахунку забезпеченості допустимої похибки прогнозу

$$p = \frac{n - m}{n} \cdot 100\%, \quad (6.29)$$

де  $p$  – забезпеченість прогнозу;

$n$  – загальна кількість перевірочних прогнозів;

$m$  – кількість прогнозів, що не виправдались.

Прогноз вважається «добрим», якщо забезпеченість прогнозу  $\geq 85\%$  і «задовільним», якщо забезпеченість становить 84-60%.

Розглянемо приклад побудови дискримінантної функції при класифікації якості вод за ІЗВ (індексом забруднення води).

Гідрохімічний індекс забруднення ІЗВ відноситься до категорії показників, що найчастіше використовуються для оцінки якості водних об'єктів. Він визначається як середнє арифметичне значення перевищення концентрацій певних речовин (азот амонійний, азот нітритний, нафтопродукти, феноли, розчинений кисень, БСК<sub>5</sub>) над ГДК:

$$ІЗВ = \frac{1}{6} \sum_{i=1}^n \frac{C_i}{ГДК_i}, \quad (6.30)$$

де  $C_i$  – середня концентрація одного з шести показників якості води;

$ГДК_i$  – гранично допустима концентрація показників якості води у відповідності із галуззю водопостачання.

Індекс ІЗВ може бути модифікованим, якщо з шести перелічених компонент залишити БСК<sub>5</sub> та О<sub>2</sub>, а інші показники обрати за найбільшим відношення до ГДК.

За таблицею 6.1 можна установити клас якості води в залежності від значення ІЗВ.

Задача полягає у побудові дискримінантної функції з використанням даних гідрохімічних спостережень за такими показниками: БСК<sub>5</sub>, О<sub>2</sub>, амоній, нітрита, нітрата, фосфати.

На першому етапі необхідно виділити дві навчальні вибірки, які розрізняються між собою ознаками. Виділяємо дві групи, які попадають у різні класи забруднення: I клас – «чисті» ( $0,3 < ІЗВ < 1,0$ ), та II клас – «помірно забруднені» ( $1,0 < ІЗВ < 2,5$ ).

На другому етапі необхідно побудувати спрощену лінійну дискримінантну функцію.

Для побудови функції знаходимо середні значення середньорічних концентрацій забруднювальних речовин за весь період спостережень



(1991-2011 рр.), для I класу та для II класу. Дисперсія приймається як така, що характеризує дві групи у цілому).

Таблиця 6.1 – Критерії оцінки якості вод за ІЗВ

Клас якості води	Характеристика класу	Величина ІЗВ
I	Дуже чиста	$\leq 0,30$
II	Чиста	0,31 – 1,00
III	Помірно забруднена	1,01 – 2,50
IV	Забруднена	2,51 – 4,00
V	Брудна	4,01 – 6,00
VI	Дуже брудна	6,01 – 10,0
VII	Надзвичайно брудна	$> 10,0$

В результаті розрахунків за (6.25) дискримінантна функція набуває виду

$$F(x) = 6,72 \frac{C_{(NO_3)}}{ГДК_{(NO_3)}} - 3,37 \frac{C_{(NO_2)}}{ГДК_{(NO_2)}} - 2,86 \frac{C_{(NH_4)}}{ГДК_{(NH_4)}} - 1,55 \frac{C_{(PO_4)}}{ГДК_{(PO_4)}} - 2,43 \frac{C_{(BCK5)}}{ГДК_{(BCK5)}} - 5,28 \frac{ГДК_{(O_2)}}{C_{(O_2)}} + 39,0 \quad (6.31)$$

На третьому етапі визначаємо число Махаланобіса, яке характеризує відстань між центрами виділених класів, і є характеристикою якості побудови дискримінантної функції

$$\Delta = \sum_{i=1}^n (\mu_{1i} - \mu_{2i})^2 / \sigma_i^2 = 11,9.$$

Оскільки  $\Delta = 11,9$ , що більше 11, то ймовірність похибки класифікації менше 5%. Це означає, що різниця між центрами класів є статистично значущою і отримана класифікація є обґрунтованою.

На четвертому етапі перевіряємо як за даною функцією виконується класифікація, використовуючи дані незалежної вибірки, що не увійшла у

побудову дискримінантної функції. Наприклад, підставляючи дані гідрохімічних спостережень за 2019 рік у вираз (6.31) отримаємо

$$F(x) = -72,5.$$

Оскільки  $F(x) < 0$ , то це означає, що якість води з ознаками ІЗВ належить до II класу – «помірно забруднені» ( $1,0 < \text{ІЗВ} < 2,5$ ), що відповідає дійсності.

Якщо при побудові дискримінантної функції використовується багато ознак (наприклад, концентрацій забруднювальних речовин або будь-яких показників якості води), то можливо видалити з них малозначущі. Ознака визнається малоінформативною, коли її видалення з розгляду приводить до незначного зменшення числа Махаланобіса (не більше ніж  $2m\%$ , при загальній кількості ознак  $m$ ).

Дискримінантний аналіз може бути застосованим до отримання розв'язувальних правил з метою визначення класів якості води. Наприклад, розв'язувальні правила, які відповідають таблиці виду 6.1, можуть бути отриманими шляхом побудови шести дискримінантних функцій з метою послідовного дихотомічного (по дві групи) розділення класів води: дуже чиста, чиста, помірно забруднена, забруднена, брудна, дуже брудна, надзвичайно брудна. Спочатку на основі матеріалів спостережень виділяється дві підмножини: при  $F_1(x) < 0$  – «дуже чиста», при  $F_1(x) > 0$  – множина, яка включає інші класи. Дані, що попали у першу групу у подальшому не використовуються. Далі, при тій же кількості даних, коли  $F_1(x) > 0$ , функція  $F_2(x) < 0$  буде показувати «чиста», а  $F_2(x) > 0$  – усі інші варіанти. За аналогією для  $F_2(x) > 0$  функція  $F_3(x)$  розділить підмножини «помірно забруднена» ( $F_3(x) < 0$ ) та усі варіанти, що залишилися при виконанні умови ( $F_3(x) > 0$ ). Функція  $F_4(x)$  допоможе виділити групи «забруднена» та інші, функція  $F_5(x)$  – «брудна» та інші, функція  $F_6(x)$  – «дуже брудна» та «надзвичайно брудна».

Підчас побудови дискримінантної функції добре виконувати попередню класифікацію за методом факторного аналізу даних.

### **Контрольні запитання:**

1. Який прогноз називається альтернативним?
2. У чому полягає задача розпізнавання образів?
3. З якою метою будується дискримінантна функція?-
4. Перерахувати види дискримінантних функцій.
5. Критерій якості розрахункової або прогностичної методики, побудованої у вигляді дискримінантної функції.

## РОЗДІЛ 7

### ВИКОРИСТАННЯ ВЗАЄМНОЇ КОРЕЛЯЦІЙНОЇ ФУНКЦІЇ ТА ЇЇ ЗАСТОСУВАННЯ ДО РОЗРАХУНКІВ САМООЧИЩЕННЯ ВОДИ НА ДІЛЯНЦІ РУСЛА

*Сукупність усіх процесів, спрямованих на відновлення початкового хімічного складу води відповідно до існуючої раніше рівноваги, називається самоочищенням водного об'єкта.* Більшість забруднювальних речовин є нестійкими і з часом виводяться з розчину під впливом різних процесів, які сприяють самоочищенню. Самоочищення і встановлення біологічної рівноваги відбувається в результаті сукупної дії гідравлічних, фізичних, хімічних і біологічних чинників. У річках процеси самоочищення обумовлені фізичними та хімічними чинниками. У озерах та ставках самоочищення відбувається переважно за допомогою живих організмів. Під фізичними чинниками самоочищення слід розуміти гідравлічні процеси: осідання нерозчинних осадів, вплив ультрафіолетового випромінювання та інше. Серед хімічних чинників зазначають біохімічне окиснення та окиснення неорганічних речовин. Сильно забруднена річка може шляхом самоочищення перейти із стану сапробної зони у мезосапробну і навіть у олігосапробну при відсутності постійного поповнення забруднювальними речовинами [32].

В процесі самоочищення відбувається поліпшення фізичних властивостей води за рахунок адсорбції завислими частками органічних речовин, важких металів, мікроорганізмів, коагуляції та седиментації завислих неорганічних і органічних речовин, мінералізації нестійкої органічної речовини. При самоочищенні вміст кисню має зростати завдяки аерації та дії водної рослинності. Патогенні бактерії будуть при цьому відмирати, наявність сапрофітних мікроорганізмів різко зменшиться. Завдяки самоочищенню невелике забруднення не може змінити природного стану водойми. Але кожна водойма має певну межу самоочисної здатності від забруднень, після якої відбувається різке погіршення всіх характеристик санітарного стану. Процеси самоочищення протікають більш сприятливо завдяки більшій проточності у річках, ніж в озерах і водосховищах. При вивченні процесів самоочищення важливе значення мають співвідношення кількості забруднювальних речовин і

об'єму водної маси, швидкість течії, турбулентності водних мас, глибини, умови вітрового перемішування, температурний режим тощо.

Характеристиками самоочищення є кількісні показники, які визначаються для хімічних та біологічних процесів. При забрудненні річок господарсько-побутовими водами, провідним процесом у самоочищенні є розкладання органічної речовини, кінцевим продуктом якого є мінеральні сполуки. В анаеробних умовах розкладання викликає посилене споживання кисню і корелює з показниками біологічного споживання кисню. Ці показники характеризують ступінь розкладання нестійкої органічної речовини. Константа швидкості розкладання  $K$  залежить від складу забруднювальних речовин і має різні значення. Для господарсько-побутових вод вона становить близько  $0,1 \text{ доб}^{-1}$  і має приблизно таке саме значення при розкладанні фітопланктону. Промислові забруднювальні речовини обумовлюють більші коливання характеристики  $K$ . Швидкості розпаду органічних речовин донних відкладів у 20-50 разів нижчі, ніж господарсько-побутових стічних вод. Для характеристики самоочищення велике значення має швидкість зміни кількості бактерій. Протягом перших 15 годин відмирає 70% від початкової величини бактеріального зараження, а на п'яту добу їх залишається лише частки відсотка.

Процеси самоочищення вивчають на спеціальних пунктах контролю. Ці дослідження виконуються в зоні забруднення річки чи водойми, де у зв'язку з надходженням забруднювальних речовин порушуються природні біохімічні процеси і концентрація забруднювальних речовин за санітарними чи іншими показниками перевищує встановлені норми. Дослідження проводяться у трьох створах (фоновому, головному і замикальному). Фоновий створ повинен бути розташований вище створу скиду забруднених вод. Концентрації забруднювальних речовин у цьому створі не повинні перевищувати гранично допустимі. Додаткові створи встановлюються між головним (контрольним) і замикальним створами. Кількість таких створів залежить від завдань спостережень, місцевих умов та складу забруднювальних речовин. Вони розміщуються нижче випуску забруднених вод з послідовним збільшенням відстані між ними (шість - дев'ять створів). Додаткові створи можуть призначатися за наявності на досліджуваній ділянці бокового припливу. Такі створи встановлюються вище і нижче притоки та в її гирлі. Визначення самоочисної здатності водних об'єктів виконується для специфічних забруднювальних речовин, наявність яких встановлено у воді під час експедиційних досліджень, а

також за такими показниками забруднення води як хімічне споживання кисню, біохімічне споживання кисню. Визначення температури, рН, вмісту розчиненого кисню є обов'язковим. Ці показники характеризують умови і хід процесів самоочищення. Спостереження за забрудненням води виконуються кілька разів на рік у характерні фази гідрологічного і гідробіологічного режимів. Тривалість спостережень визначається необхідністю отримання надійних матеріалів щодо характеристики самоочисної здатності водотоків у роки з різним ступенем водності (багатоводні, маловодні, середні) [33].

Самоочисна здатність води на ділянці обчислюється за таким рівнянням

$$CЗ = \frac{C_B - C_H}{C_B} 100\% , \quad (7.1)$$

де  $C_B$  – концентрація забруднювальної речовини у верхньому створі, мг/дм<sup>3</sup>;

$C_H$  – концентрація забруднювальної речовини у нижньому створі, мг/дм<sup>3</sup>.

Ступінь самоочищення визначають за зниженням концентрації забруднювальної речовини на певному відрізку водотоку при відсутності додаткового забруднення між точками спостережень.

$$Sm = Q(C_B - C_H) , \quad (7.2)$$

де  $C_B$  – концентрація забруднювальної речовини у верхньому створі, мг/дм<sup>3</sup>;

$C_H$  – концентрація забруднювальної речовини у нижньому створі, мг/дм<sup>3</sup>;

$Q$  – витрата води, м<sup>3</sup>/с.

Швидкість самоочищення визначається за зниженням концентрації забруднювальної речовини за одиницю часу

$$Sr = \frac{dC}{dt} = \frac{C_B - C_H}{\tau} , \quad (7.3)$$

де  $C_v$  – концентрація забруднювальної речовини у верхньому створі, моль/дм<sup>3</sup>;

$C_n$  – концентрація забрудної речовини у нижньому створі, моль/дм<sup>3</sup>;

$\tau$  – час добігання між верхнім та нижнім створами.

Під молям розуміють кількість речовини, маса якої, виражена в грамах, чисельно дорівнює її молекулярній масі, вираженій в атомних одиницях маси (а.о.м), що визначається як 1/12 частина маси атома ізотопу <sup>12</sup>C, яка дорівнює 1,66056\*10<sup>-24</sup> г. Моль поширений на будь-які види реальних і умовних частинок. Під реальними частинками розуміють молекули, атоми, іони, електрони, радикали, під умовними – еквіваленти. Відповідно можна говорити про моль молекул, моль іонів, моль електронів.

Щоб розрахувати молярну концентрацію хімічної речовини в розчині необхідно масу речовини в грамах поділити на її атомну масу.

Сумарний коефіцієнт самоочищення визначається за формулою В.Г. Стрітера

$$K_C = \frac{2.3}{\tau} \lg \frac{C_B}{C_H} = \frac{1}{\tau} \ln \frac{C_B}{C_H} , \quad (7.4)$$

де  $C_v$  – концентрація забруднювальної речовини у верхньому створі, мг/дм<sup>3</sup>;

$C_n$  – концентрація забруднювальної речовини у нижньому створі, мг/дм<sup>3</sup>;

$\tau$  – час добігання, доба.

1. **Час добігання – це час, на протязі якого водна маса проходить задану відстань.** Використовується цей термін у декількох значеннях [34].

2. Час, за який маса води з різновіддалених частин басейну досягає замикального створу.

3. Час переміщення об'єму води на заданій ділянці, розраховується за формулою

$$\tau = \frac{\Delta W}{\Delta Q} . \quad (7.4a)$$

4. Час переміщення фазовооднорідних (відповідних) рівнів або витрат води, який визначається як різниця за часом їх настання у верхньому та нижньому створах.

Значення  $K_c$  отримують в установах Державної гідрометеорологічної служби, фонові концентрації розраховують за рівнянням типу  $C = f(Q)$ , а сумарні коефіцієнти самоочищення визначають по завчасно встановленим зв'язкам типу  $K = f(Q)$  або  $K = f(t)$ , або  $\kappa = f(Q, t)$ . За відсутності таких зв'язків розраховують середньоарифметичні значення відповідних коефіцієнтів за багаторічними даними за той місяць, для якого складається прогноз.

Якщо відомі коефіцієнт самоочищення та час добігання від верхнього створу до нижнього, можна установити концентрацію забруднювальної речовини у нижньому створі, із завчасністю яка дорівнює  $\tau$

$$C_H = C_B 10^{-\frac{\tau K_C}{2.3}} = C_B e^{-\tau K_C}, \quad (7.5)$$

де  $C_B$  – концентрація забруднювальної речовини у верхньому створі;

$C_H$  – концентрація забруднювальної речовини у нижньому створі;

$\tau$  – час добігання;

$K_C$  – сумарний коефіцієнт самоочищення (за В.Г.Стріттером).

*Визначення часу добігання між верхнім та нижнім створами можна здійснити на основі розрахунків взаємної кореляційної функції:*

$$r_{xy}(\tau) = \frac{K_{xy}(\tau)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n-\tau} (x_i - \bar{x})(y_{i+\tau} - \bar{y})}{\sigma_x \sigma_y (n - \tau_{зсув} - 1)}, \quad (7.6)$$

де  $r_{xy}(\tau)$  – взаємна кореляційна функція при зсуві у часі  $\tau_{зсув}$ ;

$\bar{x}, \bar{y}$  – середні арифметичні значення рядів  $X, Y$ ;

$\sigma_x, \sigma_y$  – середні квадратичні відхилення двох рядів спостережень довжиною  $n$ .



При цьому ряд спостережень у нижньому створі  $Q_H$  зсувається у часі відносно ряду спостережень у верхньому створі  $Q_B$ , тобто формулу (7.6) можна представити у вигляді:

$$r_{xy}(\tau) = \frac{K_{Q_B Q_H}(\tau)}{\sigma_{Q_B} \sigma_{Q_H}} = \frac{\sum_{i=1}^{n-\tau} (Q_{Bi} - \overline{Q_B})(Q_{Hi+\tau} - \overline{Q_H})}{\sigma_{Q_B} \sigma_{Q_H} (n - \tau_{зсув} - 1)}, \quad (7.7)$$

де  $Q_{Bi}$  – витрати води у верхньому створі в час  $i$ ;  
 $Q_{Hi}$  – витрати води у нижньому створі в часі  $i+\tau$ ;  
 $\sigma_{Q_B}, \sigma_{Q_H}$  – середні квадратичні відхилення для витрат води у верхньому та нижньому створах, відповідно.

Час добігання мас води від верхнього створу до нижнього визначався як такий, при якому  $r_{xy}(\tau)$  досягала свого максимального значення (рис.7.1, рис. 7.2).

Під рядами випадкових величин  $X$  та  $Y$  слід розуміти ряд стоку (витрат води) у верхньому створі  $Q_B$  та нижньому –  $Q_H$ .

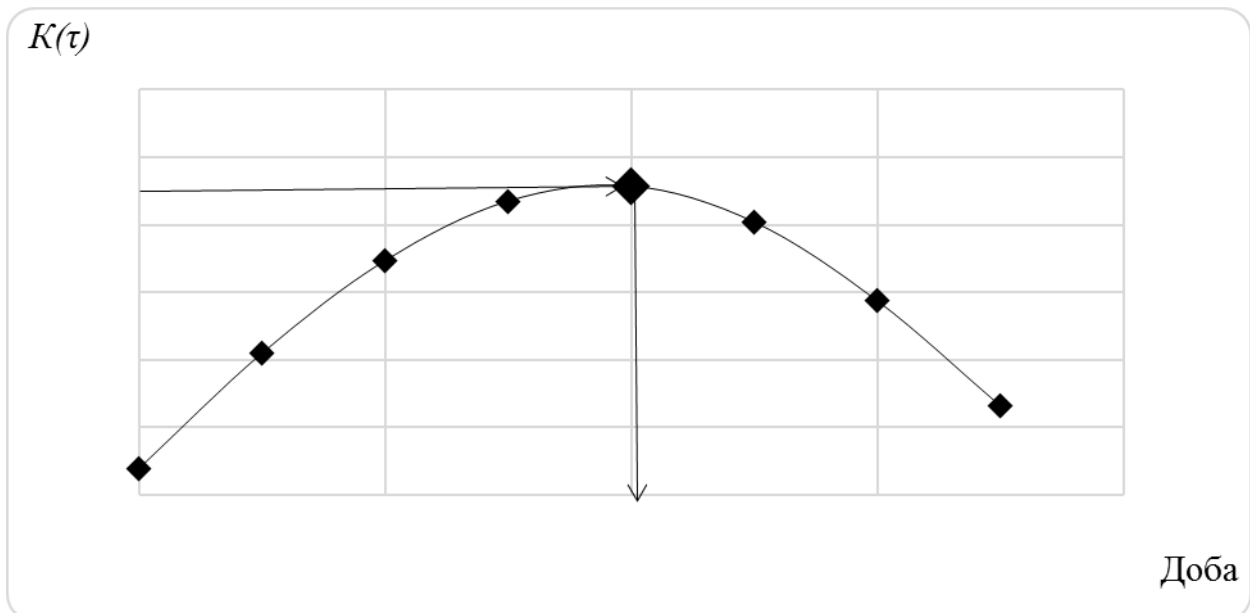


Рисунок 7.1 – Хід взаємної кореляційної функції при зсуві у часі ряду стоку р. Псел – м. Суми відносно ряду стоку р. Псел – с. Крупець [35]

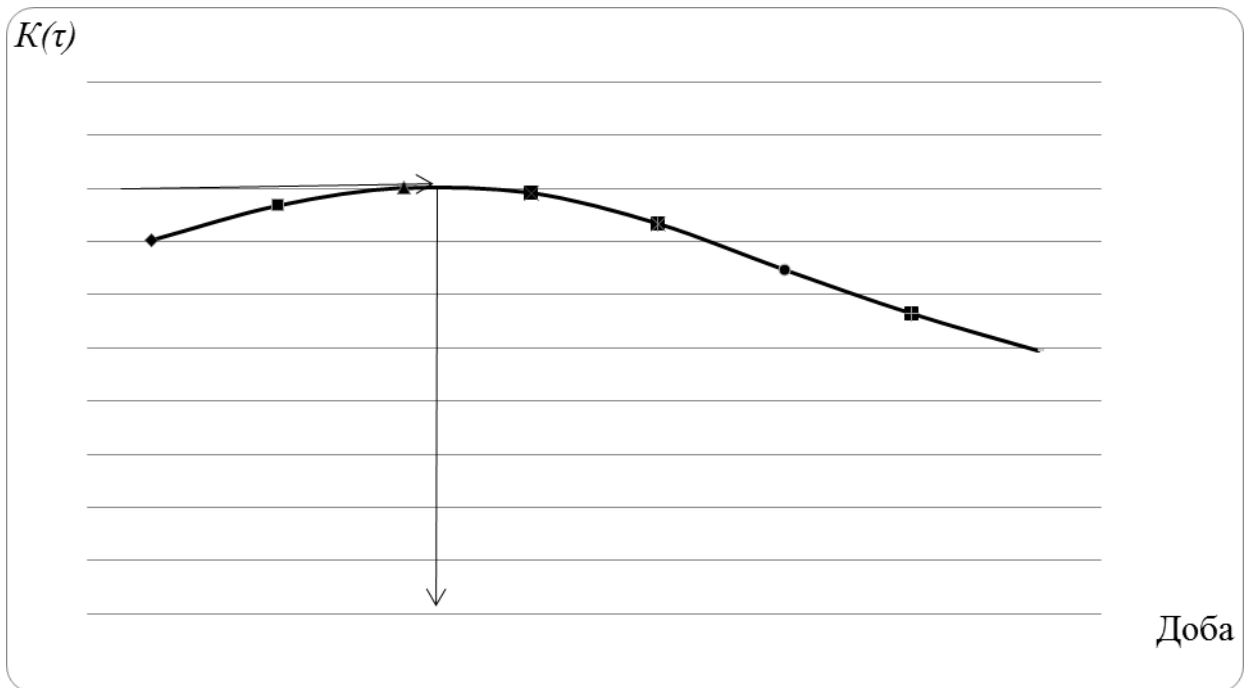


Рисунок 7.2 – Хід взаємної кореляційної функції при зсуві у часі ряду стоку р. Ворскла – с. Чернеччина відносно ряду стоку р. Ворскла – с. Козинка [36]

### Контрольні запитання:

1. Що розуміють під самоочищенням водних об'єктів?
2. Як визначається швидкість самоочищення?
3. Який зв'язок характеризує взаємна кореляційна функція?
4. За якою ознакою визначається час добігання від верхнього створу до нижнього?
5. Як зростання часу добігання з верхнього створу до нижнього впливає на самоочищення води на ділянці річки?

## ВИКОРИСТАННІ ДЖЕРЕЛА

1. Угода про асоціацію між Україною, з однієї сторони, та Європейським Союзом, Європейським співтовариством з атомної енергії і їхніми державами-членами, з іншої сторони. Редакція від 30.11.2015. URL: [https://zakon.rada.gov.ua/laws/show/984\\_011#Text/](https://zakon.rada.gov.ua/laws/show/984_011#Text/) (дата звернення: 23.04.2023)
2. Директива 2000/60/ЄС Європейського Парламенту і Ради "Про встановлення рамок діяльності Співтовариства в галузі водної політики" від 23 жовтня 2000 року. URL: [https://zakon.rada.gov.ua/laws/show/994\\_962#Text](https://zakon.rada.gov.ua/laws/show/994_962#Text) / (дата звернення: 27.04.2023)
3. Директива Ради від 12 грудня 1991 року щодо захисту вод від забруднення, спричиненого нітратами з сільськогосподарських джерел. URL: [https://zakon.rada.gov.ua/laws/show/987\\_002-91#Text](https://zakon.rada.gov.ua/laws/show/987_002-91#Text) / (дата звернення: 28.04.2023)
4. Водний режим та гідроекологічні характеристики Куяльницького лиману / за ред. Лободи Н.С., Гопченка Є.Д. Одеса: ТЕС, 2016. 332 с.
5. Водні ресурси та гідроекологічний стан Тилігульського лиману: колективна монографія / за ред. Ю.С. Тучковенка, Н.С. Лободи. Одеса: ТЕС, 2014. 276 с.
6. Вплив змін клімату на гідрологічний і гідроекологічний режими лиманів північно-західного Причорномор'я: монографія / Тучковенко Ю.С., Хохлов В.М., Лобода Н.С., Кушнір Д.В., Серга Е.М.; за ред. Ю.С. Тучковенка. Одеса: Одеський державний екологічний університет, 2022. 202 с. ISBN 978-966-186-227-1. URL: <http://eprints.library.odku.edu.ua/id/eprint/10929/>
7. Чугай А.В., Сафранов Т.А. Методи оцінки техногенного впливу на довкілля : навч. посіб. Одеса: Видавець Букаєв В. В., 2021. 118 с.
8. Кліматичні зміни та їх вплив на сфери економіки України / Степаненко С.М., Польовий А.М., Лобода Н.С. та ін.; за ред. С.М. Степаненка, А.М. Польового. Одеса: "ТЕС", 2015. 520 с.
9. Школьнік Є.П., Лоева І.Д., Гончарова Л.Д. Обробка та аналіз гідрометеорологічної інформації / Одес. держ. еколог. ун-т. Одеса, 1999. 600с.
10. Kendall M.G. The estimation of parameters sn linear autoregressive time series. *Econometricsa Suppl.*, 1947. 17, 44 p.

11. Методика екологічних оцінки якості поверхневих вод за відповідними категоріями / В.Д. Романенко, В.М. Жукинський, О.П. Оксіюк та ін. Київ: СИМВОЛ-Т, 1998. 28 с.
12. Anatolyev S. Durbin-Watson statistic and random individual effects // *Econometric Theory (Problems and Solutions)*. 2002-2003.
13. Кулачок К.В., Лобода Н.С. Критерії ризику в гідроекології (на прикладі гідрохімічних даних річки Дністер - смт.Біляївка). *Збірник тез за матеріалами студентської наукової конференції молодих вчених ОДЕКУ (06-10 травня 2019р.)*. Одеса: ТЕС, 2019. С. 113-115.
14. Richard H. McCuen. Modeling Hydrologic Change. Statistical methods. *Department of Civil and Environmental Engineering University of Maryland*. Levis Publishers, 2003. 433 p.
15. Лобода Н.С., Катинська І.В. Методи багатовимірної аналізу при вирішенні гідроекологічних задач: конспект лекцій. Одеса: Одеський державний екологічний університет, 2022, 134 с. URL: <http://eprints.library.odeku.edu.ua/id/eprint/10634/>
16. Лобода Н.С., Катинська І.В. Методичні вказівки до практичних занять з дисципліни «Методи багатовимірної аналізу при вирішенні гідроекологічних задач». Одеса: ОДЕКУ, 2023.
17. Школьнік Є.П., Гончарова Л.Д., Миротворська Н.К. Методи обробки та аналізу гідрометеорологічної інформації (збірник задач і вправ). Одес. держ. еколог. ун-т. Одеса: 2000. 419с.
18. Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy // *Official Journal of the European Communities*. 22.12.2000.L. 327, vol. 43. 72 p.
19. Визначення основних антропогенних навантажень та їхніх впливів на стан поверхневих вод : методичні рекомендації / О. Ярошевич, Е. Осійський, М. Скоблей та ін. Київ, 2018. 58 с.
20. Лобода Н.С., Катинська І. В. Визначення антропогенних навантажень та екологічних ризиків в басейні р. Кривий Торець (за програмою підтримки ЄС Водної політики України). *Український гідрометеорологічний журнал*. 2020. №25. С.81-92.
21. Методичні рекомендації щодо визначення основних антропогенних навантажень та їхніх впливів на стан поверхневих вод / Вихрист С., Мудра К., Осійський Е. та ін. Держводагенство, 2018. 21 с.

22. Вікіпедія URL: <https://uk.wikipedia.org/wiki/%D0%A0%D0%B8%D0%B7%D0%B8%D0%BA> / (дата звернення: 03.05.2023)
23. Юрасов С.М., Сафранов Т.А., Чугай А.В. Оцінка якості природних вод: навчальний посібник. Одеса: Екологія, 2012. 168 с.
24. Методичні рекомендації щодо оцінки ймовірності ризикових подій внаслідок забруднення водних об'єктів та ґрунтів української частини Нижньодунайського регіону. Одеса: ФОП Шилов М.В., 2016.
25. Буркинський Б.В., Рубель О.Є. Оцінка ризиків для здоров'я людини та навколишнього середовища від джерел забруднення ґрунту та вод. Звіт "Інвентаризація, оцінка та зменшення впливу антропогенних джерел забруднення в Нижньодунайському регіоні України, Румунії, республіки Молдова, 2007-2013" (MIS ETC CODE 995). НАН України, Інститут проблем ринку та еколого-економічних досліджень. Одеса, 2016. 84 с.
26. Muller G. 1969. Index of geoaccumulation in sediments of the rhine river. *Geojournal*. 1969. 2. P. 108-118.
27. Кулачок К.В., Лобода Н.С. Якісне і кількісне оцінювання екологічних ризиків загального та нітратного забруднення річки Ягорлик (с.Артирівка). *Матеріали міжнародної наукової конференції молодих вчених*. Одеса: ОДЕКУ, 2020. С.85-87
28. Лобода Н.С., Кулачок К.В. Методичні підходи до оцінки екологічних ризиків на базі використання комплексних показників якості води. *Збірник наукових праць VII –й всеукраїнський з'їзд екологів з міжнародною участю (Екологія/Ecology-2019)*. Україна, Вінниця, 25-27 вересня, 2019. С.75.
29. Loboda N., Daus M. Development of a method of assessment of ecological risk of surface water pollution by nitrogen compounds. *Eastern-European Journal of Enterprise Technologies*. 2021. Vol.5. №10 (113): Ecology, P. 15-25. ISSN 1729-3774. <https://doi.org/10.15587/1729-4061.2021.243058>.
30. Хільчевський В.К. Гідрохімічний словник. Київ: ДІА, 2022. 208 с.
31. Лобода Н.С. Методи статистичного аналізу у гідрологічних розрахунках і прогнозах: навчальний посібник. Одеса: Екологія. 2010. 184 с.

32. Осадчий В.І. Процеси формування хімічного складу поверхневих вод. Київ: Ніка Центр, 2013. 240 с.
33. Хільчевський В.К., Осадчий В.І., Курило С.М. Основи гідрохімії. Київ: Ніка-Центр, 2012. 312с.
34. Лобода Н.С. Гідрологічні прогнози: конспект лекцій. Одеса: ТЕС, 2009. 172 с.
35. Loboda N.S., Pilipyuk V.V. Hydrochemical composition of Psyol and Vorskla river waters under conditions of antropogenic influence. *European Applied Sciences*. 2014. № 7. P. 57-60.
36. Loboda N.S., Pilipyuk V.V. Evaluation of ability for natural purification of the Psyol and Vorskla rivers. *International Journal of Research In Earth & Environmental Sciences*. 2015. №1. P. 28-32.

## ПРЕДМЕТНИЙ ПОКАЖЧИК

### А

Альтернативна гіпотеза, 8, 19, 23, 25, 43, 53, 73  
Альтернативний прогноз, 69, 72

### Б

Багатовимірний аналіз, 69  
Біогенні речовини, 59

### В

Взаємна кореляційна функція, 88, 89  
Водна рамкова директива, 58

### Г

Гіпотеза 7  
Гіпотеза статистична 7, 20  
Гіпотеза статистична непараметрична 7  
Гіпотеза статистична параметрична 7

### Д

Дискримінантна функція, 74-82  
Дискримінантний аналіз, 69, 71, 72, 78, 82  
Довірча ймовірність, 11, 47  
Довірча область, 48  
Довірчий інтервал, 46-53  
Довірчий інтервал для математичного сподівання 47, 48  
Довірчий інтервал для коефіцієнта кореляції, 51

### Е

Екологічні ризики, 62, 63, 67  
Ексцес, 40, 41  
Етап навчання, 71  
Етап розпізнавання, 71, 77

### З

Забезпеченість випадкової величини, 56  
Забезпеченість допустимої похибки прогнозу, 80  
Забезпеченість прогнозу, 80  
Закон Гаусса, 33, 41, 42

Закон розподілу  $\chi^2$ , 13, 19, 20, 47  
Закон розподілу нормальний 12, 15, 19, 33, 34, 36, 37, 41, 43, 47, 53  
Закон розподілу випадкової величини, 7, 33, 39, 47, 56, 64  
Закон розподілу Стьюдента, 14, 16, 47  
Закон розподілу Фішера-Снедекора, 13, 18

## Й

Ймовірність появи випадкової величини 56  
Ймовірність перевищення випадкової величини 56

## І

Інверсійний критерій Вілкоксона, 22  
Інверсія, 22, 25, 26  
Індекс забруднення води (ІЗВ), 64, 65, 80

## К

Категорії ризику, 59  
Категорія «без ризику», 59  
Категорія «можливо під ризиком», 59  
Категорія «під ризиком», 59  
Квадратична дискримінантна функція, 77  
Коефіцієнт асиметрії, 34, 39, 41  
Коефіцієнт варіації, 35, 40  
Коефіцієнт самоочищення, 87, 88  
Критерій Аббе, 28, 30  
Критерій випадковості за коефіцієнтом Кендалла, 25  
Критерій Вілкоксона, 22, 23  
Критерій Гаусса, 41, 42  
Критерій Дарбіна-Уотсона, 30, 31  
Критерій Діксона-Томпсона, 43, 51  
Критерій Махалобіса, 79  
Критерій рангової кореляції, 25  
Критерій статистичний, 8, 9  
Критерій Стьюдента, 16, 17, 20, 23  
Критерій Фішера, 19, 21  
Критерій випадковості, 25  
Критерій значущості, 11, 12  
Критерій ризику недосягнення екологічних цілей, 60  
Критерій оцінки якості вод за ІЗВ, 65, 81  
Критерій ризику, 58-60  
Критична область, 8-12, 16, 23



## Л

Лінійна дискримінантна функція, 78, 79, 80  
Лінійна парна регресія, 46, 49, 51, 55  
Лінійний тренд, 49

## М

Метод семантичного диференціала, 64  
Модифікований індекс ІЗВ, 65, 80

## Н

Навчальна вибірка, 73, 80  
Навчаюча сукупність, 72  
Належність «викидів» до генеральної сукупності, 14-16, 43-45  
Непараметричні критерії, 22  
Непараметричні гіпотези, 7  
Нормальний закон розподілу, 12, 15, 19, 33, 34, 36, 37, 41, 43, 47, 53  
Нульова гіпотеза, 7-9, 11, 12, 16-20, 22, 25, 28-30, 41, 43, 45, 53, 54

## О

Оцінка виникнення ризику недосягнення екологічних цілей, 58

## П

Параметричні гіпотези, 7  
Перетворення Фішера, 51, 52, 54  
Помилка 1-го роду, 8, 9  
Помилка 2-го роду, 8, 9  
Порог, 74, 75  
Порогові значення, 59  
Процентіль, 56, 58, 59

## Р

Ранговий критерій Вілкоксона, 23  
Ризик, 16, 17, 41, 42, 59, 62-66, 69  
Ризик екологічний, 16, 56, 59, 62, 63, 67  
Ризик недосягнення екологічних цілей, 58, 60  
Рівень значущості, 9, 11, 12, 16, 17, 21, 23, 29, 31, 46, 48  
Розв'язувальне правило, 74-76, 82  
Розпізнавання образів, 69, 70, 73

## С

Самоочисна здатність водних об'єктів, 84-86  
Самоочищення водного об'єкта, 84-86

Семантичне узгодження, 64, 66

Статистична значущість параметрів рівнянь лінійної регресії, 49, 53-55

## **Т**

Теорія розпізнавання образів, 69, 70, 73, 75

Тренд, 25-29, 49

## **У**

Умовне математичне сподівання, 49, 50

## **Ф**

Функція дискримінантна, 74-82

Функція дискримінантна квадратична 77

Функція дискримінантна лінійна, 78, 79, 80

Функція дискримінантна спрощена, 76, 79, 80

## **Х**

Характеристики самоочищення, 85

## **Ц**

Центровані значення випадкової величини, 35

## **Ч**

Час добігання, 87, 88

Число інверсій, 22, 23

Число Міллера, 64

## **Ш**

Швидкість самоочищення, 86

## **Щ**

Щільність розподілу випадкової величини, 34, 36

## **Я**

Якісний показник ризику, 62, 64

Навчальне електронне видання

**ЛОБОДА Наталія Степанівна**  
**КАТИНСЬКА Ірина Вікторівна**

**ПРИКЛАДНІ АСПЕКТИ ЗАСТОСУВАННЯ МЕТОДІВ МАТЕМАТИЧНОЇ  
СТАТИСТИКИ У ГІДРОЕКОЛОГІЧНИХ ДОСЛІДЖЕННЯХ**

Навчальний посібник

**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](mailto:info@odeku.edu.ua)

Свідоцтво суб'єкта видавничої справи

ДК № 5242 від 08.11.2016