МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
«ДНІПРОВСЬКА ПОЛІТЕХНІКА»

Д.В. Рудаков, О.О. Сдвижкова

МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ ПРИРОДНИЧИХ СИСТЕМ

Навчальний посібник

Дніпро
НТУ «ДП»
2022

УДК 519.248:550.8.053
Р83

Рекомендовано до друку
вченою радою Національного технічного університету «Дніпровська політехніка» як навчальний посібник для студентів вищих навчальних закладів напрямів 10 Природничі науки та 18 Виробництво та технології
(протокол № 6 від 23.03.2021).

Рецензенти:
В.Б. Говоруха – д-р фіз.-мат. наук, професор, зав. кафедри вищої математики та фізики Дніпровського державного аграрно-економічного університету;
Я.Б. Петрівський – д-р техн. наук, професор, проректор з навчально-виховної роботи Рівненського державного гуманітарного університету.

Рудаков Д.В.
Р 83 Математичне моделювання природничих систем: навч. посіб. / Д.В. Рудаков, О.О. Сдвижкова ; М-во освіти і науки України, Нац. техн. ун-т «Дніпровська політехніка». – Дніпро: НТУ «ДП», 2022. – 178 с.

ISBN 978-966-350-762-0

Викладено основи математичних моделей та методів, які необхідні для здійснення кількісних оцінок, аналізу та прогнозу стану об’єктів природокористування. Розглянуто застосування розподілів випадкових величин, кореляційного та дисперсійного аналізів геологічних об’єктів з перевіркою відповідних статистичних гіпотез та методи оптимізації при вирішенні прикладних задач природокористування. Містить інструкції до практичних робіт, які дозволяють краще засвоїти принципи моделювання та статистичного аналізу природничих систем.

Для студентів рівня «магістр» спеціальностей 103 Науки про Землю та 184 Гірництво.

ISBN 978-966-350-762-0 УДК 519.248:550.8.053

ЗМІСТ

ПЕРЕДМОВА

Використання математичних методів та моделей стає невід'ємною складовою сучасних досліджень природничих систем та обробки результатів пов'язаних з ними натурних і експериментальних робіт. Це потребує від сучасних фахівців відповідних ґрунтовних знань та відповідної кваліфікації. Загальним інструментом кількісних оцінок для всіх природничих наук залишаються статистичні методи, які зараз реалізовані в багатьох програмних комплексах, у тому числі Microsoft Excel, що суттєво поширює можливості практичного застосування.

Разом з тим потужні методи статистики та математичного моделювання ще недостатньо використовуються фахівцями у природничих науках, зокрема, через об'єктивні обмеження при відборі проб, проведенні натурних експериментів, доступності для досліджень. Тому стає необхідним викладення математичних моделей та методів з наведенням численних прикладів, які б дозволили майбутнім фахівцям природничого напряму без розширеної математичної підготовки коректно формулювати статистичні гіпотези, проводити їх перевірку, використовувати поняття теорії ймовірності та математичної статистики з розумінням особливостей геологічної інформації для кількісного аналізу результатів геологічних, гідрогеологічних та геофізичних досліджень, а також оцінок впливу на довкілля, прогнозу стану об'єктів природокористування.

Це потребує знань для обґрунтованого вибору розподілів випадкових величин, які описують властивості гірського масиву та процеси у ньому, коректного формулювання та перевірки статистичних гіпотез, застосування кореляційного та дисперсійного аналізів геологічних об'єктів, побудови рівнянь регресії та аналізу їх статистичної значущості. Необхідним для сучасного фахівця в галузі природничих наук стають методи оптимізації, які зараз широко застосовуються при вирішенні прикладних задач природокористування.

Розділи 2.4, 3, 4.1, 4.3 та 5 написані проф. Рудаковим Д.В, розділи 1, 2.1–2.3 та 4.2 – проф. Сдвижковою О.О., розділи 2.4 та 2.5 – спільно обома авторами. Автори будуть вдячні за всі зауваження щодо поліпшення викладеного матеріалу та структури посібника.

ПЕРЕЛІК ОСНОВНИХ ПОЗНАЧЕНЬ

A – асиметрія розподілу випадкової величини;

A^* – статистична оцінка асиметрії;

$D(X)$ – дисперсія розподілу випадкової величини;

D^* – статистична оцінка дисперсії;

E – ексцес розподілу випадкової величини;

E^* – статистична оцінка ексцесу;

$f(x)$ – диференціальна функція або щільність розподілу випадкової величини;

$f^0(x)$ – нормована щільність розподілу за нормальним законом;

F – значення статистики, що перевіряється критерієм Фішера;

$F(X)$ – інтегральна функція розподілу випадкової величини;

$F^0(t)$ – нормована інтегральна функція нормального закону;

H_0 – нульова гіпотеза;

l – довжина, м;

m_i – частота;

$M(X)$ – математичне сподівання випадкової величини x;

n – кількість вимірів або об'єм вибірки, безрозмірна;

P – ймовірність події;

R – індекс кореляції;

r – коефіцієнт кореляції між двома вибірками;

S – оцінка середньоквадратичного відхилення (стандартне відхилення);

S^2 – оцінка дисперсії вибірки;

t – нормована випадкова величина, розподілена за нормальним законом; випадкова величина, розподілена за законом Стьюдента;

x, y, z – декартові координати, м;

x_i, y_i, z_i – можливі значення випадкових величин;

$\overline{X}$ – середнє значення випадкової величини X;

$\overline{X}^*$ – вибіркове середнє;

V^* – коефіцієнт варіації;

w_i – відносна частота;

W – статистика Вілкоксона;

X, Y, Z – випадкові величини;

Z – функція, зворотна функції нормального розподілу;

$\beta_1 = \left(\frac{\mu_3}{\sigma_3} \right)^2$ – нормований показник асиметрії;

$\beta_2 = \frac{\mu_4}{\sigma_4}$ – нормований показник ексцесу;

$\Gamma(\eta)$ – гамма-функція;

Δ – різниця між величинами;

η – відносна варіація;

λ – параметр показникового розподілу;

μ_k – центральний момент *k*-го порядку випадкової величини;

ν – число степенів вільності;

ν_k – початковий момент *k*-го порядку випадкової величини;

$\sigma(X)$ – середньоквадратичне відхилення випадкової величини;

χ^2 – випадкова величина «хі-квадрат»;

$\Phi(t)$ – функція Лапласа;

θ – азимут, град;

ω – частота повторюваності напрямів вітру;

α – рівень значущості;

τ – статистика критерію Граббса – Смирнова;

τ_α – табличне значення критерію Граббса – Смирнова.

Розділ 1
ПОНЯТТЯ МОДЕЛІ ТА МОДЕЛЮВАННЯ. КЛАСИФІКАЦІЯ МОДЕЛЕЙ

Моделюванням називають дослідження різних процесів, явищ, об'єктів на основі створення їх моделей. Модель є відтворенням досліджуваного процесу (явища, об'єкта).

Натурний експеримент, тобто дослідження явища, властивостей та поведінки об'єкта в певних умовах з використанням самого об'єкта, є важливою складовою проектування та управління об'єктами. Однак у багатьох випадках натурне моделювання є економічно недоцільним або неможливим через неготовність самого об'єкта або внаслідок інших причин. *Модель* (від лат. *modulus* – міра, зразок, норма) – це об'єкт-замінник, створений з метою відтворення за певних умов суттєвих властивостей об'єкта-оригіналу.

Основними цілями моделювання є вивчення основних властивостей об'єкта або явища, а також прогнозування поведінки об'єкта-оригіналу в реальних умовах.

Усі моделі можна умовно поділити на матеріальні та логічні. Матеріальна (фізична) модель зазвичай реалізується за допомогою спеціальних технічних пристроїв. Логічні (словесні) моделі являють собою сукупність гіпотез, передумов і рівнянь, тобто вони є абстрактними побудовами. Основні характеристики моделей – цілеспрямованість, скінченність, простота, повнота, адекватність.

Цілеспрямованість моделі. Модель завжди будується з певною метою. Ця мета впливає на те, які властивості об'єкта або явища вважаються істотними, а які – ні. Задача моделювання полягає в тому, щоб для заданого об'єкта підібрати такий опис, який би у найбільш повній мірі відображав оригінал з точки зору заданої мети моделювання.

Скінченність моделі. Зазвичай модель відтворює лише скінченну кількість властивостей та відношень, і через це модель завжди є більш простою, ніж оригінал.

Повнота моделі полягає в тому, що вона має відображати всі істотні, з точки зору мети моделювання, властивості оригіналу.

Адекватність моделі. Необхідною умовою для переходу від дослідження об'єкта до дослідження моделі й подальшого перенесення результатів на об'єкт дослідження є вимога адекватності моделі та об'єкта. Адекватність – це відтворення моделлю з необхідною повнотою всіх властивостей об'єкта, важливих для цілей даного дослідження. Оскільки будь-яка модель простіша за

оригінал, ніколи не можна говорити про абсолютну адекватність, при якій модель за всіма характеристиками відповідає оригіналу.

Моделювання включає такі етапи, як створення, дослідження та використання моделей об’єктів.

Теорія моделювання – розділ науки, що вивчає способи дослідження властивостей об'єктів (оригіналів) на основі заміщення їх іншими об’єктами (моделями).

Одним з найбільш універсальних видів логічного моделювання є *математичне* моделювання, що ставить у відповідність досліджуваному фізичному процесу систему математичних співвідношень, рівнянь, нерівностей, що описують властивості досліджуваного процесу, об'єкта або системи. Розв’язання такої системи дозволяє отримати відповідь на питання про поведінку об'єкта. Метод *математичного моделювання* вдало поєднує в собі основні переваги відомих теоретичних і експериментальних методів дослідження. Робота не з самим об'єктом, а з його моделлю, дає можливість відносно швидко і без істотних витрат досліджувати його властивості та поведінку в будь-яких можливих ситуаціях.

Існує декілька класифікацій моделей. Відповідно до різних критеріїв можна виділити такі основні види моделей:

– динамічні або статичні;

– аналітичні, імітаційні чи комп’ютерні;

– детерміновані або стохастичні (ймовірнісні).

Динамічні моделі відтворюють поведінку об’єктів та процесів, що змінюються у часі, тобто нестаціонарних. *Статичні* моделі описують стаціонарний стан об’єкта чи процесу, тобто стан у деякий момент часу або стан, що практично не змінюється протягом тривалого проміжку часу.

Зокрема, *статичні* моделі в геології описують властивості досліджуваних об’єктів, що не змінюються з часом, за результатами вивчення на основі вибіркових даних з подальшим узагальненням. Прикладами є геологічний розріз або карта, електронні атласи, просторовий розподіл певних властивостей породного масиву, наприклад, вмісту хімічних елементів у рудному тілі, проникності та пористості. Картування передбачає виконання низки проміжних дій, у тому числі інтерполяції за точками спостережень, свердловинами, виділення меж, зон та районів. Картування – це основа для виконання прогнозів з використанням графічних моделей. Побудована карта є просторовим прогнозом певних властивостей у досліджуваній області. За допомогою картування можуть бути оконтурені, наприклад, зсувонебезпечні зони, зони підтоплення, місця виходу рудникових газів на поверхню землі та ін.

Разом з тим усі геологічні об’єкти певним чином трансформуються з часом, у зв’язку з чим виникає необхідність оновлення геологічної інформації та застосування *динамічних* моделей. Прикладом динамічних моделей у геології, тобто моделей, які відтворюють зміни об’єкта протягом часу внаслідок впливу зовнішніх та внутрішніх чинників, є формування родовищ корисних копалин у геологічному часі, осідання ґрунту з часом через навантаження чи його замочування, зміни рівня підземних вод при осушенні чи затопленні кар’єрів та шахт. Спеціальний випадок – моделі з періодичними змінами параметрів протягом року, наприклад, коливання рівня ґрунтових вод та зсуви, активізація яких залежить від сезону.

Дослідження математичних моделей проводять за допомогою аналітичних, чисельних та якісних методів.

Аналітичні моделі відтворюють процеси функціонування об’єкта у вигляді аналітичних залежностей: алгебраїчних, диференціальних, інтегральних рівнянь або їх систем, логічних умов. Прикладом простої аналітичної моделі може бути закон Гука або рівняння вільних коливань.

Аналітичні методи полягають у пошуку явних залежностей між характеристиками. Однак такі залежності можна отримати лише для обмеженої кількості переважно простих моделей.

Чисельні методи дозволяють кількісно досліджувати математичні моделі, для яких застосування аналітичних методів неможливо або недоцільно. Розв’язання диференціальних або інших рівнянь чисельними методами проводиться для конкретних вихідних даних і має додаткову похибку.

Коли явища складні та різноманітні, досліднику доводиться йти на істотні спрощення моделей, у результаті чого аналітична модель стає занадто грубим наближенням до дійсності. Тому в багатьох випадках дослідники змушені використовувати імітаційне моделювання.

Імітаційне моделювання передбачає подання моделі у вигляді алгоритму, реалізованого комп’ютерною програмою, яка дозволяє відтворити поведінку об’єкта. Імітаційні моделі розглядаються як експерименти, що проводяться на комп’ютерах з математичними моделями, які імітують поведінку реальних об'єктів. Вибір на користь імітаційного моделювання визначається найбільш повними можливостями для дослідження об'єкта.

З розвитком ЕОМ значного поширення набуло комп’ютерне структурно-функціональне моделювання. Під комп’ютерною моделлю найчастіше розуміють:

– умовний образ об’єкта, системи або процесу, описаних за допомогою комп’ютерних таблиць, схем, діаграм, графіків, рисунків, анімаційних

фрагментів, що відтворюють структуру та взаємозв'язки між елементами об'єкта чи системи;

– програму чи програмний комплекс, що уможливлює виконання послідовності обчислень з подальшим графічним відображенням їх результатів.

Детермінована модель дає аналітичне уявлення або опис об'єкта, за яким для даної сукупності вхідних даних буде однозначно отримано єдиний результат. Детерміновані моделі відтворюють функціональні зв'язки між аргументом і залежними змінними. Вони записуються у вигляді рівнянь, у яких певному значенню аргументу відповідає тільки одне значення функції (залежної змінної). Детерміновані моделі описують поведінку об'єкта з позицій повної визначеності в сьогоденні й майбутньому, тобто не містять істотної випадковості. Прикладами таких моделей є процеси (явища), які задовільно описуються формулами фізичних законів або співвідношеннями, що являють собою технологічні процеси тощо.

Детерміновані моделі можуть використовуватися для вивчення властивостей геологічних об'єктів як функцій від просторових змінних (координат точки виміру), у яких існують певні закономірності. Детерміновані моделі також використовуються для розрахунків елементів геолого-технічних (техногенно змінених геологічних) систем на основі фундаментальних фізичних законів, коли властивості їх елементів вважаються змінними, але заздалегідь відомими. Прикладами є моделі напружено-деформованого стану породного масиву (схилу, дамби) та/або просідання ґрунту, течій підземних вод, формування родовищ у геологічному часі.

Імовірнісні моделі враховують вплив випадкових факторів на поведінку об'єкта. Якщо об'єкт моделювання стаціонарний і піддається випадковим впливам, то модель є статистичною. *Статистичними* моделями називаються математичні співвідношення та вирази, що містять принаймні одну випадкову компоненту, тобто таку змінну, значення якої не можна передбачити точно для одиничного спостереження. Такі моделі краще враховують невизначеність вихідних даних, помилки та випадкові коливання експериментальних даних.

Стохастичні моделі використовують для моделювання нестаціонарних імовірнісних процесів. Стохастичні моделі визначаються як такі, у яких параметри, умови функціонування і характеристики об'єкта описані випадковими величинами і функціями, вхідна інформація частково чи повністю представлена випадковими величинами. Наприклад, породний масив можна розглядати як середовище з випадково розподіленими властивостями. Зокрема, міцність окремого структурного елементу породного масиву є випадковою величиною, оскільки її формування здійснювалося під впливом багатьох факторів. Тоді неоднорідний породний масив стає системою структурних

елементів, властивості яких змінюються у просторі та підпорядковуються ймовірнісним законам. Стійкість гірничих виробок, що проведені у неоднорідному породному масиві, треба оцінювати з урахуванням ймовірнісної природи міцності гірських порід. Зокрема, розвиток деформацій контуру горизонтальної підземної гірничої виробки у часі та вздовж траси виробки описується випадковими функціями, тобто досліджується на основі стохастичної моделі. Відзначимо, що стохастичні моделі набагато складніше детермінованих.

Класифікація моделей за якоюсь однією ознакою не може охопити всі їх види, бо модель є багатогранною і відображає лише ті властивості системи, які становлять інтерес для дослідника. Завдяки розвитку і впровадженню програмних засобів для роботи з графічними об'єктами, проведення обчислень, обробки різнорідної інформації, різниця між описаними типами моделей у природничих науках стає досить умовною.

Зокрема, прогноз природних явищ або поведінки об'єкта часто грунтується на використанні відомих аналітичних моделей, але з урахуванням стохастичної складової. На практиці часто замість стохастичних моделей застосовують детерміновані, у яких випадкова величина замінюється її середнім значенням або математичним сподіванням. При цьому вплив випадкових факторів досліджують завдяки залученню імітаційного моделювання. Можна також досліджувати результат прогнозу як функцію від випадкової величини з відомим розподілом, що дозволяє оцінювати розкид прогнозних значень, що є більш коректним при порівнянні з даними спостережень. Наприклад, потужність водоносного горизонту та його проникність чи пористість можуть апроксимуватися на основі спостережень як випадкові величини з певними розподілами. Тоді прогнозне положення рівня підземних вод є функцією від цих випадкових величин, а адекватність моделі фактичним даним можна оцінювати за параметрами розподілів.

Оскільки стохастичні моделі відіграють дуже важливу роль у прогнозуванні явищ та поведінки об'єктів, нижче, у розділах 2 і 3, наведені основні принципи створення таких моделей. Першим кроком для цього є збір та обробка даних про явище або об'єкт (підрозділи 2.1, 2.2), на основі чого вивчаються закони розподілу випадкових величин (підрозділи 2.3, 2.4), визначаються ймовірності реалізації тих чи інших випадкових подій у процесі функціонування об'єкта або розвитку природнього явища (підрозділи 2.5, 3).

Контрольні питання до розділу 1

1. Що називають моделлю об’єкта?
2. Які основні цілі моделювання? В чому полягає задача моделювання?
3. Що називають математичним моделюванням?
4. Назвіть основні види моделей.
5. Наведіть приклади статичних та динамічних моделей у геології.
6. Які моделі називають аналітичними?
7. Що називають імітаційним моделюванням?
8. Поясніть різницю між детермінованими і стохастичними моделями.

Література до розділу 1

1. Стеценко І.В. Моделювання систем: навч. посіб. / І.В. Стеценко; М-во освіти і науки України, Черкас. держ. технол. ун-т. – Черкаси : ЧДТУ, 2010. – 399 с.

2. Томашевський В.М. Моделювання систем. / В.М. Томашевський. – Київ: BHV, 2005. – 352 с.

3. Рудаков Д.В. Моделювання в гідрогеології / Д.В. Рудаков; М-во освіти і науки України, Нац. гірн. ун-т. – Дніпропетровськ: НГУ, 2011. – 88 с.

4. Каждан А.Б. Математическое моделирование в геологии и разведке полезных ископаемых: учеб. пособие / А.Б. Каждан, О.И. Гуськов, А.А. Шиманский. – Москва: Недра, 1979. – 168 с.

5. Гавич И.К. Теория и практика применения моделирования в гидрогеологии. / И.К. Гавич. – Москва: Недра, 1980. – 358 с.

Розділ 2
МЕТОДИ ЗБОРУ ІНФОРМАЦІЇ ТА ДАНИХ ПРО СИСТЕМУ. ПОБУДОВА СТАТИСТИЧНИХ МОДЕЛЕЙ

2.1. Випадкова величина та закон її розподілу

Поняття випадкової величини. *Випадковою* називають величину, яка в результаті випробувань набуває будь-якого значення. Заздалегідь воно невідоме і зумовлене випадковими причинами.

Якщо значення випадкової величини можна перелічити, вона називається *дискретною*. Приклад – кількість відбійних молотків, що виходять із ладу протягом зміни. Тут випадкова величина може набувати тільки цілих значень.

Якщо випадкова величина в ході випробування набуває будь-яких значень з деякого інтервалу, вона називається *неперервною*. До неперервних випадкових величин належать: значення вмісту металу в руді за результатами випробування, час безвідмовної роботи механізму, продуктивність праці робітника, похибки вимірювань і т. ін. Майже всі величини, що характеризують властивості гірських порід (межа міцності, модуль деформацій, питома вага), є неперервними випадковими величинами.

Надалі випадкові величини будемо позначати великими літерами (X, Y і т. д.), а їх можливі значення – відповідними малими ($x_1, x_2,\ldots,x_n; y_1, y_2, \ldots, y_n$).

Розглянемо дискретну випадкову величину X з її можливими значеннями $x_1, x_2,\ldots, x_n$. Кожного з цих значень величина X набуває з деякою ймовірністю, а саме: значення $X = x_1$ – з імовірністю p_1; значення $X = x_2$ – з імовірністю p_2, …, значення $X = x_n$ – з імовірністю p_n.

Події «$X=x_2$», «$X=x_2$»,…,«$X=x_n$» несумісні, тобто не можуть реалізуватися в одному випробуванні та вичерпують усі його можливі результати. В алгебрі подій [1] доведено, що такі події утворюють так звану «повну групу» і сума ймовірностей цих подій дорівнює одиниці, тобто

$$\sum_{i=1}^{n} p_i = 1. \qquad (2.1)$$

Ця сумарна ймовірність певним чином розподілена між окремими значеннями випадкової величини X. Випадкова величина буде цілком описана з імовірнісної точки зору, якщо ми будемо точно знати, як саме розподілена сумарна ймовірність, тобто яка ймовірність кожної з подій «$X=x_2$», «$X=x_2$»,…,«$X=x_n$». Відповідність між значеннями випадкової величини і ймовірностями, з якими ці значення з'являються в результаті випробування,

називається *законом розподілу*. Про випадкову величину говорять, що вона підпорядкована даному закону.

Найпростішою формою завдання закону розподілу є таблиця, що містить значення випадкової величини і відповідні їм імовірності (табл. 2.1). Така таблиця називається *рядом розподілу.*

Таблиця 2.1

Ряд розподілу випадкової величини X

X	x_1	x_2	……	x_n
p	p_1	p_2	……	p_n

Приклад 2.1. На руднику протягом 100 робочих змін фіксувалася кількість відбійних молотків, які виходили з ладу з різних причин. Результати спостережень наведено в табл. 2.2, що являє собою ряд розподілу для випадкової величини X – кількості молотків, що вийшли з ладу протягом робочого часу.

Таблиця 2.2

Приклад розподілу в табличному вигляді

Кількість молотків, що вийшли з ладу протягом робочого часу	Кількість спостережень m	Відносні частоти w
0	16	0,16
1	56	0,56
2	24	0,24
3	3	0,03
4	1	0,01
	$\sum_{i=1}^{n} m_i = 100$	$\sum_{i=1}^{n} w_i = 1$

Щоб додати ряду розподілу наочність, будують так звану полігональну криву (рис. 2.1): по осі абсцис відкладають значення випадкової величини, по осі ординат – імовірності цих значень (у даному випадку відносні частоти).

З наведеного прикладу видно, що в умовах даної шахти найбільш імовірним є вихід з ладу одного відбійного молотка протягом робочого часу.

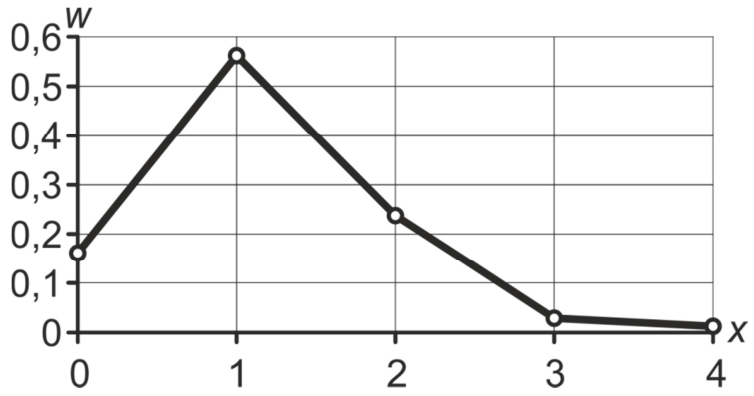


Рис. 2.1. Полігональна крива розподілу випадкової величини
X – кількості відбійних молотків, що вийшли з ладу протягом зміни

Інтегральна функція розподілу. Ряд розподілу вичерпно характеризує дискретну випадкову величину. Однак ця характеристика не є універсальною. Для неперервної випадкової величини побудувати такий ряд неможливо, оскільки не можна перелічити всі її значення. Крім того, як ми переконаємося надалі, ймовірність того, що неперервна випадкова величина набуває якогось конкретного значення *x,* дорівнює нулю. Для неперервної випадкової величини має сенс говорити тільки про ймовірність того, що вона набуває значення з деякого, нехай навіть дуже малого, інтервалу.

Тому для кількісної характеристики розподілу неперервної випадкової величини розглядають не ймовірність події «$X = x$», а ймовірність події «$X < x$». Зрозуміло, що ймовірність цієї події залежить від значення x, тобто є функцією x, що називається *функцією розподілу* випадкової величини і позначається $F(x)$

$$F(x) = P(X < x). \tag{2.2}$$

Вираз $P(X < x)$ означає «ймовірність події «$X < x$». Функція розподілу є універсальною формою закону розподілу. Вона також придатна для опису неперервних і дискретних величин.

Функція розподілу має такі *властивості*:

1. З визначення інтегральної функції $F(x)$ як імовірності виходить, що її значення належать відрізку [0,1]: $0 \le F(x) \le 1$.

2. Інтегральна функція $F(x)$ – неспадна функція свого аргументу, тобто

$$F(x_2) \ge F(x_1), \text{ якщо } x_2 \ge x_1.$$

З останньої формули випливає важливий висновок: імовірність того, що випадкова величина отримає значення з інтервалу (x_1, x_2), дорівнює приросту функції на цьому інтервалі:

$$P(x_1 \le X < x_2) = F(x_2) - F(x_1). \tag{2.3}$$

3. Якщо можливі значення випадкової величини належать інтервалу (a, b), то

1) $F(x) = 0$ при $x \le a$ (ймовірність події «$X \le a$», а, отже, і значення функції в точці $x = a$ дорівнює нулю, оскільки ця подія неможлива);

2) $F(x) = 1$ при $x \ge b$ (ймовірність події «$X \ge b$», а, отже, і значення функції в точці $x = b$ дорівнює одиниці, оскільки ця подія достовірна).

З формули (2.3) випливає, що ймовірність того, що неперервна випадкова величина X набуває одного визначеного значення $x = x_1$, дорівнює нулю.

Цей факт цілком відповідає вимогам практичних задач. Наприклад, важливо визначити ймовірність того, що розміри деталей не виходять за дозволені межі, але не ставиться питання про ймовірність їх збігу з номінальним значенням.

Для дискретної випадкової величини графік інтегральної функції має східчастий вигляд. Наприклад, на рис. 2.2 зображена інтегральна функція, побудована для ряду розподілу, заданого табл. 2.2.

Коли величина X набуває дискретних значень 0, 1, 2, 3, 4, інтегральна функція змінюється стрибкоподібно, причому величини стрибків дорівнюють ймовірностям цих значень (відповідно 0,16; 0,56; 0,24; 0,03; 0,01). Сума всіх стрибків функції $F(x)$ дорівнює одиниці.

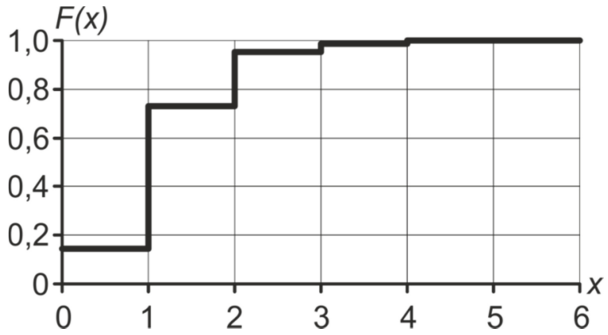


Рис. 2.2. Графік інтегральної функції розподілу для випадкової величини X – числа відбійних молотків, що вийшли з ладу протягом робочого часу

Зі збільшенням можливих значень X кількість стрибків збільшується, а стрибки зменшуються. Східчаста лінія при цьому наближається до плавної кривої. Для неперервної випадкової величини графік інтегральної функції стає неперервною лінією (рис. 2.3).

Зрозуміло, що у випадку, коли можливі значення неперервної випадкової величини заповнюють всю числову вісь, справедливі граничні співвідношення, що випливають із визначення інтегральної функції: $F(-\infty) = 0$; $F(+\infty) = 1$.

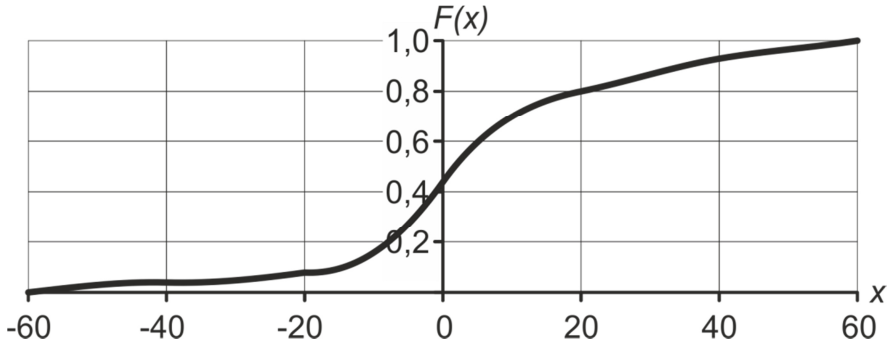


Рис. 2.3. Графік інтегральної функції для неперервної випадкової величини

Приклад 2.2. Випадкова величина X задана функцією розподілу (рис. 2.4)

$$F(x)=\begin{cases} 0 & при \quad x\le 2, \\ 0{,}5x-1 & при \quad 2<x\le 4, \\ 1 & при \quad x<4. \end{cases}$$

Знайти ймовірність того, що в результаті випробувань випадкова величина набуде значення: а) з інтервалу (2,5; 3,0); б) не менше 3.

Розв'язок

а) $P(2{,}5 < X < 3{,}0) = F(3{,}0) - F(2{,}5) = 0{,}5 - 0{,}25 = 0{,}25$;

б) події «$X \ge 3$» та «$X < 3$» є протилежними, тому, використовуючи функцію розподілу, знайдемо спочатку ймовірність того, що $x < 3$:

$P(X < 3) = F(3) = 0{,}5$, тоді $P(X \ge 3) = 1 - P(X < 3) = 1 - 0{,}5 = 0{,}5$.

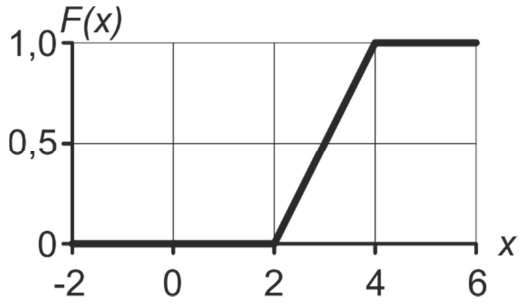


Рис. 2.4. Графік заданої інтегральної функції розподілу (приклад 2.2)

Диференціальна функція розподілу (щільність розподілу). Дослідимо неперервну випадкову величину X з функцією розподілу $F(x)$. Обчислимо ймовірність потрапляння цієї випадкової величини на інтервал від x до $x + \Delta x$:

$$P(x < X < x + \Delta x) = F(x + \Delta x) - F(x). \qquad (2.4)$$

Розглянемо відношення цієї ймовірності до довжини ділянки, тобто середню ймовірність, що приходиться на одиницю довжини цієї ділянки, і будемо наближати Δx до нуля, тобто перейдемо до границі при $\Delta x \to 0$:

$$\lim_{\Delta x \to 0} \frac{F(x+\Delta x) - F(x)}{\Delta x}.$$

Цей вираз є не чим іншим, як похідною функції розподілу $F'(x)$. Її називають *диференціальною функцією розподілу* та позначають $f(x)$. Таким чином,

$$f(x) = F'(x) = \lim_{\Delta x \to 0} \frac{F(x+\Delta x) - F(x)}{\Delta x}. \qquad (2.5)$$

Функція $f(x)$ показує, як щільно розподіляються значення випадкової величини в інтервалі від x до $x+\Delta x$. Звідси походить й інша її назва – *щільність розподілу* випадкової величини. Щільність розподілу, так само як і інтегральна функція розподілу, є однією з форм закону розподілу. Ми побачимо надалі, що вона більш наочна, хоча і не така універсальна. Вона існує тільки для неперервної випадкової величини.

З виразів (2.4) – (2.5) випливає, що ймовірність потрапляння значення випадкової величини на проміжок $(x, x+\Delta x)$ з точністю до нескінченно малих визачається як

$$P(x < X < x + \Delta x) = f(x)dx\,. \qquad (2.6)$$

Величину $f(x)dx$ називають елементом імовірності. Геометрично її можна зобразити площею елементарного прямокутника з висотою, що дорівнює $f(x)$, і основою dx (рис. 2.5). Очевидно, що ймовірність влучення випадкової величини в довільний інтервал від x_1 до x_2 дорівнює сумі елементів імовірності на цій ділянці, тобто інтегралу

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} f(x)dx\,. \qquad (2.7)$$

Геометрично ймовірність влучення випадкової величини на ділянку (x_1, x_2) дорівнює площі криволінійної трапеції, обмеженої кривою $f(x)$ і прямими $x = x_1$, $x = x_2$.

Функція розподілу $F(x)$ для щільності розподілу $f(x)$ є первісною. Тому її називають *інтегральною* функцію розподілу. За визначенням

$$F(x) = P(X < x) = P(-\infty < X < x)\,,$$

тоді за формулою (2.7) отримаємо

$$F(x) = \int_{-\infty}^{x} f(x)dx\,. \qquad (2.8)$$

Геометрично $F(x)$ – не що інше, як площа під кривою щільності розподілу, що знаходиться ліворуч від точки x.

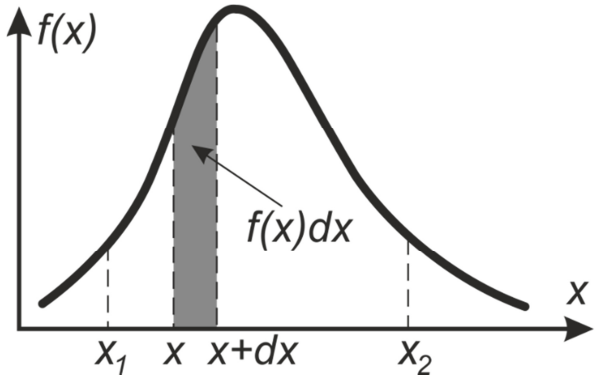


Рис. 2.5. Графік диференціальної функції (щільності розподілу)

Диференціальна $f(x)$ (щільність розподілу) та інтегральна $F(x)$ функції – різні форми закону розподілу випадкової величини X. Знаючи хоча б одну з цих функцій, ми можемо відповісти на найбільш важливе практичне запитання: з якою ймовірністю досліджувана випадкова величина набуває значення з того чи іншого інтервалу. З урахуванням (2.7) і формули Ньютона – Лейбніца ця ймовірність визначається так:

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} f(x)dx = F(x_2) - F(x_1)\,. \qquad (2.9)$$

Далі буде показано, що саме ця формула є основою розв'язання більшості статистичних задач.

При цьому істотне значення мають властивості щільності розподілу:

1. Функція $f(x) \geq 0$; геометрично це означає, що крива розподілу завжди лежить не нижче осі абсцис.

2. Інтеграл від функції $f(x)$ на всьому інтервалі визначення x дорівнює 1:

$$\int_{-\infty}^{+\infty} f(x)dx = 1\,. \qquad (2.10)$$

Дійсно, цей інтеграл є ймовірністю того, що випадкова величина X набуває значення з інтервалу $(-\infty, +\infty)$. Така подія є достовірною, а її ймовірність дорівнює одиниці. Геометрично це означає, що вся площа під кривою розподілу дорівнює одиниці.

Приклад 2.3. Визначимо щільність розподілу для випадкової величини, заданої в попередньому прикладі 2.2 інтегральною функцією:

$$F(x)=\begin{cases} 0 & при \quad x \le 2, \\ 0,5x-1 & при \quad 2 < x \le 4, \\ 1 & при \quad x > 4. \end{cases}$$

Диференціюючи $F(x)$, одержимо:

$$f(x)=F'(x)=\begin{cases} 0 & при \quad x \le 2, \\ 0,5 & при \quad 2 < x \le 4, \\ 0 & при \quad x > 4. \end{cases}$$

Графік отриманої диференціальної функції зображений на рис. 2.6. На всьому інтервалі зміни випадкової величини X щільність розподілу є постійною. Далі буде показано, що такий вид розподілу називається *законом рівномірного розподілу* (рис. 2.6).

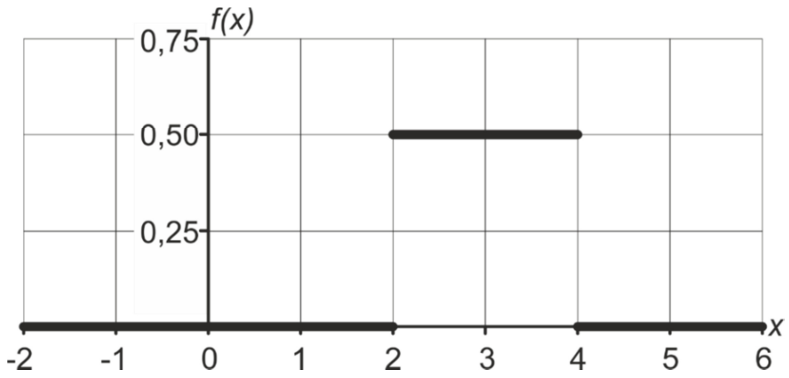


Рис. 2.6. Графік диференціальної функції (щільність рівномірного розподілу)

Контрольні питання до підрозділу 2.1

1. Яку величину називають випадковою?
2. Які випадкові величини називають дискретними і які безперервними? Наведіть приклади.
3. Що називають законом розподілу випадкової величини?
4. Яким чином може бути заданий закон розподілу?
5. Що називають інтегральною функцією розподілу? Назвіть властивості цієї функції. Який вигляд має її графік?
6. Що таке диференціальна функція розподілу? Чому її називають щільністю розподілу?
7. Напишіть, як пов'язана інтегральна функція розподілу з диференціальною функцією.
8. Як визначити ймовірність потрапляння випадкової величини в заданий інтервал? Що являє собою ця ймовірність на графіках інтегральної і диференціальної функцій?

2.2. Побудова статистичного розподілу кількісної ознаки

Вище було зазначено, що прогноз поведінки гірських порід навколо підземної споруди можна здійснювати на підставі аналізу деяких кількісних ознак. Зокрема, такими є фізико-механічні властивості масиву. Внаслідок неоднорідності різного масштабу, що притаманна породному масиву, більшість з кількісних ознак є випадковими величинами. Наприклад, у результаті випробування зразків гірської породи на стиск ми одержуємо різні значення міцності навіть для зразків однієї літологічної різниці, що зумовлено непрогнозованими факторами. Таким чином, дана кількісна ознака – міцність породи на стиск – є випадковою величиною, яка може набувати значення з певного інтервалу. Дослідника, що виконує геомеханічні розрахунки, цікавить, які значення досліджуваної ознаки є найбільш імовірними, як оцінити наявний розкид цих значень, як він вплине на остаточні результати. Для відповіді на ці питання дослідник має підібрати для отриманих даних теоретичний закон розподілу у вигляді інтегральної чи диференціальної функцій розподілу. Першим кроком на шляху підбору теоретичного розподілу є статистична обробка даних, тобто їх *статистичний аналіз.*

При статистичному аналізі процесів гірничого виробництва або геологічних дослідженнях використовують *вибіркові дані*. З обстежуваної *генеральної сукупності* однорідних об'єктів, кількість яких N, відбирають випадково певну кількість об’єктів (n одиниць) **,** тобто формують *вибірку*.

Нехай вивчається деяка кількісна ознака X, що властива генеральній сукупності однорідних об'єктів. Наприклад, досліджується міцність деякої гірської породи на одноосьовий стиск. Зазвичай ця характеристика встановлюється шляхом випробування в лабораторних умовах зразків породи об’ємом V, виготовлених з проби, що у свою чергу відібрана в досліджуваному масиві гірської породи. Таким чином, застосовується вибірковий метод: з генеральної чисельності зразків $N = V/v$ (V – увесь об’єм досліджуваного масиву гірської породи) відбирається тільки n зразків. Величина n називається *об’ємом вибірки*.

Для того, щоб за даними вибірки можна було достатньо впевнено судити про важливі ознаки генеральної сукупності, необхідно, аби об'єкти вибірки правильно її представляли. Коротко ця вимога формулюється так: вибірка повинна бути *репрезентативною*. Вибірка є репрезентативною, якщо її створювати випадково, тобто так, щоб усі об'єкти генеральної сукупності мали рівні шанси потрапити у вибірку.

Нехай у результаті вимірів отримані значення (варіанти) кількісної ознаки X: $x_1, x_2, x_3, \ldots, x_n$. Якщо ці значення записані в тій послідовності, у якій вони

спостерігалися в дійсності, вони називаються *незгрупованими* даними. Якщо таких даних кілька десятків чи сотень, наш розум не в змозі охопити зміст явища, що спостерігається. Для виявлення характерних рис досліджуваної ознаки доцільно згрупувати дані.

Нехай у результаті дослідів кількісної ознаки X значення x_1 з'явилося m_1 раз, x_2 – з'явилося m_2 разів, значення x_r – з'явилося m_r раз. Числа m_i $(i = 1,...,r)$, що показують, скільки разів спостерігається кожне зі значень ознаки, називаються *частотами.* Зрозуміло, що сума частот повинна дорівнювати об'єму вибірки

$$\sum_{i=1}^{r} m_i = n\,. \tag{2.11}$$

Величини $w_i = \dfrac{m_i}{n}$ називають *відносними частотами*. Очевидно, що

$$\sum_{i=1}^{r} w_i = 1\,. \tag{2.12}$$

Якщо розташувати варіанти в порядку, що зростає чи спадає, і зазначити для кожної варіанти її частоту, одержимо розподіл ознаки у вигляді впорядкованого *варіаційного ряду* (табл. 2.3). Як бачимо, варіаційний ряд аналогічний ряду розподілу дискретної випадкової величини, що поданий у табл. 2.1. Раніше наведена табл. 2.2, яка характеризує статистичний розподіл кількості відбійних молотків, які вийшли з ладу за певний період, являє собою приклад варіаційного ряду.

Таблиця 2.3

Загальний вигляд варіаційного ряду

Значення кількісної ознаки X	x_1	x_2	x_3	...	x_n
Частоти m_i	m_1	m_2	m_3	...	m_n
Відносні частоти w_i	w_1	w_2	w_3	...	w_n

Побудова інтервального ряду. Гістограма частот. У табл. 2.4 як приклад наведено фрагмент журналу досліду з визначення межі міцності зразків алевроліту на одноосьовий стиск у вигляді незгрупованих даних. Але, якщо для цих даних побудувати варіаційний ряд, він вийде досить громіздким (табл. 2.5).

Для кількісної ознаки, що являє собою *неперервну випадкову величину*, варто побудувати *інтервальний ряд*. Для цього:

1) серед незгрупованих даних вибирають найбільше та найменше значення x_{max} та x_{min};

2) відрізок (x_{max}, x_{min}) поділяють на інтервали (розряди) довжиною l:

$$l = \frac{x_{max} - x_{min}}{k}, \tag{2.13}$$

де k – кількість інтервалів; це число можна вибрати довільно, а можна скористатися формулою

$$k = 1 + 3{,}2 \cdot \lg n; \tag{2.14}$$

3) визначають частоту потрапляння значень величини X у кожний з інтервалів;

4) обчислюють середину кожного інтервалу і надалі розглядають її як варіанту.

Таблиця 2.4

Фрагмент журналу досліду з визначення межі міцності алевроліту

№ зразка	Міцність зразка на стиск, МПа
1	52,8
2	48,4
3	60,0
…	...
31	56,3
32	49,7
33	56,4
і т. д...	...

Таблиця 2.5

Варіаційний ряд, побудований на основі журналу дослідів

Варіанти x_i, МПа	Частота, m_i	Варіанти x_i, МПа	Частота, m_i	Варіанти x_i, МПа	Частота, m_i
48,4	1	51,2	1	53,4	3
48,9	1	51,4	1	54,3	4
49,2	1	51,8	1	54,5	1
49,3	2	52,2	2	64,6	1
49,7	1	52,4	1	55,4	1
50,3	2	52,7	2	56,4	3
50,8	2	52,8	4	57,2	1
51,0	2	53,0	1	60,0	1

Таким чином, в інтервальному ряді частоти належать не до окремих значень ознаки, а до середин інтервалів. Наприклад, дані, що наведені в табл. 2.5, розбиті на 6 інтервалів довжиною $l = 2$ (табл. 2.6). Наочним

графічним зображенням інтервального ряду є *гістограма частот*. На осі абсцис відкладають відрізки, які дорівнюють довжинам інтервалів. На кожному з відрізків будують прямокутник, висота якого складає:

$$h_i = \frac{w_i}{l_i}, \quad i = 1,...,k\,. \tag{2.15}$$

Тоді площа кожного *i*-го прямокутника дорівнює відносній частоті w_i, а площа всієї гістограми – одиниці (рис. 2.7).

Наприклад, на рис. 2.8 побудована гістограма відносних частот для величини X – міцності на стиск алевроліту (за даними табл. 2.6).

Таблиця 2.6

Інтервальний ряд для кількісної ознаки – межі міцності на стиск

Інтервали міцності зразка на стиск $u_{i-1} - u_i$, МПа	Значення середини інтервалу u_i	Частота m_i	Відносна частота w_i
48 – 50	49	6	0,15
50,1 – 52	51	9	0,225
52,1 – 54	53	13	0,325
54,1 – 56	55	7	0,175
56,1 – 58	57	4	0,1
58,1 – 60	59	1	0,025
		$\sum_{i=1}^{r} m_i = 40$	$\sum_{i=1}^{r} w_i = 1{,}0$

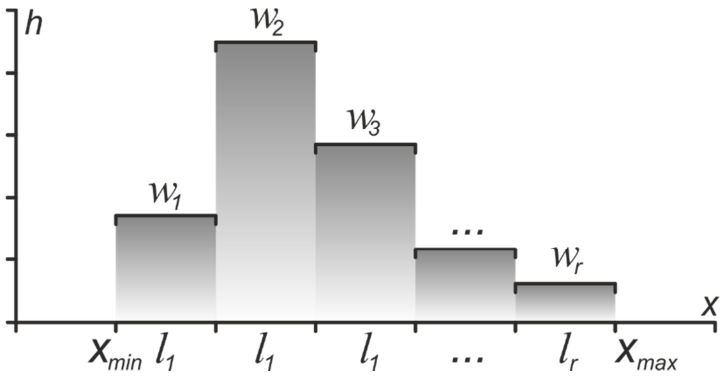


Рис. 2.7. Гістограма відносних частот

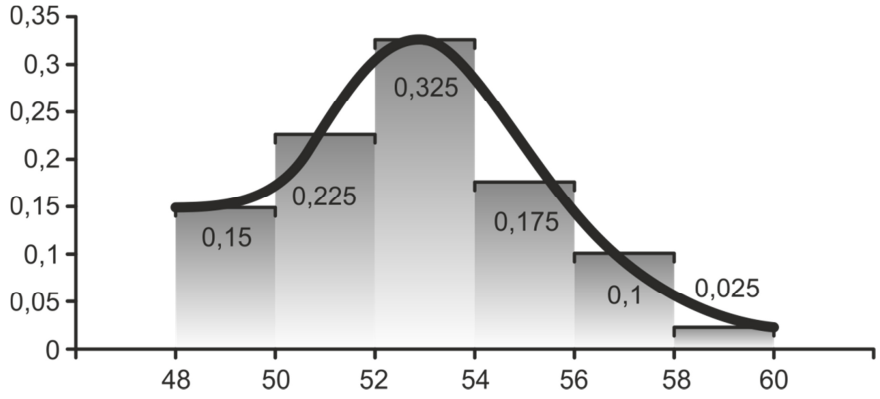


Рис. 2.8. Гістограма відносних частот межі міцності на стиск алевроліту (за даними табл. 2.6)

Проведемо приблизно через середини сторін прямокутників лінію *f**. Очевидно, що при збільшенні числа дослідів можна вибирати все дрібніші інтервали. При цьому лінія *f** буде все більше наближатися до деякої теоретичної кривої, що обмежує площу, яка дорівнює одиниці. Можна переконатися, що ця теоретична крива являє собою графік щільності розподілу досліджуваної кількісної ознаки *X*, тобто графік функції $f(x)$.

Подальший аналіз полягає в підборі аналітичного виразу для функції $f(x)$. При цьому дослідник аналізує вигляд гістограми та лінії *f**, що в першому наближенні вказує на вигляд графіка щільності розподілу *f(x)*. Далі буде показано, що для вибору функції $f(x)$ має значення фізична сутність досліджуваної величини. Але на першому етапі підбору аналітичного виразу для *f(x)* головну роль відіграють так звані числові характеристики випадкової величини, що будуть розглянуті далі.

Числові характеристики випадкової величини. Моменти розподілу. Визначення їх за даними досліду. Числові характеристики – це числа, що характеризують істотні риси розподілу випадкової величини, наприклад, деяке середнє значення, навколо якого групуються можливі значення випадкової величини; яке-небудь число, що характеризує ступінь варіації можливих значень відносно середнього. У теорії ймовірностей застосовується велика кількість числових характеристик. Розглянемо лише ті з них, що використовуються найбільш часто для характеристики кількісних ознак.

Математичне сподівання випадкової величини іноді називають просто середнім значенням. Нехай *X* – *дискретна* випадкова величина, що набуває значень $x_1, x_2, \ldots, x_n$ з ймовірностями $p_1, p_2, \ldots, p_n$. Математичним сподіванням $M(X)$ чи середнім значенням $x_{ср}$ називають величину, що дорівнює сумі добутків її можливих значень на ймовірності, з якими ці значення з'являються:

$$M(X)= \bar{x} = x_1 p_1 + x_2 p_2 + \dots x_n p_{n...} = \sum_{i=1}^{n} x_i p_i . \qquad (2.16)$$

Якщо X – *неперервна* випадкова величина з щільністю розподілу $f(x)$, то її математичне сподівання виражається вже не сумою, а інтегралом

$$M(X) = \int_{-\infty}^{+\infty} xf(x)dx . \qquad (2.17)$$

Із визначення математичного сподівання випливають його властивості:

а) математичне сподівання постійної величини X дорівнює цій величині:

$$M(C)=C; \quad C=const;$$

б) математичне сподівання добутку незалежних випадкових величин дорівнює добутку їх математичних сподівань; наприклад, для трьох незалежних величин X, Y, Z:

$$M\left(X,Y,Z\right) = M\left(X\right)M\left(Y\right)M\left(Z\right),$$

звідси виходить, що постійний множник можна виносити за знак математичного сподівання:

$$M\left(CX\right) = M\left(C\right)M\left(X\right) = CM\left(X\right);$$

в) математичне сподівання алгебраїчної суми випадкових величин дорівнює сумі математичних сподівань цих величин:

$$M\left(X+Y+Z\right) = M\left(X\right) + M\left(Y\right) + M\left(Z\right).$$

Для характеристики центру розподілу використовуються також такі характеристики, як мода і медіана.

Мода випадкової величини *Mo* вказує на її найбільш імовірне значення (для дискретної величини) або на значення з максимальною щільністю розподілу (для неперервної величини). Для ряду чисел мода – це число, яке зустрічається в даному ряду частіше за інші. Ряд чисел може мати більше однієї моди, а може не мати моди зовсім. Модою ряду 32, 26, 18, 26, 15, 21, 26 є число 26, бо воно зустрічається 3 рази. В ряду чисел 5,24; 6,97; 8,56; 7,32 і 6,23 моди немає.

Медіана випадкової величини *Me* вказує на її значення для ймовірності 0,5, тобто на таке значення, для якого ймовірності появи більших та менших значень однакові

$$P(X \le Me) = P(X > Me). \qquad (2.18)$$

Медіаной упорядкованого ряду чисел з непарним числом членів називається число, записане посередині, а медіаной упорядкованого ряду чисел з парним числом членів називається середнє арифметичне двох чисел, записаних посередині. Медіаной довільного ряду чисел називається медіана

відповідного упорядкованого ряду. Медіана ряду 4, 1, 2, 3, 3, 1 дорівнює 2,5. Розглянемо далі характеристики розкиду випадкової величини.

Відхиленням називають різницю між значенням випадкової величини і її математичним сподіванням: $|x - M(X)|$.

Дисперсією (розсіюванням) називають математичне сподівання квадрата відхилень випадкової величини:

$$D(X) = M\left(|x - M(X)|^2\right). \tag{2.19}$$

Формула для визначення дисперсії має вигляд:

а) для дискретної випадкової величини

$$D(X) = \sum_{i=1}^{n} \left(x_i - M(X)\right)^2 p_i; \tag{2.20}$$

б) для неперервної випадкової величини з щільністю розподілу *f(x)*

$$D(X) = \int_{-\infty}^{+\infty} \left(x - M(X)\right)^2 f(x)dx. \tag{2.21}$$

На практиці обчислення дисперсії за формулами (2.20) і (2.21) трохи громіздкі. З урахуванням властивостей математичного сподівання формулу (2.19) можна перетворити до вигляду

$$D(X) = M\left(X^2\right) - \left[M(X)\right]^2. \tag{2.22}$$

Тоді для дискретної випадкової величини

$$D(X) = \sum_{i=1}^{n} x_i^{\,2} p_i - \left[M(X)\right]^2, \tag{2.23}$$

а для неперервної випадкової величини

$$D(X) = \int_{-\infty}^{+\infty} x^2 f(x)dx - \left[M(X)^2\right]. \tag{2.24}$$

Оскільки при піднесенні в квадрат змінюється розмірність випадкової величини, для характеристики розкиду використовують величину, яка обчислюється таким чином:

$$\sigma(X) = \sqrt{D(X)} \tag{2.25}$$

і називається *середнім квадратичним* або *стандартним відхиленням*.

Початковим моментом k-го порядку називають математичне сподівання величини X^k, тобто

$$\nu_k = M(X^k). \tag{2.26}$$

Так, для дискретної величини формула має вигляд

$$\nu_k = \sum_{i=1}^{n} x_i^k p_i , \tag{2.27}$$

а для неперервної величини

$$\nu_k = \int_{-\infty}^{+\infty} x^k f(x)dx . \tag{2.28}$$

Зокрема, початковий момент першого порядку є не чим іншим, як математичним сподіванням: $\nu_1 = M(X)$, початковий момент другого порядку є математичне сподівання квадрата випадкової величини: $\nu_2 = M(X^2)$ і т. д.

Центральним моментом k-го порядку називають математичне сподівання величини $(X - M(X))^k$:

$$\mu_k = M\left((X - M(X))^k\right). \tag{2.29}$$

Зокрема, з (2.29) випливає, що центральний момент першого порядку (тобто математичне сподівання відхилень) дорівнює нулю. Центральний момент другого порядку являє собою дисперсію:

$$\mu_2 = M\left((X - M(X))^2\right) = D(X). \tag{2.30}$$

З формули (2.30) випливає, що дисперсія легко може бути виражена через початкові моменти першого і другого порядків:

$$D(X) = \mu_2 = \nu_2 - \nu_1^2 . \tag{2.31}$$

Третій і четвертий центральні моменти також можна виразити через початкові моменти :

$$\mu_3 = \nu_3 - 3\nu_2\nu_1 + 2\nu_1^3 , \tag{2.32}$$

$$\mu_4 = \nu_4 - 4\nu_3\nu_1 + 6\nu_2\nu_1^2 - 3\nu_1^4 .$$

Моменти вищих порядків використовуються для опису закону розподілу. З ними пов'язані

1) *асиметрія*

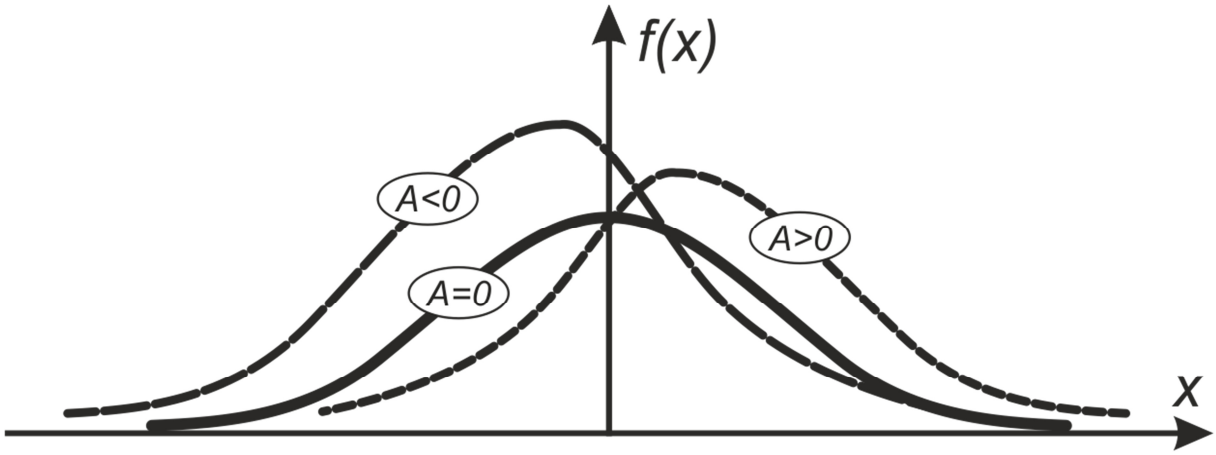
$$A = \frac{\mu_3}{\sigma^3} = \frac{\mu_3}{\mu_2^{3/2}} , \tag{2.33}$$

що характеризує несиметричність кривої щільності розподілу випадкової величини *X* відносно її математичного сподівання (рис. 2.8);

2) *ексцес*

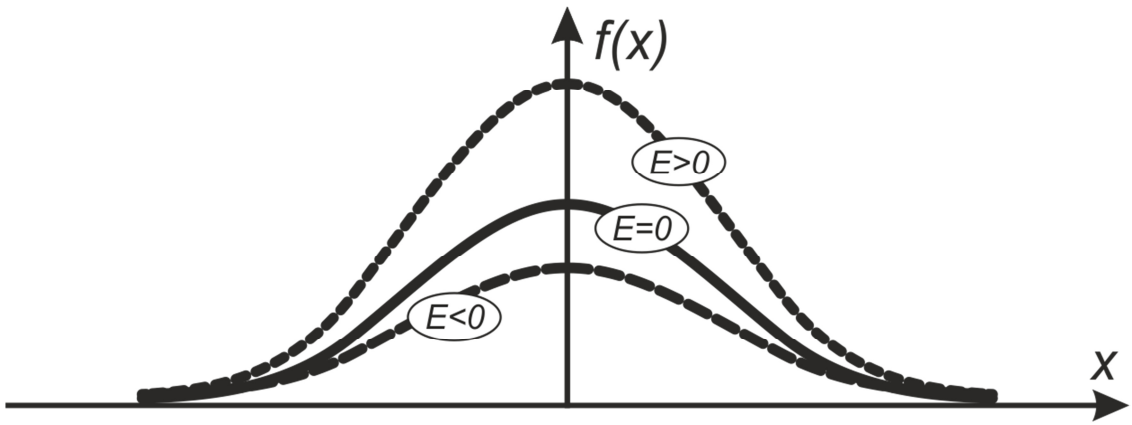
$$E = \frac{\mu_4}{\sigma^4} - 3 = \frac{\mu_4}{\mu_2^2} - 3 , \tag{2.34}$$

що характеризує гостровершинність графіка диференціальної функції розподілу (рис. 2.9).



$A < 0$ – лівостороння асиметрія, $A > 0$ – правостороння асиметрія

а)



б)

Рис. 2.9. Вигляд кривих щільності розподілу випадкової величини:
а) при різній асиметрії, б) при різному ексцесі

Визначення числових характеристик кількісної ознаки за даними досліду. Параметри статистичного розподілу. Усі моменти розподілу випадкової величини, розглянуті вище, мають свої статистичні аналоги. Вони називаються *моментами статистичного розподілу* і визначаються за даними дослідів (вибірками).

Якщо дані згруповані у варіаційний ряд, подібний до наведеного в табл. 2.3, статистичні моменти визначаються за тими самими формулами, що і їх теоретичні аналоги для дискретної випадкової величини, але із заміною ймовірності p_i на відносну частоту w_i (табл. 2.7). Статистичні параметри будемо позначати «*».

Наприклад, статистичний аналог математичного сподівання, тобто статистичний момент першого порядку, називається *вибірковою середньою*, позначається $\overline{X}*$ і визначається формулою:

$$\overline{x}* = \nu_1^* = \sum_{i=1}^{r} x_i w_i \ , \tag{2.35}$$

r – кількість значень випадкової ознаки, що спостерігаються у вибірці.

Статистичним аналогом дисперсії є *вибіркова дисперсія*, що визначається як статистичний центральний момент другого порядку:

$$D^* = \mu_2 = \sum_{i=1}^{r} x_i^2 w_i - \left(\overline{x}*\right)^2 \ . \tag{2.36}$$

Тут варто зробити зауваження.

1. Статистичні величини ν_k^*, μ_k^***,** у тому числі середнє вибіркове $\overline{x}^*$ і дисперсія D^*, є для теоретичних величин μ_k, ν_k (у тому числі $\overline{x}, D(x)$ **)** так званими точковими оцінками, тобто такими, що виражаються одним числом.

2. У повному курсі математичної статистики доводиться, що середнє вибіркове дає *незміщену* оцінку дійсної величини математичного сподівання. На відміну від неї вибіркова дисперсія дає *зміщену* оцінку дійсної дисперсії досліджуваної кількісної ознаки. *Незміщену* оцінку дає величина, що обчислюється таким чином:

$$S^2 = \frac{n}{n-1} D^* . \tag{2.37}$$

Її називають *виправленою дисперсією*. Оцінкою середнього квадратичного відхилення є корінь з виправленої дисперсії (стандартне відхилення).

Статистичні оцінки асиметрії та ексцесу пов'язані зі статистичними центральними моментами третього і четвертого порядків.

Якщо дослідні дані сформовані у вигляді інтервального ряду, у ролі варіанти x_i виступають середини інтервалів (u_i).

Моменти статистичного розподілу можуть бути визначені також і за незгрупованими даними (табл. 2.7).

Приклад 2.4. Визначити моменти статистичного розподілу для даних випробувань межі міцності на стиск зразків алевроліту, що наведені в табл. 2.4 і згруповані в інтервальний ряд (табл. 2.5).

1. Визначимо початковий момент першого порядку, тобто вибіркове середнє:

$$\overline{x}^* = \nu_1^* = 49 \cdot 0{,}15 + 51 \cdot 0{,}225 + 53 \cdot 0{,}325 + 55 \cdot 0{,}175 + 57 \cdot 0{,}1 + 59 \cdot 0{,}025 = 52{,}85 \,\text{МПа}.$$

Таблиця 2.7

Статистичні моменти розподілу

Назва статистичного моменту	Незгруповані дані (n – об'єм вибірки)	Згруповані дані (r – кількість варіант)
1. Початковий момент k-го порядку *Окремий випадок*: початковий момент першого порядку – середнє вибіркове	$\nu_k^* = \frac{1}{n}\sum_{i=1}^{n} x_i^k$ $\overline{x}^* = \nu_1^* = \frac{1}{n}\sum_{i=1}^{n} x_i$	$\nu_k^* = \sum_{i=1}^{r} x_i^k \ w_i$ $\overline{x}^* = \nu_1^* = \sum_{i=1}^{r} x_i \ w_i$
2. Центральний момент k-го порядку *Окремі випадки:* – центральний момент другого порядку – вибіркова дисперсія; – центральний момент третього порядку; – центральний момент четвертого порядку	$\mu_k^* = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}^*)^k$ $D^* = \mu^*{}_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\overline{x}^*)^2$ $\mu_3^* = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}^*)^3$ $\mu_4^* = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}^*)^4$	$\mu_k^* = \sum_{i=1}^{r}(x_i - \overline{x}^*)^k \cdot w_i$ $D^* = \mu_2^* = \sum_{i=1}^{r} x_i^2 w_i - (\overline{x}^*)^2$ $\mu_3^* = \sum_{i=1}^{r}(x_i - \overline{x}^*)^3 \cdot w_i$ $\mu_4^* = \sum_{i=1}^{r}(x_i - \overline{x}^*)^4 \cdot w_i$
3. Виправлена дисперсія	$S^2 = \frac{n}{n-1} D^*$	
4. Статистична оцінка середнього квадратичного відхилення (стандартне відхилення)	$s = \sqrt{S^2} = \sigma\sqrt{\frac{n-1}{n}}$	
	$\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n}}$	$\sigma = \sqrt{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 w_i}$
5. Відносна варіація, коефіцієнт варіації	$\eta^* = \frac{S}{\overline{x}^*}$; $V^* = \eta^* \cdot 100\%$.	
6. Статистична оцінка асиметрії	$A^* = \frac{\mu_3^*}{S^3} = \frac{\mu_3^*}{\mu_2^{*\,3/2}}$	
7. Статистична оцінка ексцесу	$E^* = \frac{\mu_4^*}{S^4} - 3 = \frac{\mu_4^*}{\mu_2^{*\,2}} - 3$	

2. Визначимо характеристики варіації значень кількісної ознаки:

- статистичний початковий момент другого порядку:

$$\nu_2^* = 49^2 \cdot 0{,}15 + 51^2 \cdot 0{,}225 + 53^2 \cdot 0{,}325 + 55^2 \cdot 0{,}175 + 57^2 \cdot 0{,}1 + 59^2 \cdot 0{,}025 = 2799{,}6 \text{ МПа}^2,$$

статистичний центральний момент другого порядку (вибіркова дисперсія):

$$D^* = \mu_2^* = \nu_2^* - (\nu_1^*)^2 = 2799{,}6 - 52{,}85^2 = 6{,}48 \text{ МПа}^2;$$

- виправлена дисперсія

$$S^2 = \frac{40}{40-1} \cdot 6{,}48 = 6{,}64 \text{ МПа}^2;$$

- стандартне відхилення

$$S = \sqrt{S^2} = 2{,}58 \text{ МПа};$$

- відносна варіація, коефіцієнт варіації

$$\eta^* = \frac{2{,}58}{52{,}85} = 0{,}05; \; V^* = 0{,}05 * 100 = 5\%.$$

3. Визначимо величини, що характеризують асиметрію та ексцес:

- статистичні початкові моменти третього і четвертого порядків:

$$\nu_3^* = 49^3 \cdot 0{,}15 + 51^3 \cdot 0{,}225 + 53^3 \cdot 0{,}325 + 55^3 \cdot 0{,}175 + 57^3 \cdot 0{,}1 + 59^3 \cdot 0{,}025 = 148648{,}3,$$

$$\nu_4^* = 49^4 \cdot 0{,}15 + 51^4 \cdot 0{,}225 + 53^4 \cdot 0{,}325 + 55^4 \cdot 0{,}175 + 57^4 \cdot 0{,}1 + 59^4 \cdot 0{,}025 = 7911190;$$

- статистичні центральні моменти третього і четвертого порядків:

$$\mu_3^* = \nu_3^* - 3\nu_2^*\nu_1^* + 2\left(\nu_1^*\right)^3 = 148648{,}3 - 4 \cdot 2799{,}6 \cdot 52{,}85 + 2 \cdot 52{,}85^3 = 4{,}72,$$

$$\mu_4^* = \nu_4^* - 4\nu_3^*\nu_1^* + 6\nu_2^*\left(\nu_1^*\right)^2 - 3\left(\nu_1^*\right)^2 =$$

$$= 7911190 - 4 \cdot 148648{,}3 \cdot 52{,}85 + 6 \cdot 2799{,}6 \cdot 52{,}85^2 - 3 \cdot 52{,}85^4 = 104{,}76;$$

- статистичні оцінки асиметрії та ексцесу:

$$A^* = \frac{4{,}72}{2{,}58^3} = 0{,}28; \; E^* = \frac{104{,}76}{2{,}58^4} - 3 = -0{,}63.$$

Аналізуючи статистичні моменти досліджуваної ознаки – міцності гірської породи на одноосьовий стиск – можна зробити такі висновки:

– середнє значення ознаки складає 52,85 МПа;

– значення ознаки коливаються навколо свого середнього з невеликою варіацією, що складає 5%;

– розподіл ознаки має невелику правобічну асиметрію, його можна вважати майже симетричним;

– значення ексцесу також близько до нуля;

– за виглядом гістограми можна припустити, що розподіл даної ознаки описується деякою функцією $f(x)$, графік якої має максимум у точці $x = 52{,}85$, крива майже симетрична відносно вертикальної прямої, проведеної через цю точку.

Для прогнозу появи значень ознаки з тією чи іншою ймовірністю потрібно підібрати для функції $f(x)$ аналітичний вираз. Існує ряд функцій, що задовільно описують розподіл достатньо вивчених випадкових величин. Вони називаються теоретичними законами розподілу. Далі будуть наведені деякі з них.

Контрольні питання до підрозділу 2.2

1. У чому сутність вибіркового методу? Поясніть це на прикладі вивчення властивостей гірських порід.
2. Яку основну вимогу повинна задовольняти вибірка з генеральної сукупності?
3. Як групують статистичні дані? Який вигляд має варіаційний ряд?
4. Що являє собою інтервальний ряд?
5. Що є графічним зображенням інтервального ряду? Які висновки може зробити дослідник після побудови гістограми частот. Які подальші кроки він може почати для вивчення розподілу кількісної ознаки?
6. Що характеризує математичне сподівання випадкової величини? Наведіть формули для визначення математичного сподівання дискретної і безперервної випадкових величин.
7. Що характеризує дисперсія випадкової величини?
8. Наведіть формули для визначення дисперсії дискретної і безперервної випадкових величин.
9. Що таке початковий момент розподілу k-го порядку? Яку числову характеристику дає початковий момент 1-го порядку? Що таке центральний момент k-го порядку? Яку числову характеристику дає центральний момент 2-го порядку?
10. Що характеризують асиметрія та ексцес випадкової величини? Наведіть формули для визначення цих характеристик.
11. Як визначити числові характеристики випадкової величини за даними досліду? Наведіть формули для середньої вибіркової і вибіркової дисперсій для згрупованих і незгрупованих даних.
12. Що таке виправлена дисперсія? Наведіть формулу для її визначення. Що таке коефіцієнт варіації випадкової величини?

2.3. Ідентифікація ймовірнісного закону розподілу кількісної ознаки

2.3.1 Рівномірний закон розподілу ймовірностей

Неперервна випадкова величина X називається розподіленою рівномірно на відрізку $[a, b]$, якщо її щільність розподілу ймовірностей постійна на даному відрізку, тобто задана формулою

$$f(x)=\begin{cases}0, & x\notin[a,b],\\ C, & x\in[a,b].\end{cases}$$

Для того, щоб дана функція дійсно описувала закон розподілу ймовірностей випадкової величини, вона має задовольняти умову (2.10), тобто

$$C\int_a^b dx=1.$$

За цією умовою знайдемо значення постійної C:

$$C\int_a^b dx=C(b-a),\ \text{звідси}\ C(b-a)=1\Rightarrow C=\frac{1}{b-a}.$$

Отже, рівномірний розподіл має вигляд

$$f(x)=\begin{cases}0, & x\notin[a,b],\\ \dfrac{1}{b-a}, & x\in[a,b].\end{cases}\qquad(2.38)$$

Суть рівномірності полягає в тому, що відповідно формулі (2.9) який би внутрішній проміжок фіксованої довжини l ми би не розглянули, ймовірність того, що випадкова величина набуде значення саме з цього проміжку, буде тією самою.

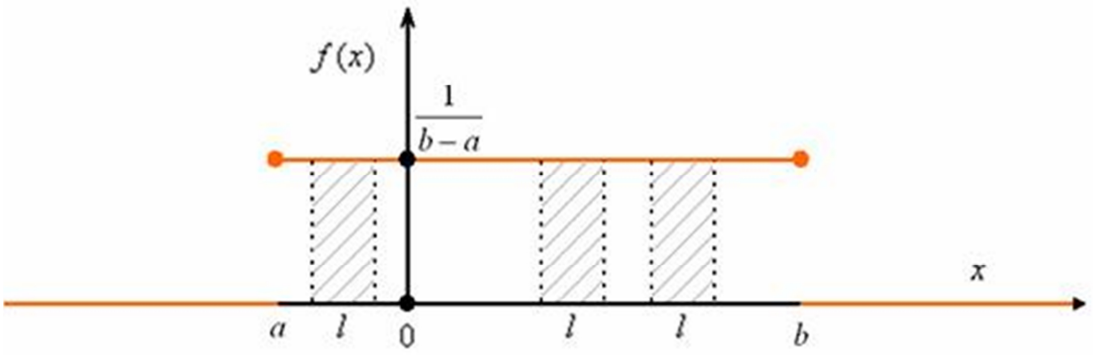


Рис. 2.10. Диференційна функція (щільність) рівномірного розподілу

Прикладом рівномірно розподіленої неперервної випадкової величини X є помилка при округленні відліку до найближчого цілого ділення шкали

вимірювального приладу. Рівномірний розподіл в інтервалі [0, 1] називається *стандартним* і використовується для генерації випадкових чисел.

Математичне сподівання рівномірно розподіленої випадкової величини обчислюється за формулою (2.17)

$$M(X)=\int_a^b xf(x)dx=\frac{1}{(b-a)}\int_a^b xdx=\frac{a+b}{2}. \qquad (2.39)$$

Цей результат є очевидним, оскільки щільність розподілу є константою. Дисперсію згідно з формулою (2.24) обчислимо так:

$$D(X)=\int_a^b x^2 f(x)dx-(M(X))^2=\frac{1}{(b-a)}\int_a^b x^2dx-\frac{(a+b)^2}{4}=\frac{(b-a)^2}{12}. \qquad (2.40)$$

Середнє квадратичне відхилення рівномірно розподіленої величини отримаємо згідно з (2.25)

$$\sigma(X)=\sqrt{\frac{(b-a)^2}{12}}=\frac{b-a}{2\sqrt{3}}. \qquad (2.41)$$

2.3.2 Експоненціальний закон розподілу

Така випадкова величина, як інтервал часу між незалежними подіями, що відбуваються з середньою інтенсивністю λ, описується щільністю розподілу (диференціальною функцією), графік якої наведено на рис. 2.11.

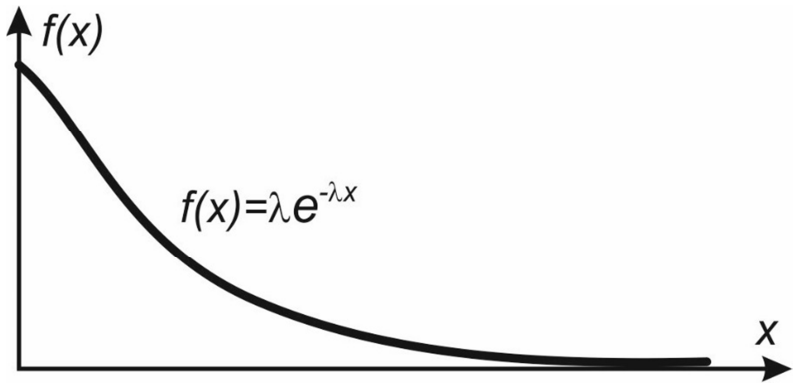


Рис. 2.11. Диференціальна функція (щільність) експоненціального розподілу

Видно, що ця крива являє собою експоненту, що спадає на проміжку $(0,+\infty)$, тобто для зазначеного інтервалу $f(x)=C\,e^{-\lambda x}$. Для того, щоб дана функція дійсно описувала закон розподілу ймовірностей випадкової величини, вона повинна задовольняти умову (2.10), отже,

$$C\int_0^{+\infty} e^{-\lambda x}dx = 1.$$

За цією умовою знайдемо значення постійної C.

$$C\int_0^{+\infty} e^{-\lambda x}dx = -\frac{1}{\lambda}C\lim_{b\to\infty}\left(e^{-\lambda b} - e^{0}\right) = \frac{C}{\lambda}\ ;\ \text{звідси}\ \frac{C}{\lambda} = 1 \Rightarrow C = \lambda .$$

Таким чином, щільність розподілу експоненціального закону має вигляд:

$$f(x) = \lambda\ e^{-\lambda x}. \tag{2.42}$$

Величину λ називають *параметром розподілу*. Для опису розподілу конкретної кількісної ознаки параметр розподілу має бути визначений на основі статистичних даних. Для цього використовують *метод моментів*. Він полягає в прирівнюванні начальних і центральних емпіричних і теоретичних моментів одного порядку:

$$\nu_k^* = \nu_k;\quad \mu_k^* = \mu_k . \tag{2.43}$$

Ми побачимо далі, що теоретичні моменти для того чи іншого розподілу часто пов'язані з параметром даного розподілу. Тому, насамперед, варто визначити теоретичні моменти для даного розподілу. Знайдемо математичне сподівання експоненціального розподілу, тобто початковий момент першого порядку. За формулою (2.17) отримаємо

$$M(X) = \int_0^{+\infty} x\,f(x)dx = \lambda\int_0^{+\infty} xe^{-\lambda x}dx = \left\{\begin{array}{ll} x = u & du = dx \\ e^{-\lambda x} = dV & V = -\dfrac{1}{\lambda}e^{-\lambda x}\end{array}\right\} =$$

$$= \lambda\lim_{b\to\infty}\left(-\frac{1}{\lambda}xe^{-\lambda x} + \frac{1}{\lambda}\int_0^b e^{-\lambda x}dx\right) = \lambda\lim_{b\to\infty}\left(-\frac{1}{\lambda}be^{-\lambda b} - \frac{1}{\lambda^2}e^{-\lambda b} + \frac{1}{\lambda^2}\right) = \frac{1}{\lambda}.$$

Таким чином, для експоненціального розподілу:

$$M(X) = \frac{1}{\lambda}. \tag{2.44}$$

Аналогічно за формулою (2.24), виконуючи два рази інтегрування за частинами і переходячи до границі функції, знайдемо дисперсію експоненціального розподілу, тобто центральний момент другого порядку

$$D(X) = \lambda\int_0^{+\infty} x^2e^{-\lambda x}dx - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}. \tag{2.45}$$

Визначимо середнє квадратичне відхилення відповідно до формули (2.25)

$$\sigma(X) = \sqrt{\frac{1}{\lambda^2}} = \frac{1}{\lambda}. \qquad (2.46)$$

Як бачимо, параметр розподілу λ тісно пов'язаний з числовими характеристиками випадкової величини. Крім того, з (2.44) і (2.45) виходить, що для експоненціального закону притаманна рівність середнього квадратичного відхилення і математичного сподівання. Цей факт допоможе у подальшому підбирати закон розподілу для емпіричних даних.

Застосуємо тепер метод моментів для визначення λ. Прирівняємо теоретичний і емпіричний моменти першого порядку, тобто математичне сподівання і середнє вибіркове:

$$M(X) = \overline{x}^*.$$

Оскільки для показового закону $M(X) = 1/\lambda$, отримаємо $\overline{x}^* = 1/\lambda$, звідки $\lambda = 1/\overline{X}^*$.

Таким чином, для знаходження параметра експоненціального розподілу достатньо мати середнє вибіркове за емпіричними даними.

Приклад 2.5. Нехай відомі результати перевірки деякого приладу на випробному стенді. Випробувано 100 зразків, вивчено час безвідмовної роботи приладу. Він склав: для 80 приладів менше 10 годин; для 10 приладів від 10 до 20 годин; для 5 приладів від 20 до 30 годин і для 5 приладів від 30 до 40 годин. Визначити ймовірність того, що час роботи приладу складе не менше 20 годин.

Інтервальний ряд для цих даних наведено в табл. 2.8.

Таблиця 2.8

Приклад інтервального ряду даних випробувань

Час безвідмовної роботи, год	0 – 10	10 – 20	20 – 30	30 – 40
Середини інтервалів, год	5	15	25	35
Кількість приладів (частота ознаки), шт.	80	10	5	5
Відносна частота	0,8	0,1	0,05	0,05

Спираючись на вигляд гістограми (рис. 2.12), у грубому наближенні можна висловити припущення про те, що дана кількісна ознака – час безвідмовної роботи приладу – розподілена за експоненціальним законом.

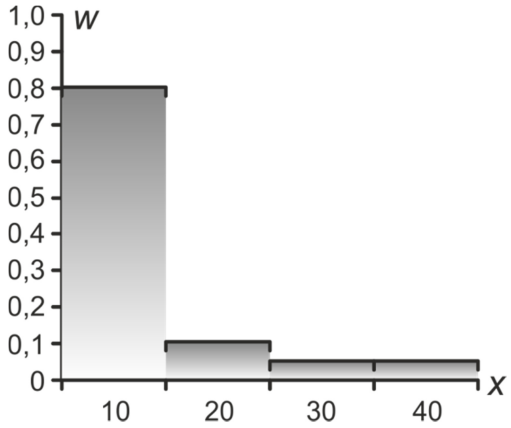


Рис. 2.12. Гістограма відносних частот для розподілу з табл. 2.8

Визначимо емпіричні моменти розподілу:

– середнє вибіркове

$$\overline{x}^{*} = 5 \cdot 0,8 + 15 \cdot 0,1 + 25 \cdot 0,05 + 35 \cdot 0,05 = 4 + 1,5 + 3 = 8,5;$$

– вибіркова дисперсія

$$D^{*} = 25 \cdot 0,8 + 225 \cdot 0,1 + 625 \cdot 0,05 + 1225 \cdot 0,05 - (8,5)^2 =$$
$$= 20 + 22,5 + 92,5 - 72,25 = 62,75;$$

– виправлена дисперсія і середнє квадратичне відхилення

$$S^2 = \frac{n}{n-1} \cdot 62,75 = \frac{100}{99} \cdot 6,75 = 63,38;$$
$$\sigma^{*} = \sqrt{S^2} = 7,9.$$

Бачимо, що математичне сподівання і середнє квадратичне відхилення є близькими за значеннями. Це також свідчить на користь експоненціального закону розподілу з диференціальною функцією $f(x) = \lambda\ e^{-\lambda x}$. Визначимо за допомогою методу моментів параметр розподілу λ:

$$\lambda = \frac{1}{\overline{x}^{*}} = \frac{1}{8,5} = 0,117 \approx 0,12\ .$$

Тепер, маючи теоретичний закон розподілу даної кількісної ознаки, можна прогнозувати ймовірність появи певного значення цієї випадкової величини. Отже, імовірність того, що час безвідмовної роботи приладу буде не менше 20 годин, визначимо за формулою (2.7)

$$P\,(20 < t < +\infty) = 0,12 \int_{20}^{+\infty} e^{-0,12t} dt = 0,12 \cdot \left(\frac{1}{-0,12} \right) \lim_{b \to +\infty} (e^{-b} - e^{-2,4}) = 0,091\ .$$

Експоненціальний закон відіграє велику роль у теорії надійності механізмів і машин, а також широко використовується в геології, наприклад

для опису розподілу параметрів малоамплітудних порушень вугільного пласта в межах шахтного поля.

2.3.3 Нормальний закон розподілу

Звернемося тепер до гістограми, що зображена на рис. 2.8. Крива, що проведена через середини сторін багатокутників, у першому наближенні показує, який вигляд має диференціальна функція розподілу *f(x)*. У даному випадку крива має форму, майже симетричну відносно вертикальної прямої $x = \bar{x}^*$. Характерно те, що велика частина значень кількісної ознаки групується поблизу точки $x = \bar{x}^*$, що є точкою максимуму кривої.

Велика кількість випадкових величин має подібну щільність розподілу ймовірностей. Її називають *нормальним законом* або *законом Гаусса* (рис. 2.13). Вираз для диференціальної функції розподілу закону Гаусса має вигляд:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}. \qquad (2.47)$$

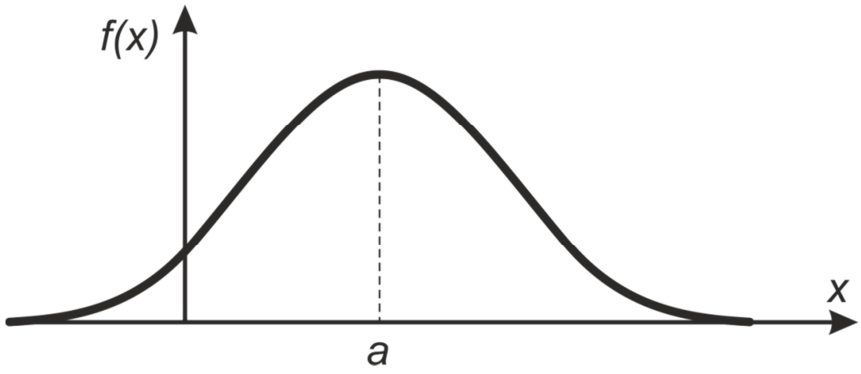


Рис. 2.13. Графік диференціальної функції (щільності) нормального розподілу

Перш за все виникає питання, чи є функція (2.47) щільністю розподілу. Перевіримо рівність (2.10). Для цього зробимо заміну $\frac{x-a}{\sigma} = t \Rightarrow dt = \frac{1}{\sigma}dx$ та візьмемо до уваги інтеграл Ейлера – Пуассона

$$\int_{-\infty}^{+\infty} e^{\frac{-t^2}{2}} dt = \sqrt{2\pi}\ .$$

Тоді отримаємо

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \cdot e^{-\frac{(x-a)^2}{2\sigma^2}} dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{\frac{-t^2}{2}} \sigma dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{\frac{-t^2}{2}} dt = 1.$$

Таким чином, рівність (2.11) виконується, площа під кривою (2.47) дорівнює одиниці. Максимальна ордината кривої, що дорівнює $1/\sigma\sqrt{2\pi}$, відповідає точці $x = a$. При віддаленні від точки a щільність розподілу зменшується і при $x \to \pm\infty$ крива асимптотично наближається до осі абсцис.

З'ясуємо значення параметрів a і σ, що входять до виразу (2.47). Для цього знайдемо числові характеристики нормальної величини. Відповідно до (2.17) математичне сподівання нормальної величини обчислимо за формулою

$$M(X) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} x\, e^{-\frac{(x-a)^2}{2\sigma^2}}\, dx .$$

Використовуючи попередню заміну та інтеграл Ейлера – Пуассона, отримаємо $M(X) = a$. Аналогічно можна показати, що $D(X) = \sigma^2, \sigma(X) = \sigma$.

Отже, параметр a являє собою математичне сподівання і характеризує центр розподілу; параметр σ характеризує розсіювання випадкової величини відносно центра розподілу та обумовлює крутість кривої: чим більше σ, тим більше розсіювання, тим менше крутість кривої.

Можна показати, що для нормальної випадкової величини центральний момент 3-го порядку дорівнює нулю ($\mu_3 = 0$), а центральний момент четвертого порядку $\mu_4 = 3\sigma^4$.

Це означає, що асиметрія та ексцес для нормального закону дорівнюють нулю

$$A = \frac{\mu_3}{\sigma^3} = 0 \; ; \; E = \frac{\mu_4}{\sigma^4} - 3 = 0 .$$

Зауваження. Закон розподілу, що виражається диференціальною функцією

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

називається *загальним нормальним* розподілом. Для нього $M(X) = a, \sigma(X) = \sigma$. Часто при розв'язуванні задач переходять до так званої *нормованої* величини $\frac{x-a}{\sigma} = t$. Для неї щільність розподілу має вигляд:

$$f^0(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} . \qquad (2.48)$$

і називається *нормованим нормальним* розподілом.

Математичне сподівання і середнє квадратичне відхилення для нормованої величини відповідно $M(t) = 0$; $\sigma(t) = 1$.

Інтегральна функція нормального розподілу має вигляд:

а) для загального розподілу

$$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{(x-a)^2}{2\sigma^2}}\,dx, \tag{2.49}$$

б) для нормованого розподілу

$$F_0(t)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-\frac{t^2}{2}}dt. \tag{2.50}$$

Інтеграли (2.49), (2.50) не виражаються в елементарних функціях. Але для інтеграла (2.50) існують таблиці, де наведені значення функції $F_0(t)$ для різних значень аргументу *t* (додаток А). Використовують також функцію Лапласа:

$$\Phi(t)=\frac{1}{\sqrt{2\pi}}\int_{0}^{t} e^{-\frac{t^2}{2}}\,dt\ , \tag{2.51}$$

перевагою якої є непарність $\Phi(-t)=-\Phi(t)$. Функції (2.50) і (2.51) пов'язані співвідношенням $\Phi(t)=F_0(t)-0{,}5$.

Відповідно формулі (2.7) ймовірність потрапляння випадкової величини в інтервал (x_1, x_2) визначається так:

$$P(x_1<X<x_2)=\int_{x_1}^{x_2} f(x)dx=\frac{1}{\sigma\sqrt{2\pi}}\int_{x_1}^{x_2} e^{-\frac{(x-a)^2}{2\sigma^2}}\,dx.$$

Зробимо заміну змінних: $t=\frac{x-a}{\sigma};\ dx=\sigma dt;\ t_1=\frac{x-a}{\sigma};\ t_2=\frac{x-a}{\sigma}.$

Тоді отримаємо:

$$P(x_1<X<x_2)=\frac{1}{\sqrt{2\pi}}\int_{t_1}^{t_2} e^{-\frac{t^2}{2}}\,dt=F_0(t_2)-F_0(t_1)=F_0\left(\frac{x_2-a}{\sigma}\right)-F_0\left(\frac{x_1-a}{\sigma}\right).$$

Тут $F_0(t)$ – інтегральна функція (2.50). Можна показати, що при такій самій заміні може бути використана і функція Лапласа (2.51). Таким чином, для нормального закону розподілу ймовірність потрапляння випадкової величини в заданий інтервал визначається так:

$$P(x_1<X<x_2)=F_0\left(\frac{x_2-a}{\sigma}\right)-F_0\left(\frac{x_1-a}{\sigma}\right)=\Phi\left(\frac{x_2-a}{\sigma}\right)-\Phi\left(\frac{x_1-a}{\sigma}\right). \tag{2.52}$$

Імовірність заданого відхилення. Правило трьох сигм. Знайдемо ймовірність того, що відхилення нормальної випадкової величини *X* від її математичного сподівання *a* не перевищує величини δ, тобто знайдемо ймовірність події $P(|X-a|<\delta)$. За формулою (2.52) отримаємо:

$$P(|X-a|<\delta)=P(a-\delta<X<a+\delta)=\Phi\left(\frac{a+\delta-a}{\sigma}\right)-\Phi\left(\frac{a-\delta-a}{\sigma}\right)=$$
$$=\Phi\left(\frac{\delta}{\sigma}\right)-\Phi\left(-\frac{\delta}{\sigma}\right)=2\Phi\left(\frac{\delta}{\sigma}\right). \quad (2.53)$$

Тут використано факт, що функція (2.51) непарна. Проаналізуємо такі випадки:

1. Нехай $\delta=\sigma$. Тоді

$$P(|X-a|<\sigma)=2\Phi\left(\frac{\sigma}{\sigma}\right)=2\Phi(1)=0,68.$$

Значення функції $\Phi(1)=0,34$ визначене за таблицею з додатку А.

2. Нехай $\delta=2\sigma$. Тоді

$$P(|X-a|<2\sigma)=2\Phi\left(\frac{2\sigma}{\sigma}\right)=2\Phi(2)=0,95.$$

Тут $\Phi(2)=0,478$.

3. Нехай $\delta=3\sigma$. Тоді

$$P(|X-a|<3\sigma)=2\Phi\left(\frac{3\sigma}{\sigma}\right)=2\Phi(3)=0,997.$$

Тут $\Phi(2)=0,4986$.

Бачимо, що з імовірністю 0,997, тобто з *практичною вірогідністю*, значення нормальної величини потрапляють в інтервал

$$a-3\sigma<X<a+3\sigma. \quad (2.54)$$

Цей факт називають правилом трьох сигм. За його допомогою можна попередньо судити про те, чи є дана випадкова величина розподіленою за нормальним законом, чи ні. Таким чином, однією з підстав для вибору гіпотези про нормальний закон розподілу є виконання умов:

1) 99,7 % значення випадкової величини належать інтервалу

$$a-3\sigma<X<a+3\sigma;$$

2) асиметрія й ексцес близькі до нуля ($A\cong 0$, $E\cong 0$).

Нормальний закон розподілу відіграє особливу роль у теорії ймовірностей і статистики. У теорії ймовірностей доводиться так звана центральна гранична теорема, відповідно до якої сума великої кількості незалежних випадкових величин, підпорядкованих будь-якому розподілу, приблизно підпорядковується нормальному закону. Ця закономірність тим точніше, чим більша кількість випадкових величин складає цю суму.

Наприклад, помилка вимірів – це випадкова величина, що є сумою елементарних помилок, кожна з яких викликана дією певної причини, яка не

залежить від інших. Якщо всі ці помилки відіграють у сумі рівномірно малу роль, то сама сума розподілена за нормальним законом.

Оскільки умови, що визначають нормальний розподіл, зустрічаються часто, даний закон розподілу ймовірностей здобув широке застосування. Зокрема, узагальнюючи результати випробувань гірських порід, дослідники часто дотримуються гіпотези про нормальний розподіл, параметри якого мають ясний фізичний зміст і легко визначаються за характеристиками вибірки.

Приклад 2.6. Вище як приклад вивчався статистичний розподіл такої кількісної ознаки, як міцність на одноосьове стискання гірської породи. Крива f^*, що вирівнює гістограму на рис. 2.8, має приблизно такий самий вигляд, як і нормальна щільність розподілу (рис. 2.13). На користь нормального розподілу свідчать і близькі до нуля значення асиметрії та ексцесу, а також виконання правила трьох сигм. Дійсно, максимальне і мінімальне значення випадкової величини ($x_{min} = 48{,}4$; $x_{max} = 60{,}0$) потрапляють в інтервал $(52{,}8 - 2{,}58 \cdot 3;\ 52{,}8 + 2{,}58 \cdot 3)$.

Тому в першому наближенні можна висунути гіпотезу про те, що значення досліджуваної кількісної ознаки – міцності на стиск деякої гірської породи – підпорядковуються нормальному закону розподілу.

Визначимо параметри цього розподілу на основі дослідних даних, використовучи метод моментів. Дорівняємо початкові статистичний і теоретичний моменти розподілу: $M(X) = \bar{x}^*$. Оскільки для нормального закону $M(X) = a$, отримаємо $\bar{x}^* = a$, тобто $a = 52{,}8$.

Далі, щоб знайти ще один невідомий параметр σ, дорівняємо центральні статистичний і теоретичний моменти другого порядку: $D(X) = S^2$. Оскільки для нормального закону $D(X) = \sigma^2$, то

$$S^2 = \sigma^2 \Rightarrow \sigma = S = 2{,}58 .$$

Таким чином, для даної кількісної ознаки щільність розподілу ймовірностей описується функцією:

$$f(x) = \frac{1}{2{,}58\sqrt{2\pi}}\, e^{-\frac{(x-52{,}8)^2}{2 \cdot 2{,}58^2}} .$$

Визначимо, наприклад, імовірність того, що міцність на стиск буде не менше, ніж 50 МПа:

$$P\left(X \geq 50\right) = 1 - P(X < 50) = 1 - F_0\left(\frac{50 - 52{,}8}{2{,}58}\right) = 1 - F_0\left(-1{,}08\right) = 1 - 0{,}1401 = 0{,}8599 .$$

Зауваження. У прикладі 2.6 теза про теоретичний закон розподілу ознаки була прийнята в першому наближенні. У відповідальних розрахунках потрібна перевірка статистичної гіпотези за так званими критеріями згоди, які будуть наведені далі.

2.3.4 Логарифмічно нормальний розподіл

Якщо кількісна ознака має значно асиметричний розподіл, гіпотеза про нормальний закон є некоректною. Однак нормальному закону можуть підпорядковуватися логарифми значень досліджуваної ознаки. У цьому випадку розподіл ознаки називають *логарифмічно нормальним* (скорочено *логнормальним*)**.** Як приклад на рис. 2.14 наведено розподіли потужності пластів вугілля і логарифмів значень потужності [2].

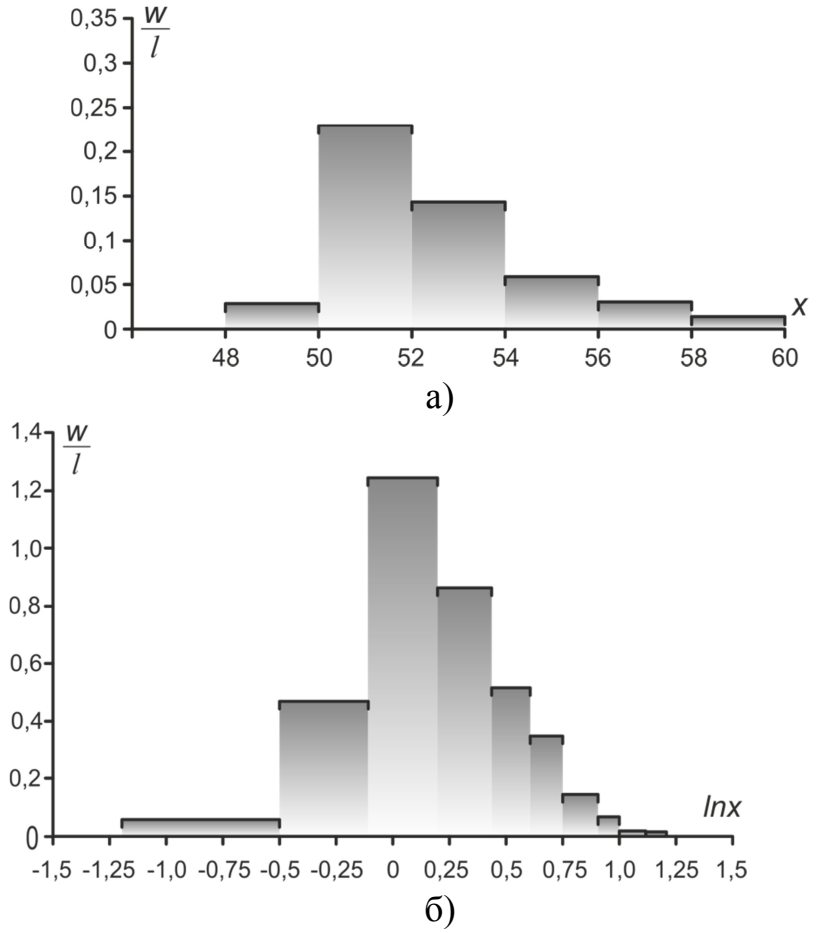


Рис. 2.14. Гістограма відносних частот (а) та логарифмів (б) потужностей пласта

Логарифмічно нормальний розподіл описує випадкову величину *X,* логарифм якої має нормальний розподіл з параметрами *a* і σ. Щільність розподілу величини $Z = \ln(X)$ має вигляд (2.47):

$$f(z)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(Z-a)^2}{2\sigma^2}}.$$

Тоді випадкова величина $X=e^Z$ розподілена за логарифмічно нормальним законом:

$$f(x)=\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(\ln x-a)^2}{2\sigma^2}}. \qquad (2.55)$$

Важливо відзначити, що параметри a і σ для нормально розподіленої величини $Z=\ln(X)$ відповідно дорівнюють математичному сподіванню і середньому квадратичному відхиленню, але для логнормальної величини X це просто параметри форми і масштабу. Однак вони теж пов'язані з числовими характеристиками випадкової величини. Математичне сподівання і дисперсія для логнормальної величини виражаються через параметри розподілу таким чином:

$$M(X)=e^{a+\frac{\sigma^2}{2}}; \qquad (2.56)$$

$$D(X)=e^{2a+\sigma^2}(e^{\sigma^2}-1). \qquad (2.57)$$

Форма кривої щільності логнормального розподілу залежить від співвідношення параметрів a і σ (рис. 2.15). Для неї характерна істотна правобічна асиметрія. Покажемо, як визначити параметри логнормального розподілу a і σ за методом моментів, використовуючи основні співвідношення: $M(X)=\overline{x}^*$, $D(X)=S^2$.

З урахуванням формул (2.56) та (2.57) утворюємо систему відносно двох невідомих a і σ

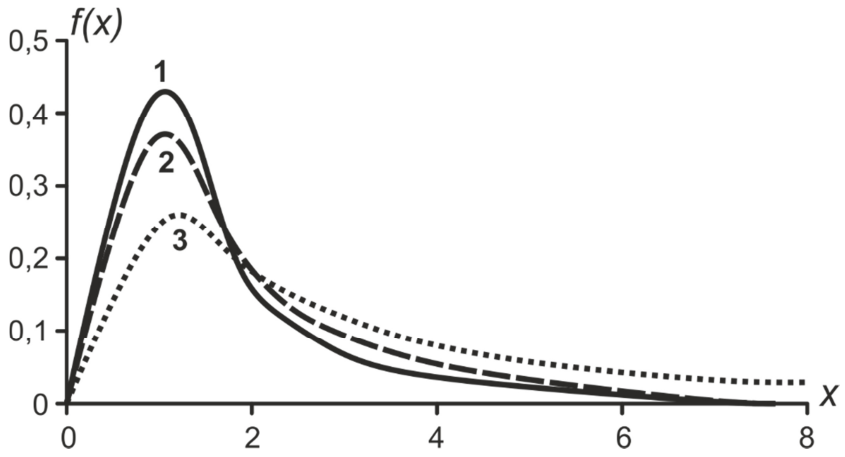
$$\begin{cases}\overline{X}^*=\exp\left(\sigma^2/2+a\right),\\ S^2=\exp\left(\sigma^2+2a\right)\left[\exp\left(\sigma^2\right)-1\right].\end{cases}$$

Логарифмуючи обидві частини рівнянь, виключимо параметр μ з рівнянь та візьмемо до уваги, що $\sqrt{S^2}\big/\overline{x}^2=\eta^2$. Тоді отримаємо:

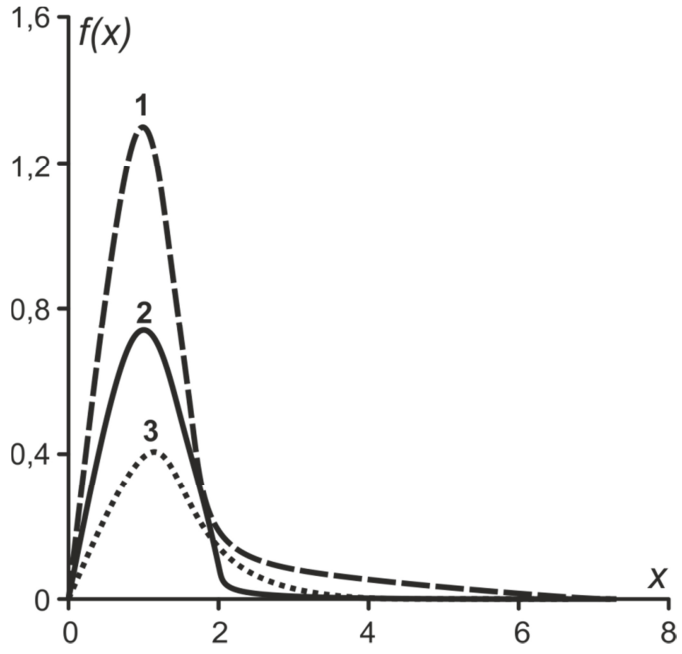
$$\begin{aligned}\sigma&=\sqrt{\ln\left(\eta^2+1\right)},\\ a&=\ln\overline{x}^*-\frac{1}{2}\ln\left(\eta^2+1\right).\end{aligned} \qquad (2.58)$$

Формули (2.58) є остаточними для визначення параметрів логнормального розподілу за вибірковими даними.

Логарифмічно нормальний розподіл ймовірностей є досить широко розповсюдженою статистичною моделлю опису явищ і процесів у науках про Землю. Цим розподілом описується вміст елементів і мінералів у вивержених гірських породах, розміри частинок осадових порід, розміри частинок при подрібненні твердих тіл зосередженою силою, величини граничних руйнівних напружень для деяких типів порід тощо.



а)



б)

Рис. 2.15. Криві щільності логарифмічно нормального розподілу з різними значеннями параметрів: а) $\sigma^2 = 1$ при $a = 0$ (1), $a = 0{,}3$ (2) та $a = 1$ (3); б) $a = 0$ при $\sigma^2 = 0{,}1$ (1), $\sigma^2 = 0{,}3$ (2), $\sigma^2 = 1$ (3)

Значний обсяг випробувань зразків осадових, магматичних і метаморфічних гірських порід на розтягнення–стискання показали, що

найбільш прийнятною статистичною моделлю для межі міцності є саме логарифмічно нормальний розподіл.

Логарифмічно нормальний розподіл можна вивести як статистичну модель випадкової величини, значення якої утворюються в результаті множення великої кількості елементарних помилок аналогічно тому, як нормальний розподіл є моделлю їх суми.

Властивості логарифмічно нормального закону виходять з властивостей відповідного нормального розподілу. Крім того, цей розподіл має найважливішу особливість: розподіл добутку *n* незалежних позитивних випадкових величин з логарифмічно нормальними розподілами знову підпорядковується цьому розподілу. Для нього має місце аналог центральної граничної теореми: розподіл добутку незалежних позитивних випадкових величин при деяких спільних умовах наближається до логарифмічно нормального закону при необмеженому зростанні числа співмножників.

2.3.5 Гамма-розподіл

Диференціальна функція гамма-розподілу має вигляд:

$$f(x)=\begin{cases}\dfrac{\lambda^{\eta}}{\Gamma(\xi)}x^{\xi-1}\cdot e^{-\lambda x}, \ \textit{при} \quad x\ge 0; \quad \lambda>0, \quad \xi>0,\\ 0, \qquad\qquad \textit{в інших випадках.}\end{cases} \tag{2.59}$$

Тут $\Gamma(\xi)$ – гамма-функція, яка визначається як інтеграл

$$\Gamma(\xi)=\int_0^{\infty}x^{\xi-1}\,e^{-x}dx\,. \tag{2.60}$$

Залежно від параметрів η і λ крива $f(x)$ набуває різних форм (рис. 2.16).

Математичне сподівання і дисперсія гамма-розподілу пов'язані з параметрами розподілу співвідношеннями:

$$M(X)=\frac{\xi}{\lambda};\quad D(X)=\frac{\xi}{\lambda^2}. \tag{2.61}$$

При $\xi=1$ гамма-функція дорівнює одиниці:

$$\Gamma(1)=\int_0^{\infty}e^{-x}dx=-e^{-x}\Big|_0^{+\infty}=-\left(\frac{1}{e^{\infty}}-e^{0}\right)=1,$$

тоді гамма-розподіл має вигляд:

$$f(x)=\begin{cases}\lambda e^{-\lambda x} \text{ при } \quad x\geq 0; \quad \lambda>0,\\ 0 \qquad \text{в інших випадках.}\end{cases}$$

Це не що інше, як щільність експоненціального розподілу. Таким чином, експоненціальний розподіл є частковим випадком гамма-розподілу.

Гамма-розподіл, також як і експоненціальний розподіл, має широке застосування в природничих науках. Експериментально встановлено, що таким розподілом можуть описуватися такі величини, як вміст корисного компонента в пробах гірських порід.

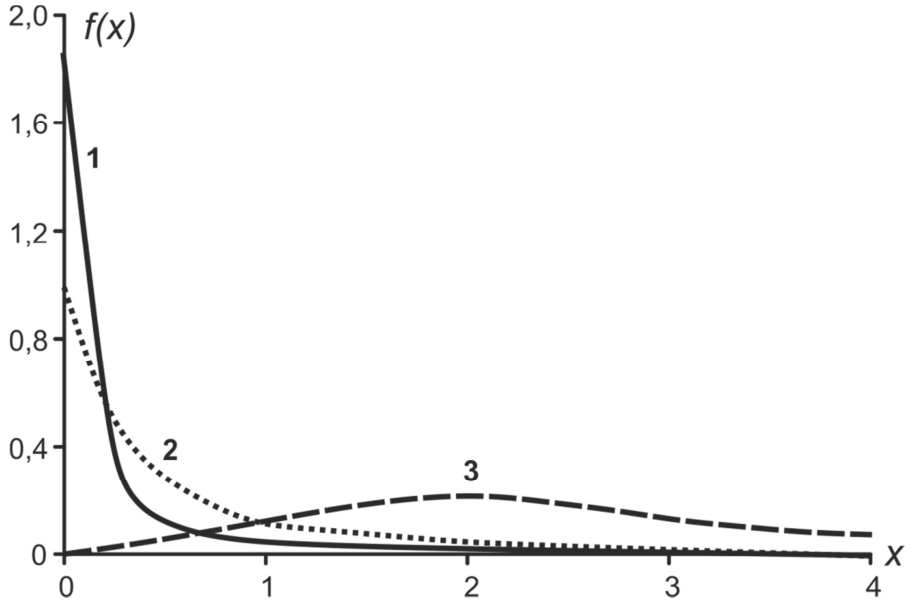


Рис. 2.16. Криві щільності гамма-розподілу з різними параметрами: 1) $\xi = 0{,}5$; $\lambda = 1$; 2) $\xi = 1$, $\lambda = 1$; 3) $\xi = 3$; $\lambda = 1$

2.3.6 Вибір теоретичного розподілу за експериментальними даними

Для вибору теоретичного розподілу за експериментальними даними використовують графік К. Пірсона [12] (рис. 2.17), на якому за віссю абсцис відкладають нормований показник асиметрії

$$\beta_1 = \left(\frac{\mu_3}{\sigma^3}\right)^2, \tag{2.62}$$

а по осі ординат – показник ексцесу

$$\beta_2 = \frac{\mu_4}{\sigma^4}. \tag{2.63}$$

Можна переконатися, що для нормального розподілу $\beta_1 = 0$, а $\beta_2 = 3$. Таким чином, на графіку Пірсона нормальному розподілу відповідає точка з координатами (0, 3). Для експоненціального розподілу $\beta_1 = 4$ і $\beta_2 = 9$ на графіку йому відповідає точка з координатами (4, 9). Для інших розподілів, таких як

логарифмічно нормальний та гамма-розподіл, показники асиметрії й ексцесу змінюються залежно від параметрів розподілу, тому на графіку їм відповідають вже не точки, а лінії. Деякі розподіли займають цілу ділянку. Для того, щоб підібрати розподіл для статистичних даних, емпіричні точки з координатами $(\beta_1^*; \beta_2^*)$, наносять на графік Пірсона. Якщо ці точки групуються поблизу лінії або точки, що відповідає тому чи іншому розподілу, висувають гіпотезу про те, що статистичні дані підпорядковуються саме цьому розподілу. Далі прийняту гіпотезу слід підтвердити за одним з критеріїв згоди, що будуть розглянуті далі.

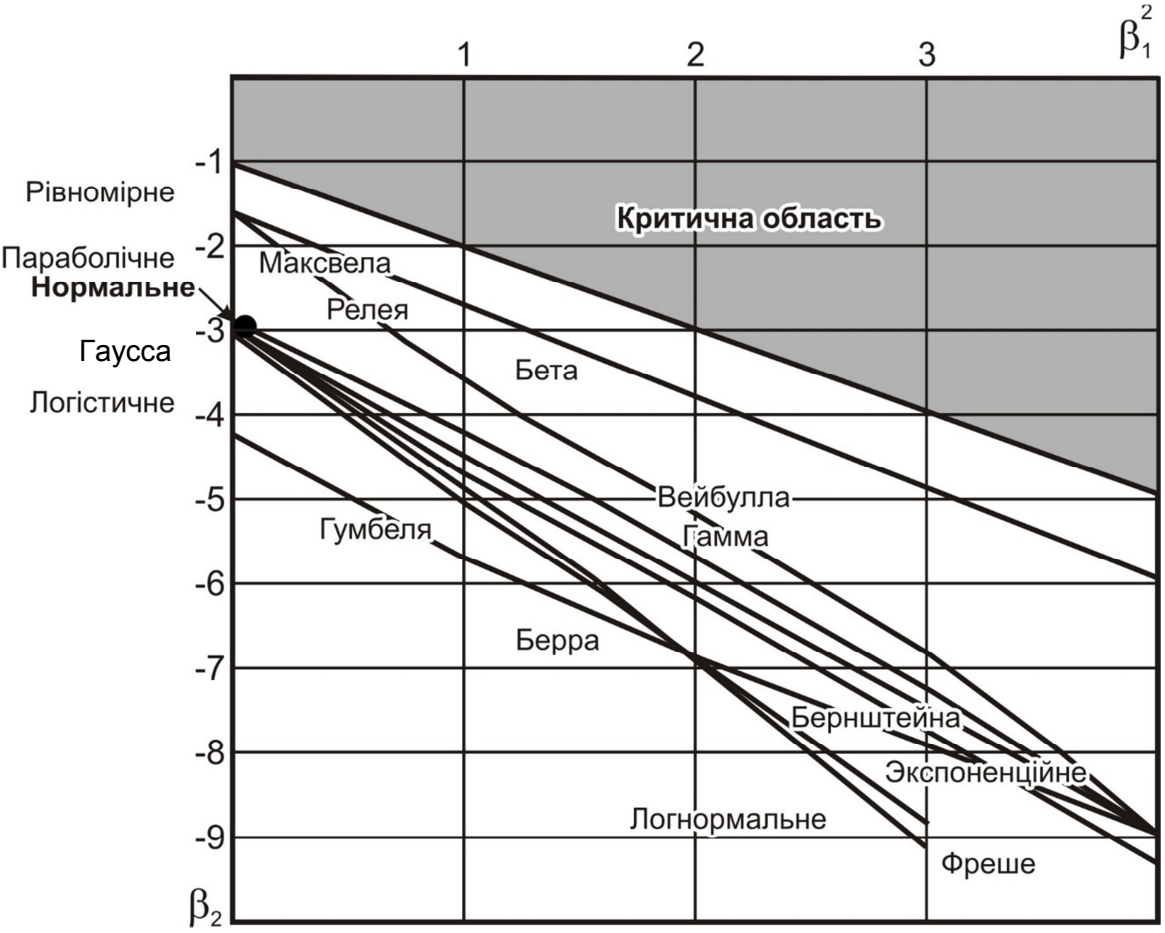


Рис. 2.17. Графік Пірсона для вибору теоретичних розподілів

Приклад 2.7. Підберемо за графіком Пірсона закон розподілу кількісної ознаки, що розглянута в підрозділі 2.2.4 (приклад 2.4). Для наведених статистичних даних нормований показник асиметрії

$$\beta_1 = \left(\frac{\mu_3}{\sigma^3}\right)^2 = \left(\frac{4,72}{2,58^3}\right) = 0,28^2 = 0,0784.$$

Нормований показник ексцесу

$$\beta_2 = \frac{\mu_4}{\sigma^4} = \frac{104,76}{2,58^4} = 2,37 .$$

Таким чином, для даного статистичного розподілу на графіку Пірсона отримаємо точку (0,078, 2,37), що розташована поблизу точки (0, 3), яка відповідає нормальному закону розподілу. Тоді додатково до виконання правила трьох сигм (приклад 2.6) маємо ще одну підставу висунути гіпотезу про підпорядкування розглянутих у прикладі 2.4 статистичних даних нормальному розподілу.

2.3.7 Деякі теоретичні розподіли, що застосовуються в математичній статистиці

Розподіл «хі-квадрат»

Цей розподіл («χ» – грецька буква, вимовляється «хі») тісно пов'язаний з нормальним. Розглядається така величина

$$\chi^2 = \sum_{i=1}^{n} y_i^2 , \tag{2.64}$$

де y_i – незалежні випадкові величини, розподілені за нормальним законом з параметрами $a = 0$ при $\sigma^2 = 1$. При цьому величина χ^2 має свій закон розподілу (так званий «χ^2-квадрат» або «χ^2», вимовляється «хі-квадрат»). Можна довести, що цей закон має щільність розподілу

$$f(x) = \begin{cases} \dfrac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2}, & x > 0, \\ 0, & x \le 0. \end{cases} \tag{2.65}$$

При цьому

$$\nu = n - r, \tag{2.66}$$

де r – число зв’язків, накладених на величину y_i^2. Якщо $y_i = \dfrac{x_i - \overline{x}}{\sigma}$ і приймається, що $\overline{x} = \overline{x}^* = \sum_{i=1}^{n} x_i$, то на y_i^2 накладається один зв'язок і $r = 1$; коли приймається, крім цього, що $\sigma = \sigma^*$, то $r = 2$; параметр ν називається числом степенів вільності.

З формули (2.65) видно, що гамма-розподіл при $\xi = 1/2$ і $\lambda = \nu$ збігається з розподілом χ^2. Зі збільшенням ν розподіл χ^2 повільно наближається до нормального (рис. 2.18).

Можна показати, що

$$M(\chi^2) = \nu, \tag{2.67}$$

$$D(\chi^2) = 2\nu. \tag{2.68}$$

Уперше χ^2-розподіл було розглянуто Р. Хельмертом у 1876 р. і далі розвинуто К. Пірсоном у 1900 р. [1].

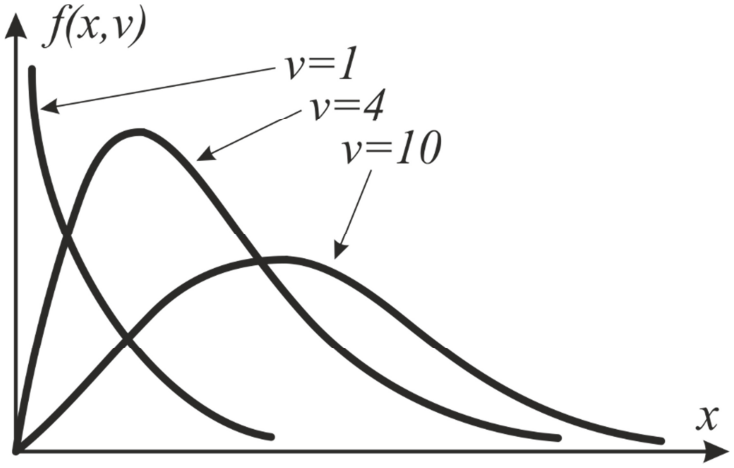


Рис. 2.18. Щільність розподілу χ^2 при різних значеннях параметра ν

Розподіл Стьюдента

Нехай Y – нормальна випадкова величина з $a = 0$ при $\sigma^2 = 1$, Z – незалежна від Y випадкова величина, розподілена за законом χ^2 з ν степенями вільності. Можна довести, що випадкова величина

$$T = \frac{Y}{\sqrt{Z/\nu}} \tag{2.69}$$

розподілена за *законом Стьюдента* із щільністю ймовірності

$$f(x) = \frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma(\nu/2)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}. \tag{2.70}$$

Цей розподіл, як і χ^2, містить один параметр ν – число степенів вільності, що обчислюється за формулою (2.66). Зі зростанням ν розподіл Стьюдента швидко наближується до нормального (рис. 2.19).

Розглянемо тепер деякі дискретні закони розподілу, що використовуються в природничих науках.

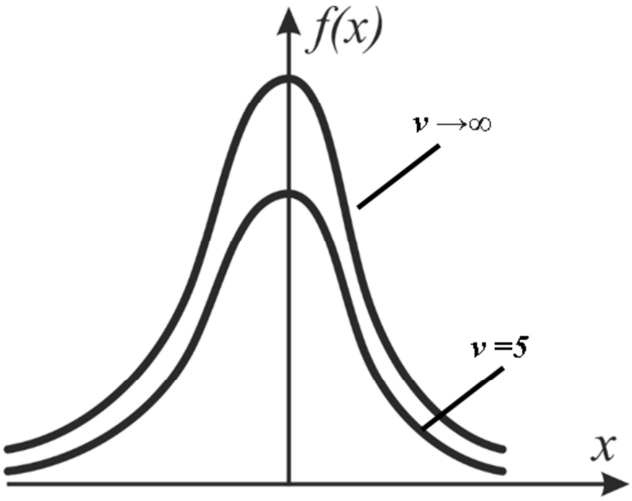


Рис. 2.19. Щільність розподілу Стьюдента

2.3.8 Деякі дискретні розподіли

Біноміальний розподіл описує дискретну випадкову величину, таку як кількість успіхів у послідовності експериментів. Розглянемо події з двома можливими результатами, один з них (А) – з імовірністю p, другий (В) – з імовірністю $q = 1 - p$.

Ймовірність події, коли в n експериментах подія А відбулася k разів, обчислюється за формулою

$$p_n(k) = C_n^k p^k (1-p)^{n-k}, \qquad (2.71)$$

де $C_n^k = \dfrac{n!}{k!(n-k)!}$, $n! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot n$.

Ймовірності, які обчислюються за формулою (2.71) при різних значеннях k, являють собою члени бінома Ньютона, завдяки чому розподіл отримав свою назву.

Випадкова величина X набуває числових значень $k = 0,1,\ldots,n$. Ймовірність події $X = k$ обчислюється за формулою (2.71).

Математичне сподівання $M(X)$ біноміального розподілу дорівнює np, а $D(X)$ дисперсія – $np(1-p)$. Даний розподіл може використовуватися для опису процентного вмісту зерен певної форми та розміру в зразках гірських порід, перевищення допустимих концентрацій у пробах води чи ґрунту та ін.

Розподіл Пуассона

Коли ймовірність p події А мала, а число незалежних експериментів n велике, так що величина $\lambda = n \cdot p$ «не велика і не мала», то ймовірність події,

коли з n разів подія А відбулася k разів, обчислюється за наближеною формулою

$$p_k = \frac{\lambda^k}{k!} e^{-\lambda},\ \lambda > 0\,. \qquad (2.72)$$

Ця формула визначає так званий розподіл Пуассона, при цьому число λ називається параметром розподілу. Числові характеристики величини X, що набуває одне зі значень $k = 0,1,...,n$ та розподілена за законом Пуассона, пов'язані з параметром розподілу таким чином: математичне сподівання – $M(X) = \lambda$, дисперсія – $D(X) = \lambda$.

Контрольні питання до підрозділу 2.3

1. Наведіть графік і аналітичний вираз щільності розподілу експоненціального закону. Яка випадкова величина описується експоненціальним законом розподілу?
2. У чому полягає метод моментів оцінки параметрів розподілу?
3. Наведіть графік і аналітичний вираз щільності розподілу нормального закону. Які випадкові величини описуються нормальним законом розподілу?
4. Як зв'язані параметри розподілу нормального закону з числовими характеристиками випадкової величини?
5. Що таке нормована нормальна величина?
6. Чим відрізняється загальний нормальний розподіл від нормованого нормального розподілу?
7. Як знайти ймовірність потрапляння нормальної випадкової величини в заданий інтервал?
8. Як визначити ймовірність заданого відхилення?
9. У чому полягає правило трьох сигм?
10. Чому дорівнює асиметрія й ексцес нормального розподілу?
11. Назвіть інші закони розподілу, що використовуються для опису кількісних ознак.
12. Як підібрати теоретичний закон розподілу для емпіричних даних?

2.4. Перевірка статистичних гіпотез

Статистичні гіпотези формулюються для перевірки схожості або відмінності певних процесів (об'єктів). Спочатку визначається вид (закон) розподілу, відповідний даній вибірці, після чого перевіряються гіпотези про рівність числових характеристик певного параметра, зокрема про:

1) однорідність досліджуваного об'єкта;

2) рівність середніх значень параметра, отриманого різними методами для того ж об'єкта;

3) рівність дисперсій двох випадкових величин за вибірковими даними.

Наприклад, статистичною гіпотезою може бути твердження щодо однаковості мінералізації у двох різних річках.

Статистична перевірка гіпотез виконується за критеріями згоди. Критерієм згоди називається значення деякої функції $f(\xi_1,\xi_2,\ldots,\xi_n)$ від випадкових величин $\xi_1,\xi_2,\ldots,\xi_n$, що характеризують гіпотезу, яка перевіряється. Функція f має бути випадковою величиною з відомим розподілом, за яким зручно перевірити гіпотезу. Статистична гіпотеза приймається чи відхиляється, якщо значення функції f більше чи менше (залежно від формулювання) за теоретичне значення f для аналогічних умов і заданої ймовірності p. Ця ймовірність відповідає практично неможливій події і називається *рівнем значущості*. Ймовірність $(1 - p)$, що визначає міру правильності висновку щодо гіпотези і відповідає практично достовірній події, називається надійною ймовірністю. В природничих науках часто використовуються надійні ймовірності 95% та рівень значущості 5%.

Якщо справедлива гіпотеза відхиляється, то робиться помилка першого роду, якщо приймається неправильна гіпотеза, то робиться помилка другого роду. Якщо ймовірність помилки другого роду позначити через β, то ймовірність відсутності такої помилки $(1 - \beta)$ називається потужністю даного критерію відносно альтернативної гіпотези.

Наприклад, перевіряється гіпотеза про те, що якість води відповідає санітарним нормам. У разі неправильного висновку щодо перевищення допустимого вмісту токсичних сполук у фактично чистій воді робиться помилка першого роду, у разі неправильного висновку щодо прийнятної якості фактично забрудненої води робиться помилка другого роду. Помилка першого роду призводить до невиправданих економічних втрат, помилка другого роду загрожує здоров'ю людей.

2.4.1 Визначення виду вибіркового розподілу. Критерій «хі-квадрат»

Закон розподілу випадкової величини встановлюється з більшою надійністю при великому обсязі вибірки за умови її однорідності. На основі обробки статистичних даних, графічного зображення статистичного розподілу у вигляді гістограми частот, аналізу точкових оцінок моментів розподілу, а також, виходячи з фізичного змісту досліджуваної кількісної ознаки, можна висунути гіпотезу H (зробити припущення) про те, що випадкова величина X підпорядковується певному закону розподілу. Розподіл, що має випадкова величина за висунутою гіпотезою, називають теоретичним (гіпотетичним).

Для остаточного висновку про закон розподілу необхідно перевірити, наскільки припущення про закон розподілу узгоджується із статистичними даними, тобто з експериментом. І навіть, якщо зроблене припущення про закон розподілу є правильним, експериментальний (статистичний) закон розподілу буде в якійсь мірі відрізняться від теоретичного. Тому потрібно вирішити задачу: чи є розходження між статистичним $f^*(x)$ та теоретичним $f(x)$ законами розподілу наслідком обмеженої кількості спостережень, тобто випадковим, чи це розходження є суттєвим і пов'язане з тим, що фактичний розподіл випадкової величини X відрізняється від гіпотетичного. Тому необхідні критерії, які дозволяють вирішити, чи узгоджуються значення випадкової величини X: $x_1, x_2, ..., x_n$ з гіпотезою щодо її щільності розподілу. Такі критерії називають критеріями згоди.

Покажемо далі порядок застосування критерію χ^2, запропонованого К. Пірсоном. Відповідно цьому критерію порівнюються емпіричні частоти та теоретичні ймовірності. Розглянемо величину:

$$\chi^2 = \sum_{i=1}^{l} \frac{(m_i - np_i)^2}{np_i}, \tag{2.73}$$

де l – кількість розрядів даного інтервального ряду; n – обсяг вибірки; m_i – інтервальні частоти (емпіричні дані); p_i – імовірності потрапляння випадкової величини в i-й проміжок, що обчислюються відповідно до гіпотетичної функції $f(x)$:

$$p(x_{i-1} \le X \le x_i) = \int_{x_{i-1}}^{x_i} f(x)dx. \tag{2.74}$$

Величина (2.73) є випадковою, тому що у різних випробуваннях вона набуває різних, заздалегідь невідомих значень. Зрозуміло, що чим менше відрізняються емпіричні та теоретичні частоти, тим менше величина (2.73).

Якщо закон розподілу критерію згоди відомий, можна вказати таке критичне значення критерію $\chi^2 = \chi^2_{кр,пр}$ або $\chi^2 = \chi^2_{кр,лв}$ (рис. 2.20), при якому відповідна ймовірність події $\chi^2 > \chi^2_{кр,пр}$ або $\chi^2 < \chi^2_{кр,лв}$ буде близькою до нуля.

Геометрично ці ймовірності являють собою площі під кривою розподілу, так звані *правосторонню* та *лівосторонню* критичні області. Ймовірність потрапляння в ці області в одиничному випробуванні дуже мала, близька до нуля. Цю ймовірність відповідно до змісту задачі зазвичай приймають такою, що дорівнює 0,01; 0,025; 0,05. Як зазначено вище, таку ймовірність (що відповідає практично неможливій події) називають *рівнем значущості* та позначають α.

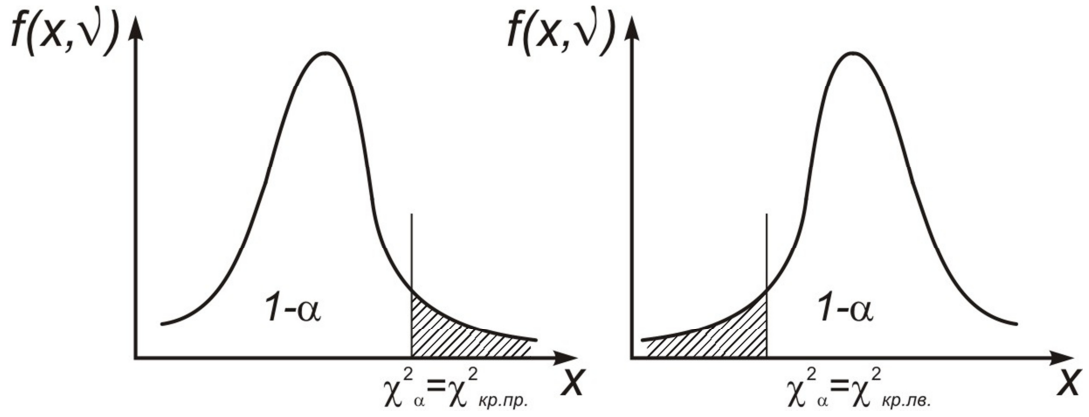


Рис. 2.20. Критичні значення критерію χ^2

Оцінювання гіпотези за критерієм полягає в порівнянні величини χ^2, яка визначена відповідно до формули (2.73) з критичним значенням ($\chi^2 = \chi^2_{кр,пр}$ або $\chi^2 = \chi^2_{кр,лв}$), що визначається залежно від обраного рівня значущості α. Якщо $\chi^2 > \chi^2_{кр,пр}$ або $\chi^2 < \chi^2_{кр,лв}$, то значення досліджуваної величини *X* потрапляє в критичну область (відповідно праву або ліву) і обрану гіпотезу про вид функції $f(x)$ слід відкинути. Якщо, навпаки, $\chi^2 < \chi^2_{кр,пр}$ або $\chi^2 > \chi^2_{кр,лв}$, то величина χ^2 знаходиться поза критичною областю і гіпотеза може бути прийнята.

Таким чином, для перевірки гіпотези необхідно визначити закон розподілу випадкової величини χ^2. Доведено, що при $n \to \infty$ закон розподілу випадкової величини (2.73), незалежно від закону розподілу генеральної сукупності *X*, наближається до розподілу χ^2 з ν степенями вільності (див. підрозділ 2.3.7). Кількість степенів вільності знаходять за формулою $\nu = l - r$, де *r* – кількість зв'язків, що накладені на частоти m_i. Наприклад, якщо $\sum_{i=1}^{l} m_i = n$ та в ролі теоретичних параметрів $\bar{x}$ і σ, які входять до теоретичної функції $f(x)$, приймаються їх точкові оцінки $\bar{x}^*$ і *S*, то приймають $r = 3$. Критичні значення величини χ^2 наведені в таблиці додатку Б залежно від значень α та ν.

Розрахунки виконують у такій послідовності:

– висувають гіпотезу про закон розподілу *f*(*x*), за вибіркою визначають параметри розподілу (наприклад, $a = \bar{x}^*$ та $\sigma = S$, якщо висунута гіпотеза про нормальний закон розподілу), а також кількість степенів вільності $\nu = l - r$ ($r = 3$ для нормального закону розподілу);

– обирають рівень значущості α; за таблицею для даних α та ν знаходять критичне значення критерію $\chi^2_{кр,пр}$;

– за формулою (2.74) визначають ймовірності p_i потрапляння випадкової величини X до інтервалів (x_{i-1}, x_i);

– за формулою (2.73) обчислюють емпіричне значення χ^2;

– порівнюють значення χ^2 з критичним.

У результаті порівняння:

а) у випадку правосторонньої критичної області гіпотеза, що висувалася, відкидається, якщо $\chi^2 > \chi^2_{кр,пр}$;

б) у випадку лівосторонньої критичної області гіпотеза, що висувалася, відкидається, якщо $\chi^2 < \chi^2_{кр,лв}$;

в) у випадку двосторонньої критичної області гіпотеза, що висувалася, буде *правильною,* якщо виконується умова:

$$p\left(\chi^2 < \chi^2_{кр.лв}\right) + p\left(\chi^2 > \chi^2_{кр.пр}\right) = \alpha. \qquad (2.75)$$

Очевидно, що критичні точки можуть бути вибрані нескінченною множиною способів. Якщо розподіл критерію симетричний відносно нуля і є підстави обирати симетричні відносно нуля точки $\chi^2_{кр,лв} < 0$ і $\chi^2_{кр,пр} > 0$, то приймаємо

$$p\left(\chi^2 < \chi^2_{кр.лв}\right) = p\left(\chi^2 > \chi^2_{кр.пр}\right).$$

Враховуючи (2.75), одержимо:

$$p\left(\chi^2 < \chi^2_{кр.лв}\right) = p\left(\chi^2 > \chi^2_{кр.пр}\right) = \frac{\alpha}{2}.$$

У додатку Б наведені лише «праві» критичні точки. Знайти «ліву» точку можна з урахуванням того, що події $\chi^2 > \chi^2_{кр,пр}$ і $\chi^2 < \chi^2_{кр,лв}$ протилежні та сума їх ймовірностей дорівнює 1:

$$p\left(\chi^2 < \chi^2_{кр.лв}\right) = 1 - p\left(\chi^2 > \chi^2_{кр.пр}\right) = 1 - \frac{\alpha}{2}. \qquad (2.76)$$

Приклад 2.8. Перевіримо гіпотезу про нормальний закон розподілу для даних, що подані вибіркою обсягом $n = 40$ та згруповані в інтервальний ряд (табл. 2.6). В прикладі 2.4 розраховані параметри статистичного розподілу цієї кількісної ознаки: середня вибіркова $\overline{x}^* = 52{,}85$ та середнє квадратичне відхилення $S = 2{,}58$. В прикладі 2.5 висунута гіпотеза про підпорядкування цих даних нормальному закону розподілу і визначені параметри цього розподілу

$a = \bar{x}^* = 52{,}8$; $\sigma = S = 2{,}58$. Знаходимо число степенів вільності $\nu = l - r = 5 - 3 = 2$. Розглянемо двосторонню критичну область за таблицею (додаток Б) для $\nu = 2$ і $\alpha/2 = 0{,}025$ і визначимо, що $\chi^2_{кр,пр} = 7{,}4$, а для $\left(1 - \frac{\alpha}{2}\right) = 0{,}975$ і того самого ν знаходимо $\chi^2_{кр,лв} = 0{,}051$.

Знайдемо тепер ймовірності p_i потрапляння значення випадкової величини в інтервали (x_{i+1}, x_i). Скористаємося функцією Лапласа відповідно до формули (2.51):

$$P_i = \Phi\left(\frac{x_i - a}{\sigma}\right) - \Phi\left(\frac{x_{i-1} - a}{\sigma}\right).$$

Отримаємо:

$$P_1 = \Phi\left(\frac{50 - 52{,}8}{2{,}58}\right) - \Phi\left(\frac{48 - 52{,}8}{2{,}58}\right) = \Phi(-1{,}08) - \Phi(-1{,}86) = -0{,}3599 + 0{,}4686 = 0{,}11;$$

$$P_2 = \Phi\left(\frac{52 - 52{,}8}{2{,}58}\right) - \Phi\left(\frac{50 - 52{,}8}{2{,}58}\right) = \Phi(-0{,}31) - \Phi(-1{,}08) = -0{,}1217 + 0{,}3599 = 0{,}24;$$

$$P_3 = \Phi\left(\frac{54 - 52{,}8}{2{,}58}\right) - \Phi\left(\frac{52 - 52{,}8}{2{,}58}\right) = \Phi(0{,}46) - \Phi(-0{,}27) = 0{,}1772 + 0{,}1217 = 0{,}3;$$

$$P_4 = \Phi\left(\frac{56 - 52{,}8}{2{,}58}\right) - \Phi\left(\frac{54 - 52{,}8}{2{,}58}\right) = \Phi(1{,}24) - \Phi(0{,}5) = 0{,}3925 - 0{,}1772 = 0{,}22;$$

$$P_5 = \Phi\left(\frac{58 - 52{,}8}{2{,}58}\right) - \Phi\left(\frac{56 - 52{,}8}{2{,}58}\right) = \Phi(2{,}01) - \Phi(1{,}28) = 0{,}4772 - 0{,}3925 = 0{,}09;$$

$$P_6 = \Phi\left(\frac{60 - 52{,}8}{2{,}58}\right) - \Phi\left(\frac{58 - 52{,}8}{2{,}58}\right) = \Phi(2{,}79) - \Phi(2{,}05) = 0{,}4974 - 0{,}4772 = 0{,}02.$$

Результати розрахунків подані в табл. 2.9.

Таблиця 2.9

Обчислення для перевірки вибірки за критерієм χ^2

Інтервали міцності зразка на стиск $u_{i-1} - u_i$, МПа	Значення середини інтервалу u_i, МПа	Частота m_i	Відносна частота w_i	Теоретичні ймовірності
48 – 50	49	6	0,150	0,11
50 – 52	51	9	0,225	0,24
52 – 54	53	13	0,325	0,30
54 – 56	55	7	0,175	0,22
56 – 58	57	4	0,100	0,09
58 – 60	59	1	0,025	0,02

За формулою (2.73) визначимо емпіричне значення критерію:

$$(\chi)^2 = \sum_{i=1}^{l} \frac{(m_i - np_i)^2}{np_i} = \sum_{i=1}^{6} \frac{(m_i - 40p_i)^2}{40p_i} =$$

$$= \frac{(6-4,4)^2}{4,4} + \frac{(9-9,6)^2}{9,6} + \frac{(13-12)^2}{12} + \frac{(7-8,8)^2}{8,8} + \frac{(4-3,6)^2}{3,6} + \frac{(1-0,8)^2}{0,8} =$$

$$= 0,58 + 0,037 + 0,08 + 0,37 + 0,04 + 0,05 = 1,157.$$

У даному випадку $\chi^2_{кр.лв} < \chi^2 < \chi^2_{кр.пр}$ $(0,051 < 1,409 < 7,4)$ Тому можна стверджувати, що гіпотеза про розподіл випадкової величини, що спостерігається X за нормальним законом з параметрами a = 52,8 і $\sigma = 2,58$ не суперечить статистичним даним.

Вище зазначалось, що *статистичні гіпотези* використовують для перевірки схожості певних процесів або об'єктів. Наприклад, практичне застосування має перевірка гіпотези про рівність середніх значень параметра, отриманого різними методами для одного й того ж об'єкта.

2.4.2 Гіпотеза про рівність середніх значень

Перевірка цієї гіпотези виконується за критеріями Стьюдента, Ван-дер-Вардена, Вілкоксона та іншими [6, 7]. Розглянемо детальніше критерій Стьюдента. Позначимо через x_1, x_2,…,x_n та y_1, y_2,…,y_m значення вибіркових даних на різних об'єктах або такі, що отримані різними методами. Вважається, що розподіли цих величин є нормальними, n та m – обсяги вибірок. Тоді величина

$$t = \frac{\left|\overline{x} - \overline{y}\right|}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} \qquad (2.77)$$

описується розподілом Стьюдента з $n + m - 2$ степенями вільності. Тут $\overline{x}$ і $\overline{y}$ – вибіркові оцінки середнього значення, S_x^2 і S_y^2 – вибіркові оцінки дисперсії. Перевірка гіпотези про рівність середніх двох вибірок полягає в обчисленні величини t та її порівнянні з табличним значенням для даної кількості степенів вільності й надійної ймовірності (додаток В). Якщо значення t перевищує табличне, то гіпотеза про рівність вибіркових середніх відхиляється, тобто різниця між середніми значеннями статистично значуща.

Приклад 2.9. У табл. 2.10 наведені значення вимірів у двох водоносних горизонтах: у першому від поверхні – ґрунтовому (свердловини № 2, 22, 6675, 5761, 6708, 6711, 6717) та неогеновому – другому від поверхні (усі інші

свердловини). За цими даними визначені середні значення та дисперсії вмісту сульфатів та сухого залишку, що наведені в табл. 2.11. Як бачимо, середні значення цих показників помітно вище у ґрунтовому горизонті.

Згідно з формулою (2.77) значення величини t для сульфатів становить 1,86, для сухого залишку – 2,03. Для надійної ймовірності 90 % критичне значення критерію Стьюдента при 7 + 13 – 2 = 18 степенях вільності дорівнює 1,73, а для надійної ймовірності 95 % і 18 степенях вільності – 2,10.

Отже, при «м'якому» критерії з ймовірністю помилки 10 % можна стверджувати, що вміст сульфатів та сухого залишку у водах ґрунтового горизонту суттєво вищий, ніж у неогеновому, у разі більш «строгого» критерію з імовірністю помилки 5 % ця гіпотеза відкидається, тобто слід вважати, що досліджувані показники приблизно однакові в обох водоносних горизонтах.

Критерії для перевірки статистичних гіпотез про однорідність досліджуваного об'єкта та про рівність дисперсій двох випадкових величин будуть розглянуті нижче у розділі 3.

Таблиця 2.10

Вміст сульфатів (C_{SO4}, мг/л) та сухого залишку ($C_{схз}$, мг/л) у пробах підземних вод зі свердловин

№ св.	2	22	6675	5761	6708	6711	6717	2а	3а	3р
C_{SO4}	240	768	120	1344	336	101	960	144	240	240
$C_{схз}$	811	1626	326	1856	629	578	2745	470	482	638

№ св.	4р	5р	7р	8р	9р	10р	13р	16	17р	18р
C_{SO4}	186	120	50	664	359	168	168	48	144	154
$C_{схз}$	450	479	271	1287	724	515	436	323	420	464

Таблиця 2.10

Параметри статистичного розподілу вмісту сульфатів та сухого залишку

Водоносний горизонт	Сульфати		Сухий залишок	
	Середнє, мг/л	Дисперсія	Середнє, мг/л	Дисперсія
Ґрунтовий	552,7	229068	1224,4	770464
Неогеновий	205,8	25266	535,3	64361

Контрольні питання до підрозділу 2.4

1. Що називають критерієм згоди? Як приймається чи відхиляється статистична гіпотеза на підставі критерію згоди?

2. Що називають рівнем значущості та надійною ймовірністю?

3. Що називається помилками першого та другого роду? Наведіть приклади.

4. Який критерій використовують для перевірки гіпотези про вид закону розподілу? Які величини порівнюються при перевірці гіпотези про вид закону розподілу?

5. Що називають критичною областю критерію?

6. Як виконується перевірка гіпотези про рівність середніх двох вибірок на основі критерію Стьюдента?

2.5. Визначення зв'язків між експериментальними даними

Зв'язок між показниками стану природного середовища, параметрами гірських порід тощо може бути функціональним або стохастичним. Функціональний зв'язок означає, що існує однозначна відповідність значень одного параметра значенням іншого параметра. Але переважна більшість показників стану навколишнього середовища є результатом ланцюга складних, часто взаємопов'язаних процесів, тому встановити такі залежності практично неможливо. Крім того, слід враховувати помилки в експериментальних даних за рахунок методики досліджень, похибок вимірювань, відбору, зберігання та аналізу проб тощо. Таким чином, через неможливість точного опису всіх причинно-наслідкових ланцюгів виникають імовірнісні зв'язки між параметрами, які досліджуються за допомогою розподілів випадкових величин.

У разі стохастичного зв'язку одному значенню показника відповідає декілька значень іншого, причому з різною ймовірністю. Для гідрохімічних показників характерні саме стохастичні зв'язки, оскільки, наприклад, концентрація деякого компонента у природних водах залежить як від чинників, характерних лише для нього, так і від чинників загальних для всіх компонентів. Кількісно зв'язки між показниками природничих процесів характеризуються тенденціями або трендами.

Зокрема, у гідрохімічних дослідженнях (наприклад, при встановленні зв'язку між мінералізацією підземних та річкових вод, вмістом певних компонентів у природних водах тощо) необхідно виявити саме таку тенденцію або декілька трендів, довести її статистичну значущість, зробити прогноз впливу однієї величини на іншу. Якщо відома поведінка одного з параметрів, то можна передбачити поведінку іншого, якщо з ним є певний зв'язок. Крім того, можна підібрати ознаки для проведення класифікації. Для цього й використовуються методи кореляційного аналізу.

Стохастична залежність між величинами. Кореляційний зв'язок. Випадкові величини X і Y називають *стохастично залежними*, якщо зміна однієї з них призводить до зміни розподілу другої. Зокрема, може змінюватися той чи інший параметр розподілу. Якщо зі зміною однієї випадкової величини зміщається центр розподілу другої, тобто її середнє значення, то стохастичний зв'язок між величинами зветься *кореляційним*.

Наведемо приклад випадкової величини Y, яка не зв’язана з величиною X функціонально, але зв’язана кореляційно. Нехай Y – об’єм видобутку вугілля, X – кількість комбайнів. З однакових за площею дільниць при рівних кількостях техніки й обслуговуючого персоналу, однаковій організації праці видобувають різну кількість вугілля на місяць, тобто Y не є функцією від X. Це зумовлено впливом випадкових факторів: варіацією фізико-механічних властивостей порід, зміною потужності вугільного пласта тощо. Разом з цим, як показує досвід, середній видобуток вугілля є функцією кількості задіяних комбайнів, тобто Y пов’язана з X кореляційно.

При стохастичному зв’язку розрізняють *кореляцію*, яка визначає, чи існує взаємозв’язок між X та Y і наскільки він суттєвий, та *регресію*, що конкретизує, яка саме залежність існує між величинами і чи можна оцінити Y по X?

Для прикладу розглянемо табл. 2.12. Числа n_{xy} у комірках таблиці показують число появ тієї чи іншої комбінації значень X і Y, де n – об’єм вибірки.

Таблиця 2.12

Приклад таблиці кореляційного зв'язку між величинами

X \ Y	30	35	40	45	50	55	n_y
18	4	6	-	-	-	-	10
28	-	8	10	-	-	-	18
38	-	-	4	35	5	-	44
48	-	-	4	12	6		22
58	-	-		1	3	2	6
n_x	4	14	18	48	14	2	100

Кожний рядок табл. 2.12 є *умовним розподілом* величини Y при заданому значенні X. Для кожного умовного розподілу можна визначити умовне середнє

$$\overline{y}_x = \frac{\sum_{k=1}^{m} n_{xy_k} y_k}{n_x}, \qquad (2.78)$$

де m – число рядків. Значення умовних середніх зведені у табл. 2.13.

Таблиця 2.13

Умовні середні, обчислені за формулою (2.78)

x	30	35	40	45	50	55
$\overline{y}_x$	18	23,7	34,67	41	46,5	58
n_x	4	14	18	48	14	2

Як бачимо, кожному значенню x відповідає одне значення $\overline{y}_x$, тобто залежність $\overline{y}_x$ від x можна вважати функціональною:

$$\overline{y}_x = f(x). \qquad (2.79)$$

Рівняння виду (2.79) визначає кореляційну залежність Y від X, його називають рівнянням регресії Y по X. Теоретичною лінією регресії служить графік функції $f(x)$. Аналогічно рівняння регресії X по Y записується у вигляді

$$\overline{x}_y = \phi(y). \qquad (2.80)$$

Якщо обидва рівняння (2.79) та (2.80) лінійні, то кореляцію між X і Y називають лінійною. В протилежному випадку говорять про нелінійну кореляцію.

Як зазначено вище, основне завдання регресійного аналізу – встановлення форми кореляційного зв'язку, тобто виду функції регресії (лінійна, квадратична, показова і т. д.).

Позначимо через x_i значення незалежної змінної (наприклад, часу, довжини) і через y_i – значення вимірюваної величини, одержані при відповідних x_i. Необхідно за парами даних (x_i, y_i), i=1,..,n, (n – кількість вимірювань) побудувати залежність $y(x)$, яка би найточніше описувала результати експерименту, при цьому мінімально відхилялася б від експериментальних значень. Для цього можна використовувати різні види залежностей. За рівнянням $y(x)$ можна передбачити значення вимірюваної величини поза межами інтервалу $[x_0, x_n]$.

Лінійна регресія. Якщо які-небудь теоретичні передумови відсутні, то можна припустити, що $f(x)$ – лінійна функція, тоді рівнянні регресії має вигляд:

$$y = b_0 + b_1 x, \qquad (2.81)$$

де параметри b_0 та b_1 – коефіцієнти регресії, які треба знайти за вибірковими даними.

Нагадаємо, що в рівнянні прямої лінії вільний член b_0 дорівнює відстані від початку координат до точки перетину прямої з віссю Oy. Коефіцієнт b_1 дорівнює тангенсу кута нахилу прямої до осі Ox (рис. 2.21). Фактично пряма проводиться через «хмару» точок, що нанесені на площину xOy і являють собою *кореляційне поле.* Якщо зв'язок функціональний (детермінований), то всі точки лежать на одній прямій (лінії регресії). Інакше зв'язок між x_i і y_i є стохастичним (імовірнісним).

Позначимо через x_i значення величини X і через y_i – значення Y, що отримані при відповідних x_i. Необхідно за парами даних (x_i, y_i), i=1,..,n (n – кількість вимірювань), побудувати залежність $y(x)$, яка має вигляд (2.81) і при цьому мінімально відхиляється від експериментальних значень.

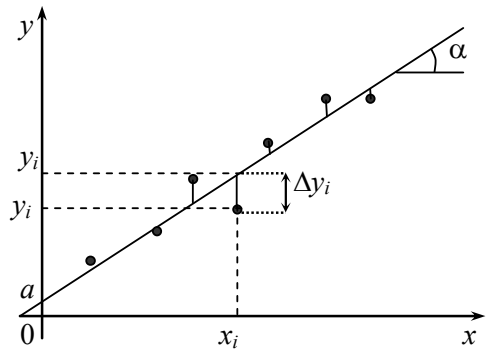


Рис. 2.21. Кореляційне поле та лінія регресії

Лінія регресії проводиться так, щоб сума квадратів відхилень емпіричних значень функції $y_i(x_i)$ від теоретичного значення $y(x_i) = b_0 + b_1x_i$ була б мінімальною. Отже, задача зводиться до мінімізації функції, що є сумою відхилень (нев'язок):

$$U(b_0,b_1)=\sum_{i=1}^{n}\left(y_i-\left(b_0+b_1x_i\right)\right)^2=\sum_{i=1}^{n}\left(\Delta y_i\right)^2\to\min. \qquad (2.82)$$

Мінімум функції двох змінних $U(b_0,b_1)$ досягається в точках, де її частинні похідні дорівнюють нулю. Тоді можна скласти систему двох рівнянь відносно невідомих параметрів b_0 і b_1:

$$\frac{\partial U}{\partial b_0}=0,\quad \frac{\partial U}{\partial b_1}=0.$$

Частинні похідні запишемо формулами:

$$\frac{\partial U}{\partial b_0}=-2\sum_{i=1}^{n}(y_i-(b_1x_i+b_0));\quad \frac{\partial U}{\partial b_1}=-2\sum_{i=1}^{n}(y_i-(b_1x_i+b_0)x_i).$$

Тоді отримаємо систему рівнянь відносно невідомих параметрів лінійної функції b_0 та b_1:

$$\begin{cases} nb_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \\ b_0 \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} x^2{}_i = \sum_{i=1}^{n} x_i y_i. \end{cases} \tag{2.83}$$

За допомогою формул Крамера отримаємо розв'язок цієї системи і знайдемо b_0 і b_1

$$b_0 = \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}, \; b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}. \tag{2.84}$$

Крім прямої регресії, може бути побудована зворотня регресія як залежність $x(y)$. Вибір типу регресії визначається тим, яка з величин розглядається як аргумент, тобто є незалежною змінною, або зручністю побудови апроксимуючої залежності. Формули для визначення коефіцієнтів зворотної регресії будуть аналогічні з відповідною заміною індексів.

Приклад 2.10. Розглянемо використання методу найменших квадратів на простому прикладі, де кожному значенню X відповідає тільки одне значення Y (табл. 2.14).

Розрахунки, наведені в табл. 2.14, використаємо для отримання коефіцієнтів системи рівнянь (2.83):

$$\begin{cases} 10b_0 + 55b_1 = 167, \\ 55b_0 + 385b_1 = 1086. \end{cases}$$

Розв'язком системи є значення коефіцієнтів $b_0 = 5{,}53$; $b_1 = 2{,}03$. Таким чином, рівняння регресії має вигляд: $y = 5{,}53 + 2{,}03\,x$.

Таблиця 2.14

Дані та результати обчислень до прикладу 2.10

	x	y	x^2	y^2	xy	$f(x)$	$(f(x)-y)^2$	$\left(y-\bar{y}\right)^2$	$\left(f(x)-\bar{y}\right)^2$
	1	**7**	1	49	7	7,5636	0,3176	94,0900	83,4738
	2	**10**	4	100	20	9,5939	0,1649	44,8900	50,4967
	3	**9**	9	81	27	11,6242	6,8864	59,2900	25,7637
	4	**15**	16	225	60	13,6545	1,8104	2,8900	9,2751
	5	**17**	25	289	85	15,6848	1,7298	0,0900	1,0306
	6	**20**	36	400	120	17,7151	5,2208	10,8900	1,0304

Продовження табл. 2.14

	x	y	x^2	y^2	xy	$f(x)$	$(f(x)-y)^2$	$\left(y-\bar{y}\right)^2$	$\left(f(x)-\bar{y}\right)^2$
	7	**18**	49	324	126	19,7454	3,0464	1,6900	9,2745
	8	**23**	64	529	184	21,7757	1,4989	39,6900	25,7627
	9	**23**	81	529	207	23,806	0,6496	39,6900	50,4952
	10	**25**	100	625	250	25,8363	0,6994	68,8900	83,4720
Сума	**55**	**167**	385	3151	1086		22,0242	362,1000	340,0747
	$\bar{x}$ =5,5	$\bar{y}$ =16,7							

Оцінку практичної значущості отриманої функціональної залежності проводять за допомогою *індексу кореляції*, який характеризує силу зв'язку між змінними і визначається формулою

$$R=\sqrt{\frac{S^2_{факт}}{S^2_{заг}}}. \qquad (2.85)$$

Тут $S^2_{заг}=\frac{1}{n-1}\sum(y_i-\bar{y})^2$ – загальна дисперсія;

$S^2_{факт}=\frac{1}{n-1}\sum_{i=1}^{n}\left(f(x_i)-\bar{y}\right)^2$ – факторна (регресійна) дисперсія.

Величина індексу кореляції знаходиться в межах від нуля до одиниці. Чим ближче його значення до 1, тим тісніше розглянутий зв'язок. Використовують також квадрат індексу кореляції – *індекс детермінації* R^2 (*коефіцієнт достовірності*), який характеризує частину загальної варіації ознаки *Y*, яка пояснюється фактором *X*. Попередні висновки щодо щільності кореляційного зв'язку на основі індексу кореляції можна зробити за шкалою Чеддока [1] (табл. 2.15).

Таблиця 2.15

Якісна оцінка сили зв'язку за шкалою Чеддока

Інтервал, якому належить значення індексу кореляції	Сила зв'язку
0,1 – 0,3	Відсутня – дуже слабка
0,3 – 0,5	Слабка – помірна
0,5 – 0,7	Помітна – середня
0,7 – 0,9	Тісна – висока
0,9 – 0,99	Дуже висока – функціональна

У наведеному прикладі 2.10, користуючись результатами табл. 2.14 (два останні стовпчики), легко підрахувати факторну (регресійну) та загальну дисперсії: $S^2_{факт}$ = 37,7861; $S^2_{заг}$ = 40,2333. Таким чином, індекс кореляції складає

$R = 0{,}96911$, а індекс детермінації – $R^2 = 0{,}9392$; виходячи з цього величини Y та X мають дуже сильний зв'язок.

У даному випадку ми виконали апроксимацію даних, тобто наближений опис масиву функцією заданого вигляду. З рис. 2.22 видно, що апроксимація даних у програмі Microsoft Excel (майстер діаграм, побудова лінії тренду) дає аналогічний результат.

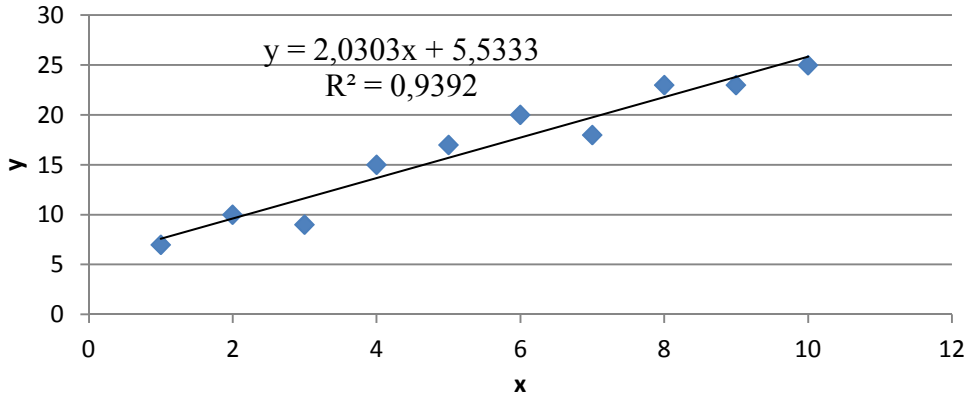


Рис. 2.22. Апроксимація даних з прикладу 2.10 за допомогою майстра діаграм Microsoft Excel

Ступінь лінійної залежності між двома змінними оцінюють також *коефіцієнтом кореляції Пірсона*. Він визначається через *коваріацію* (*кореляційний момент)*, що являє собою математичне сподівання добутків відхилень компонент від їхніх математичних сподівань. Для двох рядів незалежно отриманих результатів вимірювань x_i та y_i, $i=1,\ldots,n$, можна визначити коваріацію у вигляді

$$\operatorname{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}). \qquad (2.86)$$

Для тих же рядів даних коефіцієнт кореляції визначається у вигляді

$$r = \frac{\operatorname{cov}(x,y)}{\sigma_x \sigma_y}. \qquad (2.87)$$

Після спрощень отримаємо формулу

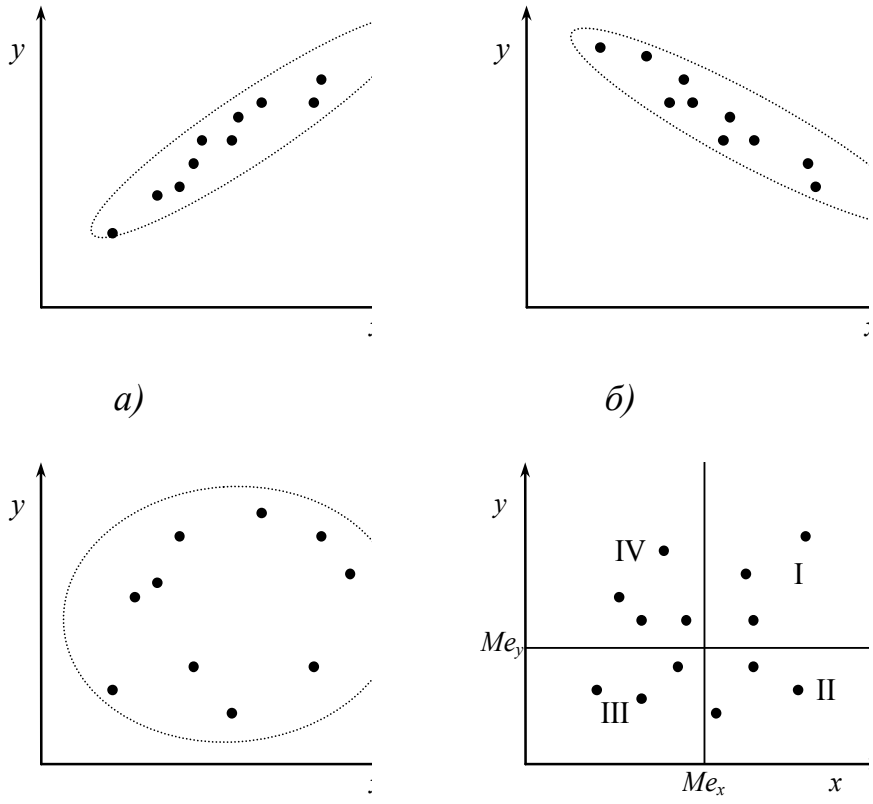
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}. \qquad (2.88)$$

Якщо існує *функціональна* залежність Y від X у вигляді $y = b_0 + b_1 x$ з коефіцієнтом $b_1 > 0$, то коефіцієнт кореляції $r = 1$. Якщо ж коефіцієнт $b_1 < 0$, то $r = -1$, тобто між X та Y існує зворотний зв’язок. У разі $r = 0$ залежність між X

та *Y* відсутня. Отже, коефіцієнт кореляції змінюється в діапазоні [–1, 1] за будь-яких результатів вимірювань. Якщо *r* ближче до 1 або до –1, але його модуль менше за граничні значення, то можна стверджувати, що існує пряма або зворотна стохастична залежність між рядами даних. У випадку лінійної залежності коефіцієнт кореляції (2.88) та індекс кореляції (2.85) збігаються.

Кореляційне поле. Слід зважати на те, що обчислення коефіцієнтів кореляції є суто технічною процедурою, яку потрібно обґрунтувати відповідним аналізом природних умов. Необхідно, щоб компоненти, між якими встановлюється кореляційний зв'язок, спостерігались одночасно або в тих самих пунктах. Так, коректні коефіцієнти кореляції між вмістом різних компонентів можна встановити за пробами в тих самих свердловинах, створах ріки тощо.

Наочне уявлення про наявність або відсутність зв'язку між різними параметрами дає кореляційне поле, приклади якого наведені на рис. 2.23. Координати точок відповідають парам значень «*x* – *y*» вмісту двох компонентів у природних водах і утворюють «хмару». За формою цієї «хмари» можна зробити висновок щодо зв'язку між компонентами.



а) *б)* *в)* *г)*

Рис. 2.23. Види кореляційного поля: *а* – прямий зв'язок між даними, *б* – зворотний зв'язок, *в* – зв'язок відсутній, *г* – схема до наближеного визначення коефіцієнта кореляції

На рис. 2.23, *а* показана позитивна кореляція, коли при зростанні одного показника зростає й значення іншого. У разі зворотної кореляції (рис. 2.23, *б*) зростання одного показника супроводжується зменшенням іншого. Хмара точок на рис. 2.23, *в* є типовою для нульової кореляції, коли зв'язок між рядами даних відсутній. Слід зважати, що кореляція залежить від масштабу змінних на осях. Чим більш вузькою є «хмара» точок, тим тіснішим є зв'язок між даними. Кількісно він характеризується коефіцієнтом кореляції.

Наближену оцінку коефіцієнта кореляції можна отримати графічно (рис. 2.23, *г*), використовуючи формулу

$$r = \frac{n_1 - n_2}{n_1 + n_2}, \tag{2.89}$$

де n_1 – кількість точок (пар спостережень) у квадрантах I та III, n_2 – кількість точок у квадрантах II та IV.

Побудова лінійного рівняння регресії за допомогою функцій Microsoft Excel. Розрахунок коефіцієнтів рівнянь регресії здійснюється за допомогою функції ЛИНЕЙН(...). Порядок розрахунків зводиться до таких дій:

1) обираємо функцію ЛИНЕЙН(...), використовуючи категорію «Статистичні» майстра функцій;

2) у вікні цієї функції задаємо область значень результуючої (*Y*) і факторної (*X*) ознак, виділяючи мишею відповідні області таблиці; параметри *константа* і *статистика* задаємо такими, що дорівнюють 1. Значення константи «1» означає, що рівняння регресії повинне мати вільний член. Значення статистики «1» означає, що, крім коефіцієнтів рівняння регресії, повинні виводитися статистичні характеристики:

- похибки коефіцієнтів регресії;
- коефіцієнт детермінації;
- стандартна похибка для оцінки *Y*;
- статистика Фішера;
- число ступенів вільності;
- факторна (регресійна) сума квадратів (розсіювання регресії);
- залишкова сума квадратів (залишкове розсіювання);

3) виділяємо область комірок (5×2), включаючи комірку, у якій був отриманий результат обчислення функції «ЛИНЕЙН(...)»; натисненням кнопки миші активізуємо рядок формул;

4) утримуючи Ctrl+Shift, натиснути Enter.

У результаті одержуємо значення коефецієнтів рівняння регресії (табл. 2.16) та статистики, що зазначені вище. Для даних прикладу 2.10 результат обчислень має вигляд (рис. 2.24).

Таблиця 2.16

Статистичні характеристики, обчислені в результаті виконання функції «ЛИНЕЙН(...)»

b_1 – коефіцієнт рівняння регресії перед ознакою X	b_0 – вільний член рівняння регресії
S_{b1} – похибка коефіцієнта регресії b_1	S_{b0} – похибка коефіцієнта регресії b_0
R^2 – коефіцієнт детермінації	S – стандартна похибка для оцінки Y
F – статистика Фішера	$df = n - m - 1$ – число ступенів вільності; m – число факторів
$Q_{факт}$ – факторна (регресійна) сума квадратів (розсіювання регресії)	$Q_{зал}$ – залишкова сума квадратів (залишкове розсіювання)

D60 f_x

	A	B
49	X	Y
50	1	7
51	2	10
52	3	9
53	4	15
54	5	17
55	6	20
56	7	18
57	8	23
58	9	23
59	10	25
60	2,03030303	5,533333333
61	0,18267475	1,133467015
62	0,939176353	1,659225814
63	123,5277931	8
64	340,0757576	22,02424242

Рис. 2.24. Результат обчислень за допомогою функції ЛИНЕЙН(...)

Оцінка рівняння регресії на адекватність. Перевірка лінійної регресії на адекватність означає з'ясування наявності залежності Y від X у вигляді $y = b_0 + b_1 x$. Приймається нульова гіпотеза H_0, яка полягає в тому, що коефіцієнт регресії $b_1 = 0$. Це означає, що рівняння регресії буде мати вигляд $y = b_1 = \bar{y}$, тобто функціональної залежності між Y та X немає. Для перевірки цієї гіпотези використовується критерій Фішера. Згідно з цим критерієм порівнюються дві дисперсії. Одна з них пов'язана з факторною дисперсією $S^2_{факт}$, тобто з дисперсією розрахункових значень $f(x_i)$, що отримані з регресійної прямої, але відрізняється ступенем вільності k_1:

$$S^2_{факт.k1} = \frac{1}{k_1}\sum_{i=1}^{n}\left(f(x_i) - \bar{y}\right)^2, \tag{2.90}$$

інша є дисперсією залишків, що характеризує відхилення фактичних значень y_i від значень $f(x_i)$, що отримані з рівняння регресії:

$$S^2_{зал.k2} = \frac{1}{k_2}\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2. \tag{2.91}$$

Число ступенів вільності k_2 для статистики $S^2_{зал.k2}$ дорівнює $n - 2$. Число ступенів вільності k_1 для статистики $S^2_{факт.k1}$ для однофакторної регресії завжди дорівнює 1. Факторна дисперсія характеризує ступінь відхилення значень $f(x_i)$, що отримані з рівняння регресії, від середнього значення Y, дисперсія залишків – відхилення точок від лінії регресії.

Відношення F зазначених дисперсій є випадкова величина, що розподілена за законом Фішера [3] зі ступенями вільності k_1, k_2:

$$F = \frac{S^2_{факт.k1}}{S^2_{зал.k2}}. \tag{2.92}$$

Значення критерію Фішера може бути записано через індекс детермінації – квадрат індексу кореляції (2.85):

$$F = \frac{R^2}{1 - R^2}(n - 2). \tag{2.93}$$

Розраховане значення порівнюється з критичним для критерія Фішера $F_{1,n-2,\alpha}$, яке обчислюють з певним рівнем значущості α. Якщо $F > F_{1,n-2,\alpha}$, нульова гіпотеза відкидається, тобто $b_1 \neq 0$, Y залежить від X. Це означає, що модель адекватна з ймовірністю $(1 - \alpha)$. У цьому випадку дисперсія залишків статистично менша від факторної дисперсії, тобто побудоване рівняння регресії краще описує зв'язок між даними, ніж середнє стале значення.

У прикладі 2.10 згідно з табл. 2.14 спостережене значення критерію Фішера за формулою (2.92) становить

$$F = \frac{340{,}074}{22{,}024} \cdot (10-2) = 123{,}527 .$$

Таке ж значення отримаємо і за формулою (2.93). Аналогічний результат дає обчислення в Excel (рис. 2.24). Критичне значення можна знайти або за допомогою таблиць критичних значень критерію Фішера, що наведені у додатках [4], або користуючись майстром функцій Excel, категорія «Статистичні», де треба вибрати функцію «=F.ОБР.ПХ(α;k1;k2)», параметрами якої є рівень значущості α та ступені вільності k_1, k_2.

Обираючи $\alpha = 0{,}05$ при ступенях вільності k_1=1, k_2=8, отримаємо значення $F_{1,8,0,05}$=F.ОБР.ПХ(0,05;1;8) = 5,317. Очевидно, що виконується умова $F > F_{1,8,0,05}$, тобто побудована в прикладі 2.10 модель заснована на лінійному рівнянні регрессії y = 2,5+1,3 x і є адекватною з ймовірністю 95%.

Як критеріальну статистику Фішера використовують також відношення загальної дисперсії $S^2_{заг} = \dfrac{\sum (y_i - \overline{y})^2}{n-1}$ та дисперсії залишків (2.91). Загальна дисперсія характеризує ступінь відхилення значень y_i від середнього значення вибірки, а дисперсія залишків, як відзначалосявище, – ступінь відхилення y_i від значень $f(x_i)$, що отримані за рівнянням регресії. Чим ближче точки розташовані до середнього значення (лінії регресії), тим менша загальна дисперсія (дисперсія залишків).

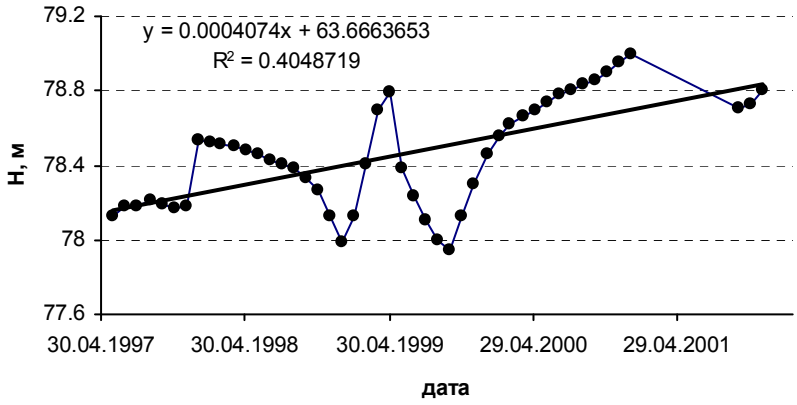
Згідно з цим підходом обчислюється величина [5]

$$F = \frac{S^2_{заг}}{S^2_{зал.k2}}, \qquad (2.94)$$

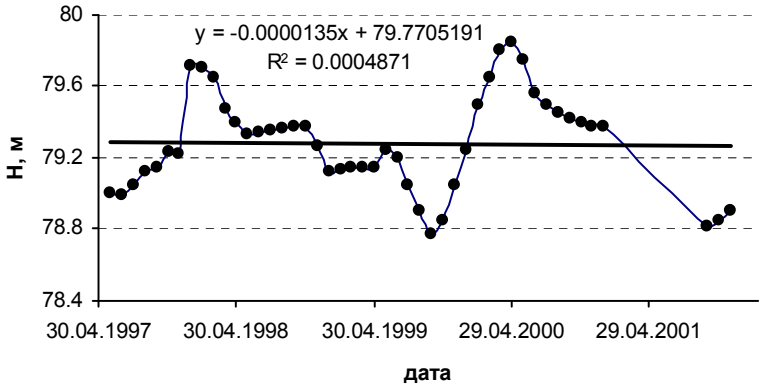
яка вважається розподіленою за законом Фішера з $n-1$ та $n-2$ ступенями вільності. Ця величина порівнюється з критичними значенням $F_{n-1,n-2,\alpha}$ з розподілу Фішера. Якщо $F > F_{n-1,n-2,\alpha}$, то з імовірністю $(1-\alpha)$ можна стверджувати, що побудоване рівняння регресії є статистично значущим. У цьому випадку дисперсія залишків менша від загальної дисперсії, тому також можна стверджувати, що побудоване рівняння регресії краще описує зв'язок між даними, ніж середнє стале значення. Якщо $F < F_{n-1,n-2,\alpha}$, то середнє значення у більшій мірі відповідає даним, ніж побудоване рівняння регресії.

Приклад 2.11. Для часових рядів коливань рівня води у двох свердловинах протягом кількох років за допомогою програми Excel побудовані лінійні регресії (рис. 2.25). Аналіз достовірності рівнянь регресії (табл. 2.17) за критерієм Фішера показує, що рівняння регресії для свердловини «2а» є статистично значущим, тоді як середнє значення рівня підземних вод у

свердловині «4р» краще описує дані, ніж рівняння регресії. Це узгоджується з суттєвою різницею значень індекса детермінації R^2 для обох рівнянь регресії.



а)



б)

Рис. 2.25. Побудова лінійної регресії часових рядів коливання рівня підземних вод у свердловинах: *а* – «2а», *б* – «4р»

Таблиця 2.17

Аналіз достовірності рівнянь регресії (рис. 2.25) за критерієм Фішера

№ св.	Середнє $\bar{y}$	Загальна дисперсія $S^2_{заг}$	Дисперсія залишків $S^2_{зал}$	F	$F_{47,46,0,05}$	Висновок
2а	78,45	0,082	0,049	1,68	1,61	Статистично значуще
4р	78,28	0,074	0,074	1,0	1,61	Статистично незначуще

Статистична значущість коефіцієнтів регресії. Повертаючись до рівняння регресії $y = b_0 + b_1x$, зазначимо, що у невеликих за обсягом сукупностях коефіцієнти регресії схильні до випадкових коливань. Тому слід

перевірити їх значущість. Для цього через дисперсію залишків $S^2_{зал.k2}$ (2.91) визначають помилки коефіцієнтів регресії за формулами:

$$S_{b0} = \sqrt{\frac{S^2_{зал,k2}}{n}\left(\frac{\sum x_i^2}{\sum\left(x_i - \bar{x}\right)^2}\right)}\ ;\ S_{b1} = \sqrt{\frac{S^2_{зал,k2}}{\sum\left(x_i - \bar{x}\right)^2}} \qquad (2.95)$$

та формують статистики:

$$t_{b1} = \frac{|b_1|}{S_{b1}};\ \ t_{b0} = \frac{|b_0|}{S_{b0}}, \qquad (2.96)$$

що розподілені за законом Стьюдента із ступенем вільності $n - 2$. Чим більший розкид значень коефіцієнтів b_0, b_1, тим більше помилки коефіцієнтів регресії S_{b0} і S_{b1}, тим менше t з розподілу Стьюдента. Величина t порівнюється з критичним значенням статистики Стьюдента $t_{n-2,\alpha}$. Якщо $t > t_{n-2,\alpha}$, то коефіцієнт регресії статистично значущо відрізняється від нуля, а розкид значень при його оцінці відносно малий. Якщо $t < t_{n-2,\alpha}$, то коефіцієнт регресії настільки малий, що правомірно прийняти його таким, що дорівнює нулю. Критичне значення можна обчислити або за допомогою таблиць у додатках [4] або користуючись майстром функцій Excel, категорія «Статистистичні», де треба вибрати функцію «=СТЬЮДЕНТ.ОБР.2Х(α;n-2)», параметрами якої є рівень значущості α та ступені вільності $n - 2$.

Для даних прикладу 2.10 похибки коефіцієнтів регресії (2.95) становлять: S_{b0}= 1,133 і S_{b1}= 0,183. Тоді статистики Стьюдента (2.96) будуть такі: t_{b0}= 4,882; t_{b1}=11,114. Критичне значення статистики Стьюдента знаходимо за допомогою функції =СТЬЮДЕНТ.ОБР.2Х(0,05;8), отже, $t_{8,0,05}$ = 2,31. Бачимо, що фактичні (спостережені) статистики значно більші за критичні; це означає, що обидва коефіцієнти регресії є статистично значущими.

Значущість індекса (коефіцієнта) парної кореляції R(r), який у випадку лінійної залежності може бути обчислений за формулами (2.85) – (2,88), також може бути перевірена за критерієм Стьюдента. Фактична (спостережена) статистика обчислюється за формулою

$$t = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \qquad (2.97)$$

та порівнюється зі значенням $t_{n-2,\alpha}$ з розподілу Стьюдента. Якщо $t > t_{n-2,\alpha}$, то гіпотеза про те, що $r = 0$ (коефіцієнт кореляції дорівнює нулю) відхиляється, тобто зв'язок між порівнюваними рядами даних існує. Інакше, можна вважати, що зв'язку між даними немає.

Для даних прикладу 2.10 спостережена статистика $t = 11{,}114$ значно перевищує критичну $t_{8,0,05} = 2{,}31$. Тому індекс кореляції є значущим, зв'язок між порівнюваними рядами даних існує.

Крім величини (2.98), для оцінки статистичної значущості коефіцієнта кореляції r може бути використана також величина

$$t = \left(\frac{1}{2} \ln\left(\frac{1+r}{1-r} \right) - \frac{|r|}{2(n-1)} \right) \sqrt{n-3} \,,$$

яка порівнюється з табличним значенням з розподілу Стьюдента.

Нехай, наприклад, за результатами 42 випробувань отриманий коефіцієнт кореляції 0,17. Тоді, порівнюючи $t \approx 1{,}2$ зі значеннями $t_{40,0,1} \approx 1{,}68$ і $t_{40,0,05} \approx 2{,}02$, переконуємося, що навіть при більш «м'якому» критерії ($\alpha = 0{,}1$) гіпотезу про наявність зв'язку між даними слід відкинути.

Ступінь кореляційної залежності між двома рядами даних можна охарактеризувати графічно (рис. 2.26). Тут $b_1 = \operatorname{tg} \alpha_1$, $b_2 = \operatorname{tg} \alpha_2$, причому $\beta = 90^\circ - (\alpha_2 - \alpha_1)$. Чим вужчий кут між прямими регресії, тим тісніший зв'язок існує між рядами даних.

Коефіцієнт кореляції надійно характеризує зв'язок між даними лише в умовах лінійного рівняння регресії, тобто «у середньому». Але, наприклад, у різних діапазонах значень досліджуваних величин характер залежності між ними може суттєво змінюватися.

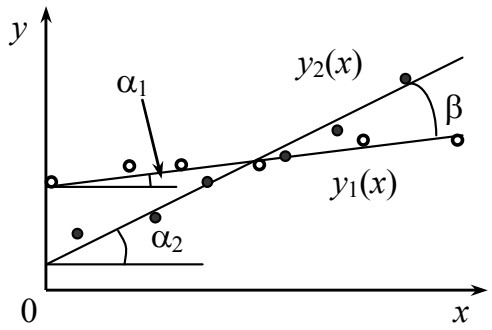


Рис. 2.26. Геометрична характеристика кореляційної залежності

Так, для слабких водних розчинів речовин існує лінійна залежність між концентрацією речовини у розчині та у сорбенті. При високих концентраціях, коли практично вся ємність сорбенту вичерпана, лінійний зв'язок між цими величинами порушується. Зміна коефіцієнта кореляції у різних діапазонах параметрів може виникнути і через неправильне групування даних, геометричні похибки тощо.

Нелінійна регресія. У деяких випадках вимірювана величина може бути більш точно передбачена за допомогою нелінійного рівняння. Як приклад розглянемо квадратичну регресію, рівняння якої має вигляд $y = b_0 + b_1 x + b_2 x^2$.

Користуючись принципом мінімуму суми квадратів відхилень уздовж осі *Oy*, складемо функцію

$$U(b_0, b_1, b_2) = \sum_{i=1}^{n} \left[y_i - \left(b_0 + b_1 x + b_2 x^2 \right) \right]^2 . \qquad (2.98)$$

Мінімум $U(b_0, b_1, b_2)$ і коефіцієнти b_0, b_1 та b_2 знайдемо, вирішуючи систему рівнянь

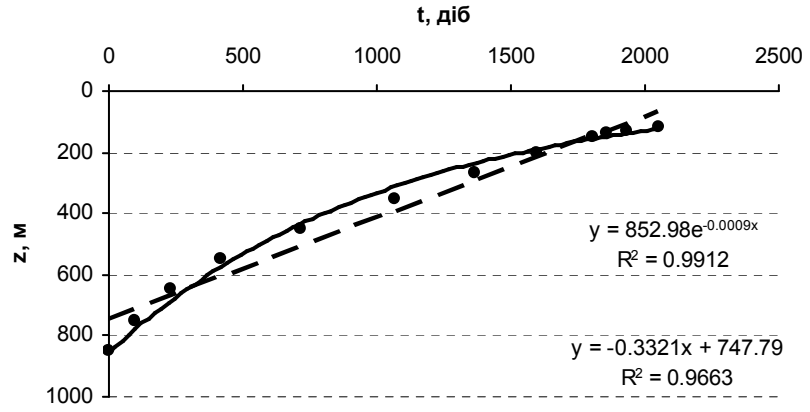
$$\frac{\partial U}{\partial b_0} = 0 , \frac{\partial U}{\partial b_1} = 0 , \frac{\partial U}{\partial b_2} = 0 .$$

Аналогічно можна побудувати рівняння регресії у вигляді многочлена більш високої степені.

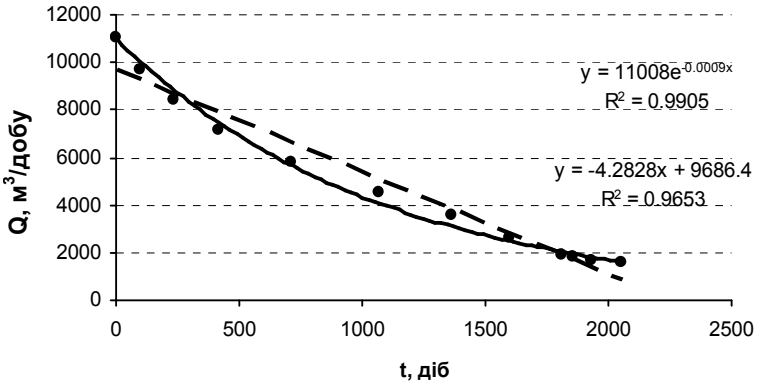
Але збільшення степені не завжди виправдане, бо ускладнює обчислення і часто не відповідає фізичній природі описуваного процесу. Тому на практиці часто використовуються експоненціальна $y = ae^{bx}$, логарифмічна $y = a + b \lg x$ і степенева $y = a x^b$ залежності.

Параметри експоненціальної регресії можна визначити на основі формули (2.89), замінюючи змінні. Нехай, наприклад, експериментальним даним відповідає залежність $y = a\, e^{bx}$. Тоді, позначивши $z = \ln y$, отримаємо рівняння $z = \ln a + b\, x$ або $z = a_0 + b\, x$. Це рівняння вже є лінійним за параметрами a_0 і b, тому для їх оцінки можна застосувати метод найменших квадратів за формулами (2.83).

Приклад 2.12. Експериментальні дані відносно динаміки затоплення шахти у центральній частині Криворізького басейну показані точками на рис. 2.27. Для апроксимації змін рівня шахтних вод (глибини від поверхні землі *z*) та водоприпливу до підземних виробок з часом використані лінійна та експоненціальна апроксимації. Вірогідність обох апроксимацій доволі висока, що демонструють високі значення індексу детермінації R^2, хоча з фізичної точки зору обґрунтованою є експоненціальна залежність, яка відтворює уповільнення динаміки затоплення з часом.



а)



б)

Рис. 2.27. Апроксимація експериментальних даних динаміки затоплення шахти у часі: *а* – глибина рівня води від поверхні землі, *б* – водоприплив (суцільна крива – експоненціальна залежність, штрихова – лінійна)

Контрольні питання до підрозділу 2.5

1. За яким принципом будується лінія регресії?
2. Поясніть геометричний зміст коефіцієнтів регресії.
3. Охарактеризуйте найбільш поширені види регресії. Як перейти від нелінійної до лінійної регресії?
4. Як перевіряється статистична значущість рівняння регресії?
5. Чому неможливо зі 100 %-ною надійністю описати всі зміни параметрів стану природного середовища?
6. Розкрийте поняття «стохастичний зв’язок між даними».
7. Що включає кореляційне поле?
8. Які висновки про зв’язок між даними можна зробити на основі візуального аналізу кореляційного поля?
9. Які висновки про зв’язок між даними можна зробити на основі аналізу коефіцієнта кореляції та його достовірності?

10. Як перевіряється статистична значущість коефіцієнта кореляції?

11. Розкрийте геометричний зміст коефіцієнта кореляції.

Література до розділу 2

1. Синайский Е.С. Высшая математика: учеб. пособие: в 2 ч. / Е.С. Синайский, Л.В. Новикова, Л.И. Заславская. – Днепропетровск: НГУ, 2006. – Ч. 2. – 452 с.

2. Елементи теорії ймовірностей та математичної статистики в гірництві: навч. посіб. / О.О. Сдвижкова, О.В. Бугрим, Д.В. Бабець, О.С. Іванов. – Дніпропетровськ: НГУ, 2015. – 102 с.

3. Рудаков Д.В. Математичні методи в охороні підземних вод: навч. посіб. / Д.В. Рудаков. – Дніпропетровськ: НГУ. – 2012. – 158 с.

4. Лук'яненко І. Економетрика / І. Лук'яненко, Л. Краснікова. – Київ: Знання, 1998. – 493 с.

5. Новікова Л.В. Теорія ймовірностей та математична статистика: навч. посіб. / Л.В. Новікова, Б.Д. Котляр, В.І. Бичков. – Київ: Техніка, 1996. – 184 с.

6. Теорія статистики: навч. посіб. / П.Г. Вашків, П.І. Пастер, В.П. Сторожук, Є.І. Ткач. – Київ: Либідь, 2001. – 320 с.

7. Вентцель Е.С. Теория вероятностей: учеб. для вузов / Е.С. Вентцель. – Москва: Наука, 1969. – 572 с.

8. Рыжов П.А. Математическая статистика в горном деле: учеб. для вузов / П.А. Рыжов. – Москва: Высш. школа, 1973. – 287 с.

9. Гмурман В.Е. Теория вероятностей и математическая статистика: учеб. для вузов / В.Е. Гмурман. – Москва: Высш. школа, 1972. – 368 с.

10. Карасев Б.В. Статистический подход к изучению природы и некоторые закономерности распределения вещества Земли: учеб. для вузов / Б.В. Карасев. – Москва: Наука, 1971. – 151 с.

11. Хан Г. Статистические модели в инженерных задачах / Г. Хан, С. Шапиро. – Москва: Мир, 1969. – 395 с.

12. Некоторые задачи статистической геомеханики: учебник / А.Н. Шашенко, С.Б. Тулуб, Е.А. Сдвижкова. – Киев: Пульсары, 2001. – 243 с.

13. Utts J.M. Seeng through statistics (revised edition) / J.M. Utts. – Brooks/Cole, 2014. – 656 p.

Розділ 3
ОСОБЛИВОСТІ ТА ПРИКЛАДИ МАТЕМАТИЧНИХ МОДЕЛЕЙ ПРИРОДНИЧИХ СИСТЕМ

3.1. Особливості геологічних моделей

Характер геологічної інформації. Геологію, на відміну від фізики, математики чи хімії, часто відносять до описових наук, що є наслідком її тривалого розвитку в умовах нестачі експериментальних даних та вимірювань через недоступність геологічних об'єктів для безпосереднього спостереження.

Експериментальною основою в геології є польові спостереження і дослідно-розвідувальні роботи, результати яких після камеральної обробки стають основою для перевірки гіпотез та верифікації математичних моделей. Оскільки в геології дуже складно провести експеримент у лабораторних умовах, поняття, які використовують в точних науках і в геології, не тотожні.

Різноманіття геологічних об'єктів і методів їх вивчення призводить до значної різнорідності отримуваної інформації, яку можна поділити на *словесну (описову)*, *графічну (картографічну)* і *цифрову*. Ще кілька десятків років тому геологічна інформація мала в основному якісний характер з переважанням словесного опису і графіків, тоді як кількісний аналіз виконував головним чином допоміжні функції. Останнім часом через нові технічні можливості вимірювань додався колосальний обсяг цифрової інформації, яка потребує осмислення та коректної інтерпретації. Але при використанні кожного з цих видів інформації виникають певні складнощі.

Словесна інформація. Якщо поняття в математиці є абстрактними і логічно виводяться на основі вихідних положень – аксіом, то поняття геологічного аналізу багатослівні й часто сформульовані образно. Теоретичні висновки геологів засновані на особистому досвіді та інтуїції, тому вони відображають не лише реальні властивості природних об'єктів та процесів, а частково і суб'єктивні уявлення авторів. Отже, існуючі в геології поняття і визначення часто неоднозначні, неконкретні, сформульовані на мові, повній образних виразів, порівнянь, аналогій.

Так, існує 39 визначень поняття «мінерал», 49 – «гірська порода», 63 – «формація», 112 – «фація» [1]. Тому для проведення кількісного аналізу геологічних об'єктів математичними методами потрібно з усього набору геологічних понять виділити основні, до яких будуть застосовні методи класифікації та математичної логіки.

Поширеною формою узагальнення знань про властивості геологічних об'єктів є класифікації та угруповання. Однак більшість з них проведена за

якісними ознаками, при цьому набір цих ознак і кількість груп у класифікаціях неоднакові. Наприклад, для поділу вивержених порід за мінеральним та хімічним складами використовується, як мінімум, п’ять різних класифікацій, які запропонували С. Мішель-Леві, Г. Розенбуш, Ф.Ю. Левінсон-Лессінг, П. Нігглі та О.М. Заварицький [2].

Картографічна інформація. Неоднозначні геологічні поняття часто беруться за основу умовних позначень при складанні графічних геологічних документів, у тому числі розрізів, планів, карт. У результаті цього картографічна геологічна інформація також є неоднозначною. Через це геологічні карти, складені в однаковому масштабі для тієї ж території в різні роки і різними дослідниками, істотно відрізняються одна від одної.

Цифрова інформація, обсяг якої суттєво збільшився за останні роки, також має деякі специфічні особливості. З точки зору вибіркового методу вивчення і складності геологічних об’єктів вона відображає їх властивості не повністю, а через технічні похибки вимірювання – не завжди достатньо точно.

Неоднозначність цифрової інформації виникає також за рахунок того, що деякі властивості геологічних об'єктів іноді можуть бути виражені різними числовими характеристиками. Наприклад, ступінь окатаності піщаних зерен і гальки характеризує вид їх транспортування і відстань до джерела. Однак, для оцінки ступеня окатаності можуть бути використані різні величини, зокрема частка від ділення радіуса кривизни найгострішого кінця піщинки або гальки на її середній радіус; відношення середнього радіуса максимальних кіл, що описують вершини всіх кутів межі в її проекції на площину, до радіуса найбільшого кола, вписаного в цю проекцію, тощо.

При вивченні корисних копалин можуть аналізуватися валовий вміст хімічних елементів, вміст їх оксидів, сульфідів або інших хімічних сполук, вміст мінералів-носіїв корисних компонентів та ін. Для більшості рудних родовищ найчастіше використовують вміст хімічних елементів, для розсипних родовищ – вміст корисних мінералів, іноді – вміст сполук металів, що мають контрастні технологічні властивості. Зокрема, при переробці олов’яних руд значно легше вилучаються в концентрати оксиди олова порівняно із сульфідами, у металургійних процесах залізних руд силікати заліза не виплавляються, а йдуть у шлаки.

Отже, для вибору оптимального параметра для кількісного аналізу слід установити, яка з технічно можливих для вимірювання характеристик найбільш повно відображає зміни властивостей, що цікаві для практики.

Типи математичних моделей у геології. Моделювання в геології вивчає закони побудови моделей геологічних об’єктів, досліджує на їх основі геологічні та фізичні процеси. Відповідно до інформації, яка використовується

для їх побудови, моделі в геології також умовно поділяють на словесні, графічні та знакові (математичні).

До *словесних моделей* належать численні класифікації, поняття і визначення, на яких побудовані геологічні теорії. *Графічними моделями* є різноманітні графічні документи, у тому числі карти, плани, розрізи, що спрощено відображають просторову структуру, особливості та властивості реальних об’єктів геологічного середовища. До *математичних моделей*, як зазначалося в розділі 1, відносять формули та співвідношення, що описують взаємозв’язки і закономірності змін властивостей геологічних об’єктів або параметрів геологічних процесів.

Вище згадувалося, що інтенсивне впровадження математичних методів та програмних засобів для обробки різнорідної інформації в геологічних дослідженнях за останні роки робить межі між описаними типами моделей досить умовними. Так, картографічна інформація за допомогою програм графічної обробки оцифровується і використовується як базис для складних просторових моделей при розв’язанні гідрогеологічних та геофізичних задач. З іншого боку, результати просторового моделювання, яке верифікується вимірами при геологічних зйомках, зображуються у вигляді карт та тривимірних зображень. Потужні комп’ютерні моделі дають можливості багатоваріантного аналізу та прогнозу геологічних властивостей, гідрогеологічних та інженерно-геологічних процесів, що робить можливим перевірку теоретичних гіпотез щодо фактичного стану масиву гірських порід.

У практиці геологічних досліджень застосовуються як *статичні*, так і *динамічні* моделі, описані в розділі 1. Динамічні моделі більш складні й потребують використання відповідного математичного апарату. Тому інколи динамічні моделі спрощують, розбиваючи весь часовий інтервал на кілька менших інтервалів. На кожному з цих інтервалів динамічні моделі замінюють статичними наближеннями, що дозволяє вивчати процес за стадіями.

Залежно від характеру зв’язку між параметрами і властивостями досліджуваних об'єктів та цілей дослідження на практиці при вивченні геологічних об’єктів можуть використовуватися комбінації *детермінованих, статистичних і стохастичних* моделей.

При практичному використанні результати детермінованих моделей геологічних об'єктів потребують коректної інтерпретації, оскільки функціональні зв’язки зберігаються лише у вузьких інтервалах, а через спрощення просторового розподілу параметрів результати моделювання відхиляються від результатів спостережень.

Статистичні моделі в геології зазвичай використовуються для:

• отримання за вибірковими даними найбільш надійних оцінок властивостей геологічних об’єктів, зокрема, середніх, дисперсій, коефіцієнтів асиметрії та ексцесу;

• перевірки гіпотез відносно подібності чи відмінності геологічних об’єктів;

• виявлення та опису залежностей між властивостями геологічних об'єктів;

• класифікації геологічних об'єктів;

• визначення обсягу вибіркових даних, необхідного для оцінки властивостей геологічних об'єктів із заданою точністю;

• оцінки ймовірності помилок за рахунок вибіркового методу.

Зважаючи на різноманіття геологічних задач, статистичні моделі застосовуються у поєднанні з деякими ймовірнісними методами (випадкові функції, часові ряди, дисперсійний аналіз), інтерполяційними поліномами, іншими методами математики.

При використанні статистичних моделей геологічного об’єкта можливі такі припущення щодо його однорідності [2]:

1. Об’єкт вважається внутрішньо однорідним, а просторові зміни властивостей є випадковими, тобто не залежать від місця виміру. Це випадок умовно однорідного (квазіоднорідного) об’єкта.

2. Об’єкт вважається внутрішньо неоднорідним, просторові зміни властивостей задаються випадковими функціями координат, тобто залежать від місця виміру.

Залежно від кількості властивостей, які розглядаються як випадкові величини, відповідні моделі поділяються на одно-, дво- та багатовимірні.

Моделі просторових геологічних змінних використовуються для:

• перевірки гіпотез про закономірності розміщення геологічних об’єктів відносно один одного;

• перевірки гіпотез про характер формування геологічних утворень;

• виділення аномалій у геологічних і геофізичних полях, гідрогеологічних та інженерно-геологічних характеристиках;

• класифікації об’єктів за особливостями їх внутрішньої будови;

• обґрунтування інтерполяції та екстраполяції при оконтурюванні геологічних об'єктів;

• вибору оптимальної густоти і форми мережі спостережень.

Стохастичні моделі визначаються як такі, у яких параметри, умови функціонування і характеристики об’єкта описані випадковими величинами і функціями, вхідна інформація частково чи повністю представлена випадковими величинами. Оскільки в стохастичних моделях один чи більше параметрів є

випадковими величинами або процесами, то для тієї ж сукупності вхідних даних може бути отримано кілька результатів.

Наприклад, масив гірських порід відносно проникності й пористості, заданих у кількох точках, розглядається у цьому сенсі випадковим полем. Тоді в результаті моделювання фільтрації підземних вод їх рівень та витрата будуть випадковими величинами, для яких можна визначити математичне сподівання, дисперсію оцінок, довірчі інтервали. Такий підхід можна реалізувати описаним далі способом.

Будується кілька сотень випадкових три- або двовимірних (залежно від розмірності моделі) полів проникності й пористості. Для кожного поля виконується розрахунок фільтрації, після чого визначаються статистичні оцінки прогнозованого рівня підземних вод, складових балансу та інших параметрів. Використання випадкових величин дозволяє моделювати раптові, погано передбачувані чинники, випадкові зміни в ході процесу, якщо вхідні параметри моделі визначені на основі недостатньої кількості даних (малої вибірки).

Етапи математичного моделювання в геології. Зважаючи на різноманіття та складність геологічних задач, їх розв’язання математичними методами є багатостадійним процесом (рис. 3.1). Безпосередньо математичному моделюванню передує створення геологічної моделі та визначення вибіркової сукупності. Тому результати геолого-математичного моделювання залежать як від правильності гіпотез, прийнятих на етапі формулювання моделі з боку фахівців з геології, так і коректності математичного методу, за яким виконано моделювання. Разом з тим рішення на більшості етапів моделювання приймаються з урахуванням особливостей геологічних об'єктів та їх властивостей, тому загальний результат та його точність більшою мірою залежить від компетенції геолога та його здатності формалізувати практичну задачу. З іншого боку, компетенція математика – вибір раціонального методу (критерію) та його реалізація, без чого загальний результат не може бути коректним.

Опробування геологічних об’єктів. Недоступність геологічних об’єктів і процесів для безпосереднього спостереження зумовила широке використання в геологічній практиці вибіркових методів, описаних у розділі 2 і заснованих на кількісному аналізі вибірок з окремих ділянок. Зразки для вибірок відбираються переважно на природних і штучних відслоненнях та у спостережних свердловинах. Площа для безпосередніх спостережень, свердловини та самі відібрані проби зазвичай незрівнянно малі порівняно з досліджуваним масивом гірських порід, на які поширюються отримані дані. Через це виникає проблема раціонального просторового розміщення пунктів локальних спостережень, часу

відбору проб підземних вод, систематизації вибіркових даних та їх поширення на прилеглу територію.

Рис. 3.1. Етапи побудови та реалізації геолого-математичної моделі

Частину геологічного об’єкта, доступну для спостереження і випробування, М. Розенфельд [3] запропонував називати випробуваною сукупністю. Ступінь відповідності досліджуваної геологічної сукупності (вибірки) випробуваній сукупності й залежить від розташування, щільності та загальної кількості точок спостережень, параметрів відібраних проб та способу вимірювання даної властивості.

Виділяють три основні системи розташування точок спостереження: рівномірне, випадкове і багатостадійне випробування [2,4].

Рівномірне випробування передбачає, що точки спостережень на території досліджуваного об'єкта розподіляються за правильною геометричною мережею, що дозволяє однаково детально вивчити всю територію об'єкта. Цей спосіб є найбільш доступним та уживаним.

Випадкове випробування зазвичай застосовується, коли

1) дослідника не цікавлять просторові закономірності змін властивостей у масиві порід або достовірно відомо, що таких закономірностей немає;

2) неможливо або важко створити мережу рівномірних спостережень, зокрема, у гірській місцевості, де проби відбираються переважно з природних відслонень та відшарувань, розміщення яких можна вважати випадковим;

3) відбираються проби для контрольних аналізів.

Багатостадійне випробування застосовується для вивчення складних геологічних об’єктів на різних масштабних рівнях їх будови. Для цього об’єкт розділяється на ділянки відповідно до елементів його неоднорідності, у яких, у свою чергу, виділяються більш дрібні елементи неоднорідності тощо. У межах кожної ділянки випробовується лише частина елементарних ділянок, за рахунок чого загальна кількість спостережень істотно скорочується порівняно з рівномірним випробуванням.

Багатостадійне випробування застосовується, наприклад, при складанні ландшафтних карт. Спочатку за результатами дешифрування космознімків масштабів 1:500000 – 1:200000 виконується районування території за типом ландшафту, далі в межах кожного з цих типів виділяються ландшафти вододілів, схилів, річкових долин і т. п. Для визначення меж елементарних ландшафтів використовуються знімки масштабу 1:50000, а їх основні характеристики (склад і потужність пухких відкладів, тип ґрунту і рослинності) оцінюються шляхом вивчення так званих ключових ділянок, тобто відносно невеликих за площею ділянок, де виявлені всі особливості даного ландшафту.

Геологічна та вибіркова (статистична) сукупності. У розділі 2 зазначалося, що в основі статистичного моделювання лежать два поняття: про генеральну сукупність – безліч можливих значень певної ознаки досліджуваного об'єкта чи явища та про вибіркову сукупність (вибірку) – сукупність спостережуваних значень цієї ознаки. Ці поняття відповідають поняттям геологічної і вибіркової сукупності.

Геологічна сукупність визначається за її елементарними складовими (об’єктами, що вивчаються), межами й видами числових вимірів. Кожній геологічній сукупності може бути поставлений у відповідність набір числових характеристик, отриманих у результаті вимірювання або аналізу властивостей окремих об’єктів. Такі набори числових характеристик називаються *вибірковими (статистичними) сукупностями*. До них можна віднести

результати визначень питомої ваги, вологості зразків ґрунтів з різних шурфів на окремому геологічному об’єкті, наприклад, масиві гірських порід, схилі кар’єру тощо.

Об’єкти і межі геологічних сукупностей встановлюються залежно від цілей і завдань досліджень. За У. Крамбейном [5], елементарні складові геологічних сукупностей можна розділити на дві групи: утворені первинними об’єктами або наборами вихідних об’єктів.

До сукупностей, утворених первинними об’єктами, належать, наприклад, сукупності копалин організмів, мінералів у шліфі або шліфах. Відповідно до кожного з таких об'єктів вимірюється одна або кілька властивостей, або оцінюються середні значення властивостей у групах досліджуваних об'єктів.

До сукупностей, утворених наборами первинних об’єктів, відносять сукупності зразків або проб, за якими визначають фізико-хімічні властивості, їх гранулометричний склад, вміст корисних або шкідливих компонентів та ін. У таких наборах властивості кожного вихідного об’єкта не вимірюються, а оцінюються середні значення окремих властивостей у наявних пробах чи зразках. Числові характеристики властивостей цієї групи сукупностей залежать від розмірів і об’ємів проб.

Геологічні дослідження переважно засновані на вивченні результатів випробувань та вимірів в окремих точках безпосередньо на місці відбору або у лабораторії після камеральної обробки. Отримані вибірки належать до елементарно малих та просторово роз’єднаних об’ємів надр (у штучному чи природному відслоненні або свердловині), у той час як зроблені висновки поширюються на весь досліджуваний об’єм.

При дослідженні складних природних об’єктів вибірковими методами з обмеженою кількістю спостережень завжди існує можливість отримати помилкові результати. Тому при використанні статистичних методів у геології необхідно оцінювати надійність висновків за вибірковими даними і значущість цих висновків.

При використанні статистичної моделі геологічні об’єкти розглядаються як сукупності нескінченно великої кількості елементарних ділянок, кожна з яких відповідає за розміром окремій пробі або місцю одиничного виміру досліджуваної властивості. В умовах обмеженої кількості спостережень такий підхід цілком правомірний, оскільки розміри проб або перетину штучних відслонень (свердловин та гірничих виробок) зазвичай незрівнянно малі порівняно з розміром досліджуваних геологічних об'єктів.

У розділі 2 були сформульовані загальні вимоги до вибірок. Крім того, при використанні статистичних методів у геології вибіркова сукупність має відповідати умовам масовості, однорідності, випадковості та незалежності [4,5].

Умова масовості полягає в тому, що обсяг вибіркової сукупності має бути достатньо великим для того, щоб виявилися статистичні закономірності. Емпіричним шляхом встановлено, що надійність статистичних оцінок різко знижується при зменшенні обсягу вибірки від 60 до 20 – 30 елементів, а при меншій кількості спостережень застосовувати статистичні методи в більшості випадків взагалі не має сенсу. При проведенні геологічних, геохімічних і геофізичних зйомок кількість спостережень, як правило, велика й умова масовості дотримується. Однак, коли для кожного спостереження потрібно проходити спеціальну гірничу виробку або свердловину, доводиться обмежуватися малими вибірками. Аналогічна ситуація виникає, коли виміри якоїсь величини проводяться рідко (раз на рік), а окремі виміри треба відкидати через похибки. Значущість статистичних оцінок у таких умовах необхідно оцінювати дуже ретельно. З умовою масовості пов’язана задача про мінімально допустимий обсяг вибірки.

Умова однорідності полягає в тому, що вибіркова сукупність має складатися з таких спостережень, які належать одному об’єкту і/та виконані однаковим способом, тобто за умов однакового розміру чи періоду відбору проб і методу аналізу чи вимірювання. Ця умова може порушуватися через помилки при визначенні меж досліджуваної геологічної сукупності або технічні й організаційні складнощі при проведенні досліджень.

Межі геологічної сукупності зазвичай задаються, зважаючи на необхідність отримання масових результатів. Вважається, що всі об’єкти, які включені до геологічної сукупності, аналогічні й внутрішньо однорідні. Хоча це припущення підтверджується не завжди, оскільки схожі за якісними ознаками об’єкти можуть іноді істотно відрізнятися за кількісними характеристиками. До того ж більшість реальних геологічних утворень мають складну будову, зокрема, через зональність і наявність неоднорідностей різного масштабу. Слід зважати й на те, що результати геологічних досліджень часто отримані в різні роки за допомогою різних технічних засобів.

У практиці геологічних досліджень умова однорідності дотримується не завжди. Тому застосування статистичних методів до розв’язання конкретної геологічної задачі має супроводжуватися аналізом можливої похибки внаслідок неоднорідності вибірки, а також перевіркою гіпотези про однорідність вибірки.

Умова випадковості передбачає непередбачуваність результату одиничного вибіркового спостереження. Складність і мінливість геологічних об'єктів, як правило, виключають можливість точної оцінки їх властивостей до проведення спостереження. Тому елемент випадковості присутній у всіх геологічних дослідженнях. Однак умова випадковості строго виконується лише тоді, коли розташування місць відбору проб або проведення замірів

досліджуваної властивості взагалі не буде якимось чином пов'язано з величиною, що характеризує цю властивість. На практиці це можна досягти за рахунок проведення спостережень на рівномірній мережі, коли всі місця спостережень намічаються заздалегідь до проведення робіт і в процесі їх виконання не коригуються.

Однак при вивченні геологічних утворень з природних відслонень ця умова може порушуватися. Наприклад, на територіях зі слабо розчленованим рельєфом природні відслонення переважно розташовуються в бортах річкових долин, які, в свою чергу, часто збігаються з розривними порушеннями або виходами порід, найбільш легко піддаються процесам ерозії. У той же час властивості міцності порід пов'язані з їх текстурними особливостями і мінеральним складом. Тому статистична обробка результатів петрографічних досліджень або випробувань їх фізико-механічних властивостей за зразками, відібраними тільки з природних відслонень, може дати викривлене уявлення про властивості порід вивченої території в цілому.

Умова випадковості може порушуватися за рахунок суб'єктивності при проведенні замірів або відборі проб. Якщо при відборі зразків з товщі гнейсів один дослідник буде віддавати перевагу прошаркам більш світлого забарвлення, а інший – прошаркам більш темного забарвлення, то отримані ними вибірки будуть істотно відрізнятися за середніми показниками мінерального складу як один від одного, так і від фактичного середнього складу досліджуваної товщі.

При проведенні геологічного розвідування часто виникає необхідність у згущенні мережі спостережень на найбільш важливих ділянках. Властивості геологічних об’єктів у межах цих ділянок і на решті досліджуваної території можуть істотно відрізнятися. Тому для дотримання умови випадковості при статистичній обробці результати спостережень на ділянці деталізації мають бути виділені в окрему вибірку.

Умова незалежності полягає в тому, що результати кожного спостереження не мають залежати від результатів попередніх і наступних спостережень, а при проведенні спостережень на площі або в об'ємі результати повинні не залежати від просторових координат. Для більшості геологічних процесів і об’єктів ця умова не дотримується. У мінливості властивостей геологічних об’єктів у просторі та параметрів геологічних процесів у часі зазвичай спостерігаються певні закономірності. Зважаючи на це, область застосування статистичних моделей обмежена об'єктами, для яких характерна відсутність будь-яких закономірностей зміни в просторі або в часі, або завданнями, при вирішенні яких ці закономірності можна не враховувати.

У геологічній практиці одновимірні статистичні моделі використовуються для вирішення двох типів задач: оцінки середніх параметрів геологічних об'єктів і статистичної перевірки гіпотез.

Контрольні питання для підрозділу 3.1

1. Охарактеризуйте геологічну інформацію за її типом.
2. Чи можливе моделювання в геології без словесної інформації? Обґрунтуйте.
3. Наведіть приклади невизначеностей цифрової інформації, яка використовується в геологічних дослідженнях.
4. Які ви знаєте приклади словесних (логічних) та графічних моделей у геології?
5. У чому полягає різниця між статичними, статистичними та стохастичними моделями?
6. Яка різниця між стохастичними та детермінованими моделями? Як це впливає на їх практичне використання?
7. Для чого використовуються моделі в геології?
8. Охарактеризуйте послідовність побудови моделей у геології.
9. Від чого залежить коректність створюваних моделей?
10. Розкрийте поняття «випробувана сукупність», «геологічна сукупність».
11. У яких випадках проводиться випадкове випробування?
12. Розкрийте зміст умов: а) масовості, б) випадковості, в) однорідності, г) незалежності при відборі проб для створення моделей у геології.

3.2. Розподіли кутових випадкових величин у задачах геології та екології

Кутові випадкові величини широко використовуються для опису процесів у природному середовищі. Так, при вивченні масивів та структури гірських порід часто виникає необхідність опису їх неоднорідності та анізотропії, пов’язаних з тріщинуватістю. Розподіл азимутів падіння тріщин у просторі зручно описувати кутовими випадковими величинами [6,7], які також зручно використовуються і для опису розподілу швидкості вітру за напрямком – рози вітрів.

Характеристики положення і розкиду кутових величин мають певні особливості. Розглянемо розподіл азимутів падіння швів тектонічних брекчій у

межах зони тріщинуватості (табл. 3.1). Якщо оцінювати математичне сподівання азимутів падіння прожилок за формулою

$$\overline{\theta} = \frac{1}{n}\sum_{j=1}^{n} n_i\theta_i \,, \tag{3.1}$$

то отримаємо $\overline{\theta}$ = 152,77°, що відповідає падінню на південь-південний схід. У той же час на рис. 3.2 чітко видно, що основна маса прожилок має падіння на північ.

Статистичними моментами кутової випадкової величини є вибірковий круговий середній напрямок, вибіркова кругова медіана та мода.

Таблиця 3.1

Заміри азимутів падіння швів тектонічних брекчій
у межах зони дроблення (приклад)

Номер i	Азимут θ, град	Частота n_i	Частість $(n_i/n)\cdot 100\%$	θ_i, град	$\cos\theta_i$	$n_i\cos\theta_i$	$\sin\theta_i$	$n_i\sin\theta_i$
1	350 – 10	142	19,51	0	1,000	142,00	0,000	0,00
2	10 – 30	102	14,01	20	0,940	95,85	0,342	34,89
3	30 – 50	42	5,77	40	0,766	32,17	0,643	27,00
4	50 – 70	34	4,67	60	0,500	17,00	0,866	29,44
5	70 – 90	22	3,02	80	0,174	3,82	0,985	21,67
6	90 – 110	4	0,55	100	–0,174	–0,69	0,985	3,94
7	110 – 130	6	0,82	120	–0,500	–3,00	0,866	5,20
8	130 – 150	15	2,06	140	–0,766	–11,49	0,643	9,64
9	150 – 170	43	5,91	160	–0,940	–40,41	0,342	14,71
10	170 – 190	30	4,12	180	–1,000	–30,00	0,000	0,00
11	190 – 210	14	1,92	200	–0,940	–13,16	–0,342	–4,79
12	210 – 230	8	1,10	220	–0,766	–6,13	–0,643	–5,14
13	230 – 250	8	1,10	240	–0,500	–4,00	–0,866	–6,93
14	250 – 270	28	3,85	260	–0,174	–4,86	–0,985	–27,57
15	270 – 290	15	2,06	280	0,174	2,60	–0,985	–14,77
16	290 – 310	42	5,77	300	0,500	21,00	–0,866	–36,37
17	310 – 330	59	8,10	320	0,766	45,20	–0,643	–37,92
18	330 – 350	114	15,66	340	0,940	107,12	–0,342	–38,99
Разом		728				353,03		–26,02

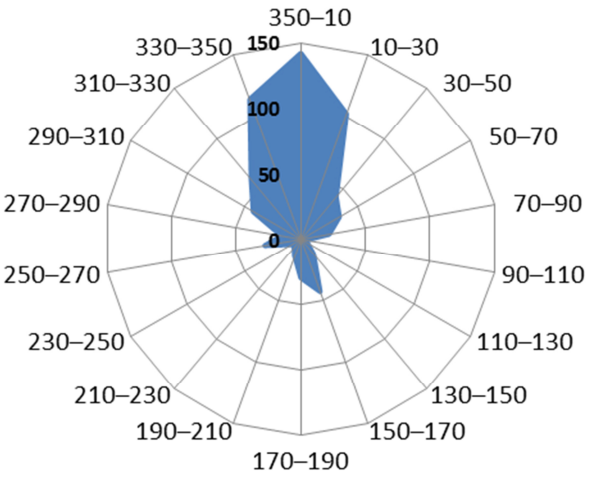


Рис. 3.2. Діаграма рози азимутів падіння швів тектонічних брекчій у межах зони дроблення

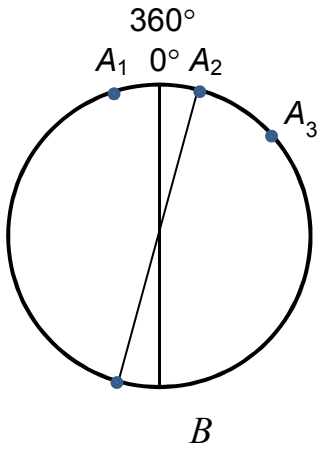


Рис. 3.3. Зображення вимірів азимутів кутів падіння прожилок у вигляді точок на окружності

Вибірковий круговий середній напрямок визначається середніми значеннями синусу та косинусу за формулами

$$\overline{C} = \frac{1}{n}\sum_{i=1}^{n} n_i \cos\theta_i\,,\ \overline{S} = \frac{1}{n}\sum_{i=1}^{n} n_i \sin\theta_i\,, \tag{3.2}$$

де θ_i – середня точка i-го інтервалу групування, причому всі інтервали мають однакову довжину; n – кількість інтервалів; n_i – частота випадкової величини на i-му інтервалі, при цьому вважається, що всі вибіркові значення в межах i-го інтервалу групування мають кут θ_i.

Довжина вектора $\overline{R}$, який відповідає середньому напрямку випадкової кутової величини θ, визначається за формулою

$$\overline{R} = \sqrt{\overline{C}^2 + \overline{S}^2}\,, \tag{3.3}$$

а відповідний кут $\overline{\theta}$ – за формулами

$$\overline{\theta} = \arccos\frac{\overline{C}}{\overline{R}};\ \overline{\theta} = \arcsin\frac{\overline{S}}{\overline{R}},\ \text{або}\ \overline{\theta} = \arctan\frac{\overline{S}}{\overline{C}}. \tag{3.4}$$

Приклад розрахунку кутового середнього напрямку показаний у табл. 3.1.

За формулами (3.2) – (3.4) отримаємо

$$\overline{C} = \frac{353{,}03}{728} = 0{,}485\,,\ \overline{S} = \frac{-26{,}02}{728} = -0{,}036\,,$$

$$\overline{R} = \sqrt{0{,}485^2 + (-0{,}036)^2} = 0{,}486;$$

$$\cos\overline{\theta} = \frac{\overline{C}}{\overline{R}} = 0{,}997\,,\ \sin\overline{\theta} = \frac{\overline{S}}{\overline{R}} = -0{,}073\,,\ \overline{\theta} = -4{,}21°.$$

Вибірковою коловою медіаною називається така точка A на окружності, що половина точок вибірки лежить по одну сторону від діаметра AB, при цьому більшість точок вибірки ближче до A, ніж до B. Якщо кількість вимірів невелика, цю характеристику можна легко знайти за графіком розподілу точок на окружності. Наприклад, для випадку на рис. 3.3 властивостям вибіркової колової медіани задовольняє точка A_2.

Для кутових вимірювань разом з модою використовують також характеристику, звану антимодою, яка відповідає значенню кута, де частота випадкової величини мінімальна.

Для деяких кутових величин, наприклад, для азимутів падіння порід в областях розвитку лінійної складчастості, властивий розподіл з двома модами, віддаленими одна від одної на 180° [6].

Вибіркова кругова дисперсія напрямків кутової випадкової величини розраховується за формулою

$$S_a^2 = 1 - \overline{R}\,. \tag{3.5}$$

Розподіл Мізеса є узагальненням нормального та рівномірного розподілів для кутової випадкової величини [4,6,7]. Його щільність виражається формулою

$$f(\theta) = \frac{1}{2\pi I_0(k)} e^{k\cos(\theta-\mu)} \tag{3.6}$$

при скінченному параметрі μ і $k > 0$, I_0 – модифікована функція Бесселя першого роду,

$$I_0(k) = \frac{1}{2\pi}\int_0^{2\pi} e^{k\cos\theta}\,d\theta = \sum_{j=0}^{\infty}\frac{1}{(j!)^2}\left(\frac{k^2}{2}\right)^j.$$

Отже, розподіл Мізеса визначається двома параметрами – μ і k, де μ є коловим середнім напрямком випадкової кутової величини, схожим з математичним сподіванням. Параметр k можна розглядати як характеристику концентрації розподілу навколо μ. Розподіл Мізеса при $k = 0$ перетворюється в рівномірний, а при $k \to \infty$ з параметрами μ і k його асимптотична поведінка відповідає нормальному розподілу з параметрами Mx = μ і $\sigma^2 = 1/k$ (рис. 3.4). Таким чином, параметр $1/k$ в розподілі Мізеса є аналогом дисперсії для нормального розподілу.

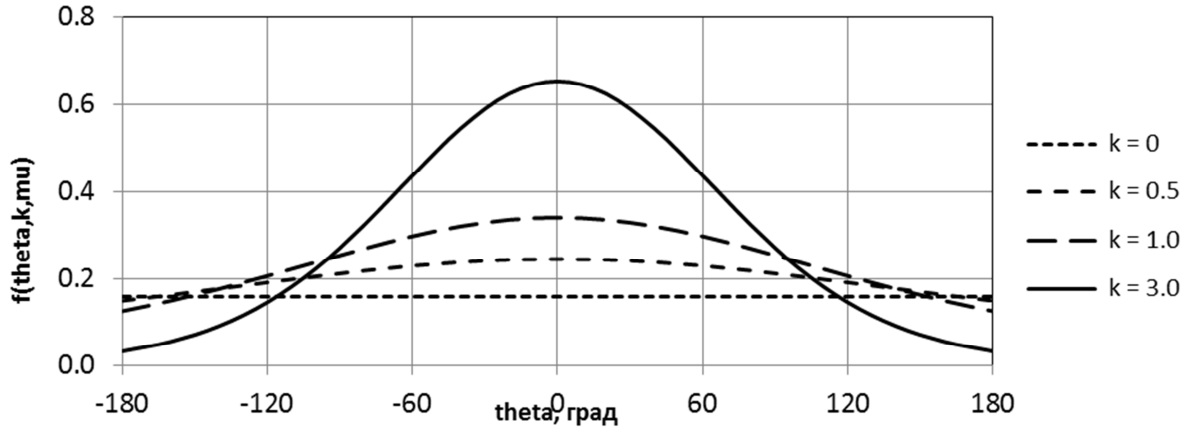


Рис. 3.4. Розподіл Мізеса для параметра $\mu = 0°$ при різних значеннях параметра k

Іншим прикладом розподілу кутової випадкової величини є роза вітрів. При дослідженні вітрових течій, які визначають перенос і дифузію домішок у приземному шарі атмосфери, виділяють мезомасштабні та мікромасштабні флуктуації вітру [8]. Мезомасштабні флуктуації характеризують зміну руху повітряних мас, що відбуваються через кілька годин – кілька діб, протягом яких сам рух можна вважати усталеним. Але в умовах такого квазіусталеного руху мають місце мікромасштабні флуктуації, характерний час яких складає від декількох секунд до декількох хвилин.

Перенос домішок в умовах постійної перебудови руху повітряних мас упродовж досить тривалого проміжку часу можна розрахувати на основі принципу зміни стаціонарних станів [8]. Відповідно до нього вважається, що протягом періоду T у даній місцевості реалізується n типів стаціонарного руху тривалістю Δt_i, причому $\sum_{i=1}^{n} \Delta t_i = T$. На кожному i-му проміжку часу розподіл концентрації домішки в атмосфері можна знайти різними аналітичним або чисельними методами [10–12] у вигляді функцій загального вигляду φ_i. Тоді середньоймовірна концентрація домішки за період T визначається за формулою

$$\bar{\varphi}(x,y,z) = \frac{1}{T}\sum_{i=1}^{n} \varphi_i(x,y,z)\,\Delta t_i, \qquad (3.7)$$

яка є статистичною моделлю. При її побудові передбачалося, що перебудова руху повітряних мас відбувається миттєво.

При наявності даних про частоту повторюваності напрямків вітру ω_i за 8 румбами тривалість періодів Δt_i визначається як $\omega_i T$, де T=365 діб. Однак розрахунок за формулою (3.7) дає фізично необґрунтовану картину з переважаючим переносом уздовж лише 8 головних напрямків. Тому більш адекватним характеру розсіювання домішок в атмосфері буде такий розподіл

напрямку вітру, щільність якого $\omega(\theta)$ будується на припущенні, що напрямок вітру безперервно змінюється при переході від одного румба до іншого:

$$\omega(\theta)=\frac{4}{\pi}\left(\omega_i+\frac{\omega_{i+1}-\omega_i}{\pi/4}\left(\theta-\theta_i\right)\right),\ \frac{\pi i}{4}\le\theta\le\frac{\pi\left(i+1\right)}{4}, \tag{3.8}$$

причому $\omega_i\ge 0,\ i=0,...,7;\ \sum_{i=0}^{7}\omega_i=1$, тому $\int_0^{2\pi}\omega(\theta)d\theta=1$.

Відповідно до неперервного розподілу швидкості вітру (3.8) формулу (3.7) треба скорегувати. Позначимо через $\varphi(x,y,z)$ стаціонарний розподіл концентрацій домішки при усталеній швидкості вітру вздовж осі Ox. У загальному випадку швидкість вітру спрямована під кутом θ до осі Ox. Використовуючи перетворення координат з поворотом на кут θ

$$\begin{cases}\xi=x\cos\theta+y\sin\theta,\\ \eta=-x\sin\theta+y\cos\theta,\end{cases} \tag{3.9}$$

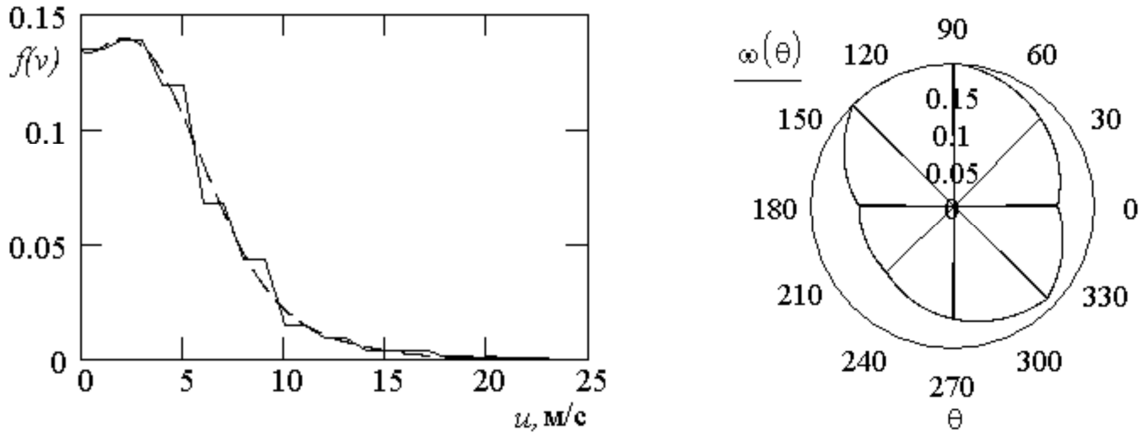
отримаємо замість (3.7) формулу для розрахунку середньоймовірної концентрації домішки у приземному шарі атмосфери у вигляді інтегралу

$$\overline{\varphi}(x,y,z)=\int_0^{v_{max}}\int_0^{2\pi}\varphi'\left(\xi(x,y,\theta),\eta(x,y,\theta),z\right)\omega(\theta)f(v,\theta)d\theta dv, \tag{3.10}$$

де $\varphi'(\xi(x,y,\theta),\eta(x,y,\theta),z)$ – концентрація домішки при усталеній швидкості вітру, спрямованій під кутом θ до осі Ox. У формулі (3.10) враховано також розподіл частоти повторюваності вітрів різної швидкості від 0 до v_{max} функцією розподілу $f(v,\theta)$.

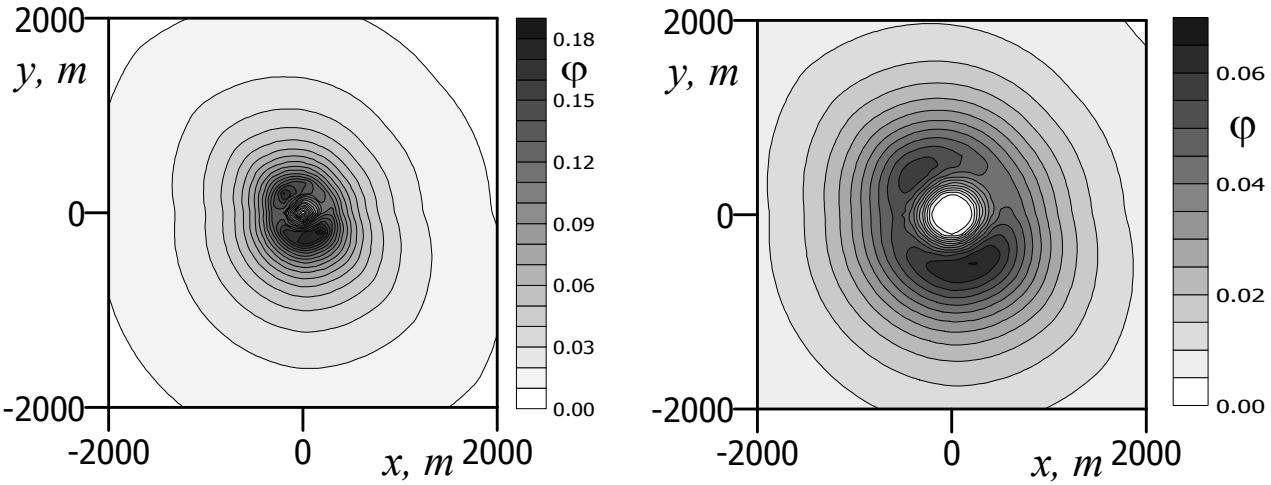
На рис. 3.5 наведено приклад розподілу швидкості та напрямку вітру, а на рис. 3.6 – відповідний розподіл середньоймовірної концентрації домішки, розрахованої за формулою (3.10) на основі тривимірної моделі розсіювання [9, с. 138]. При розрахунках прийнято, що розподіл швидкості вітру однаковий за всіма напрямками, тобто функція розподілу f не залежить від θ.

На рис. 3.6 можна виділити зони максимальної концентрації домішки, що визначаються переважаючими напрямками вітру. Зі зростанням висоти джерела викиду збільшується площа умовно «чистої» зони безпосередньо навколо джерела викиду та зменшення приземної концентрації. У наведеному прикладі концентрація домішки розрахована на основі розподілених параметрів повторюваності напрямку $\omega(\theta)$ та швидкості вітру $f(v,\theta)$.



а *б*

Рис. 3.5. Параметри вітрового режиму, використані в моделі: *а* – повторюваність швидкостей вітру (—— вимірювання, - - - апроксимація); *б* – роза вітрів



а *б*

Рис. 3.6. Розподіл середньоймовірної концентрації домішки φ (мг/м^3) на рівні землі при різній висоті джерела її викиду (м): *а* – 20, *б* – 50

Контрольні питання до підрозділу 3.2

1. Як визначається кутова випадкова величина?
2. Як обчислюються статистичні моменти кутової випадкової величини?
3. Як можна зобразити розподіл кутової випадкової величини?
4. Що можна описати розподілом Мізеса? Які його граничні випадки?
5. Чому випадкову кутову величину можна застосовувати в моделях розсіювання домішок в атмосфері?
6. Розкрийте, що означає принцип зміни стаціонарних станів атмосфери?

3.3. Аналіз однорідності вибірок

Критерії перевірки однорідності вибірок. Досвід обробки даних багатьох геологічних та гідрохімічних досліджень свідчить, що варіаційні ряди іноді включають кілька значень (до 5%), які дуже далекі від середнього значення. Враховуючи можливість помилок при одержанні результатів, слід визначитися щодо приналежності цих екстремальних значень до вибірки.

У разі великого об'єму вибірки ці значення несуттєво впливають на середнє чи дисперсію, але можуть значно змінити моменти вищого порядку: асиметрію та ексцес. Через кілька екстремальних значень емпіричний розподіл, що описується нормальним законом, може бути помилково апроксимований як логнормальний. При малих об'ємах вибірки навіть одне екстремальне значення може суттєво викривити оцінки всіх статистичних параметрів.

Часто вважається, що геологічний об'єкт є однорідним відносно досліджуваної ознаки. Це можна визначити, наприклад, на основі аналізу геологічної будови. Але на початкових стадіях досліджень дане питання важко вирішити однозначно, оскільки необхідна статистична перевірка даних відносно мінливості ознаки. Перевірка статистичної однорідності об'єкта передбачає виділення аномальних значень та подальше розділення неоднорідних вибіркових сукупностей.

Задача виявлення аномальних значень не має універсального вирішення статистичними методами. Аномальне значення часто визначається дослідним шляхом і аналізом природного об'єкта. Для відокремлення аномальних значень сукупність результатів спостережень розглядається як вибірка з різних генеральних сукупностей – «фонової» та «аномальної»; при цьому аномальні значення присутні у вибірці в невеликій кількості або відсутні взагалі.

У разі, якщо розподіл фонової генеральної сукупності вважається нормальним, відокремити аномальні значення можна за допомогою параметричних критеріїв Граббса – Смирнова та Фергюсона [2] або правила трьох сігм (підрозділ 2.3.3). Непараметричні критерії (критерій Ван-дер-Вардена, Вілкоксона [13,14], критерій згоди χ^2, наведений у підрозділі 2.4) використовуються в разі малого об'єму вибірок або в тих випадках, коли середні значення розраховані наближеними методами, наприклад за результатами спектрального аналізу. Непараметричні критерії використовуються в тих випадках, коли закон розподілу даних відрізняється від нормального або невідомий.

Критерій Граббса – Смирнова. Однорідність показників вибірки перевіряється за двома односторонніми критеріями. Для впорядкованої вибірки обчислюються величини

$$\tau_1 = \frac{x_{\max} - \bar{x}}{s} \text{ та } \tau_2 = \frac{\bar{x} - x_{\min}}{s}, \qquad (3.11)$$

де $x_{\max}$ і $x_{\min}$ – відповідно максимальне і мінімальне значення показника вибірки; $\bar{x}$ – його середнє значення; s – статистична помилка (оцінка) середнього квадратичного відхилення (табл. 2.7),

$$s = \sigma\sqrt{\frac{n-1}{n}}, \quad \sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n}}, \qquad (3.12)$$

n – кількість визначень показника або об'єм вибірки.

Чим меншими є значення τ_1 і τ_2, тим більш однорідна вибірка.

Величини τ_1 і τ_2, обчислені за рівнянням (3.11), порівнюються з табличними значеннями процентних точок критерію Граббса – Смирнова τ_α для заданої довірчої ймовірності $(1 - \alpha)$ (додаток Ж). Якщо $\tau > \tau_\alpha$, то значення, що перевіряється, вважається помилкою випробувань і визначається як таке, що не належить до вибірки, в іншому разі вибірка вважається статистично однорідною.

Приклад 3.1. Для вибірки зразків ґрунту обсягом $n = 20$, перевірених за числом пластичності x, отримані такі значення: $x_{\min} = 7$, $x_{\max} = 32$, $\bar{x} = 21{,}5$, $s = 6{,}2$. Тоді $\tau_1 = 2{,}34$, а $\tau_2 = 1{,}69$. Для довірчої ймовірності 0,9 ($\alpha = 0{,}1$) при $n = 20$ за додатком Ж знаходимо $\tau_{0,9} = 2{,}5$. Оскільки $\tau_1 < \tau_{0,9}$ та $\tau_2 < \tau_{0,9}$, можна вважати цю вибірку статистично однорідною.

Критерій Фергюсона. Вважається, що оцінка коефіцієнта асиметрії A розподілена асимптотично (тобто при великих значеннях n) нормально з математичним сподіванням 0 та дисперсією σ_A^2. Якщо розраховане значення A перевищує допустиме табличне для заданої довірчої ймовірності та ступеня вільності [15], то максимальне значення слід вважати аномальним.

Правило трьох сігм (3σ), описане в підрозділі 2.3.3, застосовують, якщо емпіричний розподіл видається симетричним, а граничні значення є явно екстремальними. Тоді з імовірністю 99,73 % можна стверджувати, що сумнівні значення (x_{min} чи x_{max}) не належать генеральній сукупності, якщо виконуються нерівності (2.54), які для статистичних оцінок набувають вигляду (див. приклад 2.6):

$$x_{min} < \bar{x} - 3\sigma, \quad \bar{x} + 3\sigma < x_{max}. \qquad (3.13)$$

Екстремальні величини вмісту компонентів є найбільш цікавими при виділенні гідрохімічних аномалій і розвідці родовищ корисних копалин, але такі числа не роблять великого впливу на визначення гідрохімічного фону.

Якщо сумнівне значення виходить за межі інтервалу $[\overline{x} - 3\sigma, \overline{x} + 3\sigma]$, то його слід виключити з вибірки і виконати повторний розрахунок статистичних параметрів. Але такий спосіб не є строгим, оскільки він не виключає помилки першого та другого родів.

Приклад 3.2 [16]. Визначимо, чи є екстремальні значення у вибірках, поданих у табл. 2.9. Середні значення вмісту сульфатів та сухого залишку для всіх свердловин становлять відповідно 327,3 та 776,5 мг/л, а середньоквадратичні відхилення – 342,2 та 630,6 мг/л. Перевіримо нерівності (3.13):

для сульфатів $x_{max} = 1344$, $x_{max} < 327{,}3 + 3 \cdot 324{,}2 = 1353{,}9$ (межа не перевищена),

для сухого залишку $x_{max} = 2745$, $x_{max} > 776{,}5 + 3 \cdot 630{,}6 = 2668{,}3$ (межа перевищена).

Зауважимо, що межа $\overline{x} + 3\sigma$ майже досягнута і для сульфатів. Отже, можна стверджувати, що ці максимальні значення є аномально високими для даної сукупності. Але такий висновок справедливий у разі нормального розподілу, якщо ж установлено правомірність логнормального закону, то відкидання крайніх значень є некоректним.

Приклад 3.3. Для вибірки, описаної в прикладі 3.1

$$x_{min} = 7 > \overline{x} - 3\sigma = 21{,}5 - 3 \cdot 6{,}3 = 2{,}6;$$
$$\overline{x} + 3\sigma = 21{,}5 + 3 \cdot 6{,}3 = 39{,}4 > 32 = x_{max}.$$

Тоді з довірчою ймовірністю 0,9973 дана вибірка вважається статистично однорідною.

Існує група так званих непараметричких критеріїв, які засновані на впорядкуванні елементів вибірки та їх ранжуванні До них належать, зокрема, критерії Ван-дер-Вардена та Вілкоксона.

Критерій Ван-дер-Вардена застосовується для перевірки гіпотези про рівність середніх, визначених за двома вибірками (А і Б); разом з тим на його основі можна зробити висновок про однорідність об'єднаної вибірки. Згідно з критерієм значення в обох вибірках впорядкуються за зростанням – ранжуються, тобто записуються в один ряд у порядку зростання. Критерій Ван-дер-Вардена являє собою величину

$$X = \sum_{1}^{h} Z\left(\frac{i}{n+1}\right), \tag{3.14}$$

де n – загальна кількість значень за двома вибірками; h – кількість спостережень у вибірці Б; i – порядковий номер кожного значення вибірки Б у загальному ряду; Z – функція, яка є зворотною функції нормального розподілу.

При $n > 20$ величина X розподілена асимптотично нормально з математичним сподіванням 0 і дисперсією σ_X^2. Перевірка гіпотези виконується у такій послідовності:

1) розрахувати всі значення аргументу $i/(n + 1)$,

2) знайти за таблицями зворотної функції нормального розподілу значення функції Z для цих аргументів,

3) підсумувати значення функції Z,

4) порівняти отримане значення критерію X з табличним для заданого рівня значущості, загального числа спостережень n і різниці між об'ємами вибірок А і Б.

Якщо розрахункове значення X за абсолютною величиною перевищує табличне, гіпотеза про рівність вибіркових середніх, а отже, і про однорідність об'єднаної вибірки відкидається.

Значення функції Z можна знайти у спеціальних таблицях та за допомогою звичайних таблиць нормованої інтегральної функції нормального розподілу (2.50) з параметрами 0, 1, використовуючи її в зворотному порядку. У разі застосування звичайних таблиць (додаток А) аргумент функції $\xi = i/(n + 1)$ прирівнюється до ймовірності p, а величина $Z(\xi)$ знаходиться за значеннями аргументу, які відповідають цим ймовірностям. Для $\xi < 0{,}5$ значення Z від'ємні, а для $\xi > 0{,}5$ – додатні. Замість таблиць нормованої інтегральної функції нормального розподілу можна скористатися функцією НОРМСТОБР у Excel.

Якщо систематичних розбіжностей між вибірками А і Б немає, то у впорядкованому ряду значення кожної вибірки будуть розташовуватися симетрично відносно середини цього ряду, що відповідає $i = n/2$ та $\xi = 0{,}5$, тоді кількість від'ємних і додатних значень Z для кожної вибірки буде приблизно однаковою, а їх алгебраїчні суми, тобто значення X-критерію, близькі до нуля. У разі несиметричності вибірок, тобто коли значення однієї групуються в нижній частині об'єднаної вибірки, а значення іншої – у верхній частині, то значення X-критерію будуть суттєво відрізнятись від нуля.

Приклад 3.4 [2]. Для зниження витрат на геологічну розвідку на одній з ділянок розсипного родовища золота частина шурфів (майже половина) була замінена свердловинами ударно-канатного буріння. Це могло вплинути на результати хімічних аналізів. Тому необхідно переконатися в тому, що результати випробування свердловин не мають систематичної помилки, тобто вибірки, отримані двома способами, дають схожі результати. Кількість свердловин і шурфів на дослідній ділянці невелика – 13 і 10 відповідно, тому для зіставлення обчислених відповідно до них вмістів золота можна

використовувати непараметричний критерій Ван-дер-Вардена. Результати випробування шурфів і свердловин наведені у табл. 3.2. За цими даними складено загальний варіаційний ряд (табл. 3.3).

У розглянутому прикладі перевіряється гіпотеза H_0 про те, що систематичних розбіжностей у визначенні вмісту золота у шурфах і свердловинах немає: $X = 0$, об'єднана вибірка однорідна. Альтернативною є гіпотеза H_1: $X \neq 0$, об'єднана вибірка неоднорідна, тобто складається фактично з двох різних вибірок. Для розрахунку величині X потрібно скористатися таблицями двостороннього критерію Ван-дер-Вардена (додаток Д).

Виконавши необхідні розрахунки, отримаємо $X = 4{,}19$. Для коректної інтерпретації результатів слід правильно встановити критичне значення та рівень значущості критерію Ван-дер-Вардена. Для цього треба порівняти помилки першого і другого родів у даному прикладі [2].

Помилка першого роду в розглянутій задачі полягає в тому, що правильна гіпотеза про відсутність систематичної розбіжності між результатами випробування у шурфах і свердловинах буде відкинута, а це не дозволить знизити витрати на подальшу розвідку за рахунок використання більш дешевих свердловин. Помилка другого роду, тобто прийняття неправильної гіпотези H_0, полягає у визнанні несуттєвою (випадковою) розбіжності між даними стосовно шурфів і свердловин, у той час як насправді вона носить систематичний характер. Через це помилка другого роду в даній ситуації може призвести до неправильної геолого-економічної оцінки родовища, що завдасть значно більший економічний збиток порівняно з додатковими витратами за рахунок помилки першого роду. Тому для зменшення ризику помилки другого роду рівень значущості при перевірці гіпотези H_0 доцільно прийняти не надто високим, наприклад 0,1.

Приймаючи в даній задачі рівень значущості $\alpha = 0{,}1$, знаходимо (додаток Д), що допустиме значення критерію Ван-дер-Вардена для числа спостережень 23 і різниці між об'ємами порівнюваних вибірок 3 становить 3,12. Оскільки $X > 3{,}12$, то гіпотеза про рівність середніх значень вмісту золота у свердловинах і шурфах відкидається, отже, об'єднану вибірку можна вважати неоднорідною. Через систематичне заниження вмісту золота у свердловинах порівняно із шурфами використовувати їх для розвідки даного родовища не рекомендується.

Таблиця 3.2

Результати випробувань на розсипному родовищі золота

Вибірка А (свердловини)		Вибірка Б (шурфи)	
№	Вміст Au, мг/м^3	№	Вміст Au, мг/м^3
1	322	1	431
2	250	2	397
3	225	3	462
4	315	4	457
5	399	5	251
6	348	6	221
7	192	7	548
8	375	8	478
9	381	9	299
10	538	10	541
11	198		
12	317		
13	293		

Таблиця 3.3

Упорядкована (ранжована) вибірка для перевірки її однорідності за критерієм Ван-дер-Вардена

№	Вміст Au, мг/м^3	Вибірка	$i/(n+1)$	$\psi(i/(n+1))$
1	192	А		
2	198	А		
3	221	Б	0,125	–1,15
4	225	А		
5	250	А		
6	251	Б	0,250	–0,67
7	293	А		
8	299	Б	0,333	–0,43
9	315	А		
10	317	А		
11	322	А		
12	348	А		
13	375	А		
14	381	А		
15	397	Б	0,625	0,32
16	399	А		
17	431	Б	0,708	0,55
18	457	Б	0,750	0,67
19	462	Б	0,792	0,81
20	478	Б	0,833	0,97
21	538	А		
22	541	Б	0,917	1,39
23	548	Б	0,958	1,73

Критерій Вілкоксона (який інколи називають критерієм Манна – Уітні) також заснований на процедурі ранжування [2,4,13,14]. Статистика Вілкоксона W і являє собою суму рангів R_i членів меншої вибірки в загальному впорядкованому ряду з обох вибірок (в об'єднаній вибірці):

$$W = \sum_{i=1}^{n_1} R_i, \; n_1 \le n_2 , \tag{3.15}$$

де n_1 і n_2 – об'єми меншої та більшої вибірок. Якщо гіпотеза H_0 про рівність середніх за сукупностями А та Б правильна ($\overline{x}_1 = \overline{x}_2$), то математичне сподівання статистики Вілкоксона (MW) і величини можливих відхилень від неї вибіркових оцінок (W) залежать тільки від об'ємів вибірок n_1 і n_2.

Критерій Вілкоксона рекомендується використовувати для вибірок невеликого об'єму, коли $n_1 < 25$ та $n_2 < 25$. Значення подвоєного математичного сподівання критерію Вілкоксона 2M(W) і його нижнього критичного значення W_1 для заданого рівня значущості α наведені в спеціальних таблицях (див. додаток И). Верхнє критичне значення критерію W_2 визначається з рівняння

$$W_2 = 2\mathrm{M}(W) - W_1 . \tag{3.16}$$

Рівень значущості для W_1 у таблицях наведений для альтернативи H_1, коли $\overline{x}_1 < \overline{x}_2$. Тому при альтернативі H_1, коли $\overline{x}_1 < \overline{x}_2$ або $\overline{x}_1 > \overline{x}_2$, рівень значущості для обчислення W_1 необхідно зменшити вдвічі.

Якщо об'єми вибірок перевищують 25, критичні значення критерію Вілкоксона можна визначити за такими наближеним формулами [4]:

$$W_1 \approx 0{,}5\left[n_1\left(n_1 + n_2 + 1\right) - 1\right] - Z_{1-\alpha/2}\sqrt{\frac{n_1 n_2\left(n_1 + n_2 + 1\right)}{12}} , \tag{3.17}$$

$$W_2 \approx n_1\left(n_1 + n_2 + 1\right) - W_1,$$

де $Z_{1-\alpha/2}$ – значення зворотної функції нормального розподілу з параметрами (0, 1).

За наявності в об'єднаній вибірці збіжних значень їм присвоюється однаковий середній ранг, що дорівнює середньому арифметичному з усіх рангів, які припадають на дану групу повторюваних значень, а формула набуває вигляду

$$W_1 \approx \frac{\left[n_1\left(n_1 + n_2 + 1\right) - 1\right]}{2} - Z_{1-\alpha/2} \times$$

$$\times \sqrt{\frac{n_1 n_2\left(n_1 + n_2 + 1\right)}{12}\left[1 - \frac{\sum_{i=1}^{k} t_i^3 - t_i}{\left(n_1 + n_2 + 1\right)\left(n_1 + n_2\right)\left(n_1 + n_2 - 1\right)}\right]^2} , \tag{3.18}$$

де k – кількість груп з повторюваними значеннями, які належать різним вибіркам; t_i – кількість значень, що збіглися в групі з номером i (i = 1,…,k). Групи повторюваних значень, що складаються повністю із значень вибірки А чи Б, можна не враховувати при обчисленні поправки.

Приклад 3.5. Використовуючи критерій Вілкоксона, перевіримо гіпотезу про рівність середнього вмісту золота у шурфах і свердловинах, необхідні дані наведені в табл. 3.4.

Результати розрахунків у вигляді фрагменту таблиці Excel показані на рис. 3.7. Розраховане значення критерію Вілкоксона $W = 151$. За додатком И для рівня значущості $\alpha/2 = 0{,}05$ і ступенів вільності $n_1 = 10$, $n_2 = 13$ знаходимо $W_1 = 92$, 2M(W) = 240, $W_2 = 148$. Таким чином, емпіричне значення критерію Вілкоксона перевищує його верхнє критичне значення; отже, з імовірністю 0,9 гіпотеза про рівність середнього вмісту золота у шурфах і свердловинах відкидається.

	A	B	C
1	Вміст Au, мг/м^3	Порядок	Ранг вибірки Б
2	192	1	0
3	198	2	0
4	221	3	3
5	225	4	0
6	250	5	0
7	251	6	6
8	293	7	0
9	299	8	8
10	315	9	0
11	317	10	0
12	322	11	0
13	348	12	0
14	375	13	0
15	381	14	0
16	397	15	15
17	399	16	0
18	431	17	17
19	457	18	18
20	462	19	19
21	478	20	20
22	538	21	0
23	541	22	22
24	548	23	23
25		276	**151**
26			
27		**W1**	**92**
28		**2MW**	240
29		**W2**	**148**

Рис. 3.7. Фрагмент обчислень у таблиці Excel для перевірки за критерієм Вілкоксона

Контрольні питання до підрозділу 3.3

1. Що означає статистична однорідність геологічного об’єкта?
2. У чому полягає різниця між «фоновою» та «аномальною» сукупностями?
3. Опишіть порядок застосування критерію Граббса – Смирнова.
4. Опишіть порядок застосування правила трьох сигм. Чому воно має таку назву?
5. Чому, на вашу думку, критерії Ван-дер-Вардена та Вілкоксона називають непараметричними?
6. Для чого виконується ранжування при застосуванні критеріїв Ван-дер-Вардена та Вілкоксона?
7. Чим відрізняються критерії Ван-дер-Вардена і Вілкоксона?

3.4. Дисперсійний аналіз

Властивості геологічних об’єктів зазвичай залежать від ряду факторів, що зумовлюють їх мінливість. Виявити ці фактори й оцінити їх вплив на мінливість властивостей чи неоднорідність досліджуваних об'єктів можна за допомогою дисперсійного аналізу.

Дисперсійний аналіз призначений для дослідження характеру впливу на вимірювану випадкову величину (відгук) одного або кількох незалежних факторів, що мають кілька градацій. При однофакторному чи двофакторному аналізі фактори, що впливають на результат, вважаються відомими, а завдання полягає в оцінюванні наявності та суттєвості цього впливу.

Дисперсійний аналіз застосовують, якщо для вибіркових груп можна допускати відповідність генеральним сукупностям з нормальним розподілом і незалежність розподілів спостережень у групах.

За допомогою дисперсійного аналізу можна вирішувати широке коло геологічних завдань [2,4], таких як:

1) перевірка гіпотез про вплив літологічних, петрофізічних, геохімічних, структурних та інших факторів на формування зрудніння;

2) виявлення елементів зональності геологічних об’єктів, зокрема, у закономірній зміні властивостей чи розподілі концентрацій різних компонентів;

3) визначення впливу способу відбору проб на їх достовірність;

4) оцінювання впливу ландшафтних умов на інтенсивність прояву різних пошукових ознак;

5) виявлення чинників, що визначають властивості міцності ґрунтів і порід.

Однофакторний дисперсійний аналіз полягає у розділенні сукупності кількох вимірів досліджуваної ознаки на m груп за якимось фактором. Якщо кількість вимірів у всіх групах (вибірках) однакова, то такий аналіз називається рівномірним; якщо ж кількість вимірів у групах (вибірках) різна, такий дисперсійний аналіз називається нерівномірним.

Спочатку розглянемо співвідношення для рівномірного аналізу. Оцінюється вплив фактора A, що має m рівнів при кількості q вимірів випадкової величини x на кожному рівні. Позначимо результати спостережень через x_{ij}, де i – номер спостереження ($i = 1,...,q$), а j – номер рівня фактора ($j = 1,...,m$) і запишемо їх у вигляді табл. 3.4.

Таблиця 3.4

Табличне представлення даних для однофакторного аналізу

Номер	Рівень фактора A			
	1-й	2-й	…	m-й
1	x_{11}	x_{12}	…	x_{1m}
2	x_{21}	x_{22}	…	x_{2m}
…	…	…	…	…
q	x_{q1}	x_{q2}	…	x_{qm}
Групові середні	$\bar{x}_1$	$\bar{x}_2$	…	$\bar{x}_m$

За цими даними розраховується дисперсія відхилень спостережень від загального середнього $\bar{x}$ та дисперсія між групами

$$S_2^2 = \frac{1}{n-m}\sum_{i=1}^{m}\sum_{j=1}^{q}\left(x_{ij} - \bar{x}\right)^2, \tag{3.19}$$

$$S_1^2 = q\sum_{i=1}^{m}\left(\bar{x}_i - \bar{x}\right)^2. \tag{3.20}$$

Якщо фактор, за яким було проведено розділення, не впливає на мінливість показника, то відношення дисперсій S_1^2 і S_2^2 буде розподілено за законом Фішера з $m - 1$ і $n - m$ ступенями вільності. Гіпотеза про вплив даного фактора відхиляється, якщо

$$\frac{S_1^2}{S_2^2} \le F_{m-1,n-m,\alpha}, \tag{3.21}$$

де $F_{m-1,n-m,\alpha}$ – табличне значення розподілу Фішера для заданої довірчої ймовірності та ступенів вільності (додаток Г).

Якщо об'єми груп (вибірок) відрізняються, то обчислення проводяться за формулами для нерівномірного дисперсійного аналізу. Позначимо через $q_1, q_2, ..., q_m$ кількість елементів у групах. Загальна сума квадратів відхилень визначається за формулою

$$C_g = \left(U_1 + U_2 + ... + U_m\right) - \frac{\left(V_1 + V_2 + ... + V_m\right)^2}{n}, \tag{3.22}$$

де $U_1, U_2, \ldots, U_m$ – суми квадратів елементів вибірок, $U_j = \sum_{i=1}^{q_j} x_{ij}^2$; $V_1, V_2, \ldots, V_m$ – суми елементів *j*-ї вибірки, $V_j = \sum_{i=1}^{q_j} x_{ij}$; n – загальна кількість елементів вибірок, $n = q_1 + q_2 + ... + q_m$.

Факторна сума квадратів відхилень обчислюється за формулою

$$C_f = \left(\frac{U_1^2}{q_1} + \frac{U_2^2}{q_2} + ... + \frac{U_m^2}{q_m}\right) - \frac{\left(V_1 + V_2 + ... + V_m\right)^2}{n}. \tag{3.23}$$

Залишкова сума квадратів відхилень розраховується за рівнянням

$$C_r = C_g - C_f. \tag{3.24}$$

Значення критерію Фішера

$$F = \frac{S_f^2}{S_r^2}, \tag{3.25}$$

де $S_g^2 = \frac{C_g}{n-1}$, $S_f^2 = \frac{C_f}{m-1}$, $S_r^2 = \frac{C_r}{n-m}$,

порівнюється з критичним для заданого рівня значущості α та числа ступенів вільності $k_1 = m - 1$ та $k_2 = n - m$. Якщо $F > F(\alpha, k_1, k_2)$, то можна стверджувати, що фактор відмінності є статистично значущим і його слід брати до уваги. Якщо $F < F(\alpha, k_1, k_2)$, то аналізований фактор не є значущим.

Приклад 3.6 [16]. Осереднені результати вимірювань вмісту азоту і фосфору в донних відкладах водоймищ на різних глибинах наведені в табл. 3.7. Значення на глибинах до 5 м одержані на основі 15 вимірів, у діапазоні глибин 5 – 15 м – 20 вимірів, на глибинах понад 15 м – 18 вимірів.

Для азоту дисперсія між групами $S_2^2 = 1066{,}55$, тоді відношення S_1^2 / S_2^2 дорівнює 26,63, що перевищує табличне значення $F_{2,50,0,05} = 19{,}47$. Отже, фактор глибини суттєво впливає на розподіл азоту в донних відкладах. Для фосфору

дисперсія між групами $S_2^2 = 17{,}46$, тоді відношення S_1^2 / S_2^2 дорівнює 53,28, що перевищує табличне значення $F_{2,50,0,05} = 19{,}47$. Отже, фактор глибини суттєво впливає також і на розподіл фосфору в донних відкладах.

Таблиця 3.5

Середні значення вмісту азоту й фосфору в донних відкладах (мг/дм^3)

Компонент	Глибина h, м			Середнє загальне $\overline{x}$	Дисперсія всієї вибірки S_1^2
	$h < 5$	$5 < h < 15$	$h > 15$		
N загальний	33,64	48,21	36,83	40,22	40,05
P загальний	1,82	2,37	3,21	2,50	0,38

Приклад застосування нерівномірного однофакторного дисперсійного аналізу продемонстровано у практичній роботі 2 (підрозділ 5.2).

Двофакторний аналіз використовується для того, щоби дослідити одночасний вплив двох факторів. Виділяється m рівнів впливу за першим фактором (позначимо його A) та q рівнів впливу – за другим (позначимо його B); тоді загальна кількість груп L дорівнюватиме $m\,q$, а вихідні дані можна представити у табличному вигляді, табл. 3.6. В ній позначено: x_{ij} – середні значення спостережної величини x у вибірці, яка відповідає i-му впливу фактора A та j-му впливу фактора $Б$; $\overline{x}_{.j}$ – середні за стовпчиками; $\overline{x}_{i.}$ – середні за рядками, $\overline{x}$ – загальне середнє.

Таблиця 3.6

Представлення даних для двофакторного дисперсійного аналізу

Рівні фактора A	Рівні фактора $Б$				Середнє
	$Б_1$	$Б_1$	…	$Б_q$	
A_1	x_{11}	x_{12}		x_{1q}	$\overline{x}_{1.}$
A_2	x_{21}	x_{22}		x_{2q}	$\overline{x}_{2.}$
…					
A_m	x_{m1}	x_{m2}		x_{mq}	$\overline{x}_{m.}$
Середнє	$\overline{x}_{.1}$	$\overline{x}_{.2}$		$\overline{x}_{.q}$	$\overline{x}$

Перевірка гіпотези про вплив двох факторів окремо і їх спільного впливу на властивості досліджуваних об’єктів проводиться за критерієм Фішера. Якщо для деяких груп наявні повторні вимірювання, тобто є кілька вимірів для

комбінацій впливу факторів *ij*, то відповідно вплив факторів оцінюється за формулами:

$$F_A = S_1^2 / S_4^2 ;\ F_Б = S_2^2 / S_4^2 ;\ F_{АБ} = S_3^2 / S_4^2 , \tag{3.26}$$

де F_A, $F_Б$, $F_{АБ}$ – статистики Фішера (значення *F*-критерію) для визначення впливу фактора А, фактора *Б* та спільного впливу факторів *А* і *Б* відповідно, дисперсії $S_1^2, S_2^2, S_3^2, S_4^2$ обчислюються згідно з табл. 3.7.

Якщо для всіх груп повторні вимірювання відсутні, то вплив факторів оцінюється за формулами:

$$F_A = S_1^2 / S_3^2 ;\ F_Б = S_2^2 / S_3^2 , \tag{3.27}$$

де всі позначення взяті з попередніх формул.

Отримані значення *F*-критерію порівнюються з критичним для заданого рівня значущості та числа ступенів вільності, на підставі чого робиться висновок щодо статистичної значущості певного фактора.

Таблиця 3.7

Формули для двофакторного дисперсійного аналізу

Тип дисперсії	Дисперсія	Кількість ступенів вільності
Факторна за фактором *А*	$S_1^2 = \frac{q}{m-1}\sum_{i=1}^{m}(\overline{x}_{i..} - \overline{x})^2 , \quad x_{i..} = \frac{1}{q}\sum_{j=1}^{q}\overline{x}_{ij}$	$m-1$
Факторна за фактором *Б*	$S_2^2 = \frac{m}{q-1}\sum_{j=1}^{q}(\overline{x}_{.j.} - \overline{x})^2 , \quad x_{.j.} = \frac{1}{m}\sum_{i=1}^{m}\overline{x}_{ij}$	$q-1$
Змішана за факторами *АБ*	$S_3^2 = \frac{1}{(m-1)(q-1)}\sum_{i=1}^{m}\sum_{j=1}^{q}(x_{ij.} - \overline{x}_{i..} - \overline{x}_{.j.} + \overline{x})^2 ,$ $x_{ij.} = \frac{1}{n}\sum_{j=1}^{n}x_{ijk}$	$(m-1)(q-1)$
Залишкова	$S_4^2 = \frac{1}{mq(n-1)}\sum_{i=1}^{m}\sum_{j=1}^{q}\sum_{k=1}^{n}(x_{ijk} - \overline{x}_{ij.})^2$	$mq(n-1)$
Загальна	$S^2 = \frac{1}{nmq-1}\sum_{i=1}^{m}\sum_{j=1}^{q}\sum_{k=1}^{n}(x_{ijk} - \overline{x})^2$	$nmq-1$

Тут x_{ijk} – елементи вибірки, яка відповідає *i*-му впливу фактора *А* та *j*-му впливу фактора *Б*.

Приклад 3.7 [2]. На основі результатів польових досліджень полі-металевого родовища (табл. 3.8) оцінити, чи впливає на вміст свинцю в рудних

жилах склад вміщувальних порід та гіпсометричне положення місць відбору проб.

Відповідно до табл. 3.8 розрахуємо дисперсії ($m = 4$, $q = 5$):

$$S_1^2 = \frac{q}{m-1}\sum_{i=1}^{m}(\overline{x}_{i\bullet} - \overline{x})^2 = \frac{5}{3}\cdot 0,422 = 0,703,$$

$$S_2^2 = \frac{m}{q-1}\sum_{j=1}^{q}(\overline{x}_{\bullet j} - \overline{x})^2 = \frac{3\cdot 0,225}{4} = 0,169,$$

$$S_3^2 = \frac{1}{(m-1)(q-1)}\sum_{i=1}^{m}\sum_{j=1}^{q}(x_{ij\cdot} - \overline{x}_{i\bullet} - \overline{x}_{\bullet j} + \overline{x})^2 = \frac{1,941}{3\cdot 4} = 0,162.$$

Таблиця 3.8

Усереднений вміст свинцю в гірських породах (у %)
на ділянці поліметалевого родовища

Позначка горизонту відбору проб, м	Склад вміщувальних порід					
	Граніти	Кварцити	Глинисті сланці	Вапняки	Доломіти	Середнє
+320	2,95	2,66	2,50	3,15	2,25	2,70
+280	2,25	2,71	2,38	2,56	1,81	2,34
+240	2,38	1,78	1,98	1,21	2,19	1,91
+200	2,18	1,65	2,46	2,10	1,32	1,94
У середньому за всіма горизонтами	2,44	2,20	2,33	2,26	1,89	2,22

Табл. 3.8 не містить повторних значень для вибірок, тобто $n = 1$, тому і S_4^2 не обчислюється, а перевірка значущості впливу факторів проводиться за формулами (3.27).

Критичні значення критерію Фішера для рівня значущості 0,05 розраховуються так:

$F((m-1), (q-1)(m-1), 0,05) = F(3, 12, 0,05) = 3,49$ для перевірки впливу фактора гіпсометричної позначки (за рядками);

$F((q-1), (q-1)(m-1), 0,05) = F(4, 12, 0,05) = 3,25$ для перевірки впливу складу вміщувальних порід (за стовпчиками).

$F_A = S_1^2 / S_3^2 = 4,37$; $F_Б = S_2^2 / S_3^2 = 0,81$.

Оскільки $F_A > F(3, 12, 0,05)$, то гіпотеза про відсутність впливу гіпсометричної позначки на вміст свинцю відхиляється і можна стверджувати, що вплив цього фактора є значущим. Але $F_Б < F(4, 12, 0,05)$, тому гіпотеза про

відсутність впливу складу порід на вміст свинцю приймається. Таким чином, можна зробити висновок: на дослідженій ділянці поліметалевого родовища статистично значущим впливом є гіпсометрична позначка, тобто глибина відбору проб.

У пакеті аналізу програми Excel реалізовані стандартні процедури для однофакторного і двофакторного дисперсійних аналізів різних видів. Ці процедури застосовні, якщо властивості вибірок не відхиляються суттєво від нормального закону розподілу, а дисперсії на різних рівнях того самого фактора однакові. Навіть у разі помірної відмінності в дисперсіях критерій залишається застосовним за умови приблизної рівності об'ємів вибірок у групах. Якщо ж застосування *F*-критерію викликає сумніви, можна скористатися непараметричними критеріями, зокрема, однофакторним дисперсійним аналізом за критерієм Краскала – Уолліса та двофакторним дисперсійним аналізом за критерієм Фрідмана [4].

Контрольні питання до підрозділу 3.4

1. Для чого використовується дисперсійний аналіз?
2. У чому полягає різниця між однофакторним та двофакторним дисперсійними аналізами?
3. Чим неоднорідний дисперсійний аналіз відрізняється від однорідного? Як це враховується в розрахунках?
4. Яка мінімальна кількість факторів та рівнів їх впливу має бути для проведення двофакторного дисперсійного аналізу?

3.5. Виявлення неоднорідностей та аномалій на площині

Поширеним завданням у геології та екології є ідентифікація розподілу точок на площині або двовимірній поверхні. Цими точками можуть бути, зокрема, місця спостережень, взяття проб, прояви руд кольорових та рідкоземельних металів, положення нафтовидобувних свердловин, місця значних водопроявів у підземних виробках тощо. Зазначимо, що рівномірність розподілу точок спостережень є одним з критеріїв застосування вибіркових методів досліджень (розділ 2). Першим етапом аналізу просторового розташування точок є перевірка гіпотези про однорідність їх розподілу на площині, що може бути зроблено за допомогою критерію χ^2 (підрозділ 2.4).

Для цього досліджувана область ділиться на підобласті однакової площі, бажано, однієї форми і підраховується кількість точок, які потрапили до кожної

підобласті. Як зазначено у [6], найбільш обґрунтованими на основі застосування цього критерію будуть висновки у разі значної кількості підобластей, а також, коли в кожній з них буде знаходитися не менше п’яти точок, хоча друга вимога не завжди є виконуваною.

Критерій перевірки гіпотези щодо рівномірного розподілу точок на площині полягає в порівнянні обчисленої статистики χ^2 з критичним значенням відповідного розподілу. Статистика χ^2 обчислюється за формулою, подібною до (2.73)

$$\chi^2 = \sum_{i=1}^{m} \frac{(q_i - p_i)^2}{p_i}, \tag{3.28}$$

де q_i – фактична кількість точок у i-й підобласті; p_i – очікувана кількість точок у i-й підобласті, яка обчислюється за виразом

$$p_i = \frac{N}{m}, \tag{3.29}$$

m – кількість підобластей; N – загальна кількість точок.

Критерій χ^2 обчислюється за $m - 2$ ступенями вільності.

Приклад 3.8. Розглянемо розташування рудних проявів на карті геологічної розвідки (рис. 3.8). Розіб’ємо всю прямокутну область на 16 квадратів однакового розміру. У табл. 3.9 показано кількість точок, які потрапили до певного квадрату, та відповідні відхилення, що необхідні для розрахунку за формулою (3.28).

Ймовірне число свердловин у кожному квадраті становить $\xi = 56/16 = 3{,}5$. Зважаючи на це, обчислена сума за формулою (3.28) дорівнює 32,57 (табл. 3.9), що перевищує критичне число 23,7 розподілу χ^2 для $m = 16 - 2 = 14$ ступенів вільності з рівнем значущості $\alpha = 0{,}05$ (додаток Б). Тому гіпотеза про рівномірність розподілу свердловин на площині відхиляється. Навіть для більш строгих рівнів значущості 0,025 та 0,001 (значення 26,1 та 29,1 відповідно) розподіл свердловин у плані можна вважати нерівномірним.

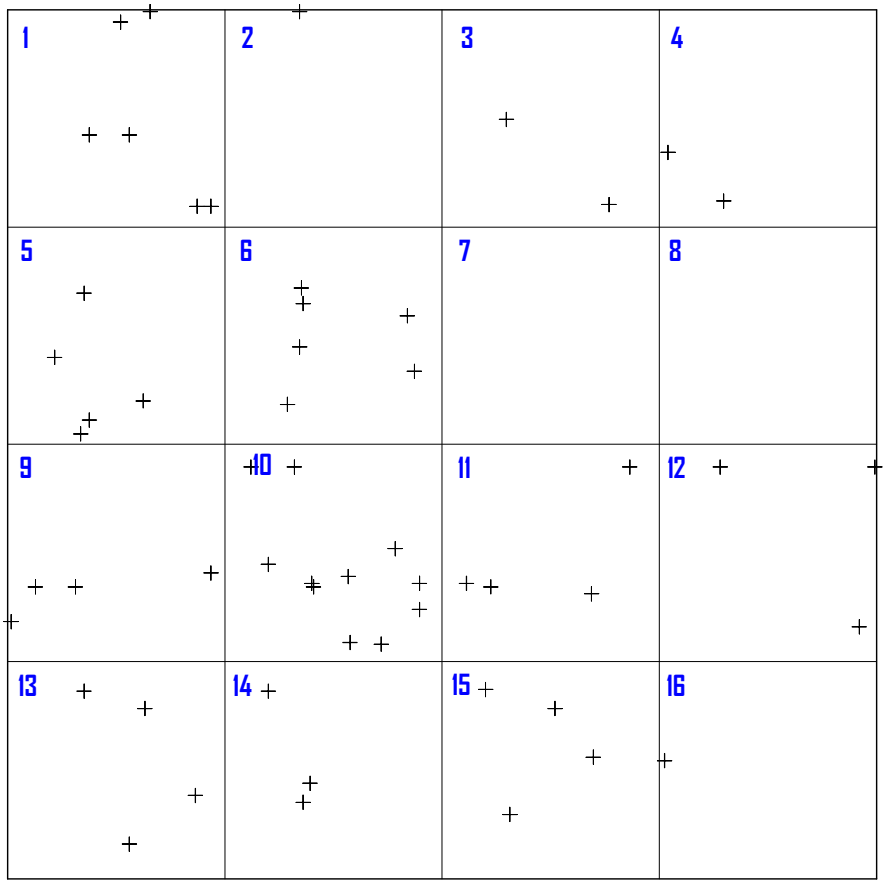


Рис. 3.8. Розподіл рудних проявів на карті геологічної розвідки

Таблиця 3.9

Перевірка гіпотези про нерівномірність розподілу свердловин на площі

№ квадрата, i	Кількість свердловин	Відхилення $(q_i - p_i)^2 / p_i$
1	6	1,79
2	1	1,79
3	2	0,64
4	2	0,64
5	5	0,64
6	6	1,79
7	0	3,50
8	0	3,50
9	4	0,07
10	11	16,07
11	4	0,07
12	3	0,07
13	4	0,07
14	3	0,07
15	4	0,07
16	1	1,79
Разом	**56**	**32,57**

Інший, відносно простий, алгоритм описано в [2]. Згідно з ним досліджувана територія розбивається на квадратні комірки однакового розміру. Частина p комірок буде містити хоча б одну точку, а інша частина $(1-p)$ залишиться порожньою. Після цього початкові комірки діляться на N нових квадратних комірок. Якщо точкові об'єкти розташовані випадково, то ймовірність того, що новий квадрат буде порожнім, є

$$p_N = (1-p)^N. \tag{3.30}$$

Для кожного $N = 4, 9, 16\ldots$ знаходимо фактичну частку порожніх комірок p'_N. Підвищена частка порожніх комірок порівняно з теоретичною ймовірністю p_N свідчить про групування об'єктів, а знижена – про регулярність їх розташування на площині.

Приклад 3.9. [17]. Для перевірки гіпотези про рівномірний розподіл мідних родовищ в Аризоні (США) досліджувана область була розділена на квадрати площею 250 км^2, які є чвертю від комірок сітки на рис. 3.9. Залежності величин p'_N та p_N на рис. 3.10 свідчать, що фактична кількість порожніх комірок істотно нижче за теоретичну, тому розподіл родовищ на території не є випадковим.

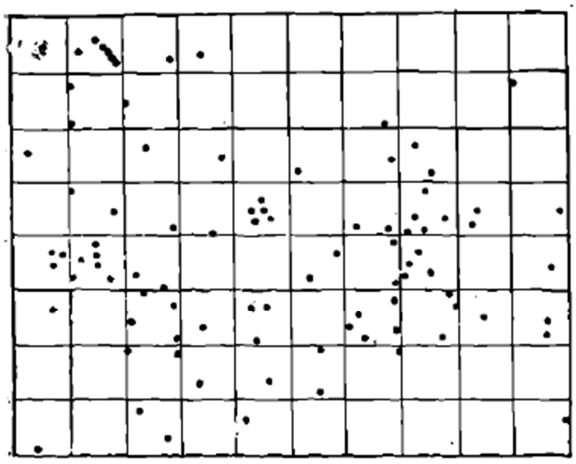


Рис. 3.9. Схема розташування родовищ міді в західній частині штату Аризона (США)

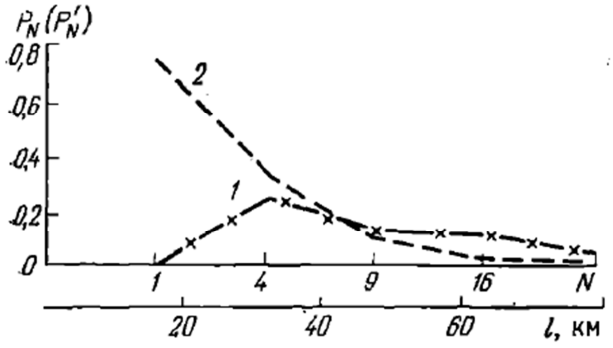


Рис. 3.10. Залежність частки порожніх комірок від довжини сторони квадрату: 1 – фактична, 2 – теоретична

Після визначення нерівномірності розподілу постає питання про локалізацію ділянок, де розташована більшість свердловин, тобто про локалізацію аномалій досліджуваної ознаки чи характеристики на площині.

Для виділення аномалій (скупчення точок) на площині чи у просторі використовуються різноманітні методи; практичне застосування деяких з них описано, зокрема у [18]. Зауважимо, що для дослідника часто важливе попереднє виділення таких аномалій, які далі можна деталізувати за допомогою більш потужних методів. Саме для попереднього визначення аномалій можуть бути використані зручні й нескладні критерії, наприклад перевірка параметрів біноміального розподілу.

Необхідно перевірити гіпотезу про те, що місця проявів якихось процесів, явищ чи ознак концентруються в певних зонах. Це може бути концентрування рудних тіл біля певних геологічних формацій, переважання підземних викидів вугілля, породи й газу та інтенсивних водопритоків до підземних виробок поблизу зон геологічних порушень, накопичення хімічних елементів у певних формах рельєфу тощо.

Для перевірки цього всю територію можна розділити на дві частини: зону впливу окремої неоднорідності площею S_1 та всю іншу частину території площею S_2. Якщо неоднорідність не впливає на просторове розташування точок прояву певної ознаки, що визначає «нульову» гіпотезу, то ймовірності потрапляння точок до обох зон пропорційні їх площі, тобто

$$p_1 = \frac{S_1}{S_1 + S_2},\ p_2 = \frac{S_2}{S_1 + S_2}. \tag{3.31}$$

Фактичну кількість точок, що потрапили до зони впливу неоднорідності, позначимо через m, загальну кількість – через n, тоді відповідна частота потрапляння становитиме

$$p_1^* = \frac{m}{n},\ p_2^* = \frac{n-m}{n}. \tag{3.32}$$

Оскільки математичне сподівання біноміального розподілу з імовірністю p_1 дорівнює $n\,p_1$, а дисперсія – $n\,p_1\,(1-p_1)$ (див. підрозділ 2.3.8), то випадкова величина

$$t_1 = \frac{p_1 - p_1^*}{\sqrt{p_1\left(1-p_1\right)/n}} \tag{3.33}$$

розподілена нормально з середнім значенням 0 та дисперсією 1. Тоді, якщо

$$t_1 > t_\alpha, \tag{3.34}$$

де t_α – аргумент функції нормованого нормального розподілу з параметрами (0, 1) (2.50), що відповідає значенню α, то з довірчою ймовірністю $(1-\alpha)$ можна вважати, що точки проявів досліджуваної ознаки концентруються навколо просторової неоднорідності.

Приклад 3.10. Необхідно встановити, чи впливає дві лінійні неоднорідності (геологічні порушення), які розташовані майже під прямим кутом, на розподіл рудопроявів на досліджуваній території (рис. 3.11). Для визначеності приймемо розмір квадратної ділянки 5 × 5 км і порівняємо частоту і ймовірність потрапляння точок до зони впливу навколо порушень у вигляді смуги шириною $l = 200$ м (по 100 м в обидві сторони) та $l = 400$ м (по 200 м в обидві сторони).

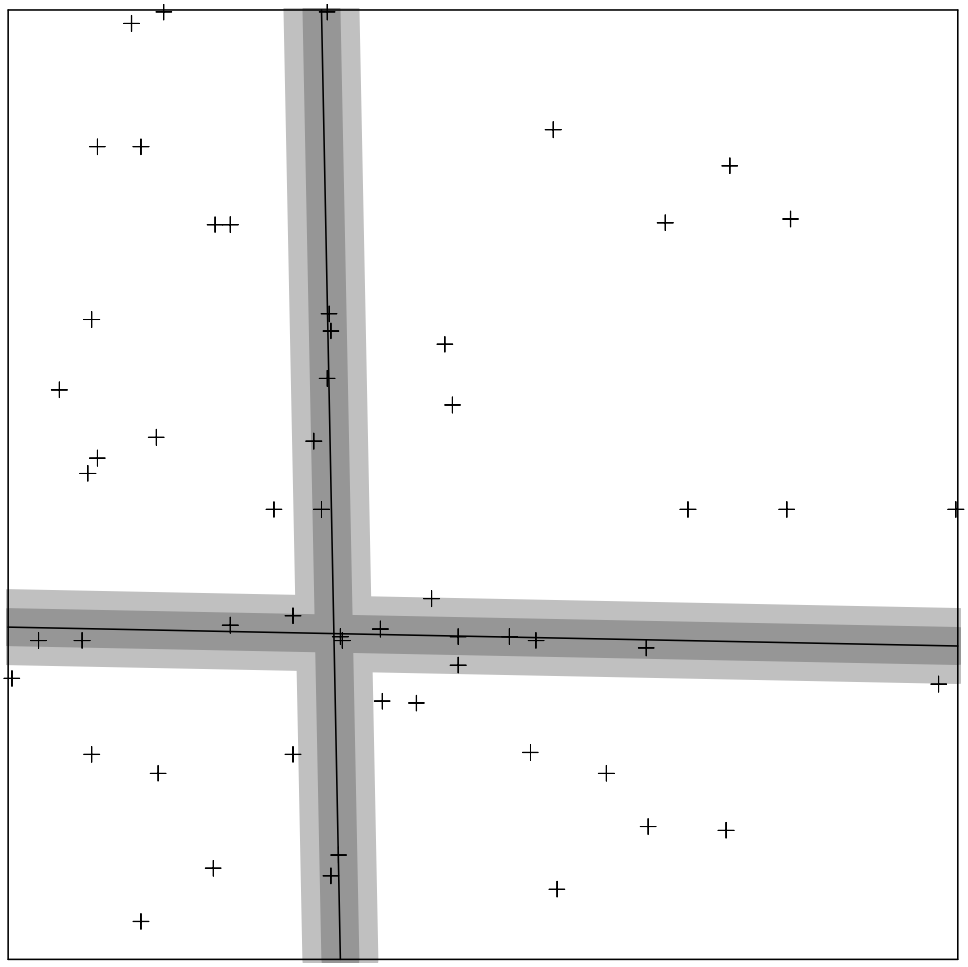


Рис. 3.11. Положення точок у зонах впливу навколо лінійних геологічних порушень (до прикладу 3.10)

Обчислені на основі табл. 3.10 значення статистики (3.33) значно перевищують критичне значення $t_\alpha = 3{,}1$ з достатньо високим рівнем значущості $\alpha = 0{,}001$ в обох випадках. Тому з довірчою ймовірністю понад 0,999 можна вважати, що точки з рудопроявами закономірно концентруються в зоні впливу геологічних порушень.

Таблиця 3.10

Результати перевірки гіпотези про біноміальний розподіл положення точок на досліджуваній площі

Зона впливу неоднорідності	Широка смуга, $l = 400$ м	Вузька смуга, $l = 200$ м
Теоретична ймовірність потрапляння до зони впливу p_1	0,1536	0,0784
Кількість точок у зоні впливу m	21 з 56	19 з 56
Фактична частота p_1^*	0,375	0,339
Статистика критерію t_1	4,59	7,26

Контрольны питання до підрозділу 3.5

1. Як застосовується критерій χ^2 для визначення аномалій на площині?
2. Опишіть процедуру перевірки гіпотези про рівномірний розподіл точок спостережень (рудопроявів тощо) на площині.
3. Як обґрунтувати статистичну значущість групування (скупчення) точок в окремих зонах досліджуваної області?

3.6. Багатовимірні статистичні моделі. Кластерний аналіз

Велика кількість спостережень і вимірів процесів у навколишньому середовищі та у геосферах записується у вигляді матриць, стовпці яких відповідають вимірюваним ознакам чи параметрам, а рядки – конкретним вимірам. Отже, розмірність матриць визначається кількістю вимірюваних параметрів та самих вимірів і може сягати кількох десятків – сотень. Обробка таких масивів даних потребує спеціальних методів, серед яких найчастіше вживаними є множинна кореляція, множинна регресія, факторний та кластерний аналіз, розпізнавання образів тощо [2, 19, 20, 21].

Множинна кореляція є одним з перших кроків обробки матриць вимірюваних даних. Математичною моделлю комплексу вимірюваних ознак є багатовимірна випадкова величина – випадковий вектор $\boldsymbol{x}$ розмірності m з компонентами $x_1, x_2, \ldots, x_m$, де m – кількість вимірюваних ознак. Вектор $\boldsymbol{x}$ може розглядатися як вектор-рядок або вектор-стовпець.

Результати n спостережень записуються в матриці X

$$X = \begin{Vmatrix} x_{11} & \dots & x_{1m} \\ x_{21} & \dots & x_{2m} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{nm} \end{Vmatrix}. \tag{3.35}$$

Статистичні властивості матриці визначаються коваріаційною матрицею

$$Q = \begin{Vmatrix} \operatorname{cov}(x_1, x_1) & \dots & \operatorname{cov}(x_1, x_m) \\ \operatorname{cov}(x_2, x_1) & \dots & \operatorname{cov}(x_2, x_m) \\ \dots & \dots & \dots \\ \operatorname{cov}(x_n, x_1) & \dots & \operatorname{cov}(x_n, x_m) \end{Vmatrix}, \tag{3.36}$$

де

$$\operatorname{cov}(x_i, x_j) = \frac{1}{n}\sum_{k=1}^{n}(x_{ik} - \overline{x}_i)(x_{jk} - \overline{x}_j), \tag{3.37}$$

$\overline{x}_i$ – середнє значення всіх елементів *i*-го стовпця матриці. Діагональні елементи матриці при $i = j$ є оцінками дисперсії.

Позначимо

$$r(x_i, x_j) = \frac{\operatorname{cov}(x_i, x_j)}{\sigma_{xi}\sigma_{xj}} \tag{3.38}$$

як коефіцієнт кореляції двох векторів x_i та x_j. Тоді можна перейти до кореляційної матриці

$$R = \begin{Vmatrix} 1 & \dots & r(x_1, x_m) \\ r(x_2, x_1) & \dots & r(x_2, x_m) \\ \dots & \dots & \dots \\ r(x_n, x_1) & \dots & 1 \end{Vmatrix}, \tag{3.39}$$

усі елементи якої є коефіцієнтами кореляції, а діагональні елементи дорівнюють одиниці. Аналіз кореляційної матриці дозволяє зробити перші висновки щодо зв'язків між досліджуваними ознаками.

Приклад 3.11. Визначимо кореляційні зв'язки між компонентами хімічного складу підземних вод у зоні впливу відстійника шахтних вод на прикладі масиву вимірювань у 25 свердловинах (рис. 3.12).

З використанням процедури «Correlation» у надбудові в MS Excel «Data Analysis» обчислимо кореляційну матрицю, а за критерієм Стьюдента перевіримо статистичну значущість її елементів (рис. 3.13).

	A	B	C	D	E	F	G	H
1	Свердл.	pH	CO2 в.	Окисн.	HCO3-	Cl-	SO42-	NO3-
2	C1	3.45	1388.00	114.74	1.00	1897.61	511.76	0.80
3	C2	3.99	121.44	27.47	1.00	1891.76	37.92	4.96
4	C3	8.31	1.00	24.89	41.50	12123.00	680.24	0.62
5	C4	6.00	1.00	7.43	34.18	2554.17	103.20	0.62
6	C5	3.22	12798.00	993.84	1.00	21021.82	298.56	0.50
7	C6	3.10	5620.00	460.56	1.00	3215.32	925.92	6.82
8	C7	3.78	308.00	32.32	1.00	1690.06	324.48	7.34
9	C8	7.59	1.00	7.76	379.61	1225.86	636.96	6.20
10	C9	4.20	41.62	23.43	1.00	1981.60	939.84	4.34
11	C10	5.78	8.80	5.17	73.24	592.02	131.52	1.81
12	C11	8.15	1.00	4.69	47.60	660.79	25.44	1.24
13	C12	7.13	1.00	5.49	41.50	851.15	320.16	0.62
14	C13	8.05	1.00	4.85	41.50	1007.49	1121.28	0.62
15	C14	8.90	1.00	16.81	102.53	3590.73	2522.40	3.72
16	C15	3.84	89.40	383.80	1.00	18713.93	24.48	0.50
17	C16	7.69	1.00	17.78	35.40	7972.71	404.16	0.30
18	C17	7.98	1.00	21.98	20.75	15112.34	24.00	4.34
19	C18	8.30	1.00	6.14	18.31	1461.96	18.24	4.96
20	C19	8.29	1.00	16.51	47.60	3476.58	1070.40	4.96
21	C20	7.33	1.00	18.09	31.74	9166.66	192.48	6.20
22	C21	7.11	1.00	4.20	93.90	180.44	16.32	3.10
23	C22	8.92	1.00	5.82	52.04	869.53	732.48	3.72
24	C23	4.45	52.80	25.86	1.00	13499.36	1191.84	0.31
25	C24	4.63	206.40	36.84	1.00	16544.52	1124.96	4.34
26	C25	4.21	242.60	27.10	1.00	3203.94	675.84	0.60

Рис. 3.12. Фрагмент робочого аркуша MS Excel з результатами гідрохімічного аналізу у 25 свердловинах. Позначення: pH – показник кислотності, CO2 в. – вміст вільного оксиду вуглецю, Окисн. – окиснюваність, HCO3-, Cl-, SO42-, та NO3- – відповідно вміст гідрокарбонатів, хлоридів, сульфатів та нітратів у воді; значення всіх параметрів, крім pH, наведені у мг/дм^3

Найбільш тісним є кореляційний зв'язок між параметрами окиснюваності та вмісту вільного оксиду вуглецю, окиснюваності та вмісту хлоридів, окиснюваності та кислотності. Менш тісний, але статистично значущий зв’язок підтверджується для пар таких параметрів, як кислотність та вміст вільного оксиду вуглецю (зворотний зв’язок), кислотність та вміст гідрокарбонатів (зворотний зв’язок), вміст вільного оксиду вуглецю та хлоридів. Взаємний зв’язок між іншими парами елементів хімічного складу підземної води є статистично незначущий.

	K	L	M	N	O	P	Q	R
1								
2		*pH*	*CO2 в.*	*Окисн.*	*HCO3-*	*Cl-*	*SO42-*	*NO3-*
3	pH	1						
4	CO2 в.	-0.4517126	1					
5	Окисн.	-0.5182349	0.942988947	1				
6	HCO3-	0.413482	-0.174151597	-0.21383	1			
7	Cl-	-0.264311	0.414036852	0.55111	-0.28596	1		
8	SO42-	0.13343784	-0.036910316	-0.08827	0.1125	-0.01817	1	
9	NO3-	0.01635796	-0.072771011	-0.13151	0.23206	-0.23509	0.07357	1

а)

	K	L	M	N	O	P	Q	R
11								
12	n=	25	t_student	2.06866				
13								
14		*pH*	*CO2 в.*	*Окисн.*	*HCO3-*	*Cl-*	*SO42-*	*NO3-*
15	pH							
16	CO2 в.	2.42818409						
17	Окисн.	2.90605251	13.58800891					
18	HCO3-	2.17788386	0.848162616	1.04976				
19	Cl-	1.31433185	2.181410503	3.16743	1.43121			
20	SO42-	0.64571995	0.177136361	0.42501	0.54296	0.08715		
21	NO3-	0.07846052	0.349925274	0.63624	1.14415	1.15995	0.35378	

б)

Рис. 3.13. Кореляційна матриця зв’язків між компонентами хімічного складу підземних вод (*а*) та перевірка статистичної значущості її елементів (*б*)

На основі аналізу кореляційної матриці можна зробити попередній висновок про те, що кислотність підземних вод суттєво зростає в разі потрапляння до водоносних горизонтів вод з високим умістом хлоридів та їх хімічної взаємодії з породами, що призводить до виділення вільного оксиду вуглецю та зменшення показника pH, тобто до зростання кислотності. Вплив сульфатів та нітратів на зміну кислотності є несуттєвим. Більш повний кореляційний аналіз потребує включення до кореляційної матриці інших компонентів хімічного складу води, зокрема, лужних металів Na та K і лужноземельних металів Ca та Mg, а також деяких специфічних елементів, що можуть впливати на кислотність, наприклад Fe.

Множинна регресія використовується для побудови функції, яка описує залежність однієї змінної від кількох незалежних змінних. Рівняння лінійної регресії залежної змінної y від незалежних змінних $x_1, x_2, \ldots, x_m$ може бути записано у вигляді

$$y = a_0 + a_1 x_1 + ... + a_m x_m = a_0 + \sum_{i=1}^{m} a_i x_i \,, \tag{3.40}$$

де a_0, a_1,…, a_m – коефіцієнти, які потрібно визначити за умови, що рівняння (3.40) найкращим способом, тобто з мінімальною сумою квадратів відхилень, описує тенденцію зміни вимірів і дозволяє оцінити спільний вплив усіх досліджуваних параметрів x_1, x_2,…, x_m на залежну змінну y.

Коефіцієнти рівняння (3.40) обчислюються за формулами

$$a_j = \frac{S_y}{S_j} \sum_{i=1}^{m} \frac{r_{iy}}{r_{ij}}, \; a_0 = \overline{y} - \sum_{j=1}^{m} a_j \overline{x}_j \tag{3.41}$$

де S_y – стандартне відхилення залежної змінної y; S_j – стандартне відхилення незалежної змінної x_j; r_{iy} – парна кореляція i-ї незалежної змінної із залежною змінною; r_{iy} – парна кореляція i-ї та j-ї незалежних змінних; $\overline{y}$ та $\overline{x}_j$ – середні значення залежної змінної y та незалежних змінних x_j.

У матричній формі рівняння (3.41) записується у вигляді

$$\left[\sum Y\right] = \left\|\sum X\right\| \cdot [A], \tag{3.42}$$

де $\left[\sum Y\right]$ – вектор-стовпець, що складається із сум квадратів і змішаних добутків змінної Y зі змінними $X_1, X_2,..., X_m$; $\left\|\sum X\right\|$ – матриця сум квадратів і змішаних добутків $X_1, X_2,..., X_m$; $[A]$ – вектор-стовпець невідомих коефіцієнтів регресії.

У загальному випадку використовується рівняння нелінійної множинної регресії, до якого залежні змінні входять як нелінійні функції, наприклад, ступеневі, експоненціальні, логарифмічні тощо.

У практичних задачах з геології та екології моделі множинної регресії використовуються для прогнозування залежної змінної, наприклад, вмісту цінного елементу, об’ємної маси руди, глибини формування мінералу, сумарного техногенного навантаження тощо за комплексом незалежних змінних: вмісту породоутворюючих елементів, об’ємних мас мінералів у рудах, вмісту елементів у мінералах, токсичних речовин в елементах навколишнього середовища та ін.

Приклад 3.12. Використовуючи дані прикладу 3.11, визначимо лінійну множинну регресію для кислотності як функції, яка залежить від вмісту різних елементів у відібраних пробах підземних вод. Зауважимо, що даний приклад є більшою мірою демонстративним, бо при побудові регресії не буде враховуватися залежність від вмісту лужних Na та K і лужноземельних металів Ca та Mg, що може скоригувати отримане рівняння регресії.

З використанням процедури «Regression» у надбудові в MS Excel «Data Analysis» обчислимо коефіцієнти рівняння регресії та величини для оцінки її достовірності (рис. 3.14). Одним з найважливіших параметрів достовірності є показник множинного коефіцієнта кореляції R^2 (що виділено жирним шрифтом та кольором фону), який обчислюється за формулою

$$R^2 = 1 - \frac{|R|}{|R_{11}|}, \tag{3.43}$$

де $|R|$ – визначник кореляційної матриці, $|R_{11}|$ – алгебраїчне доповнення до кореляційної матриці, тобто визначник кореляційної матриці без стовпця значень залежної змінної.

Для оцінки статистичної значущості коефіцієнта кореляції, крім критерію Стьюдента, може використовуватися якісна шкала Чеддока, подана в підрозділі 2.5 у табл. 2.13 [21]. У разі від'ємних значень коефіцієнта кореляції всі висновки залишаються в силі, лише йдеться про зворотний зв'язок.

	K	L	M	N	O	P	Q	R	S
30	SUMMARY OUTPUT								
31									
32	*Regression Statistics*								
33	Multiple R	0.6363349							
34	R Square	**0.40492211**							
35	Adjusted R Square	0.20656281							
36	Standard Error	1.83170279							
37	Observations	25							
38									
39	ANOVA								
40		*df*	*SS*	*MS*	*F*	*Significance F*			
41	Regression	6	41.09416807	6.84903	**2.04136**	**0.11250251**			
42	Residual	18	60.39243193	3.35514					
43	Total	24	101.4866						
44									
45		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Ipper 95.0%*
46	Intercept	**6.27561052**	0.84148267	7.4578	6.6E-07	4.50772103	8.0435	4.50772	8.0435
47	CO2 в.	**0.0003243**	0.000451262	0.71865	0.48158	-0.0006238	0.00127	-0.00062	0.00127
48	Окисн.	**-0.0088577**	0.006158915	-1.43819	0.16754	-0.0217971	0.00408	-0.0218	0.00408
49	HCO3-	**0.00956237**	0.005238772	1.82531	0.08459	-0.0014439	0.02057	-0.00144	0.02057
50	Cl-	**4.7757E-05**	7.71935E-05	0.61866	0.54389	-0.0001144	0.00021	-0.00011	0.00021
51	SO42-	**0.0001359**	0.000666427	0.20392	0.8407	-0.0012642	0.00154	-0.00126	0.00154
52	NO3-	**-0.1088622**	0.164630818	-0.66125	0.51683	-0.4547387	0.23701	-0.45474	0.23701

Рис. 3.14. Фрагмент аркуша MS Excel з розрахунками рівняння лінійної регресії до прикладу 3.12

Відповідно до табл. 2.13 $R^2 = 0{,}405$, отже, можна зробити висновок про слабкий та помірний кореляційний зв'язок, тобто про те, що побудоване рівняння регресії недостатньо добре описує залежність показника кислотності від інших компонентів хімічного складу води. Про це свідчать також результати дисперсійного аналізу нижче (комірки O41, P41), де показане значення статистики Фішера перевірки значущості рівняння регресії, яке не перевищує навіть слабкий рівень значущості $\alpha = 0{,}1$ (обчислене значення 0,112). Нижче у стовпці L показані коефіцієнти рівняння регресії, яке може бути записане у вигляді:

$$y = 6{,}276 + 3{,}2 \cdot 10^{-4} x_1 - 8{,}8 \cdot 10^{-3} x_2 + 9{,}6 \cdot 10^{-3} x_3 + 4{,}8 \cdot 10^{-5} x_4 + 1{,}4 \cdot 10^{-4} x_5 - 0{,}11 x_6,$$

де змінні $x_1, x_2, \ldots, x_6$ відповідають компонентам хімічного складу води, назви яких показано у комірках K47 – K52.

Недостатня значущість рівняння регресії може бути пов'язана, зокрема, з двома чинниками: 1) неврахуванням впливу лужних та лужноземельних металів, 2) фактично нелінійною залежністю показника кислотності від інших величин.

Сутність кластерного аналізу. У практиці геології та екології часто виникають задачі класифікації та типізації певних об'єктів, ділянок, популяцій, структур, процесів тощо. Одним з методів кількісної оцінки ступеня відмінності різних об'єктів є кластерний аналіз [21], який доцільно застосовувати, якщо відсутня вихідна вибірка або попередня інформація про генеральну сукупність ознак. Методами кластерного аналізу можна, зокрема, проводити класифікацію з урахуванням ознак об'єктів і перевіряти структурованість цілої сукупності об'єктів.

Відповідно до суті методу, увесь масив спостережень (вибірка даних) розбивається на *кластери* (угрупування) таким чином, щоби в них об'єднувалися об'єкти найвищої подібності, а роз'єднані групи залишалися при цьому максимально ізольованими (різними). Як критерії подібності можна використовувати парні коефіцієнти кореляції, відстань за певною метрикою або інші коефіцієнти.

Відстані в різних метриках. Для кількісної характеристики відмінності між елементами вибірки використовується поняття метрики. Міра подібності між елементами множин (аналог відстані) називається метрикою, якщо вона задовольняє умови симетрії, нерівності трикутника, розрізнення нетотожних об'єктів і непомітності тотожних об'єктів.

Визначимо далі найбільш застосовні метрики, позначивши через X та Y два довільних різних елементи вибірки, що відповідно мають n ознак: x_i та y_i,

$i = 1, \ldots, n$. Для точок на площині або у просторі x_i та y_i інтерпретуються як координати; якщо X та Y є, наприклад, пробами зразків ґрунту або води, то x_i та y_i можуть інтерпретуватися як вміст різних елементів у відповідних зразках.

Найбільш загальною метрикою є метрика Мінковського

$$dist_M(X,Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^r \right)^{1/r}. \quad (3.44)$$

де r – показник, який рекомендується змінювати в діапазоні від 1 до 4. Якщо $r = 2$, то метрика Мінковського переходить у евклідову метрику, яка фактично є геометричною відстанню в багатовимірному просторі або найменшою відстанню (довжиною відрізку) між двома точками на площині або у просторі. Відстань між двома елементами вибірки за евклідовою метрикою визначається за формулою

$$dist_E(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \quad (3.45)$$

Якщо не всі елементи вибірки однаково вагомі, наприклад, мають різну достовірність, то використовується зважена відстань за евклідовою метрикою

$$dist_{E,w}(X,Y) = \sqrt{\sum_{i=1}^{n} w_i (x_i - y_i)^2}, \quad (3.46)$$

де вага w_i визначається окремо відповідно до інформації щодо значущості i-го вимірювання.

Манхетенська відстань є окремим випадком відстані за метрикою Мінковського при $r = 1$:

$$dist_{Md}(X,Y) = \sum_{i=1}^{n} |x_i - y_i|. \quad (3.47)$$

Ця відстань обчислюється як сумарне розходження між відповідними вимірами. Манхетенську відстань ще називають метрикою міста чи образно дистанцією таксиста, як шлях, який має подолати таксист у районі Манхетен м. Нью-Йорка вздовж вулиць, що перетинаються під прямим кутом. Тому в двовимірному просторі це не прямолінійна евклидова відстань, а сума окремих відстаней – довжин кварталів (рис. 3.15). У манхетенській метриці всі відстані вздовж вулиць між двома точками А та В однакові й не дорівнюють евклідовій відстані.

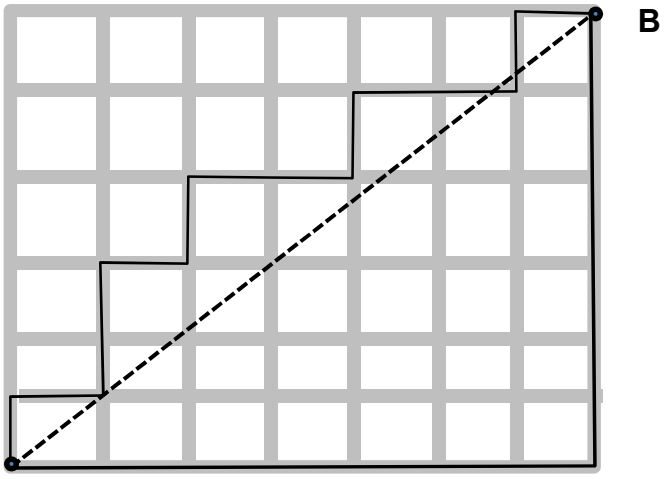


Рис. 3.15. Порівняння евклідової та манхетенської відстаней між пунктами A і B у місті з вулицями, що перетинаються під прямим кутом

Чебишевська відстань використовується для того, щоб визначити два об’єкти як різні в разі, коли вони різні за будь-якою ознакою. Різницею двох спостережень є абсолютне значення максимальної різниці відповідної ознаки

$$dist_{Ch}\left(X,Y\right)=\max_{i}\left|x_i-y_i\right|. \tag{3.48}$$

Для вибірок, які мають категоріальні (некількісні) ознаки, відстань між двома елементами можна визначити як частку або відсоток відмінності

$$dist_{dif}\left(X,Y\right)=\frac{n_{x_i\neq y_i}}{n}, \tag{3.49}$$

де $n_{x_i\neq y_i}$ – кількість ознак елементів X та Y, що не збігаються.

У літературі можна знайти приклади інших метрик та їх застосування. Подібність між об’єктами може оцінюватися за вибірковими коефіцієнтами кореляції. Причому сам коефіцієнт кореляції характеризує лише лінійний зв’язок між об’єктами, в іншому разі мірою близькості об’єктів може бути кореляційне відношення. Рангові коефіцієнти кореляції доцільно використовувати у випадку, якщо ознаки є упорядкованими за зростанням чи убуванням даних.

Принципи виділення кластерів. Найбільш поширеними методами кластерного аналізу є ієрархічні агломеративні методи [21]. Їх алгоритм передбачає послідовне об’єднання двох найбільш подібних кластерів в один; кінцевим результатом є формування кластера, який містить у собі всі об’єкти.

Спочатку кожен об’єкт визначається як окремий кластер. За матрицею відстаней (відмінностей) між об’єктами знаходять два найбільш близьких кластери, які об’єднують в один. Новостворений кластер позначається номером з найменшим значенням індексу, позначення інших кластерів не змінюються.

На другому кроці матриця відстаней з меншою вимірністю перераховується, і знову знаходяться два найбільш подібних кластери, які далі об’єднуються з подальшим перерахуванням матриці відмінностей.

Кінцевим результатом цього процесу стане утворення одного кластера, який включатиме всі об’єкти сукупності. Однак цей процес можна зупинити на певному кроці, якщо стане очевидним суттєва відмінність між новостворюваними кластерами, що проявиться високим приростом відстані між об’єднаними кластерами.

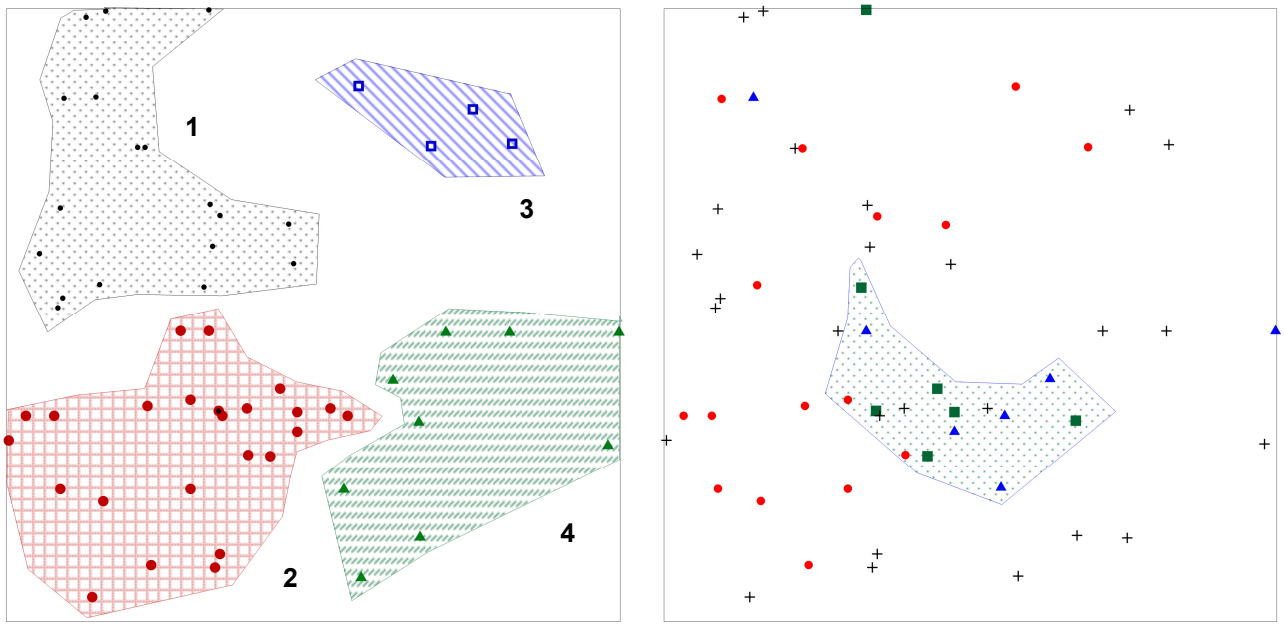
Способи обчислення відстані між кластерами є різними; найбільш поширені такі:

- *метод ближнього сусіда*, коли відстань між кластерами визначається як найменша з усіх відстаней між об’єктами, що входять до складу двох кластерів;
- *метод дальнього сусіда*, коли об'єкт приєднується до того кластера, найдальший елемент якого знаходиться ближче до нового об'єкта, ніж найдальші елементи інших кластерів;
- *метод середнього зв’язку*, коли відстань між кластерами визначається за середнім арифметичним відстаней між усіма парами об’єктів, що входять до складу цих кластерів.

Більш ускладненим є *метод Уорда*, згідно з яким на кожному кроці об’єднуються ті кластери, для яких приріст внутрішньокластерної дисперсії в результаті об’єднання буде найменшим. Метод дає можливість отримати кластери приблизно однакових розмірів з мінімальною внутрішньокласовою дисперсією.

Ці методи реалізовано в кількох програмних продуктах, зокрема, у додатковому модулі з інтелектуального аналізу даних в Excel, починаючи з 2007 р., у програмі Statistica та деяких інших. Це робить практичне застосування кластерного аналізу зручним і більш доступним для початківців.

Приклад 3.13. Розглянемо виділення кластерів рудопроявів на ділянці родовища (рис. 3.16) для того самого розташування точок відбору проб, що показано на рис. 3.8. Виділення кластерів на рис. 3.16, *а* виконано за евклідовою відстанню між точками на площині, тоді як для рис. 3.16, *б* – на основі вмісту корисного елемента в пробі.



а) *б*)

Рис. 3.16. Виділення кластерів на площині за критерієм геометричної близькості (*а*) та за вмістом рідкоземельного металу в пробі (*б*). Позначення: до 1 г/т (чорні маркери – 1), 1 – 5 г/т (червоні маркери – 2), 5 – 20 г/т (сині маркери – 3), понад 20 г/т (зелені маркери – 4)

Приклад 3.14. Визначимо геохімічні асоціації (кластери) важких металів, що осіли на поверхні ґрунту внаслідок випадінь з атмосфери навколо металургійного підприємства. Знання геохімічних асоціацій (кластерів) є важливим для оцінки впливу на довкілля, зокрема, для розуміння специфіки об’єкта, аналізу забруднення ґрунтів, підземних та поверхневих вод навколо промислових підприємств, що дозволяє правильно визначати напрямки спільної міграції та акумуляції токсичних речовин у навколишньому середовищі та обґрунтовувати захисні заходи. На рис. 3.17 подано фрагмент робочого аркуша MS Excel, на якому наведено вміст семи металів за валовим (загальним) вмістом у 22 пробах на поверхні ґрунту, відібраних у різних пунктах моніторингу навколо металургійного заводу.

Для проведення кластерного аналізу спочатку обчислимо кореляційну матрицю (рис. 3.18), яка показує існування тісного кореляційного зв’язку між умістом таких елементів, як Ni та Co (r = 0,7463), Ni та Cr (r = 0,7230), Cu та Pb (r = 0,7272), Cr та Co (r = 0,7059), Cu та Cr (r = 0,6172), Ni та Co (r = 0,7463).

	A	B	C	D	E	F	G	H
14		Fe	Co	Mn	Cu	Ni	Pb	Cr
15	1	27433	4.6	3088	89.1	41.3	102.6	72.9
16	2	26131	6.0	1861	62.5	63.9	81.9	82.1
17	3	44352	5.2	1244	52.7	65.4	57.5	83.2
18	4	45399	4.6	1320	14.5	18.0	16.9	40.9
19	5	51359	4.6	2646	47.4	71.1	56.8	70.9
20	6	70119	3.2	2615	49.8	79.0	52.3	68.5
21	7	18964	4.7	2353	31.5	31.3	63.1	67.3
22	8	20372	6.5	1327	31.7	43.8	28.2	81.5
23	9	1745	5.6	3178	54.7	63.1	50.6	79.3
24	10	15334	7.0	616	28.2	25.2	31.6	61.1
25	11	13705	5.1	1177	32.2	22.1	20.5	71.5
26	12	18797	7.2	1508	30.9	65.7	19.1	85.6
27	13	26751	14.6	1981	65.3	185.1	40.7	133.5
28	14	21448	3.5	1527	39.4	32.2	44.9	63.0
29	15	1632	6.2	1453	31.6	28.0	34.6	78.5
30	16	22736	5.2	8219	29.7	25.6	45.4	65.0
31	17	8362	7.0	1488	18.6	26.4	13.3	66.6
32	18	21579	4.6	1764	31.8	27.9	24.7	89.8
33	19	9527	3.6	416	8.7	14.8	8.7	41.7
34	20	15454	5.4	943	15.1	21.3	14.5	43.6
35	21	16026	4.0	1715	22.7	18.0	24.0	41.2
36	22	23081	5.3	1548	24.0	67.0	90.0	37.0

Рис. 3.17. Фрагмент аркуша MS Excel з даними спостережень за вмістом важких металів у ґрунті (мг/кг) навколо металургійного підприємства, що використані для кластерного аналізу

	J	K	L	M	N	O	P	Q	R
14			Fe	Co	Mn	Cu	Ni	Pb	Cr
15			*Column 1*	*Column 2*	*Column 3*	*Column 4*	*Column 5*	*Column 6*	*Column 7*
16	Fe	Column 1	1	-0.17227	0.122119	0.294445	0.320955	0.284028	0.046803
17	Co	Column 2	-0.17227	1	-0.05031	0.23984	0.746308	-0.08547	0.70589
18	Mn	Column 3	0.122119	-0.05031	1	0.24956	0.051356	0.308028	0.105867
19	Cu	Column 4	0.294445	0.23984	0.24956	1	0.563678	0.727196	0.617213
20	Ni	Column 5	0.320955	0.746308	0.051356	0.563678	1	0.305495	0.722982
21	Pb	Column 6	0.284028	-0.08547	0.308028	0.727196	0.305495	1	0.120639
22	Cr	Column 7	0.046803	0.70589	0.105867	0.617213	0.722982	0.120639	1

Рис. 3.18. Кореляційна матриця для вмісту семи важких металів у зразках ґрунту

Згідно з критерієм [22]

$$t > t_{n-2,\alpha}\,,\; t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\,, \tag{3.50}$$

де $t_{n-2,\alpha}$ – значення з розподілу Стьюдента, ці коефіцієнти кореляції є статистично значущими для рівня значущості $\alpha = 0{,}05$. Отже, можна стверджувати про існування геохімічних асоціацій між цими елементами.

На основі кореляційної матриці з використанням евклідової метрики (3.45) проведемо розрахунки матриці відстаней (рис. 3.19), що дозволить виконати кластерний аналіз. За методом найближчого сусіда визначимо так звану дендрограму геохімічних асоціацій елементів (рис. 3.20). Чим більша спорідненість між елементами, тим менша відстань між ними у вибраній метриці.

	J	K	L	M	N	O	P	Q	R
37									
38			Fe	Co	Mn	Cu	Ni	Pb	Cr
39		Fe	0	133852	125893	133715	133659	133691	133602
40		Co	133852	0	11778	170	253	208	312
41		Mn	125893	11778	0	11648	11620	11621	11536
42		Cu	133715	170	11648	0	148	86	173
43		Ni	133659	253	11620	148	0	176	159
44		Pb	133691	208	11621	86	176	0	194
45		Cr	133602	312	11536	173	159	194	0

Рис. 3.19. Матриця відстаней при кластерному аналізі геохімічних даних (до прикладу 3.14)

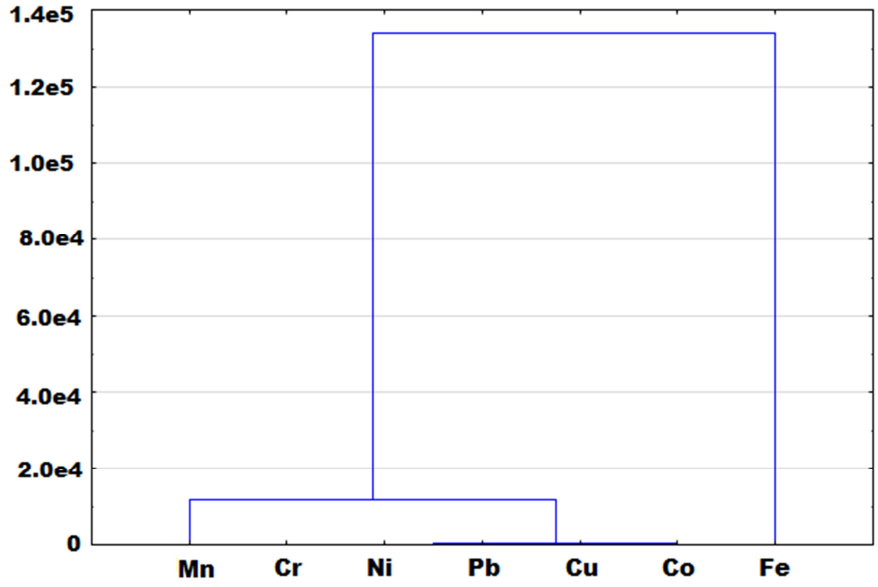


Рис. 3.20. Дендрограма геохімічних асоціацій важких металів (до прикладу 3.14)

Порівнюючи міжелементні відстані, можна стверджувати, що на розглянутому об’єкті виділяється кластер кольорових та важких металів, до якого входить 5 елементів: Ni, Co, Cr, Cu та Pb. Елемент Mn більш суттєво відрізняється від даного кластера, а найбільш відмінним від усіх названих елементів є Fe, відстані від якого до інших елементів найбільші. Така геохімічна асоціація природна для умов міграції металів у викидах підприємства чорної металургії, у яких переважають Fe та Mn, а інші метали є переважно домішками. Наступним кроком кластерного аналізу має стати географічне виділення кластерів відповідно до місць відбору проб, для чого використовуються більш потужні методи просторового аналізу даних [23].

Контрольні питання до підрозділу 3.6

1. Як формується матриця спостережень та кореляційна матриця?
2. Які висновки можна зробити на основі аналізу кореляційної матриці?
3. Як визначається рівняння множинної лінійної регресії?
4. Наведіть приклади застосування апарату множинної регресії в задачах геології та екології.
5. Для чого використовується шкала Чеддока?
6. Охарактеризуйте поняття кластера. Наведіть приклади.
7. Порівняйте різні метрики для визначення відстані між елементами вибірки.
8. За якими критеріями можна відділити кластери один від одного?
9. Що характеризує дендрограма зв’язків між досліджуваними ознаками (елементами)?

Література до розділу 3

1. Геология и математика / Ю.А. Воронин, Б.К. Алабин, С.В. Гольдин и др. – Новосибирск: Наука, 1967. – 254 с.

2. Каждан А.Б. Математическое моделирование в геологии и разведке полезных ископаемых: учеб. пособие / А.Б. Каждан, О.И. Гуськов, А.А. Шиманский. – Москва: Недра, 1979. – 168 с.

3. Rosenfeld M.A. An experimental test of visual comparison technique in estimating two-dimensional sphericity and roundness of quartz grains / M.A. Rosenfeld, J.C. Griffiths // Am. J. Sci. – Vol. 251. – 1953. – P. 553–585.

4. Мартьянова А.Е. Математические методы моделирования в геологии: учеб. пособие для студентов направления 650100 «Прикладная геология»: в 2 ч. / А.Е. Мартьянова. – Астрахань: АГТУ, 2008. – 2 ч.

5. Крамбейн У. Статистические модели в геологии / У. Крамбейн, Ф. Грейбилл. – Москва: Мир, 1969. – 398 с.

6. Дэвис Дж. С. Статистический анализ данных в геологии: в 2-х кн. / Дж. С. Дэвис. – Москва: Недра, 1990. – Кн. 2. – 427 с.

7. Мардиа К. Статистический анализ угловых наблюдений: пер. с англ. / К. Мардиа. – Москва: Физматлит, 1978. – 240 с.

8. Марчук Г.И. Математическое моделирование в проблеме окружающей среды / Г.И. Марчук. – Москва: Наука, 1982. – 320 с.

9. Рудаков Д.В. Математичні моделі в охороні навколишнього середовища: навч. посіб. / Д.В. Рудаков. – Дніпропетровськ: Вид-во Дніпропетр. держ. ун-ту, 2004. – 160 с.

10. Рудаков Д.В. Моделювання переносу домішок в атмосфері над неоднорідною поверхнею / Д.В. Рудаков // Вісник Київського ун-ту. Сер. Фіз.-мат. науки. – 2003. – № 5. – С. 87–93.

11. Численное моделирование распространения загрязнения в окружающей среде / М.З. Згуровский, В.В. Скопецкий, В.К. Хрущ, Н.Н. Беляев. – Киев: Наукова думка, 1997. – 368 с.

12. Методика расчета концентраций в атмосферном воздухе вредных веществ, содержащихся в выбросах предприятий ОНД-86. – Ленинград: Гидрометеоидат, 1987. – 93 с.

13. Холлендер М. Непараметрические методы статистики / М. Холлендер, Д. Вульф. – Москва: Финансы и статистика, 1983. – 518 с.

14. Гаек Я. Теория ранговых критериев: пер. с англ. / Я. Гаек, З. Шидак. – Москва: Наука, 1971. – 376 с.

15. Большев Л.Н. Таблицы математической статистики / Л.Н. Большев, Н.В. Смирнов. – Москва: Наука, 1983. – 416 с.

16. Рудаков Д.В. Математичні методи в охороні підземних вод / Д.В. Рудаков. – Дніпропетровськ: НГУ, 2012. – 158 с.

17. Боровко Н.Н. Статистический анализ пространственных геологических закономерностей / Н.Н. Боровко. – Ленинград: Недра, 1971. – 174 с.

18. Геостатистический анализ данных в экологии и природопользовании (с применением пакета R): учеб. пособие / А.А. Савельев, С.С. Мухарамова, А.Г. Пилюгин, Н.А. Чижикова. – Казань: Казанский гос. ун-т, 2012. – 120 с.

19. Дубров А.М. Многомерные статистические методы: учебник / А.М. Дубров, В.С. Мхитарян, Л.И. Трошин. – Москва: Финансы и статистика, 2003. – 352 с.

20. Гирко В.Л. Многомерный статистический анализ / В.Л. Гирко. – Киев: Вища школа, 1988. – 320 с.

21. Яровий А.Т. Багатовимірний статистичний аналіз: навч.-метод. посіб. для студентів мат. та екон. фахів / А.Т. Яровий, Є.М. Страхов. – Одеса: Астропринт, 2015. – 132 с.

22. Львовский Е.Н. Статистические методы построения эмпирических функций / Е.Н. Львовский. – Москва: Высш. школа, 1988. – 239 с.

23. Демьянов В.В. Геостатистика: теория и практика / В.В. Демьянов, Е.А. Савельева ; под ред. Р.В. Арутюняна ; Ин-т проблем безопасного развития атомной энергетики РАН. – Москва: Наука, 2010. – 327 с.

Розділ 4
ПОШУКИ ЕКСТРЕМУМІВ. ЗАДАЧІ ОПТИМІЗАЦІЇ

4.1. Постановка та приклади задач оптимізації природничих систем

Оптимізація являє собою процес вибору найкращого варіанта з усіх можливих. Методи оптимізації застосовують в економіці, інженерії, логістиці, де вони дозволяють вибрати найліпший розподіл ресурсів, варіант конструкції, маршрут транспортування вантажів тощо [1,2,3].

Завданням оптимізації зазвичай є визначення оптимальних значень кількох параметрів. В інженерних задачах їх називають проектними параметрами, а в економічних– параметрами плану. Проектними параметрами можуть бути, зокрема, значення розмірів об'єкта, температури, концентрації речовин, типу матеріалу тощо. Кількість проектних параметрів n визначає розмірність задачі оптимізації.

Оптимальний розв'язок знаходиться за допомогою цільової функції J, яка залежить від проектних параметрів $x_1, x_2\ldots, x_n$ [4], наприклад

$$J\left(x_1, x_2, ..., x_n\right) \to \min . \tag{4.1}$$

Вираз (4.1) означає, що розшукується такий набір значень $x_1, x_2\ldots, x_n$, при якому цільова функція J досягає мінімального значення. Задачі оптимізації можуть бути сформульовані як на пошук мінімуму, так і пошук максимуму; при цьому вони можуть бути виражені одна через одну. Наприклад, задача

$$-J\left(x_1, x_2, ..., x_n\right) \to \max \tag{4.2}$$

формально еквівалентна задачі (4.1).

Цільову функцію також іноді називають критерієм якості або критерієм оптимальності в математичній моделі. У процесі виконання завдання оптимізації повинні бути знайдені такі значення проектних параметрів, при яких цільова функція набуває мінімальних чи максимальних значень.

Зважаючи на різноманіття практичних задач, цільова функція може бути подана у вигляді формули або таблиці, яка містить окремі значення; при цьому вона обов'язково має мути бути однозначною функцією проектних параметрів.

Цільових функцій може бути декілька. Наприклад, при оптимізації водовідбору поблизу моря одночасно потрібно забезпечити максимальну витрату води та її мінімальну мінералізацію. Деякі цільові функції можуть виявитися несумісними, що потребує визначення пріоритету цільових функцій.

Завдання оптимізації можуть бути безумовними та умовними. *Безумовна оптимізація* полягає в знаходженні максимуму або мінімуму дійсної функції

(4.1) від n дійсних змінних і визначенні відповідних значень аргументів на деякій множині n-мірного простору.

Умовна оптимізація (або оптимізація з обмеженнями) виконується за певних умов (обмежень) на множині, де задані параметри x_1, x_2…., x_n. Ці обмеження задаються сукупністю деяких функцій, що задовольняють рівняння

$$f_i\left(x_1, x_2, ..., x_n\right) = 0,\ i = 1, ..., k \tag{4.3}$$

або нерівності

$$a_i \le g_i\left(x_1, x_2, ..., x_n\right) \le b_i,\ i = 1, ..., m\,. \tag{4.4}$$

Обмеження (4.3) та (4.4), що мають враховуватися при пошуку оптимального розв'язку, можуть відображати певні фізичні закони, геометричні співвідношення, наявність ресурсів, фінансові вимоги тощо.

Замість функцій g_i можуть бути використані аргументи, наприклад

$$a_i \le x_i \le b_i\,, \tag{4.5}$$

тоді нерівності (4.5) безпосередньо накладають обмеження на область зміни параметрів x_i. Наприклад, часто використовується умова того, що параметри x_i мають бути невід'ємними.

Число обмежень-рівностей може бути довільним. Іноді вдається з цих співвідношень виразити одні параметри через інші, що дозволяє зменшити кількість параметрів і розмірність задачі, і що тим самим полегшує її розв'язок. Оптимальний розв'язок при наявності обмежень може відповідати або локальному екстремуму (максимуму або мінімуму) всередині області, або значенням цільової функції на межі області. Якщо обмеження відсутні, то визначиться оптимальний розв'язок для всієї області зміни параметрів x_1, x_2…., x_n, тобто глобальний екстремум.

У разі лише одного параметра оптимізації ($n = 1$) цільова функція стає функцією однієї змінної, а її графік – кривою на площині. Для $n = 2$ цільова функція є функцією двох змінних, а її графік – поверхнею. Для цільових функцій від більшої кількості параметрів оптимізації графічна інтерпретація можлива у вигляді поверхонь, що показують зміни функції, коли лише змінюються два параметри, а інші є фіксованими.

Прикладами цільової функції, що зустрічаються в інженерних і економічних розрахунках, є міцність або маса конструкції, потужність установки, обсяг випуску продукції, вартість перевезень вантажів, прибуток і т. п.

Приклади задач оптимізації у сфері використання природних ресурсів дуже численні.

Приклад 4.1. Оцінимо можливості оптимізації роботи водозабору, яка в загальному випадку має враховувати якість підземних вод як одне з обмежень. Приклад оптимізації роботи трьох свердловин симплекс-методом детально розглянуто в підрозділі 4.3. Завдання оптимізації можна сформулювати як збільшення відбору підземних вод

$$\sum_i Q_i \to \max , \tag{4.6}$$

де Q_i – дебіти водозаборів, за умови дотримання обмежень на якість води у разі часткового забруднення водоносного горизонту

$$C \le C_{ГДК} . \tag{4.7}$$

Тут C – концентрація забруднювачів у воді, $C_{ГДК}$ – їх гранично допустима концентрація згідно із санітарними нормами.

Приклад 4.2. Зниження концентрації забруднювачів у підземних водах у разі проведення заходів їх нейтралізації. Така ситуація виникає після припинення видобутку корисних копалин, зокрема, методом підземного вилуговування або гідророзриву пласта, коли концентрація токсичних речовин на великій площі водоносних горизонтів значно перевищує $C_{ГДК}$. Після відновлення природного градієнта течії підземних вод токсичні речовини поширюються в напрямку водозабору і погіршують якість води. Для попередження цього (рис. 4.1) можна встановити одну чи кілька свердловин, щоб нейтралізувати токсичні речовини спеціальним розчином.

Задача оптимізації полягає в установленні таких параметрів нейтралізації (координат і дебіту свердловини, концентрації нейтралізатора в розчині), за яких концентрація забруднювача у воді, що відбирається водозабором C_p, буде мінімальною

$$C_p \to \min . \tag{4.8}$$

Обмеженнями у даному випадку накладаються на параметри нейтралізації, зокрема, дебіти свердловин та концентрацію речовини-нейтралізатора у розчині.

Для даного прикладу також можливе формулювання оптимізаційної задачі зі співвідношеннями (4.6) – (4.7).

Приклад 4.3. Оптимізація кута нахилу борту кар’єру, яка необхідна при видобутку корисних копалин відкритим способом. З одного боку, малий кут нахилу призводить до залишення значної частини корисних копалин у кар’єрі, з іншого – високий кут нахилу стає небезпечним через зниження стійкості борту кар’єру та зростання ризику обвалення.

Задачу оптимізації у першому наближенні можна сформулювати описаним далі способом. Необхідно максимізувати об’єм породної маси ΔV, що виймається з борту,

$$\Delta V \to \max \ , \tag{4.9}$$

де $\Delta V = l\Delta S$; l – довжина борту; ΔS – зміна площі профілю борту при зміні кута нахилу від α_1 до α_2,

$$\Delta S = \frac{1}{2}\left(hl_2 - hl_1\right) = \frac{h^2}{2}\left(\sin\alpha_2 - \sin\alpha_1\right); \tag{4.10}$$

h – висота борту (рис. 4.2).

Задачу оптимізації необхідно розв’язувати за умови обмеження на коефіцієнт стійкості схилу

$$K_c \le K_{c,\min} \ , \tag{4.11}$$

де $K_{c,\min}$ – мінімально допустимий коефіцієнт стійкості схилу (борту кар’єру), що згідно з будівельними нормативами у різних країнах зараз становить 1,2 – 1,35. Коефіцієнт стійкості схилу може бути розрахований за відповідними методиками.

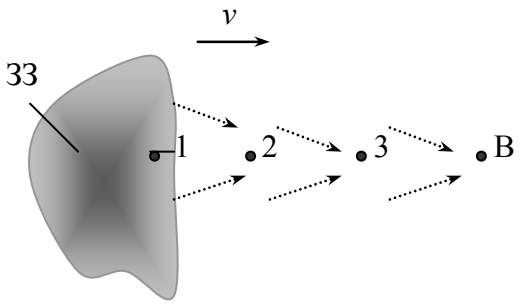


Рис. 4.1. Положення зони забруднення у водоносному горизонті (ЗЗ), водозабору В та можливих місць розташування свердловини, що відкачують воду (1, 2, 3)

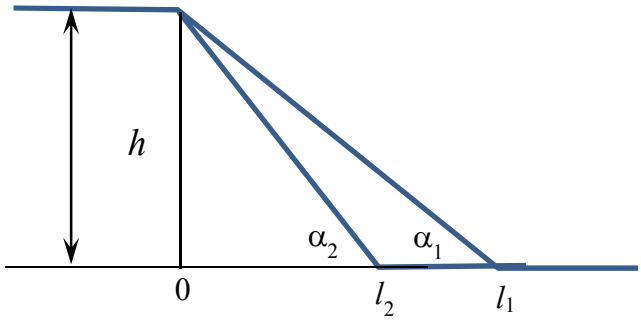


Рис. 4.2. Контур борту кар’єру при різних кутах нахилу

У практиці видобувної галузі можуть зустрічатися й інші приклади задач оптимізації, зокрема, оптимізація розміщення та дебітів свердловин на нафту й газ за критерієм максимізації видобутку [5], оптимізація форми підземної виробки за критерієм її стійкості, оптимізація розміщення вибухових речовин при проведенні буропідривних робіт з метою збільшення об’єму зруйнованої породи [6], оптимізація транспортування видобутих корисних копалин тощо.

Контрольні питання до підрозділу 4.1

1. Як визначається цільова функція?
2. Охарактеризуйте різницю між умовною та безумовною оптимізацією.
3. Чи пов’язані між собою кількість змінних задачі оптимізації та кількість її обмежень?
4. Наведіть приклади задач оптимізації з визначенням цільової функції та обмежень у галузі: а) геології та гідрогеології, б) охорони навколишнього середовища, в) видобутку корисних копалин.

4.2. Пошук екстремуму функцій, заданих аналітично

Знаходження мінімуму (максимуму) функції, заданої аналітично в деякій області зміни її аргументів, здійснюється методами математичного аналізу [7].

Неперервна *функція однієї змінної* $y = f(x)$ досягає у точці x_0 *максимуму*, якщо для всіх $x \neq x_0$ з деякого околу цієї точки $f(x) < f(x_0)$ (рис. 4.3, *а*). У випадку оберненої нерівності $f(x) > f(x_0)$ функція y має у точці x_0 *мінімум* (рис. 4.3, *б*).

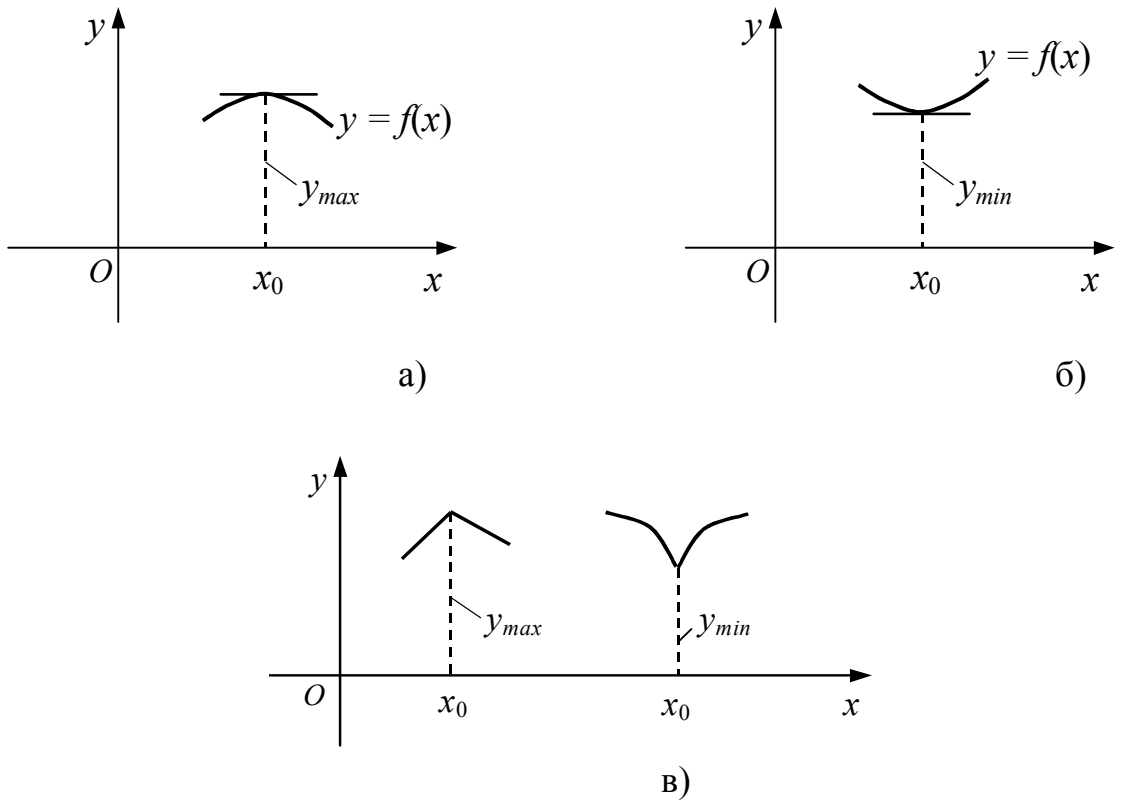


Рис. 4.3. Локальні екстремуми функції однієї змінної

Максимум (*max*) та мінімум (*min*) – це значення функції із загальною назвою – *екстремуми*. Пишуть $y_{max} = y(x_0)$ або $y_{min} = y(x_0)$, де x_0 – точка екстремуму. Окіл точки x_0, де виконується та чи інша із зазначених вище нерівностей, може бути дуже малим, тому екстремуми характеризують функцію клокально – вони визначають найбільше або найменше значення функції, але тільки в околі даної точки. Щоб підкреслити їх особливість, такі значення функції називають *локальними* екстремумами.

Якщо неперервна функція $y = f(x)$ має у точці x_0 екстремум, то в цій точці похідна y' перетворюється на нуль або не існує. Цю умову називають *необхідною* для існування екстремуму в точці x_0. Геометрично вона означає, що дотична до графіка функції у точках, де досягається екстремум, паралельна до осі *Ox* (рис. 4.3). Обернене твердження несправедливе: якщо $f'(x_0) = 0$ і дотична паралельна до осі *Ox*, то екстремуму в точці x_0 може і не бути. Точки, у яких y' перетворюється на нуль, називають *стаціонарними.* Екстремум може досягатися також у точках, де похідна y' не існує (рис. 4.3, *в*). Разом зі стаціонарними такі точки називають *критичними.* Тільки у таких точках ймовірні екстремуми функції.

Достатні умови існування екстремуму в точці x_0 формулюються так: якщо при переході через цю точку зліва направо похідна y' змінює знак з «+» на «–», то в точці x_0 функція $y = f(x)$ досягає максимуму. Зміна знаку y' з «–» на «+» означає, що в точці x_0 функція $y = f(x)$ досягає мінімуму.

Пошук екстремуму також можливий за допомогою *другої похідної*. Нехай $f'(x_0) = 0$, тобто x_0 – стаціонарна точка функції $y = f(x)$. Припустимо, що у цій точці та у деякому її околі існує неперервна друга похідна $f''(x)$. Тоді, якщо $f''(x_0) < 0$, то функція $y = f(x)$ у точці x_0 має максимум, якщо ж $f''(x_0) > 0$, то у точці x_0 досягається мінімум функції $f(x)$.

Якщо функція $y = f(x)$ неперервна на відрізку $[a, b]$, то вона набуває на цьому відрізку найбільшого (M) та найменшого (m) значення, які називають *глобальними* екстремумами (рис. 4.4). Ці значення функція набуває або в точках локальних екстремумів усередині відрізка або на кінцях останнього. Достатньо в зазначених точках обчислити значення функції та порівняти їх між собою. У результаті визначаються найбільше та найменше значення функції на відрізку і ті точки, де ці значення досягаються.

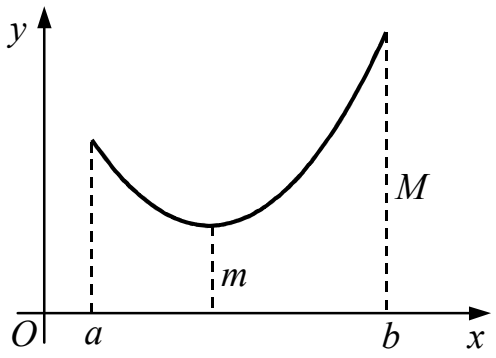


Рис. 4.4. Найбільше та найменше значення функції на відрізку

Приклад 4.4. Дослідження напруженого стану цілика порід шириною 4 м між двома підземними виробками показали, що зміна вертикальних напружень уздовж середньої лінії цілика на відрізку [0, 4] може бути описана функцією $\sigma(x) = x^4 - 8x^3 + 22x^2 - 24x + 12$. Знайти локальні та глобальні екстремуми даної функції.

Розв'язок. Дорівнюючи нулю першу похідну, отримаємо рівняння $\sigma' = 4x^3 - 24x^2 + 44x - 24 = 0$, корні якого $x_1 = 1,\ x_2 = 2,\ x_3 = 3$ є стаціонарними точками. Визначимо знак другої похідної $\sigma'' = 12x^2 - 48x + 44$ в кожній з них і отримаємо, що в точках $x_1 = 1$ і $x_3 = 3$ функція $\sigma = \sigma(x)$ має локальні мінімуми $\sigma_{\min} = \sigma(1) = \sigma(3) = 3$, а в точці $x_2 = 2$ – локальний максимум $\sigma_{\max} = \sigma(2) = 4$.

На кінцях інтервалу [0, 4] функція $\sigma = \sigma(x)$ набуває рівних значень $\sigma(0) = \sigma(4) = 12$. Таким чином функція досягає глобальних максимумів на кінцях інтервалу і глобальних мінімумів всередині інтервалу (в точках $x_1 = 1$ і $x_3 = 3$).

Приклад 4.5. Треба побудувати резервуар з квадратним дном і об'ємом 32 м^3 таким чином, щоби на оздоблення його стін і дна пішло якнайменше матеріалу.

Розв'язок. Нехай сторона квадрата, що є дном резервуара, дорівнює *a*, а висота борту – *h* (рис. 4.5). Тоді площа поверхні, яку потрібно обкласти матеріалом, складає $S=a^2+4ah$. Ця величина має бути мінімальною. Таким чином, маємо задачу оптимізації із заданим обмеженням: об'єм резервуара складатиме 32 м^3.

З формули для об'єму $V = a^2h = 32\ m^3$ визначимо *h* та підставимо у вираз для площі поверхні. Таким чином отримаємо цільову функцію однієї змінної $S = a^2 + \frac{128}{a}$, яку треба мінімізувати. Знаходячи похідну $\frac{dS}{da} = 2a - \frac{128}{a^2}$ та

прирівнюючи її до нуля, отримаємо стаціонарну точку $a = 4$. Перевіряємо достатню умову існування екстремуму в даній точці за допомогою другої похідної $\frac{d^2S}{da^2} = 2 + \frac{256}{a^3}$. Для точки $a = 4$ запишемо $\frac{d^2S}{da^2} = 2 + \frac{256}{4^3} > 0$, тобто функція має мінімум. Таким чином, оптимальні розміри резервуара при даному обмеженні такі: довжина сторони дна $a = 4$, висота борту резервуара $h = 32/4^2 = 2$.

Рис. 4.5. Ілюстрація до визначення оптимальних розмірів резервуара

Екстремуми функції двох змінних. Неперервна функція $f(x, y)$ досягає у точці $M_0(x_0, y_0)$ максимуму, якщо для всіх точок $M(x, y)$ з деякого околу точки M_0 виконується нерівність $f(M) < f(M_0)$. При зворотній нерівності функція $f(x, y)$ у точці M_0 досягає мінімуму.

Аналогічно формулюються поняття локального екстремуму, тобто максимуму або мінімуму у випадку функції n незалежних змінних.

Необхідні умови екстремуму. Нехай диференційована функція $z = f(x,y)$ у точці $M_0(x_0, y_0)$ має екстремум. Тоді в точці $M_0(x_0, y_0)$ частинні похідні функції дорівнюють нулю

$$\left.\frac{\partial f}{\partial x}\right|_{M_0} = 0, \quad \left.\frac{\partial f}{\partial y}\right|_{M_0} = 0. \tag{4.12}$$

Точку M_0, у якій виконуються умови (4.12), називають *стаціонарною.* Подібно до цього встановлюються необхідні умови екстремуму для диференційованої функції n незалежних змінних $z = f(x_1, x_2, \ldots, x_n)$. Ці умови полягають у виконанні таких n рівностей:

$$\left.\frac{\partial f}{\partial x_1}\right|_{M_0} = 0, \left.\frac{\partial f}{\partial x_2}\right|_{M_0} = 0, \ldots, \left.\frac{\partial f}{\partial x_n}\right|_{M_0} = 0. \tag{4.13}$$

Окрім стаціонарних, точками екстремуму можуть бути й такі точки, де функція f недиференційована. Наприклад, функція $z = \sqrt{x^2 + y^2}$ (геометрично це конус з вершиною на початку координат) у точці $O(0; 0)$ має мінімум, але її частинні похідні у точці O не існують. Як і в випадку функції однієї змінної, такі точки разом зі стаціонарними називають *критичними*. З їх пошуку починається дослідження функції на екстремум.

Достатні умови екстремуму функції двох змінних. Наявність критичної точки M_0 ще не означає, що в ній функція $z = f(M)$ має екстремум. Отже, умова (4.12) є необхідною, але не достатньою. Нехай M_0 – стаціонарна точка і функція $f(x, y)$ має в цій точці неперервні частинні похідні другого порядку.

Введемо позначення

$$A = \left.\frac{\partial^2 f}{\partial x^2}\right|_{M_0},\ B = \left.\frac{\partial^2 f}{\partial x \partial y}\right|_{M_0},\ C = \left.\frac{\partial^2 f}{\partial y^2}\right|_{M_0}. \qquad (4.14)$$

Розглянемо матрицю

$$H = \begin{pmatrix} A & B \\ B & C \end{pmatrix}$$

та визначник цієї матриці $\delta = AC - B^2$.

Тоді:

а) у випадку $\delta > 0$ функція $z = f(x, y)$ у точці M_0 має екстремум, при чому *максимум* при $A < 0$, *мінімум* при $A > 0$;

б) при $\delta < 0$ екстремуму в точці M_0 немає;

в) у випадку $\delta = 0$ екстремум у точці M_0 може бути, а може й не бути, потрібно додаткове дослідження.

Аналогічно вирішується питання про достатні умови екстремуму для функції, що залежить від більшої кількості змінних.

Приклад 4.6. За результатами випробування зразків гірської породи з кернів розвідувальних свердловин побудована залежність межі міцності породи на стиск (позначимо її z) від координат x, y свердловини на шахтному полі: $z = x^4 - 4x^3 + 4xy - y^2 + 1$. Знайти локальні екстремуми даної функції для визначення ділянок з найбільш міцними породами, де для проведення виробок необхідні буропідривні роботи.

Розв`язування. Відшукаємо стаціонарні точки функції z, для чого знайдемо перші похідні від z та прирівняємо їх до нуля:

$$\frac{\partial z}{\partial x} = 4x^3 - 12x^2 + 4y, \qquad \frac{\partial z}{\partial y} = 4x - 2y,$$

$$\left.\begin{array}{c} 4x^3 - 12x^2 + 4y = 0 \\ 4x - 2y = 0 \end{array}\right\} \Rightarrow \quad y = 2x,$$

$$4x^3 - 12x^2 + 8x = 0,\ 4x(x^2 - 3x + 2) = 0,$$

$$x_1 = 0,\ x_2 = 1,\ x_3 = 2,\ y_1 = 0,\ y_2 = 2, y_3 = 4.$$

Функція z має три стаціонарні точки $M_1(0, 0)$, $M_2(1, 2)$, $M_3(2, 4)$; інших критичних точок у неї немає. Перевіримо виконання достатніх умов (табл. 4.1). Екстремум є тільки у т. M_2. Оскільки тут $A < 0$, то це і є максимум. Таким чином, максимальне значення міцності породи отримано з керна свердловини з координатами $M_2(1, 2)$. Зауважимо, що координати є нормованими, тобто зведеними до певного масштабу карти шахтного поля.

Таблиця 4.1

Результати обчислень похідних у критичних точках (до прикладу 4.6)

Стаціонарна точка	$\frac{\partial^2 z}{\partial x^2} = 12x^2 - 24x$	$\frac{\partial^2 z}{\partial x \partial y} = 4$	$\frac{\partial^2 z}{\partial y^2} = -2$	$\delta = AC - B^2$	Екстремум
$M_1(0; 0)$	$A = \left.\frac{\partial^2 z}{\partial x^2}\right\|_{M_1} = 0$	$B = \left.\frac{\partial^2 z}{\partial x \partial y}\right\|_{M_1} = 4$	$C = \left.\frac{\partial^2 z}{\partial y^2}\right\|_{M_1} = -2$	$\delta < 0$	Немає
$M_2(1; 2)$	$A = \left.\frac{\partial^2 z}{\partial x^2}\right\|_{M_2} = -12$	$B = \left.\frac{\partial^2 z}{\partial x \partial y}\right\|_{M_2} = 4$	$C = \left.\frac{\partial^2 z}{\partial y^2}\right\|_{M_2} = -2$	$\delta > 0$	$z_{max} = 2$
$M_3(2; 4)$	$A = \left.\frac{\partial^2 z}{\partial x^2}\right\|_{M_3} = 0$	$B = \left.\frac{\partial^2 z}{\partial x \partial y}\right\|_{M_3} = 4$	$C = \left.\frac{\partial^2 z}{\partial y^2}\right\|_{M_3} = -2$	$\delta < 0$	Немає

Зауважимо, що знайдений екстремум функції є одночасно і локальним, і глобальним. До того ж, оскільки ніяких додаткових умов не було накладено на задану функцію, знайдений максимум є *безумовним* екстремумом.

Між тим існують задачі, у яких потрібно знайти екстремум функції за умови, що її аргументи пов`язані між собою одним або декількома рівностями. Таких зв`язків повинно бути менше, ніж незалежних змінних, інакше задача пошуку екстремуму втрачає сенс. Припустимо, розшукується екстремум функції $z = f(x, y)$ за умови

$$F(x, y) = 0. \tag{4.15}$$

Це означає, що порівнювати треба тільки ті значення функції, які відповідають точкам, що розташовані на кривій (4.15). На рис. 4.6 зображені

лінії рівня функції $f(x, y)$: $f = h_1,\ f = h_2,\ f = h_3$. При $h_1 > h_2 > h_3$ перехід від більшого рівня до меншого еквівалентний наближенню до єдиного безумовного мінімуму, який досягається у т. *K*. Разом з тим умовних екстремумів уздовж лінії (4.15) буде три – максимум у т. *B* та два мінімуми у точках *A* та *C*. Для наочності: якщо *K* – найбільш глибока точка яру, то *B* – найвища, тоді як *A* та *C* – найнижчі точки стежки на його схилах (лінія (4.15) – це проекція стежки на площину *xOy*).

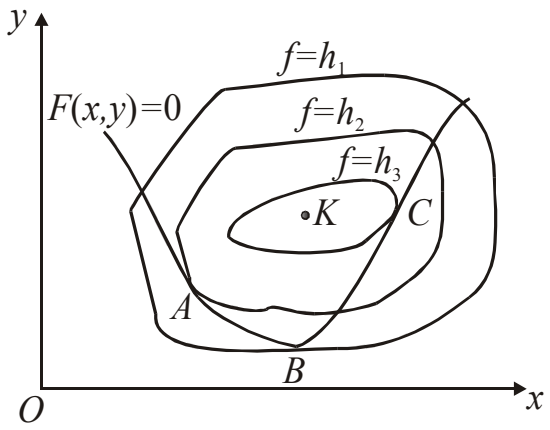


Рис. 4.6. Графічне зображення умовних екстремумів функції двох змінних

Для функції багатьох змінних $u = f(x_1, x_2, \ldots, x_n)$ умовний екстремум за наявності зв`язків

$$F_j(x_1, x_2, ..., x_n) = 0, \qquad j = 1, 2, \ldots, m, \quad m < n, \qquad (4.16)$$

знаходиться тими самими методами, що й у випадку двох змінних. За допомогою рівностей (4.16) m аргументів функції u, наприклад, $x_1, x_2, \ldots, x_m$, виключаються, після чого для отриманої функції з меншою кількістю аргументів $u = \varphi(x_{m+1}, \ldots, x_n)$ розв`язується задача на безумовний екстремум.

У наведених прикладах був застосований класичний метод дослідження функцій на екстремум за допомогою похідних. При вирішенні практичних завдань оптимізації класичний метод має обмежене застосування. Це пояснюється тим, що, по-перше, у багатьох випадках значення цільової функції знаходяться з вимірів або експериментів, а обчислення похідних є важким або неможливим, і, по-друге, навіть коли похідна задана аналітично або піддається виміру, вирішення рівнянь (4.12 – 4.13) часто викликає труднощі. Тому на практиці застосовують наближені методи пошуку екстремумів функції [8,9].

Прямі методи оптимізації функції однієї змінної. Наближені методи дозволяють знайти розв’язок цієї задачі з необхідною точністю шляхом визначення кінцевого числа значень функції $f(x)$ і її похідних у деяких точках

відрізка [*a; b*]. Методи, які використовують тільки значення функції і які не потребують обчислення її похідних, називаються *прямими методами мінімізації*. При використанні цих методів цільова функція може бути не задана в аналітичному вигляді. Алгоритми прямих методів засновані на можливості розрахунку $f(x)$ у заданих точках.

Розглянемо найбільш поширені на практиці прямі методи пошуку точки мінімуму [10,11]. Найголовнішою вимогою для функції $f(x)$, що дозволяє використовувати ці методи, є її *унімодальність*.

Функція $f(x)$ називається унімодальною на відрізку $[a, b]$, якщо вона неперервна на відрізку $[a, b]$ та існують числа α і β, $a \le \alpha \le \beta \le b$, такі, що

- якщо $a < \alpha$, то на відрізку $[a, \alpha]$ функція $f(x)$ монотонно спадає;
- якщо $\beta < b$, то на відрізку $[\beta, b]$ функція $f(x)$ монотонно зростає;
- при $x \in [\alpha, \beta]$, $f(x) = f^* = f_{min}$ на відрізку $[a, b]$.

З визначення унімодальної функції випливають такі основні властивості:

- будь-яка з точок *локального* мінімуму унімодальної функції є і точкою її *глобального* мінімуму на відрізку $[a, b]$;
- функція, унімодальна на відрізку $[a, b]$, є унімодальною і на будь-якому меншому відрізку $[c, d] \in [a, b]$;
- нехай $a \le x_1 \le x_2 \le b$, тоді, якщо $f(x_1) \le f(x_2)$, $x^* \in [a, x_2]$; якщо $f(x_1) > f(x_2)$, $x^* \in [x_1, b]$, де x^* – одна з можливих точок мінімуму функції $f(x)$.

Далі в розглянутих нижче методах будемо вважати функцію $f(x)$ унімодальною на відрізку $[a; b]$.

Метод рівномірного пошуку (сканування або перебору) є найпростішим з прямих методів. Розіб'ємо відрізок $[a, b]$ на n рівних частин точками $x_i = a + i\,(b - a)$, $i = 0, \ldots, n$. Обчислимо значення $f(x)$ в точках x_i, шляхом порівняння знайдемо точку x_m $(0 \le m \le n)$, у якій функція $f(x)$ набуває найменшого значення. Тоді точкою мінімуму є $x^* = x_m$, а мінімумом функції є $f^* = f(x_m)$.

Похибка є визначення точки x^* не перевершує величини $(b - a)/n$. Тоді при заданому є необхідна кількість точок ділення відрізка $[a; b]$ визначається так:

$$n \ge (b - a) / \varepsilon.$$

Приклад 4.7. Розглянемо функцію $f(x) = x + 2/x$ на відрізку [0,5, 3,5]. Знайдемо мінімум функції з похибкою $\varepsilon \le 0{,}5$. Число точок ділення відрізка визначимо так: $n > (b - a)/\varepsilon = 6$. Результати обчислень значень функції в точках ділення відрізка (табл. 4.2, рис. 4.7) показують, що точкою мінімуму є $x_m = x_2 = 1{,}5$; значення мінімуму: $f^* = f(1{,}5) = 2{,}83$.

Таблиця 4.2

Значення функції $f(x) = x + 2/x$ у точках ділення відрізка [0,5, 3,5]

Номер індексу i	0	1	2	3	4	5	6
Значення аргументу x_i	0,5	1	**1,5**	2	2,5	3	3,5
Значення функції $f(x_i)$	4,5	3,0	**2,83**	3,0	3,3	3,67	4,07

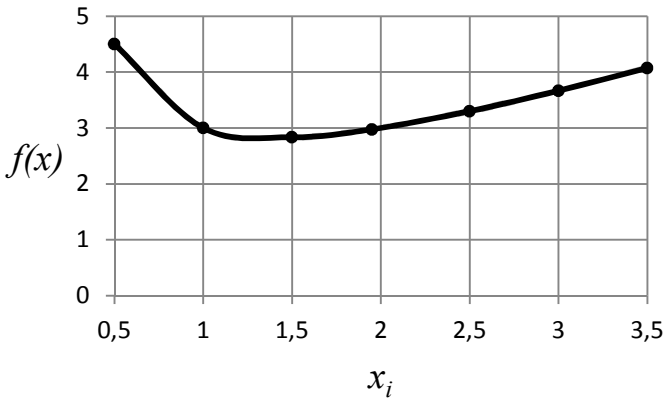


Рис. 4.7. Графік функції $f(x) = x + 2/x$ на відрізку [0,5, 3,5]

Відзначимо, що в цьому методі всі $n + 1$ точки вибираються заздалегідь і для більш ефективного пошуку мінімуму краще використовувати методи послідовного пошуку, у яких для обчислення чергової точки x_i використовується інформація, що отримана на більш ранній стадії розрахунків.

До таких методів належить *метод половинного ділення або дихотомії.* У цьому методі відрізок, де шукають точку мінімуму, звужується на кожному кроці. Спочатку знаходять дві точки (c_1 і d_1), що розташовані симетрично на відстані $\delta > 0$ від середини відрізка [a, b]. На першому кроці приймають $a_1 = a$, $b_1 = b$ і обчислюють зазначені точки

$c_1 = (a_1 + b_1 - \delta)/2$, $d_1 = (a_1 + b_1 + \delta)/2$.

Далі знаходять значення функції $f(x)$ у точках c_1 і d_1 та порівнюють їх:

- якщо $f(c_1) \le f(d_1)$, то на наступному кроці $a_2 = a_1$, $b_2 = d_1$;
- якщо $f(c_1) > f(d_1)$, то на наступному кроці $a_2 = c_1$, $b_2 = b_1$;

Далі обчислюють $c_2 = (a_2 + b_2 - \delta)/2$, $d_1 = (a_2 + b_2 + \delta)/2$ і знов порівнюють $f(c_2)$ з $f(d_2)$ до тих пір, поки не буде виконана умова $\varepsilon_i = (b_i - a_i)/2 \le \varepsilon$.

Назва методу пов'язана з тим, що при малих δ довжини інтервалів на кожному кроці зменшуються майже в два рази.

Приклад 4.8. Як і в попередньому прикладі розглянемо функцію $f(x) = x + 2/x$ на відрізку [0,5, 3,5]. Знайдемо мінімум функції з похибкою $\varepsilon \le 0{,}5$. Нехай $\delta = 0{,}1$.

Крок 1: $a_1 = 0{,}5$; $b_1 = 3{,}5$; $c_1 = (0{,}5 + 3{,}5 - 0{,}1)/2 = 1{,}95$;

$d_1 = (0{,}5 + 3{,}5 + 0{,}1)/2 = 2{,}05$; $f(c_1) = 2{,}976 < f(d_1) = 3{,}026$.

Крок 2: $a_2 = a_1 = 0{,}5$; $b_2 = d_1 = 2{,}05$; перевіряємо похибку на даному кроці: $\varepsilon_2 = (2{,}05 - 0{,}5)/2 = 0{,}775 > 0{,}5$. Похибка більше заданої, тому продовжуємо обчислення: $c_2 = (0{,}5 + 2{,}05 - 0{,}1)/2 = 1{,}225$, $d_2 = (0{,}5 + 2{,}05 + 0{,}1)/2 = 1{,}325$; $f(c_2) = 2{,}858 > f(d_2) = 2{,}834$.

Крок 3: $a_3 = c_2 = 1{,}225$; $b_3 = d_1 = 2{,}05$; перевіряємо похибку на даному кроці: $\varepsilon_2 = (2{,}05 - 1{,}225)/2 = 0{,}4125 < 0{,}5$. Оскільки похибка менше заданої, завершуємо розрахунки і приймаємо, що точкою мінімуму є $x_m = (2{,}05 + 1{,}225)/2 = 1{,}638$, а значення мінімуму $f^* = f(1{,}638) = 2{,}859$.

Схема звуження відрізку [0,5, 3,5] подана на рис. 4.8.

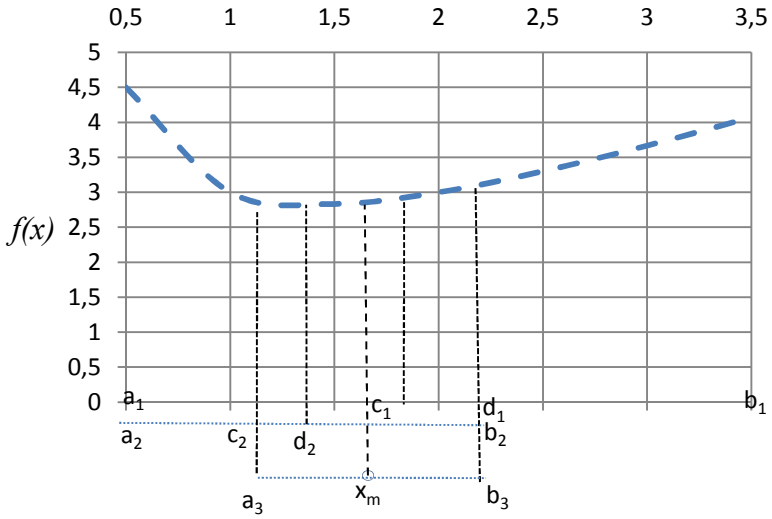


Рис. 4.8. Схема пошуку точки мінімуму за методом дихотомії

Зауважимо, що метод дихотомії дозволяє для отримання точки мінімуму із заданою похибкою звертатися до обчислення значення функції чотири рази, замість семи в методі перебору. Ще більш ефективний алгоритм дає метод ділення відрізка в пропорціях золотого перетину.

Метод золотого перетину. Нагадаємо, що точка ділить відрізок у пропорції золотого перетину, якщо відношення всієї довжини відрізка до довжини більшої його частини дорівнює відношенню довжини більшої його частини до довжини меншої. Термін «золотий перетин» увів Леонардо да Вінчі. Нехай точки c, d знаходяться симетрично відносно середини відрізка $[a, b]$ і ділять його в пропорції золотого перетину, тобто:

$$(b - a)/(d - a) = (d - a)/(b - d); \quad (b - a)/(b - c) = (b - c)/(c - a).$$

З цих виразів випливає, що

$$c = \frac{3-\sqrt{5}}{2}(b-a)+a;\ d = \frac{\sqrt{5}-1}{2}(b-a)+a.$$

На першому кроці приймаємо $a_1 = a;\ b_1 = b;\ c_1 = c;\ d_1 = d$ (рис. 4.9). Обчислюємо значення функції в точках c_1 *і* d_1 та порівнюємо їх:

– якщо $f(c_1) \le f(d_1)$, то на наступному кроці $a_2 = a_1$, $b_2 = d_1$, $d_2 = c_1$;

$$c_2 = \frac{3-\sqrt{5}}{2}(b_1 - a_1) + a_1;$$

– якщо $f(c_1) > f(d_1)$, то на наступному кроці $a_2 = c_1$, $b_2 = b_1$, $c_2 = d_1$;

$$d_2 = \frac{\sqrt{5}-1}{2}(b_1 - a_1) + a_1.$$

Знов порівнюємо значення функції в точках c_2 і d_2 і продовжуємо обчислення. На кожному кроці знаходимо похибку $\varepsilon_i = (b_i - a_i)/2$. Ітерації закінчуємо, якщо ε_i не перевищує заданої точності ε.

Приклад 4.9. Розглянемо знов функцію $f(x) = x + 2/x$ на відрізку $[a, b] = [0{,}5,\ 3{,}5]$. Знайдемо мінімум функції методом золотого перетину з похибкою $\varepsilon \le 0{,}5$.

Крок 1: $a_1 = 0{,}5$; $b_1 = 3{,}5$; $c_1 = \frac{3-\sqrt{5}}{2}(3{,}5 - 0{,}5) + 0{,}5 = 1{,}646$;

$d_1 = \frac{\sqrt{5}-1}{2}(3{,}5 - 0{,}5) + 0{,}5 = 2{,}354$; $f(c_1) = 2{,}861 < f(d_1) = 3{,}204$.

Крок 2: $a_2 = a_1 = 0{,}5$; $b_2 = d_1 = 2{,}354$; $d_2 = c_1 = 1{,}646$;

$c_2 = \frac{3-\sqrt{5}}{2}(2{,}354 - 0{,}5) + 0{,}5 = 1{,}208$. Перевіряємо похибку на даному кроці: $\varepsilon_2 = (2{,}354 - 0{,}5)/2 = 0{,}927 > 0{,}5$. Похибка більше заданої,тому порівнюємо значення функції: $f(c_2) = 2{,}864 > f(d_2) = 2{,}861$ та продовжуємо обчислення (табл. 4.3, рис. 4.9).

Таблиця 4.3

Результати обчислень за методом золотого перетину (до прикладу 4.9)

№ кроку	a_i	b_i	Інтервал	c_i	d_i	$f(c_i)$	$f(d_i)$	ε	Порівняння $f(c_i)$ та $f(d_i)$
1	0,5	3,5	3,0	1,646	2,354	2,861	3,204		$f(c_1) < f(d_1)$
2	0,5	2,354	1,854	1,208	1,646	2,864	2,861	0,927	$f(c_1) > f(d_1)$
3	1,208	2,354	1,146	1,646	1,916	2,861	2,96	0,573	$f(c_1) < f(d_1)$
4	1,208	1,916	0,708					0,354 < 0,5 – точність досягнута	

На четвертому кроці задана точність досягнута, тому обчислення завершуємо і приймаємо, що точкою мінімуму є $x_m = (1{,}208 + 1{,}916)/2 = 1{,}562$; значення мінімуму: $f^* = f(1{,}562) = 2{,}842$.

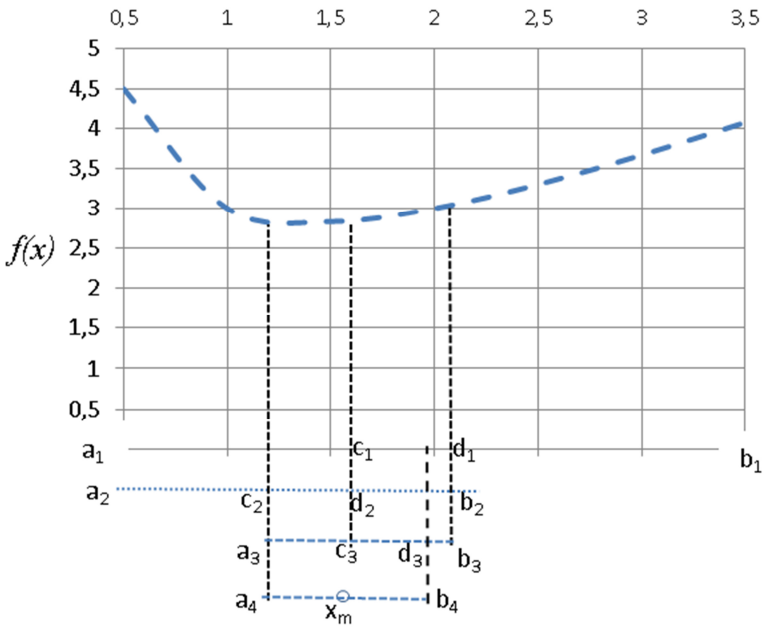


Рис. 4.9. Схема пошуку точки мінімуму за методом золотого перетину

На принципі безпосереднього обчислення значень функції в визначених точках основані й інші прямі методи мінімізації: метод Фібоначчі, метод парабол (квадратичної апроксимації), метод ламаних та інші [10].

Контрольні питання до підрозділу 4.2

1. Дайте визначення мінімуму і максимуму функції однієї змінної.
2. Сформулюйте необхідну та достатню умови існування екстремуму в точці для функції однієї змінної.
3. Що називають глобальними екстремумами функції однієї змінної на відрізку?
4. Сформулюйте необхідну та достатню умови існування екстремуму в точці для функції двох змінних.
5. Що розуміють під прямими методами мінімізації ?
6. Сформулюйте визначення унімодальності функції.
7. Опишіть метод рівномірного пошуку (сканування або перебору).
8. Опишіть метод половинного ділення (дихотомії).
9. Опишіть метод золотого перетину.

4.3. Застосування симплекс-методу для оптимізації водовідбору свердловинами

У гідрогеологічній практиці часто вникають задачі оптимального відбору підземних вод для господарсько-питного водопостачання. До таких задач відносять, зокрема:

1) оптимізацію режиму водовідбору з максимізацією дебітів в умовах обмежень на пониження рівня підземних вод,

2) раціональне розміщення свердловин при проектуванні водозаборів,

3) оптимізацію режиму водовідбору з максимізацією дебітів в умовах обмежень на вміст концентрації токсичних речовин у підземних водах.

У разі використання аналітичних співвідношень теорії фільтрації та масопереносу в підземних водах ці задачі можуть бути легко вирішені методами лінійного програмування, яке є розділом теорії оптимізації, у якому розглядаються задачі оптимізації з лінійними цільовими функціями та обмеженнями. У загальному випадку постановка *задачі лінійного програмування* може бути записана у вигляді [1, 2, 4]

$$J = \sum_{i=1}^{n} c_i x_i \to \min \tag{4.17}$$

за обмежень

$$\sum_{i=1}^{n} a_{ji} x_i = b_j, \ \ j = 1,...,m; \tag{4.18}$$

$$x_i \le x_{i,\max}, \tag{4.19}$$

де n – кількість невідомих, m – кількість обмежень.

У разі $n = 2$ задачу можна вирішити графічно, при $n > 2$ треба використовувати інші методи. Одним з поширених методів вирішення задач лінійного програмування є симплекс-метод, застосування якого розглянемо на прикладі задачі максимізації водовідбору зі свердловин у необмеженому водоносному шарі за умови обмежень на максимально допустиме зниження рівня підземних вод у кожній з них.

У загальному вигляді зниження підземних вод унаслідок водовідбору з n свердловин у напірному водоносному горизонті може бути розраховано так:

$$S_j = \sum_{i=1}^{n} a_{ji} Q_i, \tag{4.20}$$

де a_{ji} – функції впливу (або коефіцієнти у стаціонарному випадку), які враховують фільтраційні властивості водоносного шару, тип меж та відстані до них, відстані між свердловинами. З точки зору фізики коефіцієнти a_{ji}

інтерпретуються як зниження рівня в *j*-й свердловині під впливом експлуатації *i*-ї свердловини одиничного дебіту.

Тоді задача максимізації водовідбору може бути записана у вигляді

$$J = \sum_{i=1}^{n} Q_i \to \max, \tag{4.21}$$

$$\begin{cases} a_{11}Q_1 + a_{12}Q_2 + \ldots + a_{1n}Q_n \le S_1, \\ a_{21}Q_1 + a_{22}Q_2 + \ldots + a_{2n}Q_n \le S_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}Q_1 + a_{n2}Q_2 + \ldots + a_{nn}Q_n \le S_n; \end{cases} \tag{4.22}$$

$$0 \le Q_i \le Q_{i,\max};\ i = 1,\ldots,n,$$

де S_i – зниження в *i*-й свердловині. В задачах лінійного програмування зазвичай визначається мінімум цільової функції, тому будемо шукати мінімум функції – $J(Q_1,Q_2,\ldots,Q_n)$.

Згідно з алгоритмом *симплекс-методу* на першому етапі визначається так званий опорний розв'язок, який відповідає системі обмежень, а на другому – оптимальний розв'язок, для якого величина $-J(Q_1,Q_2,\ldots,Q_n)$ досягає мінімуму [1,4]. До лівої частини кожної з нерівностей (4.26) додаються невідомі y_i, після чого нерівності розглядаються як рівняння. Для кожного стовпчика обчислюють відношення S_i/a_{im} та обирають найменше з них; тоді елемент a_{im} та *m*-й рядок називають розв'язувальними.

На кожному кроці матриця системи обмежень та цільова функція перетворюються модифікованим жордановим виключенням:

а) для розв'язувального елемента $a'_{km} = 1/a_{km}$,

б) для розв'язувального рядка $a'_{ki} = a_{ki}/a_{km}$,

в) для розв'язувального стовпчика $a'_{jm} = a_{jm}/a_{km}$,

г) для всіх інших елементів таблиці, у тому числі рядка, що відповідає цільовій функції $a'_{ji} = a_{ji} - a_{jm}a_{ki}/a_{km}$.

Штрихом позначені елементи нової матриці, елементи якої обчислюються з попередньої. Після цього міняють місцями символи змінних у *m*-му стовпчику та *k*-му рядку, тобто Q_k міняється місцями з y_m. Ці операції повторюються до отримання невід'ємних значень цільової функції *J*.

Розглянемо більш детально випадок трьох свердловин ($n = 3$), які експлуатуються в необмеженому в плані горизонті напірних вод з провідністю $T = 75$ м^2/добу, який гідравлічно ізольований від інших водоносних горизонтів. Така задача з іншою кількістю свердловин була розглянута в [12]. Схема розташування свердловин радіусом $r_s = 0{,}1$ м у точках з координатами, м:

(100,100), (200,500), (500,200) показана на рис. 4.10. Потрібно знайти такий розподіл дебітів Q_1, Q_2, Q_3, який забезпечує максимальний водовідбір (сумарний дебіт), при цьому зниження рівня підземних вод S не перевищить S_{max} = 15 м, тобто за умови $S < S_{max}$. Радіус впливу свердловин R = 1000 м.

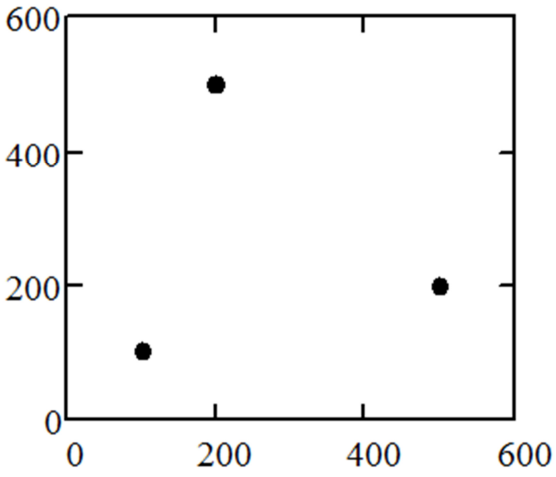


Рис. 4.10. Розташування свердловин для водовідбору, який потрібно оптимізувати

Елементи матриці, яка визначається співвідношеннями (4.21) – (4.22), обчислюються за формулами розрахунку зниження рівня підземних вод [13]

$$a_{ij} = \frac{0,366}{T} \lg \frac{R}{r_{ij}}, \tag{4.23}$$

де r_{ij} – відстань між свердловинами з індексами i та j, $r_{ij} = r_0$ при $i = j$; r_0 – радіус свердловини.

Цільова функція має вигляд

$$J = -Q_1 - Q_2 - Q_3 \to \min, \tag{4.24}$$

а обмеження для координат свердловин та параметрів водоносного горизонту, які конкретизовано за формулою (4.23), записуються у вигляді

$$\begin{cases} 0,0195Q_1 + 1,8179 \cdot 10^{-3} Q_2 + 1,8784 \cdot 10^{-3} Q_3 \le 15, \\ 1,8179 \cdot 10^{-3} Q_1 + 0,0195Q_2 + 1,8784 \cdot 10^{-3} Q_3 \le 15, \\ 1,8784 \cdot 10^{-3} Q_1 + 1,8784 \cdot 10^{-3} Q_2 + 0,0195Q_3 \le 15. \end{cases} \tag{4.25}$$

Для зручності та зменшення похибок при розрахунку нормуємо систему нерівностей (4.25), поділивши всі елементи матриці на коефіцієнти при діагональних елементах,

$$\begin{cases} Q_1 + 0{,}0931Q_2 + 0{,}0962Q_3 \le 768{,}414, \\ 0{,}0931Q_1 + Q_2 + 0{,}0962Q_3 \le 768{,}414, \\ 0{,}0962Q_1 + 0{,}0962Q_2 + Q_3 \le 768{,}414. \end{cases} \quad (4.26)$$

Подальше вирішення системи нерівностей симплекс-методом потребує виконання значної кількості рутинних арифметичних операції, які мають лише суто математичний інтерес. Детальний опис таких обчислень, який можна знайти у підручниках з методів оптимізації, у тому числі лінійного програмування [1, 2, 3, 4], не є метою даного посібника. Скористаємося тим, що процедура симплекс-методу реалізована в стандартному пакеті в середовищі Excel «Поиск решения». Приклад практичного використання цієї процедури розглянуто у практичній роботі 4 розділа 5.

Виконавши всі необхідні операції в пакеті «Поиск решения» в Excel для задачі (4.24), (4.26), отримаємо оптимальний розподіл дебітів свердловин: $Q_1 = 646{,}29$ м^3/добу, $Q_2 = 646{,}29$ м^3/добу, $Q_3 = 644{,}07$ м^3/добу. На рис. 4.11 показано розподіл знижень рівня підземних вод, що відповідає оптимальним дебітам свердловин, за яких зниження в усьому водоносному горизонті не перевищують 15 м, при цьому максимальне зниження розраховується на відстані r_0 від осі свердловини. У результаті відбору води формуються три зони зниження з центрами у місцях розміщення свердловин. Фрагмент аркуша в Excel для пошуку оптимального розподілу дебітів показано на рис. 4.12.

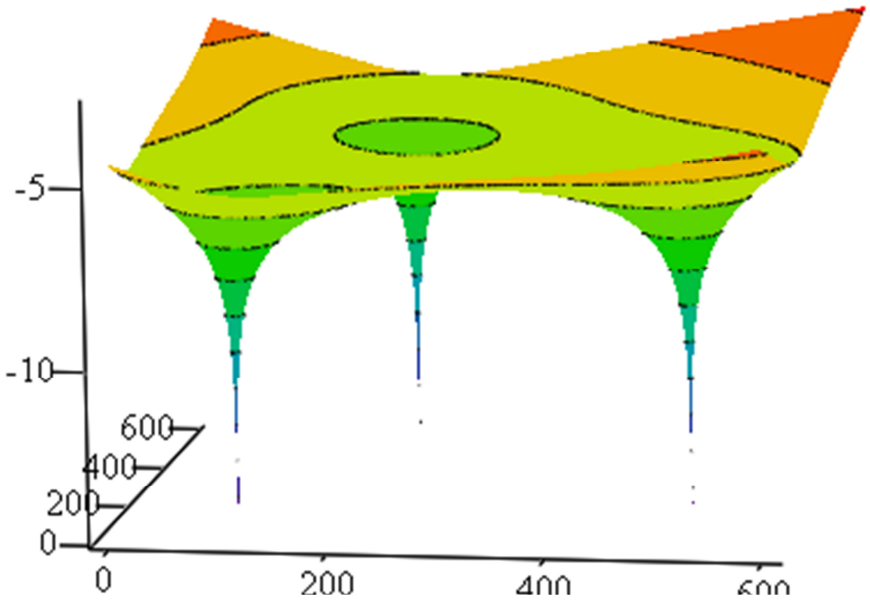


Рис. 4.11. Рівень підземних вод при роботі свердловин відносно початкового положення $z = 0$

	A	B	C	D
1	Дебіти свердловин	м3/добу		
2	Q1	646.29	За результатами розрахунку	
3	Q2	646.29		
4	Q3	644.07		
5				
6	Цільова функція, м3/добу	**1 936.6**		
7	Q1+Q2+Q3->max		S/a11	
8	Обмеження	факт		
9	a11*Q1+a12*Q2+a13*Q3<-=	768.414	768.414	свердловина 1
10	a21*Q1+a22*Q2+a23*Q3<=	768.414	768.414	свердловина 2
11	a31*Q1+a32*Q2+a33*Q3<=	768.414	768.414	свердловина 3

Рис. 4.12. Фрагмент робочого аркуша MS Excel з розрахунками оптимізації водовідбору свердловинами

Контрольні питання до підрозділу 4.1

1. Виконайте інтерпретацію змінних та коефіцієнтів у задачі оптимізації (4.24), (4.26).
2. Розкрийте суть алгоритму симплекс-методу.
3. Як виконується пошук оптимального розв’язку симплекс-методом у MS Excel?

Література до розділу 4

1. Жалдак М.І. Основи теорії і методів оптимізації / М.І. Жалдак, Ю.В. Гриус. – Черкаси: Брама-Україна, 2005. – 608 с.
2. Бейко І.В. Задачі, методи і алгоритми оптимізації: навч. посіб. / І.Б. Бейко, П.М. Зінько, О.Г. Наконечний. – Рівне: НУВГП, 2011. – 624 с.
3. Математичні методи і моделі в аграрній та природоохоронній галузях: навч. посіб. / Н.В. Попрозман, Н.А. Клименко, Л.В. Забуранна, О.І. Попрозман. – Київ: Агрармедіа Груп, 2013. – 292 с.
4. Васильев Ф.П. Методы оптимизации / Ф.П. Васильев. – Москва: Факториал-Пресс, 2002. – 824 с.
5. Мартьянова А.Е. Математические методы моделирования в геологии. учеб. пособие для студентов направления 650100 «Прикладная геология»: в 2 ч. / А.Е. Мартьянова. – Астрахань: АГТУ, 2008. – 2 ч.
6. Хоменко О.Е. Автоматизация проектирования паспортов буровзрывных работ путем оптимизации размещения шпуров / О.Е. Хоменко,

Д.В. Рудаков, М.Н. Кононенко // Форум гірників – 2011: зб. праць наук.-практ. конф. – Дніпропетровськ, 2011. – Т. 1. – С. 39 – 43.

7. Сінайський Є.С. Вища математика: навч. посіб.: у 2-х ч. / Є.С. Сінайський, Л.В. Новікова, Л.І. Заславська. – Дніпропетровськ: НГУ, 2004. – Ч.1. – 389 с.

8. Лященко М.Я. Чисельні методи: підручник / М.Я. Лященко, М.С. Голованъ. – Київ: Либідь, 1996. – 288 с.

9. Синєглазов В.М. Математичні методи оптимізації: навч. посіб. / В.М. Синєглазов, О.А. Зеленков, Ш.І. Аскеров; Нац. авіаційний ун-т. – Київ: Освіта України, 2018. – Ч. 1. – 329 с.

10. Аббасов М.Э. Методы оптимизации: учеб. пособие / М.Э. Аббасов. – Санкт-Петербург: Изд-во ВВМ, 2014. – 64 с.

11. Загорулько А.В. Чисельні методи у механіці: навч. посіб. / А.В. Загорулько. – Суми: Вид-во СумДУ, 2008. – 186 с.

12. Гавич И.К. Гидрогеодинамика: учеб. для вузов / И.К. Гавич. – Москва: Недра, 1988. – 349 с.

13. Шестаков В.М. Динамика подземных вод: учебник / В.М. Шестаков. – Москва: МГУ, 1979. – 369 с.

Розділ 5
ПРАКТИЧНІ РОБОТИ

5.1. Перевірка гіпотези про рівномірний розподіл кількості зсувів протягом року

Завдання. За даними моніторингу щомісячно реєструється загальна кількість зсувів на території району. На основі цих даних висловлено припущення про активізацію зсувів у весняний період, тобто про те, що кількість зсувів змінюється протягом року. Для підтвердження чи спростування цього слід перевірити статистичну гіпотезу щодо рівномірного розподілу кількості зсувів на місяць як випадкової величини протягом року як періоду спостережень. Аналогічна задача щодо нерівномірності оповзань борту кар'єру розглядалася в [1].

Відомості з теорії. Будемо наближено розглядати рік як цикл з періодом 2π та 12 інтервалами (місяцями), при цьому кожному місяцю відповідатиме дуга окружності з кутом $2\pi/12 = 30°$. Середини дуг, що відповідають кожному місяцю, мають кутові координати

$$\overline{\theta}_i = \frac{\pi\left(2i-1\right)}{12},\ i = 1,...,12\,, \tag{5.1}$$

розрахунок яких показано на фрагменті аркуша MS Excel (рис. 5.1).

Випадкова кутова величина рівномірно розподілена, якщо її щільність розподілу ймовірностей визначається як $f(\theta) = 1/2\pi$. У цьому разі вибіркові значення не концентруються на якомусь інтервалі та мають максимальний розкид, оскільки кутова дисперсія при цьому розподілі дорівнює 1.

Цей розподіл зустрічається у замірах орієнтування уламків у делювіальних відкладах і вулканічних брекчіях [1].

Гіпотезу про рівномірний розподіл кутової величини при малому об'ємі вибірки можна перевірити за критерієм рівномірності Релея. За вибірковими даними обчислюється величина

$$\overline{R} = \sqrt{\overline{S}_{\cos}^2 + \overline{S}_{\sin}^2}\,, \tag{5.2}$$

де $\overline{S}_{\cos} = \frac{1}{n}\sum_{i=1}^{n}\cos\theta_i$, $\overline{S}_{\sin} = \frac{1}{n}\sum_{i=1}^{n}\sin\theta_i$.

Величина $\overline{R}$ порівнюється з її критичним значенням $\overline{R}_0$ для об'єму вибірки n і прийнятого рівня значущості α (Додаток Е). При $n > 100$ величина $2n\overline{R}^2$ наближено розподілена за законом χ^2 з двома ступенями вільності [2].

Розрахунки. Відповідно до формул (5.1) та (5.2) проведемо розрахунки (рис. 5.1), згідно з якими одержимо

$$\overline{S}_{\sin} = \left(\frac{5,019}{19}\right) = 0,264; \quad \overline{S}_{\cos} = \left(\frac{-5,347}{19}\right) = -0,281; \quad \overline{R} = \sqrt{\overline{S}_{\sin}^2 + \overline{S}_{\cos}^2} = 0,386.$$

При розрахунку величин у стовпчиках D-G значення середини інтервалу переводяться з градусів у радіани.

	A	B	C	D	E	F	G
1	Місяць	Кількість зсувів	Середина інтервалу, град	sin()	ni*sin()	cos()	ni*cos()
2	I	0	15	0.259	0.000	0.966	0.000
3	II	1	45	0.707	0.707	0.707	0.707
4	III	2	75	0.966	1.932	0.259	0.518
5	IV	5	105	0.966	4.830	-0.259	-1.294
6	V	3	135	0.707	2.121	-0.707	-2.121
7	VI	1	165	0.259	0.259	-0.966	-0.966
8	VII	2	195	-0.259	-0.518	-0.966	-1.932
9	VIII	1	225	-0.707	-0.707	-0.707	-0.707
10	IX	2	255	-0.966	-1.932	-0.259	-0.518
11	X	1	285	-0.966	-0.966	0.259	0.259
12	XI	1	315	-0.707	-0.707	0.707	0.707
13	XII	0	345	-0.259	0.000	0.966	0.000
14	Разом	**19**			**5.019**		**-5.347**
15				Рівні значущості			
16	S_{sin}	0.264		0.1	0.05	0.025	
17	S_{cos}	-0.281	R0	0.348	0.394	0.438	
18	R	0.386		підтвердж.	не підтв.	не підтв.	

Рис. 5.1. Фрагмент аркуша MS Excel з розрахунками для перевірки гіпотези про рівномірний розподіл кількості зсувів протягом року

Висновок. Таким чином, значення величини $\overline{R} = 0,386$ перевищує її критичні значення з розподілу Релея для рівня значущості $\alpha = 0,1$ для об'єму вибірки $n = 19$, що згідно з Додатком Е дорівнює 0,348. На цьому рівні значущості гіпотеза про рівномірний розподіл кількості зсувів протягом року відкидається, тобто можна стверджувати, що активізація зсувів має сезонний характер. При цьому значення величини $\overline{R} = 0,386$ не перевищує її критичні значення з розподілу Релея 0,394 та 0,405 для рівнів $\alpha = 0,05$ та 0,025 для об'єму вибірки $n = 19$. Тому на цих рівнях значущості гіпотеза про рівномірний розподіл кількості зсувів протягом року не відкидається.

Слід враховувати, що при обробці та аналізі геологічних даних з відносно невеликим об'ємом вибірки використання високого рівня значущості з $\alpha < 0{,}05$ часто не є виправданим. З урахуванням цього можна відкинути гіпотезу про рівномірний розподіл та зробити висновок про сезонну активізацію зсувів.

Оформлення роботи. Звіт з виконаної роботи має містити постановку завдання з власним варіантом даних для розрахунку, скріншот з аркуша MS Excel, аналогічний рис. 5.1, та висновок з інтерпретацією отриманого результату.

5.2. Однофакторний дисперсійний аналіз зразків вугілля

Завдання. Унаслідок розвідувань на ділянці вугільного родовища відібрані зразки з кількох шарів з різним умістом вуглецю. Для вибірок різного об'єму з різних шарів визначені класифікаційні показники, зокрема, теплота згоряння вугілля. Було висловлено припущення, що теплота згоряння залежить від того, з якого вугільного пласта були відібрані зразки. Якщо кількість вибірок перевищує дві, то для аналізу будемо використовувати однофакторний дисперсійний аналіз.

Відомості з теорії. Для вирішення завдання застосуємо процедуру нерівномірного однофакторного дисперсійного аналізу (підрозділ 3.2) для трьох вибірок ($m = 3$). Позначимо через x_{1i}, x_{2i}, x_{3i} елементи вибірок (теплоту згоряння вугілля) з об'ємами q_1, q_2, q_3 відповідно.

Тоді загальна сума квадратів відхилень визначається за формулою [1]

$$C_g = \left(U_1 + U_2 + \ldots + U_m\right) - \frac{\left(V_1 + V_2 + \ldots + V_m\right)^2}{n}, \tag{5.3}$$

де $U_j = \sum_{i=1}^{q_j} x_{ij}^2$ – сума квадратів елементів j-ї вибірки, тобто вимірювань з j-го пласта, $V_j = \sum_{i=1}^{q_j} x_{ji}$ – суми елементів вибірки з j-го пласта, n – загальна кількість зразків вугілля, $n = q_1 + q_2 + \ldots + q_m$.

Факторна сума квадратів відхилень обчислюється за формулою

$$C_f = \left(\frac{U_1^2}{q_1} + \frac{U_2^2}{q_2} + \frac{U_3^2}{q_3}\right) - \frac{\left(V_1 + V_2 + V_3\right)^2}{n}. \tag{5.4}$$

Залишкова сума квадратів відхилень обчислюється за формулою

$$C_r = C_g - C_f. \tag{5.5}$$

Значення критерію Фішера обчислюється як відношення

$$F = \frac{S_f^2}{S_r^2}, \tag{5.6}$$

де $S_g^2 = \frac{C_g}{n-1}$, $S_f^2 = \frac{C_f}{m-1}$, $S_r^2 = \frac{C_r}{n-m}$.

Величина F порівнюється з критичним для заданого рівня значущості α та числа ступенів вільності $k_1 = m - 1$ та $k_2 = n - m$. Якщо $F > F(\alpha, k_1, k_2)$, то можна стверджувати, що фактор відмінності пластів є статистично значущим і його слід брати до уваги при аналізі властивостей вугілля за просторовим розташуванням пластів та походженням. Інакше цей фактор вважається незначущим.

Розрахунки. Значення елементів трьох вибірок різного об'єму вводяться в стовпчики D, C, F (рис. 5.2). Проміжні величини обчислюються у рядках 18–23. Основні величини для розрахунку статистики Фішера F у рядках 25–27 розраховуються за формулами (5.4) – (5.6). Критичне значення з розподілу Фішера у комірці E30 обчислюється функцією «F.ОБР.ПХ(B30;C30;D30)».

Правильність виконаних розрахунків можна перевірити стандартною процедурою однофакторного аналізу, яка додається до меню «Данные» в меню «Надстройки» в Excel. Результат цього розрахунку показано на рис. 5.3.

Обчислене значення статистики Фішера F 3,16 не перевищує критичне значення 3,28, отже, для даних вибірок вплив фактора статистично не значущий. Разом з тим, це значення вище за критичне для більш слабкого рівня значущості 0,1, таким чином, результат є чутливим до незначних похибок у вимірюваннях. Зокрема, у разі збільшення лише одного елемента третьої вибірки у комірці F4 з 33,0 до 33,2 значення статистики Фішера становить 3,31 і перевищує критичне для рівня значущості 0,05, що може обґрунтувати висновок про значущість фактора положення вугільного пласта на теплоту згоряння вугілля на даному родовищі.

При виконанні роботи потрібно підтвердити правильність своїх розрахунків стандартною процедурою однофакторного дисперсійного аналізу в MS Excel. Варіанти для розрахунку надаються викладачем окремо.

Оформлення роботи. Звіт з виконаної роботи має містити постановку завдання з власним варіантом даних для розрахунку, скріншоти з аркуша MS Excel, аналогічні рис. 5.2 та 5.3, висновок з інтерпретацією отриманого результату.

	A	B	C	D	E	F	G
1		Пласт 1		Пласт 2		Пласт 3	
2	№ елемента	x_{1i}	x_{1i}^2	x_{2i}	x_{2i}^2	x_{3i}	x_{3i}^2
3	1	32.7	1069.3	31.1	967.2	32.2	1036.8
4	2	33.1	1095.6	31.3	979.7	33.0	1089.0
5	3	31.5	992.3	32.6	1062.8	33.9	1149.2
6	4	31.7	1004.9	33.9	1149.2	33.0	1089.0
7	5	32.1	1030.4	33.0	1089.0	33.1	1095.6
8	6	32.3	1043.3	32.1	1030.4	32.1	1030.4
9	7	31.2	973.4	32.1	1030.4	32.6	1062.8
10	8	31.7	1004.9	32.2	1036.8	32.2	1036.8
11	9	32.2	1036.8	31.7	1004.9	33.7	1135.7
12	10	32.5	1056.3	32.2	1036.8	32.1	1030.4
13	11	32.1	1030.4			33.1	1095.6
14	12	32.9	1082.4			32.2	1036.8
15	13	31.9	1017.6				
16	14	33.0	1089.0				
17	15	31.1	967.2				
18	Кількість	15		10		12	
19	Сума	482.0	15493.8	322.2	10387.3	393.2	12888.2
20	Середнє	32.13		32.22		32.77	
21	Дисперсія	0.395238		0.664		0.39697	
22			елементів n	факторів	*n-1*	*p-1*	*n-p*
23	Загальна кількість		37	3	36	2	34
24							
25	*Cg*	18.82703		*S^2_g*	0.522973	***F***	**3.159956**
26	*Cf*	2.951027		*S^2_f*	1.475514		
27	*Cr=Cg-Cf*	15.876		*S^2_r*	0.466941		
28							
29	Фішер-тест	*a*	*k1*	*k2*	*Fcr(a,k1,k2)*		
30		0.05	2	34	**3.275898**		

Рис. 5.2. Результати розрахунків за однофакторним дисперсійним аналізом

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Пласт 1	Пласт 2	Пласт 3									
2	№ елеме	x_{1i}	x_{2i}	x_{3i}			Anova: Single Factor						
3	1	32.7	31.1	32.2									
4	2	33.1	31.3	33.0			SUMMARY						
5	3	31.5	32.6	33.9			*Groups*	*Count*	*Sum*	*Average*	*Variance*		
6	4	31.7	33.9	33.0			Column 1	15	482	32.13333	0.395238		
7	5	32.1	33.0	33.1			Column 2	10	322.2	32.22	0.664		
8	6	32.3	32.1	32.1			Column 3	12	393.2	32.76667	0.39697		
9	7	31.2	32.1	32.6									
10	8	31.7	32.2	32.2									
11	9	32.2	31.7	33.7			ANOVA						
12	10	32.5	32.2	32.1			*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
13	11	32.1		33.1			Between Groups	2.951027	2	1.475514	3.159956	0.05512	3.275898
14	12	32.9		32.2			Within Groups	15.876	34	0.466941			
15	13	31.9											
16	14	33.0					Total	18.82703	36				
17	15	31.1											

Рис. 5.3. Результати однофакторного дисперсійного аналізу стандартною процедурою в Excel

5.3. Перевірка кореляції між двома вибірками. Ранговий коефіцієнт Спірмена

Завдання. При проведенні гідрогеологічних досліджень з метою вивчення потужності та гідравлічної провідності водоносних шарів на профілі свердловин були виконані геофізичні роботи методом електророзвідки. На основі отриманих даних було висловлено припущення щодо існування кореляційного зв'язку між електричним опором порід ρ і відносною потужністю горизонту гравійно-галькових відкладів m_r, до яких віднесені водоносні горизонти. Для підтвердження чи спростування цього припущення слід визначити коефіцієнт кореляції та його статистичну значущість.

Відомості з теорії. Гіпотезу про наявність кореляційного зв'язку між значеннями електричного опору порід і відносною потужністю водоносного горизонту перевіримо за допомогою рангового коефіцієнта кореляції Спірмена та порівняємо його зі звичайним коефіцієнтом кореляції, визначивши їх статистичну значущість.

У разі малого об'єму вибірки важко перевірити гіпотезу про відповідність емпіричного розподілу, зокрема, нормальному закону. Тоді використання звичайного коефіцієнта кореляції стає під питанням. Іноді розкид однієї з величин є доволі широким і сягає 1 – 2 порядків, тоді значення іншої величини мають значно менший розкид.

У такому випадку для перевірки гіпотези про наявність кореляційного зв'язку можна використовувати ранговий коефіцієнт кореляції Спірмена [2]. Для його розрахунку вибіркові значення досліджуваних випадкових величин замінюються їх рангами в порядку зростання.

Рангом називається номер елемента вибірки в порядку зростання. Вважається, що якщо між вибірковими значеннями немає кореляційної залежності, то ранги цих величин теж будуть незалежними.

Ранговий коефіцієнт кореляції розраховується за формулою

$$r = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^{n} d_i^2 , \tag{5.7}$$

де d_i – різниця рангів відповідних значень досліджуваних величин; n – кількість пар у вибірці. У розглянутому випадку досліджувані величини є електричний опір порід ρ і відносна потужність горизонту гравійно-галькових відкладів m_r.

За наявності кількох пар з однаковими значеннями рангу коефіцієнта кореляції дещо змінюється [1]:

$$r = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^{n} \left(d_i^2 + T_1^2 + T_2^2 \right), \tag{5.8}$$

де T_1 і T_2 – поправки на повторення елементів з однаковими рангами у першій та другій вибірках відповідно. Вони обчислюються за формулою

$$T_\zeta^2 = \frac{1}{12}\sum_{i=1}^{k}(l_{\zeta,i}^3 - l_{\zeta,i}), \tag{5.9}$$

де $l_{\zeta,i}$ – кількість даних з повторюваними рангами елемента у вибірці значень параметра ζ; k – кількість груп з повторюваними рангами.

Для перевірки значущості рангового коефіцієнта кореляції можна використовувати величину $r_0 = Z(p)/\sqrt{n-1}$, де $Z(p)$ – значення зворотної функції нормального розподілу при довірчій ймовірності p. Якщо $r > r_0$, то кореляційний зв'язок можна вважати статистично значущим.

Висновки щодо існування кореляційного зв'язку на основі коефіцієнта Спірмена можна порівняти з перевіркою за звичайним коефіцієнтом кореляції. Його статистична значущість перевіряється за формулою (2.98).

Розрахунки. Відповідно до формул (5.7) – (5.9) проведемо обчислення, результати яких наведено на рис. 5.4. Значення зворотної функції нормального розподілу $Z(p)$ для довірчої ймовірності 0,95 дорівнює 1,64, а для довірчої ймовірності 0,99 – відповідно 2,33.

Значення рангу в стовпчиках C та E обчислюються за допомогою функції «РАНГ.СР» у вигляді

$$R_i = n - РАНГ.СР(x_i, X), \tag{5.10}$$

де R_i – ранг i-го елемента вибірки x_i у вибірці X з n елементами; функція «РАНГ.СР» обчислює середнє значення рангу в разі кількох однакових елементів у вибірці.

Висновок. Розраховане значення рангового коефіцієнта кореляції 0,65 перевищує критичне значення 0,494 для довірчої ймовірності $p = 0{,}95$ (рівень значущості $\alpha = 0{,}05$); отже, кореляційний зв'язок між значеннями відносної потужності та електричного опору водоносного горизонту підтверджується. Звичайний коефіцієнт кореляції 0,664 близький до рангового коефіцієнта кореляції, а відповідна статистика $t = 2{,}808$ перевищує відповідне критичне значення для розподілу Стьюдента 2,23 з довірчою ймовірністю $p = 0{,}95$.

Кореляційний зв'язок не підтверджується для більшої довірчої ймовірності $p = 0{,}99$ (рівень значущості $\alpha = 0{,}01$), як для рангового коефіцієнта кореляції 0,65, що не перевищує критичного значення $Z(0{,}99) = 0{,}703$, так і для звичайного коефіцієнта кореляції 0,664, статистика для якого 2,808 не перевищує відповідного критичного значення 3,17 з розподілу Стьюдента.

Зважаючи на те, що для геологічних даних, особливо при невеликому об'ємі вибірки рівень значущості $\alpha = 0{,}05$ є прийнятним, можна стверджувати

про тісний кореляційний зв'язок між відносною потужністю та електричним опором водоносного горизонту. Тому використання даного геофізичного методу при проведенні гідрогеологічних досліджень у даному випадку є обґрунтованим.

	A	B	C	D	E	F
1	№ свердловини	відносна потужність		*електричний опір*		d_i^2
2		значення, %	ранг	значення, %	ранг	
3	1	63	9	203	9	0
4	2	90	12	214	10	4
5	3	42	6.5	131	7	0.25
6	4	27	3	68	5.5	6.25
7	5	29	4	68	5.5	2.25
8	6	31	5	19	1	16
9	7	20	1	52	4	9
10	8	69	10	191	8	4
11	9	22	2	29	2	0
12	10	42	6.5	353	11	20.25
13	11	76	11	481	12	1
14	12	48	8	49	3	25
15	**Коефіцієнт Спірмена**				Сума	88
16	n	12	вибірка 1		вибірка 2	
17	*повторювані ранги*		*2*		*2*	
18		p=0.95	p=0.99		T_1^2	6
19	Z(p)	1.64	2.33		T_2^2	6
20	r	0.650		$Sum(d_i^2)+T_1^2+T_2^2$		100
21	r0	0.494	0.703			
22	**Звичайний коефіцієнт кореляції**					
23	r	0.664		alpha=0.05	alpha=0.02	alpha=0.01
24	t	2.808	$t_{alpha,n-2}$	2.23	2.76	3.17

Рис. 5.4. Фрагмент аркуша MS Excel з розрахунками для визначення кореляційного зв’язку між електричним опором порід і відносною потужністю водоносного горизонту

Оформлення роботи. Звіт з виконаної роботи має містити постановку завдання з власним варіантом даних для розрахунку, скріншот з аркуша MS Excel, аналогічний рис. 5.4 та висновок з інтерпретацією отриманого результату.

5.4. Оптимізація видобутку корисних копалин

Завдання. Видобування залізної руди здійснюється трьома підприємствами однієї компанії. Компанія має певні технічні та фінансові обмеження на проведення геологорозвідувальних, розкривних, підготовчих та очисних робіт, а також на рекультивацію порушених земель. Виходячи з можливих коливань вартості сировини, необхідно оптимізувати видобуток руди на трьох підприємствах для досягнення максимуму ринкової вартості сировини. Всі оцінки проводяться для періоду один рік.

Відомості з теорії. Обсяги видобутку сировини позначимо через x_1, x_2, x_3 (т), а її ринкову вартість, яка залежить від якості сировини,через a_1, a_2, a_3 (ум. од./т) відповідно.

Загальні максимально можливі витрати для компанії на всіх трьох підприємствах становлять (в ум. од.): на проведення геологорозвідувальних, розкривних, підготовчих та очисних робіт b_0, експлуатаційні витрати, у тому числі на вентиляцію та водовідлив, c_0, на рекультивацію, у тому числі закладання виробленого простору, d_0; на управління та підтримку інфраструктури f_0. Питомі витрати (ум. од./т) на проведення геологорозвідувальних робіт на кожному підприємстві позначимо через b_1, b_2, b_3; питомі експлуатаційні витрати, у тому числі на розкриття, підготовку та очисні робіти, c_1, c_2, c_3; питомі витрати на рекультивацію, у тому числі закладання виробленого простору, d_1, d_2, d_3; питомі витрати на управління та підтримку інфраструктури, f_1, f_2, f_3.

Тоді задачу оптимізації можна формалізувати у вигляді

$$J = a_1x_1 + a_2x_2 + a_3x_3 \to \max, \tag{5.11}$$

$$\begin{cases} b_1x_1 + b_2x_2 + b_3x_3 \le b_0, \\ c_1x_1 + c_2x_2 + c_3x_3 \le c_0, \\ d_1x_1 + d_2x_2 + d_3x_3 \le d_0, \\ f_1x_1 + f_2x_2 + f_3x_3 \le f_0. \end{cases} \tag{5.12}$$

Тут J – цільова функція, що дорівнює ринковій вартості сировини, видобутої на трьох підприємствах.

Отже, завдання полягає в тому, щоб знайти x_1, x_2, x_3, за яких значення цільової функції J буде максимальним. Крім того, необхідно дати прогноз доходу компанії, який визначається різницею ринкової вартості. При цьому виробничі потужності i-го підприємства і компанії в цілому обмежені: $x_{i,\max}$ та $x_{\Sigma,\max}$ на рік відповідно, тобто

$$\begin{cases} x_1 \le x_{1,\max}, \\ x_2 \le x_{2,\max}, \\ x_3 \le x_{3,\max}, \\ x_1 + x_2 + x_3 \le x_{\Sigma,\max}. \end{cases} \quad (5.13)$$

Розрахунки. Цільова функція (5.11) за умов (5.12), (5.13) максимізується симплекс-методом (підрозділ 4.3), який реалізовано в програмі MS Excel. Результати оптимізації видобутку руди для набору вихідних даних (табл. 5.1) наведені на рис. 5.5.

Таблиця 5.1

Вихідні дані для оптимізації видобутку руди

Параметри моделі	Підприємство 1	Підприємство 2	Підприємство 3	Компанія
Ринкова вартість руди	a_1, ум. од./т	a_2, ум. од./т	a_3, ум. од./т	
	95	90	85	
Максимальна виробнича потужність	$x_{1,\max}$, млн т	$x_{2,\max}$, млн т	$x_{3,\max}$, млн т	$x_{\Sigma,\max}$, млн т
	2,5	3,0	4,0	9,5
Обмеження				
Геологорозвідувальні роботи	b_1, ум. од./т	b_2, ум. од./т	b_3, ум. од./т	b_0, ум. од.
	0,7	0,6	0,5	5,4 млн
Експлуатаційні витрати: розкриття, підготовка та очисні роботи	c_1, ум. од./т	c_2, ум. од./т	c_3, ум. од./т	c_0, ум. од.
	42	40	38	360,0 млн
Рекультивація	d_1, ум. од./т	d_2, ум. од./т	d_3, ум. од./т	d_0, ум. од.
	2,0	2,1	2,2	18,0 млн
Управління та інфраструктура	f_1, ум. од./т	f_2, ум. од./т	f_3, ум. од./т	f_0, ум. од.
	0,8	0,9	1,0	8,1 млн

При розрахунку в стовпчик А заносяться описові дані – коментарі. У комірці В6 задається функція `=D3*B2+E3*B3+F3*B4`, у комірках В9 – В16 – функції, що відповідають обмеженням:

В9 `=0,5*B2+0,6*B3+0,5*B4`;

В10 `=1*B2+1,1*B3+1,1*B4`;

B11 =0,5*B2+0,6*B3+0,5*B4;

B12 =0,1*B2+0,1*B3+0,1*B4;

B13 =B2; B14 =B3; B15 =B4; B16 =B2+B3+B4.

Для проведення оптимізації треба додати пакет «Поиск решения» у «Надстройках» Excel, який стає видимим у меню «Данные». Після натискання там мишею на «Поиск решения» з'являється вікно (рис. 5.6), де задається комірка цільової функції B6 в «Изменения ячейки переменных» та комірки, де задаються обмеження «В соответствии с ограничениями»: B9 – B16 для лівої частини нерівностей (5.12), (5.13) та C9 – C16 для їх правої частини.

У комірках D1:F5 задаються вихідні дані для розрахунку зміни вартості видобутку при зміні ринкової ціни руди, причому цільова функція залежить від значень у комірках D3:F3, які у свою чергу перемножуються на коефіцієнт у комірці D5. У разі падіння ціни він стає менше 100%, при зростанні ціни – більшим за 100%.

	A	B	C	D	E	F
1	Обсяги видобутку	тонн	млн. тонн	Вартість руди, $/т		
2	X1	2 500 000.0	2.50	a1	a2	a3
3	X2	3 000 000.0	3.00	104.5	99	93.5
4	X3	3 045 454.5	3.05	Ціна, %		
5		$	млн. $	110		
6	Цільова функція	**843 000 000.0**	843.00			
7	87*X1+85*X2+86*X3->max					
8	Обмеження	факт	максимум			
9	0.7*X1+0.6*X2+0.5*X3<=	5 072 727.3	5 400 000.0	геолрозвідка, очисні роботи		
10	42*X1+40*X2+38*X3<=	340 727 272.7	360 000 000.0	експлуатаційні витрати		
11	2*X1+2.1*X2+2.2*X3<=	18 000 000.0	18 000 000.0	рекультивація		
12	0.8*X1+0.9*X2+1.0*X3<=	7 745 454.5	8 100 000.0	управління, інфраструктура		
13	X1<=	**2 500 000.0**	2 500 000.0	100.0	%	
14	X1<=	**3 000 000.0**	3 000 000.0	100.0	%	
15	X1<=	**3 045 454.5**	4 000 000.0	76.1	%	
16	X1+X2+X3<=	**8 545 454.5**	9 500 000.0			
17						
18	Дохід, млн. $	**471.45**				

Рис. 5.5. Фрагмент аркуша MS Excel з результатами оптимізації видобутку руди

Результати та висновки. Підприємства 1 та 2 з більшою ціною руди будуть повністю задіяні на максимум своїх потужностей (2,5 та 3 млн тонн

відповідно), у той час як підприємство 3 буде задіяне лише на 76,13% (3,045 млн тонн з 4 млн тонн). Найбільша вартість руди становитиме 766,36 млн ум. од., тоді як дохід компанії за винятком всіх витрат – 394,82 млн ум. од.

При зниженні ціни на 10% сумарна вартість руди зменшується до 689,73 млн ум. од. (на 10%), а дохід – до 318,18 млн ум. од. (на 19,4%). При зростанні ціни на 10% вартість руди збільшується до 843,0 млн ум. од. (на 10%), а дохід – до 471,75 млн ум. од. (на 19,5%). Дохід змінюється більше за рахунок незмінності експлуатаційних та інших витрат.

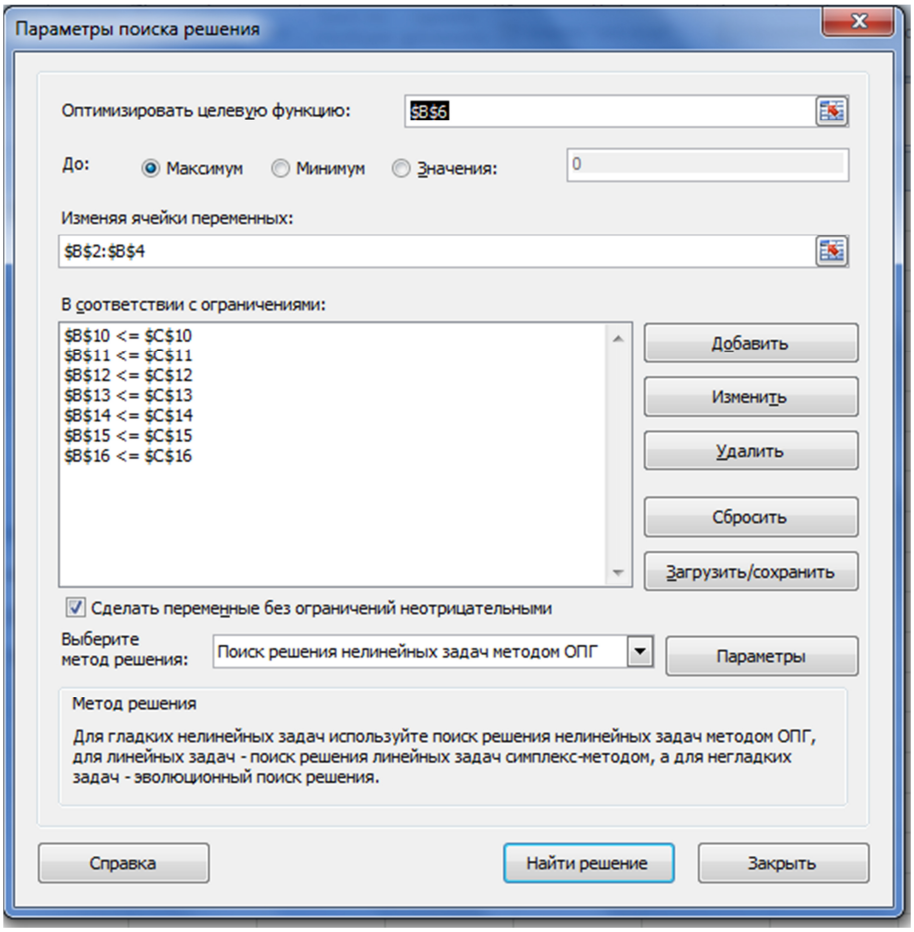


Рис. 5.6. Діалогове вікно для завдання параметрів оптимізації в програмі MS Excel

Оформлення роботи. Звіт з виконаної роботи має містити постановку завдання з власним варіантом даних для розрахунку, скріншот з аркуша MS Excel, аналогічний рис. 5.5, та висновок з інтерпретацією отриманого результату.

Результатом розрахунків має бути таблиця, яка сформована за зразком табл. 5.7, де показані параметри видобутку (вартість, дохід та обсяги видобутку на трьох підприємствах) у діапазоні змін ціни від 80 до 120% з кроком 10%.

Таблиця 5.7

Зразок оформлення результатів розрахунків практичної роботи 4

Ціна руди, %	Видобуток руди, млн т			Вартість руди, ум. од.	Дохід, ум. од.
	Підприємство 1	Підприємство 2	Підприємство 3		
80					
90					
100					
110					
120					

Література до розділу 5

1. Мартьянова А.Е. Математические методы моделирования в геологии: учеб. пособие для студентов направления 650100 «Прикладная геология»: в 2 ч. / А.Е. Мартьянова. – Астрахань: АГТУ, 2008. – 2 ч.

2. Каждан А.Б. Математическое моделирование в геологии и разведке полезных ископаемых: учеб. пособие / А.Б. Каждан, О.И. Гуськов, А.А. Шиманский. – Москва: Недра, 1979. – 168 с.

ПРЕДМЕТНИЙ ПОКАЖЧИК

Додатки

Додаток А

Значення нормованої інтегральної функції нормального розподілу

$$F_0(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{t^2}{2}} d\tau$$

x	$F_0(x)$	x	$F_0(x)$	x	$F_0(x)$	x	$F_0(x)$	x	$F_0(x)$	x	$F_0(x)$
0	0,5	-0,4	0,3446	-0,8	0,2119	-1,2	0,1151	-1,6	0,0548	-2	0,0228
-0,01	0,496	-0,41	0,3409	-0,81	0,209	-1,21	0,1131	-1,61	0,0537	-2,1	0,0179
-0,02	0,492	-0,42	0,3372	-0,82	0,2061	-1,22	0,1112	-1,62	0,0526	-2,2	0,0139
-0,03	0,488	-0,43	0,3336	-0,83	0,2033	-1,23	0,1093	-1,63	0,0516	-2,3	0,0107
-0,04	0,484	-0,44	0,33	-0,84	0,2005	-1,24	0,1075	-1,64	0,0505	-2,4	0,0082
-0,05	0,4801	-0,45	0,3264	-0,85	0,1977	-1,25	0,1056	-1,65	0,0495	-2,5	0,0062
-0,06	0,4761	-0,46	0,3228	-0,86	0,1949	-1,26	0,1038	-1,66	0,0485	-2,6	0,0047
-0,07	0,4721	-0,47	0,3192	-0,87	0,1922	-1,27	0,102	-1,67	0,0475	-2,7	0,0035
-0,08	0,4681	-0,48	0,3156	-0,88	0,1894	-1,28	0,1003	-1,68	0,0465	-2,8	0,0026
-0,09	0,4641	-0,49	0,3121	-0,89	0,1867	-1,29	0,0985	-1,69	0,0455	-2,9	0,0019
-0,1	0,4602	-0,5	0,3085	-0,9	0,1841	-1,3	0,0968	-1,7	0,0446	-3	0,0013
-0,11	0,4562	-0,51	0,305	-0,91	0,1814	-1,31	0,0951	-1,71	0,0436	-3,1	0,001
-0,12	0,4522	-0,52	0,3015	-0,92	0,1788	-1,32	0,0934	-1,72	0,0427	-3,2	0,0007
-0,13	0,4483	-0,53	0,2981	-0,93	0,1762	-1,33	0,0918	-1,73	0,0418	-3,3	0,0005
-0,14	0,4443	-0,54	0,2946	-0,94	0,1736	-1,34	0,0901	-1,74	0,0409	-3,4	0,0003
-0,15	0,4404	-0,55	0,2912	-0,95	0,1711	-1,35	0,0885	-1,75	0,0401	-3,5	0,0002
-0,16	0,4364	-0,56	0,2877	-0,96	0,1685	-1,36	0,0869	-1,76	0,0392	-3,6	0,0002
-0,17	0,4325	-0,57	0,2843	-0,97	0,166	-1,37	0,0853	-1,77	0,0384	-3,7	0,0001
-0,18	0,4286	-0,58	0,281	-0,98	0,1635	-1,38	0,0838	-1,78	0,0375	-3,8	0,0001
-0,19	0,4247	-0,59	0,2776	-0,99	0,1611	-1,39	0,0823	-1,79	0,0367	-3,9	0
-0,2	0,4207	-0,6	0,2743	-1	0,1587	-1,4	0,0808	-1,8	0,0359	0	0,5
-0,21	0,4168	-0,61	0,2709	-1,01	0,1562	-1,41	0,0793	-1,81	0,0351	0,01	0,504
-0,22	0,4129	-0,62	0,2676	-1,02	0,1539	-1,42	0,0778	-1,82	0,0344	0,02	0,508
-0,23	0,409	-0,63	0,2643	-1,03	0,1515	-1,43	0,0764	-1,83	0,0336	0,03	0,512
-0,24	0,4052	-0,64	0,2611	-1,04	0,1492	-1,44	0,0749	-1,84	0,0329	0,04	0,516
-0,25	0,4013	-0,65	0,2578	-1,05	0,1469	-1,45	0,0735	-1,85	0,0322	0,05	0,5199
-0,26	0,3974	-0,66	0,2546	-1,06	0,1446	-1,46	0,0721	-1,86	0,0314	0,06	0,5239
-0,27	0,3936	-0,67	0,2514	-1,07	0,1423	-1,47	0,0708	-1,87	0,0307	0,07	0,5279
-0,28	0,3897	-0,68	0,2483	-1,08	0,1401	-1,48	0,0694	-1,88	0,0301	0,08	0,5319
-0,29	0,3859	-0,69	0,2451	-1,09	0,1379	-1,49	0,0681	-1,89	0,0294	0,09	0,5359
-0,3	0,3821	-0,7	0,242	-1,1	0,1357	-1,5	0,0668	-1,9	0,0287	0,1	0,5398
-0,31	0,3783	-0,71	0,2389	-1,11	0,1335	-1,51	0,0655	-1,91	0,0281	0,11	0,5438
-0,32	0,3745	-0,72	0,2358	-1,12	0,1314	-1,52	0,0643	-1,92	0,0274	0,12	0,5478
-0,33	0,3707	-0,73	0,2327	-1,13	0,1292	-1,53	0,063	-1,93	0,0268	0,13	0,5517
-0,34	0,3669	-0,74	0,2296	-1,14	0,1271	-1,54	0,0618	-1,94	0,0262	0,14	0,5557
-0,35	0,3632	-0,75	0,2266	-1,15	0,1251	-1,55	0,0606	-1,95	0,0256	0,15	0,5596
-0,36	0,3594	-0,76	0,2236	-1,16	0,123	-1,56	0,0594	-1,96	0,025	0,16	0,5636
-0,37	0,3557	-0,77	0,2206	-1,17	0,121	-1,57	0,0582	-1,97	0,0244	0,17	0,5675
-0,38	0,352	-0,78	0,2177	-1,18	0,119	-1,58	0,0571	-1,98	0,0239	0,18	0,5714
-0,39	0,3483	-0,79	0,2148	-1,19	0,117	-1,59	0,0559	-1,99	0,0233	0,19	0,5753

x	$F_0(x)$	x	$F_0(x)$	x	$F_0(x)$	x	$F_0(x)$	x	$F_0(x)$
0,2	0,5793	0,6	0,7257	1	0,8413	1,4	0,9192	1,8	0,9641
0,21	0,5832	0,61	0,7291	1,01	0,8438	1,41	0,9207	1,81	0,9649
0,22	0,5871	0,62	0,7324	1,02	0,8461	1,42	0,9222	1,82	0,9656
0,23	0,591	0,63	0,7357	1,03	0,8485	1,43	0,9236	1,83	0,9664
0,24	0,5948	0,64	0,7389	1,04	0,8508	1,44	0,9251	1,84	0,9671
0,25	0,5987	0,65	0,7422	1,05	0,8531	1,45	0,9265	1,85	0,9678
0,26	0,6026	0,66	0,7454	1,06	0,8554	1,46	0,9279	1,86	0,9686
0,27	0,6064	0,67	0,7486	1,07	0,8577	1,47	0,9292	1,87	0,9693
0,28	0,6103	0,68	0,7517	1,08	0,8599	1,48	0,9306	1,88	0,9699
0,29	0,6141	0,69	0,7549	1,09	0,8621	1,49	0,9319	1,89	0,9706
0,3	0,6179	0,7	0,758	1,1	0,8643	1,5	0,9332	1,9	0,9713
0,31	0,6217	0,71	0,7611	1,11	0,8665	1,51	0,9345	1,91	0,9719
0,32	0,6255	0,72	0,7642	1,12	0,8686	1,52	0,9357	1,92	0,9726
0,33	0,6293	0,73	0,7673	1,13	0,8708	1,53	0,937	1,93	0,9732
0,34	0,6331	0,74	0,7704	1,14	0,8729	1,54	0,9382	1,94	0,9738
0,35	0,6368	0,75	0,7734	1,15	0,8749	1,55	0,9394	1,95	0,9744
0,36	0,6406	0,76	0,7764	1,16	0,877	1,56	0,9406	1,96	0,975
0,37	0,6443	0,77	0,7794	1,17	0,879	1,57	0,9418	1,97	0,9756
0,38	0,648	0,78	0,7823	1,18	0,881	1,58	0,9429	1,98	0,9761
0,39	0,6517	0,79	0,7852	1,19	0,883	1,59	0,9441	1,99	0,9767
0,4	0,6554	0,8	0,7881	1,2	0,8849	1,6	0,9452	2	0,9772
0,41	0,6591	0,81	0,791	1,21	0,8869	1,61	0,9463	2,1	0,9821
0,42	0,6628	0,82	0,7939	1,22	0,8888	1,62	0,9474	2,2	0,9861
0,43	0,6664	0,83	0,7967	1,23	0,8907	1,63	0,9484	2,3	0,9893
0,44	0,67	0,84	0,7995	1,24	0,8925	1,64	0,9495	2,4	0,9918
0,45	0,6736	0,85	0,8023	1,25	0,8944	1,65	0,9505	2,5	0,9938
0,46	0,6772	0,86	0,8051	1,26	0,8962	1,66	0,9515	2,6	0,9953
0,47	0,6808	0,87	0,8078	1,27	0,898	1,67	0,9525	2,7	0,9965
0,48	0,6844	0,88	0,8106	1,28	0,8997	1,68	0,9535	2,8	0,9974
0,49	0,6879	0,89	0,8133	1,29	0,9015	1,69	0,9545	2,9	0,9981
0,5	0,6915	0,9	0,8159	1,3	0,9032	1,7	0,9554	3	0,9987
0,51	0,695	0,91	0,8186	1,31	0,9049	1,71	0,9564	3,1	0,999
0,52	0,6985	0,92	0,8212	1,32	0,9066	1,72	0,9573	3,2	0,9993
0,53	0,7019	0,93	0,8238	1,33	0,9082	1,73	0,9582	3,3	0,9995
0,54	0,7054	0,94	0,8264	1,34	0,9099	1,74	0,9591	3,4	0,9997
0,55	0,7088	0,95	0,8289	1,35	0,9115	1,75	0,9599	3,5	0,9998
0,56	0,7123	0,96	0,8315	1,36	0,9131	1,76	0,9608	3,6	0,9998
0,57	0,7157	0,97	0,834	1,37	0,9147	1,77	0,9616	3,7	0,9999
0,58	0,719	0,98	0,8365	1,38	0,9162	1,78	0,9625	3,8	0,9999
0,59	0,7224	0,99	0,8389	1,39	0,9177	1,79	0,9633	3,9	1

Додаток Б

Критичні точки розподілу χ^2

Число ступенів вільності *k*	**Рівень значущості α**					
	0,010	**0,025**	**0,050**	**0,950**	**0,975**	**0,980**
1	6,6	5,0	3,8	0,0039	0,00098	0,00016
2	9,2	7,4	6,0	0,103	0,051	0,020
3	11,3	9,4	7,8	0,352	0,216	0,115
4	13,3	11,1	9,5	0,711	0,484	0,297
5	15,1	12,8	11,1	1,15	0,831	0,554
6	16,8	14,4	12,6	1,64	1,24	0,872
7	18,5	16,0	14,1	2,17	1,69	1,24
8	20,1	17,5	15,5	2,73	2,18	1,65
9	21,7	19,0	16,9	3,33	2,70	2,09
10	23,2	20,5	18,3	3,94	3,25	2,56
11	24,7	21,9	19,7	4,57	3,82	3,05
12	26,2	23,3	21,0	5,23	4,40	3,57
13	27,7	24,7	22,4	5,89	5,01	4,11
14	29,1	26,1	23,7	6,57	5,63	4,66
15	30,6	27,5	25,0	7,26	6,26	5,23
16	32,0	28,8	26,3	7,96	6,91	5,81
17	33,4	30,2	27,6	8,67	7,56	6,41
18	34,8	31,5	28,9	9,39	8,23	7,01
19	36,2	32,9	30,1	10,1	8,91	7,63
20	37,6	34,2	31,4	10,9	9,39	8,26
21	38,9	35,5	32,7	11,6	10,3	8,90
22	40,3	36,8	33,9	12,3	11,0	9,54
23	41,6	38,1	35,2	13,1	11,7	10,2
24	43,0	39,4	36,4	13,8	12,4	10,9
25	44,3	40,6	37,7	14,6	13,1	11,5
26	45,6	41,9	38,9	15,4	13,8	12,2
27	47,0	43,2	40,1	16,2	14,6	12,9
28	48,3	44,5	41,3	16,9	15,3	13,6
29	49,6	45,7	42,6	17,7	16,0	14,3
30	50,9	47,0	43,8	18,5	16,8	15,0

Додаток В

Критичні значення критерію Стьюдента для об’єму вибірки *N* та рівня значущості *α*

N	Двустороння критична область				
	α = 0,1	**α = 0,05**	**α = 0,02**	**α = 0,01**	**α = 0,001**
1	6,31	12,71	31,82	63,66	636,62
2	2,92	4,30	6,97	9,93	31,60
3	2,35	3,18	4,54	5,84	12,94
4	2,13	2,78	3,75	4,60	8,61
5	2,02	2,57	3,37	4,03	6,86
6	1,94	2,45	3,14	3,71	5,96
7	1,90	2,37	3,00	3,50	5,41
8	1,86	2,31	2,90	3,36	5,04
9	1,83	2,26	2,82	3,25	4,78
10	1,81	2,23	2,76	3,17	4,59
11	1,80	2,20	2,72	3,11	4,44
12	1,78	2,18	2,68	3,06	4,32
13	1,77	2,16	2,65	3,01	4,22
14	1,76	2,15	2,62	2,98	4,14
15	1,75	2,13	2,60	2,95	4,07
16	1,75	2,12	2,58	2,92	4,02
17	1,74	2,11	2,57	2,90	3,97
18	1,73	2,10	2,55	2,88	3,92
19	1,73	2,09	2,54	2,86	3,88
20	1,73	2,09	2,53	2,85	3,85
21	1,72	2,08	2,52	2,83	3,82
22	1,72	2,07	2,51	2,82	3,79
23	1,71	2,07	2,50	2,81	3,77
24	1,71	2,06	2,49	2,80	3,75
25	1,71	2,06	2,49	2,79	3,73
26	1,71	2,06	2,48	2,78	3,71
27	1,70	2,05	2,47	2,77	3,69
28	1,70	2,05	2,47	2,76	3,67
29	1,70	2,05	2,46	2,76	3,66
30	1,70	2,04	2,46	2,75	3,65
40	1,68	2,02	2,42	2,70	3,55
60	1,67	2,00	2,39	2,66	3,46
120	1,66	1,98	2,36	2,62	3,37
∞	1,65	1,96	2,33	2,58	3,29
N	**α = 0,05**	**α = 0,025**	**α = 0,01**	**α = 0,005**	**α = 0,0005**
	Одностороння критична область				

Додаток Г

Критичні значення статистики *F*-критерію Фішера для 5%-го рівня значущості

f_2 \ f_1	1	2	3	4	5	6	12	24	∞
1	164,4	199,5	215,7	224,6	230,2	234,0	244,9	249,0	254,3
2	18,5	19,2	19,3	19,3	19,3	19,3	19,4	19,5	19,5
3	10,1	9,6	9,3	9,1	9,0	8,9	8,7	8,6	8,5
4	7,7	6,9	6,6	6,4	6,3	6,2	5,9	5,8	5,6
5	6,6	5,8	5,4	5,2	5,1	5,0	4,7	4,5	4,4
6	6,0	5,1	4,8	4,5	4,4	4,3	4,0	3,8	3,7
7	5,6	4,7	4,4	4,1	4,0	3,9	3,6	3,4	3,2
8	5,3	4,5	4,1	3,8	3,7	3,6	3,3	3,1	2,9
9	5,1	4,3	3,9	3,6	3,5	3,4	3,1	2,9	2,7
10	2,0	4,1	3,7	3,5	3,3	3,2	2,9	2,7	2,5
11	4,8	4,0	3,6	3,4	3,2	3,1	2,8	2,6	2,4
12	4,8	3,9	3,5	3,3	3,1	3,0	2,7	2,5	2,3
13	4,7	3,8	3,4	3,2	3,0	2,9	2,6	2,4	2,2
14	4,6	3,7	3,3	3,1	3,0	2,9	2,5	2,3	2,1
15	4,5	3,7	3,3	3,1	2,9	2,8	2,5	2,3	2,1
16	4,5	3,6	3,2	3,0	2,9	2,7	2,4	2,2	2,0
17	4,5	3,6	3,2	3,0	2,8	2,7	2,4	2,2	2,0
18	4,4	3,6	3,2	2,9	2,8	2,7	2,3	2,1	1,9
19	4,4	3,5	3,1	2,9	2,7	2,6	2,3	2,1	1,9
20	4,4	3,5	3,1	2,9	2,7	2,6	2,3	2,1	1,8
22	4,4	3,4	3,0	2,8	2,7	2,6	2,2	2,0	1,8
24	4,3	3,4	3,0	2,8	2,6	2,5	2,2	2,0	1,7
26	4,2	3,4	3,0	2,7	2,6	2,5	2,2	2,0	1,7
28	4,2	3,3	3,0	2,7	2,6	2,4	2,1	1,9	1,7
30	4,2	3,3	2,9	2,7	2,5	2,4	2,1	1,9	1,6
40	4,1	3,2	2,9	2,6	2,5	2,3	2,0	1,8	1,5
60	4,0	3,2	2,8	2,5	2,4	2,3	1,9	1,7	1,4
120	3,9	3,1	2,7	2,5	2,3	2,2	1,8	1,6	1,3
∞	3,8	3,0	2,6	2,4	2,2	2,1	1,8	1,5	1,0

$f_1 = n_1 - 1$ – число ступенів вільності для більшої дисперсії;

$f_2 = n_2 - 1$ – число ступенів вільності для меншої дисперсії.

Додаток Д

Критичні значення статистики *X*-критерію Ван-дер-Вардена для односторонніх меж

N	α = 0,025			α = 0,001			α = 0,1		
	m = 0 m = 1	m = 2 m = 3	m = 4 m = 5	m = 0 m = 1	m = 2 m = 3	m = 4 m = 5	m = 0 m = 1	m = 2 m = 3	m = 4 m = 5
2	–	–	–	–	–	–	0,10	–	–
3	–	–	–	–	–	–	0,50	–	–
4	–	–	–	–	–	–	0,73	0,64	–
5	–	–	–	–	–	–	0,90	0,74	–
6	–	–	–	–	–	–	1,10	1,04	0,82
7	–	–	–	–	–	–	1,25	1,14	0,89
8	2,40	2,30	–	1,42	1,37	1,23	1,42	1,37	1,23
9	2,38	2,20	–	1,56	1,48	1,30	1,56	1,48	1,30
10	2,60	2,49	2,30	1,71	1,67	1,57	1,71	1,67	1,57
11	2,72	2,58	2,40	1,83	1,77	1,64	1,83	1,77	1,64
12	2,86	2,79	2,68	1,98	1,94	1,87	1,98	1,94	1,87
13	2,96	2,91	2,78	2,09	2,03	1,93	2,09	2,03	1,93
14	3,11	3,06	3,00	2,22	2,19	2,12	2,22	2,19	2,12
15	3,24	3,19	3,06	2,33	2,28	2,20	2,33	2,28	2,20
16	3,39	3,36	3,28	2,44	2,42	2,36	2,44	2,42	2,36
17	3,49	3,44	3,36	2,54	2,51	2,44	2,54	2,51	2,44
18	3,63	3,60	3,53	2,65	2,64	2,59	2,65	2,64	2,59
19	3,73	3,69	3,61	2,76	2,72	2,66	2,76	2,72	2,66
20	3,86	3,84	3,78	2,85	2,84	2,80	2,85	2,84	2,80
21	3,96	3,92	3,85	2,95	2,92	2,87	2,95	2,92	2,87
22	4,08	4,06	4,01	3,05	3,04	3,00	3,05	3,04	3,00
23	4,18	4,15	4,08	3,14	3,12	3,06	3,14	3,12	3,06
24	4,29	4,27	4,23	3,23	3,22	3,19	3,23	3,22	3,19
25	4,39	4,36	4,30	3,33	3,29	3,26	3,33	3,29	3,26
26	4,50	4,48	4,44	3,41	3,39	3,37	3,41	3,39	3,37
27	4,59	4,56	4,51	3,49	3,47	3,43	3,49	3,47	3,43
28	4,69	4,68	4,64	3,57	3,57	3,54	3,57	3,57	3,54
29	4,78	4,76	4,72	3,66	3,64	3,60	3,66	3,64	3,60
30	4,88	4,87	4,84	3,74	3,73	3,70	3,74	3,73	3,70
31	4,97	4,95	4,91	3,82	3,80	3,76	3,82	3,80	3,76
32	5,07	5,06	5,03	3,89	3,88	3,86	3,88	3,89	3,86
33	5,15	5,13	5,10	3,96	3,95	3,92	3,96	3,95	3,92
34	5,25	5,24	5,21	4,05	4,05	4,02	4,05	4,05	4,02
35	5,33	5,31	5,28	4,12	4,11	4,08	4,12	4,11	4,08
36	5,42	5,41	5,38	4,19	4,19	4,16	4,19	4,19	4,16
37	5,50	5,48	5,45	4,26	4,25	4,24	4,26	4,25	4,24
38	5,59	5,58	5,55	4,33	4,33	4,32	4,33	4,33	4,32
39	5,67	5,65	5,62	4,40	4,39	4,38	4,40	4,33	4,38
40	5,75	5,74	5,72	4,48	4,48	4,46	4,48	4,39	4,46
41	5,83	5,81	5,79	4,54	4,53	4,50	4,45	4,53	4,50
42	5,91	5,90	5,88	4,62	4,62	4,59	4,62	4,62	4,59
43	5,99	5,97	5,95	4,68	4,67	4,66	4,68	4,67	4,66
44	6,06	6,06	6,04	4,76	4,74	4,73	4,76	4,74	4,73
45	6,14	6,12	6,10	4,81	4,80	4,78	4,81	4,80	4,78
46	6,21	6,21	6,19	4,88	4,86	4,86	4,88	4,86	4,86
47	6,29	6,27	6,25	4,93	4,93	4,90	4,93	4,93	4,90
48	6,36	6,35	6,34	5,00	5,00	4,99	5,00	5,00	4,99
49	6,43	6,42	6,39	5,07	5,05	5,04	5,07	5,05	5,04
50	6,50	6,50	6,48	5,14	5,13	5,11	5,14	5,13	5,11

Додаток Е

Критичні значення статистики $\overline{R}$-критерію рівномірності Релея

n	**α**				
	0,1	**0,05**	**0,025**	**0,01**	**0,001**
5	0,677	0,754	0,816	0,879	0,991
6	0,618	0,69	0,753	0,825	0,94
7	0,572	0,642	0,702	0,771	0,891
8	0,535	0,602	0,66	0,725	0,847
9	0,504	0,569	0,624	0,687	0,808
10	0,478	0,54	0,594	0,655	0,775
11	0,456	0,516	0,567	0,627	0,743
12	0,437	0,494	0,544	0,602	0,716
13	0,42	0,475	0,524	0,58	0,692
14	0,405	0,458	0,505	0,56	0,669
15	0,391	0,443	0,489	0,542	0,649
16	0,379	0,429	0,474	0,525	0,63
17	0,367	0,417	0,46	0,51	0,613
18	0,357	0,405	0,447	0,496	0,597
19	0,348	0,394	0,436	0,484	0,583
20	0,339	0,385	0,425	0,472	0,569
21	0,331	0,375	0,415	0,461	0,556
22	0,323	0,367	0,405	0,451	0,544
23	0,316	0,359	0,397	0,441	0,533
24	0,309	0,351	0,389	0,432	0,522
25	0,303	0,344	0,381	0,423	0,512
30	0,277	0,315	0,348	0,387	0,47
35	0,256	0,292	0,323	0,359	0,436
40	0,24	0,273	0,302	0,336	0,409
45	0,226	0,257	0,285	0,318	0,386
50	0,214	0,244	0,27	0,301	0,367
100	0,15	0,17	0,19	0,21	0,26
$2n\overline{R} \cong 100\chi_2^2$	4,605	5,991	7,378	9,21	13,816

Додаток Ж

Процентні точки критерію Граббса – Смирнова

	τ_α		
N	α = 0,1	α = 0,05	α = 0,01
5	1,84	2,08	2,57
10	2,20	2,44	2,93
15	2,38	2,62	3,10
20	2,50	2,73	3,21
25	2,59	2,82	3,28
40	2,70	3,02	3,48
50	2,86	3,08	3,54
100	3,08	3,29	3,72
250	3,34	3,53	3,95
500	3,53	3,70	4,11

Додаток И

Критичні значення статистик W1 і 2MW критерію Вілкоксона для односторонніх меж

N_1	N_2	W_1		$2MW$	N_1	N_2	W_1		$2MW$
		$\alpha = 0{,}025$	$\alpha = 0{,}10$				$\alpha = 0{,}025$	$\alpha = 0{,}10$	
1	2	3	4	5	6	7	8	9	10
10	10	78	87	210	11	11	96	106	253
	11	81	91	220		12	99	110	264
	12	84	94	230		13	103	114	275
	13	88	98	240		14	106	118	286
	14	91	102	250		15	110	123	297
	15	94	106	260		16	113	127	308
	16	97	109	270		17	117	131	319
	17	100	113	280		18	121	135	330
	18	103	117	290		19	124	139	341
	19	107	121	300		20	128	144	352
	20	110	125	310		21	131	148	363
	21	113	128	320		22	135	152	374
	22	116	132	330		23	139	156	385
	23	119	136	340		24	142	161	396
	24	122	140	350		25	146	165	407
	25	126	144	360					
					13	13	136	149	351
12	12	115	127	300		14	141	154	364
	13	119	131	312		15	145	159	377
	14	123	136	314		16	150	165	390
	15	127	141	336		17	154	170	403
	16	131	145	348		18	158	175	416
	17	135	150	360		19	163	180	429
	18	139	155	372		20	167	185	442
	19	143	159	384		21	171	190	455
	20	147	164	396		22	176	195	468
	21	151	169	400		23	180	200	481
	22	155	173	420		24	185	205	494
	23	159	178	432		25	189	211	507
	24	163	183	444					
	25	167	187	456					

1	2	3	4	5	6	7	8	9	10
					15	15	184	200	465
						16	190	206	480
14	14	160	174	406		17	195	212	495
	15	164	179	420		18	200	218	510
	16	169	185	434		19	205	224	525
	17	174	190	448		20	210	230	540
	18	179	196	462		21	216	236	555
	19	183	202	476		22	221	242	570
	20	188	207	490		23	226	248	585
	21	193	213	504		24	231	254	600
	22	198	218	518		25	237	260	615
	23	203	224	532					
	24	207	229	543					
	25	212	235	560					
16	16	211	229	528	17	17	240	259	595
	17	217	235	544		18	246	266	612
	18	222	242	560		19	252	273	629
	19	228	248	576		20	258	280	646
	20	234	255	592		21	264	287	663
	21	239	261	608		22	270	294	680
	22	245	267	624		23	276	300	697
	23	251	274	640		24	282	307	714
	24	256	280	656		25	288	314	731
	25	262	287	672					
					19	19	303	325	741
18	18	270	291	666		20	309	333	760
	19	277	299	684		21	316	341	779
	20	283	306	702		22	323	349	798
	21	290	313	720		23	330	357	817
	22	296	321	738		24	337	364	836
	23	303	328	756		25	344	372	855
	24	309	335	774					
	25	316	343	792	21	21	373	399	903
						22	381	408	924
20	20	337	361	820		23	388	417	945
	21	344	370	840		24	396	425	966
	22	351	378	860		25	404	434	987
	23	359	386	880					
	24	366	394	900	23	23	451	481	1081
	25	373	403	920		24	459	491	1104
						25	468	500	1127
22	22	411	439	990					
	23	419	448	1012	24	24	492	525	1176
	24	427	457	1034		25	501	535	1200
	25	435	467	1056	25	25	536	570	1275

Навчальне видання

Рудаков Дмитро Вікторович
Сдвижкова Олена Олександрівна

МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ ПРИРОДНИЧИХ СИСТЕМ

Навчальний посібник

Редактор Ю.В. Рачковська

Підписано до друку 31.01.2022. Формат 30×42/4
Папір офсетний. Ризографія. Ум. друк. арк. 9,9.
Обл.-вид. арк. 9,9. Тираж 75 пр. Зам. №

Підготовлено до друку та видрукувано
у Національному технічному університеті «Дніпровська політехніка».
Свідоцтво про внесення до Державного реєстру ДК № 1842 від 11.06.2004.

49005, м. Дніпро, просп. Д. Яворницького, 19