Міністерство освіти і науки України
Волинський національний університет імені Лесі Українки

Супрунович С. В., Кормош Ж. О., Сливка Н. Ю.

Статистичні та хемометричні методи в хімії

Навчальний посібник

Луцьк — 2022

УДК 54:519.237-7(07)
С 89

Рекомендовано до друку вченою радою Волинського національного університету імені Лесі Українки як навчальний посібник для студентів вищих навчальних закладів (протокол № 7 від 26 травня 2022 р.)

Рецензенти:

Федосов Сергій Анатолієвич, доктор фізико-математичних наук, професор кафедри теоретичної та комп'ютерної фізики імені А. В. Свідзинського ВНУ імені Лесі Українки.
Бохан Юлія Володимирівна, кандидат хімічних наук, доцент кафедри природничих наук та методик їхнього навчання Центральноукраїнського державного педагогічного університету імені Володимира Винниченка.
Мілюкін Михайло Васильович, доктор хімічних наук, старший науковий співробітник Інституту колоїдної хімії та хімії води імені А. В. Думанського НАН України, заступник директора з наукової роботи.

С 89
Супрунович С. В., Кормош Ж. О., Сливка Н. Ю.
Статистичні та хемометричні методи в хімії : Навчальний посібник для студентів вищих навчальних закладів. Луцьк: ВНУ імені Лесі Українки, 2022. 210 с.

Навчальний посібник присвячений прикладній науковій дисципліні, що виникла на стику хімії та математики. Викладені основи статистики, планування досліджень, оптимізації експеримента. Розглянуто базові методи аналізу даних – дисперсійний, регресійний, кластерний аналізи. Висвітлено найбільш поширені методи візуалізації даних.
Для широкого кола спеціалістів з хімії та споріднених дисциплін. Корисна також студентам, вчителям та викладачам.

УДК 54:519.237-7(07)

Зміст

Передмова

Курс ”Статистичні та хемометричні методи в хімії” з’явився під назвою „Хемометрика” на хімічному факультеті ВНУ імені Лесі Українки з ініціативи Світлани Теодорівни Олексеюк, ключового на той час спеціаліста з аналітичної хімії. В той час була думка запровадити американську систему вибору предметів для вивчення, хемометрика була включена в список доступних предметів, і студенти вибрали цей предмет. Так курс «Хемометрика» став викладатись на хімічному факультеті нашого університету.

Курс „Хемометрика” у системі підготовки спеціалістів-хіміків посідає особливе місце. Він є містком між хімією та математикою. Вивчення цього курсу студентами-хіміками сприятиме становленню цілісної картини хімічної науки та формуванню системного підходу при розгляді хімічних процесів.

Пропонований посібник є основною частиною курсу „Хемометрика”, котрий упродовж багатьох років читають викладачі хімічного факультету Волинського національного університету імені Лесі Українки.

У виданні значне місце відведено розгляду методів первинної обробки експериментальних даних інформації, котрі є однією з ланок наукового процесу будь-якої дисципліни.

Задля глибшого засвоєння теоретичного матеріалу в кінці посібника наведені приклади застосування викладених підходів для вирішення завдань хімічної практики.

В кінці кожного розділу наведені питання та вправи для самопі-

дготовки. Для виконання деяких з них необхідно проведення розрахунків. Для цього рекомендується використовувати R-статистику .

Так як наш підручник орієнтований, в першу чергу, на розгляд теоретичних аспектів методів статистичної та хемометричної обробки експериментальних даних, то на практичній реалізації тих чи інших методів ми не будемо загострювати увагу. R-статистика – вільне програмне забезпечення, легко доступне для встановлення на більшості поширених операційних систем для персональних комп'ютерів. Тому ми вибрали саме цю програму для ілюстрації роботи різних методів статистичної обробки.

За необхідності в практичних завданнях будуть наводитись функції R-статистики, необхідні для їх виконання. Так як R-статистика придатна для виконання більшості хемометричних завдань, то ми рекомендуємо ознайомитись з цією програмою більш детально. Це можна зробити за допомогою відповідних посібників або дистанційних курсів . Крім того, для студентів факультету хімії, екології та фармації доступний практикум з статистичних та хемометричних методів аналізу та дистанційний курс Moodle .

Автори висловлюють щиру подяку окремим колегам за цінні поради та рекомендації, які були взяті до уваги при написанні цього посібника.

Вступ

Мета дисципліни

У своїй повсякденній роботі хімік неодмінно стикається з необхідністю обробки інформації, одержаної за допомогою приладів, від найпростіших – годинника чи термометра до досить складних, в яких застосовуються для одержання інформації найсучасніші фізичні та математичні методи — Фур'є-спектрометри. З іншої сторони, деколи необхідно підвищити ефективність роботи, мінімізувати кількість досліджень, і в той же час провести їх якісно. Тут на допомогу приходить математика.

Хемометрика — дисципліна, яка існує на стику хімії та математики. Вона об'єднує два напрями — від хімії до математики, де використовується математичний апарат для аналізу структури та обробки даних, та від математики до хімії, де математика допомагає спланувати дослідження. У зв'язку з бурхливим розвитком комп'ютерної техніки стало можливим значно зменшити час, витрачений на рутинні обрахункові операції та зосередити більше уваги на методах розрахунків і теоретичних основах тих чи інших операцій. Це дало змогу об'єднати разом усі розрізнені випадки використання математичного апарату.

Головна увага при викладанні дисципліни приділяється вивченню методів математичної статистики, які можуть знайти своє застосування в хімії. Розглядаються шляхи перетворення інформації в сучасних приладах. Лабораторний практикум дозволяє засвоїти основні методи аналізу та обробки даних.

Вивчення хемометрики повинно сприяти одержанню глибших знань про оточуючий світ, і зокрема, дозволяє на вищому рівні вирішувати проблеми, пов'язані із розвитком науки. При викладанні дисципліни постійно підкреслюється конкретний зв'язок питань, що розглядаються за програмою курсу, з питаннями, які можуть зустрітися в майбутній практичній роботі спеціаліста-хіміка. Завдання дисципліни Основними завданнями курсу є:

— сприяти розвиткові у студентів логічного мислення і діалектичного світогляду;

— добитися ґрунтовного засвоєння студентами основних методів аналізу та обробки даних;

Місце дисципліни в учбовому процесі

Курс статистичних та хемометричних методів у хімії читається для студентів факультету хімії, екології та фармації денної форми навчання спеціальностей 102 Хімія та 014 Середня Освіта (Хімія). Дисципліна читається в п'ятому семестрі після того, як студенти прослухали такі дисципліни як інформатика, теорія ймовірності, вища математика. Курс хемометрики розрахований на один семестр. Цей курс є базою для вивчення фундаментальних та спеціальних професійно орієнтованих дисциплін, таких як аналітична хімія, фізична хімія, комп'ютерна хімія та інших.

Предмет хемометрики

З давніх часів хімія була тісно пов'язана з математикою. Речовини зважувались, вимірювались об'єми рідин та газів, в прописах необхідно було більш-менш точно вказувати кількості використовуваних компонентів. Звичайно, для оперування кількостями необхідно було проводити певні математичні розрахунки. Вони спочатку були простими, включали прості арифметичні операції, але з розвитком хімії ці розрахунки ставали все складнішими й складнішими. Нарешті математика в хімії стала настільки трудомісткою та складною що постала необхідність створення окремої дисципліни, котра б оперувала

математичними аспектами хімічних дисциплін. Так з'явилась хемометрика.

> Хемометрика — наука про методи одержання важливої хімічної інформації, її організацію та представлення

Ця наука знаходиться на межі хімії та математики й включає два шляхи взаємодії цих наук: хімія → математика — сюди входять різноманітні методи обробки даних; та математика → хімія — сюди входять різні методи організації хімічних досліджень.

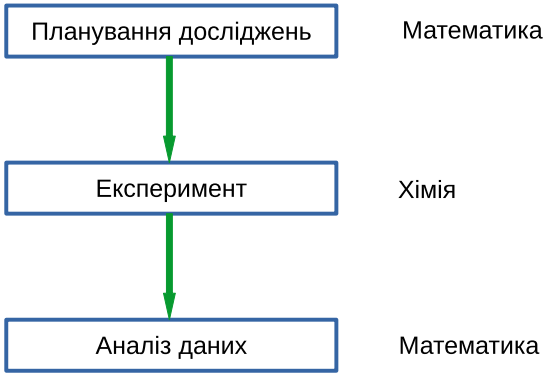


Рис. 1: Алгоритм проведення досліджень в хімії

Історія хемометрики

Початком розвитку хемометрики, певно, треба вважати 1975 р., коли К. Гаус (K. Gauss) почав використовувати метод найменших квадратів. Першим хемометриком — Уільям Госета (W. Gosset), відомого під псевдонімом Стьюдент, він в кінці 19 століття використовував методи аналізу даних працюючи аналітиком на пивоварні Гіннеса.

На початку 20 ст. з'явилась робота Карла Пірсона (K. Pearson), де був запропонований метод основних компонент. Пізніше Рональд

Фішер (R. Fisher) увів у користування ряд статистичних методів, серед яких слід відзначити метод максимальної правдоподібності та факторний аналіз. Його роботи заклали основи методів планування досліджень.

Хемометрика довгий час була тісно пов'язана з аналітичною хімією та розвивалась разом з нею. Однак з часом вона вийшла за рамки суто аналітичної хімії й поширилась на інші хімічні дисципліни. Ускладнення та узагальнення математичного апарату хемометрики дозволило його використовувати в галузях, досить далеких від аналітичної хімії.

Найбільший поштовх розвитку хемометрики дав розвиток обчислювальної техніки, котра дозволила практично використовувати методи, раніше недосяжні із-за надмірної кількості різноманітних обчислень.

Датою народження хемометрики як окремої дисципліни вважають 15 серпня 1974 року, котра зародилась внаслідок співпраці американського хіміка Брюса Ковальські та шведського викладача Сванте Волда, внука всесвітньо відомого шведського фізхіміка Сванте Ареніуса.

Хемометрика в інших науках

Зараз хемометрика використовується в самих різноманітних галузях науки, як близьких, так і досить далеких від хімії. В фізичній хімії вона використовується при дослідженнях кінетики, в органічній хімії – для передбачення властивостей різних сполук в залежності від структури (QSAR), в хімії полімерів – для дослідження антиоксидантних властивостей. В квантовій хімії широко використовуються різноманітні методи оптимізації. В структурній та теоретичній хімії важливе місце займає теорія графів.

Крім хімії хемометрика знаходить своє використання в найрізноманітніших галузях науки та народного господарства. В екології вона використовується для аналізу навколишнього середовища, у виробництві – для контролю якості продукції.

Розвиток хемометрики в світі

Хемометрика зараз швидко розвивається. Кількість наукових публікацій зростає з року в рік, і на даний час перевищує 14 тис. в рік. (за даними наукометричної бази Scopus)

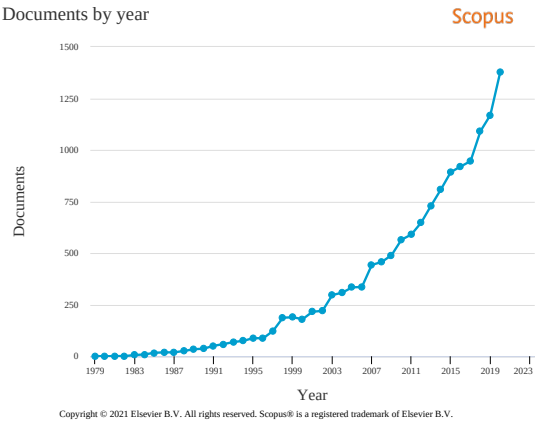


Рис. 2: Зростання кількості публікацій з хемометрики в світі

Важко уявити собі розвиток науки без проведення експериментів, отримання достовірної інформації та правильної інтерпретації даних. Аналіз даних необхідний в будь-якій області науки. Можна виділити деякі загальні принципи аналізу даних, загальні методи обробки інформації. Часто ці методи засновані на статистичних процедурах, так як експериментальні дані, як правило, включають випадкову компоненту.

При проведенні вимірювань практично неможливо врахувати і зафіксувати всі фактори, що впливають на вимірювану величину, що призводить до випадкового розкиду даних навколо істинного значення. Практична хемометрика - це комп’ютеризація, кожен крок передбачає використання програмного забезпечення. зручним засобом роботи з великими масивами чисельних даних є електронні таблиці.

Складні методи багатомірної статистики реалізовані в різних спеціалізованих програмах, наприклад таких як R-статистика, SPSS, Statistica.

Осередком вивчення хемометрики в Росії є інститут хімічної фізика РАН (ИХФ РАН). Там працює група хемометриків на чолі з Олексієм Померанцевим та Оксаною Родіоновою. Вони вивчають хемометричні методи визначення якості ліків та харчових продуктів. Потужна школа хемометрики існує в Харкові. Цю школу створив Холін.

Розділ 1

Основні положення теорії ймовірностей

Ми почнемо вивчати хемометрику з основ, а одна з основ — теорія ймовірностей. Завдання теорії ймовірностей — побудова й аналіз ймовірнісних математичних моделей реальних явищ, які враховують випадкові аспекти цих явищ. Вихідне поняття цієї теорії — поняття випадкового досліду.

> Випадковий (стохастичний) дослід — подія, що приводить до результату який неможливо однозначно передбачити знаючи всі умови випробування.

Звичайно випадковий дослід можна повторювати багато разів, але результат цього досліду кожен раз буде іншим. Як приклад випадкового досліду можна навести підкидання монети, або грального кубика. Практично всі події, що відбуваються в світі, є випадковими. Кількісною характеристикою цих подій є випадкова величина.

> Випадкова величина – це величина, яка приймає в результаті досліду певне конкретне значення з множини можли-

вих випадків, причому появу цього значення неможливо точно передбачити до його вимірювання.

Приклад 1 (Монета)
Монета, падаючи, може впасти на одну з двох сторін — орел чи решку.

Приклад 2 (Гральний кубик)
У грального кубика шість сторін. Коли він впаде, верхня грань буде мати від однієї до шести крапок, котрі відповідають числам 1, 2, 3, 4, 5, 6.

Випадковий дослід (випробування) можна проводити багато разів, але результат кожен раз буде різним. Аналогічна ситуація й в хімії, хоча й менш наочна — коли ми будемо проводити вимірювання певної величини (об'єму при титруванні, маси при зважуванні та ін.), результат кожен раз буде іншим. Якщо ми будемо проводити випробування до безкінечності, то одержимо всі можливі результати, які становлять так звану «генеральну сукупність».

Генеральна сукупність — сукупність усіх можливих результатів які можуть бути одержані за однакових умов у конкретному експерименті.

Усі можливі значення в генеральній сукупності входять в так званий простір елементарних подій.

Простір елементарних подій — множина всіх можливих наслідків випадкового (стохастичного) експерименту.

Приклади:

Приклад 3 (Монета)
У випадку монети простір елементарних подій буде включати два результати — орел, решка.

Приклад 4 (Кубик)
У цьому випадку простір елементарних подій включає шість результатів — 1, 2, 3, 4, 5, 6.

Генеральна сукупність може містити обмежену кількість результатів, як у випадку монети чи кубика, або містити необмежену кількість результатів і бути обмеженою по краях, або бути взагалі необмеженою. Відповідно до цього розрізняють дискретні та безперервні випадкові величини.

> Дискретною випадковою величиною називається така величина, число можливих значень якої або обмежене, або є нескінченою зліченою множиною.

> Неперервною випадковою величиною називається така величина, можливі значення якої неперервно заповнюють деякий інтервал числової осі (кінечний або нескінчений).

Приклад 5 (Титрування)
Простір елементарних подій обмежується об'ємом бюретки. Якщо бюретка на 50 мл, то простір елементарних подій включає всі значення, що знаходяться на відрізку $[0; 50]$.

Приклад 6 (Абсолютна температура)
Абсолютна температура не може бути нижчою 0 К, тому простір елементарних подій включає всі значення, що знаходяться на інтервалі $[0; +\infty]$.

Приклад 7 (Положення в просторі)
Положення в просторі по висоті, довготі чи широті може описуватись координатами, котрі можуть набувати будь-яких значень, тому простір елементарних подій включає весь діапазон можливих значень $[-\infty; +\infty]$.

Якщо проводити випробування до безкінечності, то можна буде помітити, що результати мають певну структуру. Цю структуру можна охарактеризувати так званою «*функцією розподілу*»:

$$F(x) = P(\omega : \xi(\omega) \leqslant x) \tag{1.1}$$

1.1. Функція розподілу

Функція розподілу ймовірностей повністю описує розподіл ймовірностей випадкової величини.

> Функцією розподілу $F(x)$ випадкової величини X називається ймовірність того, що вона набуде значення менше, ніж аргумент функції x: $F(x) = p\{X < x\}$.

Властивості функції розподілу

1. $F(x)$ – невід'ємна функція, котра набуває значень між нулем і одиницею:

$$0 \leqslant F(x) \leqslant 1 \tag{1.2}$$

2. $F(x)$ зростає на всій області визначення:

$$\forall \quad x_1 < x_2 \Rightarrow F(x_1) \leqslant F(x_2) \tag{1.3}$$

3. Нижня межа дорівнює нулю:

$$\lim_{x \to -\infty} F(x) = 0 \tag{1.4}$$

4. Верхня межа дорівнює одиниці:

$$\lim_{x \to +\infty} F(x) = 1 \tag{1.5}$$

5. $F(x)$ безперервна зліва:

$$F(x_0 - 0) = \lim_{x \to x_0 - 0} F(x) = F(x_0) \tag{1.6}$$

6. Різниця двох значень $F(x)$ дорівнює ймовірності знаходження випадкової величини в заданому інтервалі $[a, b]$:

$$F(b) - F(a) = P[a < \xi \leqslant b] \tag{1.7}$$

Функція розподілу неперервної випадкової величини хоча і є її повною характеристикою по ймовірності, має недолік – по ній досить важко судити про характер розподілу випадкової величини. Тому в хемометриці функція розподілу використовується дуже рідко. Більше практичне значення має її похідна – функція густини ймовірностей.

1.2. Функція густини ймовірностей

Густина розподілу ймовірностей безперервної випадкової величини необхідна для демонстрації зміни ймовірності в залежності від конкретного значення елементів множини. Густина розподілу – похідна від функції розподілу безперервної випадкової величини:

$$f(x) = \frac{dF(x)}{dx} \tag{1.8}$$

1.2.1. Властивості функції густини ймовірності

1. Функція густини ймовірності невід'ємна:

$$f(x) \geqslant 0 \tag{1.9}$$

2. Ймовірність будь-якої події дорівнює одиниці (умова нормування):

$$\int_{-\infty}^{+\infty} f(x)dx = 1 \tag{1.10}$$

3. Ймовірність набуття величиною x значення від a до b:

$$\int_{a}^{b} f(x)dx = P[a < \xi \leqslant b] \tag{1.11}$$

Функцію густини ймовірності частіше називають *розподілом випадкової величини*.

Питання та вправи для самостійного опрацювання

1. Охарактеризуйте поняття випадкової величини.
2. Наведіть приклад дискретної випадкової величини.
3. Наведіть приклад неперервної випадкової величини.
4. Охарактеризуйте властивості функції розподілу.
5. Охарактеризуйте властивості функції густини ймовірності.
6. Чому результат випадкового досліду кожен раз різний?
7. Навмання вводиться 4-х розрядний PIN-код мобільного телефона. Яким є простір елементарних подій?
8. Із 30-ти екзаменаційних теоретичних питань і 20-ти задач студент навмання вибирає 2 теоретичних питання і одну задачу. Яким є простір елементарних подій?
9. Для титрування використовується бюретка на 25 мл. Яким є простір елементарних подій?
10. Визначають pH розчину за допомогою універсального індикаторного паперу. Яким є простір елементарних подій?

1.3. Розподіли випадкових величин

Розподіл випадкової величини – будь яке співвідношення, що встановлює зв'язок між можливими значеннями випадкової величини та відповідними ймовірностями.

Кожен розподіл можна описати формулою що характеризує ймовірність настання тої чи іншої події. Не зважаючи на велику різноманітність процесів у природі, кількість типів розподілів, що зустрічаються, обмежена. Розглянемо основні види розподілів, що зустрічаються в практиці.

1.3.1. Біноміальний розподіл

Цим законом описується ймовірність вийняти з мішка, що містить білі та чорні кульки, певну кількість кульок одного кольору, та інші подібні задачі . В цьому розподілі використовується формула біному Ньютона, тому його називають біноміальним.

Проводяться незалежні випробування, в яких реалізується два випадки: «успіх» випадає з ймовірністю p, «невдача» – з ймовірністю $q = 1-p$. У цьому випадку дискретна випадкова величина ξ буде мати біноміальний розподіл. Ймовірність набуття нею конкретних значень має вигляд:

$$P(\xi = k) = C_n^k p^k q^{n-k} \tag{1.12}$$

де p і n – параметри:
C_n^k – біноміальний коефіцієнт ($C_n^k = \frac{n!}{k!(n-k)!}$);
p – ймовірність успіху, $p \in [0, 1]$;
q – ймовірність невдачі, $q = 1 - p$;
n –число випробувань, $n \in N$;
k – число успіхів, $k \in N$.

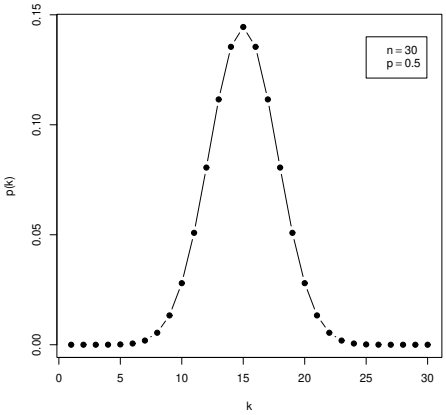


Рис. 1.1: Біноміальний розподіл

1.3.2. Геометричний розподіл

Геометричний закон розподілу зустрічається в мікробіології, генетиці, фізиці. На практиці експеримент чи дослід здійснюють до першої появи успішної події A. Тому за цим законом визначається кількість підкидань грального кубика, або кількість пострілів, щоб попасти в мішень.

Кількість проведених спроб буде цілочисельною випадковою величиною $\xi = 1, 2, \ldots, k$. Ймовірність появи події A в кожному досліді не залежить від попередніх і становить p.

$$P_k = P(X = k) = pq^{k-1}, \qquad k = 1, 2, \ldots, n \tag{1.13}$$

де $q = 1 - p$

Тобто в усіх попередніх дослідах крім k-го експеримент дає негативний результат, і лише в k-му випадку є успішним. Даний розподіл

називають геометричним, оскільки його права частина виражає величину елемента геометричної прогресії.

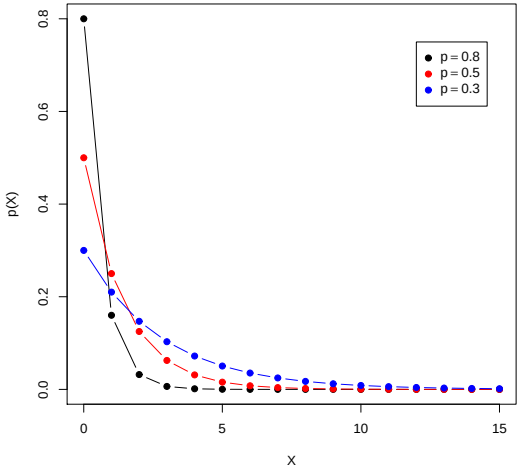


Рис. 1.2: Геометричний розподіл

1.3.3. Гіпергеометричний розподіл

Нехай в сукупності з N об'єктів містяться K об'єктів, що володіють певною ознакою. З цієї сукупності випадковим чином, і без повернення витягується n об'єктів. Тоді випадкова величина X – кількість «особливих» об'єктів k у вибірці, розподілена по гіпергеометричному закону.

$$P(X = k) = \frac{C_K^k \cdot C_{N-K}^{m-k}}{C_N^n} \tag{1.14}$$

де N – загальна кількість об'єктів у сукупності;
K – загальна кількість особливих об'єктів у сукупності;
n – загальна кількість об'єктів у вибірці;
k – кількість особливих об'єктів у вибірці;
C_K^k, C_{N-K}^{m-k}, C_N^n – біноміальні коефіцієнти ($C_n^k = \frac{n!}{k!(n-k)!}$).

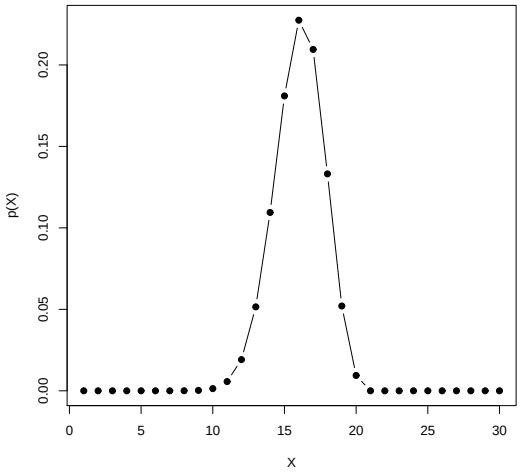


Рис. 1.3: Гіпергеометричний розподіл

1.3.4. Розподіл Пуасона

Цей розподіл названо на честь французького вченого Сімеона Дені Пуассона.

Якщо кількість випробувань n досить велика, а ймовірність p появи події A в окремо взятому випробуванні дуже мала ($p < 0.1$), то ймовірність того, що в даній серії випробувань подія A з'явиться рів-

но k раз, можна обчислити за формулою Пуасона:

$$P_k = P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \qquad k = 1, 2, \ldots, n \tag{1.15}$$

де λ – параметр.

Пуассонівський розподіл справедливий для подій, які мають малу ймовірність чи трапляються нечасто. Ним, наприклад, можна описати ймовірність того, що футболіст заб'є гол у конкретному матчі. Іноді футболіст забиває один гол, рідше два, ще рідше робить хет-трик, Пеле одного разу забив вісім. Найчастіше футболіст не забиває жодного. Ймовірність забити k голів за гру визначається параметром λ, що є середньою кількістю голів, які забиває футболіст.

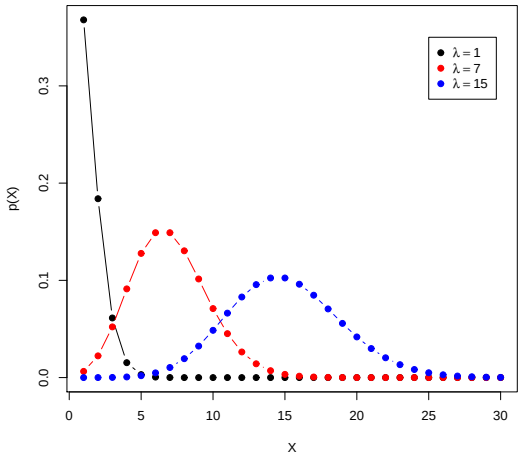


Рис. 1.4: Розподіл Пуасона для різних λ

Крім того, розподіл Пуассона знайшов широке застосування в те-

орії масового обслуговування для ймовірнісної характеристики найпростішого потоку подій.

Саме цей закон буде описувати ймовірність одержати k звінків за одиницю часу при роботі кол-центру.

1.3.5. Рівномірний розподіл

За цим законом можна оцінити, наприклад, час чекання транспорту, котрий ходить через однакові інтервали. Або похибку на прикладах, що мають циферблат і стрілку, що рухається між поділками. Рівномірний розподіл має шум в оцифрованих зображеннях звуках, відео.

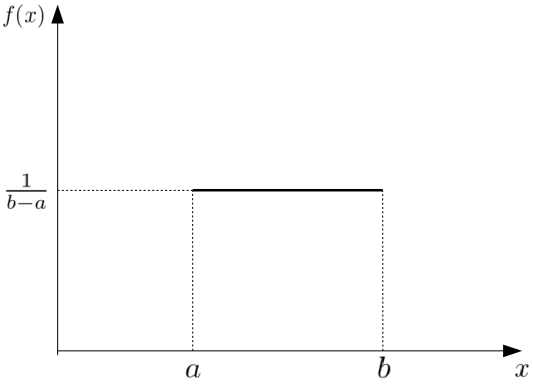


Рис. 1.5: Рівномірний розподіл

Випадкова величина X має рівномірний закон розподілу, якщо ймовірності її можливих значень однакова від експерименту до експерименту і обчислюються формулою:

$$P(x) = \begin{cases} \frac{1}{b-a}, & x \in [a,b] \\ 0, & x \notin [a,b] \end{cases} \tag{1.16}$$

1.3.6. Нормальний розподіл

Цей розподіл найчастіше зустрічається в природі. Зокрема, в хімії більшість величин мають цей розподіл.

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}, \qquad -\infty < x + \infty \tag{1.17}$$

де a, σ^2 – параметри.

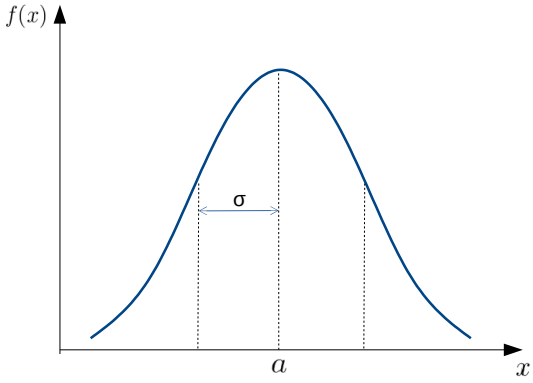


Рис. 1.6: Нормальний розподіл

1.3.7. Показниковий розподіл

Показниковий розподіл досить поширений у природі. За цим законом розподілені небесні тіла у всесвіті, доходи людей у суспільстві та ін.

$$P(x) = \begin{cases} \lambda e^{-\lambda}, & x > 0 \\ 0, & x \leqslant 0 \end{cases} \tag{1.18}$$

де λ – параметр.

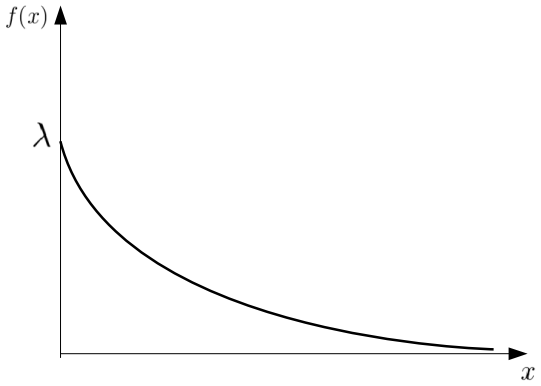


Рис. 1.7: Показниковий розподіл

Питання та вправи для самостійного опрацювання

1. Що таке розподіл випадкової величини.
2. Назвіть основні типи розподілів, що зустрічаються в природі.
3. Що характеризують параметри розподілу?
4. Охарактеризуйте властивості функції розподілу.
5. Охарактеризуйте властивості функції густини ймовірності.
6. Де зустрічається біноміальний розподіл?
7. Де зустрічається нормальний розподіл?
8. Який розподіл має довжина листя на дереві?
9. Який розподіл мають космічні тіла?
10. Охарактеризуйте властивості нормального розподілу?

1.4. Характеристики випадкових величин

Закон розподілу повністю характеризує випадкову величину з точки зору ймовірності. Однак, при вирішенні багатьох практичних задач немає необхідності характеризувати випадкову величину повністю, а достатньо тільки мати деяке загальне уявлення. В теорії ймовірності для загальної характеристики випадкової величини використовують певні величини, які носять назву числових характеристик випадкової величини. Їх призначення – у стислій формі показати найбільш суттєві особливості того чи іншого розподілу. Найбільш важливими характеристиками випадкових величин є математичне очікування, дисперсія, середньоквадратичне відхилення, коефіцієнт варіації, мода, квантилі. Для багатомірних – коваріація, коефіцієнт кореляції, коефіцієнт детермінації.

1.4.1. Математичне очікування

Математичне очікування характеризує центр тяжіння розподілу.

$$M(\xi) = \int_{-\infty}^{\infty} x f(x) dx \tag{1.19}$$

де $f(x)$ – функція густини ймовірності.

Властивості математичного очікування

1. Математичне очікування константи дорівнює самій константі:

$$M(C) = C, \qquad C = const \tag{1.20}$$

2. Математичне очікування добутку константи на випадкову величину дорівнює добутку константи на математичне очікування випадкової величини:

$$M(C\xi) = CM(\xi), \qquad C = const \tag{1.21}$$

3. Математичне очікування суми випадкових величин дорівнює сумі математичних очікувань цих випадкових величин:

$$M(\xi_1 + \xi_2 + \ldots + \xi_n) = M(\xi_1) + M(\xi_2) + \ldots + M(\xi_n) \tag{1.22}$$

4. Математичне очікування добутку випадкових величин дорівнює добутку математичних очікувань цих випадкових величин:

$$M(\xi \cdot \eta) = M(\xi) \cdot M(\eta) \tag{1.23}$$

1.4.2. Дисперсія

Дисперсія характеризує розсіювання випадкової величини.

$$D(\xi) = M[\xi - M(\xi)]^2 = \int\limits_{-\infty}^{\infty} (x - M(x))^2 f(x)dx \tag{1.24}$$

де $M(x)$ – математичне очікування;
$f(x)$ – функція густини ймовірності.

Властивості дисперсії

1. Дисперсія очікування константи дорівнює нулю:

$$D(C) = 0, \qquad C = const \tag{1.25}$$

2. Дисперсія добутку константи на випадкову величину дорівнює добутку квадрату константи на дисперсію випадкової величини:

$$D(C\xi) = C^2 D(\xi), \qquad C = const \tag{1.26}$$

3. Дисперсія суми або різниці двох незалежних випадкових величин дорівнює сумі дисперсій цих випадкових величин:

$$D(\xi + \eta) = D(\xi) + D(\eta) \tag{1.27}$$

$$D(\xi - \eta) = D(\xi) + D(\eta) \tag{1.28}$$

Дисперсія випадкової величини є досить зручною характеристикою розсіювання можливих значень випадкової величини. Проте, в неї немає наочності, так як вона має розмірність квадрата випадкової величини. Для більшої зручності треба мати характеристику, що своєю розмірністю співпадає з розмірністю випадкової величини. Такою характеристикою є середнє квадратичне відхилення.

1.4.3. Середньоквадратичне відхилення

Середньоквадратичне відхилення характеризує абсолютне розсіювання випадкової величини:

$$\sigma(\xi) = +\sqrt{D(\xi)} \tag{1.29}$$

1.4.4. Коефіцієнт варіації

Коефіцієнт варіації характеризує відносне розсіювання випадкової величини.

$$V = \frac{\sigma}{M(x)} \tag{1.30}$$

1.4.5. Квантилі

Квантиль Q_p (Рис. 1.8) – значення, яке задана випадкова величина x не перевищує з фіксованою ймовірністю p.

Деякі квантилі мають власні назви:

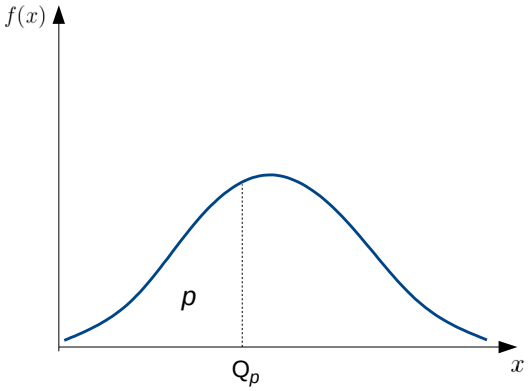


Рис. 1.8: Квантиль розподілу

- $Q_{0.25}$ – перший (нижній) квартиль, 25-й перцентиль;
- $Q_{0.5}$ – медіана;
- $Q_{0.75}$ – третій (верхій) квартиль, 75-й перцентиль;

Різниця між третім і першим квартилями $R = Q_{0.75} - Q_{0.25}$ має назву ”інтервальний розмах”.

Якщо ймовірність задана у відсотках, то квантиль називається перцентилем. Коли весь інтервал ймовірності ділять на десять частин, то відповідний квантиль називають децилем.

1.4.6. Мода

Мода (Рис. 1.9) – це таке значення випадкової величини, в якому функція густини ймовірностей набуває максимального значення.

На відміну від інших характеристик розподілів, мода може бути не одна, а кілька, в залежності від того, скільки максимумів є на графіку густини ймовірності.

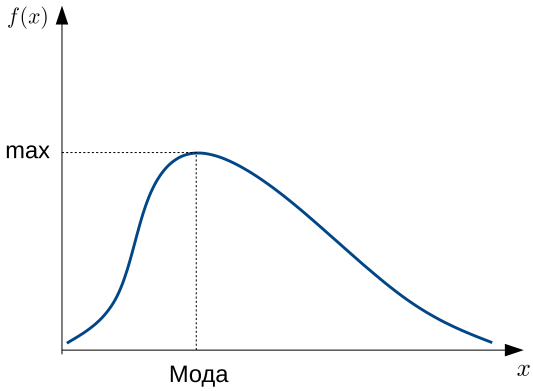


Рис. 1.9: Мода розподілу

1.5. Характеристики багатомірних випадкових величин

1.5.1. Коваріація

Коваріація – міра лінійної залежності двох випадкових величин:

$$cov(\xi, \eta) = M\left[(\xi - M(\xi))(\eta - M(\eta))\right] \tag{1.31}$$

Коваріація може набувати як додатних, так і від'ємних значень в залежності від того, як сумісно змінюються випадкові величини. Якщо зміни синхронні, то коваріація позитивна, якщо асинхронні, то негативна. Якщо коваріація близька до нуля, то величини незалежні.

1.5.2. Коефіцієнт кореляції

Коефіцієнт кореляції Пірсона – показник кореляції між двома випадковими величинами, який набуває значень від −1 до +1 включно. Він широко використовується для оцінки ступеня лінійної залежності між

двома змінними. Показник був розроблений Карлом Пірсоном (Karl Pearson) 1880-х роках.

$$r(\xi, \eta) = \frac{cov(\xi, \eta)}{\sigma_\xi \sigma_\eta} \tag{1.32}$$

Питання та вправи для самостійного опрацювання

1. Охарактеризуйте поняття випадкової величини.
2. Наведіть приклад дискретної випадкової величини.
3. Наведіть приклад неперервної випадкової величини.
4. Що таке мода розподілу.
5. Як визначається середнє значення.
6. Що характеризує дисперсія?
7. Дайте визначення квантіля?
8. Які назви мають специфічні квантилі?
9. Що таке коефіцієнт кореляції?
10. Що таке коваріація?

Розділ 2

Основи статистики

2.1. Аналіз результатів вимірювань

В хімії дані одержують в результаті вимірювань певних характеристик зразків.

> Вимірювання – експериментально визначена величина змінної, яку використовують для характеристики зразка.

Вимірювання бувають прямими (коли об'єкт безпосередньо зіставляється з носієм одиниці виміру, наприклад, вимірювання довжини лінійкою) і непрямими (коли вимірювана величина розраховується з інших виміряних величин, наприклад, вимір глибини за допомогою ехолота). В хімії більш поширені непрямі вимірювання.

> Зразок – об'єкт, над яким проводиться експеримент.

> Змінна – деяка грань (або сторона) зразка.

> Фактор – вхідна змінна, котра впливає на систему.

> Рівень фактору – конкретне, фіксоване значення фактору.

2.1.1. Типи змінних

Змінні бувають різними:

Кількісні змінні – впорядковані по величині змінні, до яких можливо використовувати арифметичні дії.

Якісні змінні – змінні, до яких неможливо використати арифметичні дії.

Якісні змінні поділяють на два типи: порядкові та номінальні.

Порядкові змінні – якісні впорядковані змінні.

Порядкові змінні ще називають ранговими. Всі об'єкти можна порівнювати за рівнем проявлення властивості, але неможливо визначити величину відмінності. Ці змінні не мають масштабної одиниці, вони не мають початку координат. Для порядкових змінних допустимі функціональні перетворення типу $x > y \Leftrightarrow f(x) > f(y)$, але арифметичні дії неможливі.

Приклад 8 (Порядкова змінна – ступінь нагріву)
Холодний, теплий, гарячий

Приклад 9 (Порядкова змінна – шкільні оцінки)
1,2, ... , 11, 12

Коли порядкові змінні позначають цифрами, то часто ігнорують природу цих змінних. Часто можна зустріти, що розраховують середнє значення порядкової змінної. Це є досить поширеною помилкою. Одержане число не можна вважати адекватною оцінкою.

Номінальні змінні – якісні не впорядковані змінні.

Приклад 10 (Номінальна змінна – розчинник)
Бензол, хлороформ, ацетон.

У номінальних змінних кожному об'єкту приписується ярлик (належність до певної групи або класу). Тут об'єкти розрізняються за проявом властивості, але не розрізняються за рівнем прояву властивості. Назва ярлика може довільно мінятись (позначатись числами, символами, словами і т. д.). Будь-яка класифікація відбувається в середовищі номінальних змінних.

Нехай маємо якийсь дослід, у якому певну величину x виміряли n раз. Одержимо набір значень $x_1, x_2, \ldots, x_n$, котрий є випадковою вибіркою з певної генеральної сукупності.

> Випадкова вибірка — вибірка, сформована випадковим вибором з генеральної сукупності.

Кількість вимірювань випадкової величини характеризує об'єм вибірки.

> Об'єм вибірки — кількість вимірювань

Для подальшої обробки експериментальних даних вони повинні задовольняти ряду вимог. Вибірка має бути репрезентативною.

> Репрезентативність — відповідність характеристик вибірки характеристикам генеральної сукупності вцілому.

Якщо про генеральну сукупність нічого не відомо, то гарантією репрезентативності буде тільки випадковий відбір значень для вибірки.

З випадкового характеру вибірки випливає, що судження про генеральну сукупність також буде випадковим! Що ж робити? Як оцінити наші результати?

> Оцінка – це статистика, що використовуються для оцінювання невідомих параметрів розподілу випадкової величини.

Оцінкою ми будемо вважати певний набір (вектор) параметрів $\vec{\theta}$, за допомогою якого ми характеризуємо генеральну сукупність. Оцінки можуть володіти рядом властивостей. Розглянемо їх.

2.2. Властивості оцінок

Для того, щоб оцінки параметрів мали хоча б якийсь практичний сенс, вони повинні задовольняти ряду характеристик, найважливішими з яких є незміщеність, ефективність та спроможність.

2.2.1. Незміщеність

Математичне очікування незміщеної оцінки рівне оцінюваному параметру:

$$M(\hat{\theta}) = \theta \tag{2.1}$$

де $\hat{\theta}$ – оцінка параметра θ;
θ – значення параметра θ;

Відповідно, якщо ця умова не задовільняється, то оцінка буде зміщеною. Використання таких оцінок може бути джерелом систематичних похибок.

2.2.2. Ефективність

Ефективна оцінка $\hat{\theta}$ має найменшу дисперсію із усіх можливих:

$$D(\hat{\theta}) \leqslant \tilde{\theta} \qquad \forall \tilde{\theta} \in \Omega \tag{2.2}$$

Так як безпосередньо визначити ефективну оцінку не можна, то цю властивість можна використовувати для порівняння оцінок, одержаних різними методами. Кращою буде та, в котрої менша дисперсія.

2.2.3. Робастність

Під робастністю в розуміють нечутливість оцінки до різних відхилень і неоднорідностей у вибірці, зумовленими тими чи іншими, у загальному випадку невідомими, причинами.

2.2.4. Спроможність

Оцінка є спроможною (обгрунтованою), якщо вона підпорядковується закону великих чисел. Тобто зі збільшенням об'єму вибірки $\hat{\theta}$ прямує по ймовірності до параметра θ :

$$\forall \epsilon > 0 \qquad \lim_{n \to \infty} P(|\hat{\theta} - \theta| > 0) = 0 \tag{2.3}$$

або іншими словами, стає більш близькою до θ.

Для знаходження оцінок можна використати метод максимальної правдоподібності.

2.3. Метод максимальної правдоподібності

Метод максимальної правдоподібності в математичній статистиці — це метод оцінювання невідомого параметра шляхом максимізації функції правдоподібності. Він ґрунтується на припущенні про те, що вся інформація про статистичну вибірку міститься у цій функції. Метод максимальної правдоподібності був впроваджений у практику на початку 20-го століття, хоча й був відомий раніше.

Нехай маємо випадкову величину ξ, котра має функцію розподілу $F(\vec{x}, \vec{\theta})$:

$$\xi \sim F(\vec{x}, \vec{\theta}) \tag{2.4}$$

де $\vec{x} = x_1, x_2, \ldots, x_n$ — вибірка;
$\vec{\theta} = \theta_1, \theta_2, \ldots, \theta_m$ — вектор параметрів.

Для безперервної ξ функція максимальної правдоподібності має вигляд:

$$L(\vec{x}, \vec{\theta}) = f(x_1, \vec{\theta}) \cdot f(x_2, \vec{\theta}) \cdot \ldots f(x_n, \vec{\theta}) \tag{2.5}$$

де $f(x_1, \vec{\theta})$ — функція густини ймовірностей ξ;

Оцінка $\hat{\vec{\theta}}$ буде оцінкою максимальної правдоподібності, коли:

$$L(\vec{x},\hat{\vec{\theta}}) > L(\vec{x},\tilde{\vec{\theta}}) \qquad \forall\tilde{\vec{\theta}} \tag{2.6}$$

Якщо існує ефективна оцінка $\hat{\vec{\theta}} = \vec{\theta}(\vec{x})$, то система рівнянь:

$$\begin{cases} \frac{\partial \ln f(\vec{x},\vec{\theta})}{\partial\theta_1} &= 0 \\ \cdots\cdots & \\ \frac{\partial \ln f(\vec{x},\vec{\theta})}{\partial\theta_1} &= 0 \end{cases}$$

має єдине рішення. Доказувати не будемо!

Приклад 11 (Використання методу максимальної правдоподібності) Нехай вибірка $x_1, x_2, \ldots, x_n$ має нормальний розподіл:

$$x_i \sim N(m,\sigma^2) \tag{2.7}$$

тоді функція максимальної правдоподібності буде мати вигляд:

$$L(x_1,x_2,\ldots,x_n,m,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x_1-m)^2}{2\sigma^2}} \cdot \ldots \cdot \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x_n-m)^2}{2\sigma^2}} \tag{2.8}$$

її логарифм:

$$\ln L(x_1,x_2,\ldots,x_n,m,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-m)^2 - \frac{1}{2}n\ln\sigma^2 - \frac{1}{2}n\ln 2\pi \tag{2.9}$$

Система рівнянь:

$$\begin{cases} \frac{\partial \ln L(x_1,x_2,\ldots,x_n,m,\sigma^2)}{\partial m} = \frac{1}{\sigma^2}\sum\limits_{i=1}^{n}(x_i-m) = 0 \\ \frac{\partial \ln L(x_1,x_2,\ldots,x_n,m,\sigma^2)}{\partial \sigma^2} = \frac{1}{2\sigma^4}\sum\limits_{i=1}^{n}(x_i-m)^2 - \frac{n}{2\sigma^2} = 0 \end{cases}$$

має рішення:

$$\begin{cases} \hat{m} = \frac{1}{n}\sum\limits_{i=1}^{n} x_i \\ \hat{\sigma^2} = \frac{1}{n}\sum\limits_{i=1}^{n}(x_i - m)^2 \end{cases}$$

Як видно з прикладу, середнє значення та дисперсія є оцінками параметрів m та σ^2 нормального розподілу. Якщо розподіл не нормальний, то розраховані за наведеними вище формулами середнє значення та дисперсія вже не будуть ефективними оцінками, й відповідні оцінки мають бути одержані іншими методами!

2.4. Характеристики вибіркових випадкових величин

Охарактеризувати особливості певної генеральної сукупності можна тільки на основі дослідження вибірки з даної генеральної сукупності. Цьому служать наступні «вибіркові» величини, котрі дають змогу оцінити ті чи інші аспекти генеральної сукупності в цілому.

2.4.1. Середнє значення

У випадку нормального розподілу (а також ряду інших розподілів) середнє значення є ефективною та спроможною оцінкою математичного очікування генеральної сукупності.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.10}$$

Завдяки буденності нормального розподілу в хімії середнє значення дуже широко використовується для оцінки «точного» значення вимірюваних величин.

Приклад 12 (Розрахунок середнього значення)
При аналізі зразків залізної руди були одержані наступні значення вмісту Fe_2O_3: 38,71%, 38,90%, 38,62%, 38,74%. Розрахувати середнє значення вмісту Fe_2O_3 в залізній руді .

Розрахунок за допомогою R-статистики:

```
> x=c(38.71,38.90,38.62,38.74)
> mean(x)
[1] 38.7425
```

Відповідь – середнє значення 38,74% (Не забуваємо про правила округлення!)

2.4.2. Вибіркова дисперсія

Вибіркова дисперсія характеризує міру розсіювання результатів вимірювань.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2.11}$$

Приклад 13 (Розрахунок дисперсії)
При мікрокристалічному дослідженні зразків сталі були одержані наступні значення кількості кристалітів перліту: вмісту хрому: 19, 21, 18, 16, 16, 21. Розрахувати дисперсію кількості кристалітів перліту.

Розрахунок за допомогою R-статистики:

```
> x=c(19, 21, 18, 16, 16, 21)
> var(x)
[1] 5.1
```

Відповідь – дисперсія 5,1

2.4.3. Вибіркове середньоквадратичне відхилення

Вибіркове середньоквадратичне відхилення характеризує величину абсолютного розсіювання експериментальних точок.

$$s = +\sqrt{s^2} \tag{2.12}$$

Приклад 14 (Розрахунок середньоквадратичного відхилення)
При аналізі зразків глини були одержані наступні значення вмісту хрому: 0,025%, 0,031%, 0,026%, 0,024%, 0,029%. Розрахувати середньоквадратичне відхилення вмісту хрому в глинах.

Розрахунок за допомогою R-статистики:

```
> x=c(0.025,0.031,0.026,0.024,0.029)
> sd(x)
[1] 0.002915476
```

Відповідь – середньоквадратичне відхилення 0,0029% (Не забуваємо про правила округлення!)

2.4.4. Вибірковий коефіцієнт варіації

Вибірковий коефіцієнт варіації характеризує величину відносного розсіювання експериментальних точок.

$$\nu = \frac{s}{\bar{x}} \tag{2.13}$$

Приклад 15 (Розрахунок коефіцієнту варіації)
Бабітові заготовки були проаналізовані на вміст Сурми. Були одержані наступні значення вмісту Сурми: 14,72%, 15,51%, 14,60%, 15,10%, 14,70%, 14,74%. Розрахувати коефіцієнт варіації вмісту Сурми в бабіті.

Розрахунок за допомогою R-статистики:

```
> x=c(14.72,15.51,14.60,15.10,14.70,14.70)
> sd(x)/mean(x)
[1] 0.02353104
```

Відповідь – коефіцієнт варіації 0,02353 (2,353%) (Не забуваємо про правила округлення!)

2.4.5. Вибіркова медіана

Медіана вибірки – значення що поділяє впорядковану за зростанням елементів вибірку на частини, що містять однакове число елементів. Половина елементів вибірки мають менше або рівне значення ніж медіана, половина – більше або рівне.

Метод визначення медіани залежить від кількості елементів у вибірці. У випадку непарної кількості медіаною буде число, що знаходиться всередині вибірки. У випадку парної – середнє значення двох середніх членів.

Приклад 16 (Непарна кількість елементів у вибірці)
Нехай маємо впорядковану вибірку 3, 5, 8, 10, 11, котра містить 5 елементів. Медіаною буде значення середнього, третього елементу - 8.

Приклад 17 (Парна кількість елементів у вибірці)
Нехай маємо впорядковану вибірку 3, 5, 8, 10, 11, 13 котра містить 6 елементів. Медіаною буде середнє від значень елементів, котрі знаходяться посередині, третього та четвертого. Це 8 і 10. Їх середнє значення - 9.

Приклад 18 (Розрахунок медіани)
Вміст Калію в перерахунку на K_2O в зразках каміння становить 2,25%, 3,31%, 4,02%, 3,24%, 6,29%, 1,23%. Визначити медіанний вміст K_2O в камінні.

Розрахунок за допомогою R-статистики:

```
> x=c(2.25, 3.31, 4.02, 3.24, 6.29, 1.23)
> median(x)
[1] 3.275
```

Відповідь – медіанний вміст K_2O в камінні 3,28% (Не забуваємо про правила округлення!)

2.4.6. Вибіркові квантилі

Не існує методу для однозначного розрахунку квантилів вибіркових даних. Функція quantile() R-статистики пропонує дев'ять (!) різних способів розрахунку квантилів. Ці методи хоча й дають близькі результати, але вони все-таки відрізняються.

Приклад 19 (Розрахунок квантілів)
В зразках кераміки був визначений наступний вміст Кальцію: 2,21%, 3,31%, 2,02%, 4,02%, 6,09%, 2,34%, 2,55%, 3,78%. Розрахувати перший та третій квартилі вмісту Кальцію в зразках кераміки.

Розрахунок за допомогою R-статистики:

```
> x=c(2.21,3.31,2.02,4.02,6.09,2.34,2.55,3.78)
> quantile(x)
0%   25%   50%   75%  100%
2.0200 2.3075 2.9300 3.8400 6.0900
```

Відповідь – перший квартіль – 2,31%, третій – 3,84% (Не забуваємо про правила округлення!)

2.4.7. Розмах вибірки

Розмах вибірки характеризує діапазон розсіювання експериментальних точок. Він дорівнює різниці між максимальним і мінімальним значеннями у вибірці.

$$R = x_{max} - x_{min} \tag{2.14}$$

Розмах – одна з найпростіших мір розсіювання у вибірці. Дає інформацію про ширину інтервалу, в якому зосереджений весь набір числових даних, геометрично – це ширина відрізка, в якому розташовуються всі значення.

Простота розрахунку, наочність та інтуїтивна зрозумілість цієї характеристики розсіювання значень є очевидною перевагою перед такими мірами розсіювання як дисперсія або стандартне відхилення. Недоліком розмаху є те, що він не містить інформації про характер розподілу результатів в інтервалі розсіювання та не стійкий до викидів, що певною мірою обмежує його використання.

Приклад 20 (Розрахунок розмаху)
При калібровці приладу одержали наступні значення напруги: 3,01 В, 2,98 В, 2,97 В, 3,01 В, 2,99 В, 3,01 В. Розрахуйте діапазон осциляцій напруги.

Розрахунок за допомогою R-статистики:

```
> x=c(3.01, 2.98, 2.97, 3.01, 2.99, 3.01)
> diff(range(x))
[1] 0.04
```

Відповідь – діапазон осциляцій напруги становить 0,04 В.

2.4.8. Міжквартильний інтервал

Міжквартильний інтервал (міжквартильний розмах) є надійною мірою розсіювання вибірки. Він дорівнює різниці між третім і першим квартилями.

$$IQR = Q_{0.75} - Q_{0.25} \tag{2.15}$$

Якщо окреме значення знаходиться між першим і третім квартилем, то можна зробити висновок, що це значення близько до центру розподілу, так як у діапазоні між цими двома показниками знаходиться половина значень. Цей показник не чутливий до викидів, так як

із розгляду виключаються по 25% крайніх значень вибірки з кожного боку. Міжквартильний розмах використовується для побудови діаграми розмаху (Рис. 2.2).

Приклад 21 (Розрахунок міжквартильного інтервалу) Дослідження якості фасовочного апарату показало наступні кількості розфасованого реактиву: 25,01 г, 24,99 г, 25,02 г, 25,02 г, 25,00 г, 25,01 г. Визначити міжквартильний розмах фасовок.

Розрахунок за допомогою R-статистики:

```
> x=c(25.01,24.99,25.02,25.02,25.00,25.01)
> IQR(x)
[1] 0.015
```

Відповідь – міжквартильний розмах становить 0,015 г.

2.5. Характеристики багатомірних випадкових величин

Основними характеристиками багатомірних випадкових величин є вибіркова коваріація та вибіроквий коефіцієнт кореляції.

2.5.1. Вибіркова коваріація

Коваріація – міра спільної варіативності двох випадкових величин.

Якщо обидві величини демонструють односпрямовану зміну, то коваріація позитивна, а якщо різноспрямовану – негативна. Якщо коваріація близька до нуля, величини незалежні. Проте інтерпретація величини коваріації не очевидна, оскільки її величина залежить від значень самої випадкової величини.

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{2.16}$$

Коваріацію можна розглядати як спільну дисперсію двох випадкових величин, а дисперсію – як окремий випадок коваріації, коли розглядається коваріація величини самої із собою.

Коваріація широко застосовується у всіх галузях, де використовуються статистичні дослідження та потрібна обробка результатів експериментів. Хоча більше значення має коваріаційна матриця, де розглядаються взаємні коваріації багатьох випадкових величин.

Коваріація має розмірність, яка є добутком розмірностей випадкових величин. Тобто величина коваріації залежить від одиниць виміру незалежних величин. Ця особливість коваріації ускладнює її використання.

Приклад 22 (Розрахунок вибіркової коваріації)
Розрахуйте коваріацію між абсолютною температурою та приведеною енергією Гібса для фтору.

T, K	298	500	600	700	800	900	1000
Ф, $\frac{Дж}{моль \cdot K}$	173,08	188,71	194,40	199,42	203,65	207,59	211,05

Розрахунок за допомогою R-статистики:

```
> T=c(298,500,600,700,800,900,1000)

> F=c(173.08,188.71,194.40,199.42,203.65,207.59,211.05)

> cov(T,F)

[1] 3085.969
```

Відповідь – коваріація становить 3086 Дж/моль.

2.5.2. Вибірковий коефіцієнт кореляції

Вибірковий коефіцієнт кореляції характеризує ступінь функціональної залежності між двома випадковими величинами.

$$\hat{r}_{xy} = \frac{s_{xy}}{s_x s_y} \tag{2.17}$$

На відміну від вибіркової коваріації, вибірковий коефіцієнт кореляції – берзрозмірна величина.

Приклад 23 (Розрахунок коефіцієнта кореляції)
Розрахуйте коефіцієнт кореляції між приведеною енергією Гібса та температурою для дейтерію.

T, K	500	600	700	800	900	1000
Ф, $\frac{\text{Дж}}{\text{моль}\cdot K}$	130,92	136,28	140,82	144,63	148,06	151,22

```
> T=c(500,600,700,800,900,1000)

> F=c(130.92,136.28,140.82,144.63,148.06,151.22)

> cor(T,F)

[1] 0.9948748
```

Відповідь – коефіцієнт кореляції становить 0,9949.

Існують й інші величини, що характеризують вибірки, але вони мають менше практичне значення, тому ми тут їх розглядати не будем.

2.6. Похибки

Як вже згадувалось вище, кожен результат вимірювання — випадкова величина, на значення якої впливає ряд незначимих факторів, котрі тяжко врахувати. Якщо вимірювання покликане оцінити істинне значення певного параметру, то в таких випадках говорять про похибки вимірювань.

Похибка (спостереження) — відхилення реального результату від істинного значення

В аналітичній практиці розрізняють різні види похибок – абсолютну та відносну. Справедливості ради слід відмітити, що точного значення ми не можемо знати, тому в більшості випадків за точне значення приймають його оцінку, звичайно це – середнє значення. Тому дані похибки мають формальний характер:

2.6.1. Абсолютна похибка

$$\Delta x = x - m \tag{2.18}$$

де x – результат вимірювання;
m – істинне значення.

2.6.2. Відносна похибка

$$\epsilon = \frac{\Delta x}{m} \tag{2.19}$$

де Δx – абсолютна похибка;
m – істинне значення.

Абсолютна та відносна похибки не можуть мати особливого сенсу, так як точне значення апріорі невідоме, а доступні тільки його оцінки. Крім формальних абсолютної та відносної похибок існують також операційні (технічні) похибки: груба похибка, або викид, та систематична похибка.

> Груба похибка (викид) — результат, що різко відрізняється від решти результатів, і, ймовірно, є помилковим.

Грубі похибки виникають внаслідок значимої дії випадкових факторів. Це можуть бути порушення методик, неуважність, нестабільність напруги в мережі, коливання будівлі внаслідок руху транспорту тощо.

Зі збільшенням об’єму вибірки зростає шанс того, що буде одержано значення, що значно відрізняється від решти, проте не є грубою

похибкою. За великої кількості вимірювань можуть з'являтися маргінали – значення, які належать даній генеральній сукупності, проте ймовірність появи яких досить низька.

> Маргінал — значення, що значно відхиляється від центру розподілу, проте не є грубою похибкою.

> Систематична похибка — похибка, викликана дією фактора, що є постійним, або закономірно змінюється в ході експерименту.

Як вже було сказано, помилки вимірювання викликаються великою кількістю різноманітних причин (факторів). Іноді в проведеної серії вимірювань вдається виділити такі причини помилок, ефект дії яких може бути розрахований.

Наприклад, якщо після вимірів виявлено неправильне регулювання приладу, яка привела до зміщення початку відліку, то все зняті показання будуть зміщені або на постійну величину, якщо шкала приладу рівномірна, чи величину, що змінюється за певним законом, якщо шкала приладу нерівномірна.

Іншим прикладом може служити зміна зовнішніх умов, наприклад, температури, якщо відомо вплив цих змін на результати вимірювань. До названих причин можна також віднести деяку недосконаліть вимірювальних приладів на межі області їх застосування, що викликає відомі помилки. Прийнято говорити, що кожна з таких причин викликає систематичну помилку.

Виявлення систематичних помилок, що викликаються кожним окремим фактором, вимагає спеціальних досліджень (наприклад, вимірювань однієї і тієї ж величини різними методами або вимірювань одним і тим же приладом деяких еталонів, відомих величин). Але як тільки систематичні помилки виявлені та їх величини розраховані, вони можуть бути легко усунені шляхом введення відповідних поправок в результати вимірювання.

Підкреслимо, що при цьому загальна помилка кожного результату залишається невідомою, так що мова йде не про виділення із загальної помилки деякої частини у вигляді систематичної помилки, а лише

про введення поправок на відомий ефект дії тих чинників, які вдалося виявити.

> Випадкова похибка — похибка, викликана дією випадкових незначних факторів.

Похибки вимірювання, що залишаються після усунення всіх виявлених систематичних помилок, або похибки результатів вимірювань, виправлених шляхом введення відповідних поправок, називаються випадковими. Випадкові похибки викликаються великою кількістю таких факторів, ефекти дії яких настільки незначні, що їх не можна виділити і врахувати окремо (при даному рівні техніки і точності вимірювань).

Випадкову похибку можна розглядати як сумарний ефект дії таких факторів. Випадкові похибки є невід'ємними, їх не можна виключити в кожному з результатів вимірювань. Але за допомогою методів теорії ймовірностей можна врахувати їх вплив на оцінку істинного значення вимірюваної величини, що дозволяє визначити значення вимірюваної величини зі значно меншою похибкою, ніж похибки окремих вимірювань. Облік впливу випадкових помилок заснований на знанні законів їх розподілу.

2.7. Точність та відтворюваність

Треба чітко розуміти, що похибки – це не помилки експерименту. Навпаки, вони є показником якості експерименту. Похибки характеризують об'єктивний рівень недосконалості приладу або неідеальності методики обробки. Їх не можна повністю усунути, але зате можна сказати, в яких рамках результату можна довіряти. Характеристиками якості методу є такі характеристики як точність, достовірність та відтворюваність.

> Достовірність – ступінь довіри до результатів вимірювань, характеризується ймовірністю того, що істинні значення вимірюваної величини знаходяться у вказаних межах.

Достовірність може бути низька при наявності похибок, про існування яких експериментатор не здогадується. Достовірність при використанні автоматизованих вимірювальних систем знижується з ростом їх складності й істотно залежить від кваліфікації персоналу.

Головним методом забезпечення достовірності є зіставлення результатів вимірювання однієї і тієї ж величини різними, не пов'язаними один з одним способами. Наприклад, для визначення деякого металу використовують колометрію та амперометрію.

Наведемо кілька прикладів, що ілюструють випадки, коли, незважаючи на застосування точних засобів вимірювань, виходять абсолютно помилкові дані, що вводять людину в оману .

Приклад 24 (Вимірювання температури повітря)
Для вимірювання температури повітря в теплиці використаний датчик температури з похибкою ±0,5 °С. Однак датчик встановлений таким чином, що в певний період дня на нього падають прямі промені сонця, які нагрівають датчик, але не змінюють температуру повітря в теплиці. При цьому похибка вимірювання температури повітря може скласти +5 °С, що дозволяє кваліфікувати результат вимірювання як недостовірний.

Приклад 25
В автоматизованій системі для вимірювання параметрів продукції використаний модуль введення з похибкою ± 0,05%, однак при налагодженні системи програміст помилково встановив частоту шумоврежекторного фільтра не 50, а 60 Гц. Проведені приймально-здавальні випробування системи не дозволили виявити цю помилку. В результаті похибка вимірювань внаслідок наведеної перешкоди з частотою 50 Гц підвищилася до ± 10% замість очікуваних ± 0,05%.

Правильність вимірювань – це характеристика якості вимірювань, яка відображає близькість до нуля систематичної похибки результатів вимірювання.

Правильність методу вимірювань має значення у випадках, коли можна прямо або опосередковано представити істинне значення вимірюваної величини.

Хоча для деяких методів вимірювань істинне значення не може бути відомо точно, існує можливість мати у своєму розпорядженні опорне значенням вимірюваної величини, наприклад, коли є в розпорядженні відповідні стандартні зразки, або коли прийняте опорне значення може бути встановлено за допомогою іншого методу вимірювання, або шляхом приготування відомого зразка.

При цьому правильність того чи іншого методу вимірювань може бути досліджена за допомогою зіставлення прийнятого опорного значення з рівнем результатів, отриманих цим методом.

Правильність, як правило, висловлюють в термінах систематичної похибки. Наприклад, при хімічному аналізі систематична похибка проявляється у випадках, коли метод вимірювання не дозволяє повністю виділити елемент або коли наявність одного елемента заважає визначенню іншого.

> Відтворюваність – характеристика якості вимірювань, яка відображає близькість один до одного результатів незалежних повторних вимірювань за певною методикою, які виконуються в різний час, в різних місцях, на іншому обладнанні.

Розглянемо відтворюваність на прикладі спектрофотометра.

Приклад 26 (Відтворюваність вимірювань спектрофотометра) Відтворюваність спектрофотометра оцінюється можливістю приладу повторити вимір спектральних коефіцієнтів відбиття або пропускання певного стабільного зразка. Вважається, що спектрофотометр має дуже високу відтворюваність, якщо результати вимірювання спектрального коефіцієнта пропускання повторюються з похибкою менш ніж $\Delta\tau(\lambda) = \pm 0.001$. Це означає, що якщо одного разу при вимірах було отримано значення $\Delta\tau(\lambda) = 0.487$ на даній довжині хвилі λ, то у всіх інших випадках виміряні значення $\tau(\lambda)$ будуть знаходитися в межах

від 0.486 до 0.488. Відносна похибка $\Delta\tau(\lambda)/\tau(\lambda)$ природно зростає зі зменшенням $\tau(\lambda)$ і може стати досить значною при малих значеннях, наприклад при $\tau(\lambda) < 0.1$. Прилад дає погану відтворюваність вимірювань, якщо $\Delta\tau(\lambda) > 0.005$. Оцінка відтворюваності повинна проводитися через різні часові інтервали. Прилад може добре повторити початковий результат при негайному повторі, але дати велику розбіжність, якщо після початку вимірювань пройшов день або більше.

Точність (прецезійність) – характеристика блискості результатів вимірювань до точного значення.

Загальний термін ”точність” використовують в стандарті ДСТУ ISO 5725 щодо обох термінів – ”правильність” і ”прецизійність”. У свій час термін ”точність” використовувався, поширюючись лише на одну складову, іменовану тепер правильністю, проте стало очевидним, що він висловлює сумарне відхилення результату від еталонного (опорного) значення, викликане як випадковими, так і систематичними причинами.

Звичайно, з огляду на вищесказане, точність є віртуальною величиною.

Звичайно намагаються уникнути грубих та систематичних похибок.

2.8. Методи виявлення викидів

При математичній обробці результатів вимірювань не слід враховувати свідомо невірні результати (викиди), або, як кажуть, результати, що містять грубі помилки. Грубі помилки виникають внаслідок порушення основних умов вимірювання або в результаті недогляду експериментатора (наприклад, при поганому освітленні замість «3» записують «8»). При виявленні грубої помилки результат вимірювання слід відразу відкинути, а саме вимірювання повторити (якщо це можливо).

Зовнішнім ознакою результату, що містить грубу помилку, є його різка відмінність за величиною від результатів інших вимірів. На цьому засновані деякі критерії виключення грубих помилок за їх величиною, але найбільш надійним і ефективним способом бракування невірних результатів залишається бракування їх безпосередньо в процесі самих вимірювань.

Проте, грубі помилки не завжди вдається виявити або розпізнати на стадії вимірювання. Тому приходиться додатково використовувати різні математичні процедури, щоб очистити дані від викидів.

Всі методи виявлення грубих похибок *суб'єктивні*, так як підозрілий результат може належати до генеральної сукупності.

З цим пов'язана одна з класифікацій викидів. Їх поділяють на жорсткі, м'які та впливові.

Жорсткі викиди, як правило, можуть бути виявлені будь-яким способом, що можна застосувати до масиву даних. Вони сильно впливають на модель. При видаленні жорсткого викиду з масиву даних характеристики моделі змінюються суттєво.

М'які викиди можуть виявлятися лише деякими з існуючих алгоритмів, а при використанні інших способів можуть не виявлятися зовсім, і будуть вважатися нормальними елементами вибірки.

Впливові спостереження дуже схожі на м'які викиди, але виявляються далеко не всіма алгоритмами. Проте вони сильно впливають на значення параметрів моделей та результати статистичного тестування. Такі спостереження часто входять в нормальні межі сукупності і відстань між ними і основним масивом даних, що використовуються в аналізі, може пояснюватися особливостями вибірки.

Так що різні алгоритми виявлення викидів володіють різною чутливістю.

2.8.1. Правило трьох сигм

При статистичних оцінках широко застосовують правило трьох сигм: відхилення значення нормально розподіленої випадкової величини x від її математичного очікування $M(x)$ не перевищує потроєного середньоквадратичного відхилення σ з ймовірністю близько 0,9973 (Рис.2.1).

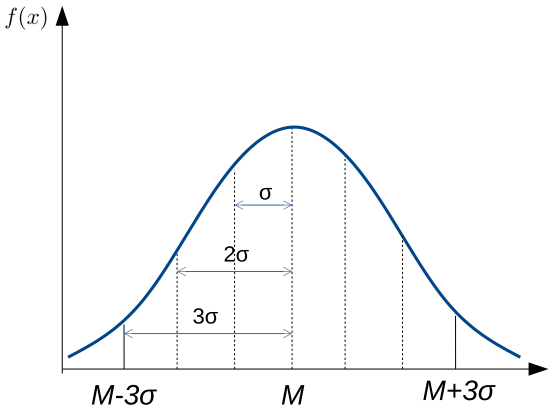


Рис. 2.1: Правило 3σ

Інакше кажучи, з ймовірністю 0,9973 значення нормально розподіленої випадкової величини знаходиться в інтервалі $[M(x) - 3\sigma; M(x) + 3\sigma]$.

На підставі цього правила для виключення з результатів вимірювань грубих помилок часто використовують критерій трьох сигм: значення нормально розподіленої випадкової величини, що відхиляються від математичного очікування $M(x)$ більше, ніж на три сигми, малоймовірні (ймовірність дорівнює $1 - 0.9973 = 0.0027$), і тому є грубими помилками. На практиці використовують оцінки математичного очікування та середньоквадратичного відхилення – середнє значення та вибіркове середньоквадратичне відхилення. Тобто значення x_i – груба помилка, якщо:

$$|x_i - \bar{x}| > 3s \tag{2.20}$$

де x_i – сумнівне значення;
s – вибіркове середньоквадратичне відхилення.

Критерій трьох сигм звичайно застосовують для швидкого набли-

женого визначення грубих помилок у вибірці. Перевагами критерію трьох сигм вважається те, що він простий, наочний і легко запам'ятовується, при його застосуванні не потрібні таблиці та складні розрахунки.

Але при збільшенні кількості вимірювань ймовірність відхилення хоча б одного значення вибірки від математичного очікування вже не дорівнює 0.0027, і залежить від об'єму вибірки n. Так, при відомих параметрах нормального розподілу ймовірність того, що при відсутності грубих помилок хоча б одне значення вийде за межі інтервалу, дорівнює $1 - 0.0973^n$. При цьому зростає ймовірність появи маргінальних значень. Тому цей критерій не слід використовувати, коли $n > 20$.

2.8.2. Метод Тьюкі

Точки за межами інтервалу $L; U$ вважаються викидами:

$$\begin{aligned} L &= Q_{0.25} - K(Q_{0.75} - Q_{0.25}) \\ U &= Q_{0.75} + K(Q_{0.75} - Q_{0.25}) \end{aligned} \tag{2.21}$$

де L – нижня межа інтервалу;
U – верхня межа інтервалу;
$Q_{0.25}$ – нижній (перший) квартиль;
$Q_{0.75}$ – верхній (третій) квартиль;
K – коефіцієнт. Для виявлення м'яких викидів використовують $K = 1.5$, жорстких – $K = 3$.

Метод Тьюки підходить для великих одновимірних вибірок, коли можна чітко виділити структурні характеристики варіаційного ряду. Для невеликих вибірок структурні характеристики тяжко визначити, або неможливо зовсім (наприклад, для вибірки $n = 4$ визначити квартилі неможливо), тоді метод Тьюкі непридатний.

Даний метод підходить для вибірок, які не мають значної асиметрії, в іншому випадку спостереження, які не є викидами, можуть бути помилково віднесені до аномальних.

На цьому методі грунтується графічний спосіб визначення викидів – так звана діаграма розмаху (Рис. 2.2).

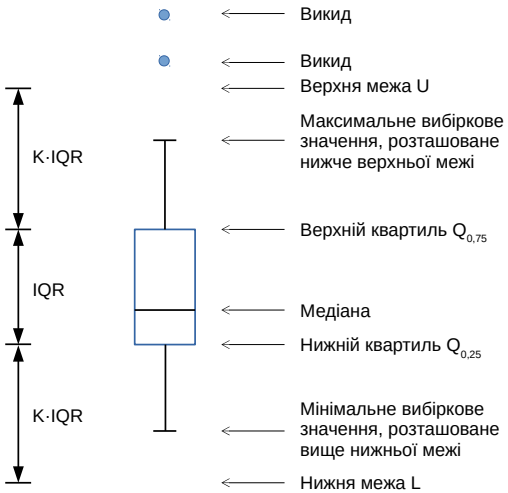


Рис. 2.2: Діаграма розмаху

2.9. Закон накопичення похибок

Закон накопичення похибок випливає з властивостей дисперсії. Якщо випадкова величина z є функцією інших випадкових величин $x_1, \ldots, x_n$:

$$z = f(x_1, \ldots, x_n) \tag{2.22}$$

то вибіркова дисперсія буде визначатись за формулою:

$$s_z^2 = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i} \right)^2 s_{x_i}^2 \tag{2.23}$$

Приклад 27
Випадкова величина z є сумою двох незалежних випадкових величин x і y.

$$z = x + y \tag{2.24}$$

$$s_z^2 = \left(\frac{\partial(x+y)}{\partial x}\right)^2 s_x^2 + \left(\frac{\partial(x+y)}{\partial y}\right)^2 s_y^2 = 1 \cdot s_x^2 + 1 \cdot s_y^2 \tag{2.25}$$

Одержуємо відому нам формулу (1.27) – дисперсія суми дорівнює сумі дисперсій.

Приклад 28 ()
Оцінити похибку при вимірюванні лінійної швидкості газу, якщо об'ємна швидкість газу в трубі становить $V = 3000$ м^3/год із стандартним відхиленням $s_V = 10$ м^3/год, поперечний переріз трубопроводу становить $F = 0.1$ м^2 із стандартним відхиленням $s_F = 1 \cdot 10^{-4}$ м^2.

Розв'язок

Знайдемо лінійну швидкість потоку, підставивши числові значення в формулу:

$$v = \frac{V}{F} = \frac{3000 \text{ м}^3/\text{год}}{1 \cdot 10^{-4} \text{ м}^2} = 30000 \text{ м/год} = 8.82 \text{ м/с} \tag{2.26}$$

Розрахуємо стандартне відхилення лінійної швидкості потоку:

$$s_v = \sqrt{\left(\frac{\partial v}{\partial V}\right)^2 s_V^2 + \left(\frac{\partial v}{\partial F}\right) s_F^2} = \sqrt{\left(\frac{1}{F^2}\right)^2 s_V^2 + \left(\frac{V^2}{F^2}\right) s_F^2} = 0.03 \text{ м/с} \tag{2.27}$$

2.10. Довірчі інтервали

Ще одним інструментом для опису вибірок є метод довірчих інтервалів. Цей метод був введений в практику американським математиком

польського походження Нейманом Ю. Ч. (відомий як Єжи Нейман). Довірчі інтервали розраховуються на основі вибіркових даних. Задається необхідний рівень достовірності. За стандарт прийнято 95% ($p = 0.95$).

Чому саме 95% (рівень значимості 5%)?

З огляду на те, що α – ймовірність помилки першого роду, здається, має сенс зробити цю область якомога меншою. Наприклад, якщо ми встановимо рівень α на рівні 10%, то існує велика ймовірність того, що ми можемо помилково відхилити нульову гіпотезу, тоді як рівень α в 1% зробить цю область зовсім маленькою. То чому б не використати 1% замість стандартних 5%?

Чим менше α-рівень, тим менша область, де відхиляється нульова гіпотеза. Отже, якщо у вас маленька область, є більше шансів, що ви не відхилите нульову гіпотезу, хоча насправді ви повинні відхилити. Це помилка другого роду.

Іншими словами, чим більше ви намагаєтеся уникнути помилки першого роду, тим більша ймовірність, що ви одержите помилку другого роду. Таким чином $\alpha = 5\%$ є компромісним рішенням .

Довірчий інтервал – область, що з заданою ймовірністю локалізує точне значення оцінюваного параметра

$$P(L \leqslant \theta \leqslant U) = p \tag{2.28}$$

де L – нижня межа довірчого інтервалу;
U – верхня межа довірчого інтервалу;
p – рівень достовірності;
$\alpha = 1 - p$ – рівень значимості.

Давайте розглянемо приклади довірчих інтервалів для різних величин.

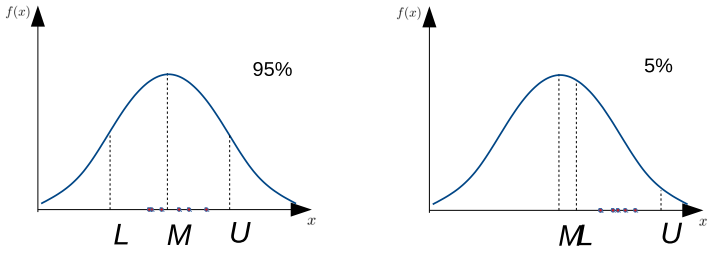


Рис. 2.3: а) Точне значення в межах довірчого інтервалу; б) Точне значення поза межами довірчого інтервалу

2.10.1. Довірчий інтервал для математичного очікування

Довірчий інтервал для математичного очікування грунтується наступних вибіркових оцінках – вибірковому середньому та вибірковому середньоквадратичному відхиленні. Для розрахунку використовується коефіцієнт Стьюдента (квантиль розподілу Стьюдента).

Цей інтервал симетричний:

$$M_{min} = \bar{x} - \frac{t_{\alpha,n-1} \cdot s}{\sqrt{n}} \tag{2.29}$$

$$M_{max} = \bar{x} + \frac{t_{\alpha,n-1} \cdot s}{\sqrt{n}} \tag{2.30}$$

де M_{min} – нижня межа довірчого інтервалу;
M_{max} – верхня межа довірчого інтервалу;
$t_{\alpha,n-1}$ – коефіцієнт Стьюдента;
s – вибіркове стандартне відхилення;
α – рівень недостовірності (рівень значимості);
$n-1$ – кількість ступенів свободи;
$\bar{x}$ – вибіркове середнє;

2.10.2. Довірчий інтервал для дисперсії

Довірчий інтервал для дисперсії грунтується на вибірковій дисперсії. Для розрахунку використовуються коефіцієнти розподілу χ^2 (хі-квадрат) (квантилі χ^2-розподілу). Розподіл χ^2 несиметричний.

$$\sigma^2_{min} = \frac{(n-1)\cdot s^2}{\chi^2_{1-\alpha/2,n-1}} \tag{2.31}$$

$$\sigma^2_{max} = \frac{(n-1)\cdot s^2}{\chi^2_{\alpha/2,n-1}} \tag{2.32}$$

де σ^2_{min} – нижня межа довірчого інтервалу;
σ^2_{max} – верхня межа довірчого інтервалу;
s^2 – вибіркова дисперсія;
$\chi^2_{1-\alpha/2,n-1}$ і $\chi^2_{\alpha/2,n-1}$ — коефіцієнти розподілу χ^2 з відповідними рівнями значимості та кількістю степеней свободи.

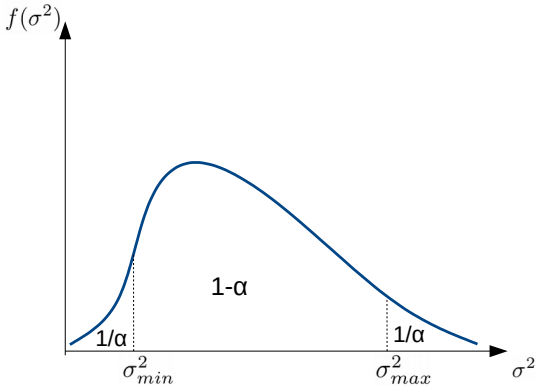


Рис. 2.4: Довірчий інтервал для дисперсії

2.11. Ступені свободи

Концепція ступенів свободи була сформульована ще в 1821 році в роботах астронома і математика Карла Фрідріха Гаусса. її сучасне визначення і використання були вперше розроблено англійським статистиком Вільямом Сілі Госсетом в його статті 1908 Biometrika «Ймовірна похибка середнього», видана під псевдонімом «Студент» . Хоча Госсет насправді не використовував термін «ступеня свободи», він пояснив цю концепцію в ході розробки того, що потім стало відомо як t-розподіл Стьюдента. Сам термін був популяризував англійський статистик і біологом Рональд Фішером.

Поняття ступеней свободи відіграє важливе значення в сучасній статистиці.

> Ступені свободи – величина, що характеризує число незалежних величин у наборі даних.

Якщо з'являються зв'язки між окремими елементами, то кількість ступенів свободи зменшується на відповідну кількість.

Приклад 29 (Як утворюються ступені свободи)

$$\underbrace{x_1, x_2, \ldots, x_n}_{\bar{x}}$$

$x_1, x_2, \ldots, x_n$ – вибірка з n елементів;

$\bar{x}$ – середнє значення.

Внаслідок розрахунку середнього значення утворюється один зв'язок між елементами. Тому кожен раз, коли використовується середнє значення $\bar{x}$, кількість ступенів свободи буде $n - 1$.

Кожен зв'язок дає змогу визначити будь-який елемент набору даних на основі всіх інших.

2.12. Статистичні гіпотези

Як проводиться дослідження? Звичайно так (в термінах статистики): з генеральної сукупності витягується репрезентативна вибірка. На підставі вивчення цієї вибірки робиться висновок про всю сукупність. Слід відмітити, що це основний метод математичної статистики і називається він вибірковим методом. Залежно від предмету дослідження можуть проводитися повторні вибірки, вибірки з декількох генеральних сукупностей, та використовуватись інші методи добору вибірок.

В результаті аналізу цих даних з'являються думки, які мають назву статистичних гіпотез.

> Статистична гіпотеза – припущення про параметри закону розподілу випадкової величини, що формуються на основі вибірки.

На цьому етапі ми будемо керуватись теорією статистичних гіпотез Рональда Фішера. Згідно Фішера існує два типа статистичних гіпотез:

- H_0 – нульова гіпотеза. Вона стверджує, що відхилення, що спостерігаються, *випадкові*;
- H_1 – альтернативна гіпотеза. Відхилення не випадкові.

Перевірка гіпотез

Оскільки нульова гіпотеза висувається на підставі вибіркових даних, то вона може виявитися як правильною, так і неправильної. *Чи вона правильна, ми не знаємо!* Тому вона підлягає статистичній перевірці.

Перевірка здійснюється за допомогою статистичних критеріїв – це спеціальні випадкові величини, які приймають різні дійсні значення. В залежності від поставленого завдання використовуються різні критерії.

Звичайно за перевіркою статистичної гіпотези розуміють розрахунок певної випадкової величини $z = z(\vec{x})$. Процедура перевірки приписує кожному значенню критерія одне з двох рішень – *прийняти* гіпотезу, чи *відхилити*.

При цьому є ризик допустити помилки двох типів:

Помилка першого роду полягає в тому, що гіпотеза буде відкинута, хоча насправді вона правильна. Імовірність допустити таку помилку називають рівнем значимості, й позначають буквою α («альфа»).

Помилка першого роду виникає з ймовірністю α, коли відхиляється вірна гіпотеза H_0, і приймається хибна H_1.

Помилка другого роду полягає в тому, що гіпотеза буде прийнята, але насправді вона неправильна. Імовірність зробити цю помилку позначають буквою β («бета»). Значення називають потужністю критерію – це ймовірність відкидання неправильної гіпотези.

Помилка другого роду виникає з ймовірністю β, коли приймається хибна гіпотеза H_0, і відхиляється вірна H_1.

Табл. 2.1: Варіанти рішень при прийнятті гіпотез

Гіпотеза H_0	Рішення	Ймовірність	Термін
Вірна	Приймається	$1-\alpha$	Достовірність
Вірна	Відхиляється	α	Рівень значимості
Не вірна	Приймається	β	Ймовірність помилки другого роду
Не вірна	Відхиляється	$1-\beta$	Потужність критерію

Рівень значимості задається дослідником самостійно, найбільш часто вибирають значення $\alpha = 0.1$, $\alpha = 0.05$, $\alpha = 0.01$. І тут виникає думка, що чим менше «альфа», тим краще. Але це тільки на перший

погляд: при зменшенні α – ймовірності відкинути правильну гіпотезу, росте β – ймовірність прийняти невірну гіпотезу (за інших рівних умов).

Тому перед дослідником стоїть завдання грамотно підібрати співвідношення ймовірностей і врахувати при цьому тяжкість наслідків, до яких призводять та чи інша помилка.

Приклад 30 (Небажана помилка першого роду)
Два препарати порівнюються за ефективністю в лікуванні одного й того ж стану. Препарат 1 легкодоступний, а препарат 2 дуже дорогий. Нульова гіпотеза — «обидва препарати однаково ефективні», а альтернативна – «препарат 2 ефективніший за препарат 1». У цій ситуації помилкою першого роду було б рішення про те, що препарат 2 є більш ефективним, хоча насправді він не кращий за препарат 1, але коштуватиме пацієнту набагато більше грошей. Це було б небажано з точки зору пацієнта, тут необхідний невеликий рівень значимості.

Приклад 31 (Небажана помилка другого роду)
Відомо, що два препарати однаково ефективні для лікування певного стану. Крім того, кожен з них однаково доступний. Однак є деякі підозри, що препарат 2 викликає серйозні побічні ефекти у деяких пацієнтів, тоді як препарат 1 використовується протягом десятиліть без жодних повідомлень про побічні ефекти. Нульовою гіпотезою є «частота побічної дії в обох препаратах однакова», а альтернативна — «частота побічних ефектів у препарату 2 вища, ніж у препарата 1». Помилкове відхилення нульової гіпотези, коли вона насправді істинна (помилка першого роду), не матиме великих наслідків для пацієнта, але помилка другого роду (тобто, неможливість відхилити нульову гіпотезу, коли насправді альтернатива істинна, що призведе до рішення, що препарат 2 не є більш шкідливим, ніж препарат 1, коли він насправді більш шкідливий) може мати серйозні наслідки з точки зору громадського здоров'я. Тут встановлення великого рівня значимості є доречним.

Помилку першого роду часто називають помилковою тривогою, або псевдопозитивним спрацьовуванням. Якщо, наприклад, аналіз крові показав наявність захворювання, хоча насправді людина здорова, або металодетектор видав сигнал тривоги, спрацювавши на металеву пряжку ременя, то прийнята гіпотеза не вірна, а отже зроблена помилка першого роду. Термін «псевдопозитивний» в даному випадку не має відношення до бажаності або небажаності самої події.

Цей термін широко використовується в медицині. Наприклад, тести, призначені для діагностики захворювань, іноді дають позитивний результат (тобто показують наявність захворювання у пацієнта), коли насправді пацієнт цим захворюванням не страждає. Такий результат називається псевдопозитивним.

В інших галузях звичайно використовують словосполучення зі схожим змістом, наприклад, «помилкове спрацьовування», «хибна тривога», "ризик постачальника" і т.д. В інформаційних технологіях часто використовують англійський термін "false positive" без перекладу.

Через можливість помилкових спрацьовувань не вдається повністю автоматизувати боротьбу з багатьма видами загроз. Як правило, ймовірність помилкового спрацьовування корелює з імовірністю пропуску події (помилки другого роду). Тобто, чим більш чутлива система, тим більше небезпечних подій вона виявляє, і, отже, запобігає. Але при підвищенні чутливості неминуче зростає і ймовірність помилкових спрацьовувань. Тому надто чутливо (параноїдально) налаштована система захисту може виродитися в свою протилежність, і призвести до того, що побічна шкода від неї буде перевищувати користь.

Відповідно, помилку другого роду іноді називають пропуском події або псевдонегативним спрацьовуванням. Людина хвора, але аналіз крові цього не показав, або у пасажира є холодна зброя, але рамка металодетектора його не виявлено (наприклад, через те, що чутливість рамки відрегульована на виявлення тільки дуже масивних металевих предметів). Дані приклади вказують на вчинення помилки другого роду. Слово «псевдонегативна» в даному випадку не має відношення до бажаності або небажаності самої події.

Термін широко використовується в медицині. Наприклад, тести,

призначені для діагностики захворювань, іноді дають негативний результат (тобто показують відсутність захворювання у пацієнта), коли насправді пацієнт страждає на це захворювання. Такий результат називається псевдонегативним.

На виробництві при контролі якості виробів помилку другого роду називають ризиком споживача.

Так як з ростом ймовірності помилки першого роду зазвичай зменшується ймовірність помилки другого роду, і навпаки, налаштування системи, що приймає рішення, повинне бути компромісним. Де саме знаходиться точка одержуваного таким чином балансу, залежить від оцінки наслідків при здійсненні обох видів помилок.

Процедура перевірки статистичних гіпотез

1. Обробка вибіркових даних і висунення основної та альтернативної гіпотез.

   До речі, до нуля нульова гіпотеза не має ніякого відношення, це просто історична назва, вона могла виявитися якою завгодно.

2. Вибір статистичного критерію K. Це безперервна випадкова величина, яка може приймати різні дійсні значення. У різних задачах критерії різні.

3. Вибір рівня значимості. Про дилему вибору цього значення сказано вище.

4. Знаходження критичного значення k_{krit} - це значення випадкової величини K, яке залежить від обраного рівня значимості та інших параметрів. Критичне значення визначає критичну область. Вона буває односторонньою (лівосторонньою чи правосторонньою) або двосторонньою (Рис.2.7).

   Критична область – це *область відхилення нульової гіпотези*. Незаштрихована область – *область прийняття гіпотези*.

5. Далі на основі вибіркових даних розраховується *спостережуване* значення критерію k_{look} і декларується висновок:

- Якщо k_{look} в критичну область *не потрапляє*, то гіпотеза H_0 на рівні значимості α приймається. Тут ми з ймовірністю α ризикуємо відкинути правильну гіпотезу. Однак не потрібно вважати, що нульова гіпотеза доведена і на 100% правильна, адже існує ймовірність β – того, ми зробили припуститися помилки 2-го роду (висунули невірну гіпотезу).
- Якщо k_{look} потрапляє в критичну область, то гіпотеза H_0 на рівні значимості α відхиляється (при цьому, якщо, наприклад $\alpha = 0.05$, то в середньому в 5-ти випадках із 100 ми відкинемо правильну гіпотезу, тобто зробимо помилку 1-го роду).

Питання та вправи для самостійного опрацювання

1. Охарактеризуйте поняття статистичної гіпотези.
2. Дайте визначення основної та альтернативної гіпотез.
3. Чому існує невизначеність?
4. Чому неможливо уникнути невизначеності?.
5. Що таке помилка першого роду?
6. Що таке помилка другого роду?
7. Наведіть алгоритм перевірки статистичних гіпотез?
8. Які використовують критерії для перевірки статистичних гіпотез?
9. Дайте визначення поняття ”рівень значимості”?
10. Дайте визначення поняття ”потужність критерію”?

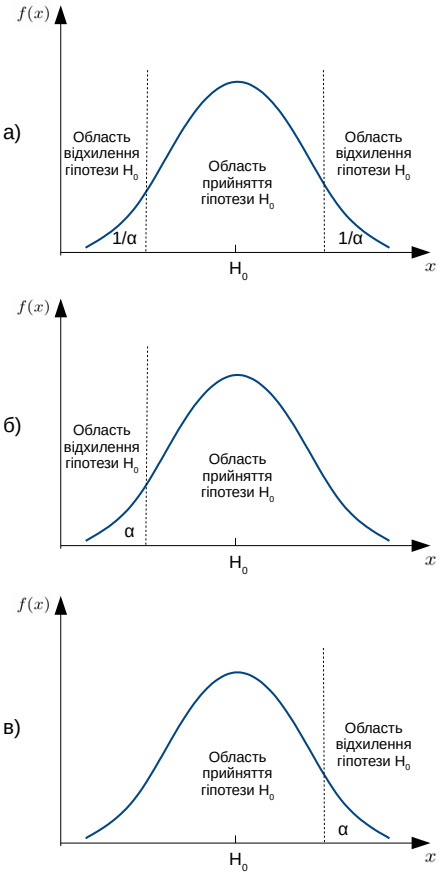


Рис. 2.5: Перевірка статистичних гіпотез з рівнем значимості α:
а) двохстороння перевірка;
б) лівостороння перевірка;
в) правостороння перевірка.

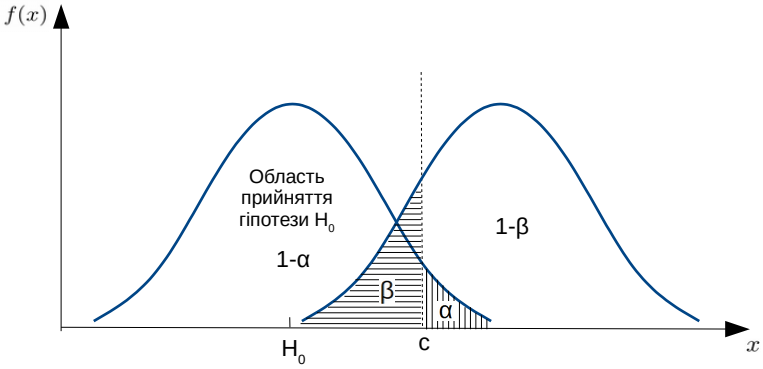


Рис. 2.6: Області існування помилок

Рис. 2.7:

Розділ 3

Перевірка значимості

Значення змінної називають статистично значущою, якщо мала ймовірність випадкового виникнення цієї або ще більш крайніх величин. Тут під крайністю розуміється ступінь відхилення тестової статистики від нульової гіпотези.

Різниця називається статистично значущою, якщо поява наявних даних (або ще більш крайніх даних) була б малоймовірно, якщо припустити, що ця різниця відсутня. *Цей вислів не означає, що дана різниця повинна бути велика, важлива, або значима в загальному розумінні цього слова.*

Загальна картина проблеми така: є вибірка з деякого простору Ω елементарних подій, і, можливо, значення на цій вибірці деяких змінних (функцій від $\omega \in \Omega$. Ймовірнісний розподіл на Ω не є відомим, а, навпаки, є головним об'єктом пошуку.

3.1. Парні дослідження

В дослідницькій та виробничій практиці часто важливо оцінити різницю між двома процесами. Висновки такого типу, як "чи дійсно каталізатор збільшує вихід продукту" є вирішальними в промисловості.

При виборі одного з двох методів аналізу треба встановити, який з них є більш достовірним, чи більш точним.

Часто необхідно з'ясувати відмінності, якщо вони є, між різними системами детектування, або між аналізами різних лабораторій одного й того ж зразка. Необхідним інструментом у цих дослідженнях є порівняльні експерименти та дисперсійний аналіз.

При порівнянні двох вибірок можливі три принципово різні варіанти:

1. Дисперсії обох вибірок однакові (на рівні гіпотези): $\sigma_1^2 = \sigma_2^2$;
2. Дисперсії двох вибірок різні (на рівні гіпотези): $\sigma_1^2 \neq \sigma_2^2$;
3. Дисперсії двох вибірок тотожно рівні (і ми впевнені в цьому): $\sigma_1^2 \equiv \sigma_2^2$;

Перші два випадки принципово відрізняються від третього в тому плані, що апріорі у нас немає інформації про співвідношення дисперсій двох вибірок. Тому є необхідність встановити цей факт.

3.1.1. Перевірка статистичної рівності дисперсій за допомогою F-тесту.

Нехай в нас є дві серії вимірювань: X_n і Y_m — відповідно n замірів величини X та m замірів величини Y.

Розраховується F-статистика, котра є співвідношення вибіркових дисперсій: $F = \frac{s_x^2}{s_y^2}$. Бажано, щоб значення F-статистики було більше одиниці (якщо це не так, міняєм вибірки місцями), в такому випадку спрощується логіка.

Порівнюєм одержану F-статистику з критичним значенням – коефіцієнтом F-розподілу – $F_{\alpha,n-1,m-1}$[1], де $(n-1)$ та $(m-1)$ – кількості

[1] В друкованій літературі для позначення індексів традиційно використовується рівень значимості α, тоді як у функціях – рівень достовірності, що відповідає параметру ймовірності відповідного квантіля.

Саме тому для позначення відповідних коефіцієнтів ми використовуємо традиційне позначення – рівень α, а в функціях – ймовірність. Надалі ми будемо надавати додаткові пояснення, коли вибір того чи іншого значення ймовірності буде не очевидним.

ступеней свободи вибірок з більшою та меншою дисперсіями.

- Якщо $F \leqslant F_{\alpha,n-1,m-1}$, то статистично дисперсії X і Y рівні;
- Якщо $F > F_{\alpha,n-1,m-1}$, то статистично дисперсії X і Y відрізняються.

Приклад 32 (Перевірка статистичної рівності дисперсійза допомогою F-тесту. Н
Визначити, чи відрізняються статистично дисперсії двох вибірок. Рівень значимості $\alpha = 0.05$.

251.3	260.4	278.7	285.2	256.2			
336	338.8	337.2	338.3	338.7	339.1	338.4	337.7

Розрахунок за допомогою R-статистики:

```
> x=c(251.3,260.4,278.7,285.2,256.2)
> y=c(336,338.8,337.2,338.3,338.7,339.1,338.4,337.7)
> var.test(x,y)

F test to compare two variances

data:  x and y
F = 208.8, num df = 4, denom df = 7, p-value = 4.681e-07
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
37.8078 1894.6546
sample estimates:
ratio of variances
208.7971
> qf(0.95,4,7)
[1] 4.120312
```

$F = 208.8 > F_{0.05,4,7} = 4.120$

Відповідь: дисперсії статистично відрізняються з достовірністю більше 95%.

Подальший розрахунок залежить від того, який висновок зроблений.

3.1.2. Випадок статистичної рівності дисперсій

Якщо дисперсії статистично рівні, то наявність значимої різниці виявляється за допомогою перевірочної t-статистики:

$$t_\nu = \frac{|\bar{x} - \bar{y}| \cdot \sqrt{n + m - 2}}{\sqrt{\frac{1}{n} + \frac{1}{m}} \cdot \sqrt{(n-1) \cdot s_x^2 + (m-1) \cdot s_y^2}} \tag{3.1}$$

де $\nu = n + m - 2$ – кількість ступенів свободи;

t_ν порівнюється з відповідним коефіцієнтом Стьюдента $t_{\alpha,\nu}$ (t-критичне) з рівнем недостовірності (α), та кількістю ступенів свободи ν.

- Якщо $t > t_{\alpha,\nu}$, то між серіями вимірювань є статистична різниця;
- Якщо $t < t_{\alpha,\nu}$, то між серіями вимірювань немає статистичної різниці.

Приклад 33 (Порівняння вибірок у випадку статистичної рівності дисперсій) Визачити, чи є статистична різниця між вибірками. Рівень значимості $\alpha = 0.05$.

222,1	249,5	222,3	252,4	247,2	232,2	222,5	
251,6	273,9	271,1	260,6	271,5	259,9	290,2	266,7

Розрахунок за допомогою R-статистики. Перевірка статистичної рівності дисперсій:

```
> x=c(222.1,249.5,222.3,252.4,247.2,232.2,222.5)
> y=c(251.6,273.9,271.1,260.6,271.5,259.9,290.2,266.7)
> var.test(x,y)

F test to compare two variances

data:  x and y
F = 1.4296, num df = 6, denom df = 7, p-value = 0.6466
```

alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.2793009 8.1424085
sample estimates:
ratio of variances
1.429629

> qf(0.95,6,7)
[1] 3.865969

$F = 1.430 < F_{0.05,6,7} = 3.866$

Висновок: дисперсії статистично не відрізняються з достовірністю більше 95%.

Розрахунок за допомогою R-статистики. Перевірка наявності статистичної різниці між вибірками за умови рівності дисперсій:

> t.test(x,y,var.equal=TRUE)

Two Sample t-test

data: x and y
t = -4.9848, df = 13, p-value = 0.0002496
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-46.91547 -18.54524
sample estimates:
mean of x mean of y
235.4571 268.1875

> qt(0.975,13)
[1] 2.160369

$|t| = 4.985 > t_{0.05,13} = 2.160$

Відповідь: вибірки статистично відрізняються з достовірністю більше 95%.

3.1.3. Випадок статистичної нерівності дисперсій

Якщо дисперсії статистично різні, то наявність значимої різниці виявляється за допомогою перевірочної t-статистики:

$$t_\nu = \frac{|\bar{x} - \bar{y}| \cdot \sqrt{n + m - 2}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}} \tag{3.2}$$

де кількість ступенів свободи розраховується за формулою:

$$\nu = \left[\frac{\left(\frac{s_x^2}{n} + \frac{s_y^2}{n} \right)^2}{\frac{1}{n-1} \cdot \left(\frac{s_x^2}{n} \right)^2 + \frac{1}{m-1} \cdot \left(\frac{s_y^2}{m} \right)^2} - 2 \right] \tag{3.3}$$

t_ν порівнюється з відповідним коефіцієнтом Стьюдента $t_{\alpha,\nu}$ (t-критичне) з рівнем недостовірності (α), та кількістю ступенів свободи ν.

- Якщо $t > t_{\alpha,\nu}$, то між серіями вимірювань є статистична різниця;
- Якщо $t < t_{\alpha,\nu}$, то між серіями вимірювань немає статистичної різниці.

Приклад 34 (Порівняння вибірок у випадку статистичної нерівності дисперсій) Визачити, чи є статистична різниця між вибірками. Рівень значимості $\alpha = 0.05$.

54	57,7	57,3	55,5	57,3	55,4	60,4	56,5
52,2	65,4	52,3	66,8	64,3	57,1	52,4	

Розрахунок за допомогою R-статистики. Перевірка статистичної рівності дисперсій:

```
> x=c(54,57.7,57.3,55.5,57.3,55.4,60.4,56.5)
> y=c(52.2,65.4,52.3,66.8,64.3,57.1,52.4)
> var.test(x,y)
```

```
F test to compare two variances

data:  x and y
F = 0.082857, num df = 7, denom df = 6, p-value = 0.004367
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.01454792 0.42411272
sample estimates:
ratio of variances
0.08285723
```

F-статистика менше одиниці. Міняємо вибірки місцями:

```
> var.test(y,x)

F test to compare two variances

data:  y and x
F = 12.069, num df = 6, denom df = 7, p-value = 0.004367
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
2.357864 68.738367
sample estimates:
ratio of variances
12.06895

> qf(0.95,6,7)
[1] 3.865969
```

$F = 12.07 > F_{0.05,6,7} = 3.866$

Висновок: дисперсії статистично відрізняються з достовірністю більше 95%.

Розрахунок за допомогою R-статистики. Перевірка наявності статистичної різниці між вибірками за умови нерівності дисперсій:

```
> t.test(x,y,var.equal=FALSE)
```

```
Welch Two Sample t-test

data:  x and y
t = -0.71973, df = 6.8706, p-value = 0.4954
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-8.081818  4.321103
sample estimates:
mean of x mean of y
56.76250  58.64286

> qt(0.975,7)
[1] 2.364624
```

$|t| = 0.7198 < t_{0.05,7} = 2.365$

Кількість ступенів свободи округлюємо.

Відповідь: вибірки статистично не відрізняються з достовірністю більше 95%.

3.1.4. Дисперсії двох вибірок тотожно рівні

Цей випадок зустрічається дуже часто на практиці. Щоб дослідити дію певного фактору на систему, частину об'єктів спостерігають в присутності цього фактору, а частину – за його відсутності при умові що всі інші фактори залишаються постійними. Це досить росповсюджений прийом при дослідженнях, на виробництві та при контролі якості. Наприклад, можна перевірити дію ліків на частоту серцебиття, каталізаторів на швидкість реакції, вплив певної стадії процесу на якість продукту, вплив кількості пластифікатора на відгук іонселективного електроду, вплив реагенту на інтенсивність поглинання в УФ-спектрі. Такі задачі звичайно називають порівнянням з холостим дослідом.

Третій випадок відрізняється від попередніх тим, що тут дисперсія апріорі однакова й немає сенсу її перевіряти. Крім того, завжди $n = m$.

Перевірочна статистика визначається формулою:

$$t_{n-1} = \frac{\bar{d}}{s_d} \cdot \sqrt{n} \tag{3.4}$$

де d_i – середнє значень $d_i = x_i - y_i$;
s_d – стандартне відхилення d_i.

t_{n-1} порівнюється з відповідним коефіцієнтом Стьюдента $t_{\alpha,n-1}$ (t-критичне) з рівнем недостовірності (α), та кількістю ступенів свободи $n - 1$.

- Якщо $t_{n-1} > t_{\alpha,n-1}$, то між серіями вимірювань є статистична різниця;
- Якщо $t_{n-1} < t_{\alpha,n-1}$, то між серіями вимірювань немає статистичної різниці.

Приклад 35 (Порівняння вибірок у парних дослідженнях)
Визачити, чи є статистична різниця між вибірками. Проводились парні дослідження. Рівень значимості $\alpha = 0.05$.

91	97,9	97	93,8	97,2	93,6	103
97	108,8	107,3	101,7	107,5	101,4	117,3

Розрахунок за допомогою R-статистики:

```
> x=c(91,97.9,97,93.8,97.2,93.6,103)
> y=c(97,108.8,107.3,101.7,107.5,101.4,117.3)
> t.test(x,y,paired=TRUE)

Paired t-test

data:  x and y
t = -9.4386, df = 6, p-value = 8.044e-05
alternative hypothesis: true difference in means is not equal to 0
```

95 percent confidence interval:
-12.142712 -7.143002
sample estimates:
mean of the differences
-9.642857

> qt(0.975,6)
[1] 2.446912

$|t| = 9.439 > t_{0.05,6} = 2.447$

Відповідь: вибірки статистично відрізняються з достовірністю більше 95%.

3.2. Дисперсійний аналіз

Будь-який об'єкт має безліч властивостей або ознак. Якщо ми вибрали певний набір вимірюваних властивостей або ознак, то вихідні дані можна представити у вигляді матриці $X = \{x_{ij}\}$ (див.стор. 158), де i - номер об'єкта, j - номер ознаки.

Які саме ознаки нас цікавлять, залежить від цілей і завдань конкретного дослідження. Вибір кінцевого набору ознак – не таке вже й просте завдання. У багатомірній системі частина ознак може виявитися неінформативними, їх відкидання не вплине на якість одержуваної інформації. Але може виявитися і навпаки, що ми не включим у розгляд важливу властивість об'єкта, котра впливає на зміну інших ознак. Питання інформативності ознаки можна вирішити і після первинного збору даних. При розробці конкретної методики вимірювань перевірка ознак на інформативність є одним з перших етапів оптимізації. Саме для цього й призначений дисперсійний аналіз.

3.2.1. Задача дисперсійного аналізу

Дисперсійний аналіз – метод в математичній статистиці, спрямований на пошук залежностей в експериментальних даних шляхом до-

слідження значущості відмінностей в середніх значеннях. На відміну від t-критерія, дозволяє порівнювати середні значення трьох і більше груп. Цей метод був розроблений Рональдом Фішером для аналізу результатів експериментальних досліджень. Для позначення дисперсійного аналізу також часто використовують англійську абревіатуру ANOVA (від англ. **AN**alysis **O**f **VA**riance).

> Дисперсійний аналіз – метод дослідження значимості відмінностей в середніх значеннях, котрий грунтується на розкладі дисперсії на компоненти – систематичну та випадкову дисперсії.

У будь-якому досліді середні значення величин, що спостерігаються, змінюються зі зміною основних факторів (кількісних та якісних), які визначають умови досліду, та випадкових факторів. Завданням дисперсійного аналізу є дослідження наявності впливу тих чи інших факторів на результати дослідів.

У дисперсійному аналізі досліджується вклад різних компонентів дисперсії (різних джерел похибок) в загальну дисперсію (використовується властивість адитивності дисперсії).

Детальний розрахунок дисперсійного аналізу ми тут наводити не будем. При бажанні можна звернутись до джерел . Сучасні статистичні пакети зразу видають результати розрахунків у вигляді таблиці дисперсійного аналізу.

Для коректних висновків дані для дисперсійного аналізу мають задовільняти наступним умовам:

- нормальний розподіл значень досліджуваних ознак;
- рівність дисперсій в порівнюваних ознак;
- випадковий і незалежний характер вибірок.

Дані, призначені для дисперсійного аналізу, звичайно представляють у вигляді таблиці, котра має назву *статистичний комплекс*.

> Статистичний комплекс – таблиця емпіричних даних.

Якщо у всіх класах градацій однакова кількість варіантів, то статистичний комплекс називається однорідним (гомогенним), якщо число варіантів різне -– різнорідним (гетерогенним). Однофакторний дисперсійний аналіз допускає використання гомогенних та гетерогенних статистичних комплексів, двофакторний – тільки гомогенних.

3.2.2. Однофакторний дисперсійний аналіз

Однофакторний дисперсійний аналіз використовується в тих випадках коли потрібно оцінити вплив (значимий чи не значимий) одного фактора на результати експерименту.

Статистичний комплекс для однофакторного дисперсійного аналізу повинен мати наступну структуру Табл. 3.2.2.

Табл. 3.1: Формат даних для однофакторного дисперсійного аналізу

Рівні фактору A			
a_1	a_2	$\dots$	a_k
y_{11}	y_{21}	$\cdots$	y_{k1}
y_{12}	y_{22}	$\cdots$	y_{k2}
$\vdots$	$\vdots$	y_{im_i}	$\vdots$
y_{1m_1}	y_{2m_2}	$\cdots$	y_{km_k}

Фактор A виміряний на k рівнях. На кожному рівні було зроблено m_i вимірювань.

Після проведення однофакторного диспесійного аналізу одержують дані, які звичайно зведені в таблицю наступної структури (Табл. 3.2.2):

Розглянемо структуру таблиці однофакторного дисперсійного аналізу. Як вже було сказано вище, детальний розрахунок сум квадратів на основі вибіркових значень ми наводити не будемо.

- В першому стовчику наводиться джерело варіації.

Табл. 3.2: Таблиця результатів однофакторного дисперсійного аналізу

Джерело варіації	Ступені свободи	Сума квадратів	Середні квадрати	F-статистика	Значимість
Фактор A	df_A	SS_A	MS_A	F_A	p_A
Залишки	df_{res}	SS_{res} MS_{res}			

- В другому стовпчику – кількість ступенів свободи. Кількість ступенів свободи для фактору A дорівнює кількості рівнів фактору A мінус одиниця: $df_A = k - 1$. Залишкова кількість ступенів свободи дорівнює різниці між загальною кількістю ступенів свободи $n - 1$ та кількісті ступенів свободи для фактору A: $df_{res} = n - df_A - 1$.
- В третьому стовчику знаходяться відповідні суми квадратів.
- В четвертому стовпчику знаходяться відповідні середні квадрати. Вони мають розмірність дисперсії. Розраховуються з даних другого та третього стовпчика – відповідні суми квадратів діляться на відповідні ступені свободи: $MS_A = \frac{SS_A}{df_A}$, $MS_{res} = \frac{SS_{res}}{df_{res}}$.
- В п'ятому стовпчику знаходиться значення F-статистики, котра рівна співвідношенню середніх квадратів для фактору A та залишку: $F_A = \frac{MS_A}{MS_{res}}$.
- В шостому стовпчику знаходиться значення, котре показує ймовірність істинності нульової гіпотези. Іншими словами, це ймовірність того, що досліджуваний фактор не впливає на результати експерименту. За цим значення можна безпосередньо оцінювати значимість F-статистики без порівняння з критичним значенням.

Для кращого розуміння наведемо процедуру формулювання висновків з однофакторного дисперсійного аналізу на основі порівняння розрахованої F-статистики з квинтилями (коефіцієнтами) F-розподілу:

- Якщо $F_A < F_{\alpha,df_A,df_{res}}$, значить фактор A є не значимим (результати експериментів від нього не залежать з достовірністю $1–\alpha$);
- Якщо $F_A > F_{\alpha,df_A,df_{res}}$, значить фактор A є значимим (результати експериментів залежать від нього з достовірністю $1–\alpha$).

Приклад 36 (Приклад однофакторного дисперсійного аналізу) Завдання — визначити, чи впливає природа хлоралкану на вихід продукту в реакції, в таблиці наведені значення виходу продукту (%).

метилхлорид	етилхлорид	пропілхлорид	бутилхлорид	пентилхлорид
11	21	22	31	35
12	22	23	30	36
11	23	25	32	34
13	22	24	31	37

Проведення однофакторного дисперсійного аналізу за допомогою R-статистики:

Встановлюємо робочу папку в R-статистиці:

```
setwd("Шлях до вашої папки")
```

Створюємо у цій папці файл disp1.txt, де вихідні дані записуємо в наступному вигляді[2]:

```
"Хлоралкан" "Вихід"
метилхлорид 11
метилхлорид 12
. . . . . .
пентилхлорид 34
пентилхлорид 37
```

Читаємо файл:

```
data1=read.table("disp1.txt",header=TRUE)
```

[2]Частина вмісту файла опущено. Показано зразок набору

Будуємо діаграму розмаху:

```
plot(data1)
```

Проводимо однофакторний дисперсійний аналіз:

```
res1=aov(Вихід~Хлоралкан,data1)
```

Виводимо результати:

```
summary(res1)
```

Одержуємо:

```
> summary(res1)
            Df Sum Sq Mean Sq F value   Pr(>F)
Хлоралкан    4 1331.0   332.7     298 4.29e-14 ***
Residuals   15   16.8     1.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Розраховуємо критичні значення F:

```
> qf(0.95,4,15)
[1] 3.055568)
```

$F = 298 > F_{0.05,4,15} = 3.056$

Відповідь: природа хлоралкану впливає на результати експериментів із достовірністю більше 95%.

3.2.3. Двофакторний дисперсійний аналіз

Двофакторний дисперсійний аналіз використовується в тих випадках коли потрібно оцінити вплив двох різних факторів на результати експерименту.

Статистичний комплекс для двофакторного дисперсійного аналізу повинен мати наступну структуру (Табл. 3.2.3):

Табл. 3.3: Статистичний комплекс для двофакторного дисперсійного аналізу

Фактори	Рівні фактору А			
Рівні фактору В	a_1	a_2	$\dots$	a_k
b_1	y_{11}	y_{21}	$\dots$	y_{k1}
b_2	y_{12}	y_{22}	$\dots$	y_{k2}
$\vdots$	$\vdots$	$\vdots$	y_{ij}	$\vdots$
b_m	y_{1m}	y_{2m}	$\dots$	y_{km}

Фактор A виміряний на k рівнях. Фактор B виміряний на m рівнях.

Після проведення двофакторного диспесійного аналізу одержують дані, які звичайно зведені в таблицю наступної струкрури (Табл. 3.2.3).

Табл. 3.4: Таблиця двофакторного дисперсійного аналізу

Джерело варіації	Ступені свободи	Сума квадратів	Середні квадрати	F-статистика	Значимість
Фактор А	df_A	SS_A	MS_A	F_A	p_A
Фактор В	df_B	SS_B	MS_B	F_B	p_B
Залишки	df_{res}	SS_{res}			

Розглянемо структуру таблиці двофакторного дисперсійного аналізу. Як вже було сказано вище, детальний розрахунок сум квадратів на основі вибіркових значень ми наводити не будемо.

- В першому стовчику наводиться джерело варіації.
- В другому стовпчику – кількість ступенів свободи. Кількість ступенів свободи для фактору A дорівнює кількості рівнів фактору A мінус одиниця: $df_A = k - 1$. Кількість ступенів свободи для фактору B дорівнює кількості рівнів фактору B мінус одиниця: $df_B = m - 1$. Залишкова кількість ступенів свободи дорівнює різниці між загальною кількістю ступенів свобо-

ди $n-1$ та кількості ступенів свободи для факторів A та B: $df_{res} = n - df_A - df_B - 1$.

- В третьому стовчику знаходяться відповідні суми квадратів.
- В четвертому стовпчику знаходяться відповідні середні квадрати. Вони мають розмірність дисперсії. Розраховуються з даних другого та третього стовпчика – відповідні суми квадратів діляться на відповідні ступені свободи: $MS_A = \frac{SS_A}{df_A}$, $MS_B = \frac{SS_B}{df_B}$, $MS_{res} = \frac{SS_{res}}{df_{res}}$.
- В п'ятому стовпчику знаходиться значення F-статистики, котра рівна співвідношенню середніх квадратів. Для фактору A – це $F_A = \frac{MS_A}{MS_{res}}$. Для фактору B – це $F_B = \frac{MS_B}{MS_{res}}$.
- В шостому стовпчику знаходяться значення, котрі показують ймовірність істинності нульової гіпотези. Іншими словами, це ймовірність того, що досліджуваний фактор не впливає на результати експерименту. За цим значення можна безпосередньо оцінювати значимість F-статистики без порівняння з критичним значенням.

Для кращого розуміння наведемо процедуру формулювання висновків з двофакторного дисперсійного аналізу на основі порівняння розрахованої F-статистики з квинтилями (коефіцієнтами) F-розподілу:

- Якщо $F_A < F_{\alpha, df_A, df_{res}}$, значить фактор A є не значимим (результати експериментів від нього не залежать з достовірністю 1–α);
- Якщо $F_A > F_{\alpha, df_A, df_{res}}$, значить фактор A є значимим (результати експериментів залежать від нього з достовірністю 1–α).
- Якщо $F_B < F_{\alpha, df_B, df_{res}}$, значить фактор B є не значимим (результати експериментів від нього не залежать з достовірністю 1–α);

- Якщо $F_B > F_{\alpha,df_B,df_{res}}$, значить фактор B є значимим (результати експериментів залежать від нього з достовірністю $1–\alpha$).

Приклад 37 (Приклад двофакторного дисперсійного аналізу) Завдання – визначити, чи впливає природа розчинника та температура на ступінь перетворення.

	30°C	40°C	50°C	60°C
гексан	0,2	0,22	0,23	0,25
ацетон	0,95	0,98	0,98	0,99
хлороформ	0,86	0,88	0,91	0,95
бензен	0,32	0,33	0,33	0,35
етанол	0,78	0,8	0,81	0,84

Створюємо у своїй папці файл disp2.txt, де вихідні дані записуємо в наступному вигляді[3]:

```
”Температура” ”Розчинник” ”Ступінь перетворення”
30°C гексан 0.2
30°C ацетон 0.95
30°C хлороформ 0.86
. . . . . .
60°C хлороформ 0.95
60°C бензен 0.35
60°C етанол 0.84
```

Читаємо файл:

```
data2=read.table(”disp2.txt”,header=TRUE)
```

Будуємо діаграму розмаху:

```
plot(data2)
```

Проводимо двофакторний дисперсійний аналіз:

```
res2=aov(Ступінь.перетворення~Температура+Розчинник,data2)
```

[3]Частина вмісту файла опущено. Показано зразок набору

Виводимо результати:

```
> summary(res2)
            Df Sum Sq Mean Sq F value   Pr(>F)
Температура  3 0.0076  0.0025   19.02 7.45e-05 ***
Розчинник    4 1.8974  0.4743 3579.94  < 2e-16 ***
Residuals   12 0.0016  0.0001
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Розраховуємо критичні значення F:

```
> qf(0.95,3,12)
[1] 3.490295
> qf(0.95,4,12)
[1] 3.259167
```

Для температури: $F = 19.02 > F_{0.05,3,12} = 3.490$.

Для розчинника: $F = 3579.94 > F_{0.05,3,12} = 3.259$.

Відповідь: температура та природа розчинника впливають на ступінь перетворення з достовірністю більше 95%.

3.2.4. Двофакторний дисперсійний аналіз із повтореннями

Двофакторний дисперсійний аналіз із повтореннями використовується в тих випадках коли потрібно оцінити вплив двох різних факторів на результати експерименту та визначити, чи існує взаємний вплив між факторами.

Статистичний комплекс двофакторного дисперсійного аналізу з повтореннями повинен мати наступний формат (Табл. 3.2.4).

Фактор A виміряний на k рівнях. Фактор B виміряний на m рівнях. В кожній точні проведено l вимірювань.

Після проведення двофакторного диспесійного аналізу з повтореннями одержують дані, які звичайно зведені в таблицю наступної струкрури (Табл. 3.2.4).

Табл. 3.5: Статистичний комплекс двофакторного дисперсійного аналізу з повтореннями

Фактори	Рівні фактору А			
Рівні фактору В	a_1	a_2	$\dots$	a_k
b_1	y_{111} y_{112} $\vdots$ y_{11l}	y_{121} y_{122} $\vdots$ y_{12l}	$\cdots$	y_{1k1} y_{1k2} $\vdots$ y_{1kl}
b_2	y_{211} y_{212} $\vdots$ y_{21l}	y_{221} y_{222} $\vdots$ y_{22l}	$\cdots$	y_{2k1} y_{2k2} $\vdots$ y_{2kl}
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$
b_m	y_{m11} y_{m12} $\vdots$ y_{m1l}	y_{m21} y_{m22} $\vdots$ y_{m2l}	$\cdots$	y_{mk1} y_{mk2} $\vdots$ y_{mkl}

Розглянемо структуру таблиці двофакторного дисперсійного аналізу з повтореннями. Як вже було сказано вище, детальний розрахунок сум квадратів на основі вибіркових значень ми наводити не будемо.

- В першому стовчику наводиться джерело варіації.
- В другому стовпчику – кількість ступенів свободи. Кількість ступенів свободи для фактору A дорівнює кількості рівнів фактору A мінус одиниця: $df_A = k-1$. Кількість ступенів свободи для фактору B дорівнює кількості рівнів фактору B мінус одиниця: $df_B = m-1$. Кількість ступенів свободи для взаємодії факторів A і B дорівнює добутку ступенів свободи цих факторів: $df_{AB} = df_A \cdot df_B$. Залишкова кількість ступенів свободи дорівнює різниці між загальною кількістю ступенів свободи $n-1$

Табл. 3.6: Таблиця двофакторного диспесійного аналізу з повтореннями

Джерело варіації	Ступені свободи	Сума квадратів	Середні квадрати	F-статистика	Значимість
Фактор A	df_A	SS_A	MS_A	F_A	p_A
Фактор B	df_B	SS_B	MS_B	F_B	p_B
Взаємодія A:B	df_{AB}	SS_{AB}	MS_{AB}	F_{AB}	p_{AB}
Залишки	df_{res}	SS_{res}			

та кількості ступенів свободи для факторів A, B та їх взаємодії: $df_{res} = n - df_A - df_B - df_{AB} - 1$.

- В третьому стовчику знаходяться відповідні суми квадратів.
- В четвертому стовпчику знаходяться відповідні середні квадрати. Вони мають розмірність дисперсії. Розраховуються з даних другого та третього стовпчика – відповідні суми квадратів діляться на відповідні ступені свободи: $MS_A = \frac{SS_A}{df_A}$, $MS_B = \frac{SS_B}{df_B}$, $MS_{AB} = \frac{SS_{AB}}{df_{AB}}$, $MS_{res} = \frac{SS_{res}}{df_{res}}$.
- В п'ятому стовпчику знаходиться значення F-статистики, котра рівна співвідношенню середніх квадратів. Для фактору A – це $F_A = \frac{MS_A}{MS_{res}}$. Для фактору B – це $F_B = \frac{MS_B}{MS_{res}}$. Для взаємодії факторів – це $F_{AB} = \frac{MS_{AB}}{MS_{res}}$
- В шостому стовпчику знаходяться значення, котрі показують ймовірність істинності нульової гіпотези. Іншими словами, це ймовірність того, що досліджуваний фактор не впливає на результати експерименту. За цим значення можна безпосередньо оцінювати значимість F-статистики без порівняння з критичним значенням.

Для кращого розуміння наведемо процедуру формулювання висновків з двофакторного дисперсійного аналізу на основі порівняння розрахованої F-статистики з квинтилями (коефіцієнтами) F-розподілу:

- Якщо $F_A < F_{\alpha,df_A,df_{res}}$, значить фактор A є не значимим (результати експериментів від нього не залежать з достовірністю $1–\alpha$);
- Якщо $F_A > F_{\alpha,df_A,df_{res}}$, значить фактор A є значимим (результати експериментів залежать від нього з достовірністю $1–\alpha$).
- Якщо $F_B < F_{\alpha,df_B,df_{res}}$, значить фактор B є не значимим (результати експериментів від нього не залежать з достовірністю $1–\alpha$);
- Якщо $F_B > F_{\alpha,df_B,df_{res}}$, значить фактор B є значимим (результати експериментів залежать від нього з достовірністю $1–\alpha$).
- Якщо $F_{AB} < F_{\alpha,df_{AB},df_{res}}$, значить взаємодія факторів A і B є не значимою (взаємодія факторів не проявляється з достовірністю $1–\alpha$);
- Якщо $F_{AB} > F_{\alpha,df_{AB},df_{res}}$, значить взаємодія факторів A і B є значимою (взаємодія факторів проявляється з достовірністю $1–\alpha$).

Приклад 38 (Приклад двофакторного дисперсійного аналізу з повтореннями) Завдання – визначити, чи впливає природа розчинника та температура на ступінь перетворення, а також, чи є взаємний вплив між температурою та розчинником.

	30°C	40°C	50°C	60°C
гексан	0,2	0,23	0,23	0,24
	0,22	0,25	0,22	0,26
	0,21	0,22	0,22	0,25
ацетон	0,95	0,96	0,96	0,97
	0,98	0,95	0,98	0,99
	0,94	0,96	0,97	0,98
хлороформ	0,86	0,85	0,86	0,87
	0,85	0,86	0,88	0,86
	0,83	0,84	0,85	0,89
бензен	0,32	0,33	0,33	0,34
	0,33	0,34	0,35	0,33
	0,31	0,34	0,34	0,31
етанол	0,79	0,79	0,81	0,81
	0,8	0,78	0,82	0,8
	0,78	0,77	0,83	0,82

Створюємо у своїй папці файл disp3.txt, де вихідні дані записуємо в наступному вигляді[4]:

```
”Температура” ”Розчинник” ”Ступінь перетворення”
30°C гексан 0.2
30°C гексан 0.22
30°C гексан 0.21
30°C ацетон 0.95
. . . . . .
60°C бензен 0.31
60°C етанол 0.81
60°C етанол 0.8
60°C етанол 0.82
```

Читаємо файл:

```
data3=read.table(”disp3.txt”,header=TRUE)
```

[4]Частина вмісту файла опущено. Показано зразок набору

Будуємо діаграму розмаху:

```
plot(data3)
```

Проводимо двофакторний дисперсійний аналіз з повтореннями:

```
res3=aov(Ступінь.перетворення~Температура*Розчинник,data3)
```

Виводимо результати:

```
> summary(res3)
                    Df Sum Sq Mean Sq  F value  Pr(>F)
Температура          3  0.005  0.0017   12.329 7.5e-06 ***
Розчинник            4  5.325  1.3313 9509.030 < 2e-16 ***
Температура:Розчинник 12  0.004  0.0003    2.181  0.0325 *
Residuals           40  0.006  0.0001
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Розраховуємо критичні значення F:

```
> qf(0.95,3,40)
[1] 2.838745
> qf(0.95,4,40)
[1] 2.605975
> qf(0.95,12,40)
[1] 2.003459
```

Для температури: $F = 12.329 > F_{0.05,3,40}$.

Для розчинника: $F = 9509 > F_{0.05,4,40}$.

Для взаємодії факторів температури та розчинника: $F = 2.181 > F_{0.05,12,40}$.

Відповідь: температура та природа розчинника впливають на ступінь перетворення з достовірністю більше 95%, існує значима взаємодія між факторами температури та розчинника з достовірністю більше 95%.

Питання та вправи для самостійного опрацювання

1. Дайте визначення дисперсійного аналізу.
2. Охарактеризуйте основні види дисперсійного аналізу.
3. Охарактеризуйте логіку дисперсійного аналізу.
4. Які критерії використовують при дисперсійному аналізі.
5. Як рахуються ступені свободи в дисперсійному аналізі.
6. Який формат повинні мати дані для проведення різних видів дисперсійного аналізу?
7. Вимоги до змінних при проведенні дисперсійного аналізу?
8. Форма видачі результатів дисперсійного аналізу?
9. Структура даних для двофакторного дисперсійного аналізу?
10. Структура даних для однофакторного дисперсійного аналізу?

Розділ 4

Аналіз багатомірних даних

Існує велика кількість підходів до аналізу багатовимірних даних. Залежно від того, яка саме відповідь нам потрібна, проводиться й вибір методу. Вкрай важливо сформулювати завдання змістовно, так, щоб мета аналізу була гранично зрозумілою, а також представити дані у вигляді, найбільш зручному для досягнення поставленої мети. Це не завжди просто зробити. Однак якщо завдання і мета чітко сформульовані, вибір конкретної методики звичайно не становить труднощів.

Значна частина багатовимірного аналізу пов'язана з простим «розгляданням» даних, описом їх деякими сумарними характеристиками і дуже часто поданням прихованих структур за допомогою відповідних графіків. такими методами можна вивчати дані, які є змінними стану якогось контрольованого технологічного процесу, наприклад, органічного синтезу (температура, вміст компонентів і т.д.), тобто деяку p-мірну характеристику n зразків.

Цілі як одновимірного, так і багатовимірного опису даних можуть бути різноманітними: від визначення середніх значень і стандартних відхилень до визначення кореляцій і побудови регресійних моделей.

Наприклад, в разі органічного синтезу природним було б подивитися, які саме змінні впливають на вихід процукту в цілому або на селективність виходу. Дані, отримані при синтезі, можна використовувати і для того, щоб відповісти на наступні питання: яка кореляція між температурою і виходом продукту? Чи впливає інтенсивність перегонки на чистоту продукту?

Задачі аналізу багатомірних даних можна умовно розділити на дві групи: аналіз зв'язку та аналіз структури:

4.0.1. Аналіз зв'язку

Аналіз зв'язку передбачає, що мають бути розрахованими величини, котрі характеризують зв'язки між окремими елементами. Звичайно це значення певних критеріїв або параметрів моделей.

- Змінні:
  - вхідні (незалежні);
  - вихідні (залежні);
- Мета:
  - виявлення ступеня та характеру зв'язку даних;
- Види аналізу:
  - дисперсійний аналіз;
  - регресійний аналіз;
  - дискримінаційний аналіз;
  - коваріаційний аналіз;

4.0.2. Аналіз структури

Аналіз структури передбачає розрахунок взаємовідносин окремих елементів між собою. І на основі цих взаємовідносин будується візуальне представлення даних.

- Змінні:
  - немає розділу змінних;
- Мета:
  - здобуття більш стислого та наочного представлення даних на основі існуючих зв’язків;
- Види аналізу:
  - факторний аналіз;
  - кластерний аналіз;
  - метод головних компонент;

4.1. Моделі та їх оцінка

На будь-яку хімічну систему (Рис. 4.1) діє багато факторів, на котрі вона реагує. Багато реакцій на вплив можна виміряти.

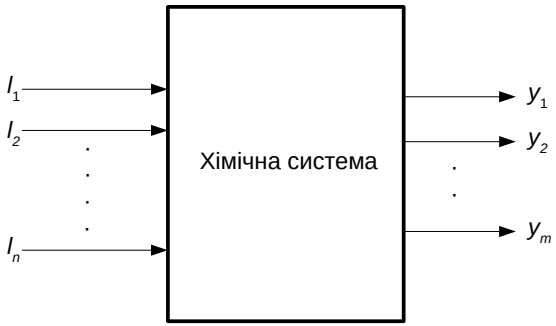


Рис. 4.1: Схема хімічної системи зі входами та виходами

Кожен результат можна представити функцією параметрів що діють на систему:

$$y_i = \varphi_i(l_1, l_2, \ldots, l_n) \tag{4.1}$$

Це — *функція відгуку*, котра описує певну *поверхню відгуку* в $n+1$ мірному просторі (в 3-х мірному — просто поверхню).

Звичайно, щоб отримати інформацію про хімічну систему, бажано знайти таку функцію. Один з методів — побудова моделей.

Моделювання – це такий метод дослідження об'єктів, коли замість оригіналу експерименти проводять на іншому об'єкті – на моделі, і результати кількісно переносять на оригінал. На підставі дослідів на моделях встановлюють найважливіші для досліджуваного об'єкта закономірності, що дає змогу передбачити його поведінку в конкретних робочих умовах. Перспективність такого підходу полягає в тому, що відкриваються можливості для виявлення найхарактерніших сторін та властивостей досліджуваного процесу.

Модель повинна відповідати наступним вимогам :

- експеримент на моделі повинен проводитись швидше, простіше, бути вигіднішим та безпечнішим, ніж дослідження самого оригіналу;
- модель має адекватно відображати поведінку оригіналу в умовах проведення експерименту;
- має бути відоме правило, за допомогою якого результати вивчення моделі будуть перенесені на оригінал.

Модель — об'єкт замінник об'єкту–оригіналу, що забезпечує вивчення деяких властивостей оригіналу.

В хімії найчастіше модель — це математичне рівняння, що пов'язує результати з експериментальними параметрами.

Моделі можуть бути *теоретичними* або *емпіричними*.

Теоретична модель — явище зрозуміле, результати контролюються відомою схемою.

Приклад 39 (Закон Ламберта-Бугера-Бера)
Оптична густина розчину пропорційна коефіцієнту поглинання, концентрації та товщині шару рідини.

$$A = \epsilon \cdot l \cdot C$$

де A – оптична густина розчину;
ϵ – молярний коефіцієнт поглинання;
l – товщина шару розчину;
C – концентрація розчину.

Емпірична модель — відгуки проявляють складну поведінку, яку не можна просто пояснити.

> Емпірична модель – модель, основу якої складають результати аналізу деякого обсягу даних, отриманих в результаті експерименту або вимірювань.

Результатом аналізу емпіричних даних, як правило, є нові формули, рівняння, закономірності, кореляційні залежності, що описують зв'язок між розглянутими величинами. Результатом також може бути деякий масив даних, що представляє собою еталон, з яким в подальшому будуть порівнюватися подібні експериментальні дані.

В цьому випадку використовують наближені моделі. Одним з методів пошуку та уточнення моделей є регресійний аналіз.

4.2. Регресійний аналіз

Регресійний аналіз використовується в наступних задачах:

- побудова калібрувальних залежностей;

  Апроксимація градуювальної функції прямою дозволяє спростити обчислення та наочно передати градуювальні характеристики.

- визначення фізико-хімічних характеристик систем;

- стиснення даних;
- апроксимація експериментальних даних функцією з метою пошуку екстремальних точок.

Регресійний аналіз — статистичний метод дослідження залежності залежної змінної від незалежних змінних (аналіз значимості коефіцієнтів регресійного рівняння та адекватності моделі)

Основою регресійного аналізу є метод найменших квадратів (МНК).

4.2.1. Метод найменших квадратів

Метод найменших квадратів грунтується на мінімізації дисперсії двомірної вибірки. Розглянемо його більш детально.

Нехай маємо експериментальні дані:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \tag{4.2}$$

де x – незалежна змінна;
y – залежна змінна;

Для коректного застосування МНК потрібно прийняти деякі допуски.

Приймемо:

1. Похибки вимірювань є тільки у залежної змінної, або похибками незалежної змінної можна знехтувати:

$$\sigma_x^2 = 0, \qquad \text{або} \qquad \sigma_x^2 \ll \sigma_y^2 \tag{4.3}$$

2. Помилки x_i й y_i незалежні;
3. Дисперсії y_i однакові:

$$\sigma_{y_1}^2 = \sigma_{y_2}^2 = \ldots \sigma_{y_n}^2 \tag{4.4}$$

4. Похибки мають нормальний розподіл:

$$\sigma_y^2 \sim N(0, \sigma^2) \tag{4.5}$$

Метод МНК грунтується на знаходженні параметрів моделі шляхом мінімізації похибки між моделлю та експериментальними даними:

$$\sigma_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (f(x_i) - y_i)^2 \Rightarrow \min \tag{4.6}$$

Розглянемо роботу МНК на прикладі простої лінійної моделі $f(x) = a \cdot x + b$. Параметрами цієї моделі є коефіцієнти рівняння a і b . Зауважимо, що $s_y^2 = s_y^2(a, b)$ — *функція параметрів* a і b .

Похибку (дисперсію), яку вносить модель, можна розрахувати за наступним рівнянням:

$$s_y^2(a, b) = \sum_{i=1}^{n} (a \cdot x_i + b - y_i)^2 \tag{4.7}$$

Знайдемо коефіцієнти a і b моделі за умови мінімуму функції, при цьому її часткові похідні по a і b будуть дорівнювати нулю:

$$\begin{cases} \frac{\partial s_y^2(a,b)}{\partial a} = 0 \\ \frac{\partial s_y^2(a,b)}{\partial b} = 0 \end{cases}$$

Знайдемо часткові похідні по a і b, використовуючи формулу для знаходження похідної складної функції $(f(\varphi))' = f'(\varphi) \cdot \varphi'$:

Одержуємо систему простих лінійних рівнянь. Зауважимо, що невідомими величинами, котрі треба знайти є a і b .

Розв'яжемо систему рівнянь будь-яким способом (надамо читачу можливість самостійно розв'язати). Розв'язками системи будуть:

$$a = \frac{n \cdot \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} x_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \tag{4.8}$$

$$b = \frac{\sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i x_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \tag{4.9}$$

Подібним чином розраховуються параметри всіх інших лінійних відносно параметрів (або таких, що зводяться до лінійних) моделей. МНК не можна використовувати (в багатьох випадках це неможливо зробити) для моделей, нелінійних відносно параметрів.

Приклад 40 (Моделі, що зводяться до лінійних)
$f(x) = e^{ax}$ логарифмуванням зводиться до $\ln f(x) = \ln a + \ln x$

Приклад 41 (Нелінійні відносно параметрів моделі)

$$f(x) = e^{ax} + e^{bx}$$
$$f(x) = a \cdot \sin bx$$

Обмеження методу найменших квадратів

Метод найменших квадратів має ряд недоліків:

- Він дає зміщену оцінку параметрів;
- Регресійна крива чутлива до викидів;
- Цей метод придатний тільки для моделей лінійних відносно параметрів.

4.2.2. Регресійний аналіз

Регресійний аналіз проводиться звичайно у випадку більш складних моделей ніж розглянута в попередньому розділі. Він має на меті спрощення рівняння моделі. Спочатку методом МНК одержують регресійне рівняння. Потім перевіряють значимість коефіцієнтів рівняння за критерієм Стьюдента:

$$t_j = \frac{|b_j|}{s_{b_j}} \tag{4.10}$$

де b_j – j-й коефіцієнт рівняння регресії;
s_{b_j} – середньоквадратичне відхилення j-го коефіцієнта;
t_j – перевірочна t-статистика.

Якщо для відповідного члена рівняння t-статистика менше коефіцієнта Стьюдента ($t < t_{\alpha, df_{\text{залишк}}}$), то цей член усувається з рівняння моделі. Процес усування змінних є ступінчастим, на кожній стадії відбраковується член рівняння з найменшою значимістю (найменшим абсолютним значенням t-статистики).

Середньоквадратичне відхилення s_{b_j} j-го коефіцієнта розраховується згідно закону накопичення похибок:

$$s_{b_j} = \sqrt{\sum_{i=1}^{n} \left(\frac{\partial b_j}{\partial y_j}\right)^2 s_i^2} \tag{4.11}$$

Для простої лінійної моделі маємо за умови однорідності дисперсій s_i^2 (умова допустимості використання МНК, $s^2 = s_i^2 = s_1^2 = \ldots = s_n^2$):

$$s_a = \sqrt{\frac{n \cdot s^2}{n \cdot \sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2}} \tag{4.12}$$

$$s_b = \sqrt{\frac{s^2 \cdot \sum_{i=1}^{n} x_i^2}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}} \tag{4.13}$$

Незначимі коефіцієнти усуваються з рівняння регресії. Звичайно виключається коефіцієнт з найменшою значимістю (значенням t-статистики). Решта коефіцієнтів перераховуються заново, так як вони закорельовані один з одним. Адекватність рівняння визначається за допомогою дисперсійного аналізу одержаної моделі. перевіряється за критерієм Фішера:

4.2.3. Алгоритм регресійного аналізу

Загальна схема регресійного аналізу.

1. Пропонується лінійна модель;
2. Методом МНК оцінюється:
   (a) Оцінки коефіцієнтів моделі;
   (b) Стандартне відхилення оцінок коефіцієнтів моделі;
   (c) t-статистика для кожного коефіцієнта;
   (d) Рівні значимості t-статистики для кожного коефіцієнта;
   (e) Залишкова похибка й кількість ступенів свободи моделі;
   (f) Коефіцієнт детермінації R^2 та приведений коефіцієнт детермінації моделі;
   (g) F-статистика моделі, відповідні їй ступені свободи та рівень значимості цієї F-статистики.
3. Знаходиться коефіцієнт з найменшою t-статистикою;
4. Проводиться перевірка значимості вибраного коефіцієнта:

- Якщо $|t_{min}| < t_{\alpha,\nu}$ – видаляємо цю складову з моделі, переходимо до пункту 1.
- Якщо $|t_{min}| > t_{\alpha,\nu}$ – аналіз закінчено, одержуємо остаточну модель.

Пункти 1–4 повторюють доти, доки всі коефіцієнти моделі не стануть значимими.

Приклад 42 (Приклад регресійного аналізу)
Завдання – оцінити задані експериментальні дані вихідною моделлю — поліномом 7-го ступеня та провести регресійний аналіз цієї моделі.

X	1	2	3	4	5	6	7	8	9
Y	22	48	61	46	35	28	37	65	89

Вводимо їх у програму:

```
x=c(1,2,3,4,5,6,7,8,9)
y=c(22,48,61,46,35,28,37,65,89)
```

В якості вихідної моделі візьмемо поліном сьомого ступеня (Модель 1).

$$f(x) = a_7x^7 + a_6x^6 + a_5x^5 + a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0$$

Далі необхідно побудувати матрицю моделі.

```
data=data.frame(x^7,x^6,x^5,x^4,x^3,x^2,x,y)
```

Проводимо регресійний аналіз. Вихідна модель - поліном 7-го ступеня.

```
res1=lm(y~x.7+x.6+x.5+x.4+x.3+x.2+x,data)
```

Виводимо результати розрахунків:

```
> summary(res1)

Call:
lm(formula = y ~ x.7 + x.6 + x.5 + x.4 + x.3 + x.2 + x, data = data)

Residuals:
1       2       3       4       5       6       7       8
0.01997 -0.15975  0.55913 -1.11826  1.39782 -1.11826  0.55913 -0.15
9
0.01997

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.066e+02  1.177e+02   1.755    0.330
x.7         -1.181e-02  6.583e-03  -1.793    0.324
x.6          4.181e-01  2.306e-01   1.813    0.321
x.5         -6.084e+00  3.291e+00  -1.849    0.316
x.4          4.667e+01  2.459e+01   1.898    0.309
x.3         -1.994e+02  1.028e+02  -1.940    0.303
x.2          4.560e+02  2.368e+02   1.926    0.305
x           -4.822e+02  2.728e+02  -1.767    0.328

Residual standard error: 2.265 on 1 degrees of freedom
Multiple R-squared:  0.9985,    Adjusted R-squared:  0.9883
F-statistic: 97.53 on 7 and 1 DF,  p-value: 0.07781
```

Розраховуємо коефіцієнт F-розподілу:

```
> qf(0.95,7,1)
[1] 236.7684
```

qf(0.95,7,1)

Розраховуємо коефіцієнт Стьюдента:

```
> qt(0.975,1)
[1] 12.7062
```

t-статистика буде мінімальною за абсолютним значенням у вільного члена рівняння (Intercept). для нього $|t| = 1.755$. Порівнюємо дану t-статистику з коефіцієнтом Стьюдента.

В даному випадку $t = 1,755 < t_{0.05,1} = 12,706$, тому цей складник рівняння є незначимим, і його усуваєм з наступної моделі (досягається модифікацією рівняння моделі).

Одержуєм наступну модель (Модель 2):

$$f(x) = a_7x^7 + a_6x^6 + a_5x^5 + a_4x^4 + a_3x^3 + a_2x^2 + a_1x$$

Проводимо регресійний аналіз моделі 2:

```
res2=lm(y~x.7+x.6+x.5+x.4+x.3+x.2+x+0,data)
```

Одержуєм результат:

```
> summary(res2)

Call:
lm(formula = y ~ x.7 + x.6 + x.5 + x.4 + x.3 + x.2 + x + 0, data = dat

Residuals:
1       2       3       4       5       6       7       8
0.56396 -1.75773  2.93910 -2.66524  0.92183  0.54772 -0.73285  0.30
9
-0.04803

Coefficients:
     Estimate Std. Error t value Pr(>|t|)
x.7  -0.000953   0.003220  -0.296    0.795
x.6   0.033199   0.101560   0.327    0.775
x.5  -0.523195   1.266372  -0.413    0.720
x.4   4.625231   7.909663   0.585    0.618
x.3 -21.856598  25.747519  -0.849    0.485
x.2  43.514793  40.670257   1.070    0.397
```

```
x    -4.356438  23.834842  -0.183    0.872

Residual standard error: 3.235 on 2 degrees of freedom
Multiple R-squared:  0.9991,    Adjusted R-squared:  0.9961
F-statistic: 329.3 on 7 and 2 DF,  p-value: 0.003031
```

Аналізуємо результати:

t-статистика буде мінімальною за абсолютним значенням у складової x. для нього $|t| = 0.193$. Порівнюємо дану t-статистику з коефіцієнтом Стьюдента.

Розраховуємо коефіцієнт F-розподілу:

```
> qf(0.975,7,2)
[1] 39.35521
```

Розраховуємо коефіцієнт Стьюдента:

```
> qt(0.975,2)
[1] 4.302653
```

В даному випадку $t = 0,193 < t_{0.05,2} = 4,3026$ тому цей складник рівняння є незначимим, і його усуваєм з наступної моделі (модель 3 не містить складника x):

```
res3=lm(y~x.7+x.6+x.5+x.4+x.3+x.2+0,data)
```

Подібним чином продовжуємо далі, доки всі члени рівняння не будуть значимі.

Остаточне рівняння буде:

$$f(x) = -0.07499x^5 + 1.6930x^4 - 12.45x^3 + 30.66x^2$$

Так як немає особливих вимог то точності розрахунків, то залишаємо у кожній складовій по чотири значущих цифр.

Необхідно також вибрати найбільш ефективну модель, котра має найбільше значення F-статистики. У нашому випадку — це третя модель ($F = 126.7$), її рівняння:

$$f(x) = -0.03779x^6 - 0.1678x^5 + 2.508x^4 - 15.44x^3 + 34.47x^2$$

4.3. Селекція моделей

Ми розглядали регресійний аналіз, котрий заключався в тому, що ми покращували модель. І це була практична задача.

Проблема в тому, що для будь-якого процесу можна запропонувати безліч різних моделей.

І яку модель з цієї безлічі моделей вибрати в якості вихідної?

Крім того, сам регресійний аналіз має ймовірнісний характер, тому на якомусь кроці цілком можливо видалити значимий параметр регресії.

Для вирішення цього питання ще в 1890 році американським геологом і педагогом був запропонований принцип "множинності моделей".

Згідно цього принципу замість нульової та альтернативної гіпотез по Фішеру ($H_0|H_1$) формується ансамбль гіпотез $H_1, H_2, \ldots, H_r$, для кожної з яких підбирається адекватна математична модель.

Потім оцінюється сила обгрунтованості кожної з гіпотетичних моделей $g_1, g_2, \ldots, g_r$.

Ця нова парадигма сформувала нову сучасну методологію "Model selection and Multimodel inference" (Burnham, Anderson, 2002), яка базується на основних принципах теорії інформації Кульбака-Лейблера (Kullback-Leibler, 1951) і включає ранжування моделей з подальшим формуванням статистичних висновків.

В цій методології використовуються інформаційні критерії (IC) якості апроксимації.

Моделі сортуються за ступенем зменшення адекватності на основі заданого критерію IC.

Інформаційні критерії

Ентропійну міру Кульбака-Лейблера D_{KL} можна трактувати як частку втраченої інформації моделі g_i в порівнянні з відображеною цією моделлю "повної інформації".

Таким чином, із ансамбля моделей G_r необхідно вибрати таку модель, котра відповідала б мінімуму інформаційних втрат.

Вперше такий критерій був запропонований японським статистиком Хіротугу Акаіке (Akaike, 1973), котрий об'єднав поняття інформаційної та статистичної теорій, і прийшов до важливого фундаментального висновку: пошук оптимальної моделі можна здійснити шляхом мінімізації величини:

$$\ln L(\hat{\theta}) - const \tag{4.14}$$

де $L(\hat{\theta})$ – функція максимальної правдоподібності моделі g_i;
$\hat{\theta}$ – коефіцієнти моделі;
X – матриця даних;
$const$ – константа, котра корегує зміщення, і залежить від кількості ступенів свободи.

Акаіке запропонував конкретне визначення відповідного інформаційного критерію:

Критерій Акаіке

$$AIC = -2 \ln L(\hat{\theta}) + 2K \tag{4.15}$$

де K – кількість параметрів моделі.

Потім був розроблений скорегований критерій Акаіке (Sugiura, 1978):

$$AIC_c = AIC + \frac{2K(K+1)}{n-K-1} \tag{4.16}$$

де n – об'єм вибірки.

Якщо $n \gg K$, то скорегований критерій Акаіке AIC_c прямує до звичайного критерію Акаіке AIC.

Для лінійної регресії формула критерію Акаіке набуває вигляду:

$$AIC_c = n \cdot \ln \frac{RSS}{n} + 2K + \frac{2K(K+1)}{n-K-1} \tag{4.17}$$

де RSS – сума квадратів залишків моделі.

Інформаційний критерій Байеса:

Інформаційний критерій Байеса грунтується на принципі максимальної апостеріорної правдоподібності:

$$BIC = -2 \ln L(\hat{\theta}) + K \ln n \tag{4.18}$$

Крім цього, описані інші інформаційні критерії:
TIC – критерій Токісакі;
DIC – критерій ненормальності;
HQ – критерій Ханана-Куіна;
FIC – фокусований критерій;
C_p – критерій Мелоуза;
$ICOMP$ – критерій інформаційної складності.

4.3.1. Показники селекції моделей

Показником селекції моделей може служити величина Δ_i, котра характеризує втрату інформації Кульбака-Лейблера порівняно з оптимальною моделлю. Ця величина дає можливість обмежити кількість моделей:

$$\Delta_i = AIC_{c_i} - AIC_{c_{min}} \tag{4.19}$$

Відносна правдоподібність $l_i(g_i)$ визначає формалізований рівень обгрунтованості моделі порівняно з найкращою моделлю:

$$l_i(g_i) = e^{-\frac{\Delta_i}{2}} \tag{4.20}$$

Ця величина дозволяє формувати твердження типу "Модель А стільки-то раз краще пояснює результати експерименту, ніж модель Б".

Термін ”статистична значимість” тут не коректно використовувати, так як він грунтується на теорії нульової та альтернативної гіпотез Фішера.

Відносна ймовірнісна міра моделі (важіль) дозволяє оцінити відносну важливість кожної незалежної змінної як суму важелів тих моделей, де ця змінна присутня:

$$\omega_i = \frac{e^{-\frac{\Delta_i}{2}}}{\sum\limits_{i=1}^{r} e^{-\frac{\Delta_i}{2}}} \qquad \forall g_i \in G_r \tag{4.21}$$

4.4. Планування досліджень

Перед тим, як проводити експеримент, необхідно мати уявлення, що, і як досліджувати. Експромти не придатні для проведеня систематичних досліджень, хоча й можуть дати ключову інформацію про те, куди рухатись далі. Тому для серйозних досліджень необхідний план експерименту.

Вибираючи план експерименту для отримання первинних експериментальних даних, слід визначити діапазони варіювання значень вхідних факторівк, методи реєстрації властивостей досліджуваної системи та аналітичні позиції.

Як правило, план експерименту вибирають інтуїтивно, на основі апріорних міркувань і досвіду попередніх досліджень. Інший підхід просто неможливий на початковій стадії досліджень, коли невідомі ні число компонентів, ні стехіометричний склад, ні стійкості хімічних форм.

Статистичні методи планування оптимальних експериментів можуть бути затребувані лише на подальших етапах дослідження, коли структура моделі визначена, а мета заключається лише в уточненні оцінок параметрів.

Хоча процедура оптимального планування експеріментів при відомій структурі моделі і заданих оцінках параметрів є цілком рутинною операцією, нам все-таки слід розглянути ці методи.

При «інтуїтивному» плануванні експериментів слід враховувати наявний в цій області досвід, що дозволяють вибрати належний метод дослідження і уникнути грубих методичних помилок.

Щоб оцінити вплив певного фактора l_i треба отримати результат y_i за різних значень l_i.

Нехай маємо:

– m_i рівнів фактора l_i;

– k спостережень у кожній точці.

Якщо факторів n, то маємо $m_i \times \ldots \times m_n \times k$ всього спостережень. Серію $m_i \times \ldots \times m_n$ називають *повним факторним планом*

Для проведення коректного експерименту необхідно провести *повний факторний експеримент*, тобто всі досліди, що входять у повний факторний план. Крім повних планів бувають і неповні, коли за певною схемою проводиться тільки частина експериментів.

Приклад 43 (Повні факторні плани)

2×2 – кількість дослідів: $2 \cdot 2 = 4$;

3×3 – кількість дослідів: $3 \cdot 3 = 9$;

$3 \times 3 \times 3$ – кількість дослідів: $3 \cdot 3 \cdot 3 = 27$

Зі збільшенням кількості факторів кількість необхідних дослідів для повного факторного експерименту швидко зростає, тому постає необхідність використання неповних планів для мінімізації кількості вимірювань.

Якщо пропущена певна кількість вимірювань, то такий план експерименту називають неповним (дробовим).

При зменшенні кількості вимірювань завжди втрачається якась інформація. Один з підходів застосовують у випадках, коли можна знехтувати взаємодією факторів — це побудова латинських квадратів.

> Латинський квадрат — таблиця розміром $n \times n$, заповнена n різних символів так, що в кожному рядку й кожному стовпчику зустрічаються всі n символів.

За принципом латинського квадрату побудовані японські кросворди — судоку.

Приклад 44 (Латинські квадрати)
Стандартні латинські квадрати;

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 3 & 1 & 2 \end{bmatrix} \quad \begin{bmatrix} A & B & C \\ B & C & A \\ C & A & B \end{bmatrix}$$

Нестандартні латинські квадрати:

$$\begin{bmatrix} 3 & 2 & 1 \\ 1 & 3 & 2 \\ 2 & 1 & 3 \end{bmatrix} \quad \begin{bmatrix} B & A & C \\ A & C & B \\ C & B & A \end{bmatrix}$$

Кількість можливих латинських квадратів швидко зростає зі збільшенням розміру квадрату (квадратів 4×4 — 576, 5×5 — 161280). Тому частіше використовують стандартні латинські квадрати.

> Стандартний (канонічний) латинський квадрат — перший рядок і перший стовпчик побудовані в алфавітному порядку, інші стовпчики і рядки будуються шляхом циклічної перестановки елементів.

4.4.1. Використання латинських квадратів

Для зменшення кількості вимірювань використовують планування експерименту за схемою латинського квадрата.

Приклад 45 (Використання латинського квадрату)
Нехай маємо два фактори A і B, кожен має відповідні рівні (a_1, a_2,a_3 для фактору A, b_1, b_2,b_3 для фактору B,). Цьому випадку буде відповідати факторний план типу 3^2. Повний факторний експеримент для цього плану потребуватиме $3^2 = 9$ спостережень.

Для зменшення кількості вимірювань накладемо на цей план латинський квадрат 3×3.

	a_1	a_2	a_3
b_1	1	2	3
b_2	2	3	1
b_3	3	1	2

В латинському квадраті кожен елемент повторюється тільки один раз в кожному рядку і в кожному стовпчику, то якими б не були властивості елемента квадрата, вони однаково вплинуть на підрахунок середнього за стовпчиками чи за рядками.

Виберемо з цього квадрату будь-яку літеру, скажем 1, і включимо до плану експерименту всі комбінації, що відповідають цій літері:

Номер	A	B	Y
1	a_1	b_1	y_1
2	a_3	b_2	y_2
3	a_2	b_3	y_3

Одержаний план є 1/3-реплікою повного факторного експерименту 3^2. Кількість спостережень зменшилась у три рази – замість 9-ти необхідно провести тільки 3 спостереження.

Приклад 46 (Використання латинського квадрату 2)
Нехай маємо три фактори A, B і C, кожен має три рівні (a_1, a_2,a_3 для фактору A, b_1, b_2,b_3 для фактору B, b_1, b_2,b_3 для фактору C). Цьому випадку буде відповідати факторний план типу 3^3. Повний факторний експеримент для цього плану потребуватиме $3^3 = 27$ спостережень.

Для зменшення кількості вимірювань використаємо латинський квадрат 3×3 наступного типу:

	a_1	a_2	a_3
b_1	C_1	C_2	C_3
b_2	C_2	C_3	C_1
b_3	C_3	C_1	C_2

Виберемо з цього квадрату всі варіанти, маємо наступний план:

Номер	A	B	C	Y
1	a_1	b_1	c_1	y_1
2	a_2	b_1	c_2	y_2
3	a_3	b_1	c_3	y_3
4	a_1	b_2	c_2	y_4
5	a_2	b_2	c_3	y_5
6	a_3	b_2	c_1	y_6
7	a_1	b_3	c_3	y_7
8	a_2	b_3	c_1	y_8
9	a_3	b_3	c_2	y_9

Одержаний план є 1/3-реплікою повного факторного експерименту 3^3. Кількість спостережень зменшилась у три рази – замість 27-ми необхідно провести тільки 9 спостережень.

4.4.2. Греко-латинські квадрати

Для 4-х і більше факторів використовують греко-латинські квадрати, котрі утворюються внаслідок накладання двох різних латинських квадратів.

A	B	C	D	E		α	β	γ	δ	ε
C	D	E	A	B		δ	ε	α	β	γ
E	A	B	C	D		β	γ	δ	ε	α
B	C	D	E	A		ε	α	β	γ	δ
D	E	A	B	C		γ	δ	ε	α	β

Aα	Bβ	Cγ	Dδ	Eε
Cδ	Dε	Eα	Aβ	Bγ
Eβ	Aγ	Bδ	Cε	Dα
Bε	Cα	Dβ	Eγ	Aδ
Dγ	Eδ	Aε	Bα	Cβ

Приклад 47 (Приклад утворення греко-латинського квадрату)

Для побудови греко-латинських квадратів використовуються не канонічні латинські квадарти. Всі комбінації літер у таких квадратах унікальні. Проте, греко-латинські квадарти можна побудувати не всіх розмірів. Наприклад, не можливо побудувати греко-латинський квадрат розміром 6, хоча вони існують для розмірів 3, 4, 5, 7, 8, 9.

Греко-латинський квадрат розміру n містить n^2 точок, а повний факторний план – n^4 точок, тому можна легко порахувати, скільки точок містить репліка, сформована за допомогою греко-латинського квадрату.

n	ПФЕ, n^4	Греко-латинський квадрат, n^2	Репліка
3	81	9	$\frac{1}{9}$
4	256	16	$\frac{1}{16}$
5	625	25	$\frac{1}{25}$
...	...	...	...
n	n^4	n^2	$\frac{1}{n^2}$

4.4.3. Симплекс-плани Шеффе

В даний час найбільше застосування одержали симплекс-гратчасті плани, запропоновані Шеффе . Ці плани забезпечують рівномірний ний розкид експериментальних точок по $q-1$-мірному симплексу. Експериментальні точки розташовуються у вузлах $\{q, n\}$-гратки на симплексі в барицентричній системі координат, де q – число компонентів суміші; n – ступінь полінома .

Симплекс-гратчасті плани є насиченими планами. По кожному компоненту є $(n+1)$ однаково розташованих рівнів $x_i = 0, 1/n, 2/n, \ldots, 1$ і беруться всі можливі комбінації з такими значеннями концентрацій компонентів. Наприклад, для квадратичної гратки $\{q, 2\}$, що забезпечує наближення поверхні відгуку поліномами другого ступеня $n = 2$, повинні бути використані такі рівні кожного з факторів: 0, $1/2$ і 1, для кубічної $n = 3$ – 0, $1/3$, $2/3$ і 1 і т.д.

4.5. Кластерний аналіз

> Кластерний аналіз — це задача розбиття множини об'єктів на групи, які називаються кластерами. Усередині кожної групи повинні виявитися «схожі» об'єкти, а об'єкти різних групи повинні бути якомога більш відмінні.

Головна відмінність кластеризації від класифікації полягає в тому, що перелік груп чітко не заданий і визначається в процесі роботи алгоритму.

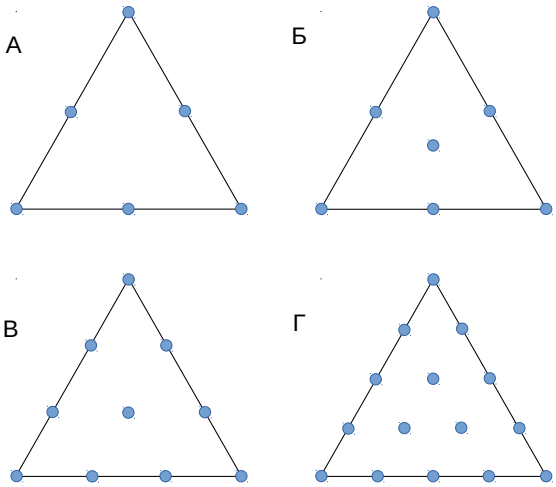


Рис. 4.2: Сімплексні гратки Шеффе в потрійній системі: квадратного $\{3, 2\}$ (А), неповного кубічного $\{3, 3*\}$ (Б), кубічного $\{3, 3\}$ (В), четвертого степеня $\{3, 4\}$ (Г)

Застосування кластерного аналізу в загальному вигляді зводиться до наступних етапів:

1. Відбір вибірки об’єктів для кластеризації;
2. Визначення множини змінних, за якими будуть оцінюватися об’єкти у вибірці;
3. При необхідності – нормалізація значень змінних;
4. Обчислення значень міри схожості між об’єктами;
5. Застосування методу кластерного аналізу для створення груп схожих об’єктів (кластерів);
6. Представлення результатів аналізу.

Після отримання та аналізу результатів можливе корегування обраної метрики і методу кластеризації до отримання оптимального результату.

Основоположником кластерного аналізу прийнято вважати Тріона (Tryon), котрий вперше в 1939 р. запропонував схему кластеризації. Взагалі під кластерним аналізом розуміють набір різних алгоритмів кластеризації.

Загальне питання кластерного аналізу — організувати дані спостережень у наочні структури, описати подібності. Сюди входять різні види класифікацій.

Приклад 48 (Класифікації об'єктів)

Біологія — класифікація живих організмів: людина — примат, ссавець, амніот, хребетний, тварина. Чим вище рівень агрегації тим менша подібність.

Астрономія — класифікація космічних об'єктів: зірки, планети, астероїди, комети.

Хімія — класифікація органічних речовин: спирти, аміни, естери, альдегіди, кетони.

Групи, до яких відносять окремі елементи називають *класами*. Класи можуть бути дискретними або безперервними.

Класифікація може проводитись за різними критеріями, тому одні й ті ж самі об'єкти, в залежності від критерію, можуть попадати в різні класи.

Приклад 49 (Залежність класифікації від критерію)

Нехай маємо речовину — етанол.

Критерій — наявність вуглецевого скелету: органічна речовина:

Критерій — агрегатний стан: рідина;

Критерій — здатність окиснювати інші сполуки: відновник.

Тому, по суті, кластерний аналіз не є точним статистичним методом.

4.5.1. Методи кластеризації

В залежності від ситуації використовують різні методи групування (кластеризації) об'єктів: об'єднання, двоходове об'єднання та метод K-середніх.

Об'єднання

Метою об'єднання, або деревовидної кластеризації, є об'єднання об'єктів у класи з використанням міри подібності (або відстані). Результатом цієї кластеризації є ієрархічне дерево.

Приклад 50 (Приклад побудови ієрархічного дерева)
Завдання – для даного набору результатів аналізу води побудувати дендрограму (класифікаційне дерево).

№	Іон	Вміст, ммоль/л
1	Na^+	2.132
2	K^+	0.064
3	Ca^{2+}	0.975
4	Mg^{2+}	0.277
5	H^+	0.026
6	NH_4^+	0.014
7	Na^-	1.125
8	NO_3^-	0.018
9	SO_4^{2-}	0.488
10	HCO_3^-	1.872

Проведення ієрархічного кластерного аналізу за допомогою R-статистики:

Створюємо у своїй папці файл clust-1.txt, де вихідні дані записуємо в наступному вигляді:

```
Іон Вміст
Na+ 2.132
K+ 0.064
Ca2+ 0.975
Mg2+ 0.277
```

```
H+ 0.026
NH4+ 0.014
Cl– 1.125
NO3– 0.018
SO42– 0.488
HCO3– 1.872
```

Читаємо файл:

```
data1=read.table("clust-1.txt",row.names='Іон',header=TRUE)
```

Розраховуємо матрицю відстаней, використовуємо метод евклідових відстаней:

```
dist1=dist(data1, method = "euclidean")
```

Проводимо ієрархічний кластерний аналіз, використовуємо повний зв'язок:

```
res1=hclust(dist1, method = "complete")
```

Виводимо зображення дендрограми:

```
plot(res1)
```

Одержуємо ієрархічну кластерну діаграму (Рис.4.3), котру інтерпретуємо в залежності від наших цілей.

Загальний алгоритм — спочатку всі об'єкти унікальні, потім критерій унікальності поступово послаблюється, схожі об'єкти об'єднуються, доки не залишиться один елемент. В результаті можна виявити кластери (гілки на ієрархічному дереві) та їх інтерпретувати згідно завдань дослідження.

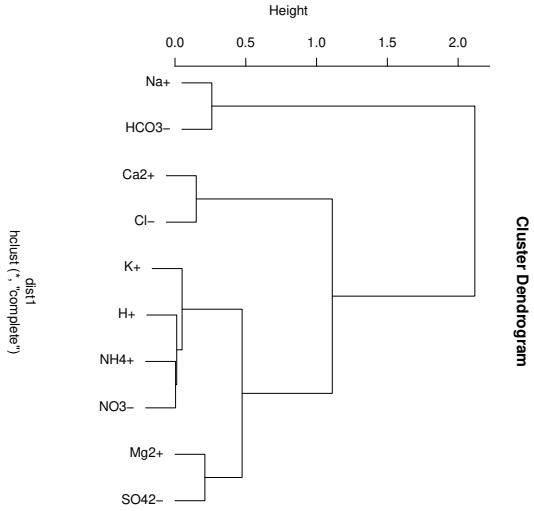


Рис. 4.3: Ієрархічна кластерна діаграма вмісту іонів у воді

Двоходове об’єднання

У цьому методі кластеризації аналізується інформація як по змінним так і по результатам.

Приклад 51 (Набір даних для використання двоходового об’єднання) Нехай маємо певну реакцію. Вихідні параметри (змінні) — температура, тиск, розчинник, концентрації реагентів та ін. Результати — вихід продукту, затрати енергії, простота виділення продукту, втрати та ін.

В результаті одержують неоднорідні кластери. Цей метод рідко використовується, але є потужним засобом розвідки.

K-середніх

Це — математичний метод. Мають бути гіпотези про кількість кластерів (задано K). Будуються K кластерів, котрі максимально відрізняються один від одного. Мінімізуються відмінності всередині кластерів. В процесі аналізу об'єкти переходять із класу в клас до досягнення максимальної значимості.

Після кластеризації можна проводити подальшу обробку даних іншими методами, щоб виявити ті чи інші особливості, зв'язки та ін.

4.5.2. Міри відстані

В залежності від задачі та експериментальних даних використовують різні методи оцінки відстані між об'єктами.

Приймемо:

n – розмірність даних, або кількість змінних, котрими характеризується кожна експериментальна точка. Тому кожна точка матиме n різних координат — $x(x_1, \ldots, x_n)$, $y(y_1, \ldots, y_n)$ тощо.

Евклідова відстань

$$d_{x,y} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{4.22}$$

На евклідову відстань сильно впливає розмірність координат. Якщо осі не співмірні (наприклад температура, тиск, концентрація та ін.), то результат буде не зовсім адекватним.

Квадрат евклідової відстані

$$d_{x,y} = \sum_{i=1}^{n}(x_i - y_i)^2 \tag{4.23}$$

У цьому випадку більш віддалені об'єкти набувають більшої ваги.

Манхеттенська відстань

$$d_{x,y} = \sum_{i=1}^{n} |x_i - y_i| \tag{4.24}$$

У випадку манхеттенської відстані, яку ще називають відстанню міських кварталів, зменшується вплив викидів (великих різниць).

Відстань Чебишева

$$d_{x,y} = \max |x_i - y_i| \tag{4.25}$$

Береться та відстань, де різниця між координатами максимальна.

Степенева відстань

$$d_{x,y} = \sqrt[r]{|x_i - y_i|^p} \tag{4.26}$$

де r і p – параметри.

Цей метод використовують для збільшення чи зменшення ваг відповідних розмірностей:

параметр p відповідає за поступове зважування за окремими координатами; параметр r — за прогресивне зважування на великих відстанях.

Якщо $r = p = 2$, то одержуємо Евклідову відстань.

Процент неузгодженості

$$d_{x,y} = \frac{N_{x_i \neq y_i}}{i} \tag{4.27}$$

Цей метод використовується для категорійних даних.

Не залежно від того, яким чином оцінюється відстань між точками, внаслідок об’єднання створюються нові об’єкти, котрі володіють своїми, притаманними лише їм, характеристиками.

4.5.3. Характеристики кластерів

Центр кластера

Центр кластера – середнє геометричне всіх його точок:

$$\vec{x}_k = \sqrt[n_k]{\prod_i \vec{x}_{ki}} \tag{4.28}$$

де k – кластер k;
n_k – кількість точок у кластері k.

Дисперсія кластера

Дисперсія кластера – міра розсіювання точок відносно центру кластера:

$$D_k = \frac{1}{n-1} \sum_{i=1}^{n} (\vec{x}_{ki} - \bar{\vec{x}}_k)^2 \tag{4.29}$$

4.5.4. Середньоквадрачичне відхилення

Середньоквадратичне відхилення об'єктів кластера:

$$s_k = \sqrt{D_k} \tag{4.30}$$

4.5.5. Радіус кластера

Радіус кластера – максимальна відстань точок кластера від його центру:

$$R_k = \max \sqrt{(\vec{x}_{ki} - \bar{\vec{x}}_k)^2} \tag{4.31}$$

4.5.6. Спірний об’єкт

Спірний об’єкт – об’єкт, котрий може бути віднесений до різних кластерів.

Розмір кластеру може визначатись як по R_k, так і по D_k.

Якщо ця умова виконується для кількох кластерів, то такий об’єкт виявляється спірним (Рис. 4.4).

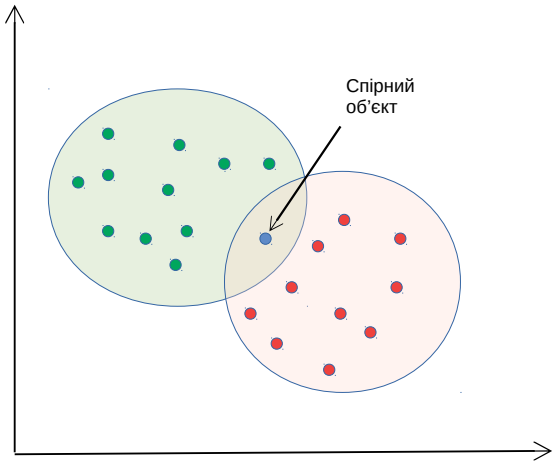


Рис. 4.4: Спірний об’єкт, котрий може бути віднесений до різних кластерів.

4.5.7. Правила об’єднання

Як об’єднати елементи на першій стадії загалом зрозуміло, але як поступати далі? Для цього розроблено ряд правил об’єднання.

Одиничний зв’язок

У методі одиничного зв’язку (метод найближчого сусіда) об’єднуються кластери, котрі мають мінімальну відстань між найближчими точками (Рис.4.5). Кластери, одержані даним методом, звичайно мають вигляд ланцюгів.

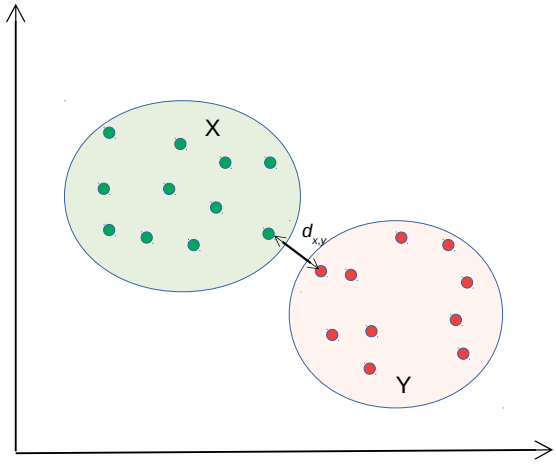


Рис. 4.5: Об’єднання кластерів методом одиничного зв’язку

Повний зв’язок

У методі повного зв’язку (метод найдальшого сусіда) об’єднуються кластери, котрі мають максимальну відстань між найбільш віддаленими точками (Рис. 4.6). Даний метод не придатний, коли кластери мають вигляд ланцюгів. У цьому випадку $d_{x,y}$ вже не буде мірою подібності.

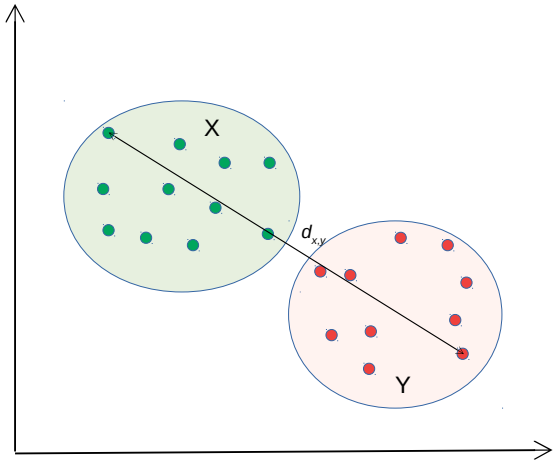


Рис. 4.6: Об’єднання кластерів шляхом повного зв’язку

Незважене попарне середнє

$$x_c = \frac{\sum\limits_{i<j} l_{ij}}{n} \qquad x_c = \frac{\sum\limits_{i<j} m_{ij}}{k} \tag{4.32}$$

де x_c та y_c – координати центрів кластерів X і Y ;
n – кількість попарних відстаней l_{ij} у кластері X;
k – кількість попарних відстаней m_{ij} у кластері Y.

Зважене попарне середнє

$$x_c = \frac{N\sum\limits_{i<j} l_{ij}}{n} \qquad x_c = \frac{K\sum\limits_{i<j} m_{ij}}{k} \tag{4.33}$$

де N – кількість точок у кластері X;
K – кількість точок у кластері Y

Дане правило діє аналогічно попередньому за виключенням того, що координати нормуються за кількістю точок.

Незважений центроїдний метод

$$x_c = \sqrt[n]{\prod_{i<j} l_{ij}} \qquad y_c = \sqrt[k]{\prod_{i<j} m_{ij}} \tag{4.34}$$

У цьому методі розраховується відстань між центрами ваги кластерів.

Зважений центроїдний метод

$$x_c = N\sqrt[n]{\prod_{i<j} l_{ij}} \qquad y_c = K\sqrt[k]{\prod_{i<j} m_{ij}} \tag{4.35}$$

Цей метод аналогічний попередньому. Він краще працює у випадках великої різниці між кластерами.

Метод Варда

У методі Варда мінімізується сума квадратів $d_{x,y}$ для всіх можливих гіпотетичних класів. Він в чомусь схожий на метод найменших квадратів. Звичайно утворюються кластери малого розміру.

4.5.8. Метод K-середніх

Метод K-середніх – це метод кластерного аналізу, мета якого є поділ m спостережень на k кластерів, при цьому кожне спостереження відноситься до того кластеру, до центру якого воно найближче. Ідея цього методу була майже одночасно сформульована Гуго Штейнгаузом і Стюартом Ллойдом в 1957 р. Сам термін «K-середніх» був вперше введений Джоном Маккуїном в 1967 р

Метод ґрунтується на мінімізації суми квадратів відстаней між кожним спостереженням та центром його кластера:

$$SS = \sum_{i=1}^{n} d\left(\vec{x}_i, m_j(x_i)\right)^2 \tag{4.36}$$

де d – міра відстані;
x_i – точка i;
$m_j(x_i)$ – центр кластеру, якому на j-тій ітерації приписана точка $\vec{x}_i$

В якості міри відстані звичайно використовується Евклідова відстань (рівняння 4.22, стор. 128).

Алгоритм K-середніх

1. Задається кількість кластерів k;
2. Випадковим чином вибирається k спостережень, які назначаються центрами кластерів;
3. Спостереження приписуються до тих кластерів, відстань до центрів яких найкоротша;
4. Розраховується новий центр кожного кластера на основі приписаних до нього спостережень;
5. Кроки 3 і 4 повторюються доти, доки центри кластерів стануть стійкими (перестануть змінюватись).

В результаті роботи алгоритму K-середніх дисперсія всередині кластерів буде мінімізована, між кластерами – максимізована.

Особливості методу K-середніх Переваги:

- простота і швидкість;
- більш наочний для великої кількості точок.

Недоліки:

- результат залежить від випадкового вибору початкових точок;

- алгоритм чутливий до викидів;
- кількість кластерів має бути задана з самого початку.

Визначення оптимальної кількості кластерів

Для визначення оптимальної кількості кластерів використовують метод ”кам’яного осипу” (метод ліктя), котрий раніше використовували лише у факторному аналізі (стор. 144).

Е інші методи (метод силуету, використання інформаційних критеріїв), але тут ми їх розглядати не будемо.

4.6. Факторний аналіз

В деяких випадках експериментальні дані, що залежать від великого числа факторів, можна описати за допомогою меншої кількості синтетичних факторів. Такі методи називають факторним аналізом.

> Факторний аналіз – сукупність статистичних методів, котрі описують експериментальні дані за допомогою невеликого числа прихованих, або латентних факторів.

Моделі факторного аналізу застосовуються при вирішенні наступних завдань:

- зниження розмірності простору ознак за рахунок зведення численних взаємозалежних змінних до деяких узагальнених неспостережуваних факторів;
- перетворення вихідних змінних до більш зручного для візуалізації або інтерпретації виду;
- класифікація об’єктів на основі стисненого простору ознак;
- створення гіпотез про структуру зв’язків об’єктів;

- інтерпретація факторів, що не піддаються безпосередньому вимірюванню.

Модель факторного аналізу – представлення вихідних змінних x_i у вигляді лінійної комбінації факторів F_j, розрахованих так, щоб з мінімальною похибкою подати представити матрицю вихідних даних X:

$$X_j = \sum_{k=1}^{p} a_{jk} F_k + U_j \tag{4.37}$$

де $F_k, \quad k = 1, 2, \ldots, p$ – латентні змінні, або *загальні фактори*;
$U_j, \quad j = 1, 2, \ldots, m$ – специфічні фактори;
a_{jk} – факторні навантаження.

Основна вимога до вихідних даних для факторного аналізу - це те, що вони повинні підкорятися припущенню про багатовимірний нормальний розподіл генеральної сукупності. Для перевірки цієї гіпотези використовують тест Бартлетта "сферичності" розподілу даних. Якщо ця гіпотеза не відкидається (тобто спостережуваний рівень значимості перевищує 5%) – немає сенсу в факторному аналізі, оскільки напрямки головних осей випадкові. На практиці припущення про багатовимірну нормальність перевірити досить важко, тому факторний аналіз найчастіше застосовується без такої процедури.

Однак якщо передбачається, що всі ознаки X_j стандартизовані ($\sigma_i = 1, m(X_j) = 0$), а фактори $F_1, F_2, \ldots, F_p$ незалежні й не пов'язані зі специфічними факторами U_j, то факторні навантаження a_{jk} збігаються з коефіцієнтами кореляції між загальними факторами та змінними X_j. Загальна дисперсія ознаки X_j розкладається при цьому на суму квадратів факторних навантажень H_i^2, яка називається спільністю, й дисперсію специфічного фактора $S_{u_i}^2$, або специфічність:

$$S_{x_i}^2 = H_i^2 + S_{u_i}^2 \tag{4.38}$$

де $H_i^2 = \sum\limits_k a_{ik}^2$

Іншими словами, спільність H_i^2 являє собою частину дисперсії змінних, поясненною факторами, а специфічність $S_{u_i}^2$ – частину дисперсії, обумовлену випадковими помилками або змінними, неврахованими в моделі. Відповідно до постановки задачі необхідно шукати такі фактори, при яких сумарна спільність максимальна, а специфічність – мінімальна.

Основним об'єктом перетворень в факторному аналізі є кореляційна матриця з коефіцієнтів кореляції Пірсона (іноді - дисперійно-коваріаційна матриця), отримана звичайним шляхом обробки масиву даних X. Виділення загальних факторів і стиснення інформації в ході факторного аналізу зводиться до відтворення, з тим або іншим ступенем точності, вихідної кореляційної матриці, тобто передбачається, що скорочена кореляційна матриця отримана з використанням тих же об'єктів, але описаних меншим числом змінних. Таким чином, слід уточнити, що фактично під стисненням інформації в факторному аналізі розуміється зменшення розмірності кореляційної матриці, а не самих даних, тим більше що відновити вихідні дані за кореляційною матрицею неможливо.

Оскільки коефіцієнти, які становлять кореляційну матрицю, можуть обчислюватися в різний спосіб, розрізняють наступні техніки факторного аналізу:

- R – техніка, коли коефіцієнти кореляції обчислюються між змінними й вихідна матриця стискається за стовпцями, тобто число ознак зменшується з m до ;
- Q – техніка, коли вивчається кореляція між об'єктами (точніше, їх станами, що описуються векторами параметрів), їх кількість зменшується з n до ;
- P – техніка, що припускає факторний аналіз результатів експериментальних досліджень, виконаних на одному й тому ж об'єкті в різні проміжки часу.

Одним з найбільш поширених прийомів пошуку факторів є метод головних компонент. Його основна відмінність від факторного аналізу

полягає в тому, що головні компоненти F_k пов'язані з спостерігаються змінними X_j лінійними функціями перетворення:

$$X_j = \sum_{k=1}^{p} a_{jk} F_k \tag{4.39}$$

$$F_k = \sum_{j=1}^{m} a_{jk} X_j \tag{4.40}$$

Метод головних компонент більш простий для розрахунків та інтерпретації, але одна з головних труднощів його використання - необхідність перетворення вихідних даних, представлених в різних одиницях виміру, в співвідносні величини. Традиційним методом перетворення є нормування за стандартними відхиленнями, коли матриця Z стандартизованих вихідних даних визначається за формулою:

$$z_{ij} = \frac{x_{ij} - x_{0\,j}}{s_j} \tag{4.41}$$

де $x_{0\,j}$ – середнє значення j-ї ознаки;
s_j – стандартне відхилення;
$j = 1, 2, \ldots, m$;
$i = 1, 2, \ldots, n$.

Для розрахунку кореляційної матриці R розміром $m \times m$, має місце просте матричне співвідношення:

$$R = \frac{1}{m-1} Z Z^T \tag{4.42}$$

Основна ідея методу головних компонент заснована на наступному припущенні: чим більше дисперсія уздовж якої-небудь осі, тим більше інформації містять значення проекцій на цю вісь. Тому цілком природно спробувати відшукати вісь з максимальною дисперсією, яку можна було б розглядати як "ординаційну" з усіма наслідками, що випливають звідси. Така вісь називається першою головною компонентою (першим головним фактором).

Пошук усієї системи взаємно перпендикулярних осей за методом головних компонент зводиться до послідовної процедури: тобто спочатку шукається перший фактор, який пояснює найбільшу частину дисперсії, потім незалежний від нього другий фактор, що пояснює найбільшу частину, що залишилася, дисперсії, і т.д.

Геометрично це виглядає наступним чином (Рис. 4.7).

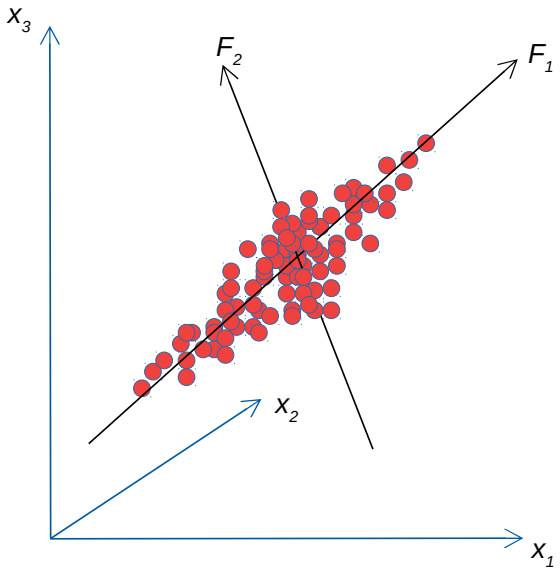


Рис. 4.7: Стиснення простору ознак при факторному аналізі

Для побудови першого фактора F_1 береться пряма, що проходить через центр координат і хмара розсіювання даних. При цьому відшукується така вісь, для якої сума квадратів відстаней всіх точок до перпендикуляра до цієї прямої була б максимальна. Це означає, що цією віссю пояснюється максимум дисперсії змінних. Знайдена вісь після нормування використовується в якості першого фактора.

Якщо ”хмара” даних витягнута у вигляді еліпсоїда (має форму

”огірка”), фактор F_1 співпаде з напрямком, в якому витягнуті об’єкти, і по ньому з найбільшою точністю можна передбачити значення вихідних змінних. Для пошуку другого фактора F_2 шукається вісь, перпендикулярна першому фактору, яка також пояснює найбільшу частину дисперсії, що не була пояснена першої віссю. Після нормування ця вісь стає другим фактором. Якщо дані представляють собою плоский еліпсоїд (”млинець”) в тривимірному просторі, два перших фактори дозволяють в точності описати ці дані. Максимально можливе число головних компонент дорівнює кількості змінних.

Основна модель методу головних компонент записується в матричному вигляді наступним чином:

$$Z = AF \tag{4.43}$$

де Z – матриця $m \times n$ стандартизованих вихідних даних;
A – матриця $m \times p$ факторів навантажень (факторне відображення);
F – матриця $p \times n$ значень факторів;
m – кількість змінних;
n – кількість об’єктів вихідної матриці;
p – кількість виділених факторів.

Очевидно, що невідомими є матриці A і F.

4.6.1. Розрахунки методу головних компонент

Розрахункові аспекти методу головних компонент зводяться до наступних кроків:

- Вирішується характеристичне матричне рівняння:

$$R = \Lambda V \tag{4.44}$$

яке в загальному випадку має m коренів λ, котрі називають власними або характеристичними числами (англ. – eigenvalue) кореляційної матриці R, кожному з яких відповідає вектор-стовпець V базисних функцій. Власними значеннями квадратної матриці

R порядку m називаються такі значення λ_k, при яких система наступних m рівнянь має нетривіальне рішення:

$$RV_k = \lambda_k V_k \tag{4.45}$$

де V_k – власні вектори матриці R, відповідні λ_k, $k = 1, 2, \ldots, m$.

- З послідовності власних значень λ_k вибирається максимальних. Матриця факторних навантажень A кожної вихідної змінної j на кожен виділений фактор k, відповідна коефіцієнтами лінійних перетворень a_{jk}, обчислюється за формулою:

$$a_{jk} = v_{ik}\sqrt{\lambda_k}, \quad j = 1, 2, \ldots, m; \quad k = 1, 2, \ldots, p \tag{4.46}$$

- Редуковану матрицю факторів F, відповідну вихідній таблиці спостережень X , в якій кількість стовпців зменшено з m до p, розраховують за формулою:

$$f_{ik} = \sum_{j=1}^{m} a_{jk} r_{ij} \tag{4.47}$$

Основна проблема розрахунків полягає в оцінці того, скільки головних компонент необхідно побудувати для оптимального представлення аналізованих вихідних факторів. Величина λ_k представляє не що інше, як частина сумарної дисперсії сукупності перетворених даних, поясненних головною компонентою F_k. Якщо змінні стандартизовані, то $\lambda_1 > \lambda_2 > \lambda_3 > \ldots$, і необхідно мати на увазі, що перші кілька членів розкладу дають основний внесок в пояснення варіації величин у вихідних даних.

Рішення про те, коли слід зупинити процедуру виділення компонент, залежить головним чином від точки зору на те, що вважати малою часткою дисперсії. Це рішення досить довільно, проте є два кри-

терії: критерій Кайзера (Kaiser) і критерій "кам'янистого осипу" Кеттелла (Cattell), які в більшості випадків дозволяють раціонально вибрати число компонент. Але аналіз складових з малими величинами власних значень навряд чи доцільний ще й тому, що вони можуть виявитися статистично недостовірними через помилки різного походження.

З огляду на те, що ілюстративною метою факторного аналізу часто є отримання факторного відображення в графічному вигляді, звичайно обмежують $p = 2$ і дають зображення простору в двомірній проекції, оскільки виконати це для трьох і більше виділених факторів проблематично.

4.6.2. Варімаксне обертання

Для інтерпретації чинників необхідно приписати кожному з них певний змістовний сенс, пов'язаний з предметною областю. Щоб зрозуміти, яка активність прихована в знайдених факторах, необхідно провести аналіз кореляцій факторних навантажень з вихідними змінними.

Для підвищення інтерпретованості факторів використовують метод варімаксного обертання, розроблений Гаррі Харманом у 1972 р. Цей метод дозволяє домогтися більшої "виразності" матриці факторних навантажень. Його суть полягає в зміні координатних осей, утворених факторами, з метою отримати більш контрастні навантаження так званої простої факторної структури. Нові фактори в результаті обертання осей шукаються у вигляді спеціального виду лінійної комбінації наявних факторів:

$$\hat{F}_i = \sum_{k=1}^{m} b_{ik} F_k \tag{4.48}$$

що максимізують "дисперсії" квадратів факторних навантажень для змінних:

$$\sum_i \left[\sum_k \frac{a_{ik}^4}{m} - \left[\sum_k \frac{a_{ik}^2}{m} \right]^2 \right] \to max \qquad (4.49)$$

Чим сильніше розійдуться квадрати факторних навантажень до кінців відрізка $[0, 1]$, тим більше буде значення цільової функції обертання і тим чіткіше інтерпретація факторів.

Для визначення оптимальної кількості факторів використовують так званий метод ”кам’яного осипу”, запропонованого в 1966 році американським психологом Реймондом Кателем .

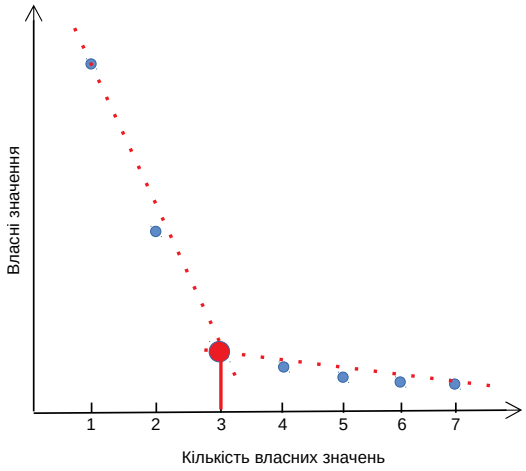


Рис. 4.8: Визначення оптимальної кількості компонент у факторному аналізі методом кам’яного осипу

Критерій кам’яного осипу полягає в пошуку точки, де спадання власних значень сповільнюється найбільш сильно. Праворуч від цієї точки знаходиться так званий ”факторная осип”. Осип – це геологічний термін для уламків, які скупчуються в нижній частині кам’янистого

схилу. Таким чином, число вибраних факторів не повинно перевищувати кількість факторів зліва від цієї точки.

4.7. Методи оптимізації

Як вже було показано в попередніх розділах, для побудови коректних моделей необхідна велика кількість експериментальних даних. Кількість необхідних експериментів швидко зростає зі збільшенням кількості факторів та їх рівнів. Крім того, побудові моделей часто перешкоджають наступні чинники:

- похибки не розподілені за нормальним законом;
- дисперсія нерівномірна;
- коректна модель нелінійна відносно параметрів.

Тому МНК не дає адекватної моделі, або зовсім не дає результату.

З іншої сторони, здебільшого має значення практичний результат, а не теорія. У таких випадках використовують методи оптимізації.

Оптимізацію використовують у наступних задачах:

- обробка результатів;
- планування експерименту (мінімізація кількості дослідів);
- коли процес описується складними функціями, котрі не мають аналітичного представлення (квантово-хімічні розрахунки).

Задача оптимізації

Звичайно задача оптимізації формулюється в наступному вигляді:

Знайти таке значення $\vec{x}_{\text{опт}}(x_{1,\text{опт}}, \ldots, x_{n,\text{опт}})$ де функція відгуку $\hat{y} = f(x_{1,\text{опт}}, \ldots, x_{n,\text{опт}})$ набуває екстремального (мінімального чи максимального) значення.

У будь-якому процесі оптимізації проводиться дослідження поверхні відгуку $f(x_1, \ldots, x_n)$. В залежності від форми поверхні відгуку оптимізація може протікати по-різному.

Величина, що характеризує рівень оптимізації процеса, називається *критерієм оптимальності*. Найчастіше це одна із функцій відгуку, що характеризує процес.

Оптимізація процесу представляє собою цілеспрямований пошук значень факторів, що впливають на процес, за яких досягається екстремум критерію оптимальності.

Класифікація поверхонь відгуку та їх придатність до оптимізації

Розглянемо різні види поверхонь відгуку на прикладі функції одної змінної (Рис. 4.9).

А Оптимізація не потрібна;

Б Насичення;

В Оптимум на нижній межі;

Г Оптимум на верхній межі;

Д Оптимум на верхній та нижній межі;

Е Типова ситуація, коли необхідна оптимізація.

Багатомірна поверхня відгуку буде комплексом тривіальних одномірних поверхонь.

Класифікація методів оптимізації

Існує ряд різних методів оптимізації, а ще більше їх варіацій. Проте, всіх їх можна розділити на групи згідно використання похідних поверхні відгуку. Таким чином є методи оптимізації:

- нульового порядку — в них не використовують похідну поверхні відгуку;

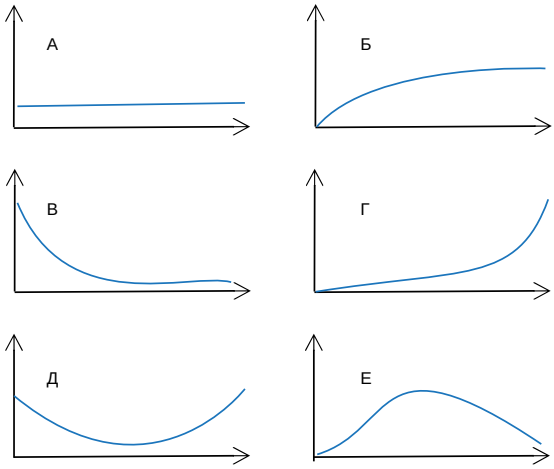


Рис. 4.9: Класифікація поверхонь відгуку в двомірному варіанті.

- першого порядку — в них використовується перша похідна поверхні відгуку;
- другого порядку — в них, крім всього іншого, використовується друга похідна поверхні відгуку.

Розглянемо більш детально ці групи методів. Хоча звичайно постає задача оптимізації багатомірної поверхні відгуку, ми будемо розглядати більш наочні, одномірні чи двомірні, приклади. Приймемо, що метою оптимізації кожного нижче розглянутого методу є знаходження *максимуму*. Незважаючи на те, що поверхня відгуку може містити багато екстремумів, будь-який метод оптимізації у випадку успіху (невдачі тут також не рідкість) завжди знаходить *один локальний екстремум!*

4.7.1. Методи оптимізації нульового порядку

Як вже було сказано, методи нульового порядку не використовують похідних поверхні відгуку в своїй роботі. Перший в світі метод оптимізації запропонував відомий флорентійський (нині Італія) вчений Фібоначчі (Fibonacci) в XIII ст.

Метод Фібоначчі

Приймемо — потрібно знайти максимум функції відгуку $y(x)$ на відрізку $[a, b]$, $x \in [a, b]$:

Метод картографування

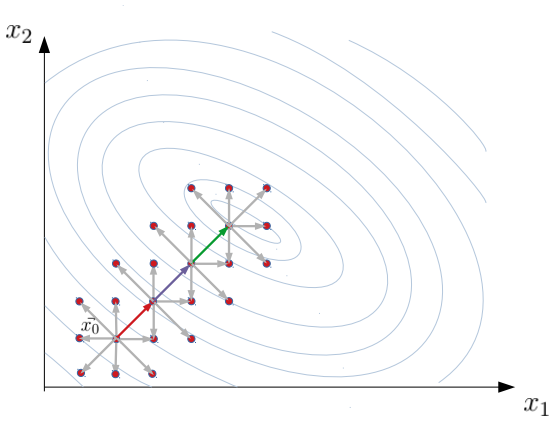


Рис. 4.10: Ілюстрація алгоритму оптимізації методом картографування

Симплекс метод

Метод був розроблений американським математиком Джорджем Данцігом у 1947 році. *Симплекс* – багатогранник у N-мірному просторі. Він має $(N+1)$ вершину. Кожна вершина відповідає певній комбінації змінних, і розглядається як вектор у N-мірному просторі. Вершину, яка відповідає найменш сприятливому відгуку, віддзеркалюють відносно гіперповерхні, утвореної рештою N вершин. В результаті цього утворюється новий симплекс, котрий відрізняється координатами однієї точки. Процедуру повторюють.

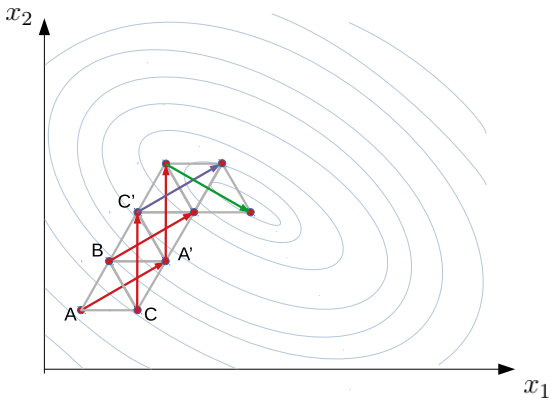


Рис. 4.11: Іллюстрація алгоритму оптимізації за методом симплексів

Метод Хука-Дживса

Метод Хука-Дживса (англ. Hooke-Jeeves) або пошук за зразком (англ. Pattern search) так само, як і метод Нелдера–Міда, призначений для пошуку безумовного локального екстремуму функції і відноситься до прямих методів, тобто спирається безпосередньо на значення функції.

Алгоритм складається з двох фаз: досліджуючий пошук і пошук за зразком.

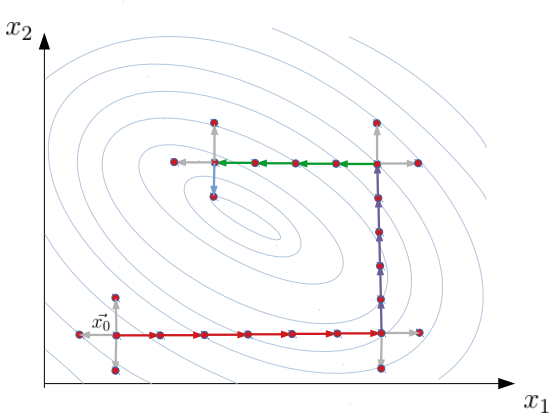


Рис. 4.12: Метод Хука-Дживса

Алгоритм:

1. Біля початкової точки $\vec{x}_0$ досліджується оточення;
2. Знаходиться напрямок $\Delta\vec{x}_i$, де $f(\vec{x})$ зменшується;
3. Рухається в знайденому напрямку до тих пір, поки $f(\vec{x})$ не перестане зменшуватись;
4. Повторюють перший крок вже для нової точки.

Особливості методу Метод Хука-Дживса гарантовано сходиться до локального мінімуму.

Загальний недолік методів нульового порядку – вони повільно сходяться.

Метод спряжених градієнтів

Алгоритм:

1. Розраховується градієнт у точці $\vec{x_0}$:

$$\vec{d} = \vec{g}(\vec{x_0}) \tag{4.50}$$

2. Рухається в розрахованому напрямку доки значення функції не стане зменшуватись:

$$\vec{x_i} = \vec{x_{i-1}} + a \cdot \vec{d} \tag{4.51}$$

де a – крок;

3. Перехід у попередню точку:

$$\vec{x}_{i+1} = \vec{x}_{i-1} \tag{4.52}$$

Перехід до пункту 1.

Особливості методу Метод спряжених градієнтів сходиться з квадратичною швидкістю (в 4–5 раз швидше, ніж метод крутого сходження).

4.7.2. Методи оптимізації другого порядку

Як вже ми згадували раніше, методи другого порядку використовують у своїй роботі значення першої та другої похідних досліджуваної функції. Тому для використання методів другого порядку необхідною умовою є подвійна диференційованість функції.

До методів другого порядку належать метод Ньютона, метод Ньютона-Рафсона, метод Марквардта та ін. Із них найбільш широко використовується метод Ньютона. Розглянемо його.

Метод Ньютона

Метод Ньютона грунтується на квадратичній апроксимації досліджуваної функції. Для цього ця функція розкладається в ряд Тейлора, обмежений квадратичними членами:

$$f(x) \approx f(x_k) + \bigtriangledown^T f(x_k) \cdot (x - x_k) + \frac{1}{2}(x - x_k)^2 \cdot \bigtriangledown^2 f(x_k) \quad (4.53)$$

Ідея методу Ньютона полягає у заміні функції $f()$ в оточенні точки її апроксимацією:

$$\phi(x) = f'(x_k) + (f''(x_k), x - x_k) \quad (4.54)$$

Точку, в якій ця апроксимація дорівнює нулю, беруть в якості наступної:

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)} \quad (4.55)$$

Формулу 4.22 називають формулою Ньютона.

У завдання пошуку мінімуму довільної квадратичної функції з позитивно визначеною матрицею Гессе других похідних ($\bigtriangledown^2 f(x_k)$) метод Ньютона дає рішення за одну ітерацію незалежно від вибору початкової точки. В цілому метод Ньютона може розходитися, якщо початкове наближення знаходиться далеко від точки мінімуму. Сходимість методу Ньютона можна гарантувати лише у випадках, коли початкове наближення перебуває у досить близькому оточенні точки мінімуму і матриця Гессе позитивно визначена і добре обумовлена. Тому на практиці цей метод зазвичай використовується в поєднанні з одним з методів, що швидко сходиться далеко від точки мінімуму.

4.7.3. Дискримінаційний аналіз

Як вже було сказано, будь-який об'єкт можна охарактеризувати деякими ознаками. Набір ознак — це *множина ознак*.

Якщо ознаки можна представити у вигляді чисел, то набір n ознак — точка в n-мірному просторі (або вектор з відповідними координатами, що виходить з початку координат). Сам об'єкт — це *образ*.

Множина образів з подібними ознаками утворюють *клас* (або область).

Таким чином постає задача — визначити, до якого класу відноситься об'єкт.

Дискримінантний аналіз допомагає виявити різницю між групами й надає алгоритм класифікації. Основне припущення – об'єкти повинні належати одному з двох (або більше) класів. Будують дискримінантні функції, якими визначають віднесення до класу .

Ознаки, що використовуються для того, щоб відрізнити один клас від іншого, називають дискримінантними змінними. Число об'єктів у загальному випадку має перевищувати кількість дискримінантних змінних приблизно вдвічі. Найчастіше дискримінантні функції є лінійними комбінаціями дискримінантних змінних. Важливо, щоб закон розподілу був нормальним.

Це дозволяє точно визначити можливість приналежності до даного класу та критерій значимості. Жодна дискримінантна змінна не може бути лінійною комбінацією інших. Неприпустимі змінні, коефіцієнт кореляції яких близький до одиниці.

В основі методів дискримінантного аналізу лежать або методи множинної регресії, або методи дисперсійного аналізу. Якщо класифікуючі змінні вважатимуться залежними від дискримінантних, то завдання аналогічне множинній регресії, лише залежна змінна вимірюється в номінальній шкалі. Але коли навпаки значення дискримінантної змінної залежить від класів, то дискримінантний аналіз є узагальненням дисперсійного аналізу.

Введемо позначення: g – число класів, – число дискримінантних змінних, n – число об'єктів класу i, - загальна кількість об'єктів. Мають виконуватись наступні співвідношення:

$$g \geqslant 2, \qquad n_i \geqslant 2, 0 < p < (n-2) \tag{4.56}$$

Дискримінантні змінні мають вимірюватися у кількісних шкалах. Канонічна дискримінантна функція є лінійною комбінацією дискри-

мінантних змінних:

$$f_{km} = u_0 + u_1 x_{1km} + u_2 x_{2km} + \ldots + u_p x_{pkm} \qquad (4.57)$$

де f_{km} – значення канонічної дискримінантної функції m-го об'єкта в групі k;
x_{ikm} – значення дискримінантної змінної x_i для m-го об'єкта в групі k;

Коефіцієнти u_i для першої функції підбираються так, щоб її середні значення для різних класів якомога більше б відрізнялися один від одного. Коефіцієнти другої функції також підбираються, але значення другої функції мають бути некорельовані зі значеннями першої. І так далі. Максимальне число дискримінантних функцій, які можна одержати, рівне кількості класів мінус один $(g-1)$, або кількості дисчантних змінних , якщо $p < g-1$.

Для визначення місцезнаходження класу можна розрахуватийого центроїд. Центроід - уявна точка, координатами якої є середні значення змінних у даному класі (вектор середніх значень для даного класу). Якщо розташування класів дійсно відрізняється, то ступінь розсіювання спостережень усередині класів буде меншою загального розсіювання.

Для оцінки розсіювання можуть бути використані коваріаційна або кореляційна матриці. Розраховують матрицю різниць міжгупової та внутрішньогрупової коваріації та розв'язують систему рівнянь, з якої визначають коефіцієнти u_i.

За знайденими коефіцієнтами для кожного конкретного об'єкта можна розрахувати значення f_{km}, й за ними провести порівняння. Для двох дискримінантних функцій зручно використовувати графічне представлення.

Якщо дискримінантних функцій більше двох, то графічне представлення не буде наочним. Якщо ми маємо справу з великою кількістю спостережень, то найпоширенішим методом графічного уявлення буде побудова гістограм за кожною функцією для кожної окремої групи.

Для визначення взаємозалежності окремих змінних та дискримінантних функцій знаходять коефіцієнти кореляції. Ці коефіцієнти називають *п*овними структурними коефіцієнтами. Вони показують, наскільки тісно пов'язані змінні з дискримінантними функціями. Коли абсолютна величина структурного коефіцієнта велика, майже вся інформація про дискримінантну функцію заключена в цій змінній.

Максимальна кількість канонічних дискримінантних функцій буде меншою за будь-яке з чисел або $(g-1)$. Деякі функції будуть або нульовими або статистично мало значущими. Корисність кожної функції може оцінюватись за різними критеріями. Так, наприклад, як критерій значущості можна вибрати частку дисперсії дискримінантної функції, яка пояснюється розбиттям на класи.

Використовуючи дискримінантні функції ми можемо проводити класифікацію, тобто передбачити клас, до якого найімовірніше належить об'єкт. Вибірку, за якою проведено поділ класів, називають навчальною. Дискримінантні функції дають нам корисну інформацію про окремі об'єкти, розбіжності між класами, про здатність змінних точно розрізняти класи.

Задачі дискримінації

Дискримінаційний аналіз може вирішувати наступні задачі:

Маємо об'єкти, кожен з яких відноситься до однієї або кількох груп.

— Треба знайти функцію, яка дозволяє поставити нові зразки в якусь із груп.

Втрачені ознаки належності об'єкта до однієї з груп. Їх треба відновити.

4.7.4. Порядок виконання дискримінантного аналізу

Процедура дискримінантного аналізу включає наступні стадії:

1. Поділ вибірки на дві частини;
2. Вибір змінних – предикторів;

3. Розрахунок параметрів дискримінантної функції;
4. Інтерпретація результатів.

Розглянемо ці етапи більш детально.

Поділ вибірки на дві частини

Поділ вибірки необхідний для перевірки адекватності одержаної дискримінантної функції. Тому вибірка поділяється навчальну та тестову сукупності. Навчальна сукупність використовується для побудови моделі, тестова для її перевірки.

Навчальна сукупність (analysis sample) – частина загальної сукупності, котру використовують для розрахунку дискримінантної функції.

Тестова сукупність (validation sample) – частина загальної сукупності, котру використовують для перевірки одержаної на основі навчальної сукупності дискримінантної функції.

Вибір змінних – предикторів

Якщо кількість змінних досить велика (наприклад, кілька сотень), то неможливо застосувати дискримінантний аналіз до всіх змінних одночасно.

Тому керуються наступними положеннями:

- на початковому етапі дискримінантного аналізу для предикторів формується кореляційна матриця. У цьому контексті вона має особливий сенс, називається загальною внутрішньогруповою кореляційною матрицею і містить середні коефіцієнти кореляції для двох або більше кореляційних матриць (кожна для однієї групи);

- крім загальної внутрішньогрупової кореляційної матриці можна побудувати коваріаційні матриці для окремих груп, для всієї вибірки або загальну внутрішньогрупову матрицю;
- крім того можна застосувати серію критеріїв між двома групами для кожної змінної або однофакторний дисперсійний аналіз, якщо число груп виявляється більше двох.

Оскільки метою дискримінантного аналізу є складання найкращого рівняння, додатковий аналіз вихідних даних ніколи не є зайвим.

Розрахунок параметрів дискримінантної функції

Тут можливі два варіанта – Метод примусового включення (direct method) та Покроковий дискримінантний аналіз (stepwise discriminant analysis).

Метод примусового включення Дискримінантну функцію розраховують при одночасному введенні всіх предикторів. У цьому випадку враховується кожна незалежна змінна

Покроковий дискримінантний аналіз Предиктори вводяться послідовно, залежно від їхньої здатності розрізняти групи. Цей метод грунтується на мінімізації коефіцієнта Вілкса (λ) після включення в рівняння регресії кожного нового предиктора.

Коефіцієнт λ –- відношення внутрішньогрупової суми квадратів до загальної суми квадратів, він характеризує ступінь впливу предиктора на дисперсію критерію. Зі значенням λ пов'язані величини F і p, що характеризують його значущість.

Цей метод краще застосовувати в ситуації, коли дослідник хоче відібрати підмножину предикторів для включення їх у дискримінантну функцію.

Інтерпретація результатів

Метою дискримінантного аналізу є складання рівняння регресії з використанням вибірки, для якої відомі як значення предикторов, так

і значення критерію. Це рівняння дозволяє за відомими значеннями предикторів визначити невідомі значення критерію іншої вибірки.

4.8. Попередня обробка даних

Дані, одержані в будь-якому експерименті, можна представити у вигляді матриці:

$$\begin{matrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ik} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mj} & \cdots & x_{mk} \end{matrix} \tag{4.58}$$

де m – кількість зразків;
k – кількість змінних (або вимірювань) на кожному зразку

Це — *матриця даних*. Завдання — виявити корисну інформацію, яка неявно міститься в даних.

Ці дані за різних причин не завжди придатні для коректного аналізу. Тому використовують різні операції перетворення вихідних даних.

> Попередня обробка даних — будь-яке перетворення вихідних даних.

Незалежно від проблем, які вирішуються, дані мають бути якісними. Тільки тоді можна очікувати на коректний результат.

> Якість даних – критерій, який визначає повноту, точність, своєчасність і можливість інтерпретації даних.

Відповідно до цього дані можна розділити на два типи:

- Якісні – піддаються інтерпретації;

- Брудні – дані низької якості, які містять пропуски, мають низьку точність, включають в себе надлишок інформації.

Забруднення даних буває різних типів:

1. Пропущені значення;
2. Дублікати;
3. Шуми;
4. Викиди.

Тип обробки залежить від характеру проблем: які задачі чи питання вирішуються, які фізичні чи хімічні фактори, яка структура та вигляд даних. Розглянемо основні операції, які використовуються при попередній обробці даних.

4.8.1. Відсутні дані

Коли в матриці даних є пусті місця виникає дуже незручна ситуація — часто такі дані неможливо обробити деякими методами. Тому використовують заповнення пропусків. Слід зауважити, що заповнення нулями в багатьох випадках теж неприпустимо. Існують наступні методи фабрикації відсутніх даних: заповнення середнім, інтерполяція, випадкове заповнення.

Заповнення середнім

У цьому випадку береться середнє значення вимірів відповідної змінної:

$$x_{ij} = \frac{1}{m-1} \sum_{e \neq i} x_{ie} \tag{4.59}$$

Інтерполяція

Для розрахунку відсутніх значень використовується інтерполяція за допомогою різних моделей, сплайнів та ін.

> Інтерполяція – це метод знаходження невідомих проміжних значень змінних значень деякої функції за наявним дискретним набором її відомих значень.

> Сплайн – функція, область визначення якої розбита на відрізки, на кожному з яких функція є деяким поліномом

> Апроксимація – заміна одних об'єктів іншими, схожими, але більш простими.

Випадкове заповнення

Випадково беруться дані з інших вимірювань.

Вилучення даних

Зразки з пропусками у вимірюваннях можна просто вилучити з матриці даних.

Слід відмітити, що незалежно від методу усунення пропусків, змінюється структура даних. У випадках заповнення середнім або інтерполяції дані «покращуються», у випадку випадкового заповнення — «погіршуються». У випадку вилучення даних кореляція кореляція може змінюватись у ту чи іншу сторону.

4.8.2. Виявлення надлишкових змінних та констант

Перед подальшою обробкою сукупність даних потрібно перевірити на наявність констант та надлишкових змінних. Для цього проводять факторний аналіз.

Факторний аналіз

Для зменшення розмірності даних перевіряють кореляційну матрицю.

З математичної точки зору факторний аналіз репрезентує вихідну матрицю даних як лінійну комбінацію ортогональних (незалежних) векторів — таким чином зменшується надлишковість у даних.

4.8.3. Трансляція

Трансляція — це зміна положення точок відносно осей (лінійне перетворення координат).

$$x'_{ij} = x_{ij} - \bar{x}_j \tag{4.60}$$

$$\bar{x}_j = \frac{1}{m} \sum_{i=1}^{m} x_{ij} \tag{4.61}$$

4.8.4. Нормування

Довжина всіх векторів приводиться до однакової довжини:

$$C_i = \sum_{j=1}^{k} x_{ij}^2 \tag{4.62}$$

$$x'_{ij} = \frac{x_{ij}}{\sqrt{C_i}} \tag{4.63}$$

При цьому перетворенні зменшується значимість дисперсії.

4.8.5. Масштабне перетворення

Зміна масштабу для більш наочного представлення даних.

$$x'_{ij} = \frac{x_{ij} - x_{j,min}}{x_{j,max} - x_{j,min}} \tag{4.64}$$

де $x_{j,min}, x_{j,max}$ – мінімальне та максимальне значення в межах класу;

Масштабне перетворення координат *дуже чутливе* до викидів.

4.8.6. Автомасштабне перетворення

В автомасштабному перетворенні даних враховується розсіювання:

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \tag{4.65}$$

де s_j – середньоквадратичне відхилення в межах змінної.

Автори [Шараф] рекомендують саме цей метод використовувати в більшості випадків.

4.8.7. Обертання

Координатні осі повертаються таким чином щоб експериментальні точки групувались навколо одної з осей.

4.8.8. Обертання власного вектора

Координати перетворюються таким чином, щоб осі були направлені в напрямку найбільших дисперсій.

4.9. Сигнали, виявлення та управління

Дослідження звичайна мають певну мету. На основі результатів досліджень роблять певні (якісні чи кількісні) висновки. Ці висновки ґрунтуються на властивостях зразків. Як правило, спостереження та фіксацію властивостей проводять за допомогою *приладів*. Результати *спостережень* наводять у вигляді *оцінок*.

Приклад 52 (Зв'язок між приладом, спостереженням та оцінкою) Ваги (прилад) використовується для визначення маси (властивість) в грамах (оцінка маси).

Вибір приладу має важливе значення для дослідження. Не завжди цей вибір є тривіальним. Властивості, що спостерігаються, мають бути «*корисними*» відгуками приладу. Іншими словами, відгук приладу повинен бути функціонально пов'язаний з властивістю, що спостерігається. Наприклад, ваги не можна використати для визначення показника заломлення невідомої величини. Показник заломлення, як властивість речовини, не є корисним відгуком ваг. Тут властивість і відгук ніяк не пов'язані між собою.

Будь-яке вимірювання, зроблене дослідником, можна назвати сигналом.

Сигнал – відгук приладу на певний вплив.

Є сенс розглядати тільки ті сигнали, котрі є інструментальним відгуком на присутність чи відсутність аналіту. Сигнал може вимірюватись як функція одного або кількох параметрів.

Вимірювання інструментальних сигналів – необхідна умова для якісних та кількісних висновків.

Розглянемо ідеальний приклад. Припустимо, що прилад реагує тільки на речовину A, якщо вона є в зразку, і ніяк не реагує, якщо її там немає. Таким чином, роблять висновок про наявність A в зразку кожен раз, коли прилад показує сигнал більше нуля. Будь-яка кількість речовини A (незалежно від того, наскільки її мало) може бути визначена, і пов'язаний з цим сигнал чітко визначений.

На жаль, ця ситуація може існувати лише в мріях дослідника (а також у головах пересічних громадян) і навряд буде коли-небудь досягнута. Будь-які незначні недосконалості аналітичної процедури (похибки в показах приладів, домішки, коливання температури, нестабільність пробовідбору, та ін.) будуть збільшувати невизначеність сигналу. До якоїсь міри це нівелюється введенням у прилади електронних схем, але вони теж вносять свій вклад у невизначеність сигналу,

так як там завжди випадкові флуктуації струму та напруги. Тому навіть за відсутності A в зразку буде спостерігатись не нульовий сигнал, його ще називають фоновим, або холостим сигналом.

> Холостий (фоновий) сигнал – сигнал, одержаний за відсутності досліджуваного компонента

Цей сигнал обов'язково присутній, так як відгук приладу за відсутності A не є фіксованою величиною, і не спричиняється виключно складом середовища в холостому досліді.

Центр розподілу фонового сигнала звичайно не знаходиться біля нульової точки. За відсутності A прилад показує відгук, котрий розподілений навколо певної величини $\bar{x}_\Phi$ – фонового значення. Фоновий сигнал обумовлений випадковими змінами в аналітичній процедурі та випадковими процесами в модулях самого приладу. Цю непідконтрольну комбінацію експериментальних та інструментальних змін називають випадковим шумом. Джерела випадкових змін діють не залежно від того, чи є в зразку A, чи ні. Із-за цього спостерігається дисперсія сигналів, як фонового, так і сигналу A.

Вплив випадкового шуму на фоновий сигнал, і на сигнал зразка має особливе значення. Для визначення розподілу фонового сигналу звичайно необхідно багато вимірювань (більше 20-ти).

Звичайно передбачається, що фоновий сигнал, і сигнал A мають нормальний розподіл. Крім того, ці сигнали мають однакову дисперсію ($\sigma_\text{A}^2 = \sigma_\Phi^2 = \sigma^2$).

4.10. Виявлення сигналу

Давайте розглянемо більш детально співвідношення сигналів фону та зразка. Представимо ці сигнали у вигляді функцій густини ймовірності. Будемо вважати, що ці сигнали мають розподіл Гауса з параметрами μ і σ^2, і відповідають вибірковим $\bar{x}$ і s_x^2.

Розглянемо можливі варіанти.

4.10.1. Сигнали повністю розділені

В даній ситуації визначення сигналу зразка повністю визначене. Незважаючи на те, що повторними вимірюваннями можна уточнити величини μ_A і σ_A^2, для якісних висновків достатньо *одного вимірювання*.

Так як величина сигналу залежить від кількості A в зразку, відстань між μ_A і μ_Φ скорочується зі зменшенням вмісту A. Це можна проілюструвати переносом сигналу зразка вздовж осі сигналу.

Коли сигнал зразка наближається до фонового, їх розподіли починають перекриватись. Величина цього перекривання визначає міру невизначеності, пов'язану з детектуванням аналітичного сигналу. Якщо сигнали фону та зразка повністю розділені, невизначеність у виявленні наближається до нуля.

Наближаючи сигнали зразка й фона неможливо уникнути їх перекривання. Це перекриття є мірою ймовірності допустити помилки у виявленні сигналу (прийняти зразок за фон, чи навпаки, див. помилки першого та другого роду, стор.67).

4.11. Співвідношення сигнал/шум

Співвідношення сигнал/шум (англ. SNR або S/N, Signal-to-noise ratio, рос. Отношение сигнал/шум) — міра сотворення сигналу спотворений шумом. Визначається як відношення інтенсивності корисного сигналу до інтенсивноссі шуму.

$$SNR = \frac{I_{\text{сигнал}} + I_{\text{шум}}}{I_{\text{шум}}} \tag{4.66}$$

4.12. Методи покращення співвідношення сигнал/шум

4.12.1. Оптимізація

Відгук будь-якого приладу залежить, окрім кількості A в зразку, також від багатьох інших факторів аналізу. Звичайно за час пробопідготовки кількість A не змінюється, таким чином, один з параметрів відгуку приладу залишається інваріантним. Тому потрібно *оптимізувати* решту експериментальних факторів для того, щоб одержати якомога вищу величину SNR.

Оптимізація експериментальних змінних залежить від методу аналізу. Вона може визначатися дією одного, або кількох параметрів. Так як експериментальні фактори можуть корелювати один з одним, то коректна оптимізація повинна взаємну зміну їх рівнів. Детальніше про методи оптимізації можна прочитати в розділі 4.7.

4.12.2. Усереднення сигналу

Якщо аналітичних сигнал легко відтворюється, то можна одержати кілька його значень та розрахувати середнє значення. Можна прийняти, що середнє значення випадкового шуму буде дорівнювати нулю. При усередненні n вимірювань співвідношення SNR збільшується в $\sqrt{n}$ раз. Тому для зменшення рівня шуму до низьких значень може бути потрібно багато вимірювань. Кількість вимірювань звичайно визначається вимогами до точності та видатками на проведення досліджень.

Найчастіше цей прийом використовується в спектральних методах аналізу, коли зразок не пошкоджується при проведенні вимірювань.

4.12.3. Фільтрування сигналу

Під фільтрацією сигналів розуміють процедуру цілеспрямованої зміни спектрального складу сигналу. Найчастіше метою фільтрації є зни-

ження рівня неінформативних складових сигналу.

Дійсний сигнал, крім корисної інформації, пов'язаної з передачею первинних повідомлень, завжди містить в собі і шумову складову:

$$z(t) = x(t) + u(t) \tag{4.67}$$

де $z(t)$ – реальний сигнал, котрий містить в собі корисний сигнал $x(t)$ і шум $u(t)$.

Величина $\Delta(t) = y(t) - x(t)$ називається залишкової похибкою фільтрації. Для того, щоб фільтрація мала місце, повинна виконуватися умова:

$$\Delta(t) \ll u(t) \tag{4.68}$$

де $x(t)$ – сигнал на вході фільтра;
$y(t)$ – сигнал на виході фільтра.

Для ідеального фільтра $y(t) = x(t)$ і похибка фільтрації дорівнює нулю, тобто відбувається 100% - виділення сигналу з адитивної суміші сигналу і шуму.

4.12.4. Модуляція сигналу

Модуляція сигналу використовується, коли характеристики шуму відомі і відрізняються від сигналу. У цьому випадку для зменшення шуму використовуються фільтри.

4.12.5. Мультиплексна спектроскопія

Мультиплексна спектроскопія – метод, що ґрунтується на одночасному детектуванні сигналу на багатьох частотах. Основна перевага цього методи перед послідовним розгортанням спектру - економія часу. Зняття спектру займає набагато менше часу, що дозволяє використати такі способи підвищення SNR як накопичення або усереднення сигналу.

Якщо багатоканальний детектор містить n каналів, то порівняно з аналогічним одноканальним детектором за той же самий час можна досягнути збільшення SNR в $\sqrt{n}$ раз.

Загальна проблема багатоканальних спектрометрів – недостатня кількість каналів. Для вирішення цієї проблеми використовуються перетворення Адамара та Фур'є.

Перетворення Адамара

Прилади, в яких застосовується перетворення Адамара, мають у своєму складі решітку, котра поміщається між джерелом сигналу та багатоканальним детектором. Решітка (маска) містить певним чином розташовані отвори, через які може вільно проходити випромінювання. При повторенні вимірювань одні отвори відкриваються, інші закриваються. В результаті одержуються дані, котрі можна описати у вигляді добутку матриці масок на матрицю відгуків індивідуальних каналів:

4.13. Метод головних компонент

Метод головних компонент (англ. Principal Component Analysis, PCA) — один з основних способів зменшення розмірності даних з втратою найменшої кількості інформації.

Метод головних компонент — один з основних способів зменшити розмірність даних, втративши найменшу кількість інформації. Описаний Карлом Пірсоном у 1901 році та доповнений і розширений Гарольдом Готелінґом в 1933 р.

Цей метод почав активно поширюватись завдяки розвитку комп'ютерної техніки. Сучасні прилади дозволяють швидко одержувати значні масиви даних. Наприклад, ІЧ-спектрометр може кожні 15 секунд знімати запис на 300 довжинах хвиль. За годину він встигає зробити 240 циклів зйомки. В результаті одержується матриця, котра містить $300 \times 240 = 72000$ значень. Так як, звичайно, ці спектри схожі, частка корисної інформації в цьому масиві даних невелика.

Головна ідея цього методу – представлення вихідних даних з використанням прихованих змінних, котрі не відомі.

Для використання цих методів мають виконуватись умови:

- Число нових змінних повинно бути суттєво меншим, ніж вихідних;
- Втрати від перетворення інформації не мають сильно перевищувати шум.

Ці методи дозволяють представити корисну інформацію в більш компактному вигляді, зручному для інтерпретації.

Суть методу PCA

$$X = TP^t + E = \sum_{a=1}^{A} t_a p_a^t + E \tag{4.69}$$

де X – вихідна матриця;
T – матриця оцінок;
P – матриця навантажень;
E – матриця залишків.
A – число головних компонент.

Число стовпчиків t_a в матриці T та число стовпчиків p_a в матриці P відповідає ефективному (хімічному) рангу матриці X.

Приклад 53

X – матриця спектрів сумішей речовин. Число рядків відповідає кількості зразків (I). Кожен рядок – спектр, знятий на J довжинах хвиль.

Можна записати:

$$X = CS^t + E \tag{4.70}$$

де C – матриця концентрацій (розмір $I \times A$);
S – матриця спектрів чистих речовин (розмір $A \times J$), $S^t = J \times A$.

Задача поділу експериментальної матриці X на "чисті" складові – предмет вивчення особливої галузі хемометрики, котра називається "розділення кривих" (curve resolution).

Для вирішення задачі можуть використовуватись різні алгоритми, які ми тут не будем розглядати.

Головна мета у всіх задачах PCA – визначення величини хімічного рангу системи, – числа головних компонентів A. Якщо PCA використовується для поділу даних на хімічно осмислені компоненти, то такий метод називають факторним аналізом.

В результаті обчислень розраховують власні значення:

$$\mu_a = 100\% \cdot \frac{\sum_{i=1}^{I} t_{ia}^2}{\sum_{i=1}^{I} \sum_{j=1}^{J} x_{ij}^2} \tag{4.71}$$

та пояснену дисперсію:

$$E_a = 100\% \cdot \left(1 - \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} e_{ij}^2}{\sum_{i=1}^{I} \sum_{j=1}^{J} x_{ij}^2} \right), \qquad a = 1, 2, \ldots, A \tag{4.72}$$

При проведенні перетворення, так як і в інших випадках, завжди втрачається частина хімічної інформації.

4.13.1. Графік оцінок

Багато інформації може дати графік оцінок.

Кожна точка на графіку представляє окремий спектр.

- Якщо точки близько – існує кореляція між спектрами, спектри схожі;

- Якщо точки лежать в сусідніх квадрантах – немає кореляції між спектрами;
- Якщо точки знаходяться в протилежних секторах – негативна кореляція.

4.14. Класифікація і дискримінація в PCA

Так як дані в хімії звичайно *мультиколінеарні* (спектри, хроматограми), то моделі, як правило, багатомірні, тому необхідно використовувати PCA або PLS.

PCA і PLS дуже візуально дуже схожі, проте між ними є суттєва різниця:

- PCA використовується для класифікації та дослідницького аналізу даних, визначені головні компоненти описують джерела мінливості в порядку зменшення значимості;
- PLS використовується для кількісної оцінки даних, прогнозування та моделювання поточних та майбутніх зразків, латентні змінні є важливим фактором у визначенні ефективності моделі.

4.14.1. Метод SIMCA

Метод SIMCA (англ. Soft Independent Modeling of Class Analogy – Формальне незалежне моделювання аналогі класів) був запропонований шведським хемометриком Сванте Волдом.

В основі цього методу лежить припущення, що об’єкти одного класу мають схожі властивості, але володіють індивідуальними особливостями, які слід вважати шумом.

Унікальність цього метода визначається наступними особливостями:

- По-перше, кожен клас моделюється відокремлено, незалежно від інших;

- По-друге, SIMCA класифікація є багатозначною – кожен зразок може бути одночасно віднесений до декількох класів;
- По-третє, в SIMCA є унікальна можливість встановити значення помилки 1-го роду і побудувати відповідний класифікатор.

Алгоритм методу SIMCA

1. Моделювання методом PCA з різним числом головних компонент A;
2. Розрахунок відстані від об'єкту до класу. Вона визначається як середньоквадратичне значення залишків e, котрі виникають при проекції об'єкту на клас:

$$d = \sqrt{\frac{1}{J-A}\sum_{j=1}^{J} e_j^2} \tag{4.73}$$

3. Порівняння відстаней із середньоквадратичними залишками всередині класів:

$$d_0 = \sqrt{\frac{1}{(J-A-1)(J-A)}\sum_{ij} e_{ij}^2} \tag{4.74}$$

4. Розрахунок розмаху. Він визначає відстань від об'єкта до центра класу, і, по-суті, є квадратом відстані Махаланобіса:

$$h = \frac{1}{I} + \sum_{a=1}^{A} \frac{\tau_a^2}{t_a^t t_a} \tag{4.75}$$

де τ_a – проекція нового зразка на головну компоненту a;

t_a – вектор рахунків усіх навчальних зразків у класі.

Вцілому зовні метод SIMCA дуже схожий на кластерний аналіз.

Питання та вправи для самостійного опрацювання

1. Перерахуйте методи аналізу багатомірних даних.
2. Дайте визначення моделі.
3. Вимоги до регресійних функцій.
4. Алгоритм послідовного регресійного аналізу.
5. Які критерії використовуються в регресійному аналізі?
6. Поясніть сутність факторного аналізу?
7. Охарактеризуйте методи попередньої обробки даних?
8. Дайте визначення кластерному аналізу.
9. Які існують алгоритми кластеризації?
10. Які використовуються міри відстаней у кластерному аналізі?

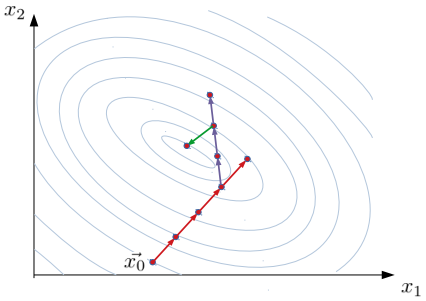


Рис. 4.13: Метод спряжених градієнтів

Розділ 5

Вибір ознак для класифікації

A priori ми не знаємо, які змінні будуть значимими в задачі класифікації. Найкращою стратегією завжди є вимірювання багатьох змінних, з наступним розрахунком важелів, для того, щоб визначити найбільш дискрімінуючі змінні.

Зважування ознак заключається у визначенні кількісної міри дискримінуючої здатності змінних в рамках класифікаційних гіпотез, висунутих стосовно даних.

Матриця невідповідностей – це матриця $N \times N$, де N – кількість прогнозованих класів. Ось декілька визначень, які потрібно пам'ятати для матриці невідповідностей:

Точність (акуратність відсутність систематичних помилок): частка від загальної кількості правильних прогнозів.

Позитивна прогностична цінність (точність - міра випадкових помилок): частка позитивних випадків, які були правильно визначені.

Негативна прогностична цінність: частка негативних випадків, які були правильно визначені.

Чутливість: частка справжніх позитивних випадків, які були пра-

вильно визначені.

Специфічність: частка фактично негативних випадків, які були правильно визначені.

Питання та вправи для самостійного опрацювання

1. Охарактеризуйте проблему класифікації.
2. Назвіть принципи відбору ознак для класифікаціх.
3. Опишіть використання дисперсійного важіля.
4. Опишіть використання важіля Фішера.
5. Які критеріх відбору ознак?
6. Вимоги для даних для класифікації?
7. Використання кластерного аналізу для класифікації.
8. Деревовидна класифікація.
9. Лінійна дискримінація.
10. Формулювання висновків у задачах класифікації.

Розділ 6

Теорія графів

За останні десятиліття в теоретичній хімії набули поширення поняття топології та теорії графів. Вони корисні при пошуку кількісних співвідношень ”структура - властивість» і «структура-активність», а також у вирішенні теоретично-графічних і комбінаторно-алгебраїчних задач, що виникають в ході збору, зберігання й обробки інформації зі структури та властивостей речовин.

Графи служать, перш за все, засобом зображення молекул. При топологічному описі молекули її зображують у вигляді молекулярного графа, де вершини відповідають атомам, а ребра – хімічним зв’язкам (теоретикчно-графічна модель молекули). Звичайно в такому поданні розглядають тільки скелетні атоми, наприклад, вуглеводні з «стертими» атомами водню.

Валентність хімічних елементів накладає на вершини певні обмеження. У розгалужених алканів ступені вершин (r) не можуть перевищувати чотирьох.

> Граф – множина об’єктів, що пов’язані між собою заданою множиною бінарних зв’язків

Графи можна представити в матричному вигляді, що є зручним при роботі з ними на комп’ютерах. Ця матриця називається *м*атрицею

суміжності вершин простого графа. Це квадратна матриця $= [a_{ij}]$, $a_{ij} = 1$, якщо вершини i та j з'єднані ребром, $a_{ij} = 1$, якщо безпосереднього зв'язку немає.

Граф також можна охарактеризувати за допомогою матриці відстаней. Це квадратна матриця $D = [d_{ij}]$. Елементи d_{ij} характеризують мінімальне число ребер (найкоротшу відстань) між вершинами i та j.

Вигляд матриць A і D залежить від способу нумерації вершин (або ребер), що спричиняє незручність при роботі з ними. Тому часто для характеристики графа використовуються інваріанти графа -– топологічні індекси. Вони можуть бути самими різноманітними, залежно від цілей, яких планується досягти.

> Топологічний індекс – числове значення, що характеризує структуру молекули.

Топологічні індекси використовуються для побудови залежностей «структура-властивість». Теоретико-графова методологія включає наступні етапи:

- Вибір об'єктів дослідження (навчальна вибірка) та аналіз стану чисельних даних по властивості для даного кола сполук.
- Відбір топологічних індексів з урахуванням їхньої дискримінаційної, кореляційної здатності.
- Вивчення графічних залежностей "Властивість – топологічний індекс молекули".
- Встановлення функціональної залежності $P = f(\text{топологічний індекс})$.
- Зіставлення розрахованих значень з експериментальними.
- Передбачення властивостей ще вивчених і ще не синтезованих сполук.

Топологічні індекси також використовуються у побудові адитивних схем розрахунку та прогнозування. Вони можуть застосовуватися при розробці нових лікарських засобів, при оцінці канцерогенної активності деяких хімічних речовин, для передбачення відносної стійкості нових (не синтезованих) сполук і т.д.

Слід пам'ятати, що вибір топологічного індексу нерідко носить випадковий характер. Вони можуть не відображати важливі структурні особливості молекул, а розрахункові схеми не мати грунтовного теоретичного фундаменту й погано піддаватися фізико-хімічній інтерпретації.

Особливості графів молекул

Якщо ми подивимось Рис. 6.1, то скажемо, що це – вуглецеві скелети етану, бутану, ізобутану і циклобутану. А те, що вони намальовані порізному, не має значення.

А у циклобутану точки можна і не ставити, малюючи, наприклад, молекули циклогексану, бензолу і його аналогів. Графи такого типу ще називають молекулярними графами (МГ).

Нам залишається додати, що в теорії графів точки найчастіше називають вершинами, а лінії, що їх з'єднують, – ребрами. Які ще особливості графів і, відповідно, МГ необхідно відзначити. Для графа «байдуже», як пара його вершин з'єднана ребром, важливо тільки знати, є воно чи ні. Тому графи з кратними ребрами називають мультиграфами. І, таким чином, мультиграфом представляють тут МГ з подвійними або потрійними зв'язками (Рис. 6.2).

Таким чином, наведені тут МГ відрізняються від графа тільки тим, що їх вершини відображають атоми вуглецевих скелетів, іншими словами без атомів гідрогену, так як їх додавання значно ускладнює МГ. Це давно вже зрозуміли хіміки-органіки, які звичайно не знають теорію графів, але широко застосовують МГ. Ребра ж символізують зв'язку між деякими з атомів вуглецю.

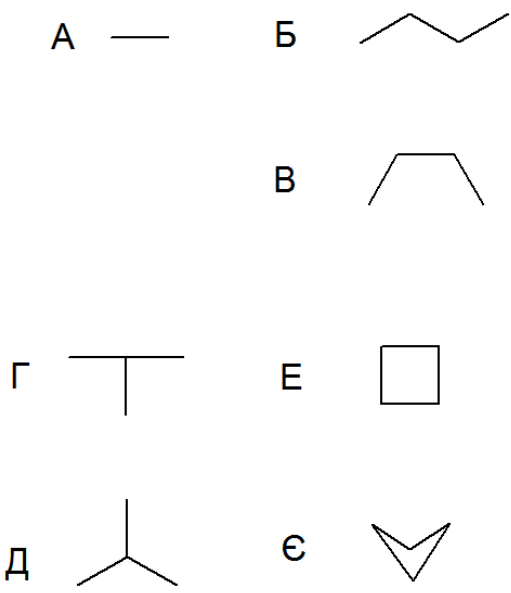


Рис. 6.1: Молекулярні графи етану (А), бутану (Б, В), ізобутану (Г, Д), циклобутану (Е, Є).

6.1. Структура графа

Що ще потрібно знати про графи (МГ)?

Вершини графа, з'єднані ребром, називаються суміжними, з'єднані вершина і ребро називаються інцидентними. Число інцідентних одній і тій ж вершині ребер називається її ступенем або валентністю. Обидва варіанти майже рівноправні в самій теорії графів, а «один із засновників сучасної теорії графів» Вільям Тат в своїй книзі застосовує тільки термін «валентність» і пише що «Термін" валентність "навія-

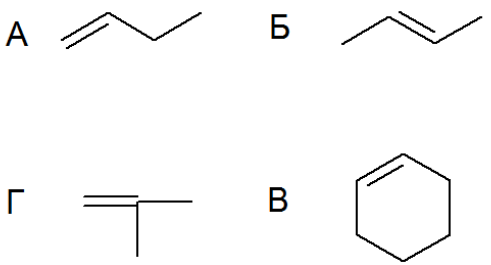


Рис. 6.2: Мультиграфи бутену-1 (А), бутену-2 (Б), циклопропену (В), 2-метилпропену (Г),

ний хімічними аналогіями». Тому тут застосування цього терміну тим більше виправдано.

Вершини, які не мають ребер (наприклад, МГ метану), називаються ізольованими, валентності 1 - висячими, валентності 2 -двухвалентними (звичйно в МГ таких вершин більшість), валентності 3 і 4 - вузловими. А в МГ їх відповідно варто називати первинними, вторинними, третинними і четвертинними вершинами або ж вуглецевими атомами, як називають їх хіміки.

Іноді в процесі вивчення видаляють з графа деякі ребра (зв'язку) або вершини. Останні обов'язково видаляються з усіма їх зв'язками, що призводить до відповідного зменшення валентності кожної з суміжних з нею вершин, що залишаються в графі. Частину називають підграфом вихідного графа.

Слідуючи цьому підходу, видалимо середній зв'язок з МГ бутану. Частина, що залишилася – підграф. Але «дістатися» від одного кінця цього графа до іншого за допомогою зв'язків неможливо, хоча «в пам'ять» про МГ бутану цей підграф – один граф. В теорії графів такі графи називаються незв'язними, а його зв'язкові частини – компонентами.

Якщо «придивитися» з хімічної точки зору, то отриманий так підграф МГ бутану складається з двох МГ етану (Рис. 6.1). А зв'язний граф, таким чином, складається з однієї компоненти. Граф, що складається з одних ізольованих вершин, називається повністю незв'язним, а протилежний йому граф, у якого кожна вершина з'єднана ребрами з усіма іншими, називається повним. Цілком зрозуміло, що всі МГ звичайних органічних молекул є зв'язковими графами, навіть МГ метану, що складається з однієї ізольованої вершини.

6.2. Графи об'ємних молекул

Для побудови графів об'ємних молекул використовують діаграми Шлегеля.

> Діаграма Шлегеля – проекція випуклого багатогранника на простір меншої розмірності через точку за однією з його граней.

Одержана фігура комбінаторно еквівалентна вихідному багатограннику. Діаграма названа по імені Віктора Шлегеля, який запропонував в 1886 році цей метод для вивчення комбінаторних і топологічних властивостей багатогранників. У тривимірному просторі діаграма Шлегеля є проекцією тривимірного багатогранника в плоску фігуру.

Приклад 54 (Графи тетраедру, куба, октаедру)

Діаграми Шлегеля було запропоновано використовувати для моделювання структур фулеренів та карборанів .

Питання та вправи для самостійного опрацювання

1. Дайте визначення графу.
2. Що таке молекулярний граф.

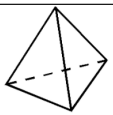
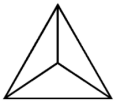
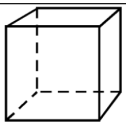
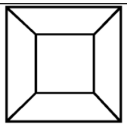
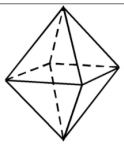
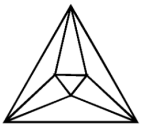
Фігура	Об'ємний вид	Граф
Тетраедр		
Куб		
Октаедр		

Рис. 6.3: Графи тетраедра, куба та октаедра

3. Вимоги до молекулярних графів.
4. Принципи побудови графів об'ємних органічних молекул.
5. Побудуйте граф тетраедра.
6. Побудуйте граф ікосаедра.
7. Які переваги використання графів для опису структур карборанів та фулеренів?
8. Що таке діаграма Шлегеля?
9. Наведіть графи бутану, трет-бутану, нафталену?
10. Наведіть графи всіх ізомерів карборану.

Розділ 7

Візуалізація даних

> Візуалізація даних – це представлення даних у вигляді, що забезпечує найбільш ефективну роботу людини з їх вивчення

Підсистема візуалізації даних є важливою складовою якісних систем інтелектуального аналізу даних, особливо орієнтованих на обробку великих обсягів інформації. Візуалізація може використовуватись на всіх етапах процесу обробки даних:

- Візуалізація вихідних даних. Цей етап корисний для оцінки ступеня відповідності очікуванням та придатності даних до аналізу, висування гіпотез про закономірності та необхідні процедури первинної обробки.
- Візуалізація вибірки, завантаженої у систему обробки.
- Візуалізація результатів первинної обробки.
- Візуалізація проміжних результатів.
- Візуалізація остаточних результатів.

Візуалізація даних забезпечує на наступні характеристики сприйняття даних:

- стисліcть (англ. concision) – здатність одночасного відображення великої кількості різнотипних даних;
- відносність (англ. relativity) і близькість (англ. proximity) – здатність демонструвати в результатах запиту кластери, відносні розміри груп, схожість та відмінність груп, випадаючі значення (англ. outliers);
- концентрацію та контекст (англ. focus with context) – взаємодія з деяким обраним об'єктом з можливістю перегляду його стану та зв'язків із контекстом;
- масштабованість (англ. zoomability) – здатність легко і швидко переміщатися між мікро- та макропредставленням;
- орієнтацію на «праву півкулю» – надання користувачеві не тільки заздалегідь встановлених методів роботи з даними (які забезпечують його свідомі та сплановані підходи до пошуку потрібної інформації), а й підтримка його інтуїтивних, імпровізаційних когнітивних процесів ідентифікації закономірностей.

Так як візуалізація має на меті вплив на людей, то для її досягнення слід керуватися наступними правилами:

- Потрібно вибрати правильну діаграму, залежно від того, яка у вас мета.
- Переконатися, що ідея вашої діаграми підходить аудиторії.
- Використати коректний дизайн оформлення діаграми.

Діаграма (грец. Διάγραμμα (diagramma) – зображення, малюнок, креслення) – графічне представлення даних відрізками чи геометричними фігурами, що дозволяє швидко оцінити співвідношення кількох величин.

Якщо діаграма вибрана невдало, людина, яка її переглядає, може заплутатися або помилково інтерпретувати дані. Тому важливо визначитися, які дані ви хочете візуалізувати та з якою метою.

7.1. Види діаграм

Для представлення даних можна використати наступні типи діаграм: Стовпчаста діаграма, Гістограма, Лінійний графік, Подвійна вісь діаграма, Діаграма площ, Гістограма з накопиченням, Діаграма Мекко, Кругова діаграма, Діаграма розсіювання, Пухирчаста діаграма, Схема водоспаду, Діаграма воронки, Маркерна діаграма, Теплова карта. Розглянемо їх більш детально.

Стовпчаста діаграма

Стовпчаста діаграма (Рис. 7.1) використовується для порівняння між різними елементами або може показувати порівняння елементів у часі.

Ствопчикова діаграма ідеально підходить для порівняння кількох наборів даних. Горизонтальні стовпчикові діаграми звичайно використовують, коли потрібно порівняти велику кількість показників або візуально виділити явну перевагу одного з них. А вертикальні стовпчикові діаграми добре ілюструють, як змінювалися показники у різні періоди.

Поради щодо створення стовпчастих діаграм:

- Використовуйте послідовні кольори по всій діаграмі, вибираючи акцентні кольори, щоб виділити важливі точки даних або зміни з часом.
- Використовуйте горизонтальні підписи, щоб покращити читабельність.
- Почніть вісь у з нуля, щоб належним чином відобразити значення на вашому графіку.

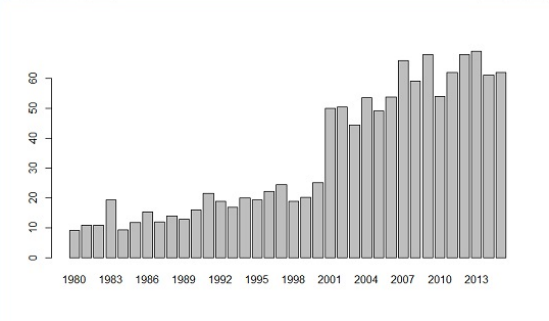


Рис. 7.1: Стовпчикова діаграма

Сендвічева діаграма

Сендвічеву діаграму (Рис. 7.2), або горизонтальну стовпчасту діаграму, слід використовувати, щоб уникнути загромадження діаграми, коли одна мітка даних довга, або якщо у вас є більше 10 елементів для порівняння. Цей тип візуалізації також можна використовувати для відображення від'ємних чисел.

Поради щодо створення сендвічевих діаграм:

- Використовуйте послідовні кольори по всій діаграмі, вибираючи акцентні кольори, щоб виділити важливі точки даних або зміни з часом.
- Використовуйте горизонтальні підписи, щоб покращити читабельність.
- Почніть вісь у з нуля, щоб належним чином відобразити значення на вашому графіку.

Гістограма

Гістограму (Рис. 7.3) часто плутають зі стовпчиковою діаграмою через візуальну схожість, але все ж у цих графіків різні цілі. Гістограма

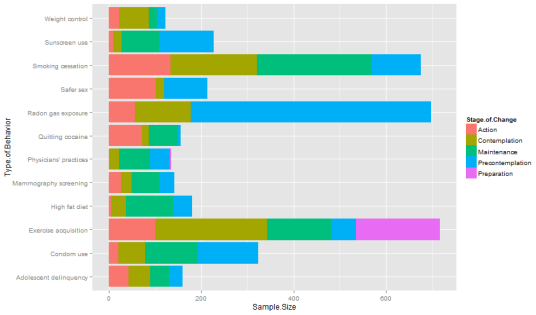


Рис. 7.2: Сендвічева діаграма

показує, як розподіляються дані у межах безперервного інтервалу чи певного періоду часу. Вертикальна вісь цього графіка звичайно має розмірність частоти, а горизонтальна -— інтервал чи часовий період.

Стовпчикова діаграма на відміну від гістограми, не пов'язана з безперервним інтервалом - тут кожен стовпчик означає окрему категорію.

Поради щодо створення гістограм:

- В гістограмі важливо коректно визначити розмір інтервалу.
- Чим більше стовпчиків у гістограмі, тим краще проявляється структура даних.
- Орієнтуйте центр розподілу в центр діаграми

Лінійний графік

Лінійний графік (Рис.7.4) показує тенденції або прогрес з часом і може використовуватися для відображення багатьох різних категорій даних. Його треба використовувати його під час створення діаграм безперервного набору даних.

Поради щодо створення лінійних графіків:

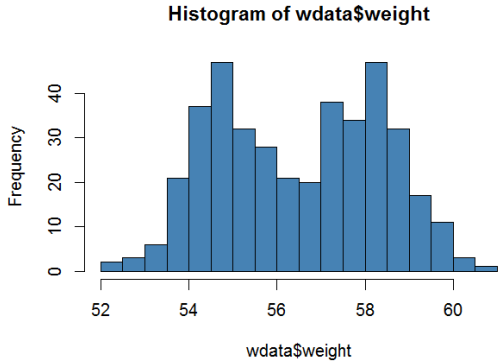


Рис. 7.3: Гістограма

- Використовуйте лише суцільні лінії.
- Не малюйте більше чотирьох ліній, щоб уникнути розсіювання уваги.
- Використовуйте коректну висоту діаграми, щоб лінії знаходились десь на рівні $2/3$ висоти осі Y.

Діаграма з двома шкалами

Діаграма з двома Y-шкалами спільною віссю X (Рис. 7.5). Вона використовує два набори даних, один з яких базується на безперервному ряду значень, а інший, який краще підходить для групування за категоріями.

Цей тип діаграм найкраще підходить для візуалізації кореляції або її відсутності між наборами даних.

Поради щодо створення діаграм з двома шкалами:

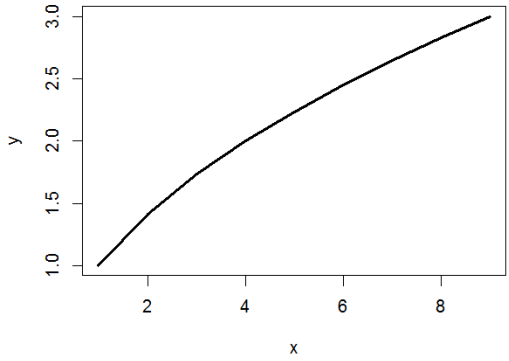


Рис. 7.4: Лінійний графік

- Використовуйте вісь ординат з лівого боку для основної змінної, оскільки мозок від природи схильний спочатку дивитися ліворуч.
- Використовуйте різні стилі графіків, щоб проілюструвати два набори даних.
- Виберіть контрастні кольори для двох наборів даних.

Площова діаграма

Площова діаграма (Рис. 7.6) -— в основні своїйу лінійна діаграма, але простір між віссю X і лінією заповнюється кольором або растром. Це корисно для відображення зв'язків від частини до цілого, це допомагає аналізувати як індивідуальну, так і загальну інформацію про тенденції.

Поради щодо створення площової діаграми:

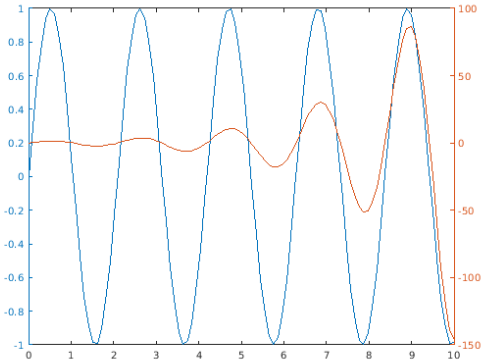


Рис. 7.5: Діаграма з двома шкалами

- Використовуйте прозорі кольори, щоб не перекривалась інформація.
- Не показуйте більше чотирьох категорій, щоб уникнути загромадження.
- Організуйте різноманітні дані у верхній частині діаграми, щоб їх було легко читати.

Стогова діаграма

Стогова діаграма (Рис. 7.7) використовується для порівняння багатьох різних елементів, враховуючи їх склад.

Поради щодо створення стогової діаграми:

- Найкраще використовувати для ілюстрації відношення частини до цілого.
- Використовуйте контрастні кольори для більшої чіткості.

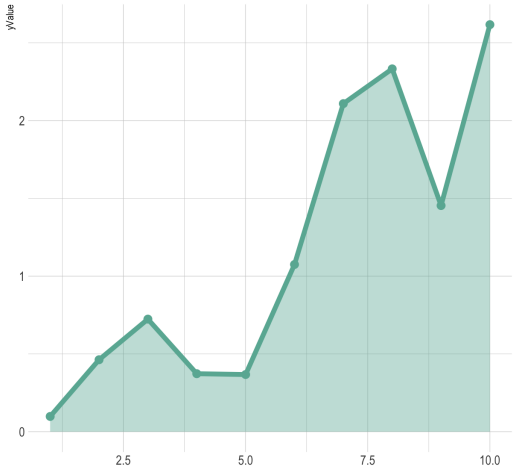


Рис. 7.6: Площова діаграма

- Зробіть масштаб діаграми достатньо великим, щоб переглядати розміри груп по відношенню один до одного.

Діаграма Мекко

Діаграма Мері Мекко (Рис. 7.8) використовується для порівняння значень, складу кожного компонента та розподілу даних між ними.

Вона подібна до стогової діаграми, за винятком того, що вісь X діаграми Мекко використовується для відображення іншого виміру значень, а не прогресу в часовій шкалі, як це часто буває зі стовпчастими діаграмами.

Порода щодо створення діаграм Мекко:

- Змінюйте висоту бруска, якщо розмір порції є важливим для порівняння.

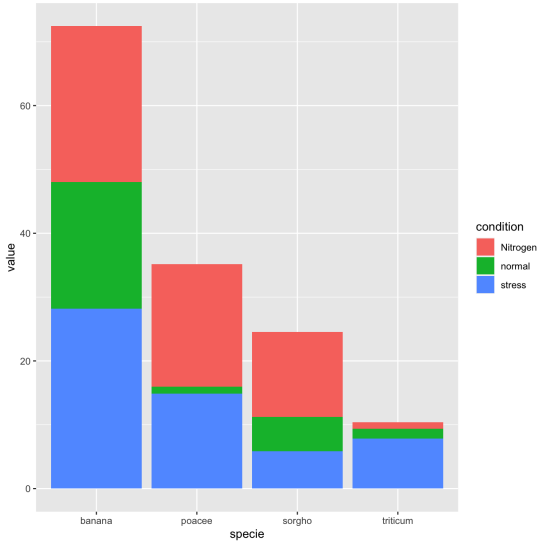


Рис. 7.7: Стогова діаграма

- Не включайте занадто багато складених значень у кожну панель. Ви можете переглянути, як представити свої дані, якщо у вас їх багато
- Розташуйте свої смуги зліва направо таким чином, щоб показати відповідну тенденцію.

Кругова діаграма

Кругова діаграма (Рис. 7.9) показує статичні значення, і те, як категорії представляють частину цілого в композиції. Кругова діаграма представляє числа у відсотках, а загальна сума всіх сегментів повинна дорівнювати 100

Поради щодо створення кругових діаграм:

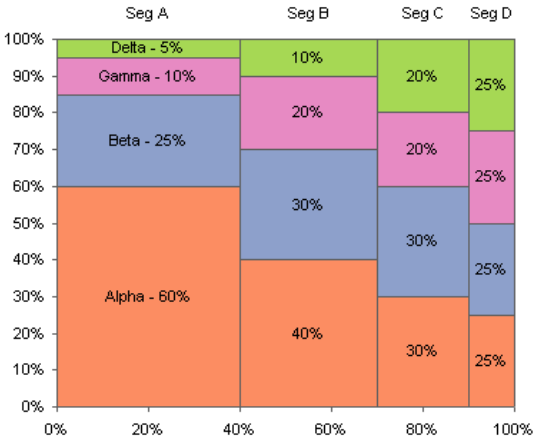


Рис. 7.8: Діаграма Мекко

- Не ілюструйте занадто багато категорій, щоб забезпечити диференціацію між фрагментами.
- Переконайтеся, що сума значень зрізів становить 100%
- Розташовуйте сектори відповідно до їхнього розміру.

Діаграма розсіювання

Діаграма розсіювання (Рис. 7.10) відображає співвідношення між двома різними змінними або виявляє тенденції розподілу. Її слід використовувати, коли є багато різних точок даних, і необхідно виділити схожість у наборі даних. Це корисно під час пошуку викидів або для розуміння розподілу даних.

Поради щодо створення діаграм розсіювання:

- Включіть якнайбільше точок.

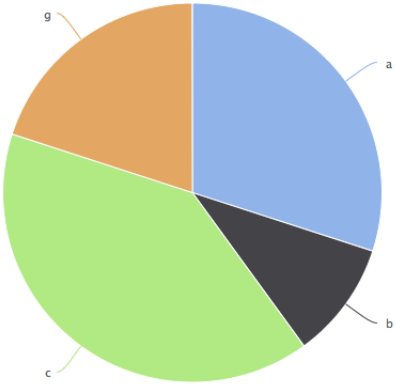


Рис. 7.9: Кругова діаграма

- Почніть вісь Y з 0 , щоб точно представляти дані.
- Якщо ви використовуєте лінії тренду, використовуйте не більше двох, щоб ваш графік був легким для розуміння.

Бульбашкова діаграма

Бульбашкова діаграма (Рис. 7.11) подібна до діаграми розсіювання в тому, що вона може показувати розподіл або взаємозв'язок. Але тут є третій набір даних, який позначається розміром бульбашки або кола.

Поради щодо створення бульбашкових діаграм

- Масштабуйте бульбашки відповідно до площі, а не діаметра.
- Переконайтеся, що мітки чіткі та помітні.
- Використовуйте тільки круглі форми.

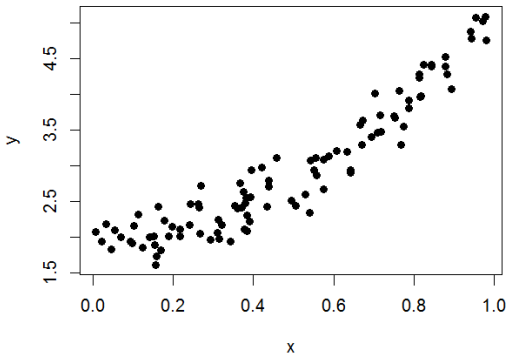


Рис. 7.10: Діаграма розсіювання

Теплова карта

Теплова карта (Рис. 7.12) показує зв'язок між двома елементами та надає інформацію про оцінки, наприклад від високого до низького або від поганого до відмінного. Інформація про рейтинг відображається за допомогою різних кольорів або насиченості.

Поради щодо створення теплової карти:

- Використовуйте простий і чіткий контур карти , щоб не відволікатися від даних.
- Використовуйте один колір різних відтінків, щоб відобразити зміни в даних.
- Уникайте використання декількох візерунків.

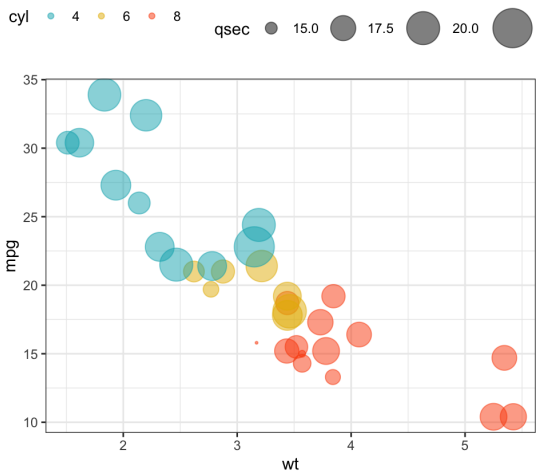


Рис. 7.11: Бульбашкова діаграма

7.2. Вибір діаграм

Як вибирати тип діаграми, як зробити їх зрозумілими, простими та привабливими.

Таблиці чи діаграми?

Використовуйте діаграму, якщо:

- потрібно передати ідею, демонструвати яку ви будете за допомогою лише кількох значень;
- необхідно показати зв’язки між багатьма значеннями.

Використовуйте таблицю, якщо:

- потрібно порівняти багато конкретних значень;

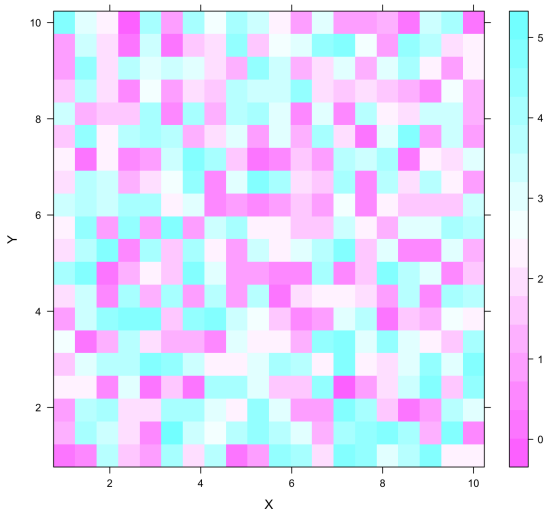


Рис. 7.12: Теплова карта

- необхідно оперувати максимально точними числами;
- дані містять багато різних розмірностей.

Діаграма допомагає досягненню наступних цілей:

- Швидко та однозначно передати ідею.
- Ніхто не любить вникати у цифри. Діаграми спрощують і допомагають потрапити в цифри.

Класифікація діаграм

Діаграми поділяються на чотири різні групи залежно від їхньої функції.

1. Показують відносини між різними числами (наприклад, кореляцію).
2. Порівнюють числа.
3. Показують склад структуру даних.
4. Показують розподіл/співвідношення величин.

В основному використовують два типи діаграм – порівняння величин та репрезентація структури величин.

Алгоритм вибору діаграми

Відправною точкою при виборі типу діаграми завжди іде ідея (message), яку ви хочете донести. Дані є визначальним чинником. Більше того, навіть не маючи даних, але визначившись із ідеєю, можна вибрати тип майбутньої діаграми.

Порядок вибору діаграми:

- визначте ідею, яку хочете донести;
- визначте потрібну функцію діаграми;
- виберіть відповідний тип діаграми;
- відформатуйте діаграму.

Загальні рекомендації щодо використання діаграм

- Шкала часу завжди має бути горизонтальною. Час має йти зліва направо.
- Розміри елементів діаграми (стовпці, рядки тощо) повинні бути завжди пропорційні значенням, які вони відображають. Стовпець числа 100 і стовпець числа 200 повинні відрізнятися у 2 рази.

- Видаліть всю інформацію зі шкал, всі зайві кольори та позначення, якщо вони не допомагають вашій ідеї.
- Колонки, рядки та ін. завжди повинні сортуватися за їх значенням. Не сортуйте за алфавітом.
- Якщо у вас є лише один тип значень, вам не потрібна легенда.
- Мітки ваших даних завжди повинні бути максимально близькими до значень, щоб уникнути плутанини.
- Не використовуйте більше 6 кольорів у діаграмі. Якщо вам потрібно більше 6 – ви помиляєтеся.
- Для порівняння одного і того ж значення в різні періоди часу використовуйте один і той же колір різної інтенсивності (наприклад, від світлого до темного).
- Найбільш універсальна палітра кольорів — чорний, білий, червоний, зелений, синій та жовтий.
- Якщо у вас кілька діаграм (таке називається серія), використовуйте єдиний стиль оформлення.
- Перевіряйте, як ваша діаграма виглядатиме, якщо її роздрукують.
- Не додавайте надто багато інформації на одну діаграму, краще розбийте її на дві.

Бібліоґрафія

1. Шараф М. А., Иллмен Д. Л., Ковальски Б. Р. Хемометрика / М. А. Шараф, Д. Л. Иллмен, Б. Р. Ковальски: Пер. з англ. – Л.: Химия, 1989, 272 с. Пер. вид.: США, 1986. – ISBN 5-7245-0361-1.

2. Kiralj R., Ferreira M. M. C. The past, present, and future of chemometrics worldwide: Some etymological, linguistic, and bibliometric investigations / Journal of Chemometrics. 20. (6-7). 2006. p.247 - 272. doi:10.1002/cem.1001.

3. Ахназарова С. Л., Кафаров В. В. Методы оптимизации эксперимента в химической технологии / С. Л. Ахназарова, В. В. Кафаров. — 2-е изд. — М.: Высш. шк., 1985, 327 с.

4. Монтгомери Д. К. Планирование эксперимента и анализ данных / Д. К. Монтгомери. – Л.: Судостроение, 1980, 384 с.

5. Самолов Н. А. Моделирование в химической технологии и расчет реакторов: Учеб.пособие / Н.А. Самолов : Уфа ООО «Монография», 2005. – 224 с.

6. Гунич С. В., Янчуковская Е.В. Математическое моделирование и расчет на ЭВМ химико-технологических процессов. Примеры и задачи: учеб. пособие / С.В. Гунич, Е.В. Янчуковская. – Иркутск: Изд-во ИрГТУ, 2010. — 216 с.

7. Товажнянский Л.Л. Вычислительная математика и программирование в химической технологии / Л.Л. Товажнянский. – Х.:НТУ «ХПИ». 2005. – 258 с.

8. Солтис М.М., Закордонський В.П. Теоретичні основи процесів хімічної технології / М.М. Солтис, В.П. Закордонський — Львів: ЛНУ ім Івана Франка., 2003, 430 с.

Додаткова література :

9. Холин Ю.В. Количественный физико-химический анализ комплексообразования в растворах и на поверхности химически модифицированных кремнеземов: содержательные модели, математические методы и приложения / Ю.В. Холин — Харьков: Фолио, 2000, 288 с.

10. Холоднов В.А., Хартманн К., Чепикова В.Н., Андреева В.П.. Системный анализ и принятие решений. Компьютерные технологии моделирования химико-технологических систем с материальными и тепловыми рециклами : учебное пособие./ В.А. Холоднов, К. Хартманн. СПб.: СПбГТИ (ТУ), 2006. – 160 с.

11. Загальна хімічна технологія:підручник/ В.Т. Яворський, Т.В. Перекупко, З.О. Знак, Л.В. Савчук. - Львів: видавництво НТУ «Львівська політехніка», 2005. – 552 с

12. Комп'ютерні системи проектування. Теорія і практика./ Ред.: М.В. Лобур. -Л.: Львів- Політехніка. 2004. - 183 с

13. Семенов С.А. Планирование эксперимента в химии и химической технологии / Учебно-методическое пособие. М.: ИПЦ МИТХТ, 2001 г., - 93 с.

14. Семенов С.А. Планирование эксперимента в химии и химической технологии. Часть 2. / Учебно-методическое пособие. М.: ИПЦ МИТХТ, 2005 г.- 29 с.

15. Семенов С.А. Планирование эксперимента в химии и химической технологии. Часть 3. / Учебно-методическое пособие. М.: ИПЦ МИТХТ, 2005 г.- 28 с.

Інтернет:

16. Померанцев А. Л. Четвертая парадигма [електронний ресурс] / А. Л. Померанцев, д.ф.-м.н., Российское хемометрическое общество [www.chemometrics.ru]. – http://www.chemometrics.ru/materials/articles/paradigm/ [22.02.2015]

17. Шитиков В. Работа с пакетом MuMIn. Часть 1: Селекция моделей / R: Анализ и визуализация данных. - URL: https://r-analytics.blogspot.com/2018/01/mumin-1.html (дата звернення: 5.2.2021)

18. Мастицкий С., Шитиков В. R: Анализ и визуализация данных. - URL: http://r-analytics.blogspot.com (дата звернення: 5.2.2021)

19. Introduction to R / ©2021 DataCamp Inc. URL: https://learn.datacamp.com/courses/free-introduction-to-r (дата звернення: 5.2.2021)

20. Miller J. N., Miller J. C. Statistics and Chemometrics for Analytical Chemistry / Harlow : Prentice Hall, Pearson Education Limited,. 6th Edition. - 278 p. - ISBN: 978-0-273-73042-2

21. Иванчей И., Карпов А. Анализ данных в R / ©2013-2021. Stepik. URL: https://stepik.org/course/129/syllabus (дата звернення: 5.2.2021)

22. Brown S.D. Thechemometrics revolution re-examined / Journal of Chemometrics. 2017;31:e2856. doi: 10.1002/cem.2856.BROWN

23. Холін Ю. В., Пушкарьова Я. М., Пантелеймонов А. В., Некос А. Н. Хемометричні методи в розв’язанні задач якісного хімічного аналізу та класифікації фізико-хімічних даних : монографія / Х. : ХНУ імені В. Н. Каразіна, 2016. – 184 с. - ISBN 978-966-285-411-4

24. Емелин А. Статистические гипотезы / ©mathprofi.ru, Александр Емелин, 2010-2020 http://mathprofi.ru/statisticheskie_gipotezy.html [14.09.2020]

25. YukhymCommunity / ©2020 YukhymCommunity URL: https://yukhym.com [14.09.2020]

26. Дребущак Т. Н. Введение в хемометрику: Учеб. пособие / Новосиб. гос. ун-т. Новосибирск, 2013. 88 с.

27. Виноградова М.Г. Теория графов в химии // Международный журнал прикладных и фундаментальных исследований. – 2010. – № 12. – С. 140-142; URL: https://applied-research.ru/ru/article/view?id=1031 (дата звернення: 23.10.2020).

28. Брюске Э. Я. Химику о теории графов: графы в химической номенклатуре // Вестник российских университетов. Математика. 2003. №5. с.840-847. URL: https://cyberleninka.ru/article/n/himiku-o-teorii-grafov-grafy-v-himicheskoy-nomenklature (дата звернення: 23.10.2020).

29. Химические приложения топологии и теории графов : пер. с англ. / Под ред. Р. Кинга. - М.: Мир, 1987. 560 с.

30. Соколов В.И., Станкевич И.В. Фуллерены – новые аллотропные формы углерода: структура, электронное строение и химические свойства // Успехи химии. – 1993. – Т. 62, No 5. – С. 455–470.

31. Cattell R. B. The Scree Test For The Number Of Factors // Multivariate Behavioral Research. Routledge, 1966, v.1. n.2. 245-276

32. Scopus. ©2021 Elsevier URL: https://www.scopus.com (дата звернення: 5.2.2021).

33. R : A language and environment for statistical computing // R Core Team. R Foundation for Statistical Computing, Vienna, Austria. 2020. URL: https://www.R-project.org (дата звернення: 5.2.2021).

34. Шипунов А. Б. , Балдин Е. М., Волкова П. А., Коробейников А. И. , Назарова С. А., Петров С. В., Суфиянов В. Г. Наглядная статистика. Используем R, авторы Наглядная статистика. Используем R! / 13 июля 2014 г. 297 с. URL: https://cran.r-project.org/doc/contrib/Shipunov-rbook.pdf (дата звернення: 06.02.2021).

35. Супрунович С.В. Методичні рекомендації для лабораторних робіт з курсу «Статистичні та хемометричні методи в хімії» : практикум / Луцьк : ВНУ імені Лесі Українки. 2020. 87 с.

36. Супрунович С. В. Статистичні та хемометричні методи в хімії : дист. курс LMS Moodle. ВНУ імені Лесі Українки. URL: http://194.44.187.60/moodle/course/view.php?id=735 (дата звернення: 06.02.2021).

37. Денисенко В. Суммирование погрешностей измерений в системах автоматизации // Современные технологии автоматизации. - 2012. - № 1. - С. 92-100. - ISSN 0206-975X URL: https://www.cta.ru/cms/f/443123.pdf (дата звернення: 21.08.2021).

38. Джадд Д., Вышецки Г. Цвет в науке и технике / М.: «Мир», 1978. - 592 с.

39. Заляжных В. В. Статистическая обработка результатов испытаний (измерений) / ©В. В. Заляжных. URL: http://arhiuch.ru (дата звернення: 21.08.2021).

40. Тьюки Дж. Анализ результатов наблюдений: Разведочный анализ / М: Мир. 1981. – 694 с.

41. Эсбенсен К. Анализ многомерных данных. Избранные главы / Черноголовка: Изд-во ИПХФ РАН. 2005. - 160 с. ISBN 5-901675-50-9

42. Дёрффель К. Статистика в аналитической химии / М.: Мир. 1994. - 268 с. ISBN 5-03-002799-8

43. Шитиков В.К., Розенберг Г.С., Зинченко Т.Д. Количественная гидроэкология: методы системной идентификации. – Тольятти: ИЭВБ РАН, 2003. – 463 с. ISBN 5-93424-109-5 URL: http://www.ievbras.ru/ecostat/Kiril/Library/Book1/Content0/Content0.htm (дата звернення: 21.08.2021).

44. Oetting J. Data Visualization 101: How to Choose the Right Chart or Graph for Your Data / ©2020 HubSpot, Inc. URL: https://blog.hubspot.com/marketing/types-of-graphs-for-data-visualization (дата звернення: 21.08.2021).

45. Student. The Probable Error of a Mean / Biometrika. Vol.6, Iss.1. March 1908, p. 1–25. doi:10.1093/biomet/6.1.1

46. Statistics How To / ©2022 Statistics How To. URL: https://www.statisticshowto.com/ (дата звернення: 21.08.2021).

47. Smith M. K. Common mistakes in using statistics: spotting and avoiding them / Department of Mathematics The University of Texas at Austin. URL: https://web.ma.utexas.edu/users/mks/statmistakes/errortypes.html. 12.5.2011.

48. Планування експерименту на сімплексі : навч. посіб. / Уклад.: Гриців В.І., Денисюк Р.О. Житомир: Житомирський держ. унів. ім. Івана Франка, 2013. 42с. URL: http://eprints.zu.edu.ua/18444/1/19_Planuvannya_eksperymentu_2013_42.pdf (дата звернення: 21.08.2021).

Перелік ілюстрацій

Перелік таблиць

Навчальне видання

Супрунович С. В., Кормош Ж. О., Сливка Н. Ю.

Статистичні та хемометричні методи в хімії

Навчальний посібник

Друкується в авторській редакції