Міністерство освіти і науки України
Національний університет "Чернігівський колегіум" імені Т.Г. Шевченка

Бондар О.С.

СТАТИСТИЧНІ МЕТОДИ В ХІМІЇ ТА ФАРМАЦІЇ

Навчально-методичний посібник

Чернігів 2024

УДК 54:519.237-7(07)
Б93

Укладач:
к.т.н., доцент кафедри фізики та астрономії Національного університету «Чернігівський колегіум» імені Т.Г.Шевченка Бондар Олена Сергіївна

Бондар О.С.

Б93 Статистичні методи в хімії та фармації. Навчально-методичний посібник / О.С. Бондар. Чернігів: НУЧК, 2024. 110 с

Затверджено вченою радою природничо-математичного факультету Національного університету «Чернігівський колегіум» імені Т.Г.Шевченка. Протокол № 12 від 27.05.2024 р.

Рецензенти:
к.б.н., доцент кафедри хімії, технологій та фармації, доцент Ткаченко Світлана Валентинівна;

к.фарм.н., доцент кафедри хімії, технологій та фармації Вороніна-Тузовських Юлія Василівна

Навчально-методичний посібник «Статистичні методи в хімії та фармації» складено для здобувачів освіти, які навчаються за освітньо-професійною програмою Хімія, Середня освіта (Хімія), Фармація. Запропоновано основи статистики, розглянуто застосування статистичних методів у ході проведення науково-дослідних експериментальних робіт та подання результатів досліджень. Наведено основний теоретичний матеріал, розглянуто приклади, запропоновано практичні роботи для перевірки знань. Посібник також буде корисний вчителям хімії, учням старших класів природничого профілю закладів загальної середньої освіти та всім, хто цікавиться основами статистики.

ЗМІСТ

ВСТУП

Статистична обробка та статистичний аналіз результатів є важливою частиною виконання здобувачами вищої освіти експериментальних робіт: лабораторних, курсових, бакалаврських та магістерських. Опанування методів математичної обробки даних є важливою частиною теоретичної підготовки майбутнього фахівця (хіміка, фармацевта та ін.), оскільки дозволяє критично ставитися до отриманих даних, усвідомлювати, які висновки є достовірними, обчислювати похибку результатів вимірювання, встановлювати зв'язок між визначеними фізико-хімічним величинами.

Застосування статистичних методів в хімії та фармації є важливим компонентом при виборі оптимальних методів якісного і кількісного аналізу речовин, при розробці біологічно активних препаратів, є ключовим елементом оптимального планування досліджень. Статистичні методи дозволяють визначити рівняння зв'язку вхідних і вихідних параметрів, аналізувати параметри технологічного процесу, побудувати математичну модель процесу, або, іншими словами, встановити взаємну залежність між різними факторами і технологічними результатами процесу.

При статистичній обробці розглядається *статистична сукупність* експериментальних даних. Окремі одиниці статистичної сукупності називаються її членами, або *варіантами*. Загальна кількість варіант, що складають дану сукупність, називається її *обсягом*.

Варіанти - це окремі значення, які характеризують досліджуваний об`єкт та за якими утворюється статистична сукупність (об'єм розчину, що був використаний на титрування, вміст ліпідів в організмі, активність ферментних систем, вміст йонів у зразку природної води).

Характерні риси або предмети, за якими один об'єкт відрізняється від іншого, називаються *ознаками*. Статистична сукупність може бути утворена і

за кількома ознаками, складатися з декількох однорідних груп, що об'єднуються в єдиний комплекс щодо прийнятих в досліді умов. У таких випадках вона називається *статистичним комплексом.*

Яку б форму і зміст не приймала статистична сукупність, вона являє собою явище масове, для якого характерна наявність індивідуальності у його компонентів. Аналіз масових явищ вимагає від дослідника певних знань: вміння правильно узагальнювати і аналізувати результати масових спостережень; робити науково обґрунтовані висновки. Аналіз причинно-наслідкових відносин показує, що результати одиничних спостережень, як правило, не збігаються один з одним, варіюють від випадку до випадку в певних межах. Варіювання, або внутрішньогрупова мінливість - явище універсальне, характерне для всіх систем.

Початок систематичного дослідження задач, що відносяться до масових випадкових явищ і поява відповідного математичного апарату відносяться до XVII століття.

З історії розвитку статистичних методів

На початку XVII століття знаменитий італійський фізик Галілео Галілей вже намагався піддати науковому дослідженню помилки фізичних вимірювань,

розглядаючи їх як випадкові і оцінюючи їхню ймовірність. До цього ж часу відносяться перші спроби створення загальної теорії страхування, заснованої на аналізі закономірностей в таких масових випадкових явищах, як захворюваність, смертність, статистика нещасних випадків тощо.

В середині XVII століття завдяки дослідженням французьких математиків Б. Паскаля (1623 - 1662), П. Ферма (1601 - 1665) та нідерландського вченого Х. Гюйгенса (1629 - 1695) в області теорії азартних ігор виникла теорія ймовірності, сформувалися поняття

ймовірність, математичне сподівання, були встановлені їх основні властивості і прийоми їх обчислення.

Швейцарський математик Я. Бернуллі (1654 - 1705) довів одне з найважливіших положень теорії ймовірності - закон великих чисел. Англійського математик А. Муавра (1667—1754) ввів і для простого випадку обґрунтував нормальний закон (закон Гаусса). В 1795 К. Гауссом було запропоновано для обробки статистичних даних метод найменших квадратів. Важливий внесок у застосування статистичних методів зробив аналітик пивоварні Гіннес У. Госсета (W. Gosset), який публікував свої праці під псевдонімом Стьюдент.

На початку XX ст. з'явилась робота К. Пірсона (K. Pearson) «On lines and planes of closest fit to systems of points in space», де запропоновано метод головних компонент. Пізніше з'явилися роботи Р. Фішера (R. Fisher), де запропоновано методи факторного аналізу і методи планування експерименту.

В 1974 р. завдяки працями американського вченого Брюса Ковальскі (B. Kowalski) та шведського вченого Сванте Волда (S. Wold) з`явилася наука, яка безпосередньо аналізує хімічні дані – хемометрика. В статті «Chemometrics: what do we mean with it, and what do we want from it?» визначається її завдання - одержання хімічно важливої інформації з хімічних даних, організація і представлення цієї інформації. Бурхливий розвиток хемометрики відбувався у 70-х рр. XX століття завдяки появі у вчених доступу до потужної розрахункової техніки. Це дозволило втілити складні алгоритми обробки даних, особливо аналіз багатофакторних експериментів.

РОЗДІЛ 1. Групування результатів спостережень та їх графічне зображення

1.1. Таблиці та ряди розподілу

Результати дослідження фіксуються зазвичай в первинних документах - протоколах дослідів, лабораторних щоденниках, робочих журналах і т.п. Зібраний фактичний матеріал потім піддається статистичній обробці. Мета обробки - витяг з маси фактів укладеної в них інформації, отримання на підставі проведеного дослідження об'єктивних і переконливих висновків.

Перший крок на шляху статистичної обробки - групування зібраних даних відповідно до завдань і умов дослідження. Найбільш раціональною формою групування служать статистичні таблиці. У них зазвичай зводяться результати спостережень.

Статистичні таблиці бувають складні і прості, і їх будова залежить від того, за якими ознаками і за якою їх кількістю групується матеріал, а також від завдань, які вирішуються групуванням зібраного матеріалу. Приклад порівняно простого групування – обсяги виробництва амоніаку та сульфатної кислоти в Україні за 2003-2009 р. за даними Держкомстату України представлені у вигляді таблиці (табл 1.1).

Таблиця 1.1

Виробництво амоніаку та сульфатної кислоти за 2003-2009 рр., тис.тонн

Продукт	Обсяг виробництва (*тис. тон)*							Середнє значення
	2003	2004	2005	2006	2007	2008	2009	
Амоніак	4775	4779	5214	5147	5139	4890	3032	4710,85
Кислота сульфатна	1133	1425	1606	1493	1657	1479	890	1383,39

Прикладом складних таблиць, що ілюструють залежність однієї з варіюючих ознак від змін іншої, служать кореляційні таблиці, а також таблиці дисперсійних комплексів.

Найбільш просту форму статистичного групування представляють ряди розподілу, які будуються на основі операції ранжування, тобто шляхом розташування варіант (окремих числових значень варіюючої ознаки) в зростаючому або спадному порядку. Наприклад, є такий сукупність 20 вимірювань ознаки: 2, 5, 3, 6, 4, 7, 4, 5, 6, 6, 5, 9, 5, 6, 10, 8,12, 9, 7, 6.

Видно, що ознака варіює від 2 до 12 одиниць. Розташуємо цю сукупність у зростаючому порядку:

2, 3, 4, 4, 5, 5, 5, 5, 6, 6, 6, 6, 6, 7, 7, 8, 9, 9, 10, 12.

Вийде ранжирований ряд значень ознак.

При розподілі членів сукупності в ряд переслідуються певні цілі. Одна з них - розкриття закономірності варіювання досліджуваної ознаки. Тому до рядів розподілу пред'являються певні вимоги:

1) вони повинні бути легкодоступні для огляду;

2) добре ілюструвати закономірність варіювання.

Ранжований ряд погано задовольняє цим вимогам. Якщо ті ж варіанти розташувати у вигляді подвійного ряду, враховуючи їх повторюваність в загальному строю, сукупність розподілиться таким чином (табл 1.2):

Таблиця 1.2

Варіанти	2	3	4	5	6	7	8	9	10	11	12
Повторюваність варіант (р)	1	1	2	4	5	2	1	2	1	0	1

Такий упорядкований ряд розподілу, в якому вказана повторюваність варіант, що належать до даної сукупності, називається *варіаційним рядом.* Повторюваність варіант в сукупності називають *вагами або частотами.*

У статистиці ознаки позначають прописними буквами *X, Y, Z* ... їх числові значення - малими x_p, x_2, x_3, ... або y_p, y_2 y_3 ... Їх частоти позначаються латинською буквою р. Загальна кількість варіант, що входять до складу даної сукупності, називають обсягом сукупності і позначають літерами n або N.

$$\sum p = n \text{ [1]}.$$

Відносна частота або частість обчислюється як частка від ділення p / n і виражається в частках або відсотках.

$$\sum \frac{p}{n} = 1 [2].$$

Заміна абсолютних значень ознаки (частот) частостей полегшує зіставлення одного варіаційного ряду з іншим і робить більш виразними характерні риси варіювання.

1.2. Класифікація ознак та побудова варіаційних рядів

Хімічні ознаки поділяються на якісні та кількісні. До якісних належать такі ознаки, як забарвлення розчину, наявність осаду, виділення газу. Якщо мова йде про вимірювані або обчислювальні величини - це будуть кількісні ознаки. У варіаційні ряди розподіляються тільки кількісні ознаки, а якісні ознаки зазвичай розглядають в альтернативній формі. Кількісні ознаки можуть бути рахунковими (варіюють дискретно) та мірними (варіюють безперервно).

Одним з способів класифікації ознак є побудова варіаційних рядів. Існує два види варіаційних рядів: безінтервальні та інтервальні. Прикладом безінтервального варіаційного ряду може служити розподіл результатів, отриманих під час визначення показника кислотності зразків водопровідної води шляхом вимірювання pH водних проб з різних точок відбору (табл. 1.3).

Таблиця 1.3

Значення pH, (x)	6,6	6,7	6,8	6,9	7,0	7,1	7,2	7,3	7,4
Кількість одержаних результатів (p)	1	8	5	18	27	22	16	13	3

В даному випадку ознака варіює слабо. Але багато ознак варіюють у дуже широких межах і розподіл їх в безінтервальний ряд не досягає мети: ряди виходять занадто розтягнутими, погано доступним для огляду, що не

відображають чітко закономірності варіювання. Наприклад, вміст Феруму (%) в 50 зразках гірської породи (табл 1.4):

Таблиця 1.4

Вміст Fe, % (x)	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
Кількість зразків породи (p)	1	1	1	1	1	1	6	5	6	5	3	2	5	5	2	3	1	0	1

У таких випадках найкращий результат виходить від розподілу сукупності в інтервальний варіаційний ряд. Для цього вся варіація, ознаки від мінімуму до максимуму варіанти, розбивається на рівні інтервали (від і до) або класи. Потім всі варіанти розподіляють по класах і частоти *p* будуть частотами класів. Для обчислення середніх величин на таких дискретних варіаційних рядах використовуються не класові інтервали, а їх середні, рівні півсумі верхньої і нижньої меж класу.

Число класів залежить від завдання дослідження і характеру зібраного матеріалу. Ширина класового інтервалу позначається не тільки на характері розподілу варіант по класах, а й на точності середніх характеристик. Установка широких класових інтервалів спотворює типові риси варіювання і веде до зниження точності числових характеристик ряду. При виборі надмірно вузьких інтервалів точність узагальнюючих числових змінних підвищується, але ряд виходить занадто розтягнутим і не дає чіткої картини варіювання. Щоб визначити величину класового інтервалу для побудови доступного для огляду варіаційного ряду Г.А. Стерджесом (Sturges, 1926) запропоновано наступну формулу:

$$i = \frac{x_{max} - x_{min}}{1 + 3{,}32 \lg_n} \text{ або } \quad i = \frac{x_{max} - x_{min}}{5 \cdot \lg_n}$$

Формулу [4] рекомендується використовувати при наявності в сукупності великого числа членів (n > 100).

Отже, розглянемо побудову варіаційного ряду: x_{min} = 9, x_{max} = 27; величина класового інтервалу буде наступною: i = 3 [3] і i = 2 [4]. Візьмемо

$i = 3$. При поділу варіації на класи, межі першого класу встановлюємо так, щоб мінімальна варіанту потрапила приблизно в середину цього класу. Якщо нижня (мінімальна) межа 1-го класу дорівнює 8, то вийде 8 класів з нижніми межами, рівними 8, 11, 14, 17, 20, 23, 26, 29. Щоб однозначно вирішити питання про приналежність варіанти до якогось класу, його верхню межу зменшують на 0,1 або 0,01, що і дає необхідне розмежування класів (табл. 1.5).

Таблиця 1.5

Класи по вмісту Fe у зразках руди	Середнє значения класів (x)	Частоти *(p)*

В результаті виходить інтервальний варіаційний ряд з переривчастим варіюванням. При побудові варіаційного ряду не допускається подвійний облік однієї і тієї ж варіанти.

1.3. Графіки розподілу

Щоб надати більшу наочність закономірності варіювання ознак, варіаційні ряди прийнято зображати графічно у вигляді гістограми, або полігону, а також у вигляді кумуляти або огіви.

Гістограма – графік розподілу частот, виходить, якщо в системі координат по осі абсцис відкласти межу класів, а по осі ординат – їх частоти.

У випадку з розподілом класів вмісту Феруму у зразках руди гістограма буде виглядати, як показано на рис. 1.1.

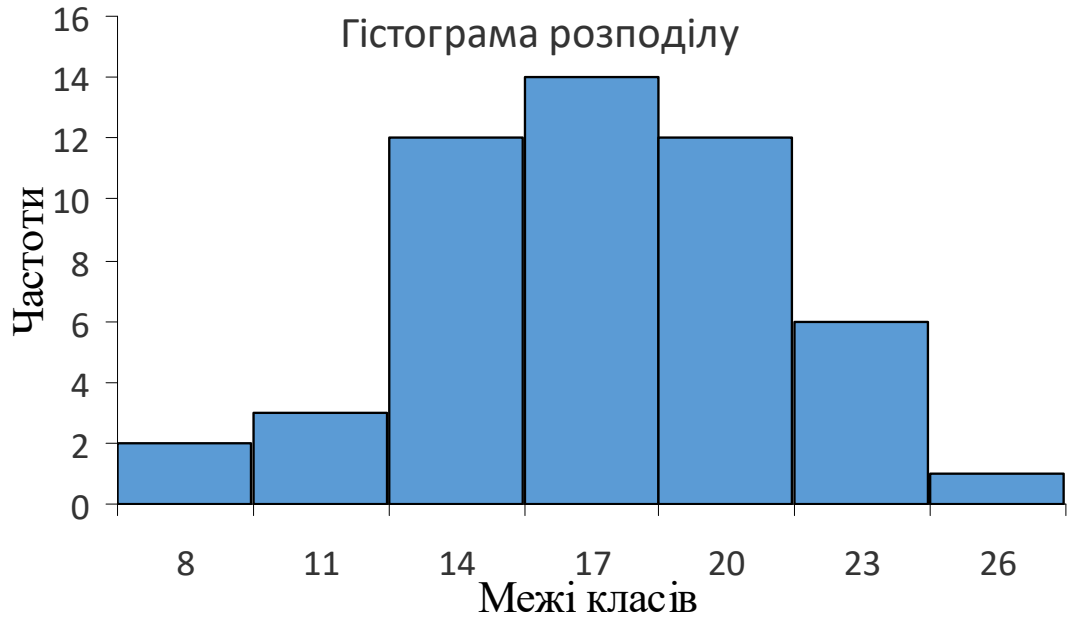


Рис . 1.1. Гістограма розподілу класів зразків вмісту Fe у зразках руди

Гістограма зображує закономірності розподілу варіант по класах варіаційного ряду, тобто при безперервному варіюванні ознаки. Прямокутники відповідають класам, а їх висота – частотам варіаційного ряду.

Якщо з середніх точок прямокутників гістограми опустити перпендикуляри на вісь абсцис, а самі точки з'єднати між собою, вийде графік дискретного варіювання, званий *полігоном розподілу* (рис. 1.2).

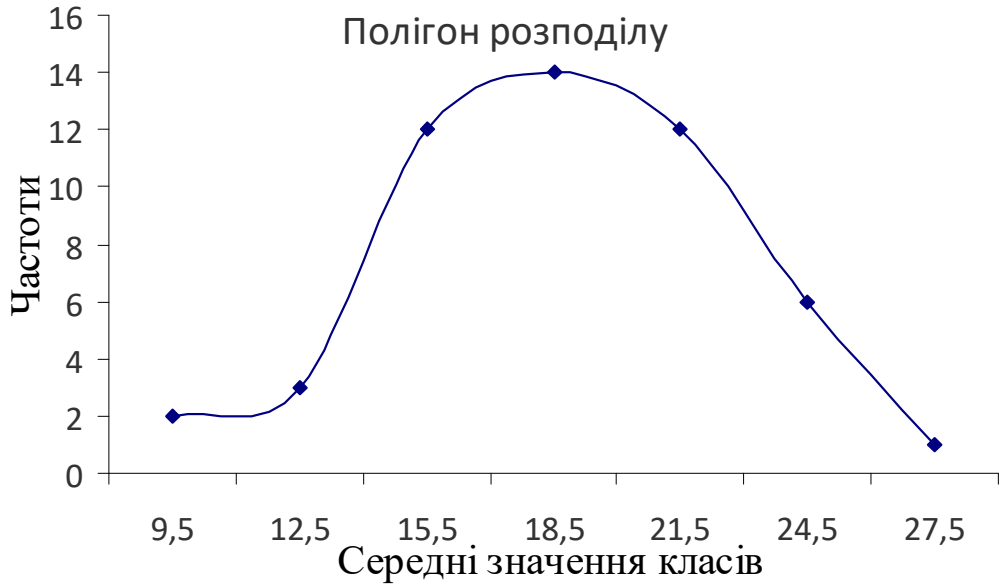


Рис. 1.2. Полігон розподілу класів вмісту Феруму у зразках руди

Полігон розподілу можна побудувати незалежно від гістограми, завдаючи на вісь абсцис середні значення класів. А коли виникає необхідність, можна полігон перетворити в гістограму.

В інших випадках графік варіаційного ряду будується в вигляді *кумуляти* (від слова cumulo - накопичувати). Для побудови кумуляти розподілу вмісту Феруму у зразках руди скористаємося даними табл. 1.6. При цьому по осі абсцис відкладають значення класових варіант, а по осі ординат - накопичені частоти (рис. 1.3).

Таблиця 1.6

Середнє значення класів	Частоти (P)	Накопичення частоти
9,5	2	
12,5	3	
15,5	12	
18,5	14	
21,5	12	
24,5	6	
27,5	1	
Сума	50	

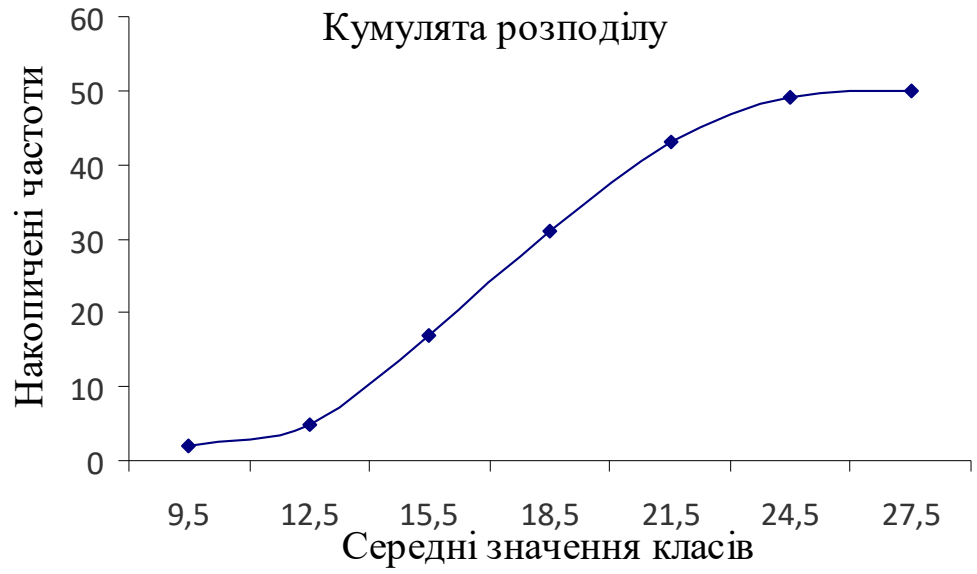


Рис. 1.3. Кумулята розподілу класів вмісту Феруму у зразках руди

Якщо ряд накопичених частот нанести на вісь абсцис, а значення варіант розташувати по осі ординат, то такий графік має назву *огіва* (рис. 1.4).

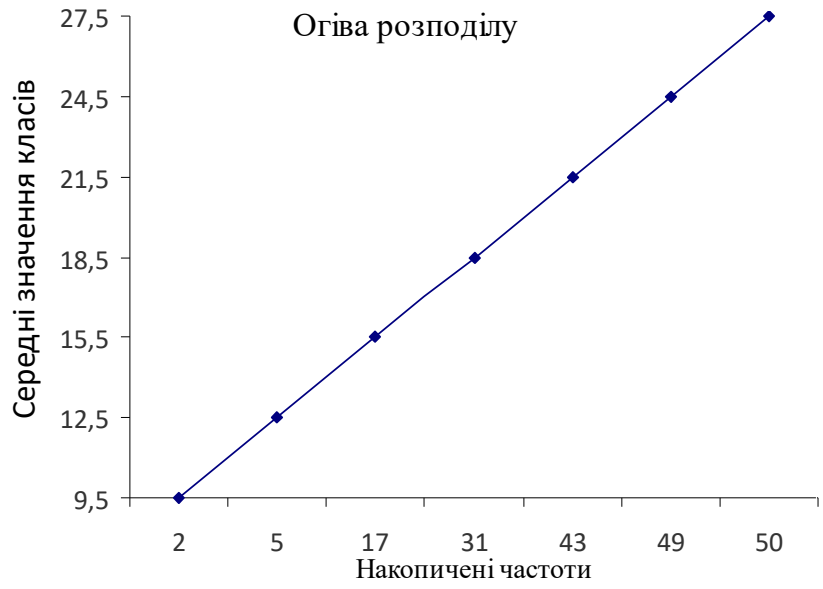


Рис. 1.4. Огіва розподілу класів вмісту Феруму у зразках руди

Значення графіків полягає в їх наочності. Але вони не дають точної характеристики варіюючої ознаки, так як залежать від прийнятих масштабів. Точну характеристику варіюючих ознак дають статистичні показники, про які йтиметься далі.

ПРАКТИЧНА РОБОТА № 1

1. Побудувати ранжований ряд
2. Побудувати варіаційний ряд
3. Побудувати графіки розподілу (гістограму, кумуляту, огіву, полігон розподілу)

Варіант 1

Вміст алюмінію у сплаві, %:
72 73 71 71 55 72 73 65 75 73 86 65 67 63 75 73 77 75 56 75 60 78 71 69 66 64 73 63 63 74 63 69 66 68 69 58 71 75 61 67 69 72 56 68 71 64 72 73 74 76 65 63 72 80

Варіант 2

Кількість амінокислотних залишків у молекулі поліпептиду:
69 64 72 71 74 76 70 73 71 69 55 72 71 65 75 73 60 76 71 69 63 63 72 78 66 64 71 63 63 72 63 69 64 68 69 56 71 75 71 86 65 64 63 73 73 77 73 56 59 67 69 70 56 68

Варіант 3

Вміст діючої речовини у препараті, мг:
98 87 109 109 120 99 84 99 102 88 102 96 101 114 105 102 100 101 99 99 104 101 96 111 101 96 95 92 105 110 99 100 92 98 91 84 93 102 115 113 112 118 107 116 107 102 95 97 96 99 105 98 105 100 102 112 119 111

Варіант 4

Вміст ефірної олії у суцвіттях м’яти, %:
5,21 4,21 4,61 4,10 4,81 4,92 5,01 4,99 5,05 4,01 5,02 3,83 4,70 4,82 4,41 5,40 5,41 5,22 4,59 5,06 5,23 5,18 4,50 4,83 3,85 4,93 4,77 4,40 5,04 4,97 4,49 4,30 4,98 4,99 4,69 4,62 5,45 4,26 4,63 5,61 4,51 5,19 4,31 5,28 4,96 5,62 4,84 4,79 5,00 4,38 5,20 5,29 4,53 4,95 4,85 4,47

Варіант 5

Вихід продукту реакції, %:
69 64 72 71 74 76 70 73 71 69 55 72 71 65 75 73 60 76 71 69 63 63 72 78 66 64 71 63 63 72 63 69 64 68 69 56 71 75 71 86 65 64 63 73 73 77 73 56 59 67 69 70 56 68

Варіант 6

Концентрація солі у розчині, мг/л:

104 95 100 98 99 108 100 105 105 102 112 126 100 102 111 104 96 104 116 105 105 102 101 99 106 101 100 111 111 120 102 86 99105 96 97 94 107 112 101 102 92 102 91 86 95 106 115 117 112 120 109 118 107 104 111 87 112

Варіант 7

Вміст діючої речовини у препараті, мг:
11 11 7 7 17 10 4 7 8 15 7 15 5 11 11 10 14 14 10 6 7 11 12 9 13 11 11 14 13 12 7 9 15 7 7 9 13 9 8 14 13 15 11 10 14 9 11 13 10 11 7 9 9 7

Варіант 8

Втрата маси зразку після корозивних досліджень, мг:
21 23 23 20 18 21 25 19 29 24 23 21 27 21 20 20 16 12 24 21 20 15 21 22 18 19 17 20 20 22 21 21 20 20 21 13 15 14 21 20 22 18 21 18 13 15 22 16 21 14 18 21 25 13

Варіант 9

Концентрація солі у розчині, мг/л:
28 25 26 28 21 22 24 22 24 23 28 23 23 24 23 23 23 23 21 35 34 33 30 28 34 37 32 31 23 28 33 22 32 34 24 32 24 27 23 23 33 27 34 21 32 24 36 27 29 22 23 24 22 24

Варіант 10

Концентрація лугу у розчині, мг/л:
21 22 24 26 23 22 24 21 28 30 21 20 21 22 23 18 29 24 20 24 20 22 16 23 24 23 22 21 21 30 20 25 18 22 21 19 22 23 23 19 21 22 27 21 21 21 23 26 16 21 21 25 24 17

Варіант 11

Концентрація кислоти у розчині, мг/л:
56 55 66 58 55 54 53 55 64 62 53 52 50 70 53 57 69 51 54 63 35 28 66 60 50 39 55 59 55 72 52 57 44 46 45 54 45 48 42 48 34 63 60 64 42 56 55 37 50 47 52 74 57 50

Варіант 12

Вміст кисню у газовій суміші, %:
25 16 19 13 32 20 17 22 32 14 19 13 22 17 29 32 17 11 14 17 22 21 11 24 13 27 33 15 14 23 13 34 13 17 19 34 17 15 21 24 31 14 25 34 29 14 22 25 27 14 17 22 25 28

РОЗДІЛ 2. ЙМОВІРНІСТЬ. РОЗПОДІЛ ТА ЙОГО ВИДИ

2.1. Імовірність і її властивості

Поняття імовірності є числовою характеристикою можливості настання випадкової події.

Хімічна наука також широко використовує концепцію ймовірності. Так, випадковою подією може бути: утворення побічного продукту реакції, виявлення сполуки з високим рівнем біологічної активності. Теоретичне значення відносної частоти очікуваної події називається його ймовірністю. Подія - це той результат, який виходить при кожному випробуванні. Якщо при кожному випробуванні подія неминуче настає, вона називається достовірною. Якщо немає - неможливою. Якщо і може наступити, і не може - це випадкова подія. Події, які при випробуванні в постійних умовах повторюються багаторазово, називаються масовими.

Ймовірністю називається відношення числа випадків або випадків m сприяють настанню очікуваної події A, до числа всіх можливих і несумісних в даному випробуванні результатів n, тобто:

$$P(A) = \frac{m}{n} [5]$$

Для зручності, ймовірність очікуваної події прийнято позначати через p, а ймовірність протилежної події через q, тоді:

$P(A) = p$ [6], $P(\acute{A}) = q$ [7], $p + q = 1$ [8],

де (A ¯) - ймовірність протилежної події (*не A*).

Все, що можна підрахувати або виміряти, називається величиною. Величини (або "ознаки") діляться на постійні і змінні. Постійною називається величина, яка в заданих умовах не змінює свого значення. Змінна - це така величина, яка в даних умовах здатна приймати різні числові значення.

Змінна величина називається випадковою, якщо в заданих умовах вона може приймати то одні, то інші значення.

Випадкова величина в N повторних випробуваннях може приймати самі різні значення, але в кожному окремому випробуванні вона приймає

завжди тільки одне з можливих значень. Яке значення прийме випадкова величина в результаті кожного випробування, заздалегідь сказати неможливо. Тому, характеризувати випадкову величину можна лише з певною ймовірністю, тобто вказуючи ймовірність її можливих значень.

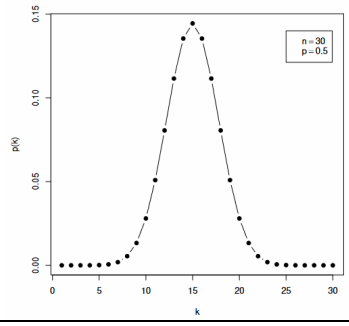
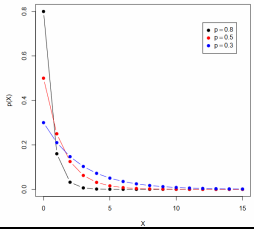
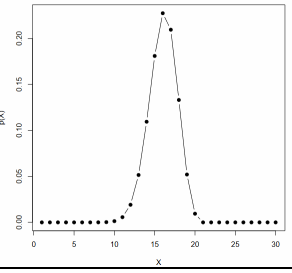
2.2. Розподіл та його види

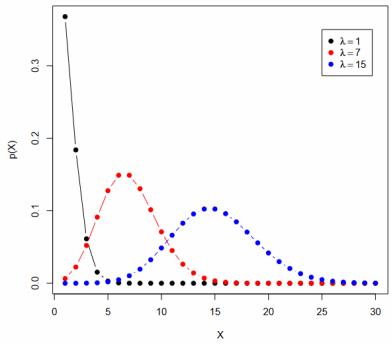
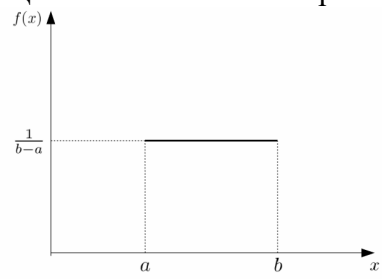
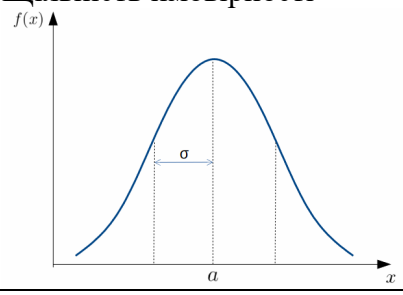
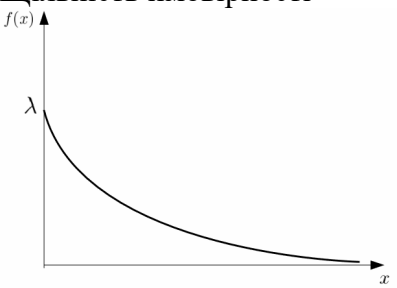
Розподіл випадкової величини – будь яке співвідношення, що встановлює зв'язок між можливими значеннями випадкової величини та відповідними ймовірностями.

Основні види розподілу представлено в табл. 2.1

Таблиця 2.1.

Види розподілу

Назва	Характеристика	Вигляд графіка
Біноміальний розподіл	Дискретний розподіл імовірностей із параметрами n і p для кількості успішних результатів, що мають двійкове значення у послідовності із n незалежних експериментів, для кожного з яких ставиться питання "так або ні". Імовірність виникнення успішного результату для кожного випробування задається параметром p, а імовірність виникнення не успішного результату відповідно дорівнюватиме $q = 1 - p$.	Залежність ймовірності успіху (p) від число успіхів (x) 
Геометричний розподіл	Розподіл при якому дискретна випадкова величина X має геометричний розподіл з параметром p, якщо вона збігається з кількістю випробувань до першого успіху в нескінченній послідовності випробувань Бернуллі з імовірністю успіху в одному випробуванні. Зустрічається в мікробіології, генетиці, фізиці.	Залежність ймовірності появи події (p) від кількості спроб, якає успішною (k) 
Гіпергеомет-ричний розподіл	Розподіл, що моделює кількість успішних вибірок без повернення зі скінченної сукупності. описує ймовірність того, що у вибірці з n різних об'єктів, витягнутих із сукупності, рівно k об'єктів є особливими. Загалом, якщо випадкова величина X відповідає гіпергеометричному розподілу з параметрами N, D та n, то ймовірність отримання рівно k успіхів.	Залежність ймовірності появи події (px) від кількості особливих об'єктів у вибірці 

Розподіл Пуасона	Розподіл, справедливий для подій, які мають малу ймовірність чи трапляються нечасто. 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 рівно k.	Залежність ймовірності появи події (p) від ймовірність того, що в даній серії випробувань подія (x) для різних параметрів λ 
Рівномірний розподіл	Розподіл випадкової величини, коли вона з однаковою ймовірністю може приймати будь-яке значення в заданих межах. При цьому ймовірність будь-якого інтервала залежить тільки від його довжини	Щільність ймовірності 
Нормальний розподіл	Розподіл, який вникає тоді, коли випадкова величина являє собою суму великого числа незалежних випадкових величин, кожна з яких відіграє незначну роль в утворенні всієї суми. Більшість величин в природі, фармації та хімії мають цей розподіл	Щільність ймовірності 
Показниковий (експоненційний) розподіл	Абсолютно неперервний розподіл, що моделює час між двома послідовними завершеннями однієї і тієї ж події. За цим законом розподілені небесні тіла у всесвіті, доходи людей у суспільстві та ін.	Щільність ймовірності 

У розподілі емпіричних сукупностей для більшості природних ознак є особливість – переважне накопичення варіант в центральних класах і поступове зменшення їх числа в міру віддалення від серединної точки варіаційного ряду. Ця особливість становить одну з характерних рис варіювання ознак. У різних випадках проявляється одна і та ж закономірність: в масі відносно однорідних одиниць (варіант) переважна

більшість складають варіанти середнього розміру, і чим далі вони відхиляються від середнього рівня ознаки, тим рідше зустрічаються в даній сукупності. Отже, між значеннями ознаки та їх зустрічаємістю існує певний зв'язок. Наочним виразом зв'язку з цим служить варіаційний ряд і його графік - варіаційна крива, яка для нормального розподілу має вигляд, представлений на рис 2.1.

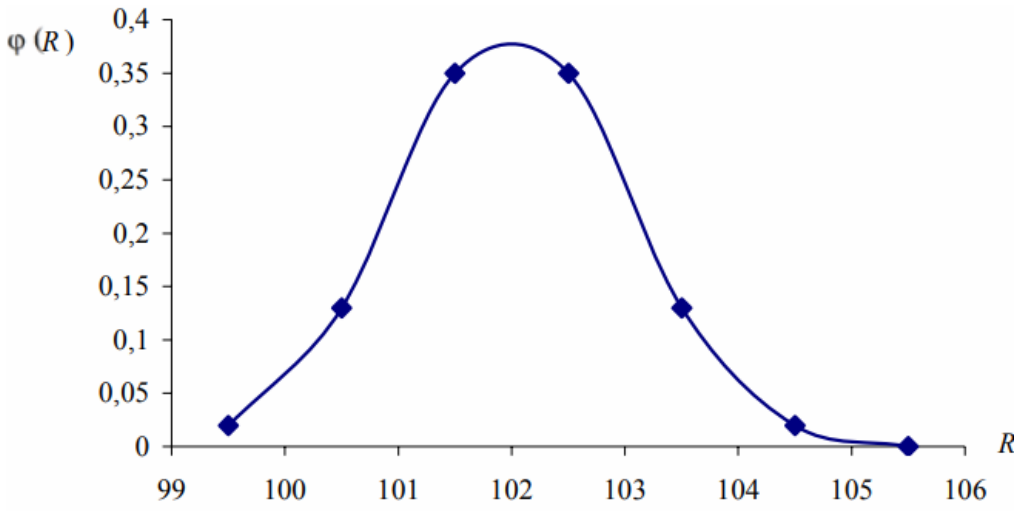


Рис. 2.1. Залежність функції щільності ймовірності від електричного опору.

Найбільш важливим та часто спостережуваним у природі розподілом випадкових величин є нормальний розподіл (розподіл Гауса). Цей вид розподілу відкрито трьома вченими в різний час: Муавром у 1733 році в Англії, Гауссом у 1809 році в Німеччині та Лапласом у 1812 році у Франції. Нормальний розподіл є найбільш теоретично вивченим та зручним для математичного аналізу даних.

Нормальний розподіл – розподіл ймовірностей випадкової величини, що виникає тоді, коли дана випадкова величина являє собою суму великого числа незалежних випадкових величин, кожна з яких грає в утворенні всієї суми незначну роль. Результати будь-яких повторних вимірювань, проведених над одним і тим же об’єктом, мають нормальний розподіл. Наприклад, закону нормального розподілу підкоряється вихід продукту реакції в одних і тих же умовах.

Нормальний розподіл має такі характеристики:

1. Крайні значення ознаки в ньому досить рідкісні, а значення, близькі до середньої величини – досить часті.

2. Графічно нормальний розподіл має вигляд кривої дзвоноподібного типу (рис. 2.2)

3. Крива розподілу симетрична щодо центру.

4. Середнє арифметичне значення, мода і медіана у нормальному розподілі рівні.

5. Форму розподілу можна описати двома параметрами: середнім арифметичним значенням і стандартним відхиленням.

Нормальний розподіл завжди виникає тоді, коли дана випадкова величина являє собою суму великого числа незалежних випадкових величин, кожна з яких грає в утворенні всієї суми незначну роль.

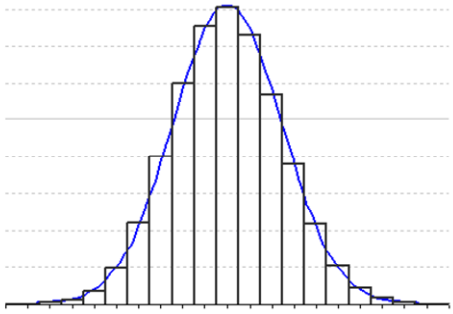


Рис. 2.2. Крива нормального розподілу

Крива нормального розподілу (або крива Гаусса) характеризується тим, що існує певне центральне максимальне значення досліджуваного параметра, і частота зустрічі інших значень тим менше, чим далі це значення від центрального.

РОЗДІЛ 3. СЕРЕДНІ ВЕЛИЧИНИ

Однією з найважливіших узагальнюючих характеристик варіюючих ознак є середня величина. Характеризуючи ту чи іншу групу результатів, кажуть, наприклад, про середній вихід продукту реакції, середній вміст йонів у воді, про середню швидкість біохімічної реакції і т.д. Значення середніх полягає в їх властивості нівелювати індивідуальні відмінності, в результаті чого виступає більш-менш стійка числова характеристика ознаки - не окремих представників, а цілої групи статистичних одиниць.

3.1. Середня арифметична

а) Проста середня арифметична

$$x = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum x_1}{n} [9]$$

б) Зважена середня арифметична

Якщо в сукупності спостережень окремі варіанти повторюються *p* раз, то середня арифметична обчислюється за формулою [10] з урахуванням повторюваності (або "ваг") окремих варіант (табл 3.1).

$$0x = \frac{x_1 p_1 + x_2 p_2 + x_3 p_3 + \ldots + x_n p_n}{p_1 + p_2 + p_3 + \ldots p_n} = \frac{\sum x_1 p_1}{p_1} [10]$$

Так як середня обчислюється з урахуванням частот або «ваг» окремих варіант, то вона називається зваженою середньою.

Таблиця 3.1

Вміст заліза в зразку руді, % (варіанти) *(x)*	7	8	9	10	11	12	13
Кількість зразків руди (повторюваність) (*p*):	1	1	2	7	3	3	1

$x = (7\times1 + 8\times1 + 9\times2 + 10\times7 + 11\times3 + 12\times3 + 13\times1)/18 = 185/18 = 10{,}28$ (%)

Аналогічним способом розраховується і загальна середня ($\bar{x}$) з суми приватних середніх.

$$x = \frac{x_1 n_1 + x_2 n_2 + x_3 n_3 + \ldots + x_k n_k}{n_1 + n_2 + n_3 + \ldots n_n} = \frac{\sum x_1 n_1}{n_1} [11]$$

3.2. Скорочений спосіб обчислення середньої арифметичної (спосіб умовної середньої)

Обчислення середньої арифметичної способом зваженої середньої не завжди зручно, особливо на сукупностях великого обсягу і при наявності багатозначних чисел, коли обчислювальна робота стає особливо трудомісткою. У таких випадках простіше розрахувати середню арифметичну спрощеним способом - способом умовної середньої.

В такому випадку одна з варіант умовно приймається за середню величину, що позначається через *A*. Зазвичай в якості умовної середньої береться варіанти (або клас) з найбільшою або близькою до неї частотою, хоча це не обов'язково. Потім розраховують відхилення варіант (або класів) від цієї умовної середньої і знаходять середню арифметичну за такою формулою:

$$x = A + \frac{\sum pa}{n} [12]$$

де А - умовна середня, $a = x\text{-}A$ - відхилення варіанти від умовної середньої.

Наприклад, скористаємося наведеної формулою для визначення середнього значення pH 18 досліджених розчинів (табл. 3.2):

Таблиця 3.2

Варіанта (x):	7	8	9	10	11	12	13
Повторюваність (p):	1	1	2	7	3	3	1
a:	-4	-3	-2	-1	0	+1	+2
pa:	-4	-3	-4	-7	0	+3	+2

$$\sum pa = +5 - 18 = -13$$

$$x = 11 + \frac{-13}{18} = 11 - 0{,}72 = 10{,}28$$

Якщо ж сукупність досить велика, то дані розрахунки простіше проводити за допомогою таблиць. Наприклад, при вивченні жорсткості (ммоль/дм^3) 100 зразків води отримали значення від 8,99 до 14,7. Розбивши

варіацію на класи відповідно до формул [3] або [4], побудуємо таблицю і рознесемо відповідні дані експерименту (табл. 3.3).

Таблиця 3.3

Середні значення класів або класові варіанти	Частоти	Відхилення класових варіант від умовної середньої	Похідні відхилень на частоти	$a' = \frac{x - A}{i}$	a'
(x)	*(P)*	*(a=x- A)*	*(pa)*		
8,9	2	-2,8	-5,6	-4	-8
9,6	3	-2,1	-6,3	-3	-9
10,3	9	-1,4	-12,6	-2	-18
11,0	17	-0,7	-11,9	-1	-17
A = 11,7	25	0	0	0	0
12,4	23	+0,7	+ 16,1	+ 1	+23
13,1	10	+1,4	+ 14,0	+2	+20
13,8	7	+2,1	+14,72	+3	+21
14,5	4	+2,8	11,2	+4	+ 16
Сума	100	-	+ 19,6	-	+28

$$x = A + \frac{\sum pa}{n} = 11,7 + \frac{19,6}{100} = 11,896 \approx 11,90\left(\text{ммоль/дм}^3\right)$$

Якщо замість, $a = x\text{-}A$ використати $a' = \frac{x - A}{i}$, де i - величина класового інтервалу (i = 0,7) і формулу:

$$\acute{x} = A + i\frac{\sum pa'}{n} \quad [13]$$

то результат обчислень виявиться таким самим:

$$x = 11,7 + 0,7\left(\frac{28}{100}\right) = 11,7 + 0,196 = 11,896 \approx 10,90\left(\text{ммоль/дм}^3\right)$$

Порівнюючи перший і другий спосіб розрахунку середньої, бачимо, що другий спосіб набагато простіше.

РОЗДІЛ 4. ПОКАЗНИКИ ВАРІАЦІЇ

Середня арифметична - найважливіша статистична характеристика, але вона нічого не говорить про величину варіювання ознаки. Одним з показників варіації служать ліміти, що показують *min* і *max* варіанти сукупності, а також розмах варіації (*R*), що представляє собою різницю між максимальною і мінімальною варіантами сукупності, тобто $R = -x_{max}-x_{min}$. Але вони не здатні характеризувати істотні риси варіювання, а, крім того, можуть сильно змінювати своє значення.

4.1. Дисперсія і середнє квадратичне відхилення

Найбільш зручною мірою варіювання служить середній квадрат відхилень або дисперсія:

$$\sigma^2 = \frac{\sum\left(x_i - \bar{x}\right)^2}{n-1} \qquad [14]$$

Враховуючи повторювальність відхилень:

$$\sigma^2 = \frac{\sum pa^2}{n-1} \qquad [15]$$

Середнє квадратичне відхилення обчислюється за формулою:

$$\sigma = \sqrt{\frac{\sum\left(x_i - \bar{x}\right)^2}{n-1}} \qquad [16]$$

Величина n-1 має назву числа ступенів свободи. Середнє квадратичне відхилення (стандартне) - величина іменована і виражається в тих же одиницях виміру, що і ознака. Чим сильніше варіює ознака, тим більше і величина середнього квадратичного відхилення і навпаки. Обчислимо середнє відхилення для рядів *x* і *y* (табл. 4.1, 4.2).

Алгоритм розрахунку середнього квадратичного відхилення для ряду *x:*

Таблиця 4.1

(x):	10	15	20	25	30	35	40	45	50	$\acute{x}$= 30
(a):	20	15	10	5	0	5	10	15	20	
(a^2):	400	225	100	25	0	25	100	225	400	$\sum a^2$=1500

$$\sigma_x = \sqrt{\frac{1500}{9-1}} = \sqrt{187.5} = 13.7$$

Алгоритм розрахунку середнього квадратичного відхилення для ряду *y:*

Таблиця 4.2

(y):	10	28	28	30	30	30	32	32	50	$\acute{y}$= 30
(a):	20	2	2	0	0	0	2	2	20	
(a^2):	400	4	4	0	0	0	4	4	400	$\sum a^2$=816

$$\sigma_y = \sqrt{\frac{816}{9-1} = \sqrt{102}} = 10.1 .$$

Середні квадратичні відхилення для рядів *x* і *y* відрізняються один від одного і характеризують специфіку варіювання ознаки. При великій кількості спостережень різниця між *n* і *n-1* істотно не позначається на величині показника варіації, тому при *n > 30* замість *n-1* можна брати значення *n*.

Середня арифметична і середнє відхилення дають повну кількісну характеристику будь-якій емпіричній сукупності, що розподіляється по нормальному закону. Середня арифметична відображає дію на ознаку основних причин, що визначають типовий для популяції рівень його розвитку, тоді як середнє відхилення характеризує варіювання значень цієї ознаки навколо центру розподілу, тобто середньої арифметичної. Середнє квадратичне відхилення є мірою ступеня впливу на ознаку різних другорядних причин, що викликають його варіювання.

Таким чином, ці показники, хоча і відображають різні сторони варіюючих ознак, тісно пов'язані між собою.

4.2. Обчислення середнього квадратичного відхилення (спосіб умовної середньої)

Формула розрахунку середнього квадратичного відхилення методом умовної середньої набуває такого вигляду:

$$\sigma = \sqrt{\frac{\sum pa^2}{n} - \left(\frac{\sum pa}{n}\right)^2}$$ [17], де a=x-A

Покажемо застосування цієї формули на прикладі розподілу жорсткості (ммоль/дм^3) зразків води використовуючи наступний алгоритм (табл. 4.3).

Таблиця 4.3

Класові варіанти *(x)*	Частоти *(p)*	Відхилення класових варіант від умовної середньої *(a = x - A)*	Похідні відхилень на частоти *(pa)*	pa^2
8,9	2	-2,8	-5,6	15,86
9,6	3	-2,1	-6,3	13,23
10,3	9	-1,4	-12,6	17,64
11,0	17	-0,7	-11,9	8,33
A = 11,7	25	0	0	0
12,4	23	+0,7	+ 16,1	11,27
13,1	10	+ 1,4	+ 14,0	19,60
13,8	7	+2,1	+14,7	30,87
14,5	4	+2,8		31,36
Сума	100	-	+19,6	147,98

Підставивши отримані значення з таблиці в формулу [17] отримаємо значення середнього квадратичного відхилення:

$$\sigma = \sqrt{\frac{147.98}{100} - \left(\frac{19.6}{100}\right)^2} = \sqrt{1.48 - 0.04} = \sqrt{1.44} = 1.20(\text{ммоль/дм}^3)$$

Якщо відхилення класових варіант від умовної середньої *A* віднести до величини класового інтервалу, тобто замість, $a = x\text{-}A$ взяти $a' = \frac{x_i - A}{i}$, то розрахунки за формулою:

$$\sigma = i\sqrt{\frac{\sum pa^2}{n} - \left(\frac{\sum pa}{n}\right)^2} \quad [18]$$

значно спростяться, що видно з наступного прикладу (табл 4.4):

Таблиця 4.4

Класові варіанти (x)	Частоти (p)	$a' = \frac{x - A}{i}$	(pa')	pa'^{2}
8,9	2	-4	-8	32
9,6	3	-3	-9	27
10,3	9	-2	-18	36
11,0	17	-1	-17	17
A = 11,7	25	0	0	0
12,4	23	+ 1	+23	23
13,1	10	+2	+20	40
13,8	7	+3	+21	63
14,5	4	+4	+16	64
Сума	100	-	+28	302

$$\sigma = 0.7\sqrt{\frac{302}{100} - \left(\frac{28}{100}\right)^2} = 0.7\sqrt{2.94} = 0.7 \times 1.715 = 1.20(\text{ ммоль/дм}^3)$$

На вибірках невеликого обсягу, особливо коли вони не згруповані в варіаційний ряд, середньоквадратичне відхилення обчислюється тим же способом за такою формулою:

$$\sigma = \sqrt{\frac{n}{n-1}\left[\frac{\sum a^2}{n} - \left(\frac{\sum a}{n}\right)^2\right]} \quad [19],$$

де $a=x-A$.

Наприклад, вихід продукту реакції (%) в 9 паралельних дослідах (табл.4.5):

Таблиця 4.5

x:	74	70	77	72	79	88	76	80	86	
a:	0	-4	+3	-2	+5	+14	+2	+6	+12	$\sum a = 36$
a^2:	0	16	9	4	25	196	4	36	144	$\sum a^2 = 434$

звідки:

$$\sigma = \sqrt{\frac{9}{9-1}\left[\frac{434}{9} - \left(\frac{36}{9}\right)^2\right]} = \sqrt{\frac{9}{8}(32.22)} = \sqrt{36.25} = 6.02$$

4.3. Коефіцієнт варіації

Середнє квадратичне відхилення є основним показником варіабельності ознак. Він не залежить від числа спостережень, і тому може використовуватися для порівняльної оцінки варіювання однорідних ознак. Разом з тим широкому використанню середнього квадратичного відхилення в якості запобіжного порівняння варіабельності ознак заважає те, що цей показник є величиною іменованою. Знаючи що середнє відхилення ряду розподілу жорсткості зразків природньої води становить 1,20 ммоль/дм^3, а варіювання вмісту Мангану в земній корі характеризується значенням середнього квадратичного відхилення 1,35% не можна порівнювати варівання цих ознак.

Щоб середньоквадратичне відхилення могло бути використано в якості порівняння варіабельності ознак, воно повинно бути неіменованим. Для цього використовують коефіцієнт варіації:

$$CV = \frac{\sigma}{\bar{x}} \times 100\% \text{ або } CV = \frac{\sigma}{\bar{x}} \text{ [20]}$$

Так,

$$CV_1 = \frac{1{,}20}{11{,}9} \cdot 100\% = 10{,}1\%$$

$$CV_2 = \frac{1{,}35}{114{,}74} \cdot 100\% = 1{,}2\%$$

Із зіставлення цих показників видно, що перша ознака варіює сильніше, ніж друга. При нормальному розподілі коефіцієнт варіації зазвичай не перевищує 45-50% і часто буває набагато нижче цього рівня. У випадках же асиметричних розподілів він може бути досить високим, що досягає іноді 100% і вище.

4.4. Нормоване відхилення

Нормоване відхилення – показник, що визначає на скільки та чи інша варіанта відхиляється від середнього рівня варіюючої ознаки. У найпростішому вигляді нормоване відхилення виглядає так:

$$t = \frac{x_i - \bar{x}}{\sigma} \quad [21]$$

Наприклад, відомо, що середня потужність установки для синтезу амоніаку становить 164,8 т/год при σ = 5,8 т/год. Оцінимо, наскільки потужність нової установки (171,2 т/год) відрізняється від середньої:

$$t = \frac{171,2 - 164,8}{5,8} = +1,1$$

Оскільки будь-яка варіанта, що належить до сукупності, що розподіляється по нормальному закону, може відхилятися від середньої до трьох *a*, знайдена величина вказує на незначне збільшення потужності установки для синтезу амоніаку в порівнянні із середнім рівнем потужності.

ПРАКТИЧНА РОБОТА 2

Розрахувати середнє квадратичне відхилення методом умовної середньої

Варіант 1

100 95 105 92 90 101 105 100 105 98 102 104 106 100 105 100 106 105 103 100

Варіант 2

100 108 100 105 103 100 102 105 99 115 108 109 105 100 94 103 107 110 105 107

Варіант 3

111 108 98 108 110 95 105 115 98 96 100 95 97 98 94 98 100 105 100 102

Варіант 4

90 95 95 102 95 100 102 105 103 102 96 98 112 105 100 95 97 98 94 105

Варіант 5

105 107 96 102 85 100 98 97 90 103 99 115 92 98 110 95 88 93 92 95

Варіант 6

166 164 173 163 163 174 163 169 166 168 169 158 171 167 165 168 169 171 170 154

Варіант 7

89 102 102 101 111 87 109 111 120 99 86 99 100 98 103 106 96 100 102 96

Варіант 8

98 94 98 100 105 100 102 96 90 95 102 95 105 100 102 96 90 95 102 95

Варіант 9

172 178 166 164 171 163 163 172 163 169 164 170 165 171 172 165 166 168 170

Варіант 10

101 99 101 111 87 109 111 120 99 86 99 100 100 102 96 90 92 102

Варіант 11

92 95 100 96 98 118 107 123 107 103 101 99 101 111 115 112 99 112

Варіант 12

69 64 72 71 74 76 70 73 71 69 55 72 71 55 72 73 65 75 73 86

РОЗДІЛ 5. ВИБІРКОВИЙ МЕТОД

5.1. Вибірка і її репрезентативність

Щоб отримати вичерпну інформацію про стан тієї чи іншої статистичної сукупності, потрібно врахувати весь її склад без винятку. Однак в силу різних обставин, не завжди доводиться вдаватися до суцільного обстеження досліджуваних сукупностей. По-перше, тому, що ця робота пов'язана з великими витратами праці, часу, матеріальних ресурсів. По-друге, маючи на увазі практичну неможливість або недоцільність повного врахування всіх членів сукупності. Природні об’єкти, як правило, недоступні суцільному статистичному опису. Внаслідок цього замість суцільного обліку всіх членів досліджуваної сукупності аналізу піддається якась її частина і по ній судять про стан сукупності в цілому.

Сукупність, з якої відбираються варіанти для спільного вивчення, називається генеральною, а відібрана з генеральної сукупності частина її членів носить назву вибірки (N і n відповідно).

Сутність вибіркового методу полягає в тому, щоб за властивостями частини (вибірки) можна судити про численні характеристики цілої (генеральної сукупності). Основу вибіркового методу складає той внутрішній зв'язок, який існує в популяціях між одиничним і загальним, частиною і цілим. Досвід показує, що правильно вироблена вибірка досить добре уявляє або репрезентує структуру і стан генеральної сукупності. Репрезентативність вибірки залежить, перш за все, від способу відбору варіант. У будь-якому випадку вибірка повинна бути типовою і цілком об'єктивною. І цьому сприяє випадковий відбір варіант або принцип рандомізації (за принципом лотереї та існуючої для цих цілей таблиці випадкових чисел).

5.2. Репрезентативність вибіркових показників

Характеристики генеральної сукупності, такі як середня величина (M), дисперсія (σ^2) і середньоквадратичне відхилення (σ) - являють собою величини постійні. По відношенню до них відповідні вибіркові

характеристики, що оцінюють генеральні параметри (x, σ^2 та σ), є величинами випадковими: вони можуть збігатися і не збігатися з величиною генеральних параметрів. Тому виникає питання про репрезентативність вибіркових показників. Можливі відхилення вибіркових показників від генеральних називаються помилками репрезентативності. Ці помилки не технічні, а статистичні, що виникли в результаті недостатньої точності, з якою вибірка репрезентує генеральну сукупність. Розміри вибіркових помилок залежать головним чином від обсягу вибірки і від розміру варіювання ознаки. На розмірах вибіркових помилок позначаються також і способи відбору варіант з генеральної сукупності.

5.2.1. Помилка середньої арифметичної

Групові середні завжди будуть відрізнятися одна від одної, тому що є випадковими величинами. Але в той же час вони будуть варіювати навколо одного і того ж центру розподілу - генеральної середньої (М), яка є величиною постійною. А так як вибіркові середні варіюють в $\sqrt{n}$ разів менше, ніж окремі варіанти однієї і тієї ж генеральної сукупності, то вибіркова помилка середньої або помилка репрезентативності обчислюється за формулами:

$$m_{\bar{x}} = \frac{\sigma}{\sqrt{n}}, \quad \text{або} \quad m_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}} \quad [22].$$

При *n<30* користуються наступними формулами:

$$m_{\bar{x}} = \frac{\sigma}{\sqrt{n-1}}, \quad \text{або} \quad m_{\bar{x}} = \sqrt{\frac{\sum\left(x_i - \bar{x}\right)^2}{n(n-1)}} \quad [23],$$

або

$$m_{\bar{x}} = \sqrt{\frac{1}{n-1}\left(\frac{\sum x^2}{n} - \bar{x}^2\right)} \quad [24],$$

або

$$m_{\bar{x}} = \sqrt{\frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n(n-1)}} \quad [25].$$

Для прикладу візьмемо наступні вісім варіант 2, 4, 3, 7, 5, 6, 4, 5 (табл 5.1).

Таблиця 5.1

(x):	2	4	3	7	5	6	4	5	$\Sigma x = 36$
(x^2):	4	16	9	49	25	36	16	25	$\Sigma x^2 = 180$

$\bar{x} = \frac{36}{8} = 4{,}5$, $\bar{x}^2 = 4{,}5^2 = 20{,}25$ звідки $m_{\bar{x}} = \frac{1}{7}\left(\frac{180}{8} - 20{,}25\right) = 0{,}32$, $m_{\bar{x}} = \sqrt{0{,}32} = 0{,}57$.

Якщо середня арифметична обчислюється у спрощений спосіб, і її помилка визначається тим же способом за формулою:

$$m_{\bar{x}} = \sqrt{\frac{1}{n-1}\left[\frac{\sum a^2}{n} - \left(\frac{\sum a}{n}\right)^2\right]} \quad [26]$$

де $a=x-A$ (відхилення від умовної середньої).

Застосуємо цю формулу до того ж прикладу (табл 5.2):

Таблиця 5.2

(x):	2	4	3	7	5	6	4	5	A=4
(a):	-2	0	-1	+3	+1	+2	0	+1	$a = +4$
(a^2):	4	0	1	9	1	4	0	1	$\Sigma a^2 = 20$

Знаходимо середню: $\bar{x} = A + \frac{\sum a}{n} = 4 + \frac{4}{8} = 4.5$ і її помилку:

$$m_{\bar{x}}^2 = \frac{1}{7}\left[\frac{20}{8} - \left(\frac{4}{8}\right)^2\right] = 0{,}32 \text{ або } m_{\bar{x}} = \sqrt{0{,}32} = 0{,}57.$$

Вибіркова помилка виражається в тих же одиницях вимірюваня, що і супроводжувані нею показники. Вона має два знака: плюс і мінус і характеризує відхилення вибіркових показників як в сторону великих (+), так і в бік менших (-) значень по відношенню до генерального параметру (М).

Середня арифметична і її помилка записується так: $x \pm m_s$, в даному прикладі цей запис виглядає у вигляді $x \pm m_x = 4{,}5 \pm 0{,}57$.

5.2.2. Властивості середньої помилки. Закон великих чисел

Вибіркова помилка характеризує варіювання вибіркових показників навколо їх генеральних параметрів; вона має ті ж властивості, що і середнє відхилення. Лише одна властивість специфічна для вибіркової помилки: вона зменшується при збільшенні числа спостережень (n). Це властивість вибіркової помилки обумовлена дією статистичного закону великих чисел. В цьому законі виражається внутрішній зв'язок між числом випробувань і наближенням вибіркової середньої до свого генерального параметру – математичного сподівання.

Значення вибіркової помилки: вона вказує на точність, з якою визначається середня величина.

Величина середньої помилки залежить не тільки від обсягу вибірки, а й від розмаху варіювання ознаки: чим більше розмах варіації, тим більше буде і величина вибіркової помилки і навпаки.

Поряд із зазначеними причинами на величині середньої помилки позначається і спосіб відбору варіант з генеральної сукупності.

5.2.3. Помилка при різних способах відбору варіант з генеральної сукупності

Залежно від характеру і методики дослідження відбір варіант з генеральної сукупності може проводитися по-різному. Існує два основних способи відбору: повторний і бесповторний випадковий відбір. Повторний відбір проводиться, коли відібрані варіанти повертаються в ту ж сукупність, з якої вже були взяті. Тому вони можуть бути відібрані повторно.

Якщо ж відбираються варіанти назад в генеральну сукупність не повертаються, відбір називається бесповторного або безповоротним. У першому випадку відбір варіант не впливає на склад генеральної сукупності, і ймовірність кожної варіанти потрапити до вибірки не змінюється. У другому випадку кожен попередній відбір, впливає на результат подальшого відбору, тому що змінюється склад генеральної сукупності і ймовірність

варіант потрапити до складу вибірки. Тому, помилка середньої арифметичної обчислюється за формулою:

$$m_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}} = \sqrt{\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)}$$ [27]

Наприклад, із загального часла криниць по Україні (5000), методом випадкового бесповторного відбору взято 500 проб води. Середня мінералізація води виявилася 160 ppm, вибіркова дисперсія σ^2 = 66,3 ppm. Звідси вибіркова помилка середньої арифметичної мінералізації води дорівнює:

$$m_{\bar{x}} = \sqrt{\frac{66,3}{500}\left(1-\frac{160}{5000}\right)} = \sqrt{0,128} = 0,358\,\text{ppm}$$

Якби відбір варіант проводився з цієї сукупності повторним випадковим способом, то помилка вибіркової середньої [22] була б рівною:

$$m_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{66,3}{500}} = \sqrt{0,13} = 0,36\ \text{ppm}$$

Повторний і бесповторний випадковий відбір може проводитися по-різному, в залежності від того, як організовується спостереження над досліджуваним об'єктом, тому можна виділити такі різновиди відбору:

а) типовий або груповий відбір, який називається також районованим, може бути пропорційним або непропорційним;

б) серійний або гніздовий;

в) механічний.

Всі ці види відбору спрямовані на підвищення репрезентативності вибірки, хоча вони і порушують принцип рандомізації (випадковості), тому що проводяться по заздалегідь наміченою схемою.

У разі типового відбору генеральна сукупність розчленовується на окремі і однакові (при пропорційному відборі) за складом групи або райони, з яких випадковим способом проводиться відбір якоїсь кількості варіант. Потім відібрані з кожної групи варіанти об'єднуються в одну вибіркову сукупність і піддаються спільній статистичній обробці. Вибіркова помилка середньої арифметичної у випадках типового пропорційного повторного

відбору визначається за формулою $m_{\bar{x}} = \sqrt{\frac{\bar{\sigma}^2}{n}}$ [28], а при бесповторному – за формулою $m_{\bar{x}} = \sqrt{\frac{\bar{\sigma}^2}{n}\left(1 - \frac{n}{N}\right)}$ [29], де σ^2 - середня зважена з вибіркових дисперсій.

При серійному відборі, генеральна сукупність ділиться на серії, які і відбираються в необхідній (намічуваному дослідником) кількості для спільної обробки. Серії можуть бути рівночисельними і неоднаковими за кількістю складових варіант. Наприклад, при вивченні вмісту азоту у зразках повітря (%), їх розділено на 6 груп. Результати вимірювання представлено в табл 5.3.

Таблиця 5.3

N групи (серії)	1	2	3	4	5	6
середнє груп (x)	71,9	72,8	73,5	76,1	73,2	74,8
дисперсії (σ^2)	16,8	18,3	20,8	17,1	23,9	22,6

Щоб визначити помилку середньої, спочатку треба обчислити середню величину міжсерійної дисперсії:

$$\bar{\sigma}_i^2 = \frac{\sum\left(x_i - \bar{x}\right)}{n_i} = \frac{(71{,}9 - 74{,}0)^2 + (72{,}8 - 74{,}0)^2 + K + (74{,}8 - 74{,}0)^2}{6} = \frac{12{,}98}{6} = 2{,}16\,\%$$

Звідки з [29] $m_{\bar{x}} = \sqrt{\frac{2{,}16}{6}\left(\frac{50-6}{50-1}\right)} = \sqrt{0{,}32} = 0{,}57\,\%$

Вміст азоту у зразках повітря: 74,0±0,57 %.

При механічному відборі генеральна сукупність розбивається на кілька рівних частин або груп. Потім з кожної групи випадковим чином відбирають по одній одиниці. Механічний відбір може проводитися і за іншою схемою, коли в вибірку потрапляє кожна десята, сота і.п. одиниця генеральної сукупності.

5.2.4. Помилки інших вибіркових показників

У дослідницькій роботі може виникнути необхідність об'єднати ряд вибіркових середніх з їх помилками або знайти похідні середніх, супроводжуваних помилками і т.д. У таких випадках поступають таким чином.

1. При розрахунку середньої арифметичної з декількох незалежних середніх з їх помилками, вибіркова помилка обчислюється за формулою:

$$m_{\bar{X}S} = \frac{1}{n_i}\sqrt{m_1^2 + m_2^2 + \ldots + m_n^2} \quad [30]$$

Наприклад, на трьох рівновеликих вибірках отримані середні:

$\bar{x}_1 = 10.2 \pm 0.12$, $\bar{x}_2 = 11.5 \pm 0.18$, $\bar{x}_3 = 13.1 \pm 0.09.$

$$\bar{x}_S = \frac{10,2 + 11,5 + 13,1}{3} = 11,6$$

Знаходимо помилку цього результату:

$$m_{\bar{X}S} = \frac{1}{3}\sqrt{0,12^2 + 0,18^2 + 0,09^2} = \frac{1}{3}\sqrt{0,045} = \frac{0,21}{3} = 0.07 \qquad \bar{x}_S \pm m_{\bar{X}S} = 11,6 \pm 0,07$$

2. Якщо обчислювальна сума декількох середніх арифметичних супроводжуються з їх помилками, то вибіркова помилка суми обчислюється за формулою:

$$m_{\bar{X}S} = \sqrt{m_1^2 + m_2^2 + \ldots + m_n^2} \quad [31]$$

На тому ж прикладі:

$\sum \bar{x} = 10,2 + 11,5 + 13,1 = 34.8$; $m_{\bar{X}} = \sqrt{0,12^2 + 0,18^2 + 0,09^2} = \sqrt{0,045} = 0,21$

3. Помилка похідних двох вибіркових середніх з їх помилками визначається за формулою:

$$m_n = \bar{x}_1 \times \bar{x}_2 \sqrt{\left(\frac{m_1}{\bar{x}_1}\right)^2 + \left(\frac{m_2}{\bar{x}_2}\right)^2} \quad [32]$$

Наприклад: $\bar{x}_1 = 10,3 \pm 0,11$, $\bar{x}_2 = 8,2 \pm 0,12$

$$\mathrm{m_n} = 10,3 \times 8,2\sqrt{\left(\frac{0,11}{10,3}\right)^2 + \left(\frac{0,12}{8,2}\right)^2} = 84,46 \cdot 0,18 = 15,2$$

4. Помилка часткового від ділення середніх арифметичних з їх помилками визначається за такою формулою:

$$m_{ch} = \frac{\bar{x}_1}{\bar{x}_2}\sqrt{\left(\frac{m_1}{\bar{x}_1}\right)^2 + \left(\frac{m_2}{\bar{x}_2}\right)^2} \quad [33]$$

Для розглянутого раніше прикладу:

$$m_{ch} = \frac{10,3}{8,2}\sqrt{\left(\frac{0,11}{10,3}\right)^2 + \left(\frac{0,12}{8,2}\right)^2} = 1,26 \cdot 0,18 = 0,22$$

5. Помилка різниці вибіркових середніх $\left(\bar{x}_1 - \bar{x}_2 = D\right)$ двох незалежних і рівновеликих розподілів (тобто $n_1 = n_2$) рівняється:

$$m_D = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n}} = \sqrt{m_1^2 + m_2^2} \quad [34].$$

6. Помилка різниці вибіркових середніх $\left(\bar{x}_1 - \bar{x}_2 = D\right)$ двох незалежних, але нерівновеликих вибірок (тобто $n_1 \neq n_2$) дорівнює:

$$m_D = \sqrt{\sigma_S^2 \times \frac{n_1 + n_2}{n_1 \cdot n_2}} = \sqrt{\sigma_S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \quad [35]$$

де $\sigma_S^2 = \frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n_1 + n_2 - 2} = \frac{n_1(n_1 - 1)m_1^2 + n_2(n_2 - 1)m_2^2}{n_1 + n_2 - 2} = \frac{\sum a_1^2 + \sum a_2^2}{n_1 + n_2 - 2}$

і $a = \left(x - \bar{x}\right)$. Так, що

$$m_D = \sqrt{\frac{\sum a_1^2 + \sum a_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{\frac{\sum a_1^2 + \sum a_2^2}{n_1 + n_2 - 2} \times \frac{n_1 + n_2}{n_1 \times n_2}} \quad [36].$$

7. Вибіркова помилка різниці середніх $\left(\bar{x}_1 - \bar{x}_2 = D\right)$ сполучених розподілів, тобто які знаходяться в залежності один від одного або від якоїсь загальної причини, обчислюється за формулою:

$$m_{1+2} = \sqrt{m_1^2 + m_2^2 - 2rm_1m_2}\text{, де}$$

r - коефіцієнт кореляції, що показує ступінь спряженості двох рядів розподілів (про нього будемо говорити пізніше).

Але вибіркову помилку різниці середніх арифметичних сполучених розподілів можна обчислити, не використовуючи коефіцієнта кореляції за такими аналогічним формулами:

$$m_d = \sqrt{\frac{1}{n-1}\left(\frac{\sum d^2}{n} - \bar{d}^2\right)} = \sqrt{\frac{\sum\left(d-\bar{d}\right)^2}{n(n-1)}} \quad [37],$$

$$m_d = \sqrt{\frac{\sum d^2 - \left(\frac{\sum d}{n}\right)^2}{n(n-1)}} \quad [38],$$

де *d*- різниця між відповідними варіантами сполучених рядів *X* і *У*, тобто *d* = *x-y, d* - середня різниця, тобто $\frac{\sum d}{n} = \bar{d}$, де *n* - загальне число парних спостережень.

5.2.5. Показник точності оцінки параметрів

Сама по собі абсолютна величина вибіркової помилки як показник іменований мало придатна для випадків порівняльної оцінки точності, з якою визначені середні результати спостережень по відношенню їх до генеральних параметрів.

Наприклад, є середні $\bar{x}_1 = 86{,}1 \pm 0{,}7$, $\bar{x}_2 = 17{,}4 \pm 0{,}2$.

За абсолютною величиною їх помилок важко сказати, яка середня визначена більш точно, оскільки середні з їх помилками виражені різними одиницями виміру. Щоб отримати певне уявлення про точність, з якою визначено той чи інший середній результат, прийнято використовувати так званий показник точності C_s:

$$Cs = \frac{m_{\bar{x}}}{\bar{x}} \cdot 100\% \quad [39]$$

Коли відомо значення коефіцієнта варіації (*CV*), показник точності можна визначити за такою формулою:

$$Cs = \frac{CV}{\sqrt{n}} \quad [40].$$

Під точністю визначення вибіркової середньої розуміється ступінь наближення її до середньої генеральної сукупності. Чим точніше визначено середній результат, тим менше буде *Cs*, і, навпаки, при менш точному середньому результаті показник *Cs* виявиться більше. Точність вважається

достатньою, якщо *Cs* не перевищує 3-5%. Так, для наведених раніше середніх, показники точності в обох випадках будуть наступними:

$$Cs_1 = \frac{0,7}{86,1} \cdot 100\% = 0,81\% \cdot \qquad Cs_2 = \frac{0,2}{17,4} \cdot 100\% = 1,15\%$$

Звідси також видно, що перша середня визначена більш точно, ніж друга.

РОЗДІЛ 6. СТАТИСТИЧНІ ГІПОТЕЗИ ТА ЇХ ПЕРЕВІРКА

Інформація, яку дістають на підставі вибірки, реалізованої із генеральної сукупності, може бути використана для формулювання певних суджень про всю генеральну сукупність. Наприклад, розпочавши виготовлення препарату нового типу від алергії, відбирають певну кількість зразків і піддають певним тестам. За результатами тестів можна зробити висновок про те, чи кращий новий препарат від препарату старого типу, чи ні. А це, у свою чергу, дає підставу для прийняття рішення: виготовляти їх чи ні. Такі рішення називають ***статистичними.***

Статистичні рішення мають ймовірнісний характер, тобто завжди існує ймовірність того, що прийняті рішення будуть помилковими. Головна перевага прийняття статистичних рішень полягає в тому, що в межах ймовірнісних категорій можна об'єктивно виміряти ступінь ризику, що відповідає тому чи іншому рішенню. Будь-які статистичні висновки, здобуті на підставі обробки вибірки, називають ***статистичними гіпотезами.***

Дані вибіркових спостережень часто становлять основу для прийняття одного з кількох альтернативних рішень (продукція може бути бракованою або якісною, технологічний процес порушується або ні, точність обробки виробу в межах норми, нижча від норми або вища від неї і т.д.). Із загальнометодологічного погляду йдеться про висунення деякої гіпотези, яку відхиляють або приймають після проведення деякого експерименту. Якщо експеримент має статистичний характер, кажуть, що гіпотеза є статистичною.

Статистичною називають гіпотезу про властивості (ознаки) генеральної сукупності, що перевіряється на основі вибірки. У статистиці виділяють два основні типи гіпотез:

1. гіпотези про закон розподілу ймовірностей випадкової величини;
2. гіпотези про значення параметрів розподілу випадкової величини.

Статистичні гіпотези першого типу називають непараметричними, а другого типу — параметричними.

6.1 Нульова гіпотеза. Рівні значущості і довірчі ймовірності

Оцінювання певної ознаки генеральної сукупності здійснюється на основі цієї ж ознаки в вибірковій сукупності із врахуванням помилки репрезентативності. А по відношенню властивостей генеральної сукупності висувається деяка гіпотеза про величину середньої, дисперсії, характер розподілу, форму і тісноту зв'язку між досліджуваними змінними. Перевірку гіпотези проводять на основі виявлення узгодження фактичних і теоретичних даних.

Гіпотезу, що підлягає перевірці, називають *основною*. Оскільки ця гіпотеза припускає відсутність систематичних розбіжностей (нульові розбіжності) між невідомим параметром генеральної сукупності і величиною, що одержана внаслідок обробки вибірки, то її називають *нульовою гіпотезою* і позначають H_0. Висуваючи нульову гіпотезу, експериментатор виходить з припущення, що спостережувана мінливість ознаки залежить не від дії організованого фактора, а визначається другорядними, нерегульованими в досвіді випадковими причинами.

Сформульована гіпотеза потребує перевірки. Щоб її прийняти або відкинути, потрібні підстави. Підстави дає теорія ймовірностей, що дозволяє пов'язувати статистичні гіпотези з певною ймовірністю. Дотримуючись закону нормального розподілу, можна стверджувати, що в 95% випадків вибіркова середня ($\bar{x}$) не відхиляється від середньої (M) генеральної сукупності більше, ніж на *2t*, де $t = \frac{\acute{x} - M}{\sigma}$. І тільки в 5% випадків, вважаючи відхилення в + і - напрямках від (M), вибіркова середня вийде за ці межі. Це означає, що ймовірність отримати у вибірці середній результат, який відхилиться від генерального параметра на *2t,* дорівнює лише 0,05. Якщо ж мова йде про відхилення від (M) тільки в одну сторону, ймовірність буде

вдвічі менше (P = 0,025). Так ось, відсоток таких малоймовірних випадків, які суперечать прийнятій гіпотезі, ставить її під сумнів, називається рівнем значущості гіпотези. У хімічних дослідженнях зазвичай приймається 5%-й рівень значимості, якому відповідала би ймовірність $P_1 = 0,05$. У випадку особливо точних досліджень приймається 1% -й або 0,1%-й рівні значущості, яким відповідає $P_2 = 0,01$ і $P_3 = 0,001$.

Таким чином, ймовірність, якою вирішено знехтувати при оцінці генеральних параметрів поданням вибіркових спостережень, виражається прийнятим рівнем значущості.

Імовірність же зворотних випадків, коли гіпотеза заслуговує довіри, називається довірчою ймовірністю. Зазвичай в дослідницькій практиці приймаються три пороги довірчої ймовірності: $P_1 = 0,95$; $P_2 = 0,99$; $P_3 = 0,999$.

Кожен поріг, або рівень довірчої ймовірності, зв'язується з певною величиною нормованого відхилення (табл 6.1):

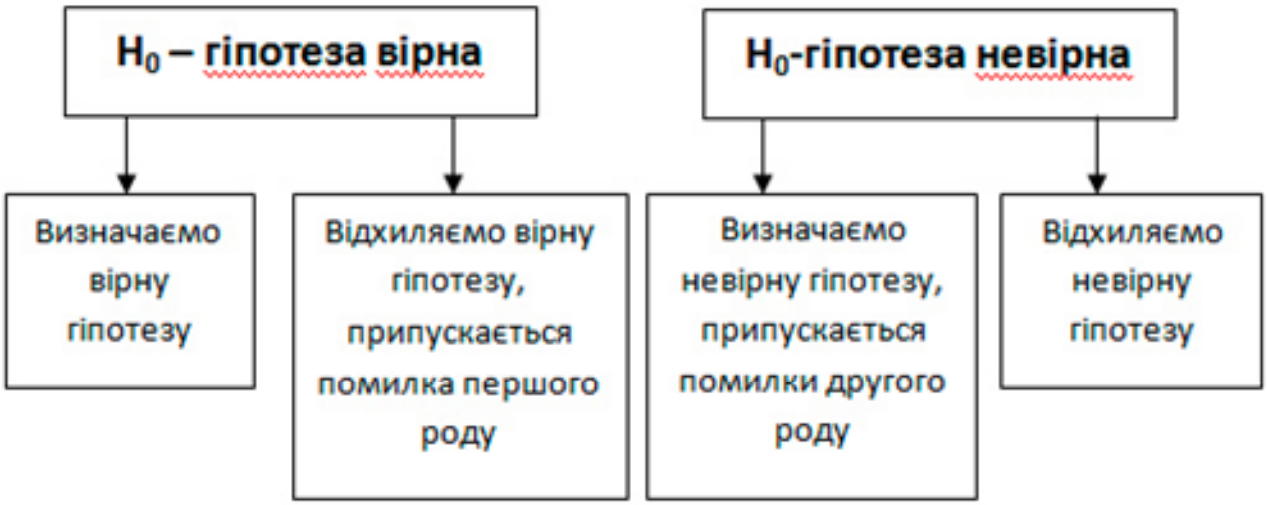
Таблиця 6.1

Довірча ймовірність *(P)*	Нормоване відхилення (t)
0,95	1,96
0,99	2,58
0,999	3,29

Величина довірчої ймовірності або рівень значущості при перевірці гіпотез встановлюється самим дослідником в залежності від ступеня точності, з якою проводиться дослідження і відповідальності висновків, що випливають з нього. Якщо $P \geq 0,05$ або ж $P < 0,95$, то відкидати нульову гіпотезу немає підстав. Коли ж $P < 0,05$ або $P > 0,95$ нульова гіпотеза відкидається.

Висунута гіпотеза може бути правильною або неправильною, у зв'язку із чим виникає необхідність перевірити її та довести, яка із гіпотез є вірною, але одночасно при перевірці гіпотез можуть бути допущені помилки першого (type 1 error) і другого роду (type 2 error). Наприклад, можна відкинути

нульову гіпотезу, коли вона насправді є вірною (так звана помилка 1-го роду) або можна прийняти нульову гіпотезу, коли вона насправді є невірною (так звана помилка 2-го роду).



6.2. Довірчий інтервал і його межі

Межі, в яких з тією чи іншою ймовірністю знаходиться параметр генеральної сукупності, називаються довірчими, а інтервал, укладений між цими межами, носить назву довірчого. У загальній формі можна наступним чином встановити довірчий інтервал для невідомого параметра М генеральної сукупності:

$$-t \le \frac{x - M}{\sigma} \le +t \quad [41]$$

Так як ймовірність відхилення кожен вид від центру розподілу визначається функцією нормованого відхилення. Перетворивши цей вираз, отримуємо:

$$x - t\sigma \le M \le x + t\sigma \quad [42]$$

Це і є довірчий інтервал, в якому знаходиться величина генерального параметра М. Тут $x - t\sigma$ і $x + t\sigma$- довірчі межі, t - нормоване відхилення, яке визначається порогом довірчої ймовірності.

Так з ймовірністю P = 0,95 (відповідає t = 1,96) можна стверджувати, що невідомий генеральний параметр М нормального розподіляється сукупності знаходиться в інтервалі:

$$x - 1{,}96\sigma \le M \le x + 1{,}96\sigma$$

Визначення можливих значень генеральних параметрів за величиною вибіркових показників носить загальну назву оцінки генеральних параметрів. Критерієм оцінки служить стандартна величина нормованих відносини (t_{st}) з якою порівнюється фактичне значення цього критерію ($t_{ф}$). Відносно генеральної середньої М цей критерій виражається наступними аналогічними відношеннями:

$$t_{ф} = \frac{\bar{x} - M}{m_{\bar{x}}}, \text{ або } t_{ф} = \frac{\bar{x} - M}{\sigma}\sqrt{n} \quad [43]$$

При $t_{ф} < t_{st}$ нульова гіпотеза зберігається. Якщо ж $t_{ф} \ge t_{st}$ нульову гіпотезу слід відкинути. Наприклад, при аналізі 95 зразків гірських порід Чернігівської області середній вміст Мангану становив 6,2 % при σ = 0,43 %. Чи можна на підставі цього результату зробити висновок, що вміст Магнану у гірський породах Чернігівської області вище ніж середній по Україні (6,4%)? Нормуючи відомі величини, знаходимо:

$$t_{ф} = \frac{6.2 - 6.4}{0.43}\sqrt{95} = -4.5$$

Для довірчої ймовірності P = 0,99 t_s = 2,58. Так як $t_{ф} \ge t_{st}$ нульова гіпотеза відкидається.

Можна також оцінити достовірність (тобто не випадковість) відмінностей, що спостерігаються між середніми $\bar{x}_1$ та $\bar{x}_2$

При порівнянні статистичних показників один з одним слід враховувати, на яких сумах - залежних або незалежних - вони отримані. Якщо варіанти однієї ознаки *X* розподіляються незалежно від значень іншої ознаки *Y*, вони називаються незалежними. Якщо ж значення однієї ознаки в тій чи іншій мірі пов'язані з відповідними значеннями іншої ознаки, вони залежні один від одного.

6.3. t- розподіл Стьюдента

Один з найпоширеніших методів оцінки достовірності результатів є t-розподіл Стьюдента. Цей метод розроблено Вільямом Госсетом (1876-1937) для оцінки якості пива на пивоварних заводах Guinness в Дубліні (Ірландія). Стаття Госсета «The probable error of a mean» («Вірогідна помилка середнього») вийшла у 1908 році у журналі «Biometrika» під псевдонімом «Student» (Студент).

Існує кілька теорій, які намагаються це пояснити, чому Госсет використав псевдонім. Перша і, мабуть, найбільш поширена, каже, що основна причина в тому, що Guinness раніше мав збитків від витоку інформації через публікації співробітників, і компанія заборонила своїм співробітникам публікувати статті, незалежно від їхньої теми. Також існує думка, що публікація Госсета під псевдонімом Стьюдент дозволила приховати від компанії Guinness, що її співробітник опублікував статтю. Інша версія говорить про те, що Госсет домовився з пивоварнею про її публікацію (проте зміст статті не був би корисним для конкуренції), але компанія попросила використати псевдонім, щоб інші співробітники не знали про цю публікацію. Також псевдонім Стьюдент міг бути використаний для Guinness, тому що компанія хотіла зберегти в таємниці ім`я дослідника, який працює на неї. Це було зроблено, щоб не було жодних доказів промислової переваги, яка була досягнута завдяки йому.

У багатьох хімічних дослідженнях обсяг вибіркової сукупності не перевищує 20-30 спостережень. Такі вибірки називають малими.

Коли вибірки незалежні, різниця між генеральними параметрами оцінюється по різниці вибіркових середніх ($\acute{x}_1 - \acute{x}_2 = D$). Число ступенів свободи в таких випадках визначається за формулою:

$$k = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2 \quad [44]$$

Якщо ж порівнювані вибірки залежні одна від одної, то різницю між параметрами слід обчислювати не по різниці вибіркових середніх, а за

середньою різниці між парними варіантами *(x - y = d)* сполучених розподілів. У цьому випадку число ступенів свободи визначаються за формулою:

$$k = n - 1 \text{ (або } k = n - 2)$$

Нульова гіпотеза відкидається при $t_{\phi} = \frac{\overline{x_1} - \overline{x_2}}{m_D} \geq t_{st}$ для відповідних *P* і *k*.

Переходимо до розгляду відповідних прикладів, на яких легше продемонструвати значення критерію *t* в оцінці генеральних параметрів за даними вибіркових спостережень.

6.3.1 Випадок незалежних вибірок

Вибірки називаються незалежними (незв'язними), якщо процедура експерименту і отримані результати вимірювання деякої властивості у випробовуваних зразках не впливають на особливості протікання цього ж експерименту і результати вимірювання цієї ж властивості у випробовуваних (респондентів) іншої вибірки.

Розглянемо наступний прикладі. Вивчався вплив йонів Кобальту на збільшення живої ваги кроликів. Дослід проводився на двох групах тварин - дослідної та контрольної. Вихідна вага особин не виходила за межі 500-600 г. Дослід проводився півтора місяці. Обидві групи тварин були на одному і тому ж кормовому раціоні. Експериментальна группа особин отримувала препарат у вигляді водного розчину по 0,06 г кобальт хлориду на 1 кг живої ваги. За час експерименту тварини дали наступні збільшення у вазі (табл 6.2):

Таблиця 6.2

Контроль:	504	560	580	600	420	530	490	580	470	n=9
Експеримент	580	692	700	621	640	561	680	630	-	n=8

Дані величини варіюють незалежно: кожна величина приймає те чи інше значення незалежно від значення іншої величини. Групуємо результати експерименту в таблицю 6.3.

Таблиця 6.3

Збільшення маси (г)		Відхилення від $\bar{x}$		Квадрати відхилень	
експеримент	контроль	експеримент	контроль	експеримент	контроль
580	504	58	22	3364	484
692	560	54	34	2916	1156
700	420	62	106	3844	11236
621	600	17	74	289	5476
640	580	2	54	4	2916
561	530	77	4	5929	16
680	490	42	36	1764	1269
630	580	8	54	64	2916
-	470	-	54	-	3136
$\sum = 5104$	$\sum = 7434$	-	-	$\sum = 18174$	$\sum = 28632$
$\bar{x}_1 = 638$	$\bar{x}_2 = 526$	-	-	$\sum = 46806$	

Різниця середніх значень в контрольній і експериментальній групі становить: 638-526 = 112 г. Визначаємо помилку цієї різниці.

$$m_D = \sqrt{\frac{\Sigma a_1^2 + \Sigma a_2^2}{n_1 + n_2 - 2} \times \frac{n_1 + n_2}{n_1 \cdot n_2}} = \sqrt{\frac{46806}{8+7} \times \frac{9+8}{9 \cdot 8}} = \sqrt{736.8} = 27.13г.$$

Критерій достовірності:

$$t_\phi = \frac{\overline{x_1} - \overline{x_2}}{m_D} = \frac{112}{27.13} = 4.1$$

Для рівня значущості $P = 0{,}01$ і для числа ступенів свободи $k = 9 + 8 - 2 = 15$ знаходимо по таблиці стандартних значень t-критерію Стьюдента (додаток, табл. 1) значення t_{st}, рівне 2,95, $t_ф > t_{st}$, (фактичне значення значно перевершує критичне значення), отже, нульову гіпотезу потрібно відкинути і визнати статистично достовірної різницю в прирості маси кроликів в експериметнтальній групі.

Розглянемо ще один аналогічний приклад. На двох групах лабораторних мишей – експериментальноі та контрольної – вивчався вплив

хіміко-терапевтичного препарату на розвиток організму тварин. В результаті випробувань отримано наступні відмінності у масі тварин в грамах (табл.6.4):

Таблиця 6.4

контрольні	70	78	60	80	60	60	68	$\bar{x}_1 = 68$ г
експериментальні	80	75	62	70	68	71	-	$\bar{x}_2 = 71$ г

D=71 - 68 =3,0 г

Визначимо помилку цієї різниці. Спочатку розрахуємо суми квадратів відхилень варіант від їх середніх за формулою:

$$\Sigma a^2 = \Sigma\left(x_i - \bar{x}\right)^2 = \Sigma x^2 - \frac{(\Sigma x)^2}{n} \quad [45]$$

Контроль: $\Sigma a_1^2 = \left(70^2 + 78^2 + 80^2 + ... + 68^2\right) - \frac{476^2}{7} = 32808 - 32368 = 440$

Експеримент: $\Sigma a_2^2 = \left(80^2 + 75^2 + 62^2 + ... + 71^2\right) - \frac{426^2}{6} = 30434 - 30246 = 188$

Знаходимо об'єднаний середній квадрат відхилень: $\sigma_s^2 = \frac{440 + 188}{7 + 6 - 2} = \frac{628}{11} = 57.1$,

звідки помилка різниці середніх обчислюється так:

$$m_D = \sqrt{\sigma_s^2 \cdot \frac{n_1 + n_2}{n_1 \cdot n_2}} \quad [46]$$

$$m_D = \sqrt{57.1 \cdot \frac{13}{42}} = \sqrt{17.7} = 4.2г.$$

Критерій достовірності відмінностей $t_ф$ = 3,0/4,2 = 0,71. По таблиці стандартних значень критерію *t*-Стьюдента (додаток, табл. 1) для рівня значущості $P = 0{,}05$ і $k = 11$ знаходимо $t_{st} = 2{,}2$, $t_ф < t_{st}$ отже, нульова гіпотеза зберігається, тобто різницю між генеральними середніми цих груп статистично недостовірна і, отже, дія хіміко-терапевтичного препарату статистично недостовірна.

Коли відома генеральна середня (М), то різниця між нею і вибіркової середньої ($\bar{x}$) оцінюється помилково вибіркової середньої, тому що генеральна середня помилки не має:

$$t = \frac{\bar{x} - M}{m_{\bar{x}}} \quad t = \frac{\bar{x} - M}{\sigma}\sqrt{n} \quad [47]$$

Наприклад, в результаті дослідження 12 зразків встановлено збільшення антиоксидантної активнівсті ферментної системи при введені нового препарату на 4,16 ± 0,025%. Вихідна антиоксидантна активність ферментної системи 4,09%. Чи достовірна різниця?

$$t = \frac{4{,}16 - 4{,}09}{0{,}025} = \frac{0{,}07}{0{,}025} = 3{,}2$$

Для рівня значущості $P = 0{,}01$ и $k = n\text{-}1 = 12\text{-}1 = 11$ $t_{st} = 3{,}11$. $t_ф < t_{st}$ нульова гіпотеза відкидається.

6.3.2 Випадок залежних вибірок

Вибірки називається залежними якщо процедура експерименту і отримані результати вимірювання деякої властивості, проведені на одній вибірці, впливають на іншу.

Найбільш типовий приклад залежних вибірок - повторне вимірювання властивості (властивостей) на одній і тій же вибірці після впливу (ситуація «до-після»). В цьому випадку вибірки (одна - до експериментального впливу, інша - після експериментального впливу) залежні в максимально можливій мірі, так як вони включають одних і тих же випробовуваних. За таких умов, описаний раніше спосіб оцінки генеральних параметрів виявляється неточним. Розглянемо наступний приклад.

Вивчався вплив двох інгібіторів на корозію сталі. Проведено 6 паралельних дослідів. Визначалася маса пластинки після витримки у інгібовних розчинах кислоти у грамах (табл. 6.5), враховуючи, що початкова маса всіх пластинок була однаковою.

Таблиця 6.5

Речовина	1	2	3	4	5	5	Середне значення
Інгібітор 1	31,3	24,0	24,6	28,6	29,1	30,1	$\bar{x}_1$=27,9
Інгібітор 2	31,6	24,2	24,8	29,1	31,0	31,0	$\bar{x}_2$=28,4

Видно, що маси пластинок, витриманих у розчинах кислоти з інгібітором I, трохи вище, ніж у розчинах з інгібітором II, $D = \acute{x}_1 - \acute{x}_2 = 0{,}5$ *г.* Чи можна покластися на цю різницю, чи достовірна вона?

Якщо використовувати вже розглянутий нами підхід до незалежних вибірок, то $t_ф = 1{,}6$. Для $k = 10$ и $P = 0{,}05$ $t_{st} = 2{,}23$. $t_ф < t_{st}$, отже, нульову гіпотезу відкинути не можна. Різниця – не достовірна.

Якщо ж порівнювати не середні, а варіанти, тобто оцінювати генеральні параметри по середній різниці варіант з урахуванням пов'язаності між ними, вийде наступний результат (табл 6.6).

Таблиця 6.6

Показник	Маса пластинок після досліду в грамах по повторюваностях						Середнє
	1	2	3	4	5	6	
Інгібітор I	31,1	24,0	24,6	28,6	29,1	30,1	27,9
Інгібітор II	31,6	24,2	24,8	29,8	29,9	31,0	28,4
Різниця (d)	0,5	0,2	0,2	0,5	0,8	0,9	-
Квадрат різниці (d^2)	0,25	0,04	0,04	0,25	0,64	0,81	$\sum d^2 = 2.03$

Помилка середньої різниці, визначається за формулою:

$$\overline{d} = \overline{x_1} - \overline{x_2} = 28.4 - 27.9 = 0.5$$

$$m_d = \sqrt{\frac{1}{n-1}\left(\frac{\Sigma d^2}{n} - \overline{d^2}\right)} = \sqrt{\frac{1}{5}\left(\frac{2{,}03}{6} - 0{,}5^2\right)} = \sqrt{0{,}018} = 0{,}13г.$$

Для P = 0,05, k = 6-1= 5, $t_{st} = 2{,}57$, $t_ф > t_{st}$, отже, нульова гіпотеза відхиляється, і різниця визнається статистично достовірною.

Наведений приклад служить наочним підтвердженням того, що статистичні методи можна застосовувати огульно, не погодивши з змістом експериментального матеріалу.

Як критерій достовірності середньої різниці може служити також і відношення:

$$t = \frac{\overline{d}}{\sigma_d}\sqrt{n}$$ [48], де $\overline{d} = \frac{1}{n}\Sigma d_i$, а

$\sigma_d = \sqrt{\frac{1}{n-1}\left(\Sigma\left(d_1 - \overline{d}\right)^2\right)}$, яке оцінюється по таблиці Стьюдента.

6.4. Порівняння дисперсії. F-розподіл Фішера

Критерій оцінки значущості відмінності двох дисперсій, розрахованих по двох вибірках з генеральних сукупностей, які характеризуються розподілом, близького до нормального запропоновано Рональдом Ейлмером Фішером (F-критерій або критерій Фішера). В своїй книзі «Планування експериментів» (1935) ним також запропонована методологія планування експерименту. Для прикладу він описав, як перевірити гіпотезу про те, що певна жінка може лише на смак визначити, молоко чи чай було спочатку налито в чашку (експеримент «Леді дегустує чай»). Таке завдання допомогло проілюструвати найважливіші ідеї планування експерименту:

- Порівняння.
- Рандомізація
- Реплікація.
- Блокування.
- Ортогональність.
- Факторіальні експерименти.

Аналіз планування експерименту був побудований на основі аналізу різниці, колекції моделей, в яких спостерігається різниця розділена на компоненти завдяки різним факторам, які оцінюються та/або тестуються.

Критерій Фішера розраховується за формулою:

$F = \frac{\sigma_1^2}{\sigma_2^2}$, причому $\sigma_1^2 \rangle \sigma_2^2$, тому що $F \geq 1$

Критерій Фішера функціонально пов'язаний з ймовірністю. Він залежить від числа ступенів свободи $k_1=n_1-1$ и $k_2=n_2-1$ дисперсій, що порівнюються.

Характерним для F-критерію є те, що він повністю визначається вибірковими дисперсіями і не залежить від генеральних параметрів, тому що передбачається, що обидві дисперсії з однієї і тієї ж генеральної сукупності. Для визначення довірчих меж для критерію Фішера існують таблиці стандартних значень. У цій таблиці ступені свободи для більшої дисперсії беруться по горизонталі, для меншої дисперсії - по вертикалі. На перетині цих стовпців знаходиться стандартне значення критерію Фішера для відповідного значення Р (Р = 0,05 або Р = 0,01).

Нульова гіпотеза виходить з визнання рівності дисперсій. Якщо емпіричні значення ($F_ф$) менше теоретичних (F_{st}) для відповідного рівня значущості і ступенів свободи k_1 і k_2, різниця розглядається як випадкова. Якщо ж $F_ф \geq F_{st}$ нульова гіпотеза відкидається, різниця між порівнюваними величинами визнається статистично достовірною. Наприклад, порівнюються маси продукту реакції, одержанного двома різними способами:

перший спосіб: $n_1 = 100$, σ_1 - 58,3 мг;

другий спосіб: $n_2 = 200$, $\sigma_2 = 59,3$ мг.

Чи достовірна різниця?

Обчислюємо значення критерію Фішера:

$$F_ф = \frac{(59,3)^2}{(58,3)^2} = \frac{3516}{3398} = 1,03$$

По таблиці Фішера для Р = 0,05 та k = 200 - 1 = 199 и k_2 = 100 - 1 = 99, знаходимо F_{st}= 1,31. $F_ф < F_{st}$, нульова гіпотеза зберігається, розбіжності між вибірками за цією ознакою виявляються статистично недостовірними.

ПРАКТИЧНА РОБОТА 3

Здійснити довірливу оцінку різниці даних одержаних у контролі та експерименті, використовуючи t критерій Стьюдента для рівня значущості P=0,95

Варіант 1

контроль	5,8	4,9	6,4	7,9	5,4	5,4	4,6
експеримент	3,8	3	3,3	3,4	3,5	3,9	4,1

Варіант 2

контроль	23	24,2	21,8	24,9	21,6	26,1	22,9
експеримент	18	19,7	20	18,6	18,9	19,4	19,3

Варіант 3

контроль	6,9	7,8	7,9	6,2	6,4	7,0	7,2
експеримент	8,2	8,3	8,5	9,7	8,6	8,9	9,0

Варіант 4

контроль	8,5	8,5	8,6	9,8	9	8,1	8,7
експеримент	5,9	5,9	6,7	4,9	5,7	6,3	

Варіант 5

контроль	3,2	4	3,5	3,3	3,9	4,1	3,5
експеримент	5,7	6,2	6,0	5,8	6,1	6,0	

Варіант 6

контроль	2,5	1,2	2,1	2,0	1,8	1,8	
експеримент	4,1	3,8	3,6	3,5	3,4	3,9	4,0

Варіант 7

контроль	15,6	16,0	13,9	15,1	13,8	14,6	14
експеримент	11,6	12,3	12,0	11,9	11,2	12,8	

Варіант 8

контроль	15,1	13,8	12,8	11,2	14	14,9	12,8
експеримент	3,2	3,4	4	3,8	4	2,8	3,6

Варіант 9

контроль	25,9	24,5	29,1	21,9	24,6	22,8	29,4
експеримент	20,0	19,6	21,4	20,8	20,4	20,7	

Варіант 10

контроль	9	8,5	6,7	8,1	8,5	8,6	9,8
експеримент	5,9	2	5,9	3,5	3,3	4,9	

Варіант 11

контроль	15,1	13,8	14,8	15,2	14	14,9	
експеримент	4,1	3,5	4,9	3,9	3,3	4,1	4,8

Варіант 12

контроль	9	8,7	8,5	9,1	8,9	8,8	9
експеримент	4,9	3,9	4,4	4	4,1	4,9	

ПРАКТИЧНА РОБОТА 4

Дві незалежні групи дослідників при визначенні вмісту селену (%) в шести пробах отримали такі дані. Перевірте, чи є статистично значуща різниця (за F-критерієм Фішера) між середніми значеннями вмісту селену для результатів, отриманих двома незалежними групами.

Варіант 1

I	9,31	9,42	9,35	9,37	9,4	9,38	9,41	9,36
II	9,51	9,43	9,61	9,63	9,44	9,55	9,49	9,52

Варіант 2

I	10,53	10,54	10,35	10,44	10,4	10,58	10,61	10,65
II	10,67	10,75	10,87	10,79	10,8	10,84	10,82	10,81

Варіант 3

I	11,33	11,44	11,37	11,4	11,43	11,41	11,38	11,5
II	11,67	11,75	11,87	11,79	11,8	11,84	11,82	11,81

Варіант 4

I	8,21	8,38	8,36	8,42	8, 33	8,31	8,35	8,41
II	8,55	8,65	8,86	8,69	8,66	8,74	8,62	8,51

Варіант 5

I	12,71	12,58	12,66	12,72	12,53	12,41	12,35	12,56
II	12,78	12,95	12,86	12,85	12,96	12,72	12,92	12,99

Варіант 6

I	12,33	12,44	12,37	12,40	12,43	12,41	12,38	12,5
II	12,55	12,65	12,86	12,69	12,66	12,74	12,62	12,51

Варіант 7

I	6,31	6,42	6,35	6,37	6,4	6,38	6,41	6,36
II	6,55	6,65	6,86	6,69	6,66	6,74	6,62	6,51

Варіант 8

I	7,34	7,36	7,38	7,38	7,32	7,37	7,39	7,37
II	7,45	7,40	7,41	7,43	7,43	7,45	7,42	7,40

Варіант 9

I	6,67	6,75	6,87	6,79	6,8	6,84	6,82	6,81
II	5,78	5,95	5,86	5,85	5,96	5,72	5,92	5,99

Варіант 10

I	7,15	7,18	7,09	7,05	7,12	7,16	7,17	7,1
II	7,21	7,30	7,23	7,26	7,21	7,28	7,25	7,33

Варіант 11

I	10,23	10,34	10,37	10,30	10,33	10,31	10,38	10,41
II	10,77	10,85	10,87	10,89	10,90	10,94	10,92	10,91

РОЗДІЛ 7. ОЦІНКА ЗАКОНІВ РОЗПОДІЛУ

7.1. Оцінка вискакуючих варіант

Бувають випадки, коли окремі крайні варіанти сильно відхиляються від сусідніх з ними варіант варіаційного ряду. Тоді виникає сумнів у тому, чи належать вони до даної генеральної сукупності. Причини таких явищ можуть бути різними: по-перше, можливі технічні помилки, допущені при утворенні вибіркової сукупності, по-друге, "вискакування" варіант може бути наслідком сильної варіабельності ознаки, тобто явищем цілком нормальним.

Критерієм оцінки вискакуючих варіант служить нормоване відхилення:

$$t = \frac{x - \bar{x}}{\sigma\sqrt{\frac{n+1}{n}}} \quad [49]$$

Нульова гіпотеза в даному випадку говорить про те, що "вискакуюча" варіантаналежить до тієї ж генеральної сукупності і відкидається, коли $t_{ф} < t_{st}$.

Наприклад, при титруванні п'яти проб, на титрування було використано наступні об’єми кислоти:

8,3 7,9 9,1 6,8 і 12,1 мл.

Викликає сумнів варіанту 12,1, сильно відрізняється від середнього значення x = 8,84, σ = 2,0 мл. Оцінимо її "вискакування" на достовірність. Знайдемо нормоване відхилення $t = \frac{12,1 - 8,84}{2,0} = \frac{3,26}{2,0} = 1,63$. По таблиці "Критичні значення величини нормованого відхилення при оцінці сумнівних варіант", знаходимо значення t_{st} для P = 0,95 и n = 5, t_{st}= 1,92. Так як $t_{st} > t_{ф}$ нульова гіпотеза зберігається, відкидати цю варіанту при розрахунку середнього об’єму кислоти не можна.

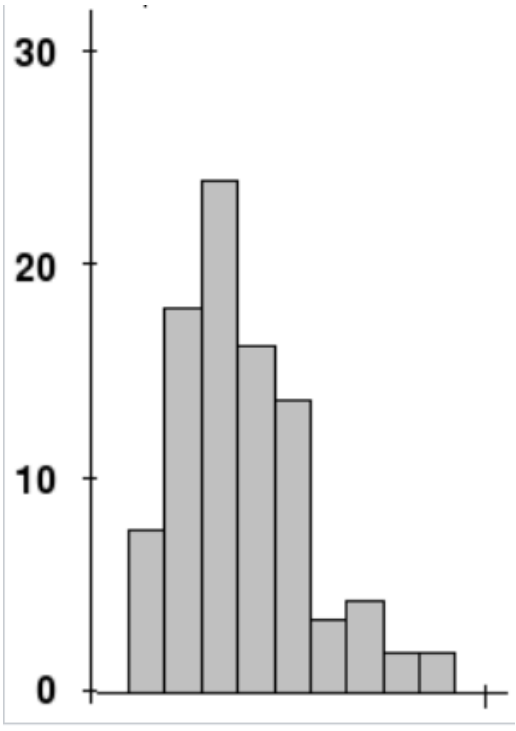
7.2. Наближені оцінки закону розподілу. Обчислення асиметрії і ексцесу

Середня арифметична, σ^2 та σ самі по собі не містять інформації про закон розподілу. Не всі ознаки розподіляються по нормальному закону: деякі

виявляють явну асиметрію, можливі й інші випадки відхилення від нормального закону розподілу. Тому, перш ніж використовувати той чи інший критерій оцінки генеральних параметрів, слід скласти уявлення про закон розподілу досліджуваної ознаки. Наближена оцінка закону розподілу може бути отримана за допомогою коефіцієнтів асиметрії та ексцесу. Варіаційні ряди можуть бути: скошеними (позитивна і негативна асиметрія), гостро-і плосковерхівними (позитивний і негативний ексцес).

Асиметрією γ (коефіцієнт асиметрії Фішера) теоретичного розподілу ймовірностей випадкової величини називають відношення центрального моменту третього порядку μ_3 до куба середнього квадратичного відхилення σ^3

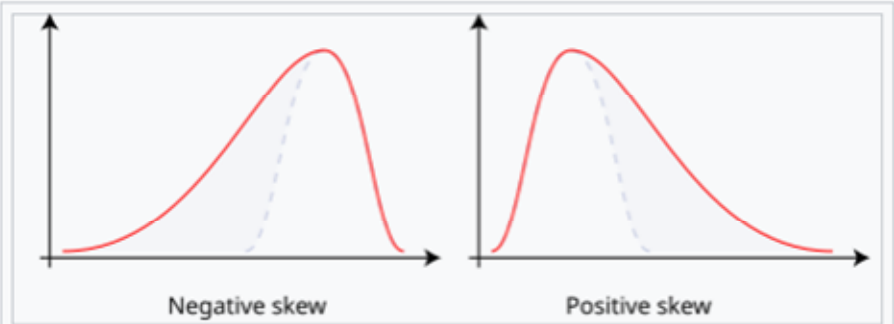
$$\gamma_1 = \mu_3 / \sigma^3$$



Приклад експериментальних даних з ненульовою асиметрією

Асиметрія додатна, якщо «довша частина» розподілу знаходиться праворуч від математичного сподівання; асиметрія від'ємна, якщо «довша частина» кривої знаходиться ліворуч від математичного сподівання. На практиці, знак асиметрії визначають за положенням кривої відносно моди:

якщо «довша» частина кривої знаходиться правіше моди, то асиметрія додатня, якщо лівіше — від'ємна.



Крива ліворуч має від'ємну асиметрію, а крива праворуч — додатню.

Мірою скошеності рядів розподілу служить коефіцієнт асиметрії, позначається символом As:

$$As = \frac{\Sigma p(x_i - \bar{x})^3}{n\sigma^3} = \frac{\Sigma pa^3}{n\sigma^3}$$ [50]

При строго симетричному розподілі As = 0, т.як $\Sigma(x_i - \bar{x})^3 = 0$. Коефіцієнт асиметрії величина відносна; він коливається від 0 до 1. Якщо As ≤ 0,2 асиметрія вважається незначною, при As > 0,5 скошенність розподілу виявляється вже сильною (рис 7.1).

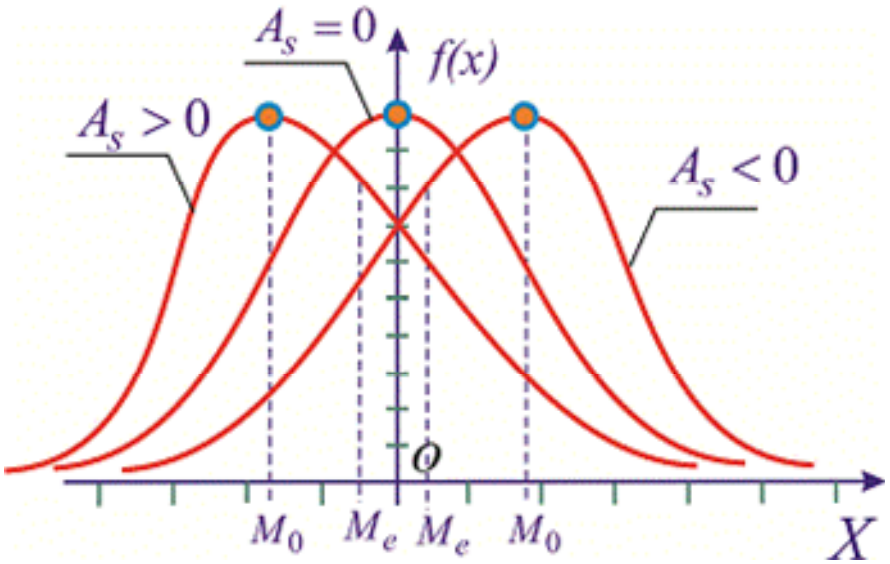


Рис 7.1. Види асиметрії розподілу

Робоча формула для обчислення коефіцієнта асиметрії за способом умовної середньої наступна:

$$As = \frac{\frac{\Sigma pa^3}{n} - 3b\frac{\Sigma pa^2}{n} + 2b^3}{\sigma^3} \quad [51]$$

де $b = \frac{\Sigma p(x_i - A)}{n}$ - умовний момент 1 порядку,

$\frac{\Sigma pa^2}{n} = \frac{\Sigma p(x_i - A)^2}{n}$ - умовний момент 2 порядку,

$\frac{\Sigma pa^3}{n} = \frac{\Sigma p(x_i - A)^3}{n}$ - умовний момент 3 порядку.

Обчислимо за цією формулою коефіцієнт асиметрії для розподілу 100 зразків води за жорсткостю, ммоль/дм^3 (табл 7.1.)

Таблиця 7.1

Класові варіанти (x)	Частоти (P)	A=(x_i-A)/i	pa	Pa^2	pa^3	pa^4
8,9	2	-4	-8	32	-128	512
9,6	3	-3	-9	27	-81	243'
10,3	9	-2	-18	36	-72	144
11,0	17	-1	-17	17	-17	17
A = 11,7	25	0	0	0	0	0
12,4	23	+ 1	+23	23	+23	23
13,1	10	+2	+20	40	+80	160
13,8	7	+3	+21	63	+ 189	567
14,5	4	+4	+ 16	64	+256	1024
Сума	100	-	+28	302	+250	2690

$$\frac{\Sigma pa^3}{n} = \frac{250}{100} = 2.50; \quad \frac{\Sigma pa^2}{n} = \frac{302}{100} = 3.02; \quad b = \frac{+28}{100} = 0.28;$$

$$3b = 0{,}84; \quad 3b\frac{\Sigma pa^2}{n} = 2.537; \quad b^3 = 0{,}0222; \quad 2b^3 = 0{,}044; \quad \sigma^3 = 5.0.$$

Підставивши в формулу [51], отримуємо

As = (2,50 - 2,54 + 0,04)/5,0 ≈ 0,01.

Отримана величина настільки мала, що дає вагому підставу вважати цей розподіл симетричним.

Ексцесом теоретичного розподілу називають характеристику, що обчислюється за формулою

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$$

Коефіцієнта ексцесу - числова характеристика розподілу ймовірностей дійсної випадкової величини. Коефіцієнт ексцесу характеризує «крутість», тобто, стрімкість підвищення кривої розподілу у порівнянні з нормальною кривою

Коефіцієнт ексцес *Ex* обчислюється за формулою:

$$Ex = \frac{\Sigma p\left(x_i - \bar{x}\right)^4}{n\sigma^4} - 3 = \frac{\Sigma p a^4}{n\sigma^4} - 3 \text{ [52]}$$

Для строго симетричних розподілів As ~ 0. позитивний ексцес має знак (+), а негативний (-). Гранична межа негативного ексцесу -2, а позитивний ексцес може бути будь-якої величини. Позитивний ексцес вважається незначним, якщо Ex <0,5 (рис.7.2)

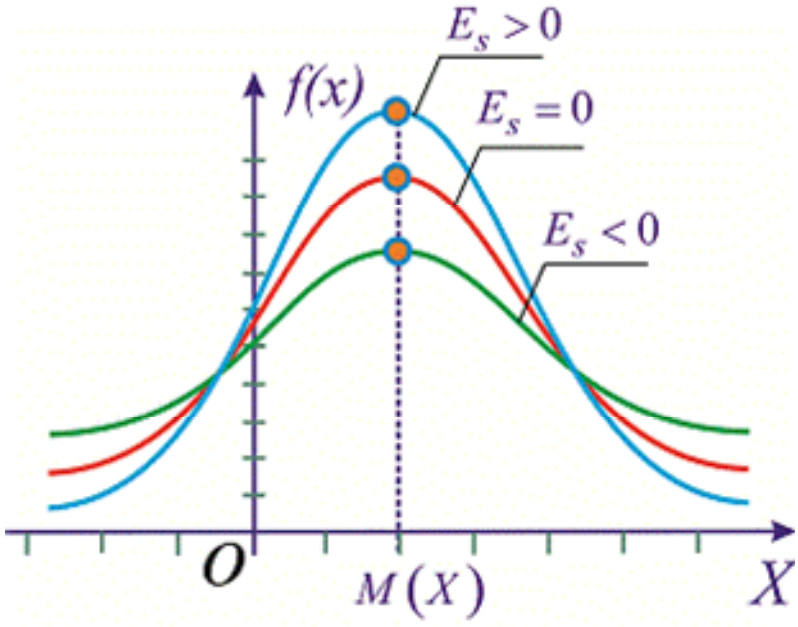


Рис. 7.2 Види ексцесу розподілу.

Для розрахунку ексцесу за способом умовної середньої скористаємося формулою:

$$Ex = \frac{\frac{\Sigma pa^4}{n} - 4b\frac{\Sigma pa^3}{n} + 6b^2\frac{\Sigma pa^2}{n} - 3b^4}{\sigma^4} - 3 \quad [53]\text{, де } a = (x - A),\ b = \frac{\Sigma p(x_i - A)}{n}.$$

$$\frac{\Sigma pa^4}{n} = \frac{2690}{100} = 26.90 \qquad \frac{\Sigma pa^3}{n} = \frac{250}{100} = 2.50 \qquad \frac{\Sigma pa^2}{n} = \frac{302}{100} = 3.02$$

Скористаємося таблицею 7.1. отримаємо:

$$b = \frac{\Sigma pa}{n} = \frac{+28}{100} = 0.28;\ \ 4b = 1{,}12;\ \ 6b^2 = 0.47;\ \ 3b^4 = 0.005;\ \ 4b\frac{\Sigma pa^3}{n} = 2.8;$$

$$6b^2\frac{\Sigma pa^2}{n} = 1.42;\ \ \sigma = 1.715;\ \ \sigma^4 = 8.64.$$

Підставивши значення, отримаємо:

Ex = [(26,90 – 2,8 + 1,42 - 0.005)/8,64] - 3 = 2,95-3 = -0,05

Отримане значення дозволяє вважати відсутність ексцесу у даного розподілу.

Як і інші вибіркові показники, коефіцієнти асиметрії та ексцесу є величинами випадковими. Щоб відрізнити уявну асиметрію і ексцес від дійсних, необхідна статистична оцінка достовірності вибіркових показників асиметрії та ексцесу. Асиметрія може виникати від як внаслідок угруповання матеріалу у варіаційні ряди (помилкова асиметрія), так бути пов’язана з особливостями набору даних.

7.3. Критерій відповідності емпіричних і теоретичних розподілів. Критерій χ^2

Статистична оцінка розбіжностей, які спостерігаються між двома і більше емпіричними розподілами або співставлення емпіричного розподілу з теоретичним (нормальним, рівномірним тощо) проводиться за допомогою особливих критеріїв відповідності. Один з них, т.з. критерій χ^2, запропонований Карлом Пірсоном в 1901 р. Критерій Пірсона дає більш точні значення для великих вибірок (n> 30).

Для знаходження величини критерію χ^2 необхідно:

1. За кожним класом варіаційного ряду знайти d.

2. Звести різницю в квадрат, і розділити її на p '.

3. Підсумувати отримані відношення для всіх класів варіаційного ряду.

Отримана величина і буде значенням χ^2, яка завжди позитивна.

$$\chi^2 = \Sigma \frac{(p - p')^2}{p'}$$ [54],

де p - емпірична частота, p'- відповідна теоретична частота. Якщо p-p'=d.

Якщо Σ(p-p') = 0, то $\chi^2 = 0$. отже, має місце повна відповідність фактичних частот очікуваним (або обчисленим). Якщо ж $\chi^2 \neq 0$, то необхідно оцінити достовірність відмінностей, порівнявши зі стандартним значенням χ^2.

Якщо $\chi^2_l \geq \chi^2_{st}$ - нульова гіпотеза відкидається.

Продемонструємо застосування критерію χ^2 на прикладі розподілу (табл. 7.2) перевіривши гіпотезу про відповідність до нормального розподілу.

Таблиця 7.2

Варіанти (x)	Частоти		Різниця (d)	Квадрат різниці (d^2)	d^2/p'
	емпіричні (p)	вирахувані (p')			
111	3	1.7	1.3	1,69	0,99
112	9	10.0	1.0	1,00	0,10
113	31	34,3	-1	10,89	0,32
114	71	67,9	j, j	9,61	0,01
115	82	77,6	3,1	19,36	0,25
116	46	51,1	4.4	26,01	0,51
117	19	19,5	5.1 ,5	0,25	0,01
118	6	4.8	1.2	1,44	0,30
Сума	267	267,2	-	-	2,49

Звідки χ^2 = 2,49. При оцінці емпіричних розподілів, які відповідають нормальному закону, число ступенів свободи k = n-3. По таблиці стандартних

значень χ^2 для числа ступенів свободи k=n-3=8-3=5 χ^2_{st} =11,1. $\chi^2_{ф} < \chi^2_{st}$ - нульова гіпотеза зберігається, розбіжності між емпіричними частотами і обчисленими за нормальним законом слід визнати випадковими.

7.4. Поняття трансгресії

При розподілі незалежних вибірок в варіаційні ряди нерідко доводиться спостерігати, що частина варіант цих вибірок розподіляються по одним і тим же класах, хоча між середніми арифметичними цих рядів існує статистично достовірна різниця.

Ряди, у яких частина класів виявляється загальною, а між середніми арифметичними виявляється статистично достовірна різниця, називаються трансгресуючими рядами. Сам факт неповного розмежування варіаційних рядів носить назву *трансгресії.*

Ступінь трансгресії може бути дуже різною. Вимірювання величини трансгресії є важливим елементом статистичного аналізу, тому що дає відповідь на питання про те, чи належать або не належать розглянуті вибірки однієї і тієї ж генеральної сукупності.

РОЗДІЛ 8. КОРЕЛЯЦІЙНИЙ АНАЛІЗ

8.1. Поняття кореляції і завдання кореляційного аналізу

Виявлення зв’язку між змінними є важливою задачею статистичного аналізу. Виділяють 3 типи залежностей між змінними:

- функціональна залежність (визначає значення змінної Y від X однозначно);
- кореляційна залежність (визначає середнє значення змінної Y від X);
- стохастична залежність (визначає розподіл змінної Y від X).

Найбільш загальною є стохастична залежність. Кореляційна залежність є стохастичною, функціональна – розглядається як окремий випадок кореляційної. Залежність між змінними випадковими величинами X і Y, при якій кожному значенню однієї з них відповідає не якесь конкретне значення, а певна групова середня іншої величини, тобто $y_x = f(x_i)$ или $x_y = f(y_i)$ називається кореляційною або просто кореляцією.

Термін "кореляція" (conelation) увів у статистику англійський біолог і статистик Френсіс Гальтон наприкінці XIX ст. До нього, у XVIII ст. відомий французький палеонтолог Жорж Кюв'є, фахівець із копалин останків тварин, ввів у науковий обіг так званий закон кореляції, який він використовував для вивчення зв'язку частин та органів живих істот. За допомогою закону кореляції можна було відновити вигляд викопної тварини, маючи в розпорядженні лише частину її останків. Подальший розвиток кореляційний аналіз отримав у працях Карла Пірсона (1857–1936) та Джорджа Юла (1871 – 1951), які розробили та ввели у науковий обіг термін "парний коефіцієнт кореляції", який до цього дня є одним із основних інструментів, що дозволяють вивчати взаємозв'язок кількох ознак.

Зміст кореляційного аналізу полягає у аналізі зв'язків, існуючих між випадковими величинами; оцінюванні за вибірковими даними коефіцієнтів

кореляції, перевірці їх значущості, оцінюванні близкозті виявленого зв`язку до лінійного та побудові довірчого інтервалу для коефіцієнтів кореляції.

У напрямку кореляція буває позитивною (прямою), і негативною (зворотною), а за формою - лінійною (прямолінійною), нелінійною або криволінійною (рис 8.1).

Кореляція вважається лінійною, коли напрямок зв'язку між ознаками X і Y графічно та аналітично виражається прямою лінією і навпаки. При позитивній кореляції, групові середні однієї ознаки зростають зі збільшенням значень іншої ознаки.

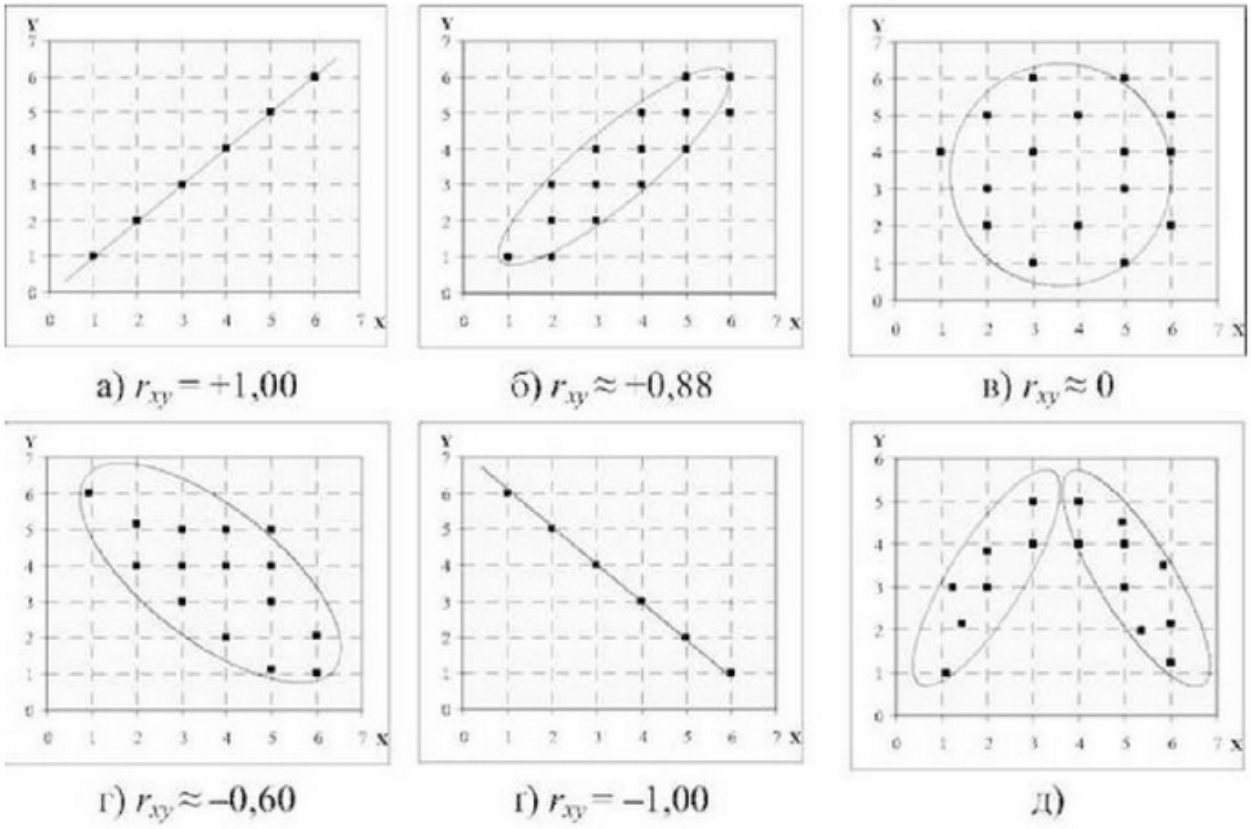


Рис. 8.1. Схематичне зображення величини та напрямку коефіцієнта кореляції: а) повна позитивна кореляція; б) сильна позитивна кореляція; в) нульова кореляція; г) помірна негативна кореляція; д) нелінійна кореляція.

При негативній кореляції групові середні однієї ознаки зменшуються при збільшенні значень іншої ознаки. Наприклад, зі збільшенням концентрації кислоти зростає швидкість корозії сталі. У всіх випадках завдання кореляційного аналізу залишаються одні і ті ж: встановлення форми

і напрямки зв'язку, що існує між варіруючими ознаками, вимір її сили або тісноти з подальшою оцінкою достовірності емпіричних показників зв'язку.

Щоб виміряти ступінь спряженості між ознаками X і Y, необхідно зіставити відповідним чином їх значення один з одним. Коефіцієнт кореляції обчислюємо за формулою:

$$r = \frac{\Sigma\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{n\sigma_x\sigma_y} \text{ [55]}$$

Коефіцієнт кореляції величина відносна і виражається в частках одиниці. Формула

$$r = \frac{n\Sigma xy - \Sigma x\Sigma y}{\sqrt{n\Sigma x^2 - \left(\Sigma x\right)^2} \times \sqrt{n\Sigma y^2 - \left(\Sigma y\right)^2}} \text{ [56]}$$

дозволяє окремо не обчислювати середнє відхилення, що помітно полегшує розрахунок коефіцієнта кореляції.

8.2. Основні властивості коефіцієнта кореляції

Коефіцієнт кореляції служить для вимірювання сили або тісноти лінійного зв'язку між значеннями ознак X і Y. Коефіцієнт кореляції лежить в межах -1 < r < +1. При наявності позитивної зв'язку між варіруючими ознаками коефіцієнт кореляції має знак (+), при наявності зворотного або негативного зв'язку r має знак (-). Знак коефіцієнта кореляції вказує на напрям – прямий чи зворотній взаємозв`язок між змінними. Коли r = 0, це означає відсутність кореляції, при r = 1 очевидний функціональний зв'язок між ознаками. Таким чином, при r > 0 цей показник характеризує не тільки наявність, але і ступінь пов'язаності між значеннями варіюючих ознак. Чим сильніше спряженість, тим вище коефіцієнт кореляції і навпаки. А знак при r дозволяє визначати напрямок зв'язку.

Зазвичай вважається, що r < 0,3 вказує на слабкий зв'язок, при 0,3 < r < 0,5 зв'язок визнається помірним. Якщо ж 0,5 < r < 0,7 - кореляція вважається значною, при 0,7 < r < 0,9 - сильною і при r > 0,9 дуже сильною, близькою до функціонального зв'язку.

8.3. Довірча оцінка коефіцієнта кореляції

Так як вибірковий коефіцієнт кореляції є величиною випадковою, і може виявитися відмінним від 0 навіть при незначному варіюванні ознак. Звідси виникає необхідність розглядати коефіцієнт кореляції в якості оцінки генерального параметра (ρ). Нульова гіпотеза стосовно оцінки генерального ρ величиною емпіричного коефіцієнта кореляції (r) полягає в припущенні, що ρ = 0, тобто між випадковими величинами X і Y кореляція відсутня. Для перевірки нульової гіпотези служить t-критерій Стьюдента.

Помилка коефіцієнта кореляції обчислюється за формулою (для n <100):

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$ [57]

Нульова гіпотеза відхиляється, якщо $t_ф > t_{st}$ для k = n-2 і взятого рівня значущості (P). Це означає, що в генеральній сукупності ρ ≠ 0 і, отже, вибірковий коефіцієнт кореляції достовірно відрізняється від 0, і між X і Y існує кореляційний зв'язок. При $t_ф < t_{st}$ нульова гіпотеза зберігається, відхилення вибіркового коефіцієнта кореляції від 0 вважається чисто випадковим.

Наприклад, на вибірці n = 36 отриманий r - 0,46. Потрібно оцінити достовірність цієї величини. критерій достовірності:

$$t_ф = \frac{0.46\sqrt{36-2}}{\sqrt{1-0.46^2}} = \frac{2.682}{0.888} = 3.0$$

По таблиці Стьюдента для k = 36 - 2 = 34 і P = 0,01 знаходимо $t_{st} = 2,58$, $t_ф > t_{st}$. Нульова гіпотеза відхиляється.

8.4. Метод Z

На нечисленних вибірках оцінка коефіцієнта кореляції описаним способом може виявитися недостатньо точною. В цому випадку доцільно застосовувати запропонований Фішером метод Z. Р. Фішер запропонував замість коефіцієнта кореляції для його оцінки використовувати пов'язану з ним допоміжну величину Z:

$$Z = \frac{1}{2}\ln\frac{1+r}{1-r} \text{ або } Z = 1.15129\lg\frac{1+r}{1-r} \text{ [58]}$$

Визначити показник Z за відомим коефіцієнтом кореляції можна за допомогою таблиці, складеної Фішером, де вказані значення Z, що відповідають різним величинам коефіцієнта кореляції. Критерієм достовірності показника служить такий вираз:

$$t_z = Z\sqrt{n-3} \text{ або } t_z = \frac{Z}{\sigma_Z} = Z\sqrt{n-3} \text{ [59]}$$

Цей критерій може застосовуватися як для малих, так і для великих вибірок; він використовується у всіх випадках, коли замість коефіцієнта кореляції береться відповідне йому значення Z.

Для оцінки достовірності і встановлення довірчого інтервалу, по якому з достатньою ймовірністю можна судити про величину коефіцієнта кореляції генеральної сукупності, надходять у такий спосіб: за значенням емпіричного коефіцієнта кореляції в таблиці знаходимо значення Z і визначаємо величину помилки показника за формулою:

$$\sigma_Z = \frac{1}{\sqrt{n-3}} \text{ [60]}$$

За формулою [55] знаходимо емпіричне значення, яке порівнюємо зі стандартним по таблиці Стьюдента для прийнятого P і до n - 2.

Наприклад, на вибірці n = 28 отримуємо r = 0,52. По таблиці знаходимо Z = 0,576. Обчислюємо помилку:

$$t_Z = Z\sqrt{n-3} = 0.576 \times \sqrt{25} = 2.88$$

По таблиці Стьюдента для A = 28 - 2 = 26 і P = 0,05 знаходимо t_{st} = 2,06. Оскільки $t_ф > t_{st}$ нульова гіпотеза не зберігається.

$$\sigma_Z = \frac{1}{\sqrt{28-3}} = \frac{1}{5} = 0.20$$

За величиною $\Delta Z = t\sigma_z = 1,96 \cdot 0,20 = 0,392$ знаходимо межу довірчого інтервалу для показника Z:

нижня межа = Z - ΔZ = 0,576 - 0,392 = 0,184;

верхня межа = Z + ΔZ = 0,576 + 0,392 = 0,968.

Користуючись таблицею, переводимо значення Z в коефіцієнт кореляції і знаходимо його довірчі межі: нижня = 0,18; верхня = 0,74.

Це означає, що величина коефіцієнта кореляції в генеральній сукупності знаходиться між межами 0,18 < r <0,74. Можна сказати, що емпіричний коефіцієнт кореляції r = 0,52 визначений з достатньою точністю.

8.5. Мінімальна кількість спостережень для планованої точності коефіцієнта кореляції

Статистична недостатність емпіричного коефіцієнта кореляції ще не доводить, що зв'язку між варіруючими ознаками немає. При достатній кількості спостережень цей зв'язок може виявитися достовірно. Розрахувати необхідний обсяг вибірки для планованої точності коефіцієнта кореляції можна за формулою:

$$n = \frac{t_Z^2}{Z^2} + 3\text{, [61]}$$

де *n* – бажана кількість спостережень, *t* - величина, задана за прийнятим порогом довірчої ймовірності (краще для P = 0,99).

Наприклад, для n = 0,25 та n=20, Z = 0,2554, t_z= $0,2554\sqrt{17}$ = 1,05. P = 0,05, k = 20 - 2 = 18, t_{st} = 2,10. Нульову гіпотезу відкинути не можна. Відомо, що довірчої ймовірності P = 0,95 відповідає нормоване відхилення t = 1,96, тому

$$n = \frac{1,96^2}{0,2554^2} + 3 = \frac{3,842}{0,065} + 3 = 59 + 3 = 62$$

Отже, щоб задовольнити поставленому завданню, необхідно провести не менше 62 спостережень.

8.6. Оцінка різниці між коефіцієнтами кореляції

Метод Z дозволяє оцінити достовірність різниці між емпіричними коефіцієнтами кореляції, обчисленими на незалежних вибіркових сукупностях. Помилка різниці $Z_1 – Z_2$ визначається за формулою:

$$m_{D_z} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}} \quad [62]$$

критерій достовірності різниці за формулою:

$$t_D = \frac{Z_1 - Z_2}{m_{D_z}} \quad [63]$$

Наприклад, визначаючи залежність між концентрацією препарату (мг) і його антиоксидантною активністю в одному випадку на вибірці n = 50 отримано r = +0,56, а в іншому - на вибірці n = 44, r = + 0,48. Різниця r_1-r_2 = 0,56 - 0,48 = 0,08. З'ясуємо, чи випадкова ця розбіжність. По таблиці знаходимо значення Z_1 = 0,633, Z_2 = 0,523. критерій достовірності:

$$t_D = \frac{0.633 - 0.523}{\sqrt{\frac{1}{50 - 3} + \frac{1}{44 - 3}}} = \frac{0.11}{\sqrt{0.0456}} = 0.52$$

Так як для P = 0,95 t_{st} = 1,96. На основі того, що t_D = 0,52, m. e. $t_D < t_{st}$, робимо висновок про випадковість різниці між коефіцієнтами кореляції.

8.7. Обчислення коефіцієнта кореляції на малих вибірках

На вибірках невеликого обсягу коефіцієнт кореляції обчислюють, не вдаючись до розподілу вибіркового матеріалу в варіаційні ряди і угруповання його в кореляційну таблицю. Для прикладу проаналізуємо дані про концентрацію вихідної речовини і вихід продукту (табл. 8.1). Передбачається, що між концентрацію вихідної речовини і виходом продукту існує прямолінійний позитивний зв'язок. Обчислимо для цих даних коефіцієнт кореляції. Для цього можна скористатися робочими формулами для малих вибірок:

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} \times \sqrt{n\Sigma y^2 - (\Sigma y)^2}} \quad [64], \text{ або}$$

$$r = \frac{D_x + D_y - D_d}{2\sqrt{D_x \times D_y}} \text{ , де } D_x = \Sigma x^2 - \frac{(\Sigma x)^2}{n};\ D_y = \Sigma y^2 - \frac{(\Sigma y)^2}{n};\ D_d = \Sigma d^2 - \frac{(\Sigma d)^2}{n};$$

d = x - y; n - число парних спостережень.

Щоб використовувати ту чи іншу формулу необхідно попередньо знайти допоміжні значення (табл. 8.1).

Беремо з даної таблиці підсумкові цифри і знаходимо середнє арифметичне

$$\bar{x} = \frac{237.4}{20} = 11.87 \text{ мг/л}; \; \bar{y} = \frac{14.06}{20} = 0.703 \text{ %}.$$

Визначаємо суми квадратів відхилень:

$$D_x = 2861.60 - \frac{237.4^2}{20} = 43.662; \; D_y = 9.9598 - \frac{14.06^2}{20} = 0.0756;$$

$$D_d = 2535.7218 - \frac{223.34^2}{20} = 41.6840$$

Таблиця 8.1

$C_{поч}$, мг/л (X)	Вихід продукту, % (Y)	X·Y	X^2	Y^2	x-y=d	d^2
10,0	0,70	7,000	100,00	0,4900	9,30	86,4900
10,8	0,73	7,884	116,64	0,5329	10,07	101,4049
11,3	0,75	8,475	127,69	0,5625	10,55	111,3025
10,0	0,70	7,000	100,00	0,4900	9,30	86,4900
10,1	0,65	6,565	102,01	0,4225	9,45	89,3025
11,1	0,65	7,215	123,21	0,4225	10,45	109,2025
11,3	0,70	7,910	127,69	0,4900	10,60	112,3600
10,2	0,61	6,222	104,04	0,3721	9,59	91,9681
13,5	0,70	9,545	182,25	0,4900	12,80	163,8400
12,3	0,63	7,749	151,29	0,3969	11,67	136,1889
14,5	0,70	10,150	210,25	0,4900	13,80	190,4400
11,0	0,65	7,150	121,00	0,4225	10,35	107,1225
12,0	0,72	8,640	144,00	0,5184	11,28	127,2384
11,8	0,69	8,142	139,24	0,4761	11,11	123,432
13,4	0,78	10,452	179,56	0,6084	12,62	159,2644
11,4	0,70	7,980	129,96	0,4900	10,70	114,4900
12,0	0,60	7,200	244,00	0,3600	11,40	129,9600
15,6	0,85	13.260	1243,36	0,7225	14,75	1217,5625
13,0	0,80	10,400	169,00	0,6400	12,20	148,8400
12,1	0,85	9,075	146,41	0,5625	1L35	128,8225
∑ 237,4	14,06	167,939	2861,60	9,9598	223,34	2535,7218

Підставляємо знайдені значення в формулу і визначаємо коефіцієнт кореляції:

$$r = \frac{20 \times 167.939 - 237.4 \times 14.06}{\sqrt{20 \times 2861.60 - (237.4)^2} \times \sqrt{20 \times 9.9598 - (14.06)^2}} = \frac{20.936}{\sqrt{873.24} \times \sqrt{1.51}} = \frac{20.936}{36.346} = +0.58$$

$$r = \frac{43.662 + 0.0756 - 41.6840}{2\sqrt{43.662 \times 0.0756}} = \frac{2.054}{2\sqrt{3.309}} = \frac{2.054}{3.616} = +0.58$$

Отримана величина (r = 0,58) вказує на наявність значного позитивного зв'язку між концентрацію вихідної речовини і виходом продукту, Оцінимо достовірність отриманої величини. Для r = 0,58 знаходимо по таблиці значення Z = 0,663. Звідси:

$$t_Z = Z\sqrt{n-3} = 0.663\sqrt{20-3} = 0.663 \times 4.12 = 2.73$$

За таблицею Стьюдента для P = 0,05 і Z = 20 - 2 = 18 і знаходимо t_{st} = 2,10. Оскільки $t_Z > t_{st}$ нульова гіпотеза відкидається, величина r = 0,58 виявляється достовірною.

8.8. Кореляційне відношення

Для вимірювання криволінійної залежності між змінними величинами X і Y коефіцієнт кореляції непридатний. У таких випадках використовується інший показник - кореляційне відношення, що позначається грецькою буквою η (ета). На відміну від коефіцієнта кореляції, який характеризує залежність між X і Y з точки зору прямої пропорційності, кореляційне відношення описує її двосторонньо. Розберемо на прикладі. Візьмемо кілька парних значень двох змінних величин X і Y (табл 8.2):

Таблиця 8.2

X:	2	4	6	8	4	6	2	6
Y:	4	8	8	7	4	10	6	12

Ранжируємо цю сукупність за X (табл. 8.3)

Таблиця 8.3

X:	2	2	4	4	6	6	6	8
Y:	4	6	4	8	10	8	12	7
$\bar{Y}$*:*	5		6		10			7

З урахуванням повторюваності (табл 8.4)

Таблиця 8.4

X:	2	4	6	8
$\bar{Y}_x$:	5	6	10	7

де $\acute{Y}_x$ - приватні чи групові середні, відповідні однаковим значенням X. Якщо ж ранжувати сукупність по Y, то вийде наступне (табл. 8.5):

Таблиця 8.5

Y:	4	4	6	7	8	8	10	12
X:	2	4	2	8	6	4	6	6
$\bar{X}$:	3		2	8	5		6	6

З урахуванням повторюваностей (табл. 8.6):

Таблиця 8.6

Y:	4	6	7	8	10	12
$\bar{X}_y$:	3	2	8	5	6	6

Таким чином залежність між змінними X і Y виражається по різному в залежності від того, за значеннями який з них ранжирується сукупність по X або Y. Тому кореляційне відношення виражається не одним, а двома показниками $\eta_{y/x}$ і $\eta_{x/y.}$ Вони обчислюються за такими формулами:

$$\eta_{y/x} = \sqrt{\frac{\sigma_{yx}^2}{\sigma_y^2}} \text{ та } \eta_{x/y} = \sqrt{\frac{\sigma_{xy}^2}{\sigma_x^2}} \text{ [65]}$$

де $\sigma_{yx}^2 = \frac{\Sigma\left(\bar{y}_x - \bar{y}\right)^2}{n}$ - середній квадрат відхилень приватних або групових середніх ($\acute{y}_x$) від загальної середньої ($\acute{y}$), тобто часткова дисперсія,

$\sigma_y^2 = \frac{\Sigma\left(\bar{y}_i - \bar{y}\right)^2}{n}$ – загальна дисперсія сукупності.

Відповідно:

$$\sigma_{xy}^2 = \frac{\Sigma\left(\bar{x}_y - \bar{x}\right)^2}{n} \text{ та } \quad \sigma_x^2 = \frac{\Sigma\left(\bar{x}_i - \bar{x}\right)^2}{n} \text{ [66]}$$

Як і коефіцієнт кореляції, кореляційне відношення - величина відносна, η має значення від 0 до 1: чим сильніше зв'язок, тим вище значення η. При

відсутності кореляції η = 0. При цьому кореляційне відношення - величина завжди позитивна.

Показники кореляційного відношення зазвичай не рівні між собою, тобто $\eta_{y/x} \neq \eta_{x/y.}$. Лише при строго лінійного зв'язку між Х і У здійснюється рівність $\eta_{y/x} = \eta_{x/y.}$. Ця особливість кореляційного відношення дозволяє характеризувати будь-яку кореляційну залежність між варіруючими ознаками - і лінійну і криволінійну. Чим ближче зв'язок між ознаками наближається до прямолінійного функціонального зв'язку, тим ближче за абсолютною величиною показники кореляційного відношення один до одного.

На малих вибірках кореляційне відношення обчислюється за розглянутою вище формулою без угруповання в варіаційні ряди і в кореляційні таблиці.

Розглянемо обчислення кореляційного відношення на прикладі: з'ясовувалася наявність кореляційного зв'язку між концентрацією каталізатору, % (Y) і масою продукту реакції, г (X) (табл. 8.7). Обчислимо для цих даних кореляційне відношення.

Таблиця 8.7

С, % **(y)**	Маса, г **(x)**	$\bar{x}_y$	$\bar{x}_y - \bar{x}$	$(\bar{x}_y - \bar{x})^2$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
5,0	24,0	26,25	4,55	20,70	6,8	46,24
6,2	26,0	26,25	4,55	20,70	2,3	5,29
5,4	26,0	32,00	1,20	1,44	1,2	1,44
5,0	28,5	28,50	2,30	5,29	4,8	23,04
6,7	28,5	28,50	2,30	5,29	0,2	0,04
8,0	29,0	32,50	1,70	2,89	0,7	0,49
5,7	31,0	32,50	1,70	2,89	2,7	7,29
5,4	31,0	32,50	1,70	2,89	1,7	2,89
6,1	31,0	31,00	0,20	0,04	0,2	0,04
5,5	31,5	32,00	1,20	1,44	1,2	1,44
5,8	32,0	33,50	2,70	7,29	2,7	7,29
5,3	32,0	31,00	0,20	0,04	0,2	0,04
6,8	32,5	26,00	4,80	23,01	4,8	23,04
5,6	32,5	34,00	3,20	10,24	3,2	10,24
6,4	32,5	32,50	1,70	2,89	1,7	2,89
7,5	33,0	28,50	2,30	5,29	2,3	5,29
5,5	33,5	33,25	2,45	6,00	3,2	10,24
6,0	33,5	33,25	2,45	6,00	1,7	2,89

6,3	34,0	33,00	2,20	4,84	2 2	4,84
6,8	34,0	29,00	1,80	3,24	1,8	3,24
Сума	-	-	-	132,44	-	158,20

$$\overline{x} = 616 / 20 = 30.8 \text{ г}, \ \overline{y} = 121.0 / 20 = 6.05$$

визначаємо кореляційне відношення концентрації каталізатору за масою продукту реакції:

$$\eta_{x/y} = \sqrt{\frac{132.44}{158.20}} = \sqrt{0.84} = 0.92$$

Таким же способом визначаємо кореляційне відношення маси продукту реакції за концентрацією каталізатору (табл. 8.8).

Підставивши в формулу [65] значення з таблиці отримаємо:

$$\eta_{y/x} = \sqrt{\frac{9.17}{12.57}} = \sqrt{0.73} = 0.85$$

Таким чином, $\eta_{y/x}$= 0,85, $\eta_{x/y}$= 0,92. це вказує на досить сильну залежність, що існує між концентрацією каталізатору та масою продукту реакції.

Таблиця 8.8

С, % **(y)**	Маса, г **(x)**	y'_y	$y'_y - y'$	$(y'_y - y')^2$	$y_i - y'$	$(y_i - y')^2$
5,0	24,0	5,00	1,05	1,1025	1,05	1,1025
6,2	26,0	5,80	0,25	0,0625	0,15	0,0225
5,4	26,0	5,80	0,25	0,0625	0,65	0,4225
5,0	28,5	5,85	0,20	0,0400	1,05	1,1025
6,7	28,5	5,85	0,20	0,0400	0,65	0,4225
8,0	29,0	8,00	1,95	3,8025	1,95	3,8025
5,7	31,0	5,73	0,32	0,1024	0,35	0,1225
5,4	31,0	5,73	0,32	0,1024	0,65	0,4225
6,1	31,0	5,73	0,32	0,1024	0,05	0,0025
5,5	31,5	5,50	0,55	0,3025	0,55	0,3025
5,8	32,0	5,55	0,50	0,2500	0,25	0,0625
5,3	32,0	5,55	0,50	0,2500	0,75	0,5625
6,8	32,5	6,27	0,22	0,0484	0,75	0,5625 ,

5,6	32,5	6,27	0,22	0,0484	0,45	0,2025
6,4	32,5	6,27	0,22	0,0484	0,65	0,4225
7,5	33,0	7,50	1,45	2,1025	1,45	2,1025
5,5	33,5	5,73	0,32	0,1024	0,55	0,3025
6,0	33,5	5,73	0,32	0,1024	0,05	0,0025
6,3	34,0	6,55	0,50	0,2500	0,25	0,0625
6,8	34,0	6,55	0,50	0,2500	0,75	0,5625
Сума	-	-		9,1722	-	12,5700

Оцінимо достовірність отриманих величин за критерієм:

$$t = \eta\sqrt{\frac{n-2}{1-\eta^2}} \quad [67]$$

для прийнятого рівня значущості (Р) і відповідного числа ступенів свободи (k = n = 2), Якщо $t_ф \geq t_{st}$ нульова гіпотеза відхиляється

$$t_{y/x} = 0.85\sqrt{\frac{20-2}{1-0.85^2}} = 0.85\sqrt{66.7} = 6.94\,; \qquad t_{x/y} = 0.92\sqrt{\frac{20-2}{1-0.92^2}} = 0.92\sqrt{112.5} = 9.71$$

По таблиці Стьюдента для Р = 0,01 і k = 18, t_{st} = 2,28, $t_ф \geq t_{st}$ нульова гіпотеза відхиляється, отримані значення $\eta_{y/x}$, $\eta_{x/y}$ статистично достовірні.

ПРАКТИЧНА РОБОТА 5

Перевірити достовірність даних одержаних в паралельних експериментах при вивченні залежності початкової швидкості реакції від концентрації речовини за методом Z.

Варіант 1

C, моль/л	$v \cdot 10^6$, моль/(л·с)	$v \cdot 10^6$, моль/(л·с)
0,246	5,076	5,373
0,181	4,545	4,871
0,161	4,464	4,663
0,120	4,118	4,118
0,095	3,632	3,77
0,076	3,333	3,269
0,063	2,778	2,915
0,060	2,762	2,828
0,052	2,513	2,591

Варіант 2

C, моль/л	$v \cdot 10^6$, моль/(л·с)	$v \cdot 10^6$, моль/(л·с)
0,420	6,110	5,435
0,290	5,622	5,155
0,184	4,895	4,608
0,153	4,568	4,329
0,089	3,770	3,534
0,064	2,943	2,857
0,053	2,611	2,519
0,041	2,191	2,137
0,039	2,098	2,053

Варіант 3

C, моль/л	$v \cdot 10^6$, моль/(л·с)	$v \cdot 10^6$, моль/(л·с)
0,25	5,23	6,494
0,17	4,8	5,682
0,144	4,4	4,673
0,11	3,94	4,167
0,085	3,49	3,690
0,074	3,24	3,413
0,065	3,1	3,145
0,061	2,91	2,857
0,058	2,82	2,762

Варіант 4

C, моль/л	$v \cdot 10^6$, моль/(л·с)	$v \cdot 10^6$, моль/(л·с)
0,520	6,412	7,143
0,392	6,081	6,081
0,284	5,633	6,711
0,223	5,248	5,848
0,118	4,092	4,329
0,085	3,462	3,623
0,063	2,910	3,021
0,041	2,177	2,212
0,028	1,611	1,650

Варіант 5

C, моль/л	$v \cdot 10^6$, моль/(л·с)	$v \cdot 10^6$, моль/(л·с)
0,28	5,571	5,128
0,19	4,950	4,739
0,154	4,579	4,348
0,12	4,118	3,968
0,09	3,770	3,690
0,084	3,444	3,378
0,063	2,914	2,994
0,06	2,827	2,857
0,056	2,706	2,688

ПРАКТИЧНА РОБОТА 6

Розрахувати коефіцієнт кореляції (для малих вибірок) та здійснити довірливу оцінку використовуючи t критерій Стьюдента

Варіант 1

Для побудови калібрувального графіку одержано такі значення показника заломлення для серії стандартних розчинів цукру

0	0,2	0,5	0,8	1,2	1,5	2	2,5	3	3,5	4	4,5
1,333	1,3335	1,3343	1,335	1,336	1,3367	1,338	1,339	1,341	1,343	1,345	1,346

Варіант 2

Для побудови калібрувального графіку одержано такі значення показника заломлення для серії стандартних розчинів цукру

0	0,6	0,9	1,28	1,53	1,8	2,1	2,7	3,1	3,8	4,3	4,9
1,333	1,328	1,329	1,332	1,333	1,335	1,336	1,338	1,339	1,343	1,345	1,349

Варіант 3

Для побудови калібрувального графіку одержано такі значення показника заломлення для серії стандартних розчинів органічної речовини

0	10	20	30	40	50	55	60	65	70	75	80
1,333	1,344	1,356	1,368	1,38	1,392	1,4	1,4047	1,4098	1,4167	1,4223	1,4301

Варіант 4

Для побудови калібрувального графіку одержано такі значення показника заломлення для серії стандартних розчинів органічної речовини

0	10	20	30	40	50	55	60	65	70	75	80
1,333	1,352	1,372	1,392	1,412	1,432	1,444	1,452	1,4672	1,479	1,498	1,514

Варіант 5

Для побудови калібрувального графіку одержано такі значення показника заломлення для серії стандартних розчинів органічної речовини

0	10	20	30	40	50	55	60	65	70	75	80
1,333	1,346	1,36	1,374	1,388	1,402	1,4	1,416	1,4328	1,44	1,449	1,463

ПРАКТИЧНА РОБОТА 7

Розрахувати кореляційне відношення у по х

Варіант 1

(*x*)	5,8	3,8	3,4	7,9	4,1	5,4	4,6	5	4,6	1,8	5,2	8,5	6,4	4,2	7	9	3,3	5,3	2,9	8
(*y*)	4,9	3	3,3	6,4	3,5	3,9	5,4	4,1	4,1	2	4,9	5,4	3,9	4	5,9	5,9	3,3	4,9	3,4	6,3

Варіант 2

(*x*)	11,6	12,3	13,9	15,1	13,8	12,8	11,2	14	14,9	11,8	12,5	11,4	16,1	10,7	12	11,1	12,7
(*y*)	3	5,9	3,3	4,1	3,5	4,9	3,9	3,3	4,1	4,9	3,4	5,4	4	5,9	2	6,3	3,9

Варіант 3

(*x*)	23	24,2	21,8	24,9	21,6	26,1	22,9	33,7	25,1	22,6	27,7	25,1	28,7	35,9	24,5	29,1	21,9
(*y*)	2	3,5	4,9	4,1	3,9	4	3	6,3	4,1	4,9	5,4	3,3	5,9	6,4	3,9	5,9	3,4

Варіант 4

(*x*)	6,9	7,8	8,5	9,7	7,9	6,2	6,4	8,2	8,3	9,1	7,8	7,9	9	8,5	6,7	8,1	8,5	8,6	9,8	9
(*y*)	5,4	3	3,3	3,4	3,9	4,9	3,9	6,4	4	4,1	4,9	5,4	4,1	5,9	2	5,9	3,5	3,3	4,9	6,3

Варіант 5

(*x*)	2,9	2,2	3,4	2,1	2,9	3,2	3,4	4	3,8	4	2,8	1,9	2,9	2,5	1,2	2,1	3,5	1,8	1,8	3
(*y*)	5,4	3	3,3	3,4	3,9	4,9	3,9	6,4	4	4,1	4,9	5,4	4,1	5,9	2	5,9	3,5	3,3	4,9	6,3

РОЗДІЛ 9. РЕГРЕСІЙНИЙ АНАЛІЗ

9.1. Поняття регресії

Поняття регресії вперше запропоновано в 1886 р. англійським дослідником Френсісом Гальтоном. Після знайомства з книгою Чарльза Дарвіна «Походження видів» в 1859 р. він зацікавився тим фактом, що люди із покоління в покоління не сильно відрізняються за зовнішнім виглядом і природними здібностями. Це привело його до вивчення спадковості, зокрема вивчення залежності росту дітей від росту батьків. За логікою діти мають бути дуже схожі на своїх батьків. Високі батьки повинні мати високих дітей, а низькорослі батьки – дітей низького росту. Таким чином, через декілька поколінь ми мали б, з одного боку, рід велетнів, з другого — рід карликів. У результаті проведення великої кількості дослідів над тваринами та статистичних досліджень Ф. Гальтон переконався, що такої тенденції немає, а, навпаки, нащадки дуже високих або дуже низьких батьків у середньому мають менш високий або відповідно менш низький зріст. Цей рух назад до середнього Ф. Гальтон назвав регресією (to regress — рухатися у зворотному напрямку).

Коефіцієнт кореляції та кореляційне відношення дозволяють вимірювати ступінь спряженості між ознаками, визначати напрямок і форму існуючого між ними зв'язку. Але вони не дають уявлення про те, наскільки в середньому може змінитися варіруюча ознака при зміні на одиницю виміру іншої, пов'язаної з ним ознаки. Функція, що дозволяє за величиною однієї ознаки (x) знаходити середні (очікувані) значення іншої ознаки, пов'язаної з X кореляційно, називається *регресією*. Статистичний аналіз регресії отримав назву *регресійного аналізу*. *Регресійний аналіз* — розділ математичної статистики, присвячений методам аналізу залежності однієї величини від іншої. На відміну від кореляційного аналізу не з'ясовує істотний зв'язок, а займається пошуком моделі цього зв'язку, вираженої у функції регресії. Регресійний аналіз використовується в тому випадку, якщо відношення між

змінними можуть бути виражені кількісно у виді деякої комбінації цих змінних.

Мета регресійного аналізу полягає у:

- прогнозуванні значень залежної ознаки за певним значенням незалежної ознаки;
- визначенні ступеня детермінованості варіації залежної змінної за допомогою незалежної
- визначення внеску окремих незалежних змінних у варіацію залежної.

Регресійний аналіз може бути застосованим за умови, що всі ознаки є кількісними та відповідають закону нормального розподілу. У випадку багатофакторного аналізу не повинні існувати сильні лінійні зв`язки між незалежними ознаками.

Призначенням регресійного аналізу є отримання за експериментальними даними математичного рівняння (моделі), що описує зміну деякої величини у залежно від х.

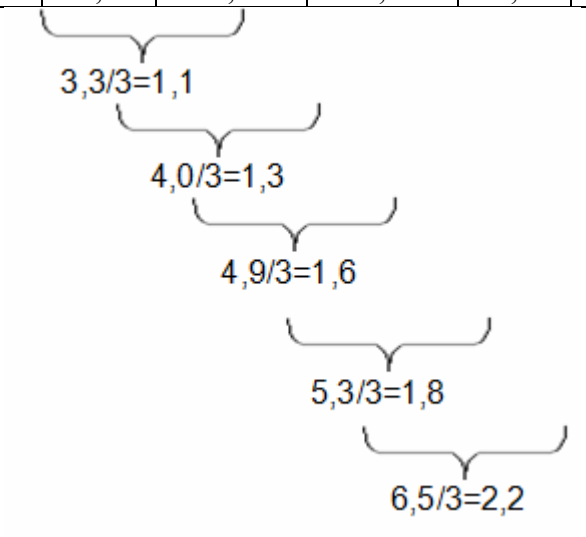
Найпростішою формою регресійного аналізу є вирівнювання рядів значень. Під вирівнюванням мається на увазі спосіб заміни ламаної лінії або ряду - регресії, динаміки, розподілу на плавно поточну, згладжену лінію, або звільнений від вагаючих значень чисельний ряд. Існують різні способи вирівнювання рядів.

1) Графічний спосіб. Даний спосіб не вимагає обчислювальної роботи. Після нанесення емпіричного ряду на графік, на око визначають середні точки лінії регресії, які потім з'єднуються. Недоліком способу є те, що він не виключає вплив індивідуальних властивостей дослідника на результат вирівнювання.

2) Спосіб ковзаючої середньої. Більш точний результат дає вирівнювання емпіричних рядів послідовним визначенням середніх арифметичних з двох або трьох сусідніх значень ряду. Наприклад, є дані про зміну маси продуктів ферментативної реакції з часом (табл. 9.1). Знаходимо суми та середні арифметичні:

Таблиця 9.1

Час, хв	0	1	2	1 J	4	5	6
Маса продукту (г):	0,7	1,0	1,6	1,4	1,9	2,0	2,6



Отримуємо усереднені значення ряду: 1,1; 1,3; 1,6; 1,8; 2,2. Спосіб ковзної середньої простий і особливо зручний в тих випадках, коли емпіричний ряд представлений багатьом числом членів і втрата двох з них (крайніх) помітно не позначається на його загальній структурі. Цінність цього методу в тому, що він дозволяє себе модернізувати: усереднені величини можна отримувати з двох, трьох і більшого числа членів емпіричного ряду.

9.2. Лінійна регресія

Для математичного виразу зв'язку між змінними X і Y служить рівняння загального вигляду Y = f(x), де символом f(x) позначається підбираєма форма рівняння, яка більш-менш повно виражає функціональну залежність середньої величини однієї змінної від значень іншої змінної величини X. Такого роду математичні рівняння називаються кореляційними або регресійний.

Залежність між ознаками може бути найрізноманітнішою. В більшості випадків емпіричні регресії виражаються простим рівнянням лінійної залежності:

$$\bar{y}_x = a + bx$$

Тут y_x групова середня арифметична, або очікуване значення змінної Y, відповідне заданому значенню змінної X; a й b - параметри рівняння; a - служить вільним членом, b - є показником пропорційності, який називають коефіцієнтом регресії.

Розрахунок емпіричного рівняння регресії проведемо за формулами:

$$b_{y/x} = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma y^2 - n\bar{y}^2} \text{ [68]; } a = \bar{y} - b\bar{x} \text{ [69]}$$

Для з'ясування залежності між тиском у системі (атм) та виходом продукту (%) в реакції A→B (табл. 8.2), попередньо розрахуємо допоміжні величини.

n=15; $\bar{x}$=162.247; $\bar{y}$=84.26; $\bar{x^2}$=26323.9917; $\bar{y^2}$=7099.7476.

За підсумковими даними таблиці за формулами знаходимо:

$$b_{y/x} = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma y^2 - n\bar{y}^2} = \frac{205669.55 - 15 \times 162.247 \times 84.26}{397213.75 - 15 \times 26323.9917} = \frac{622.0667}{2353.8745} = 0.264$$

$$a = \bar{y} - b\bar{x} = 84.26 - 0.264 \times 162.247 = 41.44$$

$$\bar{y}_x = 0.264x + 41.44$$

Таблиця 9.2

X	Y	XY	X^2	Y^2	Y_x	$\acute{Y}_x - Y$	$(\acute{Y}_x - Y)^2$
148,5	81,5	12102,75	2205225	6642,25	80,7	3,56	12,6736
150,5	82,0	1234 1,00	22650,25	6724,00	812	3,06	9,3636
152,5	80,7	12306,75	23256,25	6512,49	81,7	2,56	6,5536
154,5	81,0	12514,50	2387025	6561,00	822	2,06	4,2436
156,5	82,1	12848,65	2449225	6740,41	82,6	1,66	2,7556
158,5	83,8	13282,30	2512225	7022,44	83,3	0,96	0,9216
160,5	83,9	13465,95	2576025	7039,21	83,6	0,66	0,4356
162,5	83,9	13633,75	26406,25	7039,21	84,3	0,04	0,0016
164,5	85,0	13982,50	2706025	7225,00	84,8	0,54	0,2916
166,5	86,1	14335,65	2772275	7413,21	852	0,94	0,8836
168 J	86,7	14608,95	2839225	7516,89	85,7	1,44	2,0736
170 5	86,1	14680,05	2907025	7413,21	86,3	2,04	4,1616
172,5	86,4	14904,00	2975625	7464,96	87,0	2,74	7,5076
174,5	85,9	14989,55	3045025	7378,81	87,3	3,04	9,2416

176 5	88,8	15673,20	31152,25	7885,44	88,0	3,74	13,9876
2437,5	1263,9	205669,55	397213,75	106758,53	1263,9	29,04	75,0960

Додатково до графічного зображення регресії можна визначити міру лінійності:

$$\gamma = \eta^2 - r^2 \text{ [70]}$$

Коефіцієнт кореляції визначаємо за формулою:

$$r = \frac{\Sigma xy - \frac{\Sigma x \times \Sigma y}{n}}{\sqrt{\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right) \times \left(\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right)}}$$

$$\Sigma\left(y_i - \bar{y}\right)^2 = \Sigma y^2 - \frac{(\Sigma y)^2}{n} = 106578.53 - \frac{1263.9^2}{15} = 82.0$$

$$\Sigma\left(x_i - \bar{x}\right)^2 = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 397213.75 - \frac{2437.5^2}{15} = 1120.0$$

$$r = \frac{205669.55 - 1/15\left(2437.5 \times 1263.9\right)}{\sqrt{1120 \times 82}} = 0.943$$

$$r^2 = 0.889\,;\ \eta^2_{y/x} = \frac{\Sigma\left(\bar{y}_x - \bar{y}\right)^2}{\Sigma\left(\bar{y}_i - \bar{y}\right)} = \frac{75.0960}{82.00} = 0.914$$

Визначимо міру лінійності ($\gamma = \eta^2 - r^2$): $\gamma = 0{,}914 - 0{,}889 = 0{,}025$.

Вибіркова помилка дорівнює:

$$m_\gamma = \frac{2\sqrt{\gamma - \gamma^2\left(2 - \eta^2 - r^2\right)}}{\sqrt{n}} \text{ [71]}$$

$$m_\gamma = \frac{2\sqrt{0.025 - 0.025^2\left(2 - 0.914 - 0.889\right)}}{\sqrt{15}} = \frac{2\sqrt{0.025 - 0.025^2 \times 0.197}}{\sqrt{15}} =$$

$$\frac{2\sqrt{0.024877}}{\sqrt{15}} = \frac{2 \times 0.16}{3.87} = \frac{0.32}{3.87} = 0.087$$

Критерій $t_\gamma = \frac{\gamma}{m_\gamma}$. При $t_\gamma < 3$ (у відповідних випадках $t_\gamma < 2{,}5$) кореляція між ознаками оцінюється практично прямолінійною.

$$t_\gamma = \frac{0{,}025}{0{,}087} = 0{,}28\text{, т.е } \ t_\gamma < 1.$$

Отже, і графічно та аналітично підтверджується початкове припущення про лінійність регресії даної залежності.

ПРАКТИЧНА РОБОТА 8

Здійснити регресійний аналіз, визначивши коефіцієнти рівняння прямої, міру лінійності оцінити її достовірність.

Варіант 1

X	36,8	43,5	33,4	34,5	46,8	31	26,9	35	26,9	50
Y	52,5	62,2	43,4	41,5	64,1	56,3	47,7	55,5	35,6	56,8

Варіант 2

X	52,5	62,2	43,4	41,5	64,1	56,3	47,7	55,5	35,6	59,2
Y	30	48,3	19,6	24,5	49,4	45	31,1	37	17	45,2

Варіант 3

X	36,8	39,7	33,4	42,2	48,2	31	26,9	35	27,7	50
Y	28	26,6	31	26,9	20,9	33,2	38	29,5	35,3	20,6

Варіант 4

X	43,5	39,5	26,5	33,5	36,5	30,5	38,2	28,4	32,5	34,5
Y	13,5	18,5	23,5	18,5	18,2	22	16,7	23,5	21,5	19,2

Варіант 5

X	6	10,3	12	9	14,4	6,6	14,3	12	10,8	14,4
Y	18,9	18	17,6	18,5	17	19	17,3	18	18	17,7

Варіант 6

X	1,1	2,1	3,2	4,1	5,9	6,8	7,9	9,6	9,8	10,9
Y	0,5	1,8	1,6	2,3	2,8	3,9	3,7	4,8	4,3	4,7

Варіант 7

X	9,6	17,7	13,5	21,9	15,1	11,4	21,1	16,4	21,5	19,3
Y	19,0	18,9	17,4	19,2	19,8	10,7	20,2	19,1	22,1	20,5

Варіант 8

X	10	12	14	16	18	20	22	24	26	28
Y	14,3	16,5	18,6	19,1	19,5	20,6	20,6	20,7	21,9	22,0

Варіант 9

X	33	44	34	32	22	31	29	37	30	34
Y	20	17	17	13	13	15	17	13	22	21

Варіант 10

X	15,5	16,5	16,0	14,0	15,0	15,0	14,0	13,5	14	14,5
Y	21,5	22,0	21,0	18,0	20,5	19,0	20,0	17,0	18,6	19,5

Варіант 11

X	18,4	19,0	19,0	20,0	21,8	21,8	22,2	22,4	23,0	25,8
Y	25	20	24	23	24	24	22	28	29	26

РОЗДІЛ 10. ДИСПЕРСІЙНИЙ АНАЛІЗ

10.1. Суть методу і його основні завдання

Дисперсійний аналіз був розроблений Рональдом Ейлмером Фішером для статистичної обробки результатів агрономічних дослідів. У 1921 р. він запропонував оцінювати результати дослідів не по середнім арифметичним, а шляхом порівняння вибіркових дисперсій, їх відносин з критичним значенням критерію F, тому метод отримав назву дисперсійного аналізу. У 1925 р., після виходу книги Фішера «Статистічні методи для дослідників», метод почали застосовувати у експериментальній психології, а згодом він був поширений на інші галузі науки. У випадках комплексної оцінки результатів спостережень він виявився більш "економним" і досить ефективним в порівнянні з іншими статистичними методами.

Дисперсійний аналіз проводиться як на малих, так і на великих вибірках, на однорідному і біологічно неоднорідному матеріалі, коли в одному і тому ж комплексі об'єднуються результати спостережень, проведених на особинах різної статі, віку, видової або расової приналежності і т.д. Цей метод є одним з найбільш потужних і ефективних методів статистичного аналізу; він дозволяє вирішувати найрізноманітніші завдання, які постають перед дослідниками. Мета дисперсійного аналізу полягає у встановленні наявності впливу фактора на досліджуваний процес за рахунок вибору найбільш значущих факторів і оцінки їх впливу на досліджуваний процес.

Існує декілька класифікацій методів дисперсійного аналізу.

1) За кількістю ознак, що аналізуються виділяють:

- Однофактроний (Analysis of variance (ANOVA)), що передбачає дослідження впливу одного або кількох якісних факторів на одну залежну кількісну змінну.

- Багатофактроний (Multivariate analysis of variance (MANOVA)), що передбачає порівняння багатовимірних вибіркових середніх. Як

багатовимірна процедура, вона використовується, коли є дві або більше залежних змінних, і часто супроводжується тестами значущості, що включають окремі залежні змінні окремо.

2) За принципом аналізу:

- параметричний (для аналізу нормально розподілених ознак у групах)

- непараметиричний (для аналізу кількісної ознаки незалежно від його розподілу у групах).

3) За даними, що аналізуються:

- дані, одержані у незалежних вибірках (в т.ч. дані однократних спостережень;

- дані, одержані у залежних вибірках (в т.ч. дані повторних спостережень).

Ознаки, що змінюються під впливом тих чи інших причин, називаються результативними, причини, що діють на результативні ознаки, прийнято називати факторами. Наприклад, маса, об'єм речовини - це ознаки. Такі засоби впливу, як дози лікарських або токсичних речовин, дози каталізаторів і т.п. відносяться до категорії факторів. Зазвичай кожен із чинників представлений деякою кількістю груп, званих *градаціями.*

Вибіркова сукупність, організована певним чином для вивчення ефективності дії організованих факторів на результативну ознаку, називається *статистичним*, або *дисперсійним комплексом*. Структура такого комплексу визначається числом градацій, на які поділяються організовані фактори і враховується в досвіді ознака. Форма комплексу дається таблицею, в якій число стовпців дорівнює числу градацій одного або декількох організованих факторів, а по рядках відкладаються градації результативної ознаки, залежно від числа чинників, що враховуються, розрізняють одно- дво- і багатофакторні дисперсійні комплекси.

Оцінка достовірності впливу організованого фактора на результативну ознаку проводиться за критерієм Фішера (F), який потім оцінюється по таблиці Фішера для відповідних ступенів свободи і прийнятого рівня

значущості P = 0,05 або P = 0,01, Нульова гіпотеза, тобто, припущення про відсутність впливу організованого фактора на результативну ознаку, відкидається за умови, якщо $F > F_r$ якщо ж $F_ф < Fs_l$ спостерігаються між груповими середніми розбіжності визнаються статистично недостовірними.

9.2. Дисперсійний аналіз однофакторних комплексів малих груп

Дисперсійний аналіз однофакторних комплексів малих груп проводиться за такою схемою:

1. Знаходять середні величини, x всього комплексу і приватні чи групові середні x, по градаціях фактора A.

2. Визначають загальну суму квадратів відхилень (D_y):

$$D_y = \Sigma(x_i - \bar{x})^2 \quad [72]$$

3. Обчислюють міжгрупову суму квадратів (D):

$$D_x = n\,\Sigma(\bar{x}_i - \bar{x})^2 \quad [73]$$

при різних числах варіант в градаціях фактора:

$$D_x = \Sigma[n_i(\bar{x}_i - \bar{x})^2] \quad [74]$$

4. Знаходять внутрішньогрупову суму квадратів (D_z):

$$D_z = \Sigma\Sigma(x_i - \bar{x}_i)^2 \quad [75]$$

Розрахунки спрощуються якщо використовувати такі робочі формули:

$$D_y = \Sigma x^2 - \frac{(\Sigma x)^2}{n};\ D_x = \Sigma\frac{(\Sigma x_i)^2}{n_A} - \frac{(\Sigma x)^2}{N};\ D_z = \Sigma x^2 - \Sigma\frac{(\Sigma x_i)^2}{nA},$$

де x - варіанти, що входять до складу дисперсійного комплексу;

x_i - варіанти, що входять до складу градацій;

$\acute{x}$ - загальна середня арифметична;

$\acute{x}_A$ и $\acute{x}_i$ - групові або приватні середні арифметичні;

N - загальне число варіант;

n_A і n_i - числа варіант по градаціях і в групах комплексу,

Так як $D_y = D_x + D_z$, то $D_z = D_y - D_x$.

5. Визначивши суми квадратів відхилень, встановлюють числа ступенів свободи (k):

для загальної дисперсії $К_y = N-1$,

для груповий дисперсії $К_x = a-1$,

для внутрішньогрупової дисперсії $К_z = N-a$,

де а - число градацій фактора А.

6. Обчислюємо

загальну дисперсію $$\sigma_Y^2 = \frac{D_y}{N-1} = \frac{D_y}{K_y} \quad [76]$$

міжгрупову $$\sigma_X^2 = \frac{D_x}{a-1} = \frac{D_x}{K_x} \quad [77]$$

внутрішньогрупову: $$\sigma_Z^2 = \frac{D_z}{N-a} = \frac{D_z}{K_z} \quad [78]$$

7. Оцінимо достовірність:

$$F_\phi = \frac{\sigma_x^2}{\sigma_z^2}$$

Розглянемо застосування даного алгоритму на наступному прикладі: при вивченні впливу доз каталізатора на масу продукту реакції одержано наступні результати (табл. 9.1).

Таблиця 9.1

Дози каталізатора (%)	Маса продукту (мг) по повторюваностям						n_i	Сума	Середнє
	1	2	3	4	5	6			
15	8,0	8,4	9,0	8,6			4	34,0	8,5
20	8,2	9,0	10,0	10,2	9,2	10,0	6	56,4	9,4
25	11,0	13,0		12,0			3	36,0	12,0
30	7,5	8,5					2	16,0	8,0
Сума	-	-	-	-	-	-	15	142,4	9,5

$\bar{x} = 9.5$ мг, $\Sigma n = N = 15$, $\Sigma x = 142.4$.

При групуванні даних в таблицю 9.2 розрахунки полегшуються:

Таблиця 9.2

Показники	Градації фактора А (дози каталізватора)				Сума
	1 (15)	2(20)	3(25)	4(30)	
Маса продукту (Х)	8,0; 8,4	8,2; 9,0	11,0	7,5	a= 4
	9,0; 8,6	10,0;10,0	13,0	8,5	
		9,2; 10,0	12,0		
n_A	4	6	3	2	$\sum n_A = N=15$
$\sum x_i$	34,0	56,4	36,0	16,0	$\sum\sum x_i = \sum x= 142,4$
$(\sum x_i)^2$	1156,00	3180,96	1296,00	256,00	-
$(\sum x_i)^2/n_A$	289,00	530,16	432,00	128,00	$\sum(\sum x_i)^2/n_A=1379,16$
$\sum x_i^2$	289,52	532,88	434,00	128,50	$\sum x^2 =13804,9$

$$D_y = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 1384.9 - \frac{142.4^2}{15} = 1384.9 - 1351.85 = 33.05$$

$$D_x = \Sigma \frac{(\Sigma x_i)^2}{n_A} - \frac{(\Sigma x)^2}{N} = 1379.16 - 1351.85 = 27.31$$

$$D_z = \Sigma x^2 - \Sigma \frac{(\Sigma x_i)^2}{nA} = 1384.9 - 1379.16 = 5.74$$

K_y =15-1=14, K_x =4-1=3, K_z =14-3=11,

$$\sigma_X^2 = \frac{D_x}{K_x} = \frac{27.31}{3} = 9.1$$

$$\sigma_Z^2 = \frac{D_z}{K_z} = \frac{5.74}{11} = 0.52$$

$$F_\phi = \frac{\sigma_x^2}{\sigma_z^2} = \frac{9.1}{0.52} = 17.7$$

Визначаємо по таблиці Фішера значення F_{st}: k_x = 3 (по горизонталі), k_Z = 11 (по вертикалі), F_{st} = 6,2, $F_\phi > F_{st}$ нульова гіпотеза відкидається, статистично достовірно, що різні дози каталізатора з різною силою впливають на масу продукту ферментативної реакції.

ПРАКТИЧНА РОБОТА 9

Провести дисперсійний аналіз впливу концентрації препарату на масу продукту реакції

Варіант 1

	Маса продукту реакції (г) по повторюваностям				
	1	2	3	4	5
10%	0,6	0,6	0,8	0,7	
15%	0,9	1,1	0,8	0,9	0,7
20%	1		1	0,9	1
25%	1,2	1,4	1,1	1	1,2
33%	1,7	1,4	1,3	1,5	1,3

Варіант 2

	Маса продукту реакції (г) по повторюваностям				
	1	2	3	4	5
10%	0,2	0,2	0,2	0,3	0,2
15%	0,2		0,3	0,3	0,3
20%	0,6	0,5	0,4	0,3	
25%	0,7	0,6	0,4	0,4	0,8
35%	0,8	0,7	0,4	0,5	0,9

Варіант 3

	Маса продукту реакції (г) по повторюваностям				
	1	2	3	4	5
10%	1,8	1,6	1,2	1,4	1,9
15%	1,4	1,8	1,4	1,3	1,3
20%	1,8		1,2	1,4	1,9
25%	1,1	1,3	1,5	1,2	1,7
35%	1,2	1,3	1,5		1,3

Варіант 4

	Маса продукту реакції (г) по повторюваностям				
	1	2	3	4	5
10%	1,4	1,51	1,5	1,5	1,51
15%	1,6	1,6	1,6	0,6	1,62
20%	1,7	1,7		0,73	1,79
25%	1,8	0,8	1,8	0,83	0,85
35%	1,4	1,4	1,3		1,4

Варіант 5

	Маса продукту реакції (г) по повторюваностям				
	1	2	3	4	5
10%	6,4	6,5	6,5	6,5	6,5
15%	6,6	6,6	6,6	5,6	6,6
20%	6,7		5,7	5,7	5,8
25%	7,8	6,8	2,8	4,8	
35%	6,4	1,4	6,8	4,2	4,4

Варіант 6

	Маса продукту реакції (г) по повторюваностям				
	1	2	3	4	5
10%	1,1	0,8	1,1	0,7	0,6
15%	0,9	1,3	1,2	1,1	
20%	1,3	1,4		1,5	1,6
25%	1,5	1,5	1,4	1,7	1,9
35%	2,1	2	2,3	2,5	2,2

Варіант 7

	Маса продукту реакції (г) по повторюваностям				
	1	2	3	4	5
10%	0,2	0,1	0,6	0,2	0,6
15%	0,6	0,8	0,5	0,8	0,6
20%	1,1	0,8		0,9	
25%	1,3	1,2	1,3	1,5	1,1
35%	1,8	1,8	1,6	2,1	1,8

Варіант 8

	Маса продукту реакції (г) по повторюваностям				
	1	2	3	4	5
10%	0,67	0,67	0,91	0,87	1,0
15%	1,32	1,57	1,46	1,22	1,11
20%	1,47	1,66	1,55	1,47	1,33
25%	1,43	1,6	1,56	1,54	1,54
35%	1,77	1,77			1,6

Варіант 9

	Маса продукту реакції (г) по повторюваностям				
	1	2	3	4	5
10%	5,4	5,3	5,2	5,45	5,25
15%	5,91	5,7	5,8	5,92	5,91
20%	6,4	6,4	6,5	6,4	6,32
25%	6,6	6,52	6,67		6,5
35%	6,6		6,77	6,7	6,7

Варіант 10

	Маса продукту реакції (г) по повторюваностям				
	1	2	3	4	5
10%	7,98	8,02	8,01		7,99
15%	7,74	7,69	7,63	7,73	7,72
20%	7,55	7,59	7,48	7,42	7,53
25%	7,33	7,37	7,24	7,31	7,29
35%	6,62	6,90	6,87	6,95	

Варіант 11

	Маса продукту реакції (г) по повторюваностям				
	1	2	3	4	5
10%	7,57	7,55	7,49	7,58	7,69
15%	7,61	7,69		7,65	
20%	7,85	7,83	7,81	7,85	7,88
25%	7,95	7,96	7,99	7,89	7,93
35%	8,11	8,10	8,30	8,16	8,09

ДОДАТКИ

Таблиця 1.

Критичні точки t-критерію Стьюдента при різних рівнях значущості α

Стандартні значення критерію Стьюдента

Число ступенів волі $u=n_1+n_2-2$	Критерій Стьюдента t_{st} при імовірності безпомилкового заключення p			
	0.1	0.05	0.02	0.01
1	6.314	12.706	31.821	63.657
2	2.920	4.303	6.965	9.952
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499
8	1.860	2.306	2.896	3.355
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.169
11	1.796	2.201	2.718	3.106
12	1.782	2.179	2.684	3.055
13	1.771	2.160	2.650	3.012
14	1.761	2.145	2.624	2.977
15	1.732	2.131	2.602	2.947
16	1.746	2.120	2.583	2.921
17	1.740	2.110	2.567	2.898
18	1.734	2.101	2.552	2.878
19	1.729	2.093	2.539	2.861
20	1.723	2.086	2.528	2.845
21	1.721	2.080	2.518	2.831
22	1.717	2.074	2.508	2.819
23	1.714	2.069	2.500	2.807
24	1.714	2.064	2.492	2.797
25	1.708	2.060	2.485	2.787
26	1.706	2.056	2.479	2.779
27	1.703	2.052	2.473	2.771
28	1.701	2.048	2.467	2.763
29	1.699	2.045	2.462	2.756
30	1.697	2.042	2.457	2.750
∞	1.645	1.960	2.326	2.576

Таблиця 2.

Критичні значення величини нормованого відхилення при оцінці сумнівних варіант з урахуванням обсягу вибірки *n* і рівнів значущості α

n	α,%		*n*	α,%		*n*	α,%	
	5	1		5	1		5	1
4	1.71	1,73	13	2,56	2,81	23	2,84	3.16
5	1,92	1,97	14	2,60	2,86	24	2,86	3,18
6	2.07	2.16	15	2,64	2.90	25	2,88	3,20
7	2,18	2,31	16	2,67	2 95	26	2,90	3,22
8	2.27	2,43	17	2,70	2,98	27	2.91	3,24
9	2.35	2,53	18	2,73	3,02	28	2,93	3,26
10	2.41	2,62	19	2,75	3,05	29	2,94	3,28
11	2.47	2,69	20	2,78	3,08	30	2,96	3.29
12.	2,52	2,75	21	2,80	3.11			
P	0.05	0,01	–	0,05	0.01		0,05	0,01

Таблиця 3.

Значення F-критерію Фішера при рівнях значимості α = 5% (верхній рядок) і α = 1% (нижній рядок)

K_2	Число степенів свободи варіації для більшої дісперсії (K_1)									
	1	3	5	7	9	11	14	20	30	50
1	161	216	230	237	241	243	245	248	250	252
	4052	5403	5764	5928	6022	6082	6142	6208	6258	6302
3	10.3	9.28	9.01	8.88	8.81	8.76	8.71	8.66	8.62	8.58
	34.12	29.46	28.24	27.67	27.34	27.13	26.92	26.69	26.50	26.35
5	6.61	5.41	5.05	4.88	4.78	4.70	4.64	4.56	4.50	4.44
	16.26	12.06	10.97	10.45	10.15	9.96	9.77	9.55	9.38	9.24
7	5.59	4.35	3.97	3.79	3.68	3.60	3.52	3.44	3.38	3.32
	12.25	8.45	7.46	7.00	6.71	6.54	6.35	6.15	5.98	5.85
9	5.12	3.86	3.48	3.29	3.18	3.10	3.02	2.93	2.86	2.80
	10.56	6.99	6.06	5.62	5.35	5.18	5.00	4.80	4.64	4.51
11	4.84	3.59	3.20	3.01	2.90	2.82	2.74	2.65	2.57	2.50
	9.05	6.22	5.32	4.88	4.63	4.46	4.29	4.10	3.94	3.80
13	4.67	3.41	3.02	2.84	2.72	2.63	2.55	2.46	2.38	2.32
	9.07	5.74	4.86	4.44	4.19	4.02	3.85	3.67	3.51	3.37
15	4.54	3.29	2.90	2.70	2.59	2.51	2.43	2.33	2.25	2.18
	8.68	5.42	4.56	4.14	3.89	3.73	3.56	3.36	3.20	3.07
17	4.45	3.20	2.81	2.62	2.50	2.41	2.33	2.23	2.15	2.08
	8.40	5.18	4.34	3.93	3.68	3.52	3.35	3.16	3.00	2.89
19	4.38	3.13	2.74	2.55	2.43	2.34	2.26	2.15	2.07	2.00
	8.18	5.01	4.17	3.77	3.52	3.36	3.19	3.00	2.84	2.70
21	4.32	3.07	2.68	2.49	2.37	2.28	2.20	2.09	2.00	1.93
	8.02	4.87	4.04	3.65	3.40	3.24	3.07	2.88	2.72	2.58
23	4.28	3.03	2.64	2.45	2.32	2.24	2.14	2.04	1.96	1.88
	7.88	4.46	3.94	3.54	3.30	3.14	2.97	2.78	2.62	2.48
25	4.24	2.99	2.60	2.41	2.28	2.20	2.11	2.00	1.92	1.84
	7.77	4.68	3.86	3.46	3.21	3.05	2.89	2.70	2.54	2.40
27	4.21	2.96	2.57	2.37	2.25	2.16	2.08	1.97	1.88	1.80
	7.68	4.60	3.79	3.39	3.14	2.98	2.83	2.63	2.45	2.33
29	4.18	2.93	2.54	2.35	2.22	2.14	2.05	1.94	1.85	1.77
	7.60	4.54	3.73	3.33	3.08	2.92	2.77	2.57	2.41	2.27

Критерій Фішера *F*

Рівень значимості $\alpha = 0{,}01$								
$f_2 \backslash f_1$	4	7	10	16	24	40	100	∞
1	5625,0	5928,0	6056,0	6169,0	6234,0	6286,0	6334,0	6366,0
2	99,25	99,34	99,40	99,44	99,46	99,48	99,49	99,50
3	28,71	27,67	27,23	26,83	26,60	26,41	26,23	26,12
4	15,98	14,98	14,54	14,15	13,93	13,74	13,57	13,46
5	11,39	10,45	10,05	9,68	9,47	9,29	9,13	9,02
6	9,15	8,26	7,87	7,52	7,31	7,14	6,99	6,88
7	7,85	7,00	6,62	6,27	6,07	5,90	5,75	5,65
8	7,01	6,19	5,82	5,48	5,28	5,11	4,96	4,86
9	6,42	5,62	5,26	4,92	4,73	4,56	4,41	4,31
10	5,99	5,21	4,85	4,52	4,33	4,17	4,01	3,91
12	5,41	4,65	4,30	3,98	3,78	3,61	3,46	3,36
14	5,03	4,28	3,94	3,62	3,43	3,26	3,11	3,00
16	4,77	4,03	3,69	3,37	3,18	3,01	2,86	2,75
18	4,58	3,85	3,51	3,19	3,00	2,83	2,68	2,57
Рівень значимості $\alpha = 0{,}05$								
$f_2 \backslash f_1$	4	7	10	16	24	40	100	∞
1	225,0	237,0	242,0	246,0	249,0	251,0	253,0	254,0
2	19,25	19,36	19,39	19,43	19,45	19,47	19,49	19,50
3	9,12	8,88	8,78	8,69	8,64	8,60	8,56	8,53
4	6,39	6,09	5,96	5,84	5,77	5,71	5,66	5,63
5	5,19	4,88	4,74	4,60	4,53	4,46	4,40	4,36
6	4,53	4,21	4,06	3,92	3,84	3,77	3,71	3,67
7	4,12	3,79	3,63	3,49	3,41	3,34	3,28	3,23
8	3,84	3,50	3,34	3,20	3,12	3,05	2,98	2,93
9	3,63	3,29	3,13	2,98	2,90	2,82	2,76	2,71
10	3,48	3,14	2,97	2,82	2,74	2,67	2,59	2,54
12	3,26	2,92	2,76	2,60	2,50	2,42	2,35	2,30
14	3,11	2,77	2,60	2,44	2,35	2,27	2,19	2,13
16	3,01	2,66	2,49	2,33	2,24	2,16	2,07	2,01
18	2,93	2,58	2,41	2,25	2,15	2,07	1,98	1,92

Примітка: f_1 – відноситься до більшої дисперсії. f_2 – до меншої.

Таблиця 4

χ^2-Розподіл. Критичні (процентні) точки для різних значень ймовірності P і числа ступенів свободи

Стандартні значення критерію достовірності (хі-квадрат)

ν	Рівні ймовірності P			ν	Рівні ймовірності P		
	0,95	0,99	0,999		0,95	0,99	0,999
1	3,8	6,6	10,8	26	38,9	45,6	54,1
2	6,0	9,2	13,8	27	40,1	47,0	55,5
3	7,8	11,3	16,3	28	41,3	48,3	56,9
4	9,5	13,3	18,5	29	42,6	49,6	58,3
5	11,1	15,1	20,5	30	43,8	50,9	59,7
6	12,6	16,8	22,5	32	46,2	53,5	62,4
7	14,1	18,5	24,3	34	48,6	56,0	65,2
8	15,5	20,1	26,1	36	51,0	58,6	67,9
9	16,9	21,7	27,9	38	53,4	61,6	70,7
10	18,3	23,2	29,6	40	55,8	63,7	73,4
11	19,7	24,7	31,3	42	58,1	66,2	76,1
12	21,0	26,2	32,9	44	60,5	68,7	78,7
13	22,4	27,7	34,5	46	62,8	71,2	81,4
14	23,7	29,1	36,1	48	65,2	73,7	84,0
15	25,0	30,6	37,7	50	67,5	76,2	86,7
16	26,3	32,0	39,3	55	73,3	82,3	93,2
17	27,6	33,4	40,8	60	79,1	88,4	99,6
18	28,9	34,8	42,3	65	84,8	94,4	106,0
19	30,1	36,3	43,8	70	90,5	100,4	112,3
20	31,4	37,6	45,3	75	96,2	106,4	118,5
21	32,7	38,9	46,8	80	101,9	112,3	124,8
22	33,9	40,3	48,3	85	107,5	118,2	131,0
23	35,2	41,6	49,7	90	113,1	124,1	137,1
24	36,4	43,0	51,2	95	118,7	130,0	143,3
25	37,7	44,3	52,6	100	124,3	135,8	149,4

Таблиця 5.

Значення Z, що відповідають значенням вибіркового коефіцієнта кореляції $r_{xy.}$

R	Соті долі коефіцієнта кореляції *R*									
	0	1	2	3	4	5	6	7	8	9
0,0	0,000	0,010	0,020	0,030	0,040	0,050	0,060	0,070	0,080	0,090
0,1	0,100	0,111	0,121	0,131	0,141	0,151	0,161	0,172	0,182	0,192
0,2	0,203	0,213	0,224	0,234	0,245	0,255	0,266	0,277	0,288	0,299
0,3	0,310	0,321	0,332	0,343	0,354	0,365	0,373	0,388	0,400	0,412
0,4	0,424	0,436	0,448	0,460	0,472	0,485	0,498	0,510	0,523	0,536
0,5	0,549	0,563	0,576	0,590	0,604	0,618	0,633	0,648	0,663	0,678
0,6	0,693	0,709	0,725	0,741	0,758	0,776	0,793	0,811	0,829	0.848
0,7	0,867	0,887	0,908	0,929	0,951	0,973	0,996	1,020	1,045	1,071
0,8	1,099	1,127	1,157	1,188	1,221	1,256	1,293	2,092	1,376	1,422
0,9	1,472	1,528	1,589	1,658	1,738	1,832	1,946	2,092	2,298	2,647
0,99	2,647	2,700	2,759	2,826	2,903	2,995	3,106	3,250	3,453	3,800

Таблиця 6.

Критичні значення коефіцієнту асиметрії As

Об'ем вибірки, n	Рівні значущості α, %		Об'ем вибірки, n	Рівні значущості α, %	
	5	1		5	1
25	0,711	1,061	250	0,251	0,360
30	0,661	0,982	300	0,230	0,329
35	0,621	0,921	350	0,213	0,305
40	0,587	0,869	400	0,200	0,285
45	0,558	0,825	450	0,188	0,269
50	0,533	0,787	500	0,179	0,255
60	0,492	0,723	550	0,171	0,243
70	0,459	0,673	600	0,163	0,233
80	0,432	0,631	650	0,157	0,224
90	0,409	0,596	700	0,151	0,215
100	0,389	0,567	750	0,146	0,208
125	0,350	0,508	800	0,142	0,202
150	0,321	0,464	850	0,138	0,196
175	0,298	0,430	900	0,134	0,190
200	0,280	0,403	950	0,130	0,185
P	0,05	0,01		0,05	0,01

Таблиця 7

Критичні значення коефіцієнту ексцесу Ex

Об'ем вибірки, n	**Рівні значущості, α, %**		
	10	**5**	**1**
11	0,890	0,907	0,936
16	0,873	0,888	0,914
21	0,863	0,877	0,900
26	0,857	0,869	0,890
31	0,851	0,863	0,883
36	0,847	0,858	0,877
41	0,844	0,854	0,872
46	0,841	0,851	0,868
51	0,839	0,848	0,865
61	0,835	0,843	0,859
71	0,832	0,840	0,855
81	0,830	0,838	0,852
91	0,828	0,835	0,848
101	0,826	0,834	0,846
201	0,818	0,823	0,832
301	0,814	0,818	0,826
401	0,812	0,816	0,822
501	0,810	0,814	0,820
P	***0,10***	***0,05***	***0,01***

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Massart D.L. Chemometrics. NY: Elsevier, 1988. 464 p.

2. 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.

3. Рудавський Ю.К., Мокрий Є.М., Піх З.Г., Чип М.М., Куриляк І.Й. Математичні методи і хімії та хімічній технології / За ред. Рудавського Ю.К. Львів: Світ, 1993. 208 с.

4. Іщенко О.В., Михальчук В.М., Біла Н.І., Гайдай С.В., Білий О.В. Статистичні методи у хімії. Підручник для студентів хімічних спеціальностей вищих навчальних закладів. Донецьк: Видавництво ДонНУ, 2012. 504 с.

5. Статистичні та хемометричні методи в хімії: навчальний посібник / А.В. Пантелеймонов, І.В. Христенко, В.В. Іванов та ін. Х.: ХНУ імені В.Н. Каразіна, 2012. 40 с.

6. Чмиленко, Ф.О., Смітюк Н.М. Навчальний посібник з основ статистичної обробки аналітичного експерименту. Д.: РВВДНУ, 2013. 60 с.

7. Статистика: підручник / С.С. Герасименко, А.В. Головач, А.М.Єріна та ін.; за ред. С.С. Герасименка. К.: КНЕУ, 2000. 467 с.

8. Лапач С.Н., Чубенко А.В., Бабич П.Н. Статистичні методи в медико-біологічних дослідженнях з використанням EXCEL. К.: Моріон, 2001. 408 с.

9. Руденко В.М. Математична статистика. К.: Центр учбової літератури, 2012. 304 с.

10. 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.