МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ЗАКАРПАТСЬКИЙ УГОРСЬКИЙ ІНСТИТУТ ІМЕНІ ФЕРЕНЦА РАКОЦІ ІІ

КАФЕДРА МТЕМАТИКИ ТА ІНФОРМАТИКИ

Методичні вказівки

для практичних занять

з дисципліни

Статистичні основи наукових досліджень

Берегове

2021

УДК 372.8.519.2(=161.2)(439.23)

Рекомендовано до друку Вчиною радою ЗУІ ім. Ф. Ракоці ІІ

(протокол №8 від 27. серпня 2021 р.)

Рецензент:

Стойка М. В. – кандидат фіз. мат. наук, доцент, доцент кафедри математики та інформатики Закарпатського угорського інститут ім. Ф. Ракоці ІІ, м. Берегове

Методичні вказівки до практичних занять з дисципліни «Статистичні основи наукових досліджень» / укл. Кучінка К.Й., Кулін Ю.І., Роман І.Т. Беревоге: ЗУІ ім. Ф. Ракоці ІІ, 2021. –102 с.

Навчально-методичний посібник охоплює всі змістовні модулі, передбачені навчальною програмою дисципліни «Статистичні основи наукових досліджень». Містить методичні рекомендації щодо вивчення теми, методичні рекомендації щодо вирішення задач за темами, вправи для самостійної роботи. Цей посібник рекомендовано слухачам вибіркової дисципліни «Статистичні основи наукових досліджень», та всім, хто цікавиться статистичними дослідженнями.

ЗМІСТ

Програма та структура навчальної дисципліни
A tantárgy programja és felépítése

модуль 1.

Тема 1. Типи наукових досліджень. Класифікація ознак за шкалами вимірювання.
Kutatás típusai. Mérési skálák.

Тема 2. Отримання описових статистик та побудова гістограм.
Leíró statisztikák. Grafikus ábrázolás

Тема 3. Перевірка статистичних гіпотез.
Hipotézis vizsgálat.

модуль 2.

Тема 4. Дисперсійний аналіз
Varianciaanalízis

Тема 5. Кореляційний аналіз кількісних ознак.
Korreláció- és regresszió-analízis

Тема 6. Приклади здійснення кореляційного аналізу
Korreláció- és regresszió-analízis a gyakorlatban

Тема 7. Кластерний аналіз
Klaszteranalízis

План практичних занять з дисципліни

A gyakorlati foglalkozások vázlatai

№ з/п	Назва теми	Кількість годин
1	Шкали вимірювань. Отримання описових статистик та побудова гістограм. Mérési skálák. Leíró statisztikák. Grafikus ábrázolás	4
2	Первірка статистичних гіпотез. Hipotézis vizsgálat a gyakorlatban.	4
3	Дисперсійний аналіз. Varianciaanalízis	2
4	Приклади здійснення кореляційного та регресійного аналізу аналізу. Korreláció- és regresszió-analízis	2
5	Кластерний аналіз. Klaszteranalízis	2
	Всього	

Зміст практичних занять

1. Вибірковий метод (математичні показники вибірки). Класифікація ознак за шкалами вимірювання. Описова статистика

Математична статистика – це розділ прикладної математики, предметом якого є розробка раціональних прийомів і методів отримання, опису та обробки експериментальних даних з метою вивчення закономірностей масових випадкових явищ.

Основними завданнями математичної статистики є:

– визначення за статистичними даними законів розподілу випадкових величин;
– визначення за статистичними даними параметрів розподілу випадкових величин;
– визначення за статистичними даними виду зв'язку між різними явищами (об'єктами) або властивостями одного і того ж явища (об'єкту);
– визначення сили (тісноти зв'язку) між різними явищами (об'єктами) або властивостями одного і того ж явища (об'єкту);
– перевірка вірогідності статистичних гіпотез;
– розробка рекомендацій щодо проведення експерименту та обробки його результатів.

У прикладних дослідженнях зазвичай необхідно вивчити сукупність однорідних об'єктів або спостережень за якою-небудь кількісною або якісною ознакою.

Сукупність об'єктів або спостережень, всі елементи якої підлягають вивченню при статистичному аналізі, називається **генеральною сукупністю.**

Генеральна сукупність може бути скінченою або нескінченною. Так, при вивченні розподілу населення за родом занять, розглядається велика, але скінчена генеральна сукупність об'єктів. При вивченні впливу яскравості освітлення робочого місця на продуктивність праці працівника генеральна сукупність спостережень теоретично нескінченна, оскільки яскравість освітлення може змінятся безперервно у межах певного інтервалу.

Число об'єктів (спостережень) генеральної сукупності називається її **об'ємом** і позначається ***N***.

На практиці рідко є можливість досліджувати кожен елемент генеральної сукупності, оскільки це зв'язано з великими витратами засобів і часу, а іноді з псуванням або знищенням досліджуваних об'єктів. У деяких випадках дослідити всі об’єкти генеральної сукупності взагалі неможливо. Тому при статистичному аналізі, як правило, вивчається не вся генеральна сукупність, а деяка її частина.

Частина об'єктів генеральної сукупності, використовувана в ході дослідження, називається **вибіркою**. Число об'єктів (спостережень) вибірки називається її **об'ємом** і позначається ***n***.

Наприклад, продукція у кількості *N* одиниць, вироблена підприємством на протязі року, є генеральною сукупністю. Для дослідження якості продукції на практиці розглядається вибірка, що складається з *n* одиниць продукції. Ознакою якості в даному дослідженні служить відповідність вибраної одиниці товару сертифікатним вимогам.

Суть вибіркового методу в статистиці полягає в тому, що висновки, зроблені на основі вивчення вибірки, розповсюджуються на всю генеральну сукупність.

Слід зазначити, що незалежно від способу організації вибірки вона повинна правильно відображати кількісні співвідношення генеральної сукупності, тобто бути **репрезентативною**. Крім того, всі елементи генеральної сукупності повинні мати однакову ймовірність бути відібраними у вибірку, тобто вибірка повинна бути **випадковою**. Для результатів, що отримані при вибірковому дослідженні, необхідна перевірка на точність і статистичну значущість; спосіб формування вибірки та її об’єм повинні відповідати певному методу обробки даних.

Числові характеристики статистичних рядів

Поняття про оцінки параметрів

Генеральну сукупність X можна розглядати як випадкову величину. Тоді вибірка значень X – це емпіричний закон розподілу випадкової величини. Для дискретних і неперервних випадкових величин визначені числові характеристики, основними з яких є математичне сподівання, дисперсія і середнє квадратичне відхилення. Числові характеристики випадкових величин часто є параметрами їх розподілів. Аналогічно числові характеристики визначені і для статистичних рядів, це – вибіркове середнє, вибіркове середнє геометричне, вибіркова дисперсія, вибіркове середнє квадратичне відхилення і т. ін.

У прикладних задачах часто необхідно визначити за даними вибірки закон розподілу випадкової величини, що є однією із основних задач математичної статистики. При цьому вибіркове середнє вважається **оцінкою** (аналогом) математичного сподівання, вибіркова дисперсія – оцінкою дисперсії, вибіркове середнє квадратичне відхилення – оцінкою середнього квадратичного відхилення. При цьому виникає питання: наскільки правомірні такі оцінки?

Оцінки параметрів повинні відповідати таким вимогам.

Незсуненість. Це позначає, що при проведенні великої кількості спостережень (вимірювань) з вибірками одного об’єму оцінка параметру, отримана з кожної вибірки, прямує до істинного значення цього параметру генеральної сукупності.

Спроможність. Зі збільшенням об’єму вибірки оцінка прямує до значення відповідного параметру генеральної сукупності з ймовірністю, що дорівнює 1.

Достатність. Оцінка містить всю необхідну інформацію.

Ефективність. Оцінки, отримані за вибірками однакового об’єму, мають мінімальну дисперсію.

Зауваження. При використанні оцінок необхідно пам’ятати, що вони отримуються тільки при певних передмовах і, відповідно, дійсні тільки при виконанні цих передмов.

Для оцінювання параметрів розподілу за даними вибірки зазвичай

використовується метод максимальної правдоподібності. Але він застосовується тільки тоді, коли відомий закон розподілу.

Числові характеристики положення

Основною числовою характеристикою статистичного ряду є середня арифметична, звана також вибірковим середнім.

Вибірковим середнім називається величина $\overline{x}$ яка обчислюється за формулою:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i = \sum_{i=1}^{k} x_i w_i \,. \qquad (1.6)$$

У разі інтервального статистичного ряду як x_i вибирається середина *i*-го інтервалу.

Якщо вибірка містить незгруповані дані, то вибіркове середнє розраховується за формулою:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1.7)$$

Зауваження. Оскільки статистичний ряд є емпіричним законом розподілу величини *X*, то вибіркове середнє зазвичай вважається аналогом або оцінкою математичного сподівання випадкової величини *X*. Хоча це твердження безумовно вірне тільки для нормального закону розподілу.

Вибірковим середнім геометричним називається величина $\overline{x_G}$, яка обчислюється за формулою:

$$\overline{x_G} = \sqrt[n]{\prod_{i=1}^{n} x_i} \,. \qquad (1.9)$$

Середнє геометричне застосовується як центральна тенденція тоді, коли значення *X* змінюються с постійним співвідношенням між попереднім і наступним значеннями, тобто якщо $\frac{x_i}{x_{i-1}} = \frac{x_{i+1}}{x_i}$ (наприклад, збільшення

капіталовкладень, експлуатаційні витрати і т. ін.).

Модою *Мо* називається таке значення величини *X*, яке спостерігається у вибірці з найбільшою частотою. У випадку інтервального статистичного ряду мода розраховується за формулою:

$$Mo = x_{Mo} + \frac{\Delta\left(n_{Mo} - n_{Mo-1}\right)}{2n_{Mo} - n_{Mo-1} - n_{Mo+1}}, \qquad (1.10)$$

де x_{Mo} – початок інтервалу, якому відповідає найбільша частота (такий інтервал називається модальним);

n_{Mo} – частота у модальному інтервалі;

n_{Mo-1} , n_{Mo+1} – частоти в попередньому і наступному інтервалах відповідно.

Зауваження. Мода не застосовується тоді, коли гістограма або полігон частот показують наявність двох або більше вершин („піків”).

Медіаною *Ме* називається таке значення величини *X*, яке розділяє вибірку, елементи якої розташовані у порядку зростання, на дві рівні за об’ємом частини.

Якщо це вибірка значень дискретної випадкової величини, то медіаною є те її значення, яке розташовано всередині, якщо кількість членів ряду непарна: тобто це елемент з номером $\frac{n+1}{2}$. Якщо кількість елементів вибірки парна, то медіана дорівнює середньому арифметичному її членів з номерами $\frac{n}{2}$ та $\frac{n}{2}+1$.

Якщо розглядається вибірка неперервної випадкової величини, то медіана розраховується формулою:

$$Me = X_{Me} + \frac{\Delta\left(\frac{n}{2} - n_x^{\max}\right)}{n_m}, \qquad (1.11)$$

де X_{Me} – фактична нижня границя медіанного інтервалу;

$n_x^{\max}$ – сума частот, що накопичена до початку медіанного інтервалу;

n_m – частота в медіанному інтервалі.

Зауваження. На значення медіани не впливають змінення значень крайніх елементів впорядкованої вибірки, тому її часто застосовують як центральну тенденцію тоді, коли крайні елементи вибірки значно відрізняються від інших її елементів.

Числові характеристики розсіювання

Варіаційним розмахом *R* називається різниця між максимальним і мінімальним елементом вибірки:

$$R = x_{\max} - x_{\min}. \tag{1.12}$$

Вибірковою дисперсією S^2 називається середня арифметична квадратів відхилень варіант від їх вибіркової середньої:

$$S^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i = \sum_{i=1}^{k}(x_i - \bar{x})^2 w_i \tag{1.13}$$

або $$S^2 = \frac{1}{n-1}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i. \tag{1.14}$$

Дисперсія є показником розсіювання елементів вибірки відносно їх середнього значення. Вибіркова дисперсія, отримана за формулою (1.14), називається незсуненою оцінкою дисперсії генеральної сукупності.

Різниця дисперсій, отриманих за формулами (1.13) та (1.14) зазвичай невелика, однак може вплинути на точність оцінок. Тому, якщо відомо точне значення математичного сподівання, використовують формулу (1.13), в іншому випадку – формулу (1.14).

Якщо дані не згруповані, то дисперсію можна розрахувати за формулою:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{1.15}$$

Вибірковим середнім квадратичним відхиленням S називається

величина, що дорівнює кореню з вибіркової дисперсії:

$$S = \sqrt{S^2}\,. \qquad (1.16)$$

Вибіркове середнє квадратичне відхилення теж є показником розсіювання елементів вибірки відносно їх середнього значення, але, на відміну від дисперсії, воно має ті одиниці вимірювання, які мають елементи вибірки.

Коефіцієнтом варіації *v* називається величина, що дорівнює процентному відношенню вибіркового середнього квадратичного відхилення до модуля вибіркової середньої:

$$v = \frac{S}{|\bar{x}|} \cdot 100\% \ (\bar{x} \neq 0)\,. \qquad (1.17)$$

Якщо коефіцієнт варіації більший за 100%, то елементи вибірки неоднорідні і вона не може бути використана у подальших дослідженнях.

ПРИКЛАД За даними вибіркового дослідження відомі ціни x_i певного товару у різних торгівельних організаціях (табл.) Знайти всі можливі числові характеристики за даними таблиці.

Таблиця

Організація	1	2	3	4	5	6	7	8
Ціна	100	110	115	125	140	145	145	150

Розв'язок. За незгрупованими даними таблиці 1.9 можна знайти: вибіркове середнє за формулою (1.7), медіану, розмах варіації за формулою (1.12), дисперсію за формулою (1.15), вибіркове середнє квадратичне відхилення за формулою (1.16), коефіцієнт варіації за формулою (1.17). Кількість елементів вибірки $n = 8$. Вибіркове середнє $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{100+110+115+125+140+145+145+150}{8} = 128,75$.

Кількість елементів вибірки парна, тому медіана дорівнює

середньому арифметичному її членів з номерами $\frac{n}{2}$ та $\frac{n}{2}+1$: $x_{n/2}=x_4=125;$

$x_{n/2+1}=x_5=140;\ Me=\frac{x_4+x_5}{2}=\frac{125+140}{2}=132,5$.

Розмах варіації $R=x_{\max}-x_{\min}=150-100=50$.

Для розрахунку вибіркової дисперсії складемо таблицю 1.10.

Таблиця

x_i	100	110	115	125	140	145	145	150
$x_i-\bar{x}$	-28,75	-18,75	-13,75	-3,75	11,25	16,25	16,25	21,25
$\left(x_i-\bar{x}\right)^2$	826,563	351,56	189,06	14,063	126,56	264,06	264,06	451,56

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

$\frac{826,563+351,56+189,06+14,063+126,56+264,06+264,06+451,56}{8}$=310,938.

Вибіркове середнє квадратичне відхилення $S=\sqrt{S^2}$ $=\sqrt{310,938}\approx 17,63$.

Коефіцієнт варіації $v=\frac{S}{|\bar{x}|}\cdot 100\%=\frac{17,63}{128,75}\cdot 100\%\approx 13,69\%$.

ПРИКЛАД За даними вибіркового дослідження відомі ціни x_i певного товару у різних торгівельних організаціях (табл. 1.11). Визначити центральну тенденцію за даними таблиці.

Таблиця 1.11

Організація	1	2	3	4	5	6	7	8	9

Ціна	100	110	115	125	140	145	145	150	450

Розв'язок. Оскільки дані таблиці 1.11 незгруповані, то оцінкою центральної тенденції може слугувати вибіркове середнє або медіана.

Знайдемо вибіркове середнє за формулою (1.7):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{100+110+115+125+140+145+145+150+450}{9} \approx 164,4 .$$

Кількість елементів вибірки непарна, тому медіана дорівнює її члену з номером $\frac{n+1}{2}$: $Me = x_{n+1/2} = x_5 = 140$.

Зобразимо дані таблиці графічно, вкажемо на діаграмі положення середнього і медіани (рис. 1.8).

На рис. 1.8 видно, що як центральну тенденцію вибірки слід взяти медіану.

Зауваження. Цей приклад показує, що у випадках наявності у вибірці даних, які сильно відрізняються один від одного, або даних, які сильно відрізняються від всіх останніх (так званих викидів), медіана є більш усталеною оцінкою центральної тенденції, ніж вибіркове середнє.

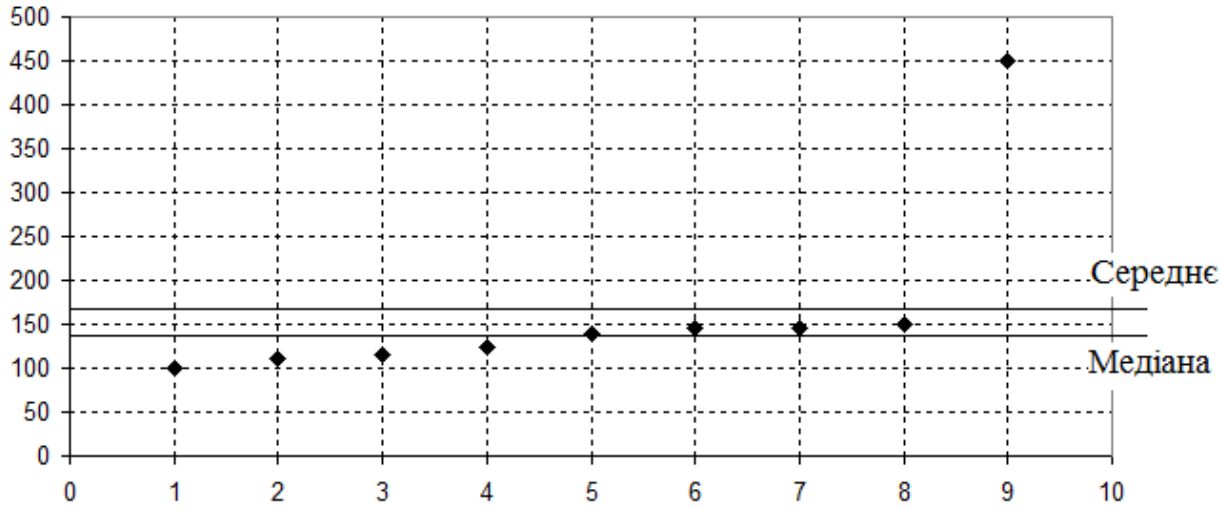


Рис. 1.8. Положення середнього і медіани відносно вибіркових даних

ПРИКЛАД За даними вибіркового дослідження відома кількість людей, що відвідували лікарню протягом року. Дані згруповані залежно від віку відвідувачів (табл. 1.12). Знайти всі можливі числові характеристики за даними таблиці.

Таблиця 1.12

Вік	20-29	30-39	40-49	50-59	60-69
Кількість відвідувань	45	36	175	361	825

Розв'язок. Позначимо X – вік відвідувачів лікарні, n – загальна кількість відвідувань, n = 1442; n_i – кількість відвідувань залежно від віку; k – кількість досліджуваних вікових груп, k = 5. Тоді відповідно даним таблиці 1.11 отримаємо інтервальний статистичний ряд (табл. 1.13).

Таблиця 1.13

$[a_i; a_{i+1}]$	20-29	30-39	40-49	50-59	60-69
n_i	45	36	175	361	825

Розрахуємо вибіркове середнє за формулою (1.6). Для зручності розрахунки оформимо у вигляді таблиці (табл. 1.14).

Таблиця 1.14

$[a_i; a_{i+1}]$	20-29	30-39	40-49	50-59	60-69	Суми
x_i	24,5	34,5	44,5	54,5	64,5	
n_i	45	36	175	361	825	1442
$x_i n_i$	1102,5	1242	7787,5	19675	53213	83019

Тоді $\overline{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i = \frac{83019}{1442} \approx 57,57$.

Моду обчислимо за формулою (1.10) враховуючи, що:

x_{Mo} – початок модального інтервалу (якому відповідає найбільша частота), x_{Mo} =60;

n_{Mo} – частота у модальному інтервалі, n_{Mo} = 825;

n_{Mo-1} , n_{Mo+1} – частоти в попередньому і наступному інтервалах відповідно, n_{Mo-1} = 361, n_{Mo+1} = 0 (оскільки модальний інтервал є останнім);

Δ - довжина інтервалу, Δ = 69-59 = 9.

Отже, $Mo = x_{Mo} + \frac{\Delta\left(n_{Mo} - n_{Mo-1}\right)}{2n_{Mo} - n_{Mo-1} - n_{Mo+1}} = 60 + \frac{10(825 - 361)}{2 \cdot 825 - 365 - 0} = 65.$

Медіану обчислимо за формулою (1.11), враховуючи, що:

X_{Me} – фактична нижня границя медіанного інтервалу, X_{Me} = 60 (оскільки всього даних 1442, то їх половина 1442:2=721; у статистичному ряді до початку останнього інтервалу міститься 617 даних, тому медіана повинна знаходитися в останньому інтервалі);

$n_x^{\max}$ – сума частот, що накопичена до початку медіанного інтервалу, $n_x^{\max}$ =617;

n_m – частота в медіанному інтервалі, n_m =825.

Отже, $Me = X_{Me} + \frac{\Delta\left(\frac{n}{2} - n_x^{\max}\right)}{n_m} = 60 + \frac{10\left(\frac{1442}{2} - 617\right)}{825} = 61,26.$

Дисперсію обчислимо за формулою (1.14). Для зручності розрахунки оформимо у вигляді таблиці (табл. 1.15).

Таблиця 1.15

x_i	24,5	34,5	44,5	54,5	64,5	Суми
n_i	45	36	175	361	825	1442
$\left(x_i - \bar{x}\right)^2$	1093,6249	1190,25	1980,25	2970,25	4160,25	11394,62

Тоді $S^2 = \frac{1}{n-1}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i = \frac{11394,62}{1442-1} = 7,9.$

Вибіркове середнє квадратичне відхилення за формулою (1.16)

дорівнює: $S = \sqrt{S^2}$ = 2,81.

Коефіцієнт варіації за формулою (1.17) дорівнює: $v = \frac{S}{|\bar{x}|} \cdot 100\% = \frac{2,81}{57,57} \cdot 100\% \approx 4,89\%$.

Довірчі інтервали і довірча ймовірність

Однією з основних задач математичної статистики є оцінка числових характеристик (параметрів) генеральної сукупності за вибірковими даними.

Для вибірки можна обчислити такі числові характеристики, як: вибіркове середнє, мода, медіана, вибіркова дисперсія та вибіркове середнє квадратичне відхилення. Для генеральної сукупності часто визначаються не самі ці параметри, а довірчі інтервали.

Довірчим інтервалом для певного параметру генеральної сукупності називається такий числовий інтервал, в межах якого знаходиться цей параметр. Ймовірність, з якою довірчий інтервал захватить істинне значення параметру, називається **довірчою ймовірністю** або **рівнем надійності** і позначається γ.

Значення довірчої ймовірності обирає дослідник залежно від того, яку ступінь точності розрахунків вимагає дослідження. Зазвичай це значення знаходиться в інтервалі від 0,9 до 0,999. Якщо вимоги точності дуже високі, то для довірчої ймовірності обирається значення 0,999; якщо підвищені – 0,99; звичайні – 0,95; знижені – 0,9.

Довірчі інтервали розраховуються з урахуванням певних вимог до генеральної сукупності. Зазвичай це вимога нормального розподілу її даних.

Довірчий інтервал для генерального середнього при відомій генеральної дисперсії

Нехай *X* – генеральна сукупність, що підкоряється нормальному закону розподілу; σ^2 – відома генеральна дисперсія; $\{x_1, x_2, ..., x_n\}$ – вибірка

з генеральної сукупності об'єму *n*; $\bar{x}$ - вибіркове середнє. Потрібно знайти довірчий інтервал для генерального середнього *a* із заданим рівнем надійності γ.

Шуканий довірчий інтервал знаходиться за формулою:

$$\bar{x} - z_{\frac{1-\gamma}{2}} \cdot \frac{\sigma}{\sqrt{n}} < a < \bar{x} + z_{\frac{1-\gamma}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \qquad (1.18)$$

де значення $z_{\frac{1-\gamma}{2}}$ знаходиться з таблиці (табл. 1.16) або за допомогою вбудованої функції Excel НОРМСТОБР(γ). Величина $z_{\frac{1-\gamma}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ є шириною довірчого інтервалу.

Таблиця 1.16

Значення $z_{\frac{1-\gamma}{2}}$

γ	0,4	0,25	0,2	0,15	0,1	0,05	0,025	0,01	0,005	0,001
$z_{\frac{1-\gamma}{2}}$	0,253	0,675	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,090

ПРИКЛАД Автомат, що фасує чай в пачки, працює зі стандартним відхиленням σ=5 г. Проведено вибірку об'ємом *n*=30 пачок. Середня вага пачки чаю у вибірки $\bar{x}$=101 г. Знайти довірчий інтервал для середньої ваги пачки чаю в генеральної сукупності із рівнем надійності γ=0,95. Знайти об'єм вибірки, якщо потрібна ширина довірчого інтервалу ± 1 грам.

Розв'язок. Оскільки γ=0,95, то $\frac{1-\gamma}{2} = \frac{1-0,95}{2} = 0,025$. З таблиці 1.16 знайдемо $z_{\frac{1-\gamma}{2}} = z_{0,025} = 1,96$.

Тоді ширина довірчого інтервалу: $z_{\frac{1-\gamma}{2}} \cdot \frac{\sigma}{\sqrt{n}} = 1,96 \cdot \frac{5}{\sqrt{30}} \approx 1,79$ і за формулою (1.18) маємо довірчий інтервал:

$$\bar{x} - 1,79 < a < \bar{x} + 1,79; \qquad 101 - 1,79 < a < 101 + 1,79; \qquad 99,21 < a < 102,79.$$

Отже, середня вага пачки чаю знаходиться в інтервалі від 99,21 до 102,79.

Знайдемо об'єм вибірки, необхідний для того, щоб ширина довірчого інтервалу дорівнювала 1 грам, тобто $z_{\frac{1-\gamma}{2}} \cdot \frac{\sigma}{\sqrt{n}} = 1$. Знайдемо *n* з отриманого рівняння: $1,96 \cdot \frac{5}{\sqrt{n}} = 1 \Rightarrow \sqrt{n} = 1,96 \cdot 5 = 9,8 \Rightarrow n = 9,8^2 = 96,04$. Отже, мінімальний об'єм вибірки для отримання довірчого інтервалу шириною 1 грам дорівнює 97 пачок.

Зауваження. У прикладі для розрахунку нового довірчого інтервалу потрібно знаходити вибіркове середнє для вибірки об'єму 97 пачок, а не середнє арифметичне середніх для вибірок об'єму 30 і 67 пачок.

Довірчий інтервал для генерального середнього при невідомій генеральної дисперсії

Нехай *X* – генеральна сукупність, що підкоряється нормальному закону розподілу; генеральна дисперсія σ^2 невідома; $\{x_1, x_2, ..., x_n\}$ – вибірка з генеральної сукупності об'єму *n*; $\bar{x}$ - вибіркове середнє; *S* – вибіркове середнє квадратичне відхилення. Потрібно знайти довірчий інтервал для генерального середнього *a* із заданим рівнем надійності γ.

Шуканий довірчий інтервал знаходиться за формулою:

$$\bar{x} - t_{\frac{1-\gamma}{2},n-1} \cdot \frac{S}{\sqrt{n-1}} < a < \bar{x} + t_{\frac{1-\gamma}{2},n-1} \cdot \frac{S}{\sqrt{n-1}}, \qquad (1.19)$$

де значення $t_{\frac{1-\gamma}{2},n-1}$ знаходиться з таблиці розподілу Стьюдента, яка є у будь-яких статистичних довідниках, або за допомогою вбудованої функції Excel

СТЬЮДРАСПОБР($\frac{1-\gamma}{2}$, *п*-1). Величина $t_{\frac{1-\gamma}{2},n-1}\cdot\frac{S}{\sqrt{n-1}}$ є шириною довірчого інтервалу.

ПРИКЛАД Автомат фасує чай в пачки. Проведено вибірку об'ємом *п*=30 пачок. Середня вага пачки чаю у вибірки $\bar{x}$=101 г, вибіркове стандартне відхилення *S*=4 г. Знайти довірчий інтервал для середньої ваги пачки чаю в генеральної сукупності із рівнем надійності γ=0,95. Знайти об'єм вибірки, якщо потрібна ширина довірчого інтервалу ± 1 грам.

Розв'язок. Оскільки γ=0,95, то $\frac{1-\gamma}{2}=\frac{1-0,95}{2}=0,025$; *п*=30, тоді *п*-1=29. За допомогою Excel знайдемо $t_{\frac{1-\gamma}{2},n-1}=t_{0,025;29}$. Натиснемо ***fx*** у командному рядку, виберемо категорію ***Статистические*** і функцію СТЬЮДРАСПОБР; задамо параметри 0,025 і 29 (див. рис. 1.9, зміст командного рядку). Отримаємо $t_{0,025;29}=2,3638$.

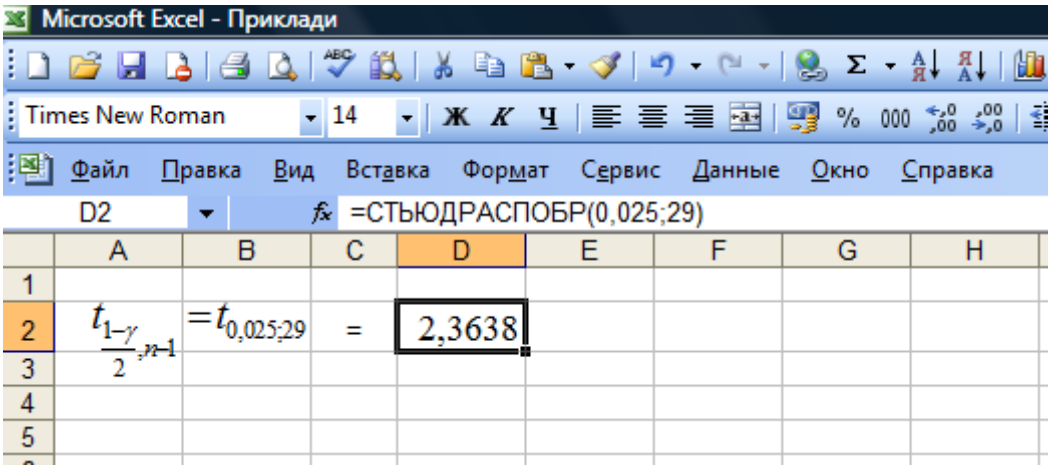


Рис. 1.9. Знаходження $t_{\frac{1-\gamma}{2},n-1}$ за допомогою Excel

Тоді ширина довірчого інтервалу $t_{\frac{1-\gamma}{2},n-1}\cdot\frac{S}{\sqrt{n-1}}=2,3638\cdot\frac{4}{\sqrt{29}}\approx 1,75$ і довірчий інтервал за формулою (1.19):

$$\bar{x}-1,75<a<\bar{x}+1,75;\quad 101-1,75<a<101+1,75;\quad 99,25<a<102,75.$$

Отже, середня вага пачки чаю знаходиться в інтервалі від 99,25 до 102,75.

Знайдемо об'єм вибірки, необхідний для того, щоб ширина довірчого інтервалу дорівнювала 1 грам, тобто $t_{\frac{1-\gamma}{2},n-1} \cdot \frac{S}{\sqrt{n-1}}$=1. Знайдемо *n* з отриманого рівняння: $2,3638 \cdot \frac{4}{\sqrt{n-1}} = 1 \Rightarrow \sqrt{n-1} = 2,3638 \cdot 4 \approx 9,46 \Rightarrow$. $\Rightarrow n-1 = 9,46^2 = 89,49 \Rightarrow n = 90,49 \approx 91$. Отже, мінімальний об'єм вибірки для отримання довірчого інтервалу шириною 1 грам дорівнює 91 пачка.

Довірчий інтервал для генеральної частки

В прикладних дослідженнях часто потрібно визначити частку об'єктів, що мають певну властивість.

Частка об'єктів генеральної сукупності, що має певну властивість, називається **генеральною часткою**. Частка об'єктів вибірки, що має певну властивість, називається **вибірковою часткою**.

Нехай *X* – генеральна сукупність, що підкоряється нормальному закону розподілу; $\{x_1, x_2, ..., x_n\}$ – вибірка з генеральної сукупності об'єму *n*; *m* – кількість елементів вибірки, що мають задану властивість; $w = \frac{m}{n}$ – вибіркова частка. Потрібно знайти довірчий інтервал для генеральної частки *W* із заданим рівнем надійності γ.

Шуканий довірчий інтервал знаходиться за формулою:

$$w - z_{\frac{1-\gamma}{2}} \cdot \sqrt{\frac{w(1-w)}{n}} < W < w + z_{\frac{1-\gamma}{2}} \cdot \sqrt{\frac{w(1-w)}{n}}, \qquad (1.20)$$

де значення $z_{\frac{1-\gamma}{2}}$ знаходиться з таблиці 1.16 або за допомогою вбудованої функції Excel НОРМСТОБР(γ). Величина $z_{\frac{1-\gamma}{2}} \cdot \sqrt{\frac{w(1-w)}{n}}$ є шириною довірчого інтервалу.

Зауваження. Формула (1.20) використовується тоді, коли $nw \geq 5, \quad n(1-w) \geq 5$.

ПРИКЛАД Проведено вибірку об'ємом *n*=2000 одиниць продукції. Серед обраних 150 одиниць виявилися бракованими. Знайти довірчий інтервал для генеральної долі бракованих виробів із рівнем надійності γ =0,95.

Розв'язок. Оскільки γ =0,95, то $\frac{1-\gamma}{2} = \frac{1-0,95}{2} = 0,025$. З таблиці 1.16 знайдемо $z_{\frac{1-\gamma}{2}} = z_{0,025} = 1,96$.

Знайдемо вибіркову частку бракованих виробів: $m = 150; w = \frac{m}{n} = \frac{150}{2000} = 0,075$.

Перевіримо можливість знаходження довірчого інтервалу:

$nw = 2000 \cdot 0,075 = 150 \geq 5, \quad n(1-w) = 2000(1-0,075) = 2000 \cdot 0,925 = 1850 \geq 5$

.

Тоді ширина довірчого інтервалу: $z_{\frac{1-\gamma}{2}} \cdot \sqrt{\frac{w(1-w)}{n}} = 1,96\sqrt{\frac{0,075(1-0,075)}{2000}} \approx 0,012$ і за формулою (1.20) маємо довірчий інтервал:

$$w - 0,012 < W < w + 0,012; \qquad 0,075 - 0,012 < W < 0,075 + 0,012;$$
$$0,063 < W < 0,087.$$

Отже, доля бракованих виробів в генеральній сукупності знаходиться в межах від 0,063 до 0,087, тобто складає від 6,3% до 8,7% від об'єму продукції.

2.Види і функції статистичних таблиць та графіків. Загальні правила їх побудови. Описова статистика

Шкали вимірювань

Статистичної обробці підлягають тільки ті ознаки об'єктів або фактори, які можна виміряти за деякою шкалою. Існують такі шкали вимірювань:

- шкала найменувань (класифікації);
- шкала порядку;
- шкала інтервалів;
- шкала відношень.

Шкала найменувань (класифікації)

Якщо дані вимірюються за шкалою найменувань, то над ними можливі тільки такі операції порівняння, як „рівні” або „не рівні”. Найменування слугує тільки для ідентифікації певного об’єкту – номер філіалу деякої фірми, номер методики навчання і т. ін.

Шкала порядку

Якщо дані вимірюються за шкалою порядку, їх можна порівняти за величиною – „більше”, „менше” або „рівні”. За такою шкалою вимірюються, наприклад, експертні оцінки, вік респондентів і т. ін.

Шкала інтервалів

Якщо дані вимірюються за шкалою інтервалів, їх можна порівняти за величиною – „більше”, „менше” або „рівні”, – та виконати операції додавання і віднімання, тобто визначити, наскільки більше або наскільки менше. Прикладом є шкали вимірювання температури.

Шкала відношень

Якщо дані вимірюються за шкалою відношень, їх можна порівняти за величиною та виконати всі арифметичні операції: додавання, віднімання, множення і ділення. За такою шкалою вимірюються вага, довжина, доход, об’єм виробництва і т. ін.

Від шкали вимірювання даних залежить можливість обчислити числові характеристики (табл. 1.17) і застосовувати певний статистичний метод.

Таблиця 1.17

Шкали вимірювань і числові характеристики

Назва шкали	**Числові характеристики**

Найменувань	Частоти
Порядку	Частоти, середнє, мода, медіана
Інтервалів	Частоти, середнє, мода, медіана, дисперсія
Відношень	Всі відомі

Визначення числових характеристик і довірчих інтервалів з використанням табличного процесору Microsoft Excel

Більшість числових характеристик у випадку незгрупованих даних можна обчислити з використанням табличного процесору Microsoft Excel. Основні вбудовані функції Excel, що застосовуються для таких розрахунків, надано у таблиці 1.18. Щоб викликати потрібну функцію, слід нажати кнопку f_x у командному рядку, обрати категорію ***Статистические*** та ім'я функції.

Крім того, часто корисні такі функції:

- НАИБОЛЬШИЙ (масив даних, *k*) – надає *k*-е найбільше значення в ряді даних;
- НАИМЕНЬШИЙ (масив даних, *k*) – надає *k*-е найменше значення в ряді даних.

Ширину довірчого інтервалу для генерального середнього можна знайти за допомогою вбудованої статистичної функції Excel ДОВЕРИТ (*альфа, станд_откл, размер*). Параметр *альфа* – це так званий рівень значущості, $\alpha = 1 - \gamma$; параметр *станд_откл* – це вибіркове середнє квадратичне відхилення *S*; параметр *размер* – це об'єм вибірки.

Таблиця 1.18

Статистичні функції Excel

Числові характеристики	**Назва функції**
Середнє	СРЗНАЧ (масив даних)
Середнє геометричне	СРГЕОМ (масив даних)

Мода	МОДА (масив даних)
Медіана	МЕДИАНА (масив даних)
Дисперсія	ДИСП (масив даних)
Середнє квадратичне відхилення	СТАНДОТКЛОН (масив даних)
Мінімальне значення	МИН (масив даних)
Максимальне значення	МАКС (масив даних)
Частота	ЧАСТОТА (масив даних; масив інтервалів)

ПРИКЛАД За даними вибіркового дослідження відома заробітна платня (у грн.) 20-и службовців певної компанії (табл. 1). Знайти за допомогою вбудованих статистичних функцій Excel всі можливі числові характеристики за даними таблиці. Знайти довірчий інтервал для генерального середнього – середньої заробітної платні службовців компанії.

Таблиця 1

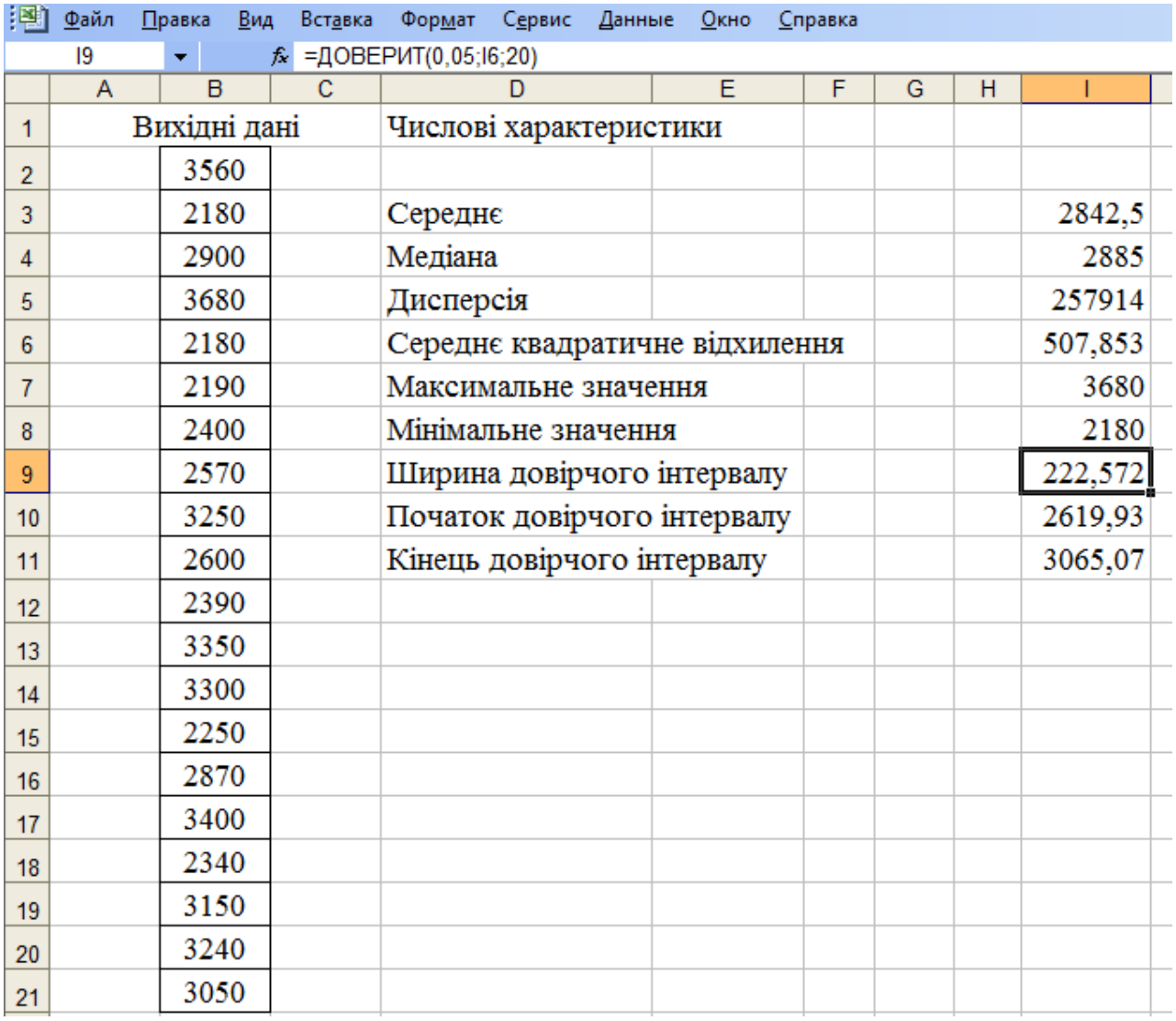
3560	2190	2390	3400
2180	2400	3350	2340
2900	2570	3300	3150
3680	3250	2250	3240
2180	2600	2870	3050

Розв’язок. Запишемо в лист Excel вхідні дані і числові характеристики, що можна знайти (рис. 1). Для знаходження характеристик введемо: в чарунку I3 формулу „=СРЗНАЧ(B2:B21)”; в чарунку I4 формулу „=МЕДИАНА(B2:B21)”; в чарунку I5 формулу „=ДИСП(B2:B21)”; в чарунку I6 формулу „=СТАНДОТКЛОН(B2:B21)”; в чарунку I7 формулу „=МАКС(B2:B21)”; в чарунку I8 формулу „=МИН(B2:B21)”.

Для знаходження довірчого інтервалу для генерального середнього

знайдемо за допомогою функції ДОВЕРИТ його ширину (див. рис. 1.10, командний рядок). Параметрами візьмемо $\alpha = 1 - \gamma = 1 - 0{,}95 = 0{,}05$; замість другого параметру надамо посилання на чарунку I6, що містить розраховане значення середнього квадратичного відхилення; *размер* – це об'єм вибірки, що дорівнює 20.

Для знаходження початку інтервалу запишемо в чарунку I10 формулу „=I3–I9"; для знаходження кінця – формулу „=I3+I9" в чарунку I11.



	A	B	C	D	E	F	G	H	I
1		Вихідні дані		Числові характеристики					
2		3560							
3		2180		Середнє					2842,5
4		2900		Медіана					2885
5		3680		Дисперсія					257914
6		2180		Середнє квадратичне відхилення					507,853
7		2190		Максимальне значення					3680
8		2400		Мінімальне значення					2180
9		2570		Ширина довірчого інтервалу					222,572
10		3250		Початок довірчого інтервалу					2619,93
11		2600		Кінець довірчого інтервалу					3065,07
12		2390							
13		3350							
14		3300							
15		2250							
16		2870							
17		3400							
18		2340							
19		3150							
20		3240							
21		3050							

Рис. 1 Розрахунок числових характеристик

Побудова гістограми засобами Microsoft Excel

Excel надає два способи побудови гістограми.

Для побудови гістограми першим способом необхідно:

1) Внести в лист Excel вхідні дані і інтервали, за якими ці дані будуть групуватися.

2) Знайти частоти попадання даних в інтервали за допомогою функції ЧАСТОТА, для чого:

- виділити діапазон чарунок (на одну більше, ніж інтервалів), в яких будуть записані частоти;
- викликати ***f_x – Статистические*** – ЧАСТОТА;
- ввести посилання на чарунки, що містять вхідні дані і інтервали;
- натиснути **Ctrl+Shift+Enter**.

3) Викликати ***Вставка – Диаграмма – Гистограмма***, появиться діалогове вікно (див. рис. 2.

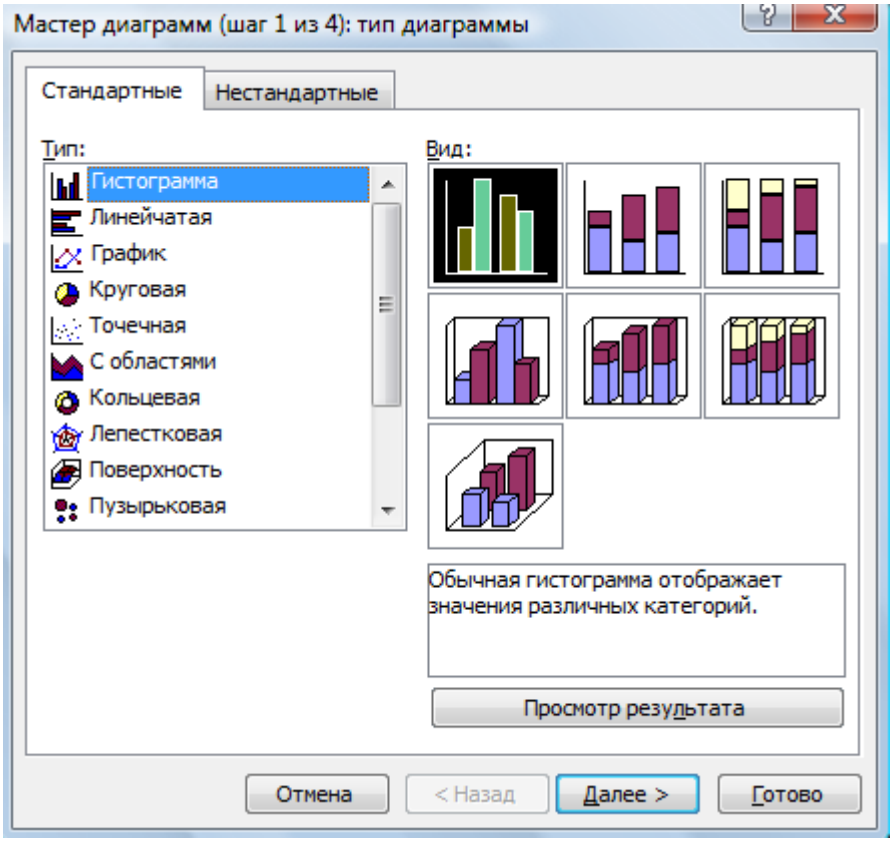


Рис. 2. Діалогове вікно майстра діаграм

4) Надати необхідні для побудови гістограми параметри:

- діапазон вхідних даних, спосіб їх групування (за рядками або стовпцями) та імена рядів даних, якщо це потрібно;

- якщо імена рядів надано, відмітити *Добавить легенду* і вказати її розміщення;
- якщо потрібно, добавити *Имена рядов*, або (та) *Имена категорій*, або (та) *Значения*;
- якщо потрібно, добавити *Заголовок*, *Линии сетки*, *Оси*, *Таблицу данных*.

Для побудови гістограми другим способом необхідно:

1) Внести в лист Excel вихідні дані.

2) Обрати в меню ***Сервис – Анализ данных – Гистограмма***, появиться діалогове вікно (див. рис. 3).

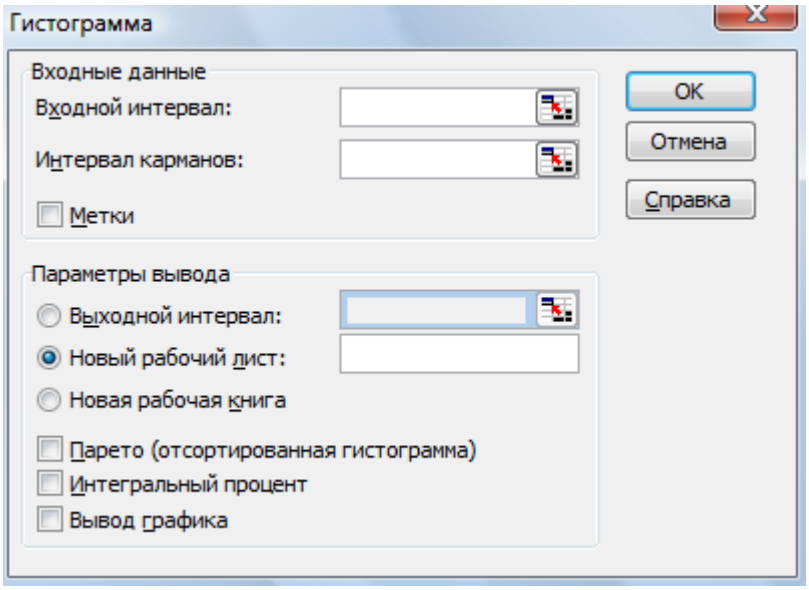


Рис. 3. Діалогове вікно для побудови гістограми

3) Задати необхідні для побудови гістограми параметри:

входной диапазон – задати посилання на чарунки, в яких знаходяться вхідні дані;

интервал карманов (параметр не є обов'язковим) – задати діапазон чарунок і набор граничних значень у порядку зростання; якщо параметр не введений, то буде автоматично створений набор відрізків, рівномірно розподілених між мінімальним і максимальним значеннями даних;

выходной диапазон – ввести посилання на верхню ліву чарунку діапазону, в який буде надано гістограму, або відмітити параметр *Новый рабочий лист* або *Новая рабочая книга*;

интегральный процент – якщо цей параметр відмічено, то будуть розраховані накопичені частоти і побудований їх графік;

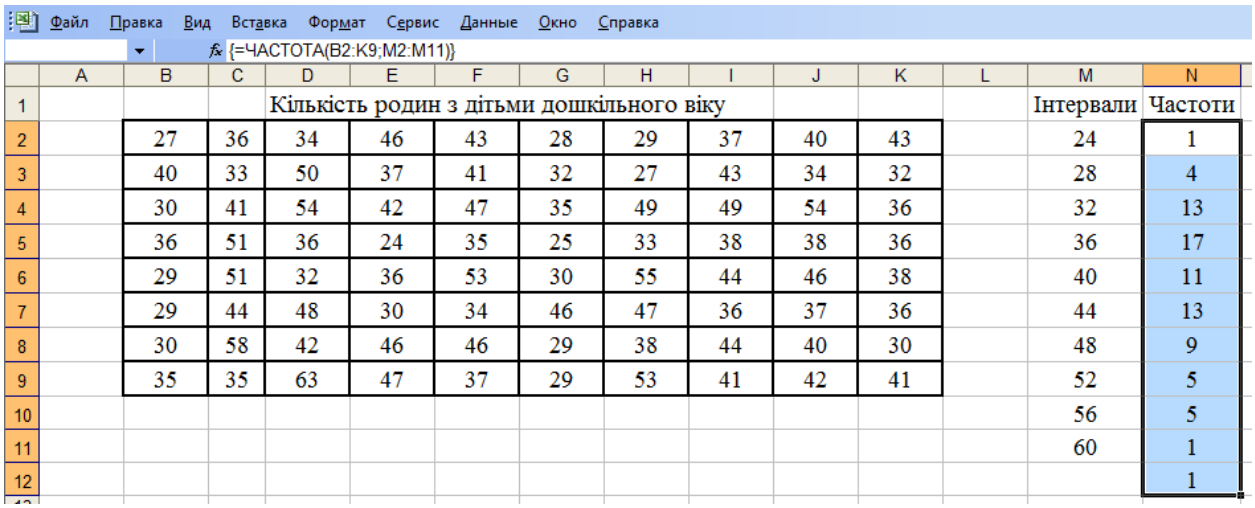
вывод графика – якщо цей параметр відмічено, то буде створено автоматичну діаграму, при цьому обов'язково задається значення *Новая книга*.

ПРИКЛАД За даними вибіркового дослідження відома кількість родин з дітьми дошкільного віку у селах деякої області (табл.2). Побудувати за допомогою Excel гістограму за даними таблиці.

Розв'язок. Запишемо в лист Excel вхідні дані завдання (рис. 2), стовпець з границями інтервалів в чарунках M2:M11 (задається тільки початок інтервалів). Розрахуємо частоти попадання в інтервали (див. зміст командного рядку на рис. 2).

Таблиця 2

27	36	34	46	43	28	29	37	40	43
40	33	50	37	41	32	27	43	34	32
30	41	54	42	47	35	49	49	54	36
36	51	36	24	35	25	33	38	38	36
29	51	32	36	53	30	55	44	46	38
29	44	48	30	34	46	47	36	37	36
30	58	42	46	46	29	38	44	40	30
35	35	63	47	37	29	53	41	42	41



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1				Кількість родин з дітьми дошкільного віку									Інтервали	Частоти
2		27	36	34	46	43	28	29	37	40	43		24	1
3		40	33	50	37	41	32	27	43	34	32		28	4
4		30	41	54	42	47	35	49	49	54	36		32	13
5		36	51	36	24	35	25	33	38	38	36		36	17
6		29	51	32	36	53	30	55	44	46	38		40	11
7		29	44	48	30	34	46	47	36	37	36		44	13
8		30	58	42	46	46	29	38	44	40	30		48	9
9		35	35	63	47	37	29	53	41	42	41		52	5
10													56	5
11													60	1
12														1

Рис. 2. Вхідні дані для побудови гістограми

Викличемо ***Вставка – Диаграмма – Гистограмма***, задамо діапазон даних, тобто розраховані частоти і вкажемо групування за стовпцями (див. рис. 3, виділення діапазону чарунок, що містять частоти).

Для зручності читання діаграми добавимо *Заголовок* та *Значения*.

Уберемо відмітку *Легенда*, оскільки імена рядів не було надано – розглядається тільки один тип даних.

Після побудови діаграми можна у разі необхідності змінити шрифти, ширину стовпців гістограми, колір стовпців і фону. Для внесення змін потрібно двічі натиснути правою кнопкою миші на відповідне поле гістограми.

Зауважимо, що на горизонтальній вісі надаються не границі інтервалів, а їх порядковий номер.

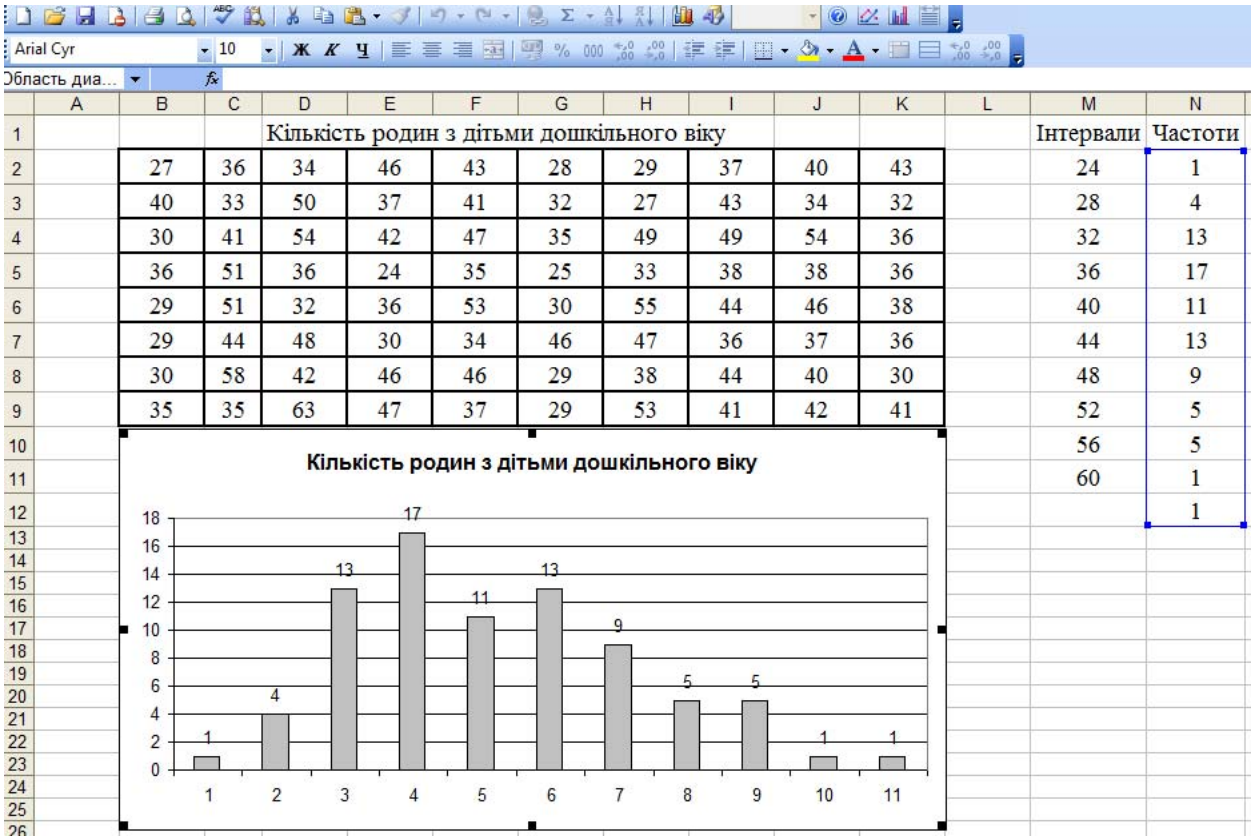


Рис. 3. Лист Excel з вхідними даними і гістограмою

Задачі

1.1. Відомі дані про безаварійну роботу автоматизованого комплексу (в місяцях) (табл. 1. 21). Побудувати статистичний ряд за даними вибірки, визначити середній час безаварійної роботи, дисперсію і середнє квадратичне відхилення часу. Побудувати полігон частот і гістограму.

Таблиця 1.21

0,000	0,000	0,002	0,006	0,023	0,084	0,382	0,810	0,003	0,864
1,033	0,912	0,093	0,324	0,194	0,522	2,336	0,057	0,654	0,250
0,877	0,276	0,037	0,537	0,183	1,306	0,752	0,198	1,623	0,875
0,185	0,274	0,613	0,356	0,645	0,676	1,079	0,500	0,902	0,191
0,250	0,348	0,320	0,182	0,458	0,936	1,204	0,576	0,303	0,522

1.2. Інтервал між потягами у метро складає 3 хв. В таблиці 1.21 надано час очікування пасажирами потягу. Скласти інтервальний статистичний ряд,

знайти середній час очікування, медіану, дисперсію і середнє квадратичне відхилення часу. Побудувати функцію розподілу величини *X* – часу очікування.

Таблиця 1.22

0,787	1,004	0,941	0,612	1,200	1,692	1,354	0,908	1,245	1,292
0,617	0,828	1,413	1,030	1,459	2,483	2,769	1,563	2,661	1,635
1,654	0,838	1,143	0,618	2,317	1,853	1,555	0,653	1,922	1,653
1,747	2,677	0,341	2,952	0,545	1,297	1,981	0,214	2,452	2,087
0,001	0,007	0,025	0,312	1,068	2,604	0,014	0,045	2,340	2,001

1.3. Відомі дані про інтервал часу між появою покупців у касовому залі деякого магазину (табл. 1.23). Побудувати інтервальний статистичний ряд, знайти всі можливі числові характеристики. Побудувати графіки функції і щільності розподілу величини X – інтервалу часу.

Таблиця 1.23

0,002	1,004	0,007	0,612	0,091	1,692	1,527	0,908	2,590	1,292
4,134	3,647	0,374	2,150	0,778	5,223	3,344	2,001	3,492	4,011
3,507	0,838	0,148	0,618	0,704	1,853	3,007	0,653	3,600	1,653
0,738	1,069	2,453	1,447	2,614	3,742	4,314	1,211	1,949	5,001
1,000	0,007	1,272	0,312	1,832	2,604	2,267	0,045	4,450	2,001

1.4. В таблиці 1.24 наведено значення прибутку 50 фірм, що належать одній корпорації (в 1000 у. од.). Знайти середнє значення прибутку, дисперсію і середнє квадратичне відхилення. Побудувати полігон частот і гістограму.

Таблиця 1.24

4,744	9,127	7,201	8,650	11,534	9,013	10,390	9,268	7,354	10,255
6,232	6,739	6,088	8,671	15,103	9,124	11,902	10,216	10,954	11,470
7,351	9,832	7,126	9,715	10,744	10,687	10,582	12,271	11,047	13,190
5,536	8,917	9,823	8,383	14,212	15,031	13,001	11,089	12,091	10,321
9,766	5,854	2,917	6,379	6,748	7,024	11,587	11,101	10,954	10,387

3. Первірка статистичних гіпотез(параметричні та непарметричні) ***Параметричні*** **та на параметричні методи** ***перевірки статистичних гіпотез***

Поняття про статистичні гіпотези

При застосуванні певних статистичних методів обробки даних вибірки часто ставляться вимоги до розподілу даних або до числових характеристик.

Статистичною гіпотезою називається будь-яке припущення про властивості досліджуваної величини, висунуте на основі статистичних даних.

За змістом статистичні гіпотези можна віднести до таких типів:

1) Гіпотези про вид закону розподілу досліджуваної величини.
2) Гіпотези про числові характеристики досліджуваної величини.
3) Гіпотези про рівність числових характеристик досліджуваних величин.
4) Гіпотези про належність досліджуваних величин до одній генеральної сукупності.
5) Гіпотези про вид моделі, що описує взаємозв'язок між досліджуваними величинами.
6) Гіпотези про належність досліджуваних величин до одного класу.

Статистичні гіпотези позначаються латинськими буквами H_0, H_1, і т.д. Гіпотеза H_0 формулюється як основна в тому сенсі, що при перевірці бажано було б встановити її справедливість. Основної гіпотезі H_0 протиставляються інші гіпотези H_1, H_2, …, які називаються альтернативними.

Прийняття основної або одній з альтернативних гіпотез здійснюється на основі дослідження статистичних даних. Дослідження проводиться за певним **критерієм**, який обирається відповідно до змісту гіпотези і виду наявних статистичних даних.

Якщо сформульовані гіпотези H_0 – основна та H_1 альтернативна (конкуруюча) і обраний критерій перевірки справедливості основної гіпотези, то прийняття H_0 позначає відкидання H_1, а відкидання H_0 позначає

справедливість H_1.

Оскільки прийняття гіпотези здійснюється на основі статистичних даних, то завжди існує ймовірність помилки.

Ймовірність відкидання гіпотези H_0, якщо вона справедлива, називається ймовірністю помилки першого роду або **рівнем значущості** і позначається α. Величина 1-α є ймовірністю прийняття справедливої гіпотези і називається **рівнем довіри**. Ймовірність прийняття гіпотези H_0, якщо вона не вірна, називається ймовірністю помилки другого роду і позначається β. Величина 1-β є ймовірністю відкидання невірної гіпотези і називається **потужністю критерію**.

Чим менше рівень значущості, тим менше ймовірність відкинути вірну гіпотезу. Зазвичай рівень значущості обирається дослідником рівним 0,1; 0,05; 0,01 або 0,001. Якщо, наприклад, обраний рівень значущості α=0,01, то риск відкинути вірну гіпотезу виникає в одному випадку із ста.

Зауваження. Перевірка статистичної гіпотези не надає точного висновку щодо її вірності або невірності. Прийняття гіпотези позначає, що на прийнятому рівні значущості вона не протиричіть статистичним даним.

Перевірка статистичних гіпотез зазвичай здійснюється за такими **етапами**:

1) Висунення припущень про вид розподілу досліджуваної величини (величин) або про її числові характеристики.
2) Формулювання статистичних гіпотез.
3) Вибір критерію перевірки відповідно до змісту гіпотез і статистичних даних.
4) Вибір рівня значущості залежно від вимог до точності результатів дослідження.
5) Розрахунок значення обраного критерію за статистичними даними.
6) Порівняння розрахованого значення критерію з його критичним значенням і прийняття або відкидання основної гіпотези.

2.2. Перевірка гіпотези про вид закону розподілу досліджуваної величини

Перевірка гіпотези про вид закону розподілу досліджуваної величини має велике значення для прикладних досліджень. Необхідність такої перевірки виникає при виборі критерію, оскільки для багатьох з них висувається вимога нормального розподілу статистичних даних. Означені гіпотези перевіряються при проектуванні систем масового обслуговування, перевірки якості продукції або праці і т. ін.

Припустимо, що з деякої генеральної сукупності X, яка розглядається як випадкова величина, обрана вибірка $\{x_1, x_2, \dots x_n\}$. За даними вибірки побудовано статистичний ряд (табл. 2.1), що містить варіанти x_i та відповідні частоти n_i, $i = \overline{1,k}$, k – кількість варіант у випадку дискретного ряду. У випадку інтервального ряду x_i – середини інтервалів, k – кількість інтервалів.

Таблиця 2.1

x_i	x_1	x_2	...	x_k
n_i	n_1	n_2	...	n_k

Отриманий на основі вибіркових даних статистичний ряд називається **емпіричним законом розподілу** величини *X*.

За даними статистичного ряду можна знайти числові характеристики, які є вибірковими параметрами закону розподілу *X*. Вид закону розподілу визначається відповідно до умов здобуття вибірки або залежно від виду графіка емпіричної щільності розподілу (гістограми) у випадку неперервної випадкової величини *X* і полігону частот, якщо величина *X* дискретна. Параметри обраного закону розподілу заміняються відповідними вибірковими параметрами.

Закон розподілу випадкової величини *X*, параметрами якого є відповідні вибіркові числові характеристики, називається **теоретичним законом розподілу**.

При здійсненні такої заміни немає впевненості, що закон розподілу обраний правильно. Тому розроблено процедуру, яка дозволяє оцінити ступінь відповідності обраного закону даним вибірки. Критерії здійснення такої перевірки називаються **критерії згоди**, найбільш відомим з яких є **критерій Пірсона** χ^2 (хи-квадрат).

Критерій Пірсона χ^2 обчислюється за формулою:

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(n_i - n_i^{'}\right)^2}{n_i^{'}}, \qquad (2.1)$$

де n_i' – частоти, отримані за теоретичним законом розподілу (теоретичні частоти).

З формули (2.1) видно, що у випадку, коли відповідні теоретичні та емпіричні частоти співпадають, χ^2=0. Тобто чим ближче χ^2 до нуля, тим краще узгоджуються вибіркові дані та обраний теоретичний закон розподілу.

Розраховане значення критерію χ^2 порівнюється з його критичним значенням $\chi^2_{\alpha,l}$, яке знаходиться за статистичними таблицями або за допомогою вбудованої статистичної функції Excel ХИ2ОБР(α, *l*). Параметрами функції ХИ2ОБР є: α – рівень значущості; *l* – ступені волі, *l* $= k - r - 1$, де *k* – кількість груп емпіричного розподілу, *r* – кількість параметрів теоретичного розподілу (наприклад, для нормального розподілу *r*=2, оскільки параметрів два – *a* і σ). Якщо $\chi^2 < \chi^2_{\alpha,l}$, то гіпотеза про закон розподілу приймається. У противному випадку гіпотеза відкидається.

Зауваження. У деяких статистичних таблицях критичне значення χ^2 надається залежно від рівня довіри γ, γ =1- α.

Отже, перевірка гіпотези про закон розподілу величини *X* здійснюється за такими **етапами**:

1) З генеральної сукупності *X* здобувається вибірка і будується статистичний ряд.

2) Висувається гіпотеза про закон розподілу випадкової величини *X.*

3) Знаходяться вибіркові параметри обраного закону розподілу.

4) Розраховуються теоретичні частоти.

5) Розраховується критерій χ^2 за формулою (2.1).

6) Обирається рівень значущості α (або рівень довіри γ) і знаходиться критичне значення $\chi^2_{\alpha,l}$ (або $\chi^2_{\gamma,l}$).

7) Порівнюються розраховане і критичне значення критерію χ^2 і робиться висновок про справедливість висунутої гіпотези.

ПРИКЛАД За наданим інтервальний статистичним рядом (табл. 2.2) знайти закон розподілу випадкової величини *X*.

Таблиця 2.2

$[a_i; a_{i+1})$	$[-2;-1,2)$	$[-1,2;-0,4)$	$[-0,4;0,4)$	$[0,4;1,2)$	$[1,2;2)$
n_i	6	11	21	7	5

Розв'язок. Для визначення виду закону розподілу побудуємо гістограму за даними таблиці 2.2 (рис. 2.1). За видом гістограми висуваємо гіпотезу про нормальний закон розподілу даної випадкової величини:

H_0 – випадкова величина X розподілена за нормальним законом;

H_1 – випадкова величина X не розподілена за нормальним законом.

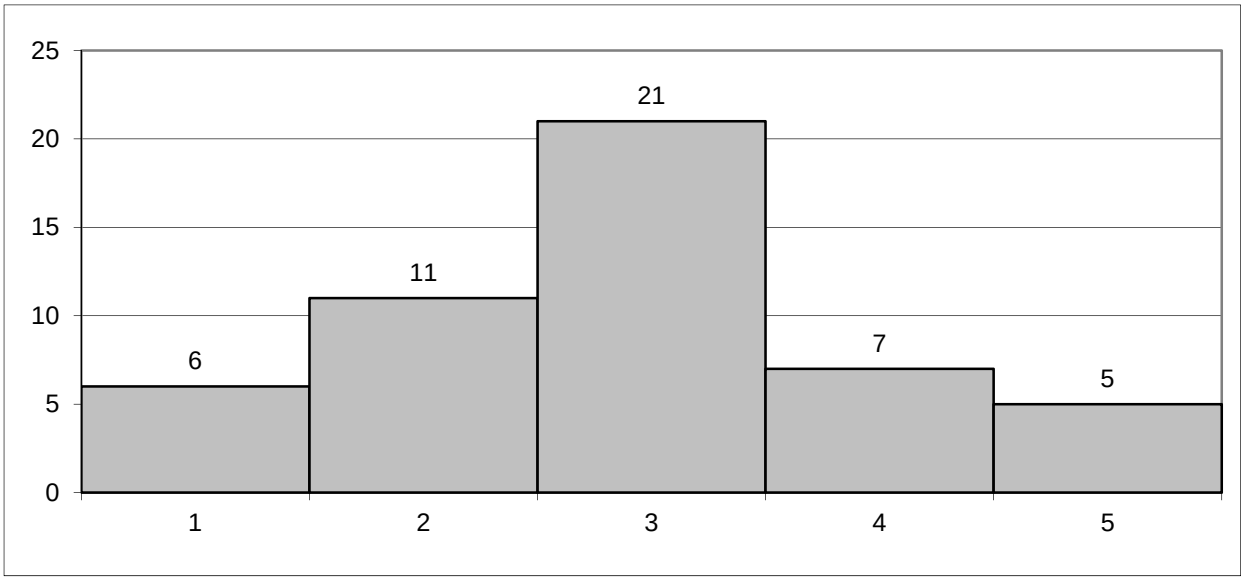


Рис. 2.1. Гістограма за даними таблиці 2.2

Щільність розподілу випадкової величини, розподіленої за нормальним законом має вид $f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}$, де a і σ – параметри розподілу. Знайдемо означені параметри, враховуючи, що $\bar{x}=a;\quad S^2=\sigma^2$. Розрахунки оформимо у вигляді таблиці (табл. 2.3).

Таблиця 2.3

$[a_i;a_{i+1})$	$[-2;-1,2)$	$[-1,2;-0,4)$	$[-0,4;0,4)$	$[0,4;1,2)$	$[1,2;2)$
n_i	6	11	21	7	5
x_i	-1,6	0,8	0	0,8	1,6
$x_i n_i$	-9,6	8,8	0	5,6	8
$\left(x_i-\bar{x}\right)^2 n_i$	13,572	5,452	0,194	5,620	14,382

Знайдемо вибіркове середнє, вибіркову дисперсію і вибіркове середнє квадратичне відхилення за формулами (1.6), (1.13) та (1.16) відповідно:

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{k}x_i n_i=\frac{1}{50}(-1,6\cdot 6-0,8\cdot 11+0\cdot 21+0,8\cdot 7+1,6\cdot 5)=-0,096;$$

$$S^2=\frac{1}{n}\sum_{i=1}(x_i-\bar{x})^2 n_i=\frac{1}{50}(13,572+5,452+0,194+5,620+14,382)=0,7844;$$

$$S=\sqrt{S^2}\approx 0,886.$$

Отже, параметрами теоретичного закону розподілу є: $\bar{x}=a=-0,096;\quad S=\sigma=0,886$.

Для знаходження значення критерію χ^2 розрахуємо теоретичні частоти n_i'. Теоретичні частоти можна знайти за формулою $n_i'=np_i$, де p_i – ймовірності попадання випадкової величини в певний інтервал. Для нормального закону розподілу означені ймовірності знаходяться за формулою $P(a_i<X<a_{i+1})=p_i=\frac{1}{2}\left[\Phi\left(\frac{a_{i+1}-\bar{x}}{S}\right)-\Phi\left(\frac{a_i-\bar{x}}{S}\right)\right]$, де Ф – функція

Лапласа, значення якої надані у статистичних таблицях. Для зручності обчислень побудуємо таблицю (табл. 2.4).

За формулою (2.1) маємо: $\chi^2 = \sum_{i=1}^{k} \frac{\left(n_i - n_i^{'}\right)^2}{n_i^{'}} \approx 4{,}16$. Знайдемо критичне значення $\chi^2_{\alpha,l}$, враховуючи, що $l = k - r - 1$=5-2-1=2. Рівень значущості α оберемо рівним 0,1. За допомогою Excel знаходимо ХИ2ОБР(0,1; 2)=4,6.

Отже, оскільки $\chi^2 < \chi^2_{\alpha,l}$, гіпотеза H_0 про нормальний розподіл приймається, гіпотеза H_1 відкидається.

Таблиця 2.4

$[a_i; a_{i+1})$	$[-2;-1,2)$	$[-1,2;-0,4)$	$[-0,4;0,4)$	$[0,4;1,2)$	$[1,2;2)$
n_i	6	11	21	7	5
x_i	-1,6	0,8	0	0,8	1,6
$x_i n_i$	-9,6	8,8	0	5,6	8
$\left(x_i - \bar{x}\right)^2 n_i$	13,572	5,452	0,194	5,620	14,382
$\frac{a_i - \bar{x}}{S}$	-2,1498	-1,2465	-0,3433	0,56	1,4633
$\Phi(\frac{a_i - \bar{x}}{S})$	-0,958	-0,785	-0,266	0,425	0,856
p_i	0,0856	0,2595	0,3455	0,2155	0,063
$n_i^{'}=np_i$	4,325	12,975	17,275	10,775	3,15
$\frac{\left(n_i - n_i^{'}\right)^2}{n_i^{'}}$	0,649	0,301	0,803	1,323	1,087

ПРИКЛАД На одній з міських АТС фіксувалася кількість телефонних

дзвінків в годину. Спостереження велися на протязі 100 годин, їх результати представлені в таблиці 2.5. Чи можна вважати навантаження на АТС стандартним?

Таблиця 2.5

Кількість викликів в годину	0	1	2	3	4	5	6	7
Кількість спостережень	6	27	26	20	10	5	5	1

Розв'язок. Навантаження на АТС можна вважати стандартним, якщо випадкова величина *X* – кількість телефонних дзвінків, що поступили, підкоряється закону розподілу Пуассона. Тобто отримання відповіді необхідно перевірити гіпотезу про закон розподілу випадкової величини. Сформулюємо гіпотези:

H_0 – випадкова величина *X* підкоряється закону розподілу Пуассона;

H_1 – випадкова величина *X* не підкоряється закону розподілу Пуассона.

Закон Пуассона має вигляд: $p_k = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k = 0,1,\ldots;\lambda > 0$ де λ - параметр розподілу. Крім того, відомо, що $\bar{x} = \lambda; \quad S^2 = \lambda$. Отже, для встановлення параметра λ потрібно знайти $\bar{x}$ або S^2.

Випадкова величина *X* – кількість викликів в годину; тоді кількість спостережень – це відповідні значенням *X* частоти n_i, а таблиця 2.5 є статистичним рядом і емпіричним законом розподілу величини *X.* Знайдемо $\bar{x}$ і S^2. Для зручності обчислення оформимо у вигляді таблиці (табл. 2.6).

Таблиця 2.6

x_i	0	1	2	3	4	5	6	7	Суми

n_i	6	27	26	20	10	5	5	1	100
$x_i\, n_i$	0	27	52	60	40	25	30	7	241
$\left(x_i - \bar{x}\right)^2 n_i$	34,85	53,68	4,37	6,96	25,28	33,54	64,44	21,07	244,19

Отже, $\bar{x} = \frac{1}{n}\sum_{i=1} x_i n_i = \frac{1}{100} \cdot 241 = 2,41$;

$S^2 = \frac{1}{n}\sum_{i=1} (x_i - \bar{x})^2 n_i = \frac{1}{100} \cdot 244,19 \approx 2,44$.

Оскільки повинна виконуватися рівність $\bar{x} = \lambda;\ S^2 = \lambda$ то як параметр можна вибрати або $\bar{x}$, або S^2, або їх середнє арифметичне. Виберемо $\lambda = \frac{\bar{x} + S^2}{2} \approx 2,426$. Таким чином, гіпотеза H_0 – це припущення, що величина *X* розподілена згідно із законом Пуассона: $p_k = \frac{e^{-2,426}\, 2,426^k}{k!}, \quad k = 0,1,\ldots$

Перевіримо правильність гіпотези за допомогою критерію Пірсона. Знайдемо теоретичні частоти, використовуючи формулу: $p_k = \frac{e^{-2,426}\, 2,426^k}{k!}, \quad k = 0,1,\ldots$ Як k візьмемо значення *X*, тобто x_i.

Відмітимо, що p_i – ймовірність того, що *X* прийме значення x_i, тобто статистично вони є відносними частотами, теоретичні частоти знаходитимемо за формулою: $n_i' = np_i$.

Для зручності при обчисленні теоретичних частот продовжимо таблицю, складену на основі статистичного ряду (табл. 2.7).

Отже, за результатами розрахунків $\chi^2 = \sum_{i=0}^{7} \frac{\left(n_i - n_i'\right)^2}{n_i'} = 5,739$.

Для даного завдання $l = k - r - 1$=8-1-1=6. Виберемо рівень значущості α=0,01 і знайдемо за допомогою таблиць або функції ХИ2ОБР табличного процесора Excel значення $\chi^2_{\alpha,l}$: $\chi^2_{0,01;6}$=16,812. Оскільки для

такого рівня довіри (0,99) $\chi^2 < \chi^2_{\alpha,l}$, то гіпотезу H_0 про розподіл Пуассона можна прийняти.

Таблиця 2.7

x_i	0	1	2	3	4	5	6	7	Суми
n_i	6	27	26	20	10	5	5	1	100
$p_i=\frac{e^{-2,426}2,426^{x_i}}{x_1!}$	0,09	0,21	0,26	0,21	0,13	0,06	0,03	0,01	
$n_i'=np_i$	8,84	21,44	26,01	21,03	12,76	6,19	2,50	0,87	
$\frac{\left(n_i-n_i'\right)^2}{n_i'}$	0,91	1,44	4,65E-06	0,05	0,59	0,23	2,49	0,02	5,739

Висновок: навантаження на АТС можна вважати стандартним.

ПРИКЛАД З метою впорядкування роботи міського суспільного транспорту фіксувався час очікування в хвилинах пасажирами тролейбусів на декількох маршрутах. Було проведено 200 вимірювань, їх результати представлені в таблиці 2.8. Чи можна вважати, що перевезення по перевірених маршрутах забезпечені раціонально?

Таблиця 2.8

Час очікування	1 – 3	3 – 5	5 – 7	7 – 9	9 – 11	11 – 13
Кількість спостережень	25	30	48	35	42	20

Розв'язок. Можна вважати, що перевезення по перевірених маршрутах забезпечені раціонально, якщо випадкова величина *X* – час очікування пасажирами транспорту підкоряється рівномірному закону розподілу. Тобто задача зводиться до перевірки гіпотези про закон розподілу випадкової величини. Сформулюємо гіпотези:

H_0 – випадкова величина *X* розподілена рівномірно;

H_1 – випадкова величина *X* не розподілена рівномірно.

Щільність розподілу випадкової величини, що підкоряється рівномірному закону має вигляд: $f(x)=\begin{cases}\dfrac{1}{b-a}, a\le x\le b\\ 0, \quad x<a, x>b\end{cases}$ де *a; b* - параметри розподілу. Крім того, відомо, що $\bar{x}=\dfrac{a+b}{2};\quad S^2=\dfrac{(b-a)^2}{12}$. Тобто для встановлення параметрів *a* і *b* потрібно знайти $\bar{x}$ і S^2, після чого розв'язати систему: $\begin{cases}\bar{x}=\dfrac{a+b}{2}\\ S^2=\dfrac{(b-a)^2}{12}\end{cases}$.

Оскільки за умов задачі випадкова *X* – час очікування транспорту, то кількість спостережень – це відповідні значенням *X* частоти n_i, а таблиця 2.8 – це інтервальний статистичний ряд і емпіричний закон розподілу *X.* Знайдемо $\bar{x}$ і S^2. Як x_i візьмемо середини відповідних інтервалів. Для зручності обчислення оформимо у вигляді таблиці (табл. 2.9).

Таблиця 2.9

$[a_i; a_{i+1})$	1 – 3	3 – 5	5 – 7	7 – 9	9 – 11	11 – 13	Суми
x_i	2	4	6	8	10	12	
n_i	25	30	48	35	42	20	200
$x_i\, n_i$	50	120	288	280	420	240	1398
$\left(x_i-\bar{x}\right)^2 n_i$	622,5	268,2	47,045	35,704	380,52	502	1856

Отже, $\bar{x}=\dfrac{1}{n}\sum_{i=1}^{6}x_i n_i=\dfrac{1}{200}\cdot 1398=6,99$;

$S^2=\dfrac{1}{n}\sum_{i=1}^{6}(x_i-\bar{x})^2 n_i=\dfrac{1}{200}\cdot 1856\approx 9,2799$.

Складемо систему для визначення параметрів рівномірного розподілу і розв’яжемо її:

$$\begin{cases} 6,99 = \dfrac{a+b}{2} \\ 9,2799 = \dfrac{(b-a)^2}{12} \end{cases} \Rightarrow \begin{cases} a+b = 13,98 \\ (b-a)^2 = 111,3588 \end{cases} \Rightarrow \begin{cases} a+b = 13,98 \\ b-a \approx 10,55 \end{cases} \Rightarrow \begin{cases} a \approx 1,715 \\ b \approx 12,265 \end{cases}.$$

Таким чином, гіпотеза H_0 – це припущення, що *X* розподілена за рівномірним законом із щільністю розподілу:

$$f(x) = \begin{cases} \dfrac{1}{12,265 - 1,715}, & 1,715 \le x \le 12,265 \\ 0, & x < 1,715, \quad x > 12,265 \end{cases}.$$

Перевіримо справедливість гіпотези H_0 за допомогою критерію Пірсона. Для знаходження теоретичних частот використаємо формулу $n_i' = np_i$, а ймовірності попадання в інтервали p_i знайдемо за формулою: $P(\alpha < X < \beta) = \dfrac{\beta - \alpha}{b - a}$.

Для зручності обчислень теоретичних частот складемо таблицю (табл. 2.10). Врахуємо, що $\beta - \alpha = 2; \quad b - a = 12,265 - 1,715 = 10,55$ для всіх інтервалів, окрім першого і останнього. Для першого інтервалу $\beta - \alpha = \beta - a = 3 - 1,715 = 1,285$; для останнього інтервалу $\beta - \alpha = b - \alpha = 12,265 - 11 = 1,265$.

Таблиця 2.10

$[a_i; a_{i+1})$	1 – 3	3 – 5	5 – 7	7 – 9	9 – 11	11 – 13	Суми
n_i	25	30	48	35	42	20	200
$\beta - \alpha$	1,285	2	2	2	2	1,265	
$p_i = \dfrac{\beta - \alpha}{10,55}$	0,122	0,19	0,19	0,19	0,19	0,12	
$n_i' = np_i$	24,360	37,915	37,915	37,915	37,915	23,981	

$\frac{\left(n_i - n_i'\right)^2}{n_i'}$	0,017	1,6522	2,6827	0,2241	0,4402	0,6609	5,677

Отже, $\chi^2 = \sum_{i=0}^{7} \frac{\left(n_i - n_i'\right)^2}{n_i'} = 5,677$.

Для даного завдання $l = k - r - 1$=6-2-1=3. Виберемо рівень значущості α =0,01 і знайдемо за допомогою таблиць або функції ХИ2ОБР табличного процесора Excel значення $\chi^2_{\alpha,l}$: $\chi^2_{0,01;3}$=11,34. Оскільки для такого рівня довіри (0,99) $\chi^2 < \chi^2_{\alpha,l}$ гіпотезу про рівномірний розподіл приймаємо.

Висновок: перевезення по перевірених маршрутах організовані раціонально.

Перевірка гіпотез про генеральні середні і дисперсії

В прикладних задачах часто виникає необхідність перевірки рівності середніх значень та дисперсій за даними двох або більше вибірок. Наприклад, коли визначається перевага одній з технологій виготовлення певної продукції, або наявність підвищення продуктивності праці після внесення змін в процес виробництва, або при перевірці якості продукції. Здійснення означеної перевірки виконується за критеріями, що обираються залежно від виду розподілу вибіркових даних і мети дослідження. Для деяких критеріїв перевірки рівності середніх значень висувається додаткова вимога про рівність генеральних дисперсій.

Перевірка гіпотези про рівність генеральних дисперсій. F-критерій (Фішера)

Перевірка гіпотези про рівність генеральних дисперсій здійснюється за F-критерієм (Фішера) тільки тоді, коли статистичні дані незалежні і розподілені за нормальним законом. Формулюються гіпотези:

H_0 – дисперсії двох нормально розподілених генеральних сукупностей рівні, тобто $S_1^2 = S_2^2$;

H_1 - дисперсії двох нормально розподілених генеральних сукупностей не рівні, тобто $S_1^2 \neq S_2^2$.

F-критерій (Фішера) розраховується за формулою:

$$F = \frac{S_1^2}{S_2^2}, \quad S_1^2 > S_2^2, \tag{2.2}$$

Гіпотеза H_0 приймається, якщо розраховане значення F менше критичного значення розподілу Фішера $F_{крит}$, взятого із рівнем значущості α і ступенями волі l_1 та l_2 для чисельнику і знаменнику відповідно: $l_1 = n_1 - 1$, $l_2 = n_2 - 1$, де n_1, n_2 – об'єми вибірок. $F_{крит}$ можна знайти за допомогою вбудованої статистичної функції Excel FРАСПОБР (α; l_1; l_2).

Зауваження. Дисперсія у чисельнику дроби у формулі (2.2) повинна бути більше дисперсії у знаменнику, тобто значення F-критерію повинно бути більше одиниці.

ПРИКЛАД Відомі дані про продуктивність праці (одиниць продукції за зміну) двох груп працівників: група 1 складається з працівників, що пройшли спеціальний навчальний курс; група 2 – із працівників, що не пройшли курсу (табл. 2.11). Враховуючи, що дані розподілені за нормальним законом, перевірити гіпотезу про рівність дисперсій.

Таблиця 2.11

	Група 1					Група 2				
Продуктивність праці	34	85	96	102	103	63	69	83	89	106

Кількість працівників	5	2	11	8	4	2	6	8	3	1

Розв'язок. Дані таблиці 2.11 є двома вибірками. Перша – вибірка значень величини X_1 – продуктивності праці робітників, що пройшли навчання, друга – вибірка величини X_2 – продуктивності праці робітників, що не пройшли навчання.

Сформулюємо гіпотези: H_0 – дисперсії генеральних сукупностей, з яких зроблено вибірки, рівні, $S_1^2 = S_2^2$; H_1 - дисперсії не рівні, $S_1^2 \neq S_2^2$. Перевіримо справедливість гіпотези H_0 за F-критерієм (Фішера).

Знайдемо за вибірковими даними оцінку дисперсії X_1 за формулою (1.14). Розрахунки оформимо у вигляді таблиці (табл. 2.12).

Таблиця 2.12

x_i	34	85	96	102	103	Суми
n_i	5	2	11	8	4	30
$x_i n_i$	170	170	1056	816	412	2624
$(x_i - \bar{x})^2 n_i$	14293,42	12,17	801,00	1689,74	965,14	17761,47

Отже, $\bar{x}_1 = \frac{1}{n}\sum_{i=1}^{k} x_i n_i = \frac{1}{30} \cdot 2624 \approx 87,47$;

$$S_1^2 = \frac{1}{n-1}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i = \frac{1}{29} \cdot 17761,47 \approx 612,46.$$

Аналогічно знайдемо оцінку дисперсії X_2. Розрахунки оформимо у вигляді таблиці (табл.. 2.13).

Таблиця 2.13

x_i	63	69	83	89	106	Суми
n_i	2	6	8	3	1	20
$x_i n_i$	126	414	664	267	106	1577

$(x_i - \bar{x})^2 n_i$	502,45	582,13	137,78	309,07	737,12	2268,55

Отже, $\overline{x_2} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i = \frac{1}{20} \cdot 1577 \approx 78,85$;

$$S_2^{\,2} = \frac{1}{n-1}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i = \frac{1}{19} \cdot 2268,55 \approx 119,40 .$$

Знайдемо значення F-критерію за формулою (2.2). Оскільки $S_1^2 > S_2^2$, то $F = \frac{S_1^2}{S_2^2} = \frac{612,46}{119,40} \approx 5,13$. Знайдемо $F_{крит}$, враховуючи, що $l_1 = n_1 - 1 = 30-1 = 29$; $l_2 = n_2 - 1 = 20-1 = 19$. Рівень значущості оберемо α =0,05. Тоді $F_{крит}$= FРАСПОБР (0,05; 29; 19)=2,077.

Оскільки $F > F_{крит}$, то гіпотезу H_0 відкидаємо і приймаємо гіпотезу H_1 – дисперсії не рівні, тобто вибірки здобути з різних генеральних сукупностей.

Висновок: навчальний курс суттєво впливає на продуктивність праці робітників.

Перевірка гіпотези про рівність генеральних дисперсій. Критерій Зігеля-Тьюкі

Якщо статистичні дані не розподілені за нормальним законом або виміряються з використанням порядкової шкали, то перевірка гіпотези про рівність генеральних дисперсій здійснюється за критерієм Зігеля-Тьюкі. Формулюються гіпотези:

H_0 – дисперсії двох генеральних сукупностей рівні, тобто $S_1^2 = S_2^2$;

H_1 - дисперсії двох генеральних сукупностей не рівні, тобто $S_1^2 \neq S_2^2$.

Перевірка виконується за даними двох вибірок за такими етапами:

1) Формується об’єднана вибірка.

2) Даним об’єднаної вибірки присвоюються ранги (порядкові номери) за правилом: найменшому значенню присвоюється ранг 1, двом найбільшим – ранги 2 і 3; наступним двом найменшим – ранги 4 і 5;

наступним найбільшим – ранги 6 і 7 і т. д. При цьому, якщо кількість елементів вибірки непарна, то її центральний елемент (тобто медіана) не отримує ніякого рангу.

3) Розраховуються суми рангів елементів вихідних вибірок R_1 *і* R_2.

4) Розраховується нормальна випадкова величина *Z* за формулою:

$$Z = \frac{2R_1 - n_1\left(n_1 + n_2 + 1\right) + 1}{\frac{n_2}{3} \cdot \sqrt{n_1\left(n_1 + n_2 + 1\right)}}, \qquad (2.3)$$

де n_1, n_2 – об'єми вибірок. При цьому R_1 – сума рангів меншої за об'ємом вибірки. Якщо $2R_1 > n_1\left(n_1 + n_2 + 1\right) + 1$, *Z* розраховується за формулою:

$$Z = \frac{2R_1 - n_1\left(n_1 + n_2 + 1\right) - 1}{\frac{n_2}{3} \cdot \sqrt{n_1\left(n_1 + n_2 + 1\right)}}. \qquad (2.4)$$

5) У випадку, коли перевіряються вибірки різних об'ємів, обчислюється скоректована нормальна випадкова величина Z' за формулою:

$$Z' = Z + \left(\frac{1}{10n_1} - \frac{1}{10n_2}\right)\left(Z^3 - 3Z\right). \qquad (2.5)$$

6) Обирається рівень значущості α.

7) За допомогою таблиці значень функції нормального розподілу або вбудованої функції Excel НОРМРАСП знаходиться ймовірність *P*(*Z*) або *P*(Z').

8) Порівнюються рівень значущості α і величина 2*P*(*Z*) (2*P*(Z')). Якщо 2*P*(*Z*)> α (або 2*P*(Z')>α), то гіпотеза H_0 про рівність генеральних дисперсій приймається.

Зауваження. Для перевірки правильності присвоєння рангів можна скористатися формулами: $R_1 + R_2 = \frac{\left(n_1 + n_2\right)\left(n_1 + n_2 + 1\right)}{2}$ у випадку парної

кількості елементів об'єднаної вибірки; $R_1 + R_2 = (n_1 + n_2)\left(\frac{n_1 + n_2 + 1}{2} - 1\right)$ у випадку непарної кількості цих елементів.

ПРИКЛАД У результаті дослідження надійності станків двох виробників отримані дані про час (в годинах) безаварійної праці (табл. 2.14). Враховуючи, що дані не розподілені за нормальним законом, перевірити гіпотезу про рівність дисперсій.

Таблиця 2.14

Виробник	Час безаварійної праці									
1	280	230	112	176	90	175	216	110	205	115
2	200	126	225	210	260	194	156	240	170	232

Розв'язок. Дані таблиці 2.14 є двома вибірками. Перша – вибірка значень величини X_1 – часу безаварійної праці станків виробника 1; друга – вибірка величини X_2 – часу безаварійної праці станків виробника 2.

Сформулюємо гіпотези: H_0 – дисперсії генеральних сукупностей, з яких зроблено вибірки, рівні, $S_1^2 = S_2^2$; H_1 - дисперсії не рівні, $S_1^2 \neq S_2^2$. Перевіримо справедливість гіпотези H_0 за критерієм Зігеля-Тьюкі.

Сформуємо об'єднану вибірку, присвоїмо її елементам ранги і знайдемо їх суму. Результати розрахунків оформимо у вигляді таблиці (табл. 2.15). Для зручності підкреслимо елементи першої вибірки.

Розрахуємо за формулою (2.3) значення Z, враховуючи, що n_1=n_2=10:

$$Z = \frac{2R_1 - n_1(n_1 + n_2 + 1) + 1}{\frac{n_2}{3} \cdot \sqrt{n_1(n_1 + n_2 + 1)}} = \frac{2 \cdot 95 - 10(10 + 10 + 1) + 1}{\frac{10}{3} \cdot \sqrt{10(10 + 10 + 1)}} \approx \frac{-19}{48,26} \approx -0,394.$$

Оберемо рівень значущості α =0,05. За допомогою вбудованої функції Excel НОРМРАСП знаходиться ймовірність $P(Z)$:

$P(Z)$=НОРМРАСП(-0,394; 0; 1; ИСТИНА)=0,3469.

Оскільки $2P(Z)=2\cdot 0{,}3469=0{,}6938 > \alpha=0{,}05$, то гіпотеза H_0 про рівність генеральних дисперсій приймається.

Висновок: дисперсії надійності станків двох виробників однакові.

Таблиця 2.15

Елементи об'єднаної вибірки	Сортована об'єднана вибірка	Ранги елементів об'єднаної вибірки	Ранги елементів першої вибірки	Ранги елементів другої вибірки
280	90	1	1	
230	110	4	4	
112	112	5	5	
176	115	8	8	
90	126	9		9
175	156	12		12
216	170	13		13
110	175	16	16	
205	176	17	17	
115	194	20		20
200	200	19		19
126	205	18	18	
225	210	15		15
210	216	14	14	

260	225	11		11
194	230	10	10	
156	232	7		7
240	240	6		6
170	260	3		3
232	280	2	2	
		Суми	95	115

Перевірка гіпотези про рівність генеральних середніх. Критерій Стьюдента

Критерій Стьюдента використовується для перевірки гіпотез про рівність генеральних середніх, якщо статистичні дані розподілені за нормальним законом. Формулюються гіпотези:

H_0 – середні двох генеральних сукупностей рівні, тобто $\overline{x_1} = \overline{x_2}$;

H_1 - середні двох генеральних сукупностей не рівні, тобто $\overline{x_1} \neq \overline{x_2}$.

Перевірка виконується за даними двох вибірок об'ємом n_1 та n_2. При цьому можливі такі випадки.

Випадок 1. Генеральні дисперсії рівні ($S_1^2 = S_2^2$). Тоді t-критерій Стьюдента обчислюється за формулою:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}. \qquad (2.6)$$

Розраховане значення t-критерію порівнюється з критичним значенням $t_{крит}$, де $t_{крит}$ – критичне значення розподілу Стьюдента з параметрами $\frac{\alpha}{2}$ і ступенем волі $l = n_1 + n_2 - 2$, яке надається в

статистичних таблицях або знаходиться за допомогою вбудованої функції Excel СТЬЮДРАСПОБР($\frac{\alpha}{2}$;*l*).

Випадок 2. Генеральні дисперсії не рівні ($S_1^2 \neq S_2^2$). Тоді t-критерій Стьюдента обчислюється за формулою:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}. \qquad (2.7)$$

Розраховане значення t-критерію також порівнюється з критичним значенням $t_{крит}$, але ступені волі розраховуються за формулою:

$$l = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 + 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 + 1}} + 2. \qquad (2.8)$$

Випадок 3. Вибірки не є незалежними, оскільки на них впливає певний фактор і його вплив не відомий, або вибірки є даними, отриманими до і після проведення певного експерименту. Тоді формується парна вибірка, і для кожної пари елементів знаходиться *d* – різність їх значень. Подальша перевірка здійснюється над вибіркою різностей. t-критерій Стьюдента обчислюється за формулою:

$$t = \frac{\overline{x_d}}{S_d / \sqrt{n-1}}. \qquad (2.9)$$

де $\overline{x_d}$ – вибіркове середнє для вибірки різностей;

S_d – вибіркове середнє квадратичне відхилення для вибірки різностей;

n – об’єм вибірки різностей.

Розраховане значення t-критерію також порівнюється з критичним значенням розподілу Стьюдента з параметрами $\frac{\alpha}{2}$ і ступенями волі $l = n - 1$.

У всіх випадках гіпотеза H_0 приймається, якщо розраховане значення t-критерію менше за абсолютною величиною критичного значення $t_{крит}$:

$$|t| < t_{крит}.$$

ПРИКЛАД Для виробництва кожної з 10 деталей за першою технологією було витрачено, у середньому, 30с. Дисперсія часу складала 1с^2. Для виробництва кожної з 16 деталей за другою технологією було витрачено, у середньому, 28с із дисперсією часу 2с^2. Чи можна вважати, що у середньому для виробництва деталей за першою технологією потрібно більше часу?

Розв'язок. За умов задачі було зроблено дві вибірки: перша – вибірка об'єму n_1=10 значень величини X_1 – часу, потрібному для виготовлення деталей за першою технологією; друга – вибірка об'єму n_2=16 значень величини X_2 – часу, потрібному для виготовлення деталей за другою технологією. Відомі вибіркові середні $\overline{x_1}$=30с та $\overline{x_2}$=28с – середній час, необхідний для виготовлення деталей за першою і другою технологіями відповідно. Відомі дисперсії часу для вибірок: S_1^2=1с^2 та S_1^2=2с^2. За питанням задачі потрібно перевірити гіпотезу про рівність генеральних середніх.

Сформулюємо гіпотези:

H_0 – середні двох генеральних сукупностей рівні, тобто $\overline{x_1} = \overline{x_2}$;

H_1 - середні двох генеральних сукупностей не рівні, тобто $\overline{x_1} \neq \overline{x_2}$.

Перед вибором критерію для перевірки потрібно встановити, чи рівні генеральні дисперсії. Скористуємось критерієм Фішера. За формулою (2.2) обчислимо значення F-критерію:

оскільки $S_2^2 > S_1^2$, то $F = \frac{S_2^2}{S_1^2} = \frac{2}{1} = 2$.

Знайдемо критичне значення розподілу Фішера $F_{крит}$: оберемо рівень значущості α =0,05; врахуємо, що ступені волі l_1=n_1–1=9 та l_2=n_2–1=15. Тоді FРАСПОБР (α; l_2; l_1)=FРАСПОБР (0,05; 15; 9)=3,006. Отже, $F < F_{крит}$, тому генеральні дисперсії можна вважати рівними.

Оскільки генеральні дисперсії рівні (випадок 1), то t-критерій Стьюдента розраховуємо за формулою (2.6):

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{30 - 28}{\sqrt{\frac{10 \cdot 1 + 16 \cdot 2}{10 + 16 - 2}\left(\frac{1}{10} + \frac{1}{16}\right)}} \approx 7{,}032.$$

Знайдемо критичне значення розподілу Стьюдента $t_{крит}$, враховуючи, що $l = n_1 + n_2 - 2 = 10 + 16 - 2 = 24$. Оберемо значення α =0,05. Тоді $t_{крит}$=СТЬЮДРАСПОБР($\frac{\alpha}{2}$;l)=СТЬЮДРАСПОБР(0,025;24)=2,39.

Отже, $|t| > t_{крит}$, тому гіпотеза H_0 про рівність генеральних середніх відкидається на рівні значущості 0,05 і приймається гіпотеза H_1.

Висновок: для вироблення деталей за першою технологією потрібно, у середньому, більше часу.

ПРИКЛАД Для виробництва кожної з 51 деталі за першою технологією було витрачено, у середньому, 30с. Дисперсія часу складала 6с^2. Для виробництва кожної з 41 деталей за другою технологією було витрачено, у середньому, 25с із дисперсією часу 3с^2. Чи можна вважати, що у середньому для виробництва деталей за першою технологією потрібно більше часу?

Розв'язок. За умов задачі було зроблено дві вибірки: перша – вибірка об'єму n_1=51 значень величини X_1 – часу, потрібному для виготовлення деталей за першою технологією; друга – вибірка об'єму n_2=41 значень

величини X_2 – часу, потрібному для виготовлення деталей за другою технологією. Відомі вибіркові середні $\overline{x_1}$=30с та $\overline{x_2}$=25с – середній час, необхідний для виготовлення деталей за першою і другою технологіями відповідно. Відомі дисперсії часу для вибірок: S_1^2=6с² та S_1^2=3с². За питанням задачі потрібно перевірити гіпотезу про рівність генеральних середніх.

Сформулюємо гіпотези:

H_0 – середні двох генеральних сукупностей рівні, тобто $\overline{x_1} = \overline{x_2}$;

H_1 - середні двох генеральних сукупностей не рівні, тобто $\overline{x_1} \neq \overline{x_2}$.

Перед вибором критерію для перевірки потрібно встановити, чи рівні генеральні дисперсії. Скористуємось критерієм Фішера. За формулою (2.2) обчислимо значення F-критерію:

$$\text{оскільки } S_1^2 > S_2^2, \text{ то } F = \frac{S_1^2}{S_2^2} = \frac{6}{3} = 2.$$

Знайдемо критичне значення розподілу Фішера $F_{крит}$: оберемо рівень значущості α=0,05; врахуємо, що ступені волі $l_1=n_1-1=50$ та $l_2=n_2-1=40$. Тоді FРАСПОБР (α;l_1;l_2)=FРАСПОБР (0,05; 50; 40)=1,66.

Отже, $F > F_{крит}$, тому генеральні дисперсії можна вважати різними.

Оскільки генеральні дисперсії різні (випадок 2), то t-критерій Стьюдента розраховуємо за формулою (2.7):

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{30 - 25}{\sqrt{\frac{6}{51} + \frac{3}{41}}} = \frac{5}{\sqrt{0,19077}} \approx 11,45.$$

Знайдемо критичне значення розподілу Стьюдента $t_{крит}$. Оберемо рівень значущості α=0,05. Обчислимо ступені волі за формулою (2.8):

$$l=\frac{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1+1}+\frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2+1}}+2=\frac{\left(\frac{6}{51}+\frac{3}{41}\right)^2}{\frac{\left(\frac{6}{51}\right)^2}{51+1}+\frac{\left(\frac{3}{41}\right)^2}{41+1}}+2\approx 93.$$

Тоді $t_{крит}$=СТЬЮДРАСПОБР($\frac{\alpha}{2}$;*l*)=СТЬЮДРАСПОБР(0,025;93)=2,27.

Отже, $|t|>t_{крит}$, тому гіпотеза H_0 про рівність генеральних середніх відкидається на рівні значущості 0,05 і приймається гіпотеза H_1.

Висновок: для вироблення деталей за першою технологією потрібно, у середньому, більше часу.

ПРИКЛАД Проводилося дослідження якості праці двох механіків з регулювання станків-автоматів. Вимірювався час роботи станків до першої зупинки для регулювання (табл. 2.16). Чи можна вважати, що механіки працюються однаково якісно?

Таблиця 2.16

Номер станка	1	2	3	4	5
Час роботи після регулювання першим механіком (год.)	60,2	62,3	61,3	60,7	63,4
Час роботи після регулювання другим механіком (год.)	59,4	58,3	62,1	63,4	60,8

Розв'язок: За умов задачі було зроблено дві вибірки: перша – вибірка об'єму n_1=5 значень величини X_1 – часу роботи станків після регулювання першим механіком; друга – вибірка об'єму n_2=5 значень величини X_2 – часу роботи станків після регулювання другим механіком. За питанням задачі потрібно перевірити гіпотезу про рівність генеральних середніх.

Сформулюємо гіпотези:

H_0 – середні двох генеральних сукупностей рівні, тобто $\overline{x_1} = \overline{x_2}$;

H_1 - середні двох генеральних сукупностей не рівні, тобто $\overline{x_1} \neq \overline{x_2}$.

Оскільки вибірки не є незалежними, то необхідно сформувати парну вибірку (випадок 3), знайти для кожної пари елементів різність їх значень *d*, обчислити для парної вибірки $\overline{x_d}$ і S_d та розрахувати t-критерій Стьюдента за формулою (2.9). Розрахунки для зручності оформимо у вигляді таблиці (табл. 2.17).

Таблиця 2.17

$\overline{x_{1i}}$	60,2	62,3	61,3	60,7	63,4	Суми
$\overline{x_{2i}}$	59,4	58,3	62,1	63,4	60,8	
$d=\overline{x_{1i}} - \overline{x_{2i}}$	0,8	4	-0,8	-2,7	2,6	3,9
$(d-\overline{x_d})^2$	0,0004	10,3684	2,4964	12,1104	3,3124	28,288

Знайдемо $\overline{x_d}$ за формулою (1.7): $\overline{x_d} = \frac{1}{n}\sum_{i=1}^{n} d = \frac{3,9}{5} = 0,78$. Знайдемо $S_d{}^2$ за формулою (1.15): $S_d{}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(d - \overline{x_d}\right)^2 = \frac{28,288}{5} = 5,6576$. Тоді $S_d = \sqrt{S_d{}^2} \approx 2,379$.

Знайдемо t-критерій Стьюдента за формулою (2.9):

$$t = \frac{\overline{x_d}}{S_d / \sqrt{n-1}} = \frac{0,78}{2,379 / \sqrt{5-1}} \approx 0,656.$$

Знайдемо критичне значення $t_{крит}$ розподілу Стьюдента. Оберемо рівень значущості α=0,05, врахуємо, що ступені волі $l = n - 1$=4. Тоді $t_{крит}$=СТЬЮДРАСПОБР($\frac{\alpha}{2}$;*l*)=СТЬЮДРАСПОБР(0,025;4)=3,495.

Отже, $|t| < t_{крит}$, тому гіпотеза H_0 про рівність генеральних середніх приймається на рівні значущості 0,05.

Висновок: якість праці двох механіків однакова.

Перевірка гіпотези про рівність генеральних середніх. Критерій Уілкоксона

Критерій Уілкоксона використовується для перевірки гіпотези про рівність генеральних середніх за статистичними даними двох вибірок тоді, коли дані не підкоряються нормальному закону розподілу.

Формулюються гіпотези:

H_0 – середні двох генеральних сукупностей рівні, тобто $\overline{x_1} = \overline{x_2}$;

H_1 - середні двох генеральних сукупностей не рівні, тобто $\overline{x_1} \neq \overline{x_2}$.

Для здійснення перевірки необхідно, щоб об'єм першої вибірки n_1 був менше об'єму другої вибірки n_2. Якщо ця умова не виконується, то слід змінити місця вибірок, тобто першу вважати другої. Перевірка виконується за такими етапами:

1) Формується об'єднана вибірка, об'єм якої $n=n_1+n_2$.

2) Даним об'єднаної вибірки присвоюються ранги (порядкові номери) за правилом: найменшому значенню присвоюється ранг 1, наступному найменшому – ранг 2 і т. д. При цьому, якщо деякі елементи вибірки співпадають, то їм присвоюються середні ранги. Для цього додаються ранги, які мали б ці елементи, якщо були б різні; розраховується їх середнє арифметичне; кожному із однакових елементів присвоюється ранг, що дорівнює розрахованому середньому арифметичному.

3) Розраховуються суми рангів елементів вихідних вибірок R_1 і R_2.

4) Обирається величина *W*: якщо вибірки рівні за об'ємом, то *W*= R_1 або *W*= R_2; якщо вибірки різні за об'ємом, то *W*= R_1 – сумі рангів меншої за об'ємом вибірки.

5) Розраховується значення W^* критерію Уілкоксона за формулою:

$$W^* = \frac{2W - n_1\left(n_1 + n_2 + 1\right) + 1}{\sqrt{\frac{n_1 n_2}{3} \cdot \left(n_1 + n_2 + 1\right)}}. \qquad (2.10)$$

6) Якщо у вибірці є дані з однаковими рангами – зв'язки, то критерій Уілкоксона W^* розраховується за формулою:

$$W^* = \frac{2W - n_1\left(n_1 + n_2 + 1\right) + 1}{\sqrt{\frac{n_1 n_2}{12} \cdot \left(n_1 + n_2 + 1\right) - \frac{\sum_{i=1}^{m} t_i\left(t_i^2 - 1\right)}{\left(n_1 + n_2\right)\left(n_1 + n_2 - 1\right)}}}, \qquad (2.11)$$

де *m* – кількість груп однакових рангів, що містять дані обох вибірок (кількість загальних зв'язок);

t_i – кількість елементів в і-тої групі (розмір зв'язок).

6) Обирається рівень значущості α.

7) Розраховується найменший рівень значущості $p_{гр}$ за формулою:

$$p_{гр} = 2 \cdot \left(1 - \Phi\left(\left|W^*\right|\right)\right), \qquad (2.12)$$

де *Ф*(*u*) – функція нормального стандартного розподілу, значення якої можна знайти в статистичних таблицях або за допомогою вбудованої функції Excel НОРМРАСП: $p_{гр}$=НОРМРАСП(W^*; 0; 1; ИСТИНА).

8) Порівнюються рівень значущості α і величина $p_{гр}$. Якщо $p_{гр} > \alpha$, то гіпотеза H_0 про рівність генеральних дисперсій приймається.

ПРИКЛАД У результаті дослідження надійності станків двох виробників отримані дані про час (в годинах) безаварійної праці (табл. 2.18).

Враховуючи, що дані не розподілені за нормальним законом, перевірити гіпотезу про рівність середніх.

Таблиця 2.18

Виробник	Час безаварійної праці									
1	280	230	112	176	90	175	216	110	205	115
2	200	126	225	210	260	194	156	240	170	232

Розв'язок. Дані таблиці 2.18 є двома вибірками. Перша – вибірка значень величини X_1 – часу безаварійної праці станків виробника 1; друга – вибірка величини X_2 – часу безаварійної праці станків виробника 2.

Сформулюємо гіпотези:

H_0 – середні генеральних сукупностей, з яких зроблено вибірки, рівні, $\overline{x_1} = \overline{x_2}$;

H_1 - середні не рівні, $\overline{x_1} \neq \overline{x_2}$.

Перевіримо справедливість гіпотези H_0 за критерієм Уілкоксона.

Сформуємо об'єднану вибірку, присвоїмо її елементам ранги і знайдемо їх суму. Результати розрахунків оформимо у вигляді таблиці (табл. 2.19). Для зручності підкреслимо елементи першої вибірки.

Оскільки вибірки рівні за об'ємом, то $W = R_1$ або $W = R_2$.

Розрахуємо значення W^* критерію Уілкоксона за формулою (2.10), якщо $W = R_1 = 89$:

$$W^* = \frac{2W - n_1(n_1 + n_2 + 1) + 1}{\sqrt{\frac{n_1 n_2}{3} \cdot (n_1 + n_2 + 1)}} = \frac{2 \cdot 89 - 10(10 + 10 + 1) + 1}{\sqrt{\frac{10 \cdot 10}{3} \cdot (10 + 10 + 1)}} \approx -1,476.$$

Таблиця 2.15

Елементи об'єднаної вибірки	Сортована об'єднана вибірка	Ранги елементів	Ранги елементів	Ранги елементів

		об’єднаної вибірки	першої вибірки	другої вибірки
280	90	1	1	
230	110	2	2	
112	112	3	3	
176	115	4	4	
90	126	5		5
175	156	6		6
216	170	7		7
110	175	8	8	
205	176	9	9	
115	194	10		10
200	200	11		11
126	205	12	12	
225	210	13		13
210	216	14	14	
260	225	15		15
194	230	16	16	
156	232	17		17
240	240	18		18
170	260	19		19
232	280	20	20	
		Суми	89	121

Оберемо рівень значущості α =0,05.

Розрахуємо найменший рівень значущості $p_{гр}$ за формулою (2.12):

якщо W^*=-1,476, то $\Phi(|W^*|)$=НОРМРАСП(1,476; 0; 1; ИСТИНА)=0,93, тоді $p_{гр} = 2 \cdot \left(1 - \Phi\left(\left|W^*\right|\right)\right)$=2(1-0,93)=0,14. Отже, оскільки $p_{гр} > \alpha$, то гіпотеза H_0 про рівність середніх приймається на рівні значущості α =0,05.

Розрахуємо значення W^* критерію Уілкоксона, якщо $W = R_2$=121:

$$W^* = \frac{2W - n_1\left(n_1 + n_2 + 1\right) + 1}{\sqrt{\frac{n_1 n_2}{3} \cdot \left(n_1 + n_2 + 1\right)}} = \frac{2 \cdot 121 - 10(10 + 10 + 1) + 1}{\sqrt{\frac{10 \cdot 10}{3} \cdot \left(10 + 10 + 1\right)}} \approx 1,245.$$

Розрахуємо найменший рівень значущості $p_{гр}$:

якщо W^*=1,245, то $\Phi(W^*)$=НОРМРАСП(1,245; 0; 1; ИСТИНА)=0,89, тоді $p_{гр} = 2 \cdot \left(1 - \Phi\left(\left|W^*\right|\right)\right)$=2(1-0,89)=0,22. Отже, оскільки $p_{гр} > \alpha$, то гіпотеза H_0 про рівність середніх приймається на рівні значущості α =0,05.

Висновок: надійність праці станків двох виробників однакова.

Перевірка статистичних гіпотез із використанням Microsoft Excel

Двохвибірковий F-тест для дисперсій

Перевірку гіпотези про рівність генеральних дисперсій за F-критерієм (Фішера) можна виконати за допомогою пакету аналізу Microsoft Excel. Для використання пакету необхідно:

1) Вибрати в меню послідовно пункти ***Сервис – Анализ данных***, після чого з’явиться вікно для вибору інструменту аналізу (рис. 2.3).

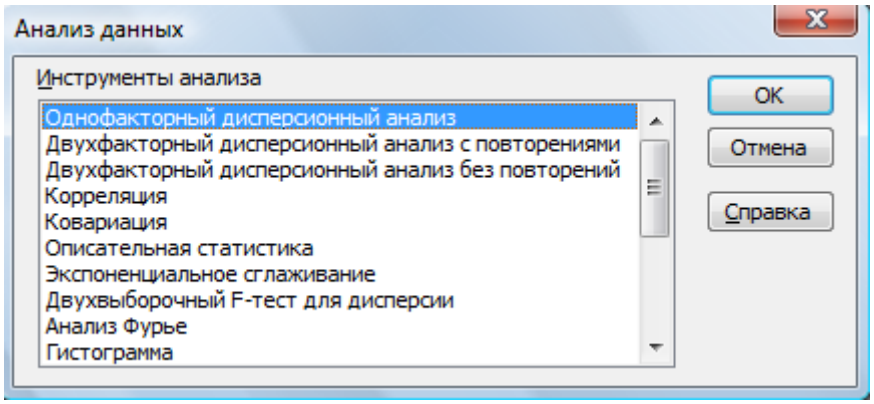


Рис. 2.3. Діалогове вікно аналізу даних

2) Вибрати у діалоговому вікні інструмент ***Двухвыборочный F-тест для дисперсии***, після чого з’явиться вікно для надання параметрів (рис. 2.4).

3) Задати всі необхідні параметри і натиснути ОК.

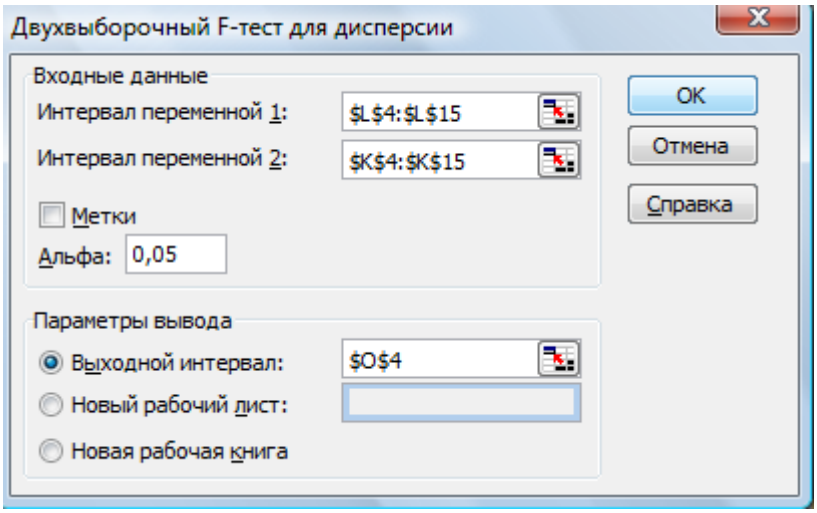
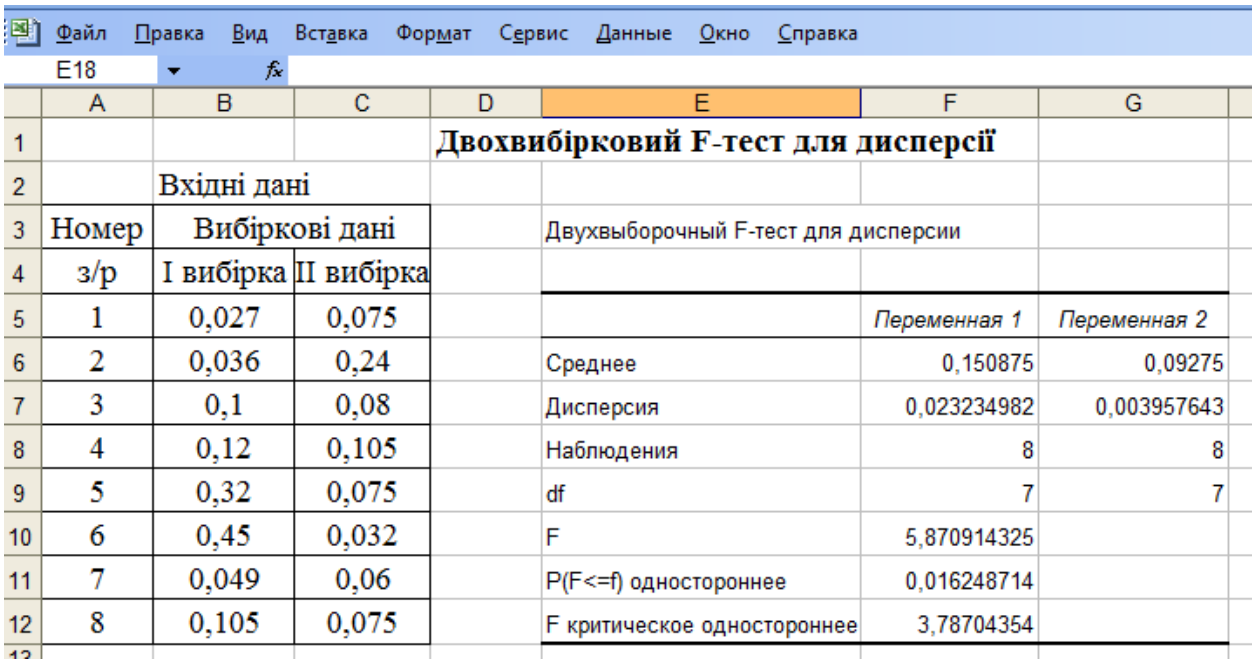


Рис. 2.4. Вікно надання параметрів

Приклад і результат роботи тесту наведено на рис. 2.5.



Файл Правка Вид Вставка Формат Сервис Данные Окно Справка

E18

	A	B	C	D	E	F	G
1					**Двохвибірковий F-тест для дисперсії**		
2		Вхідні дані					
3	Номер	Вибіркові дані			Двухвыборочный F-тест для дисперсии		
4	з/р	I вибірка	II вибірка				
5	1	0,027	0,075			*Переменная 1*	*Переменная 2*
6	2	0,036	0,24		Среднее	0,150875	0,09275
7	3	0,1	0,08		Дисперсия	0,023234982	0,003957643
8	4	0,12	0,105		Наблюдения	8	8
9	5	0,32	0,075		df	7	7
10	6	0,45	0,032		F	5,870914325	
11	7	0,049	0,06		P(F<=f) одностороннее	0,016248714	
12	8	0,105	0,075		F критическое одностороннее	3,78704354	

Рис. 2.5. Перевірка рівності генеральних дисперсій

Двохвибірковий t-тест для середніх

Перевірку гіпотез про рівність генеральних середніх за критерієм Стьюдента можна виконати за допомогою пакету аналізу даних Microsoft Excel. Для здійснення перевірки необхідно у вікні для вибору інструменту аналізу (рис. 2.3) вибрати:

1) У випадку рівних генеральних дисперсій – інструмент ***Двухвыборочный t-тест с одинаковыми дисперсиями***;

2) У випадку різних генеральних дисперсій – інструмент ***Двухвыборочный t-тест с разными дисперсиями***;

3) У випадку залежних вибірок – ***Парный двухвыборочный t-тест для средних***.

Після вибору інструменту аналізу з’явиться вікно для надання параметрів аналізу, аналогічне зображеному на рис. 2.4, в якому необхідно задати масиви чарунок із вхідними вибірковими даними і рівень значущості (стандартний рівень – 0,05). Приклад і результат роботи тесту наведено на рис. 2.6.

Файл Правка Вид Вставка Формат Сервис Данные Окно Справка

C22

	A	B	C	D	E	F	G
1				**Двохвибірковий t-тест для середніх**			
2		Вхідні дані					
3	Номер	Вибіркові дані			Двухвыборочный t-тест с различными дисперсиями		
4	з/р	I вибірка	II вибірка				
5	1	0,027	0,075			*Переменная 1*	*Переменная 2*
6	2	0,036	0,24		Среднее	0,150875	0,09275
7	3	0,1	0,08		Дисперсия	0,023234982	0,003957643
8	4	0,12	0,105		Наблюдения	8	8
9	5	0,32	0,075		Гипотетическая разность средних	0	
10	6	0,45	0,032		df	9	
11	7	0,049	0,06		t-статистика	0,996970694	
12	8	0,105	0,075		P(T<=t) одностороннее	0,172413357	
13					t критическое одностороннее	1,833112923	
14					P(T<=t) двухстороннее	0,344826713	
15					t критическое двухстороннее	2,262157158	
16							

Рис. 2.6. Перевірка рівності генеральних середніх

Задачі

2.1. При дослідженні ефективності ліків двох типів було сформовано дві групи хворих з 15 осіб. У випадку використання ліків першого типу строк одужання склав, у середньому, 11 днів з дисперсією 3 дні; другого – 8 днів з дисперсією 4 дні. Перевірити на рівні значущості 0,01 гіпотезу про перевагу другого типу ліків.

2.2. При дослідженні якості продукції, яку випускають дві фірми, було перевірено 50 одиниць продукції першої фірми і 63 одиниці другої. Середня оцінка продукції першої фірми склала 36 балів, другої – 32 бали. Дисперсія оцінок виявилася рівною 10. Визначити фірму, що виробляє більш якісну продукцію на рівні значущості 0,01ю

2.3. Середній річний об’єм виробництва 10 однотипних компаній в регіоні А склав 5900 одиниць продукції; 8 компаній того ж типу в регіоні В – 5690 одиниць. Вибіркова дисперсія об’єму виробництва в регіоні А склала

1000, в регіоні В – 1500. Перевірити гіпотезу про рівність середніх значень на рівні значущості 0,05.

2.4. Середня урожайність пшениці у фермерському хазяйстві складала 75 ц/га. Після внесення нового добриву середня урожайність підвищилася на 5 ц/га. Вибіркова дисперсія склала 2,5 ц/га. Перевірити на рівні значущості 0,01 гіпотезу про доцільність нового добриву.

2.5. Середня продуктивність праці на підприємстві складала 30 одиниць продукції за зміну з дисперсією 4. Після внесення змін в організацію праці середня продуктивність склала 34 з дисперсією 7. Перевірити на рівні значущості 0,01 гіпотезу про доцільність внесення змін в організацію праці.

2.6 – 2.15. Для виробництва кожної з n_1=53 деталей за першою технологією було витрачено, у середньому $\overline{x_1}$ секунд часу з дисперсією S_1^2. Для виробництва кожної з n_2=43 деталей за другою технологією було витрачено, у середньому $\overline{x_2}$ секунд часу з дисперсією S_2^2 (табл. 2.16). Чи можна зробити висновок, що для виробництва деталей за першою технологією потрібно, у середньому, більше часу, ніж за другою. Гіпотезу перевірити на рівні значущості α.

Таблиця 2.16

№	$\overline{x_1}$	S_1^2	$\overline{x_2}$	S_2^2	α
2.6	38	4	31	2	0,05
2.7	39	5	32	3	0,01
2.8	33	7	31	8	0,05
2.9	37	8	34	7	0,01

2.10	35	4	32	5	0,05
2.11	37	5	36	4	0,01
2.12	37	7	35	7	0,05
2.13	38	8	33	8	0,01
2.14	42	3	40	5	0,05
2.15	40	2	34	4	0,01

2.16. Знайти закон розподілу величини X – кількості відвідувачів ресторану швидкого харчування за даними таблиці 2.17.

Таблиця 2.17

Час роботи	9.00 – 11.00	11.00 – 13.00	13.00 – 15.00	15.00 – 17.00	17.00 – 19.00	19.00 – 21.00
Кількість відвідувачів	25	53	68	73	52	39

4. Дисперсійний аналіз. Varianciaanalízis

Дисперсійний аналіз призначений для дослідження задачі про вплив на величину, що вимірюється, одного чи декількох факторів. Причому в однофакторному, двофакторному і так далі аналізі фактори, що впливають на результат, вважаються відомими, і необхідно виявити тільки суттєвість цього впливу та оцінити цей вплив. Так, наприклад, досліджується вплив ваги спеціального браслета, що одягається на зап'ястя, на частоту мимовільного тремтіння м'язів рук – тремору.

Дисперсійний аналіз. Визначення та основні поняття дисперсійного аналізу в біометрії. Багатофакторний дисперсійний аналіз. Ефекти взаємодій між факторами при багатофакторному дисперсійному аналізі в біометрії.

Регресійний аналіз. Поняття регресії, види регресії. Проста і множинна лінійна модель регресійного аналізу. Знаходження оцінок невідомих параметрів регресії, перевірка їх достовірності, перевірка адекватності моделі. Довірчий інтервал для передбачених значень. Покрокова регресія. Нелінійна регресія.

Кластерний аналіз. Метод головних компонент та факторний аналіз. Основна мета та застосування факторного аналізу в біометрії. Факторний аналіз як метод редукції даних в біометрії. Факторний аналіз як метод класифікації в біометрії.

Для проведення дисперсійного аналізу необхідно:

1 Ввести дані в таблицю таким чином, щоб у кожному стовпці були дані, які відповідають одному значенню фактора, що досліджується. Від стовпця до стовпця фактор, що досліджується, повинен або зростати, або зменшуватись.
2 Вибрати ***Сервис →Анализ данных***.
3 Вибрати зі списку ***Инструменты анализа*** рядок ***Однофакторный дисперсионный анализ***. Натиснути кнопку ***ОК***.
4 У діалоговому вікні, що з'явилося, зазначити ***Входной интервал*****:** , тобто посилання на діапазон даних.
5 Зазначити вихідний діапазон, тобто ввести посилання на комірки, в які буде виведено результат аналізу. Для цього слід поставити перемикач у положення ***Выходной интервал*****:** і у полі праворуч зазначити вихідний діапазон.

6 Встановити перемикач у розділі ***Группирование*****:** у положення ***по столбцам*** (***по строкам***).

7 Натиснути кнопку ***ОК***.

В результаті вихідний діапазон буде містити результати дисперсійного аналізу: середнє, дисперсії, критерій Фішера та інші показники.

Вплив досліджуваного фактора визначається за величиною значущості критерію Фішера, який знаходиться в таблиці «***Дисперсионный анализ***» на перетині рядка «***Между группами***» та стовпця «***Р-Значение»***. У випадках коли ***Р-Значение*** < 0,95, критерій Фішера є значущим та вплив досліджуваного фактора можна вважати доведеним.

Задача.

Визначити, чи вливає рівень шуму на правильність розпізнання емоційної складової мови у дітей за результатами, які були отримані під час дослідження вікових змін слухової функції у дітей

Відношення сигнал / шум	**Без шуму**	**6 дБ**	**12 дБ**
Відсоток правильних відповідей	78,6	61,9	45,2
	95,2	97,6	97,6
	83,3	61,9	80,9
	85,7	73,8	62,4
	80,4	75,6	70,6
	90,2	68,8	69,2

Хід виконання

Ввести дані з таблиці до електронної таблиці. Оформити їх у вигляді таблиці :

	A	**B**	**C**
1	**Без шуму**	**6 дБ**	**12 дБ**
2	78,6	61,9	45,2
3	95,2	97,6	97,6
4	83,3	61,9	80,9
5	85,7	73,8	62,4
6	80,4	75,6	70,6
7	90,2	68,8	69,2

Зайти в пункт меню ***Сервис***, вибрати команду ***Анализ данных***.
У вікні, що з'явилося, зі списку ***Инструменты анализа*** вибрати рядок ***Однофакторный дисперсионный анализ***. Натиснути кнопку ***ОК***.
У діалоговому вікні, що з'явилося, у полі ***Входной интервал*** **:** зазначити діапазон залежних даних **A2:C7**; встановити перемикач у положення ***Выходной интервал*****:** та зазначити комірку **E1**.
Встановити перемикач у розділі ***Группирование*****:** у положення ***по столбцам***.
Натиснути ***ОК.***

У таблиці «***Дисперсионный анализ***» на перетині рядків «***Между группами***» та стовпця «***P-Значение***» знаходиться величина 0,15529. Величина ***P-Значение*** > 0,05, звідки критерій Фішера є незначущим та вплив фактора шуму на ефективність розпізнання довести не вдалося.

5. Приклади здійснення кореляційного та регресійного аналізу

При вивченні стохастичних зв'язків між різними ознаками економічного об'єкта головною задачею є встановлення виду кореляційної залежності результативної ознаки (*Y*) від факторної (*X*), тобто виду функціональної залежності $\overline{Y}=f(X)$. В першу чергу це пов'язано з необхідністю прогнозування досліджуваних процесів. Математико-статистичний апарат, що дозволяє встановити вид кореляційної залежності називається **регресійним аналізом**, а функція, що описує цю залежність, називається **рівнянням регресії**.

Встановлення виду кореляційної залежності

Регресійний аналіз проводиться за такими етапами:

1) Встановлення виду кореляційної залежності результативної ознаки *Y* від факторної ознаки *X*.
2) Побудова регресійної моделі.
3) Перевірка статистичної значущості побудованої моделі.

Перший етап регресійного аналізу є найважливішим, оскільки помилки у виборі виду залежності призводять до побудови регресійної моделі, що не відповідає емпіричним даним і не може використовуватися для прогнозування.

Вибіркові дані для вивчення кореляційного зв'язку між ознаками *X* та *Y* зазвичай мають вигляд пар їх значень: $(x_1;y_1)$, $(x_2;y_2)$, …, $(x_n;y_n)$, x_i – значення величини *X*, y_i – значення *Y*, *n* – кількість пар значень, $i=\overline{1,n}$. Якщо їх кількість достатньо велика, то для зручності розрахунків дані групуються і будується статистичний ряд, що містить значення *X*, відповідні середні значення *Y* та частоти (табл. 4.1).

Таблиця 4.1

x_i	x_1	x_2	…	x_k
$\overline{y}_{x_i}$	$\overline{y}_{x_1}$	$\overline{y}_{x_2}$	…	$\overline{y}_{x_k}$

n_i	n_1	n_2	…	n_k

Згруповані дані (табл. 4.1) зображуються графічно, що часто дозволяє визначити вид залежності *Y* від *X*.

Ламана лінія, що сполучає крапки з координатами $(x_i; \overline{y_{x_i}})$, називається **емпіричною лінією регресії**.

Якщо емпірична лінія регресії значно наближається до прямої лінії, то висувається гіпотеза про наявність лінійного зв'язку між досліджуваними ознаками (рис. 4.1).

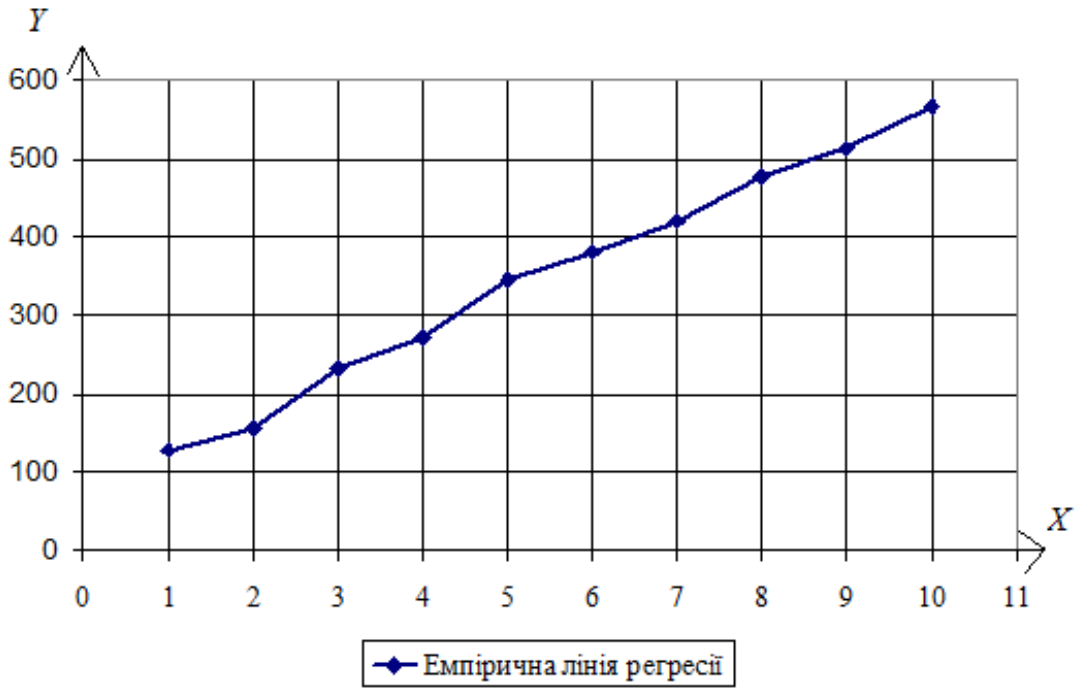


Рис. 4.1. Гіпотетична лінійна залежність

В іншому випадку висувається гіпотеза про наявність нелінійного зв'язку (рис.4.2).

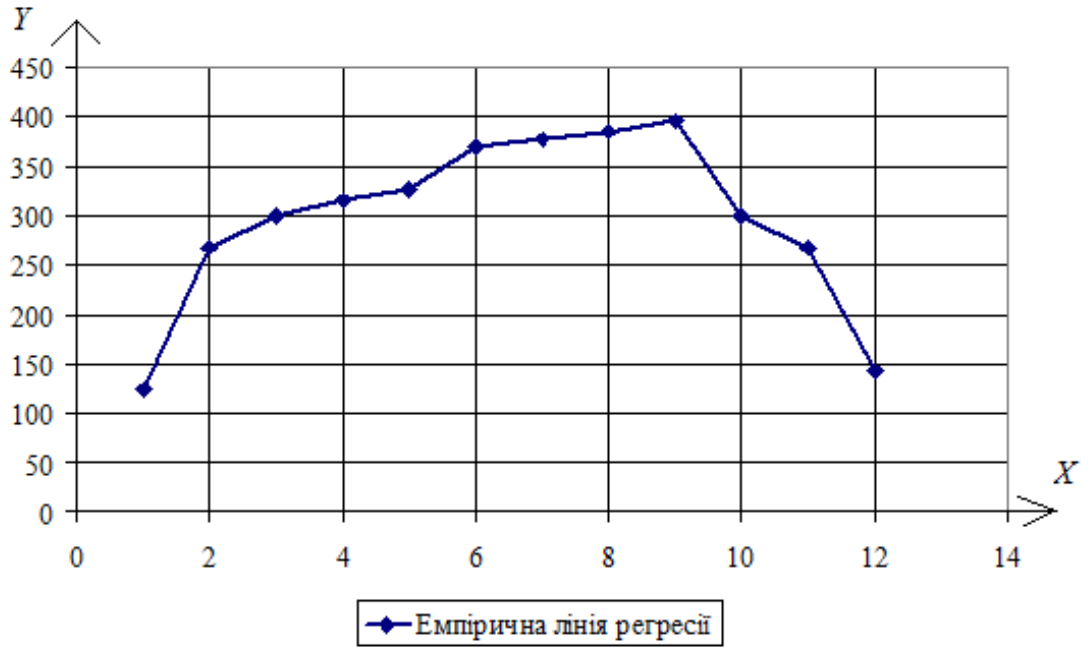


Рис. 4.2. Гіпотетична нелінійна залежність

Лінійна регресія

Якщо висунуто гіпотезу про наявність лінійної залежності результативної ознаки (*Y*) від факторної (*X*), то рівняння регресії має вид:

$$\overline{y_x} = ax + b\,, \qquad (4.1)$$

де a, b - параметри моделі.

Побудова лінійної регресійної моделі – це знаходження параметрів рівняння (4.1). Параметри рівняння регресії зазвичай знаходяться за **методом найменших квадратів**.

Ідея методу найменших квадратів

Нехай при вивчення залежності *Y* від *X* було отримано вибіркові дані: $x_1, x_2, \ldots, x_n$ – значення величини *X*, $y_1, y_2, \ldots, y_n$ – відповідні значення *Y*. За

вибірковими даними було побудовано рівняння регресії $y = ax + b$. Якщо в рівняння підставити замість *x* значення $x_1, x_2, ..., x_n$, то будуть отримані теоретичні значення *Y*: $y_{1,теор}, y_{2,теор}, ..., y_{n,теор}$, які відрізняються від $y_1, y_2, ..., y_n$. Різниця значень $y_{i,теор} - y_i$ називається помилкою регресійної моделі і позначається e_i. Якщо параметри рівняння підбираються так, щоб сума квадратів помилок була мінімальною, то говорять, що вони отримані за методом найменших квадратів.

У випадку лінійної регресії параметри рівняння регресії за методом найменших квадратів знаходяться з системи лінійних алгебраїчних рівнянь:

$$\begin{cases} a\sum_{i=1}^{k} x_i^2 n_i + b\sum_{i=1}^{k} x_i n_i = \sum_{i=1}^{k} x_i n_i \overline{y_{x_i}} \\ a\sum_{i=1}^{k} x_i n_i + b\sum_{i=1}^{k} n_i = \sum_{i=1}^{k} n_i \overline{y_{x_i}} \end{cases} \quad (4.2)$$

Якщо вибіркові дані не згруповані, то система (4.1) значно спрощується:

$$\begin{cases} a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \\ a\sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} y_i \end{cases} \quad (4.3)$$

Перевірка правильності побудови рівняння регресії здійснюється за основним варіаційним рівнянням:

$$Q = Q_p + Q_o, \quad (4.4)$$

де $Q = \sum_{i=1}^{k} \left(\overline{y_{x_i}} - \overline{y}\right)^2 n_i$ - загальна варіація, тобто сума квадратів відхилень емпіричних значень *Y* від середнього, $\overline{y} = \frac{\sum_{i=1}^{k} \overline{y_{x_i}} n_i}{n}$;

$Q_p = \sum_{i=1}^{k} \left(y_{i,теор} - \overline{y} \right)^2 n_i$ - варіація регресії, тобто сума квадратів відхилень теоретичних значень *Y* від середнього, що обумовлена регресією;

$Q_o = \sum_{i=1}^{k} \left(y_{i,теор} - \overline{y}_{x_i} \right)^2 n_i$ - варіація залишків, тобто сума квадратів відхилень теоретичних значень *Y* від емпіричних.

У випадку незгрупованих даних загальна варіація, варіації регресії і залишків знаходяться за формулами: $Q = \sum_{i=1}^{n} \left(y_i - \overline{y} \right)^2$; $Q_p = \sum_{i=1}^{n} \left(y_{i,\text{теор}} - \overline{y} \right)^2$; $Q_o = \sum_{i=1}^{n} \left(y_{i,\text{теор}} - y_i \right)^2$; а середнє значення за формулою $\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$.

Для перевірки статистичної значущості рівняння регресії розраховується F-статистика за формулою:

$$F = \frac{Q_p (n-l)}{Q_o (l-1)}, \tag{4.5}$$

де *n* – кількість наглядів, *l* – кількість груп у кореляційній таблиці або кількість параметрів моделі у випадку незгрупованих даних. Розраховане значення F-статистики порівнюється з критичним значенням $F_{кр}$ розподілу Фішера, яке можна знайти за статистичними таблицями або за допомогою вбудованої функції Excel $FРАСПОБР(\alpha, k_1, k_2)$, де $k_1 = l - 1;\ \ k_2 = n - l$ - ступені волі, α - рівень значущості.

Адекватність моделі вибірковим даним можна оцінити за коефіцієнтом детермінації R^2, що показує частину варіації значень результативної ознаки *Y*, що пояснюється рівнянням регресії. Коефіцієнт детермінації розраховується за формулою:

$$R^2 = 1 - \frac{Q_o}{Q} = \frac{Q_p}{Q}. \qquad (4.6)$$

Значення коефіцієнта детермінації знаходяться в інтервалі $[0;1]$, тобто $0 \le R^2 \le 1$. Чим ближче R^2 до 1, тим краще отримане рівняння регресії пояснює поведінку результативної ознаки. Наприклад, якщо R^2=0,98, то 98% варіації результативної ознаки *Y* пояснюється рівнянням регресії.

ПРИКЛАД Побудувати регресійну модель, що описує залежність сумарних виробничих затрат *Y* (тис. грн.) від об'ємів виробництва *X* (тис. од.). Відповідні статистичні дані надано у таблиці 4.2.

Таблиця 4.2

X	41	44	52	57	59	64	68	70	73	75
Y	670	657	713	736	778	812	833	876	911	932

Розв'язок. В таблиці 4.2 надано вибіркові дані: значення $x_i, i = 1, n$ величини *X* та відповідні значення $y_i, i = 1, n$; кількість пар – $n = 10$ невелика, тому для проведення регресійного аналізу їх можна не групувати.

Перший етап аналізу: визначимо вид залежності *Y* від *X*. Побудуємо емпіричну лінію регресії (рис. 4.3).

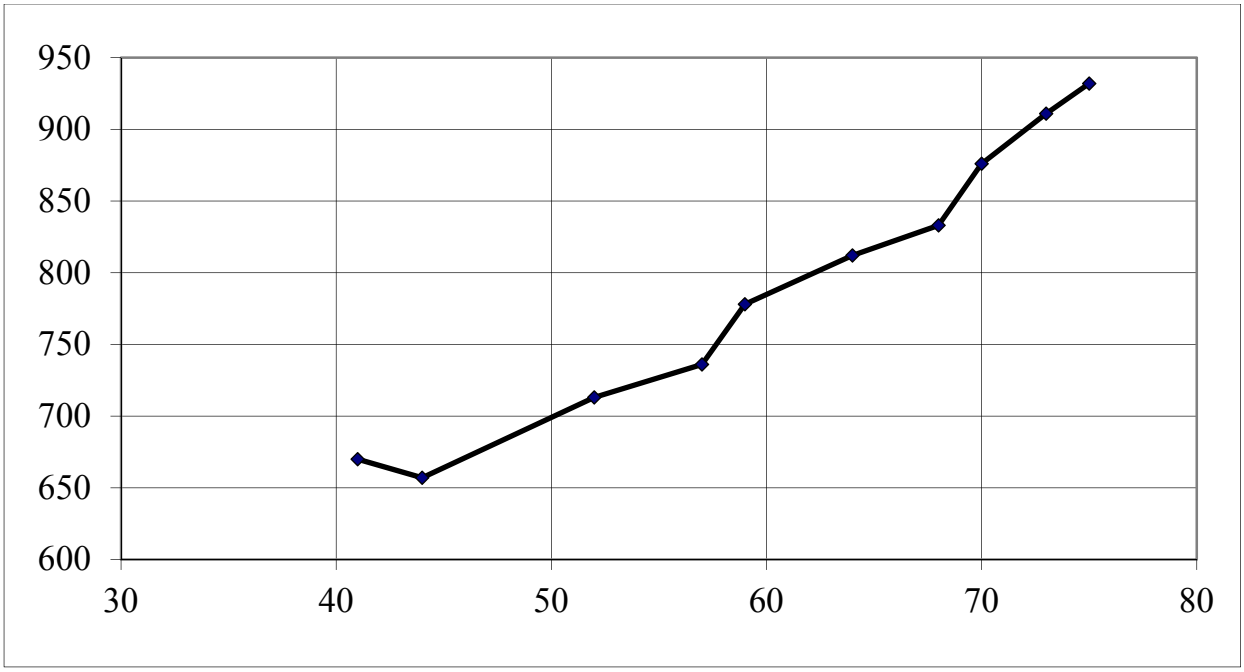


Рис. 4.3. Емпірична лінія регресії

Оскільки емпірична лінія регресії наближається до прямої лінії, то висуваємо гіпотезу про лінійну залежність *Y* від *X*, тобто рівняння регресії будемо шукати у вигляді $y = ax + b$.

Другий етап: знайдемо параметри a, b рівняння регресії, для чого складемо систему (4.3) для даних, що не згруповані. Необхідні розрахунки для зручності оформимо у вигляді таблиці (табл. 4.3).

Таблиця 4.3

Розрахункова таблиця											Суми
x_i	41	44	52	57	59	64	68	70	73	75	603
y_i	670	657	713	736	778	812	833	876	911	932	7918
x_i^2	1681	1936	2704	3249	3481	4096	4624	4900	5329	5625	37625
$x_i y_i$	27470	28908	37076	41952	45902	51968	56644	61320	66503	69900	487643

Отже, складемо систему для знаходження параметрів рівняння регресії та розв’яжемо її за правилом Крамера:

$$\begin{cases} a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \\ a\sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} y_i \end{cases} \Rightarrow \begin{cases} 37625a + 603b = 487643 \\ 603a + 10b = 7918 \end{cases}.$$

Знайдемо головний визначник системи, що складений із коефіцієнтів перед невідомими: $\Delta = \begin{vmatrix} 37625 & 603 \\ 603 & 10 \end{vmatrix} = 37625 \cdot 10 - 603^2 = 12641.$

Знайдемо допоміжні визначники, що отримуються із головного заміною відповідного стовпця коефіцієнтів на стовпець вільних членів:

$$\Delta a = \begin{vmatrix} 487643 & 603 \\ 7918 & 10 \end{vmatrix} = 487643 \cdot 10 - 603 \cdot 7918 = 101876;$$

$$\Delta b = \begin{vmatrix} 37626 & 487643 \\ 603 & 7918 \end{vmatrix} = 37626 \cdot 7918 - 48743 \cdot 603 = 3866021.$$

Знайдемо невідомі за формулами Крамера:

$$a = \frac{\Delta a}{\Delta} = \frac{101876}{12641} \approx 8{,}06; \quad b = \frac{\Delta b}{\Delta} = \frac{3866021}{12641} \approx 305{,}83.$$

Отже, шукане рівняння регресії має вигляд $y = 8{,}06x - 305{,}83$.

Третій етап: перевіримо правильність побудови моделі за рівнянням (4.4), її статистичну значущість за F-статистикою (4.5) і адекватність вибірковим даним за коефіцієнтом детермінації (4.6). Для чого знайдемо загальну варіацію, варіації регресії та залишків; необхідні розрахунки оформимо у вигляді таблиці (табл. 4.4).

Передусім знайдемо $\overline{y}$: $\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n} \approx 791{,}8$.

Таблиця 4.4

x_i	y_i	$y_{i,\text{теор}}$	$\left(y_i - \overline{y}\right)^2$	$\left(y_{i,\text{теор}} - \overline{y}\right)^2$	$\left(y_{i,\text{теор}} - y_i\right)^2$

41	670	636,26	14835,24	24193,32	1138,52
44	657	660,44	18171,04	17256,64	11,80
52	713	724,91	6209,44	4474,42	141,82
57	736	765,20	3113,64	707,31	852,92
59	778	781,32	190,44	109,77	11,04
64	812	821,62	408,04	889,17	92,52
68	833	853,86	1697,44	3850,90	434,96
70	876	869,97	7089,64	6111,17	36,31
73	911	894,15	14208,64	10475,83	283,87
75	932	910,27	19656,04	14035,10	472,20
		Суми	85579,6	82103,63	3475,974

Отже, $Q = 85579,6$; $Q_p = 82103,63$; $Q_o = 3475,974$; тоді основне варіаційне рівняння $Q = Q_p + Q_o$ для побудованої моделі має вигляд: $85579,6 = 82103,63 + 3475,974$ і є тотожністю, тому рівняння регресії побудовано правильно.

Для перевірки статистичної значущості рівняння регресії знайдемо F-статистику, враховуючи, що *n*=10, *l*=2 – оскільки шукали рівняння з двома параметрами:

$$F = \frac{Q_p(n-l)}{Q_o(l-1)} = \frac{82103(10-2)}{3475,974(2-1)} \approx 188,96.$$

Знайдемо $F_{кр}$: $F_{кр} = FРАСПОБР(0,001;2-1;10-2) \approx 25,41$. Розраховане значення F-статистики більше критичного, тому регресійна модель є статистично значущою на рівні 0,001.

Знайдемо коефіцієнт детермінації R^2: $R^2 = \frac{Q_p}{Q} = \frac{82103,63}{85579,6} \approx 0,96$.

Значення коефіцієнта детермінації свідчить, що 96% варіації результативної ознаки *Y* пояснюються рівнянням регресії.

Висновок: Сумарні виробничі затрати *Y* (тис. грн.) лінійно залежать від об'єму виробництва *X* (тис. од.). Залежність описується рівнянням $y = 8,06x - 305,83$, яке є статистично значимим на рівні значущості 0,001 та описує 96% вибіркових даних.

Нелінійна регресія

Якщо висунуто гіпотезу про наявність нелінійної залежності результативної ознаки (*Y*) від факторної (*X*), то регресійний аналіз проводиться за тими ж етапами, як й у випадку лінійної залежності. Вид рівнянь регресії і системи для знаходження їх параметрів для нелінійних залежностей, що найчастіше зустрічаються, надано у таблиці 4.5.

Таблиця 4.5

<table>
<tr><td colspan="2">Рівняння параболічної регресії:
$\overline{y_x} = ax^2 + bx + c$</td></tr>
<tr><td colspan="2">Система для знаходження параметрів</td></tr>
<tr><td>для згрупованих вибіркових даних:
$\begin{cases} a\sum_{i=1}^{k} x_i^4 n_i + b\sum_{i=1}^{k} x_i^3 n_i + c\sum_{i=1}^{k} x_i^2 n_i = \sum_{i=1}^{k} x_i^2 n_i \overline{y_x} \\ a\sum_{i=1}^{k} x_i^3 n_i + b\sum_{i=1}^{k} x_i^2 n_i + c\sum_{i=1}^{k} x_i n_i = \sum_{i=1}^{k} x_i n_i \overline{y_{x_i}} \\ a\sum_{i=1}^{k} x_i^2 n_i + b\sum_{i=1}^{k} x_i n_i + c\sum_{i=1}^{k} n_i = \sum_{i=1}^{k} n_i \overline{y_{x_i}} \end{cases}$</td><td>для незгрупованих вибіркових даних:
$\begin{cases} a\sum_{i=1}^{n} x_i^4 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i^2 y_i \\ a\sum_{i=1}^{n} x_i^3 + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \\ a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i + cn = \sum_{i=1}^{n} y_i \end{cases}$</td></tr>
<tr><td colspan="2">Рівняння гіперболічної регресії:
$\overline{y_x} = \frac{a}{x} + b$</td></tr>
</table>

Система для знаходження параметрів	
для згрупованих вибіркових даних:	для незгрупованих вибіркових даних:
$\begin{cases} a\sum\limits_{i=1}^{k}\frac{1}{x_i^2}n_i + b\sum\limits_{i=1}^{k}\frac{1}{x_i}n_i = \sum\limits_{i=1}^{k}\frac{1}{x_i}\overline{y_{x_i}}n_i \\ a\sum\limits_{i=1}^{k}\frac{1}{x_i}n_i + b\sum\limits_{i=1}^{k}n_i = \sum\limits_{i=1}^{k}\overline{y_{x_i}}n_i \end{cases}$	$\begin{cases} a\sum\limits_{i=1}^{n}\frac{1}{x_i^2} + b\sum\limits_{i=1}^{n}\frac{1}{x_i} = \sum\limits_{i=1}^{n}\frac{1}{x_i}y_i \\ a\sum\limits_{i=1}^{n}\frac{1}{x_i} + bn = \sum\limits_{i=1}^{n}y_i \end{cases}$
Рівняння показникової регресії: $\overline{y_x} = ba^x$	
для згрупованих вибіркових даних:	для незгрупованих вибіркових даних:
$\begin{cases} \lg a\sum\limits_{i=1}^{k}x_i^2 n_i + \lg b\sum\limits_{i=1}^{k}x_i n_i = \sum\limits_{i=1}^{k}x_i n_i \lg\overline{y_{x_i}} \\ \lg a\sum\limits_{i=1}^{k}x_i n_i + \lg b\sum\limits_{i=1}^{k}n_i = \sum\limits_{i=1}^{k}n_i \lg\overline{y_{x_i}} \end{cases}$	$\begin{cases} \lg a\sum\limits_{i=1}^{n}x_i^2 + \lg b\sum\limits_{i=1}^{n}x_i = \sum\limits_{i=1}^{n}x_i \lg y_i \\ \lg a\sum\limits_{i=1}^{n}x_i + n\lg b = \sum\limits_{i=1}^{n}\lg y_i \end{cases}$

Перевірка статистичної значущості нелінійної регресійної моделі також здійснюється за F-статистикою. При цьому для параболічної регресії кількість параметрів *l*=3, для гіперболічної і показникової – *l*=2.

ПРИКЛАД Дано розподіл однотипових підприємств за об’ємом виробництва *X* (тис. од.) і собівартістю одиниці продукції *Y* (грн.) (табл. 4.6). Знайти регресійну модель, що описує собівартості продукції від об’єму виробництва.

Таблиця 4.6

Y / *X*	10	15	20	25
25	-	-	1	2

50	-	2	2	-
75	-	5	3	1
100	1	3	-	-
125	3	1	1	-

Розв'язок. Для проведення регресійного аналізу за даними таблиці 4.6 побудуємо кореляційну таблицю (табл. 4.7).

Таблиця 4.7

x_i / y_j	25	50	75	100	125	n_j
10	0	0	0	1	3	4
15	0	2	5	3	1	11
20	1	2	3	0	1	7
25	2	0	1	0	0	3
n_i	3	4	9	4	5	$n=25$

За даними кореляційної таблиці побудуємо ряд, що відображає залежність середнього значення *Y* від *X* (табл. 3.2), для чого знайдемо середні значення $\overline{y}_{x_i}$ для кожного значення x_i, $i=\overline{1,5}$, і заповнимо таблицю 4.8:

$$\overline{y}_{x_1}=\frac{y_1n_{11}+y_2n_{12}+y_3n_{13}+y_4n_{14}}{n_1}=\frac{20\cdot1+25\cdot2}{3}\approx23{,}33;\ \overline{y}_{x_2}=\frac{15\cdot2+20\cdot2}{4}=17{,}5;$$

$$\overline{y}_{x_3}=\frac{15\cdot5+20\cdot3+25\cdot1}{9}=17{,}78;\ \overline{y}_{x_4}=\frac{10\cdot1+15\cdot3}{4}=13{,}75;$$

$$\overline{y}_{x_5}=\frac{10\cdot3+15\cdot1+20\cdot1}{5}=13.$$

Таблиця 4.8

x_i	25	50	75	100	125
$\overline{y}_{x_i}$	23,33	17,5	17,78	13,75	13
n_i	3	4	9	4	5

Перший етап аналізу: визначимо вид залежності *Y* від *X*. Побудуємо емпіричну лінію регресії (рис. 4.4).

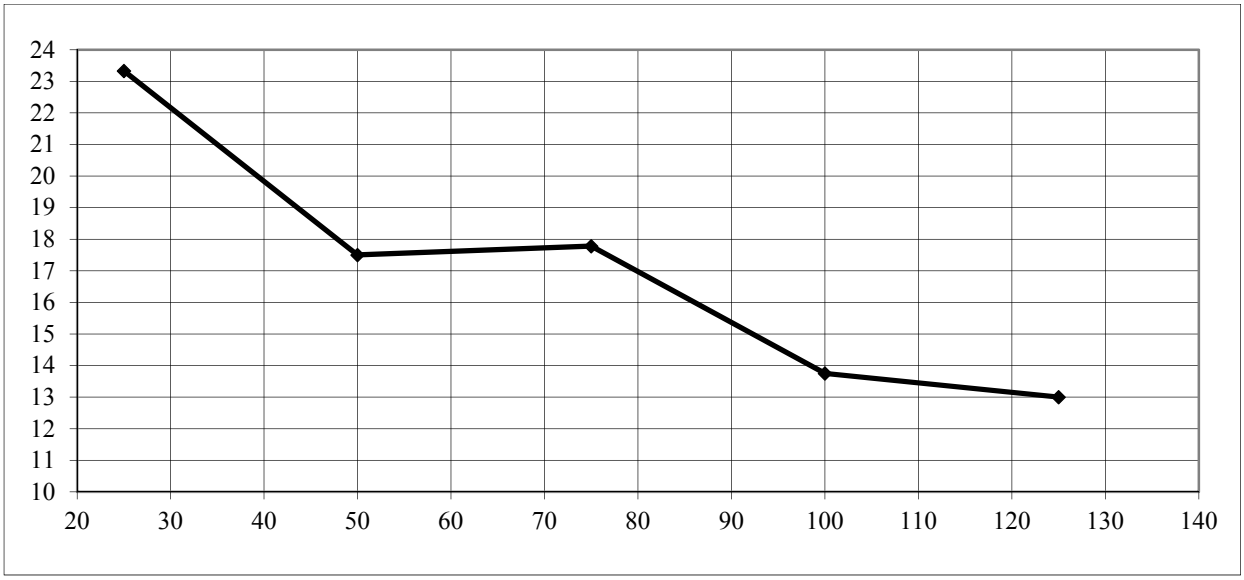


Рис. 4.4. Емпірична лінія регресії

Оскільки емпірична лінія регресії наближається до гіперболи, то висуваємо гіпотезу про гіперболічну залежність *Y* від *X*, тобто рівняння регресії будемо шукати у вигляді $\overline{y_x} = \frac{a}{x} + b$.

Другий етап: знайдемо параметри a, b рівняння регресії, для чого складемо систему для згрупованих даних. Необхідні розрахунки для зручності оформимо у вигляді таблиці (табл. 4.9), в останньому рядку якої знайдемо відповідні стовпцям суми.

Таблиця 4.9

x_i	$\overline{y}_{x_i}$	n_i	$\frac{1}{x_i}$	$\frac{1}{x_i} n_i$	$\frac{1}{x_i^2} n_i$	$\overline{y}_{x_i} n_i$	$\frac{1}{x_i} \overline{y}_{x_i} n_i$

25	23,33	3	0,04	0,12	0,0048	69,99	2,7996
50	17,5	4	0,02	0,08	0,0016	70	1,4
75	17,78	9	0,0133	0,12	0,0016	160,02	2,1336
100	13,75	4	0,01	0,04	0,0004	55	0,55
125	13	5	0,008	0,04	0,0003	65	0,52
			Суми	0,4	0,0087	420,01	7,4032

Отже, складемо систему для знаходження параметрів рівняння регресії та розв'яжемо її за правилом Крамера:

$$\begin{cases} a\sum\limits_{i=1}^{k}\frac{1}{x_i^2}n_i + b\sum\limits_{i=1}^{k}\frac{1}{x_i}n_i = \sum\limits_{i=1}^{k}\frac{1}{x_i}\overline{y_{x_i}}n_i \\ a\sum\limits_{i=1}^{k}\frac{1}{x_i}n_i + b\sum\limits_{i=1}^{k}n_i = \sum\limits_{i=1}^{k}\overline{y_{x_i}}n_i \end{cases} \Rightarrow \begin{cases} 0,0087a + 0,4b = 7,4032 \\ 0,4a + 25b = 420,01 \end{cases}.$$

Головний визначник системи:

$$\Delta = \begin{vmatrix} 0,00872 & 0,4 \\ 0,4 & 25 \end{vmatrix} = 0,00872 \cdot 25 - 0,4^2 = 0,058.$$

Допоміжні визначники: $\Delta a = \begin{vmatrix} 7,4032 & 0,4 \\ 420 & 25 \end{vmatrix} = 7,4032 \cdot 25 - 0,4 \cdot 420 = 17,076;$

$$\Delta b = \begin{vmatrix} 0,00872 & 7,4032 \\ 0,4 & 420 \end{vmatrix} = 0,00872 \cdot 420 - 7,4032 \cdot 0,4 = 0,701207.$$

Формули Крамера: $a = \frac{\Delta a}{\Delta} = \frac{17,076}{0,058} \approx 294,41$; $b = \frac{\Delta b}{\Delta} = \frac{0,7012207}{0,058} \approx 12,09$.

Отже, шукане рівняння регресії має вигляд $\overline{y_x} = \frac{294,41}{x} + 12,09$.

Третій етап: перевіримо правильність побудови моделі за рівнянням (4.4), її статистичну значущість за F-статистикою (4.5) і адекватність вибірковим даним за коефіцієнтом детермінації (4.6). Для чого знайдемо

загальну варіацію, варіації регресії та залишків; необхідні розрахунки оформимо у вигляді таблиці (табл. 4.10).

Знайдемо $\overline{y}$: $\overline{y} = \frac{\sum_{i=1}^{k} \overline{y_{x_i}} n_i}{n} \approx 16,8$.

Таблиця 4.4

x_i	y_i	n_i	$y_{i,\text{теор}}$	$\frac{\left(y_i - \overline{y}\right)^2}{n_i}$	$\frac{\left(y_{i,\text{теор}} - \overline{y}\right)^2}{n_i}$	$\frac{\left(y_{i,\text{теор}} - y_i\right)^2}{n_i}$
25	23,33	3	23,866	127,907	149,782	0,863
50	17,5	4	17,978	1,958	5,547	0,914
75	17,78	9	16,015	8,637	5,547	28,028
100	13,75	4	15,034	37,220	12,482	6,594
125	13	5	14,445	72,215	27,737	10,441
			Суми	247,936	201,096	46,840

Отже, $Q = 247,936$; $Q_p = 201,096$; $Q_o = 46,840$; тоді основне варіаційне рівняння $Q = Q_p + Q_o$ для побудованої моделі має вигляд: $247,936 = 201,096 + 46,840$ і є тотожністю, тому рівняння регресії побудовано правильно.

Для перевірки статистичної значущості рівняння регресії знайдемо F-статистику, враховуючи, що *n*=25, *l*=2 – оскільки шукали рівняння з двома параметрами:

$$F = \frac{Q_p\left(n-l\right)}{Q_o\left(l-1\right)} = \frac{201,096\left(25-2\right)}{46,840\left(2-1\right)} \approx 98,75.$$

Знайдемо $F_{кр}$: $F_{кр} = FРАСПОБР\left(0,001;2-1;25-2\right) \approx 14,20$. Розраховане значення F-статистики більше критичного, тому регресійна модель є статистично значущою на рівні 0,001.

Знайдемо коефіцієнт детермінації R^2: $R^2 = \frac{Q_p}{Q} = \frac{201,096}{247,936} \approx 0,81$.

Значення коефіцієнта детермінації свідчить, що 81% варіації результативної ознаки *Y* пояснюється рівнянням регресії.

Висновок: Залежність собівартості одиниці продукції *Y* (грн.) від об'єму виробництва *X* (тис. од.) описується рівнянням $\overline{y_x} = \frac{294,41}{x} + 12,09$, яке є статистично значимим на рівні значущості 0,001 та описує 81% вибіркових даних.

Регресія у Microsoft Excel

Пакет аналізу даних Microsoft Excel надає можливість будувати регресійні моделі, але тільки у випадку лінійної залежності результативної ознаки *Y* від факторної ознаки *X* і тільки для незгрупованих вибіркових даних.

Для побудови лінійної регресійної моделі необхідно:

1) Викликати ***Сервис – Анализ данных – Регрессия – ОК***. З'явиться вікно для надання вхідних даних (рис. 4.5).

2) У графі ***Входной интервал Y*** та ***Входной интервал X*** вказати відповідні стовпці даних; у графі ***Выходной интервал*** вказати ту клітину, починаючи з якої будуть надаватися вихідні дані – параметри рівняння регресії та результати її статистичного аналізу.

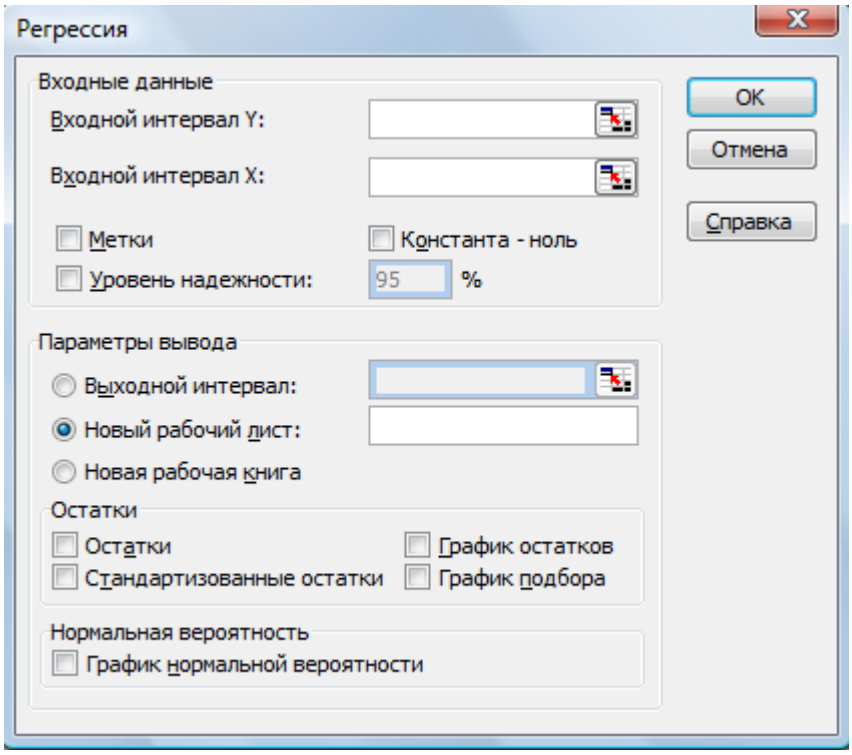


Рис. 4.5. Діалогове вікно функції ***Регрессия***

Приклад і результати роботи функції ***Регрессия*** надано на рис. 4.6.

Файл Правка Вид Вставка Формат Сервис Данные Окно Справка

I23

	A	B	C	D	E	F	G	H	I	J
1					**Регресійний аналіз**					
2		Вхідні дані			Вихідні дані					
3	№	Значення			ВЫВОД ИТОГОВ					
4	**i**	**X**	**Y**							
5	1	1	328		*Регрессионная статистика*					
6	2	2	329		Множественный R	0,972633354				
7	3	3	329		R-квадрат	0,946015642				
8	4	4	345		Нормированный R-квадрат	0,937018249				
9	5	5	352		Стандартная ошибка	5,786032717				
10	6	6	370		Наблюдения	8				
11	7	7	377							
12	8	8	385		Дисперсионный анализ					
13						*df*	*SS*	*MS*	*F*	*Значимость F*
14					Регрессия	1	3520,005952	3520,005952	105,1433059	5,01935E-05
15					Остаток	6	200,8690476	33,4781746		
16					Итого	7	3720,875			
17										
18						*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	
19					Y-пересечение	310,6785714	4,508440371	68,91043151	6,28282E-10	
20					Переменная X 1	9,154761905	0,892804231	10,25394099	5,01935E-05	

Рис. 4.6. Результати регресійного аналізу

В таблиці (рис. 4.6) у графі ***Коэффициенты*** вказані значення параметрів моделі a та b: b - в графі **Y-пересечение**, a - в графі **Переменная X1**. Отже, побудована лінійна регресійна модель має вигляд:

$$y = 69{,}15x + 310{,}68.$$

Для перевірки статистичної значущості моделі надається значення F-статистики у графі *F* : *F* = 105,14.

Коефіцієнт детермінації моделі R^2 надається у графі **R-квадрат,** R^2=0,97.

Крім того, може бути надано: графік підбору – порівняльна діаграма, що містить емпіричну і теоретичну лінії регресії; таблиця залишків – різниць емпіричних і теоретичних значень *Y* (рис. 4.7).

Регресійний аналіз

№	Вхідні дані	
i	X	Y
1	1	328
2	2	329
3	3	329
4	4	345
5	5	352
6	6	370
7	7	377
8	8	385

Вихідні дані

ВЫВОД ОСТАТКА

Наблюдение	Предсказанное Y	Остатки
1	319,8333333	8,166666667
2	328,9880952	0,011904762
3	338,1428571	-9,142857143
4	347,297619	-2,297619048
5	356,452381	-4,452380952
6	365,6071429	4,392857143
7	374,7619048	2,238095238
8	383,9166667	1,083333333

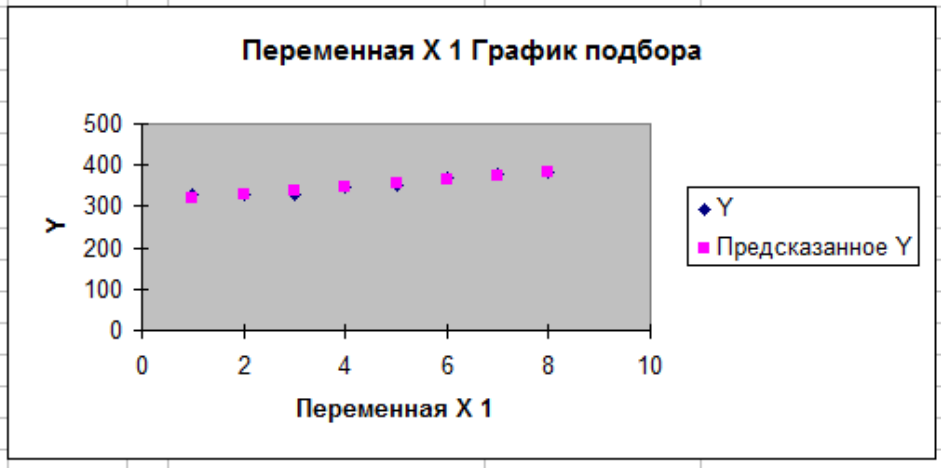


Рис. 4.7. Додаткові результати регресійного аналізу

Завдання для самостійного виконання

4.1. Відомі дані про обсяг виробництва сільськогосподарської продукції (грн.) на 1 особу АР Крим (табл. 4.11). Побудувати регресійну модель за даними таблиці, оцінити її статистичну значущість та адекватність.

Таблиця 4.11

Рік	2001	2002	2003	2004	2005	2006	2007
Обсяг виробництва с/г продукції	1000	995	949	926	1049	1112	1596

4.2. Відомі дані про чисельність науково та науково-технічних працівників, що припадають на 1000 осіб (табл. 4.12). Побудувати регресійну модель за даними таблиці, оцінити її статистичну значущість та адекватність.

Таблиця 4.12

Рік	2001	2002	2003	2004	2005	2006	2007
Чисельність науково та науково-технічних працівників	1,1	1,1	1,0	1,0	1,0	1,0	0,9

4.3. Відомі дані про інвестиції в основний капітал в розрахунку на 1 особу (табл. 4.13). Побудувати регресійну модель за даними таблиці, оцінити її статистичну значущість та адекватність.

Таблиця 4.13

Рік	2001	2002	2003	2004	2005	2006	2007
Інвестиції в основний капітал	376,8	600,2	735,7	955,2	1376,2	1704,1	2375,6

4.4. Відомі дані про обсяг інноваційної продукції в розрахунку на 1 особу (табл. 4.14). Побудувати регресійну модель за даними таблиці, оцінити її статистичну значущість та адекватність.

Таблиця 4.14

Рік	2001	2002	2003	2004	2005	2006	2007
Обсяг інноваційної продукції	38,4	139,0	264,5	172,3	313,1	469,9	282,2

4.5. Відомі дані про обсяг експорту товарів в розрахунку на 1 особу (табл. 4.15). Побудувати регресійну модель за даними таблиці, оцінити її статистичну значущість та адекватність.

Таблиця 4.15

Рік	2001	2002	2003	2004	2005	2006	2007
Обсяг експорту товарів	84,6	107,3	109,2	158,1	137,1	178,1	201,7

4.6 – 4.15. Знайти рівняння регресії, що описує залежність *Y* від *X* за даними кореляційної таблиць 4.16 – 4.17, оцінити її статистичну значущість та адекватність.

Таблиця 4.16

Y	*X*					
	10	20	30	40	50	60
5	*a*	*b*				

10		*c*	*d*			
15			*e*	*f*	*g*	
20			*h*	*k*	*m*	
25				*n*	*p*	*q*

Таблиця 4.17

	4.6	4.7	4.8	4.9	4.10	4.11	4.12	4.13	4.14	4.15
a	2	2	4	1	4	2	2	2	3	4
b	3	6	2	5	2	4	4	4	3	2
c	7	4	6	5	6	3	6	6	5	5
d	3	4	4	3	2	7	2	3	4	3
e	2	7	6	9	5	5	3	6	20	5
f	50	35	45	40	40	30	50	45	22	45
g	2	8	2	2	5	10	2	4	8	5
h	1	2	2	4	2	7	1	2	5	2
k	10	10	8	11	8	10	10	8	10	8
m	6	8	6	6	7	8	6	6	6	7
n	4	5	4	4	4	5	4	4	4	4
p	7	6	7	7	7	6	7	7	7	7
q	3	3	4	3	8	3	3	3	3	3

6. Загальне поняття кластерного аналізу

Кластерний аналіз (англ. cluster analysis) - задача розбиття заданої вибірки об'єктів (ситуацій) на підмножини, які називаються кластерами, так, щоб кожен кластер складався зі схожих об'єктів, а об'єкти різних кластерів

істотно відрізнялися. Завдання кластеризації відноситься до статистичної обробки, а також до широкого класу задач навчання без учителя.

Більшість дослідників схиляються до того, що вперше термін «кластерний аналіз» (англ. Cluster - гроно, згусток, пучок) був запропонований математиком Р. Тріоном. Згодом виник ряд термінів, які в даний час прийнято вважати синонімами терміна «кластерний аналіз»: автоматична класифікація; ботріологія.

Кластерний аналіз - це багатовимірна статистична процедура, що виконує збір даних, що містять інформацію про вибірку об'єктів, і потім упорядковуються об'єкти в порівняно однорідні групи (кластери) (Q-кластеризація, або Q-техніка, власне кластерний аналіз). Кластер - група елементів, які характеризуються загальною властивістю, головна мета кластерного аналізу - знаходження груп схожих об'єктів у вибірці. Спектр застосування кластерного аналізу дуже широкий: його використовують в археології, географії, медицині, психології, хімії, біології, державному управлінні, філології, антропології, маркетингу, соціології та інших дисциплінах. Однак універсальність застосування призвела до появи великої кількості несумісних термінів, методів і підходів, що ускладнюють однозначне використання і несуперечливу інтерпретацію кластерного аналізу.

Кластерний аналіз виконує такі основні завдання:

- Розробка типології або класифікації;
- Дослідження корисних концептуальних схем групування об'єктів;
- Породження гіпотез на основі дослідження данних;
- Перевірка гіпотез або дослідження для визначення, чи дійсно типи (групи), виділені тим або іншим способом, присутні в наявних даних.

Незалежно від предмета вивчення, застосування кластерного аналізу припускає наступні етапи:

- Відбір вибірки для кластеризації (мається на увазі, що має сенс кластеризувати тільки кількісні дані);
- Визначення безлічі змінних, за якими будуть оцінюватися об'єкти у вибірці, тобто ознакового простору;
- Обчислення значень тієї чи іншої міри подібності (або відмінності) між об'єктами;
- Застосування методу кластерного аналізу для створення груп схожих об'єктів;
- Перевірка достовірності результатів кластерного рішення.

Можна зустріти опис двох фундаментальних вимог пропонованих до даних - однорідність і повнота. Однорідність вимагає, щоб всі сутності, представлені в таблиці, були однієї природи. Вимога повноти полягає в тому, щоб безлічі I і J представляли повний опис проявів даного явища. Якщо розглядається таблиця в якій I - сукупність, а J - безліч змінних, що описують цю сукупність, то повинно повинно бути представлено вибіркою з досліджуваної сукупності, а система характеристик J повинна давати задовільне векторне подання індивідів і з точки зору дослідника.

Методи кластеризації

Загальноприйнятою класифікації методів кластеризації не існує, але можна відзначити солідну спробу В. С. Берікова і Г. С. Лобова. Якщо узагальнити різноманітні класифікації методів кластеризації, то можна виділити ряд груп (деякі методи можна віднести відразу до декількох груп і тому пропонується розглядати дану типізацію як деяке наближення до реальної класифікації методів кластеризації):

1. Імовірнісний підхід. Передбачається, що кожен аналізований об'єкт відноситься до одного з k класів. Деякі автори (наприклад, А. І. Орлов) вважають, що дана група зовсім не відноситься до кластеризації та

протиставляють її під назвою «дискримінація», тобто вибір віднесення об'єктів до однієї з відомих груп (навчальних вибірок).

- K-середніх (K-means)
- K-medians
- EM-алгоритм
- Алгоритми сімейства FOREL
- Дискримінантний аналіз

2. Підходи на основі систем штучного інтелекту. Вельми умовна група, так як методів AI дуже багато і методично вони досить різні.

- Метод нечіткої кластеризації C-середніх (C-means)
- Нейронна мережа Кохонена
- Генетичний алгоритм

3. Логічний підхід. Побудова дендрограми здійснюється за допомогою дерева рішень.

4. Теоретико-графовий підхід.

• Графовий алгоритм кластеризації

5. Ієрархічний підхід. Передбачається наявність вкладених груп (кластерів різного порядку). Алгоритми в свою чергу поділяються на агломеративні (об'єднуючи) і дивізійні (розділяючи). За кількістю ознак іноді виділяють монотетичні і політетичні методи класифікації. Ієрархічна дивізійна кластеризація або таксономія. Задачі кластеризації розглядаються в кількісній таксономії.

6. Інші методи. Не ввійшли в попередні групи.

- Статистичні алгоритми кластеризації
- Ансамбль кластерізаторов
- Алгоритми сімейства KRAB
- Алгоритм, заснований на методі просіювання
- DBSCAN та ін.

Незважаючи на значні відмінності між перерахованими методами всі вони спираються на вихідну «гіпотезу компактності»: в просторі об'єктів всі близькі об'єкти повинні відноситися до одного кластеру, а всі різні об'єкти відповідно повинні знаходитися в різних кластерах.

У даному методичному посібнику ми поетапно розглянемо процес здійснення кластерного аналізу на прикладі даних по вирощуванню сільськогосподарських культур по областях України.

ОРТОГОНАЛІЗАЦІЯ ДАНИХ

Важливою умовою достовірності результатів, отриманих в таксономічному аналізі, є «ортогональність» даних, на основі яких проводиться дослідження, тобто вихідні показники не повинні мати високих значень кореляції між собою. Іноді під час виконання практичної роботи «ортогоналізація» не потрібна,коли показники слабко корелюють між собою, але її можна виконати до будь-якого аналізу, в тому числі й таксономічного (кластерного).

В таксономічному аналізі наведемо приклад, на основі даних вирощування зернобобових культур, цукрових буряків, соняшнику, картоплі та овочів на території України (по областях) (табл.1).

Табл.1. Вирощування рослин по регіонах України

	A	B	C	D	E	F	G
1			зернобобові культури (у початково	цукрові буряки (фабричні)	соняшнику (у початково оприбуткова	Картопля	відкритого ґрунту
2		площадь					
3	в Республіка Крим	26 081	74,5792	0	2,235344	13,876	9,044899
4	Вінницька	26 513	138,4641	90,6687	11,09644	69,98831	14,13269
5	Волинська	20143	37,5664	30,9636	0,064539	54,46061	13,2304
6	Дніпропетровська	31 914	106,1697	2,26546	32,58758	17,54402	16,93614
7	Донецька	26517	87,13278	0,93525	28,54395	27,66904	18,91994
8	Житомирська	29 900	33,5017	15,8863	1,752508	44,47492	7,555184
9	Закарпатська	12 777	24,3015	0	0,367848	46,20803	20,34906
10	Запорізька	27 180	82,4283	0,0699	36,77336	10,46358	11,32082
11	Івано-Франківськ	13 928	34,1829	6,41872	0,911832	63,64159	9,606548
12	Київська	28 131	71,0462	42,3412	6,259998	63,83705	16,74665
13	Кіровоградська	24588	123,276	23,8328	37,51017	21,05499	9,329754
14	Луганська	26 517	47,6034	0,00377	20,88094	14,68869	10,74028
15	Львівська	21 833	42,2297	19,2095	0,087024	83,5341	20,51024
16	Миколаївська	24 598	109,196	0,81307	25,99398	8,683633	18,98528
17	Одеська	33 310	97,7754	0,04203	12,99009	16,63765	15,21165
18	Полтавська	28 748	134,009	72,4781	19,44483	44,2396	17,28468
19	Рівненська	20 047	39,0482	33,9801	0,214496	68,71352	11,18372
20	Сумська	23 834	75,7447	28,5726	10,2123	48,40983	7,753629
21	Тернопільська	13 823	127,367	119,207	1,721768	90,03834	15,58996
22	Харківська	31 415	106,599	27,1558	29,61324	30,97565	20,82445
23	Херсонська	28 481	88,5422	0	14,45487	9,173957	36,96989
24	Хмельницька	20 600	99,8544	73,6262	3,281553	71,49515	11,1068
25	Черкаська	20 900	136,493	49,9282	17,23445	44,18182	15,64115
26	Чернівецька	8 097	67,4571	13,5359	1,48203	77,62134	27,08411
27	Чернігівська	31 865	55,2393	15,5123	4,374706	55,00392	6,135258

Заносимо дані в середовище статистика, нормалізуємо (приводимо до інтервалу від 0 до 1) та будуємо кореляційну матрицю для показників (рис.1).

Spearman Rank Order Correlations (Spreadsheet1)
MD pairwise deleted
Marked correlations are significant at p <,05000

Variable	Var1	Var2	Var3	Var4	Var5
Var1	1,000000	0,388761	0,638462	-0,190000	0,228462
Var2	0,388761	1,000000	-0,169361	0,677444	-0,108545
Var3	0,638462	-0,169361	1,000000	-0,689231	0,094615
Var4	-0,190000	0,677444	-0,689231	1,000000	-0,036154
Var5	0,228462	-0,108545	0,094615	-0,036154	1,000000

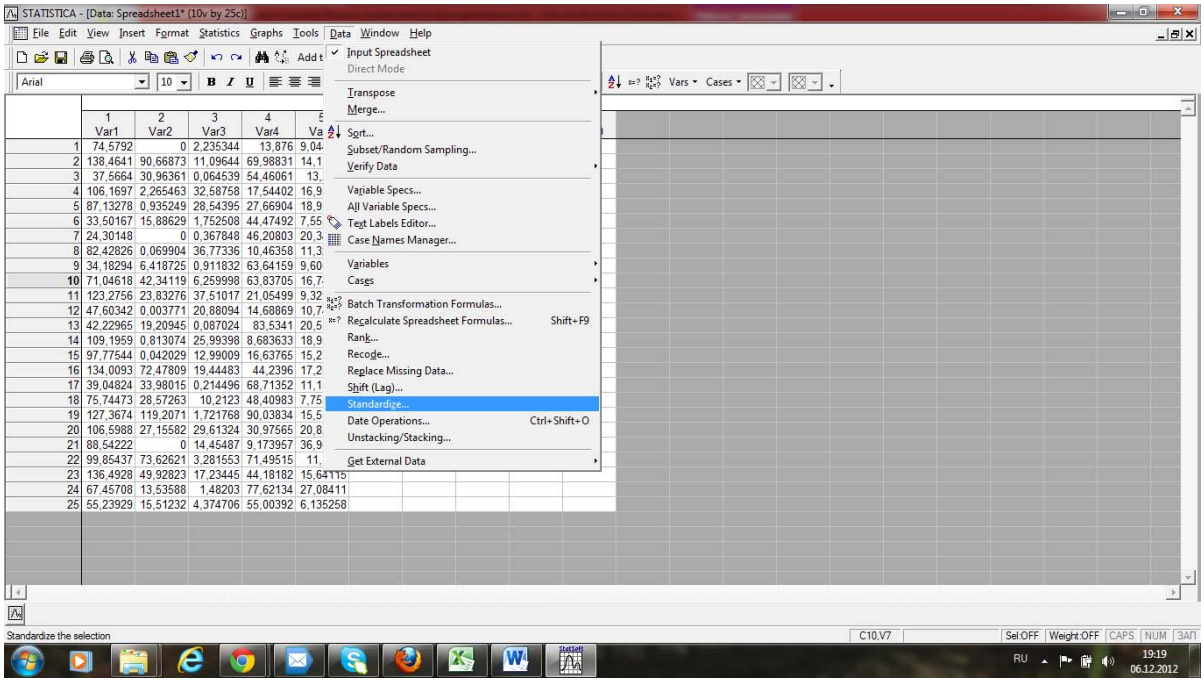
Рис.1. Приклад кореляційної матриці

Виходячи з отриманих даних, важливо сказати про те, що показники досить слабко корелюють між собою, тобто це означає те, що, наприклад, вирощування деяких видів рослинності на території України не є рівномірним і звичайно залежить від низки фізико-географічних факторів .

Вплив спотворень, викликаних високою взаємозалежністю змінних, можна також усунути, попередньо зваживши їх по компонентним навантаженням, які виділені з допомогою компонентного аналізу.

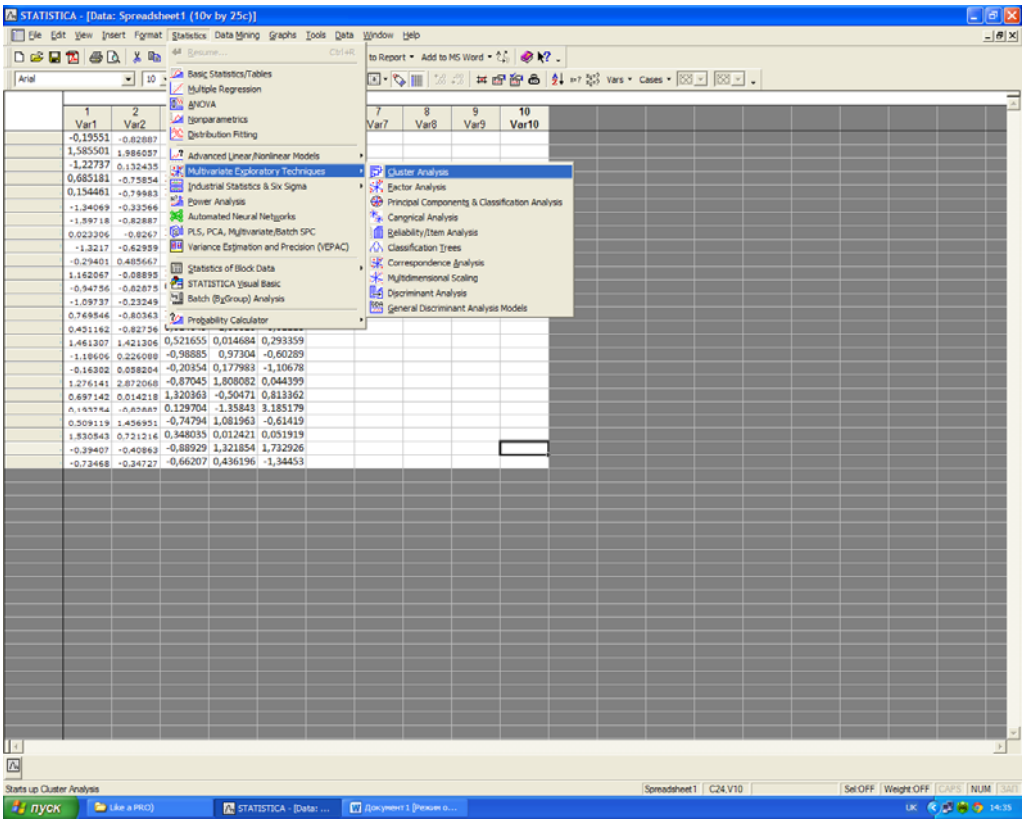
ПОБУДОВА ДЕНДРИТУ

1. Перед створенням дендриту нам необхідно скопіювати до програми Statistica дані, що стосуються врожайності кожної із культур рослинності, представлених у тонах на кілометр квадратний (т/км²). Попередню операцію переводу даних необхідно провести у програмі Microsoft Excel.
2. Скопіювавши дані до Statistica, необхідно провести нормування задля подальшої працею із даними. Робиться це наступним чином.

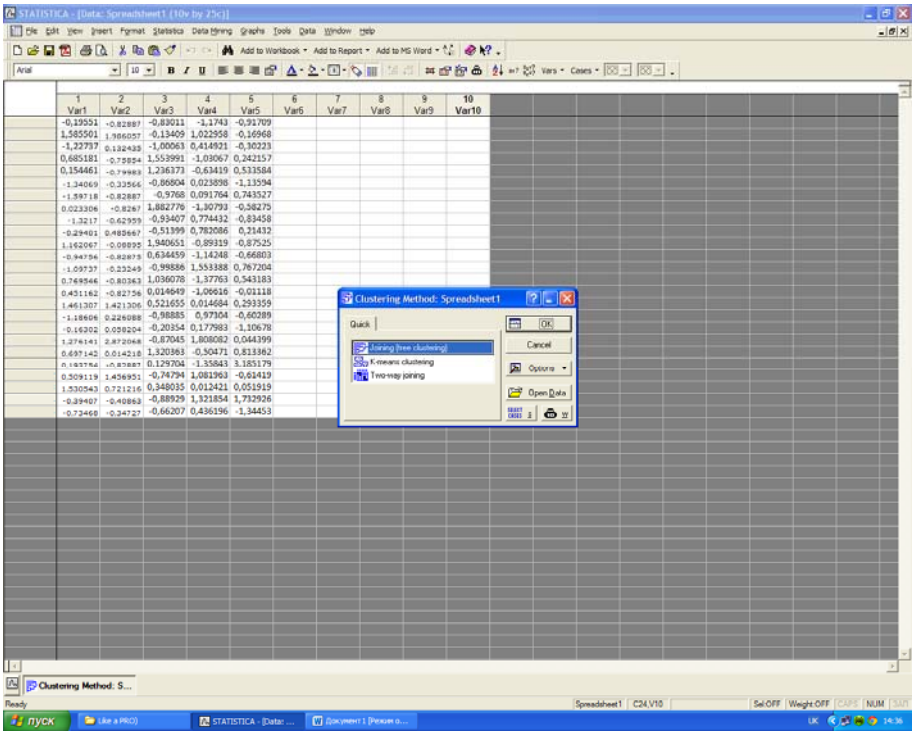


Перед тим, як провести стандартизацію, необхідно перевірити, чи обрані ті колонки, за якими проводиться дана процедура. В нашому випадку необхідно, щоб були обрані із першої по п’яту колонки (кожну колонку - ОКРЕМО).

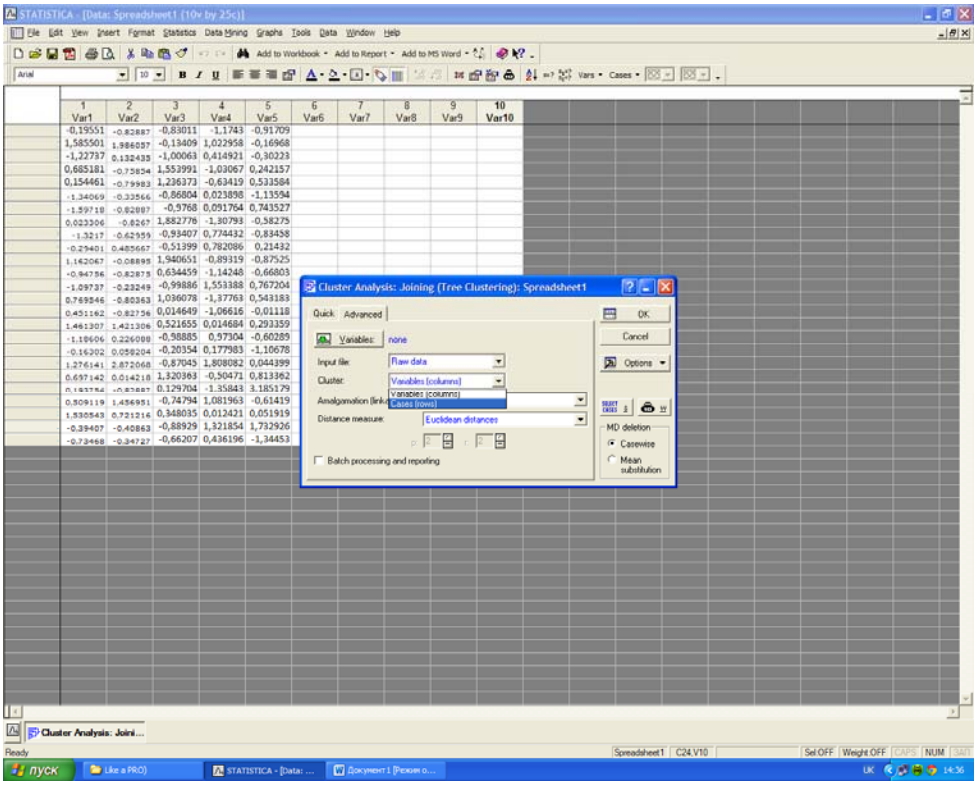
3. Після того, як дані будуть оброблені, починаємо побудову дендриту із того, що обираємо вкладку «Statistics» та серед спадаючих списків (див. малюнок нижче) обираємо ClusterAnalysis.



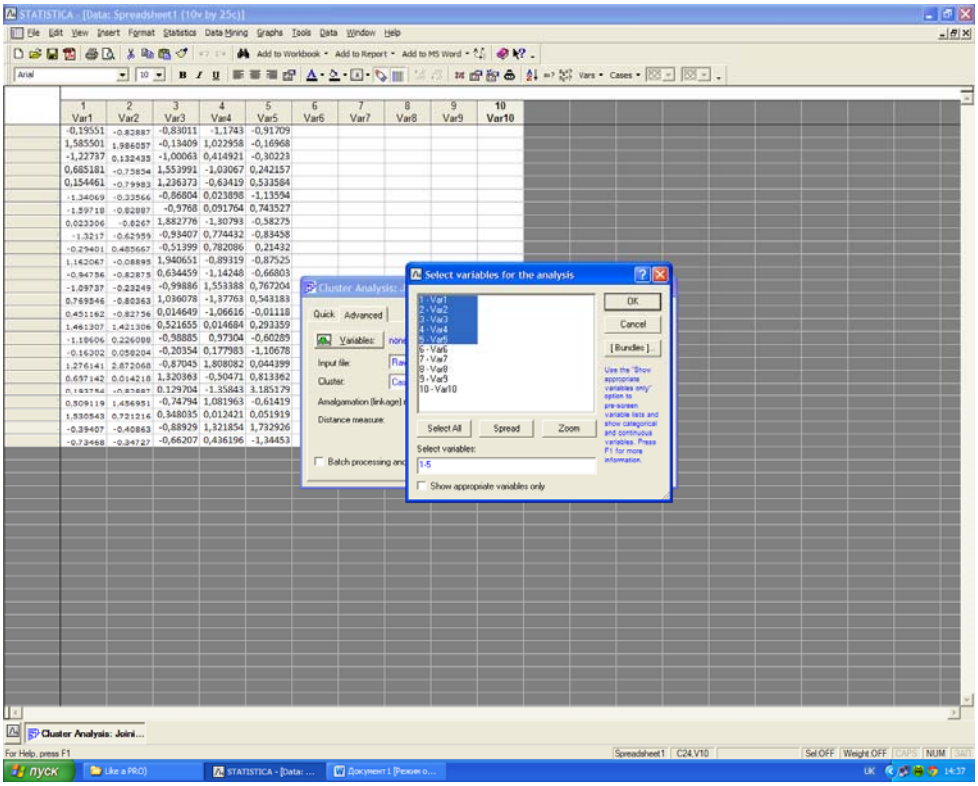
Після даної операції перед нами з’являється вікно під назвою «Cluster Metog». Обираємо об’єднану схему («Joining»), яка схожа за своїм зовнішнім виглядом на дерево.



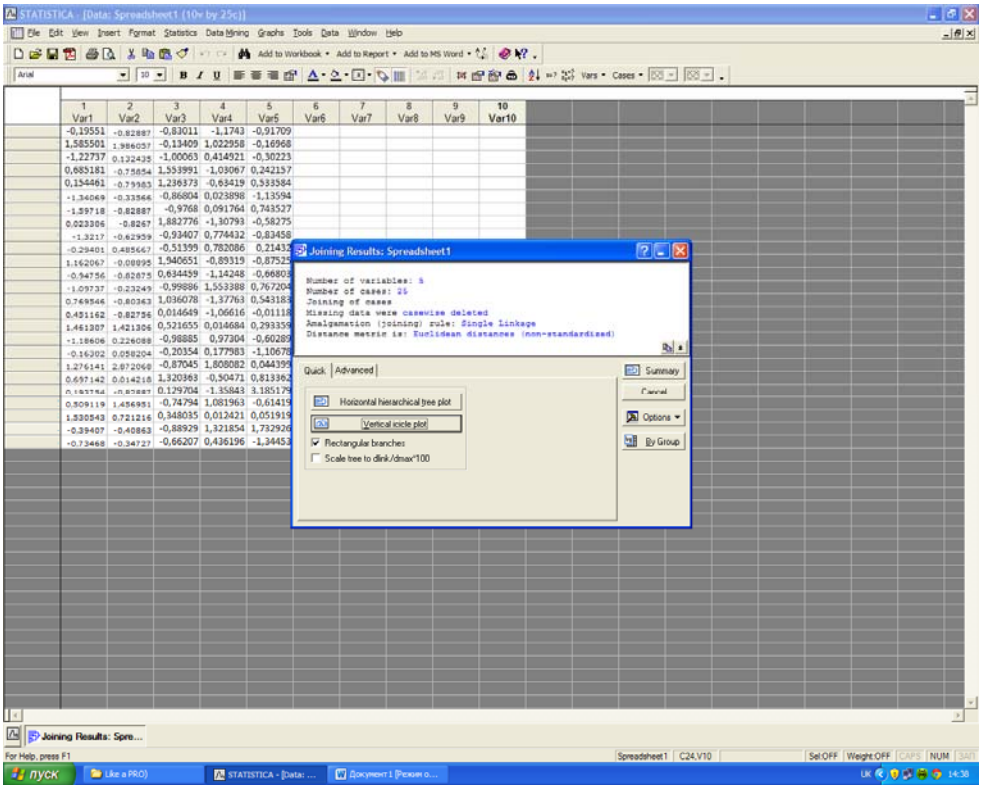
Після цього натискаємо «ОК». Але після цього перед нами знову з’являться вікно, де необхідно обрати особливості нашого дендриту. У спадаючому списку «Cluster»обираємо «Cases (rows)».



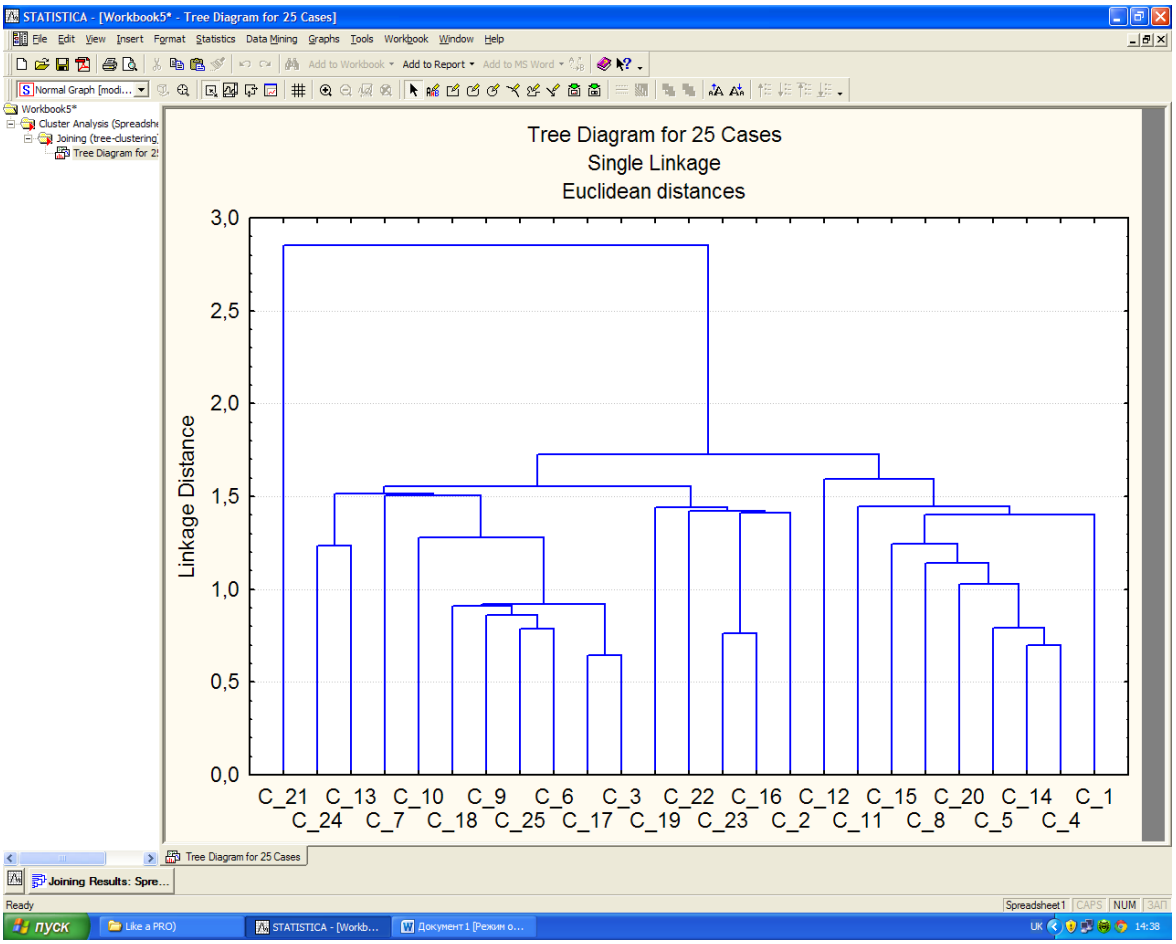
Після цього необхідно натиснути кнопку «Variables» та обрати наші п’ять колонок.



Після цього натискаємо «ОК». Але перед нами знову з’являється вікно. Тут нашою задачею є обрання вертикальної структури нашого дендриту. Обираємо її і натискаємо «ОК».



В результаті ми отримуємо дендрит, в якому наші дані поділені на 25 класів. Але така кількість класів є завеликою, тому нам оптимальніше буде виділити меншу кількість класів та проаналізувати, чому саме ті чи інші області опинилися у тому чи іншому класі: якими фізико-географічними чи суспільно-географічними чинниками це обумовлено.



КАРТОГРАФУВАННЯ РЕЗУЛЬТАТІВ КЛАСИФІКАЦІЇ

Після побудови дендриту і розробки класифікації областей (зазначаємо, що класифікувати області можна на різних рівнях із виділенням більшої чи меншої кількості класів в залежності від мети дослідження), класифікацію необхідно представити у вигляді карти або схеми. Зрозуміло, що картографічне представлення інформації буде сприйматися краще як непрофесіоналами, так і географами-спеціалістами. Для розробки карти ми виділили оптимальну кількість класів – 8, у деякі класи увійшло по декілька областей, у деяких опинилася лише 1 область (причини цьому наведені у розділі 5). Карту, представлену на рис.2,

розроблено максимально наочно, дуже спрощена географічна основа (відсутня гідрографічна мережа, мережа населених пунктів).

Рис. 2. Карта класифікації областей України за

сільськогосподарською спеціалізацією

АНАЛІЗ ОТРИМАНОЇ КЛАСИФІКАФІЇ

Після того, як ми закартографували отриману класифікацію, необхідно проаналізувати отримані результати, а саме пояснити, чому ті чи інші об’єкти опинилися в одному класі, які причини призвели до цього.

Загальні вимоги до такого аналізу наступні: він має бути об’єктивним, послідовним, логічним, максимально повним. Дослідник повинен пам’ятати, що отримана класифікація завжди буде дещо умовною і може не цілком враховувати особливості кожного конкретного досліджуваного об’єкта.