# Прикладна статистика

курс лекцій

д.ф.-м.н., проф. В.О. Коцюбинський

Прикарпатський національний університет імені Василя Стефаника
Кафедра управління та бізнес-адміністрування

м. Івано-Франківськ

2017

## Зміст навчальної дисципліни

1. Статистичні методи аналізу даних ;
2. Методи моделювання та аналізу взаємозвєязківк між характеристиками об’єкта або явища;
3. Способи обробки результатів вимірювань чи досліджень та експертного оцінювання;
4. Методи оцінки і прогнозування явищ.

Лекція 1

# Основні поняття математичної статистики

1. Поняття вибіркового методу в статистиці
2. Шкали вимірювань
3. Статистичні ряди та їх графічна інтерпретація
4. Числові характеристики статистичних рядів
5. Довірчі інтервали і довірча ймовірність
6. Визначення числових характеристик і довірчих інтервалів
   з використанням табличного процесору Microsoft Excel

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

Завданняматематичної статистики (МС):

- 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'єкта);
- визначення сили (тісноти зв'язку) між різними явищами (об'єктами) або властивостями одного і того ж явища (об'єкта);
- перевірка вірогідності статистичних гіпотез;
- розробка рекомендацій щодо проведення експерименту та обробки його результатів.

# Генеральна сукупність та вибірка

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

Генеральна сукупність може бути скінченою або нескінченною.

Кількість об'єктів (спостережень) генеральної сукупності називається об'ємом генеральної сукупності і позначається N.

Частина об'єктів генеральної сукупності, використовувана в ході дослідження, називається вибіркою.

Кількість об'єктів (спостережень) вибірки називається її об'ємом і позначається n.

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

# Генеральна сукупність та вибірка

Вибірка вона повинна правильно відображати кількісні та якісні співвідношення генеральної
сукупності, тобто бути репрезентативною.
Всі елементи генеральної сукупності повинні мати однакову ймовірність бути відібраними у вибірку, тобто вибірка має бути випадковою.
Типи ймовірнісної вибірки (відрізняються характером використаних прийомів):

- проста ймовірнісна вибірка, яка проводиться шляхом випадкового відбору об'єктів у вибірку;
- стратифікована вибірка, що використовується тоді, коли цілі та завдання дослідження вимагають відбору об'єктів для вивчення за певними груповими критеріями;
- багатоступінчаста вибірка, для якої характерно декілька послідовних змін одиниць відбору.

# Шкали вимірювань

Шкала – числова система, що відображає досліджувані властивості та ознаки об'єкта.

Шкала найменувань (класифікації, номінальна)

Якщо дані вимірюються за шкалою найменувань, то над ними можливі тільки операції порівняння : „рівні" або „нерівні". Дані номінальної шкали необхідні для ідентифікації певного об'єкту – місце розташування організації, адреса фірми, артикул товару.

Шкала порядку

Якщо дані вимірюються за шкалою порядку, їх можна порівняти за величиною „більше", „менше" або „рівні". За такою шкалою вимірюються, наприклад, вік студентів групи.

Шкала інтервалів

Якщо дані вимірюються за шкалою інтервалів, до них можна застосувати операції: порівняння – „більше", „менше", „рівні"; додавання і віднімання.

Прикладом даних, які належать до цієї шкали, є результати вимірювання температури, тиску.

Шкала відношень

Якщо дані вимірюються за шкалою відношень, їх можна порівняти за величиною та виконати всі арифметичні операції: додавання, віднімання, множення і ділення. Такою шкалою кодується вага, маса, ріст, довжина, дохід, обсяг виробництва і т. ін.

# Статистичні ряди

Припустимо, що необхідно вивчити деяку ознаку генеральної сукупності X, для чого було проведено n вимірювань цієї ознаки і складено вибірку її значень $\{x_1, x_2, \ldots, x_n\}$ об'єму n.

Різні елементи вибірки називаються варіантами.

Число $n_i$, що показує, скільки разів варіанта $x_i$ зустрічається у вибірці, називається частотою варіанти.

Число $w_i$, що дорівнює відношенню частоти варіанти $n_i$ до об'єм вибірки n, називається відносною частотою варіанти $x_i$:

$$w_i = \frac{n_i}{n}$$

Ряд варіант, розташованих в порядку зростання їх значень, називається варіаційним рядом. Послідовність, що складається із варіант і відповідних їм частот (відносних частот), називається статистичним рядом або рядом розподілу.

Ознака X є випадковою величиною, а статистичний ряд – емпіричним (тобто отриманим у результаті експерименту або спостережень) законом її розподілу.

# Статистичні ряди

Ознака X є випадковою величиною, а статистичний ряд – емпіричним (тобто отриманим у результаті експерименту або спостережень) законом її розподілу. Статистичний ряд називається дискретним, якщо він є законом розподілу дискретної випадкової величини, та інтервальним, якщо він є законом розподілу неперервної випадкової величини.

*Дискретний статистичний ряд*

| Варіанти $x_i$ | $x_1$ | $x_2$ | … | $x_k$ |
|---|---|---|---|---|
| Частоти $n_i$ (відносні частоти $w_i$) | $n_1$ $(w_1)$ | $n_2$ $(w_2)$ | … | $n_k$ $(w_k)$ |

де k – кількість варіант

*Неперервний статистичний ряд*

| Інтервали $[a_i;a_{i+1})$ | $[a_1;a_2)$ | $[a_2;a_3)$ | … | $[a_{k-1};a_k)$ |
|---|---|---|---|---|
| Частоти $n_i$ (відносні частоти $w_i$) | $n_1$ $(w_1)$ | $n_2$ $(w_2)$ | … | $n_k$ $(w_k)$ |

де k – кількість інтервалів

$$\sum_{i=1}^{k} n_i = n\,,\quad \sum_{i=1}^{k} w_i = 1\,.$$

Для побудови інтервального статистичного ряду множину значень варіант розбивають на інтервали $[a_i; a_{i+1})$, тобто проводять їх групування.
Кількість інтервалів k рекомендується розраховувати за формулою Стерджерса:

$$k = 1 + 1{,}4 \ln n$$

Довжина кожного із інтервалів розраховується за формулою

$$\Delta = \frac{x_{\max} - x_{\min}}{k}$$

де $x_{max}$, $x_{min}$ – максимальне і мінімальне значення у варіаційному ряді.

Підраховуючи кількість значень варіант, що потрапили в певний інтервал отримують частоти $n_i$

## Полігон частот, гістограма

Полігоном частот (відносних частот) називається ламана лінія, що сполучає точки з координатами:
$(x_i ; n_i)$ або $(x_i ; w_i)$ для дискретного статистичного ряду;
$(c_i ; n_i)$ або $(c_i ; w_i)$ для інтервального ряду, де $c_i$ – середина і-того інт$c_i = \frac{a_i + a_{i+1}}{2}$

Гістограмою називається ступінчаста фігура, яка складається з прямокутників з основами, що дорівнюють довжинам інтервалів, та висотами, які пропорційні частотам $n_i$ (відносним частотам $w_i$) і обчислюються як відношення частот $n_i$ (відносних частот w i) до довжин відповідних інтервалів.
Площа гістограми частот дорівнює об'єму вибірки n.
Площа гістограми відносних частот дорівнює одиниці.

*За статистичним рядом можна встановити емпіричну функцію розподілу та емпіричну щільність розподілу випадкової величини X.*

# Емпірична функція розподілу і кумулята

Емпірична функція розподілу

$$F_n(x) = \frac{1}{n}\sum_{x_i<x} n_i = \sum_{x_i<x} w_i$$

Кумулятою називається крива, що проходить через точки з координатами: $(a_i ; F_n(a_i))$

Емпіричною щільністю розподілу для інтервального ряду називається функція

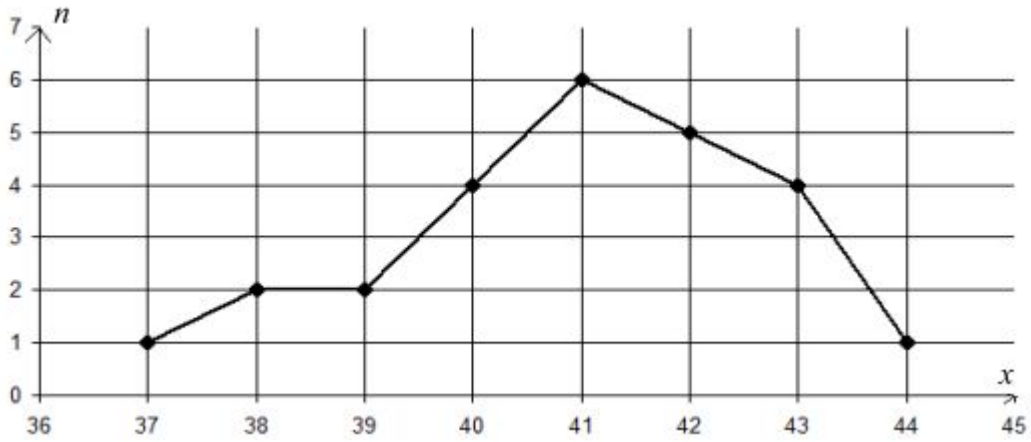
$$f_n(x) = \begin{cases} \dfrac{n_i}{n\Delta} = \dfrac{w_i}{\Delta}, \text{ якщо } a_i \le x \le a_{i+1}, i = \overline{1,k} \\ 0, \quad \text{якщо } x < a_1 \text{ або } x > a_{k+1}, i = \overline{1,k} \end{cases}$$

## Приклад 1.1. Дискретний розподіл

У результаті тестування службовців деякої компанії були отримані такі результати (у балах): 39, 41, 40, 42, 41, 40, 42, 44, 40, 43, 38, 42, 41, 43, 39, 37, 43, 41, 38, 42, 40, 41, 42, 40, 41. Побудувати дискретний статистичний ряд для випадкової величини X – оцінки службовців, полігон частот, емпіричну функцію розподілу та її графік.

Розв'язок. Для побудови дискретного статистичного ряду записуємо у порядку зростання різні значення випадкової величини X і відповідні частоти Останній стовпчик таблиці використовується для перевірки правильності побудови статистичного ряду (усього у тестуванні приймали участь 25 осіб, сума частот повинна дорівн

| $x_i$ | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | Сума |
|---|---|---|---|---|---|---|---|---|---|
| $n_i$ | 1 | 2 | 2 | 4 | 6 | 5 | 4 | 1 | $\sum_{i=1}^{k} n_i = 25$ |




Полігон частот розподілу

## Приклад 1.1. Дискретний розподіл

Для побудови емпіричної функції розподілу доповнимо таблицю двома рядками. В першому рядку обчислимо суму частот варіант, що менші $x_i$

якщо $x < x_1 = 37$, то $\sum_{x_i<37} n_i = 0$, оскільки таких значень $X$ немає;

якщо $x < x_2 = 38$, то $\sum_{x_i<38} n_i = n_1 = 1$;

якщо $x < x_3 = 39$, то $\sum_{x_i<39} n_i = n_1 + n_2 = 1 + 2 = 3$;

якщо $x < x_4 = 40$, то $\sum_{x_i<40} n_i = n_1 + n_2 + n_3 = 1 + 2 + 2 = 5$;

якщо $x < x_5 = 41$, то $\sum_{x_i<41} n_i = n_1 + n_2 + n_3 + n_4 = 1 + 2 + 2 + 4 = 9$;

якщо $x < x_6 = 42$, то $\sum_{x_i<42} n_i = n_1 + n_2 + n_3 + n_4 + n_5 = 1 + 2 + 2 + 4 + 6 = 15$;

якщо $x < x_7 = 43$, то $\sum_{x_i<43} n_i = n_1 + n_2 + n_3 + n_4 + n_5 + n_6 = 1 + 2 + 2 + 4 + 6 + 5 = 20$;

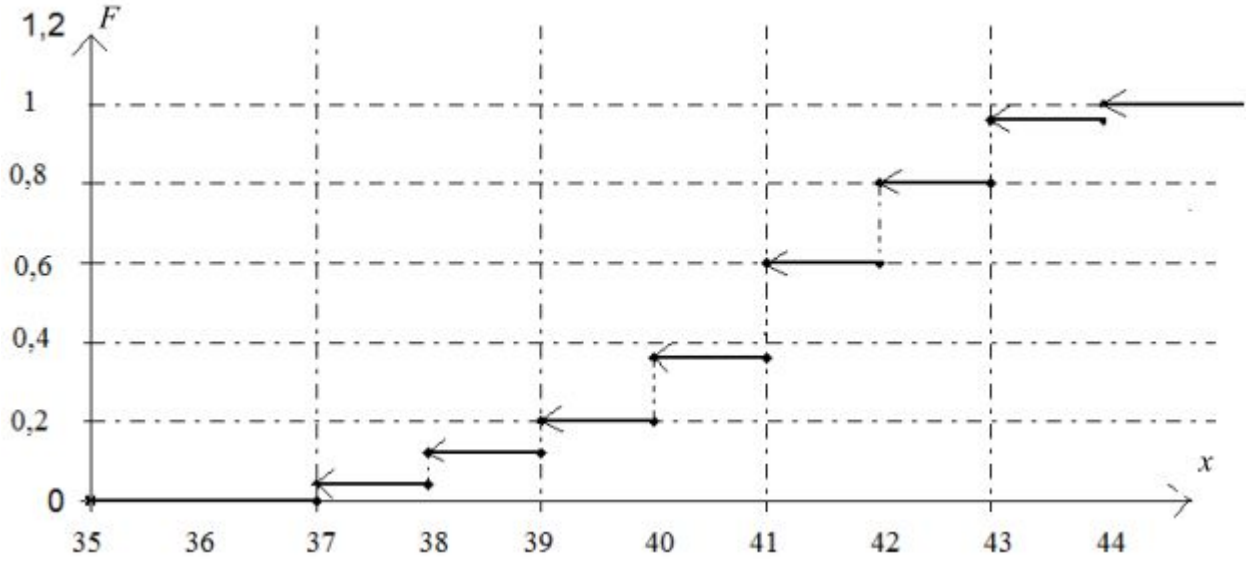
якщо $x < x_8 = 44$, то $\sum_{x_i<44} n_i = n_1 + n_2 + n_3 + n_4 + n_5 + n_6 + n_7 = 20 + 4 = 24$;

якщо $x > x_8 = 44$, то $\sum_{x_i<x} n_i = 25$ – це означає, що всі значення $X$ менші числа, більшого за 44.

# Приклад 1.1. Дискретний розподіл

$$F_n(x) = \frac{1}{n}\sum_{x_i<x} n_i = \sum_{x_i<x} w_i$$

| $x_i$ | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | |
|---|---|---|---|---|---|---|---|---|---|
| $n_i$ | 1 | 2 | 2 | 4 | 6 | 5 | 4 | 1 | |
| $\sum_{x_i<x} n_i$ | 0 | 1 | 3 | 5 | 9 | 15 | 20 | 24 | 25 |
| $F_i$ | 0 | 0,04 | 0,12 | 0,2 | 0,36 | 0,6 | 0,8 | 0,96 | 1 |




Графік емпіричної функції розподілу

## Приклад 1.2. Неперервний розподіл

За даними вибіркового дослідження було отримано розподіл родин за доходом на одного їх члена в умовних одиницях (табл. ). Побудувати інтервальний статистичний ряд, полігон частот, гістограму, полігон відносних частот, емпіричні функцію і щільність розподілу та їх графіки.

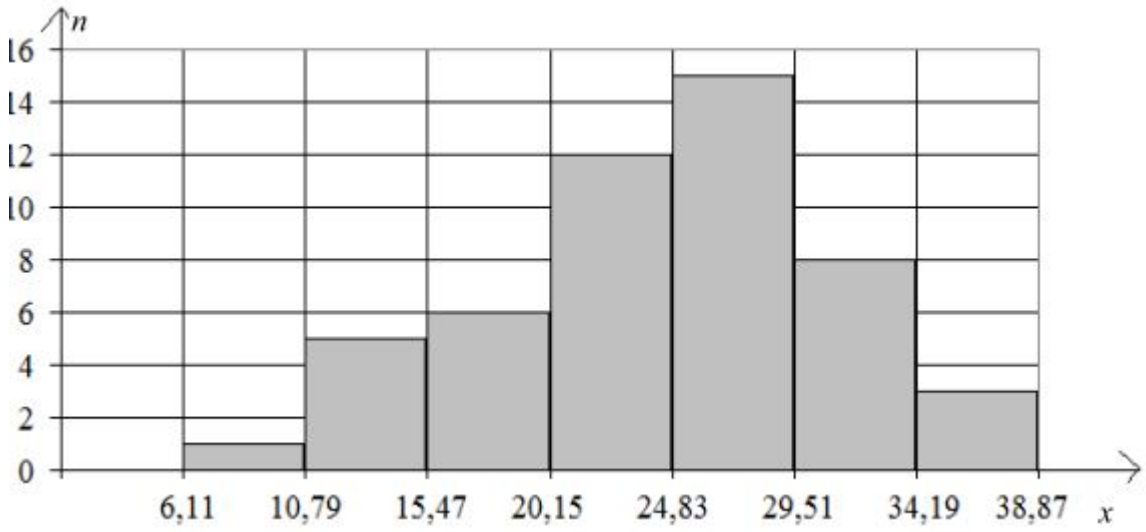
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 28,92 | 27,54 | 22,36 | 29,09 | 32,19 | 26,04 | 17,06 | 26,83 | 24,55 | 33,22 |
| 17,53 | 30,07 | 36,27 | 24,24 | 26,03 | 31,05 | 13,94 | 14,56 | 21,40 | 23,04 |
| 13,09 | 38,84 | 25,57 | 22,87 | 6,11 | 27,79 | 25,68 | 16,30 | 17,93 | 24,37 |
| 28,92 | 27,54 | 22,36 | 29,06 | 32,19 | 26,04 | 17,06 | 26,83 | 24,55 | 33,22 |
| 17,53 | 30,07 | 36,27 | 24,24 | 26,03 | 31,05 | 13,94 | 14,56 | 21,40 | 23,04 |

**Розв'язок.** Табл. містить 50 даних, тобто $n = 50$. Для побудови інтервального статистичного ряду знаходимо: кількість інтервалів за формулою $k = 1 + 1,4 \ln 50 \approx 6,477 \approx 7$; $x_{\max} = 38,84$, $x_{\min} = 6,11$; довжина кожного інтервалу за формулою $\Delta = \frac{x_{\max} - x_{\min}}{k} = \frac{38,84 - 6,11}{7} \approx 4,68$. Отже, за початок першого інтервалу обираємо $a_1 = x_{\min} = 6,11$. Тоді $a_2 = a_1 + \Delta = 6,11 + 4,68 = 10,79$. Аналогічно, $a_3 = 15,47$; $a_4 = 20,15$; $a_5 = 24,83$; $a_6 = 29,51$; $a_7 = 34,19$; $a_8 = 38,87$.

## Приклад 1.2. Неперервний розподіл

Підраховуючи кількість варіант, що попали в кожний інтервал, отримаємо інтервальний статистичний ряд

| $[a_i; a_{i+1})$ | [6,11; 10,79) | [10,79; 15,47) | [15,47; 20,15) | [20,15; 24,83) | [24,83; 29,51) | [29,51; 34,19) | [34,19; 38,87) |
|---|---|---|---|---|---|---|---|
| $n_i$ | 1 | 5 | 6 | 12 | 15 | 8 | 3 |
| $n_i/\Delta$ | 0,214 | 1,068 | 1,282 | 2,564 | 3,205 | 1,709 | 0,641 |



## Приклад 1.2. Неперервний розподіл

Для побудови полігона частот і полігона відносних частот обчислимо середини інтервалів за формулою: $c_i = \frac{a_i + a_{i+1}}{2}$. Отримаємо:

$c_1 = \frac{6{,}11 + 10{,}79}{2} = 8{,}45$; $c_2 = 13{,}13$; $c_3 = 17{,}81$; $c_4 = 22{,}49$; $c_5 = 27{,}17$; $c_6 = 31{,}85$; $c_7 = 36{,}51$.

Розрахуємо відносні частоти за формулою : $w_1 = \frac{n_1}{n} = \frac{1}{50} = 0{,}02$; $w_2 = 0{,}1$; $w_3 = 0{,}12$; $w_4 = 0{,}24$; $w_5 = 0{,}3$; $w_6 = 0{,}16$; $w_7 = 0{,}06$.

| $c_i$ | 8,45 | 13,13 | 17,81 | 22,49 | 27,17 | 31,85 | 36,51 | Перевірка |
|---|---|---|---|---|---|---|---|---|
| $n_i$ | 1 | 5 | 6 | 12 | 15 | 8 | 3 | $\sum_{i=1}^{k} n_i = 50$ |
| $w_i$ | 0,02 | 0,1 | 0,12 | 0,24 | 0,3 | 0,16 | 0,06 | $\sum_{i=1}^{k} w_i = 1$ |

# Приклад 1.2. Неперервний розподіл

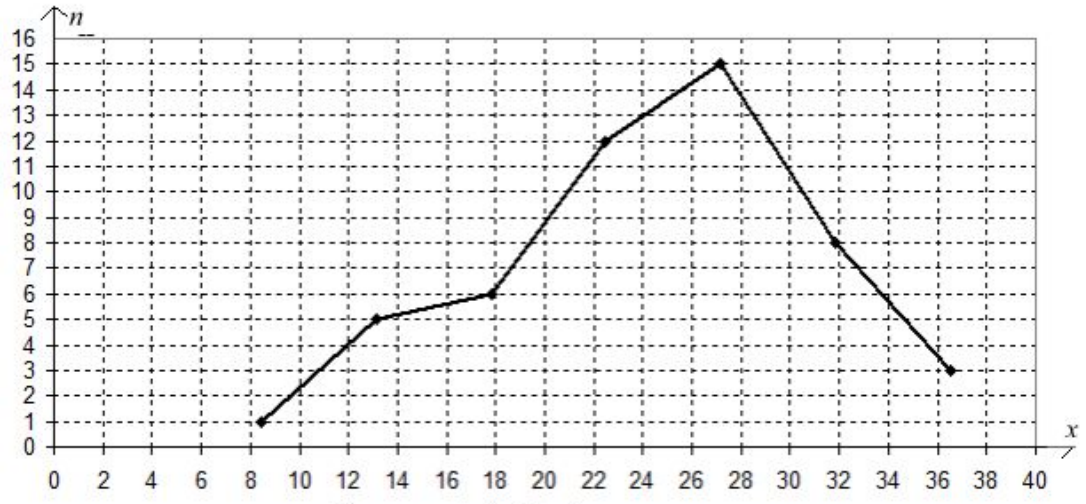
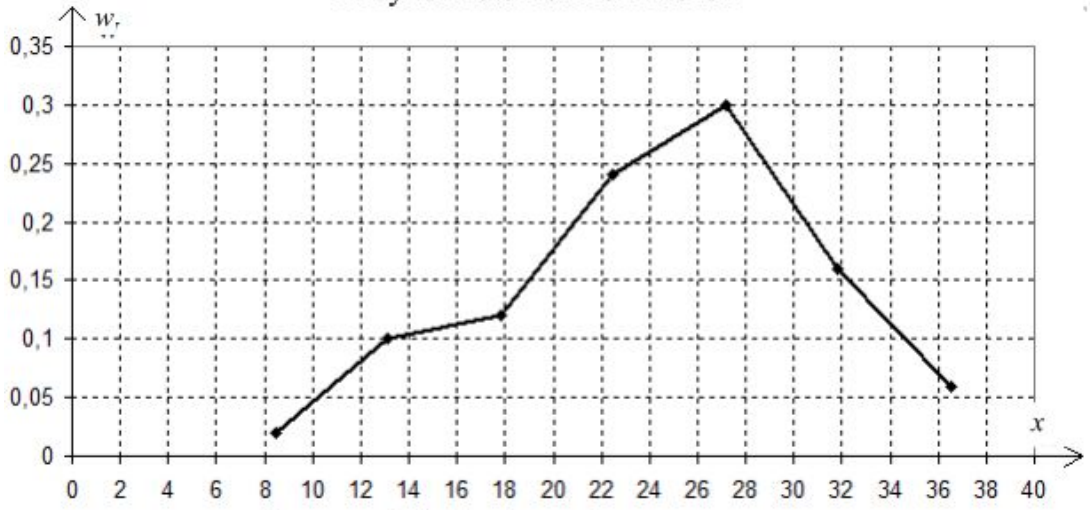



Рисунок 1.4. Полігон частот

Полігон частот




Полігон відносних частот

## Приклад 1.2. Неперервний розподіл

$$F_n(x)=\frac{1}{n}\sum_{x_i<x} n_i=\sum_{x_i<x} w_i$$

якщо $x=a_1=6{,}11$, то $F_n(x)=\sum_{x_i<6{,}11} w_i=0$, оскільки таких значень $X$ немає;

якщо $x=a_2=10{,}79$, то $F_n(x)=\sum_{x_i<10{,}79} w_i=w_1=0{,}02$;

якщо $x=a_3=15{,}47$, то $F_n(x)=\sum_{x_i<15{,}47} w_i=w_1+w_2=0{,}02+0{,}1=0{,}12$;

якщо $x=a_4=20{,}15$, то $F_n(x)=\sum_{x_i<20{,}15} w_i=w_1+w_2+w_3=0{,}02+0{,}1+0{,}12=0{,}24$;

якщо $x=a_5=24{,}83$, то $F_n(x)=\sum_{x_i<24{,}83} w_i=w_1+w_2+w_3+w_4=$

$=0{,}02+0{,}1+0{,}12+0{,}24=0{,}48$;

якщо $x=a_6=29{,}51$, то $F_n(x)=\sum_{x_i<29{,}51} w_i=w_1+w_2+w_3+w_4+w_5=$

$=0{,}02+0{,}1+0{,}12+0{,}24+0{,}3=0{,}78$;

якщо $x=a_7=34{,}19$, то $F_n(x)=\sum_{x_i<34{,}19} w_i=w_1+w_2+w_3+w_4+w_5+w_6=$

$=0{,}02+0{,}1+0{,}12+0{,}24+0{,}3+0{,}16=0{,}94$;

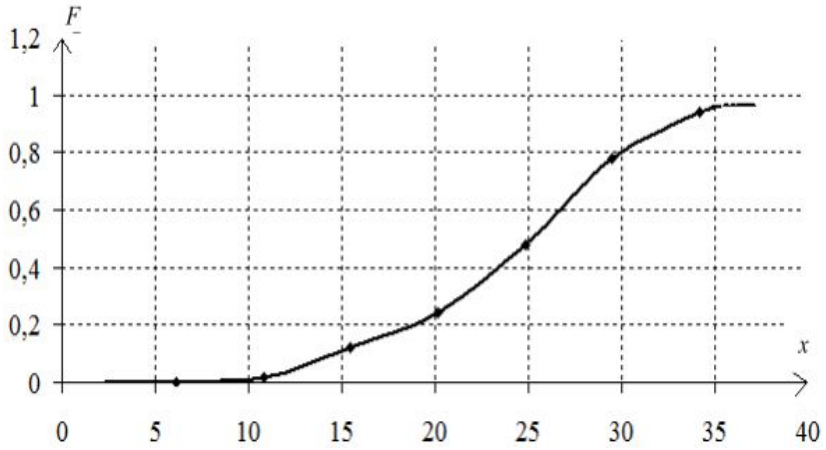
якщо $x=a_8=38{,}87$, то $F_n(x)=\sum_{x_i<38{,}87} w_i=w_1+w_2+w_3+w_4+w_5+w_6+w_7=$

$=0{,}02+0{,}1+0{,}12+0{,}24+0{,}3+0{,}16+0{,}06=1$.

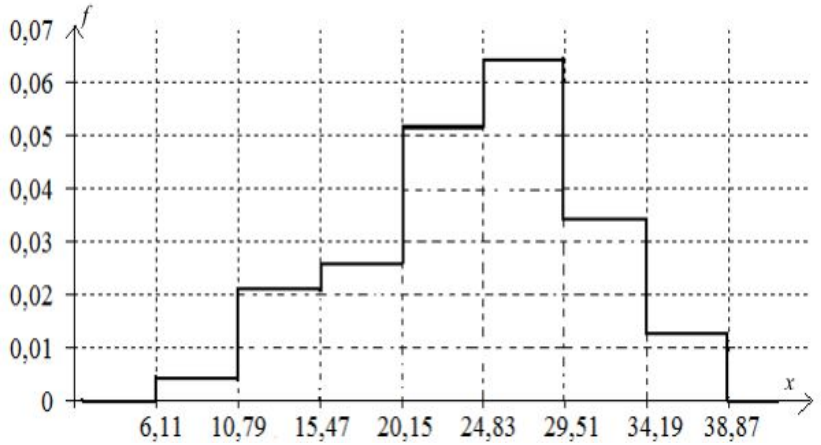
## Приклад 1.2. Неперервний розподіл

Емпіричну щільність розподілу обчислимо за формулою $f_n(x) = \frac{w_i}{\Delta}$

| $a_i$ | 6,11 | 10,79 | 15,47 | 20,15 | 24,83 | 29,51 | 34,19 | 38,87 |
|---|---|---|---|---|---|---|---|---|
| $w_i$ | 0,02 | 0,1 | 0,12 | 0,24 | 0,3 | 0,16 | 0,06 | $\sum_{i=1}^{k} w_i = 1$ |
| $F_i$ | 0 | 0,02 | 0,12 | 0,24 | 0,48 | 0,78 | 0,94 | 1 |
| $[a_i; a_{i+1})$ | [6,11; 10,79) | [10,79; 15,47) | [15,47; 20,15) | [20,15; 24,83) | [24,83; 29,51) | [29,51; 34,19) | [34,19; 38,87) | |
| $f_i$ | 0,0043 | 0,0214 | 0,0256 | 0,0513 | 0,0641 | 0,0342 | 0,0128 | |




Графік кумуляти розподілу

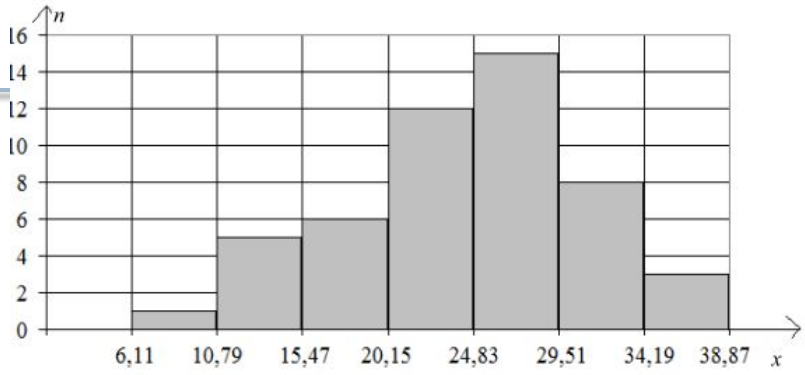



Графік емпіричної щільності розподілу

# Числові характеристики статистичних рядів

Деяку ознаку X генеральної сукупності можна розглядати як випадкову величину. Числові характеристики випадкових величин є параметрами їх розподілів.

Числові характеристики статистичних рядів
*вибіркове середнє,*
*вибіркове середнє геометричне,*
*вибіркова дисперсія,*
*вибіркове середнє квадратичне відхилення.*




Вибіркове середнє

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i = \sum_{i=1}^{k} x_i w_i$$

Вибіркове середнє геометричне

$$\overline{x_G} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Мода Mo значення величини X, яке ма у вибірці найбільшу частоту.

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

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

Δ – величина інтервала;

$n_{Mo}$ – частота у модальному інтервалі;

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

Медіаною Me називається значення величини X, що розділяє вибірку, елементи якої розташовані у порядку зростання, на дві рівні за об'ємом частини.

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

$X_{me}$ –нижня границя медіанного інтервала;
x
$n_x^{\max}$ – сума частот, що накопичена до початку медіанного інтервала;
$n_{Me}$ – частота в медіанному інтервалі.

Якщо виникає необхідність у більш дрібному поділі статистичного ряду крім медіани виділяють
квартилі $Q_1$, $Q_2$ =Me, $Q_3$ (1/4 ряду),
квінтилі $q_1$ ,..., $q_4$ (1/5 ряду), децилі $d_1$ , $d_2$ ,..., $d_9$ (1/10 ряду).

нижній квартиль – $$Q_1 = X_{Q_1} + \frac{\Delta\left(\frac{1}{4}n - n_{x_1}^{\max}\right)}{n_{Q_1}},$$

верхній квартиль – $$Q_3 = X_{Q_3} + \frac{\Delta\left(\frac{3}{4}n - n_{x_3}^{\max}\right)}{n_{Q_3}},$$

$X_Q$ – фактична нижня границя квартильних інтервалів;
$\Delta$ – величина інтервала;
$n$ – об'єм вибірки;

$n_{x_1}^{\max}$, $n_{x_3}^{\max}$ – суми частот, накопичені до початку відповідних квартильних інтервалів;

$n_{Q_1}$, $n_{Q_3}$ – частоти в квартильних інтервалах.

## Числові характеристики розсіювання

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

$$R = x_{max} - x_{min}$$

Вибірковою дисперсією $S^2$ називається середнє арифметичне квадратів відхилень варіант від їх вибіркової середньої:

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

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

Якщо дані незгруповані, то

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

Дисперсія є показником розсіювання елементів вибірки відносно їх середнього значення.
Вибіркова дисперсія називається незсуненою оцінкою дисперсії генеральної сукупності.

*Незсуненість* - 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.

## Числові характеристики розсіювання

Вибіркове середнє квадратичне відхилення S - величина, що дорівнює кореню квадратному з вибіркової дисперсі $S = \sqrt{S^2}$

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

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

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

## Приклад 1.3

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

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

**Розв'язок.** За незгрупованими даними табл. 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 (1.17), коефіцієнт варіації за формулою (1.18). 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 $\frac{n}{2}+1$: $x_{n/2}=x_4=125$; $x_{n/2+1}=x_5=140$;

$$Me=\frac{x_4+x_5}{2}=\frac{125+140}{2}=132,5;\ Q_1=\frac{x_2+x_3}{2}=112,5,\ Q_3=145.$$

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

## Приклад 1.3

| $x_i$ | 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 |

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{826{,}563 + 351{,}56 + 189{,}06 + 14{,}063 + 126{,}56 + 264{,}06 + 264{,}06}{8} +$$

$$+ \frac{451{,}56}{8} = 310{,}938.$$

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

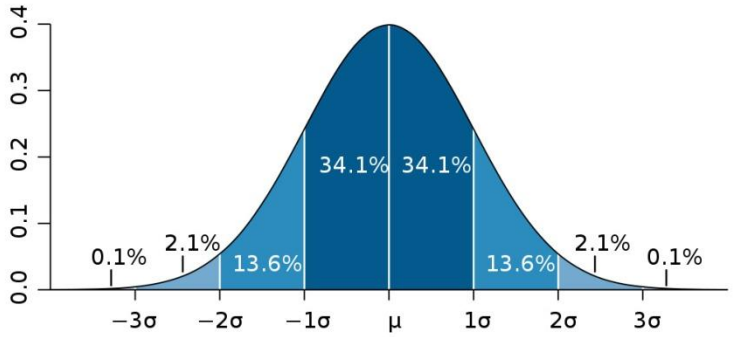
Коефіцієнт варіації $v = \frac{S}{|\bar{x}|} \cdot 100\% = \frac{17{,}63}{128{,}75} \cdot 100\% \approx 13{,}69\%$.

# Довірчі інтервали і довірча ймовірність

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

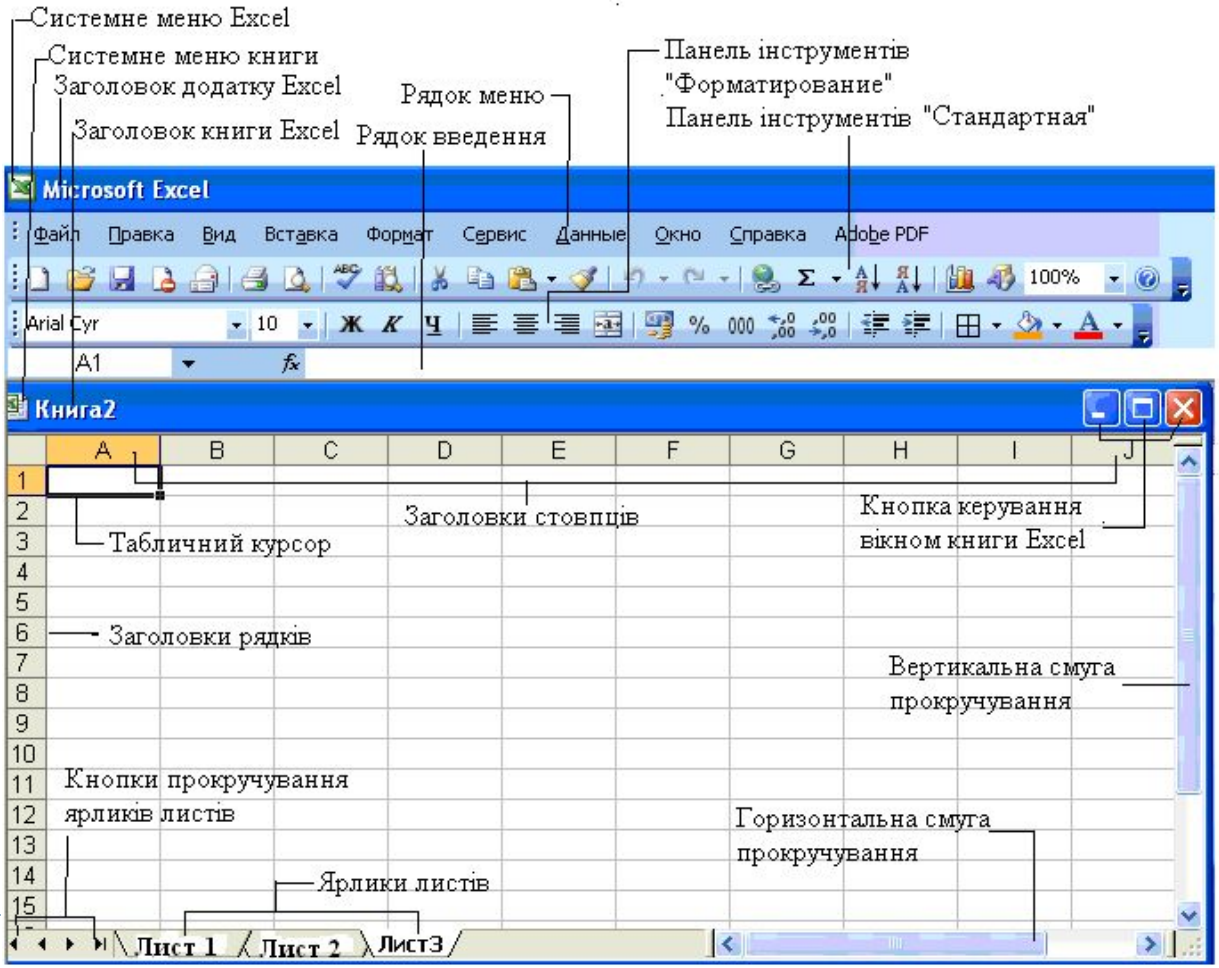
Значення довірчої ймовірності обирає дослідник залежно від того, яку ступінь точності розрахунків вимагає дослідження.
Типово значення знаходиться в інтервалі від 0,9 до 0,999.
Якщо вимоги точності дуже високі, то для довірчої ймовірності обирається значення 0,999; якщо підвищені – 0,99;
звичайні – 0,95;
знижені – 0,9.
Довірчі інтервали розраховуються з урахуванням певних вимог до генеральної сукупності. Типова вимога - нормальний розподіл даних.

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



Табличний процесор – це засіб для автоматизації розрахунків при роботі з табличними даними.

Microsoft Excel – це засіб для роботи з електронними таблицями, що містить апарат для обробки даних у вигляді набору функцій аналізу даних,



# Робота з функціями Excel

Функції – це заздалегідь визначені формули, що виконують обчислення за заданими величинами (аргументами) в зазначеному порядку.

1) математичні функції;
2) статистичні функції;
3) логічні функції;
4) фінансові функції;
5) функції дати і часу;
6) вкладені функції;
7) функції роботи з базами даних;
8) текстові функції;
9) функції посилання та масивів

**=ім'я функції (параметр/и)**

# Математичні функції

СУММ – додає аргументи.
КОРЕНЬ – повертає додатне значення квадратного кореня.
COS, SIN, TAN – тригонометричні функції cos, sin і tg.
ACOS, ATAN – зворотні тригонометричні функції arccos, arctg.
ГРАДУСЫ – перетворює радіани в градуси.
LN – натуральний логарифм числа.
ABS – модуль числа.
ПИ – повертає число Пі (π=3.14).
ЗНАК – повертає знак числа.
ПРОИЗВЕД – повертає добуток аргументів.
СТЕПЕНЬ – повертає результат піднесення до степеня.
ОКРУГЛ – закруглює число до заданої кількості десяткових розрядів.
ОСТАТ – повертає остачу від ділення.
СЛЧИС – повертає випадкове число в інтервалі від 0 до 1.
РИМСКОЕ – перетворює число в арабському записі до числа в римському як текст.
СУММЕСЛИ – повертає суму вмістимого комірок, яке задовольняє заданому критерію;
СУММКВ – повертає суму квадратів аргументів.
МОБР, МУММНОЖ, МОПРЕД – зворотна матриця, добуток та визначник матриці.

# Статистичні функції

| Название | Обозна-чение | Название в сводной таблице | Метод вычисления | Формула Excel |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |
| Размах вариации | R | **Интервал** | Разница максимального и минимального значений | **МАКС (интервал) – МИН (интервал)** |
| Объём выборки | n | **Счёт** | Количество статистических единиц | **СЧЕТ (интервал)** |
| Медиана | Me | **Медиана** | Центральное значение отсортированной выборки | **МЕДИАНА (интервал)** |
| Мода | Mo | **Мода** | Наиболее часто встречающееся значение | **МОДА (интервал)** |
| Среднее | $\overline{x}$ | **Среднее** | Среднее арифметическое | **СРЗНАЧ (интервал)** |
| Среднее линейное отклонение | d | - | Средний модуль отклонения от среднего значения | **СРОТКЛ (интервал)** |
| Дисперсия | D | **Дисперсия** | Средний квадрат отклонения от среднего значения | **ДИСП (интервал)** |

# Статистичні функції

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| Среднее квадратичное отклонение | S | - | Среднее квадратическое отклонение от среднего значения | **СТАНДОТКЛОНП (интервал)** |
| Среднее квадратичное отклонение (несмещённая оценка) | S | **Стандартное отклонение** | Среднее квадратическое отклонение от среднего значения с поправкой на объём выборки | **СТАНДОТКЛОН (интервал) – несмещённая оценка** |
| Коэффициент осцилляции | $V_R$ | - | $V_R = \frac{R}{\bar{x}} \cdot 100$ | - |
| Линейный коэффициент вариации | $V_d$ | - | $V_d = \frac{d}{\bar{x}} \cdot 100$ | - |
| Коэффициент вариации | $V_\sigma$ | - | $V_\sigma = \frac{\sigma}{\bar{x}} \cdot 100$ | - |
| Коэффициенты асимметрии | A | **Асимметрич-ность** | $A_s = \frac{\sum (x_i - \bar{x})^3}{n\sigma^3}$ | - |
| Коэффициент эксцесса | E | **Эксцесс** | $E_x = \frac{\sum (x_i - \bar{x})^4}{n\sigma^4} - 3$ | - |

# Приклад застовування Excel

За даними вибіркового дослідження відома заробітна платня (в ум.од.) 20-ти службовців певної компанії. 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.

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

| **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) |

# Статистичні функції

НОРМРАСП(x; Среднее; Стандартное_откл; Интегральная),

НОРМРАСП(x; a; σ; 0) $= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-a)^2}{2\sigma^2}}$

НОРМРАСП(x; a; σ; 1) $= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{(t-a)^2}{2\sigma^2}} dt$

СТЬЮДРАСП(x; степени свободы; 1)

Ступені свободи - число ступенів свободи, що характеризує розподіл.

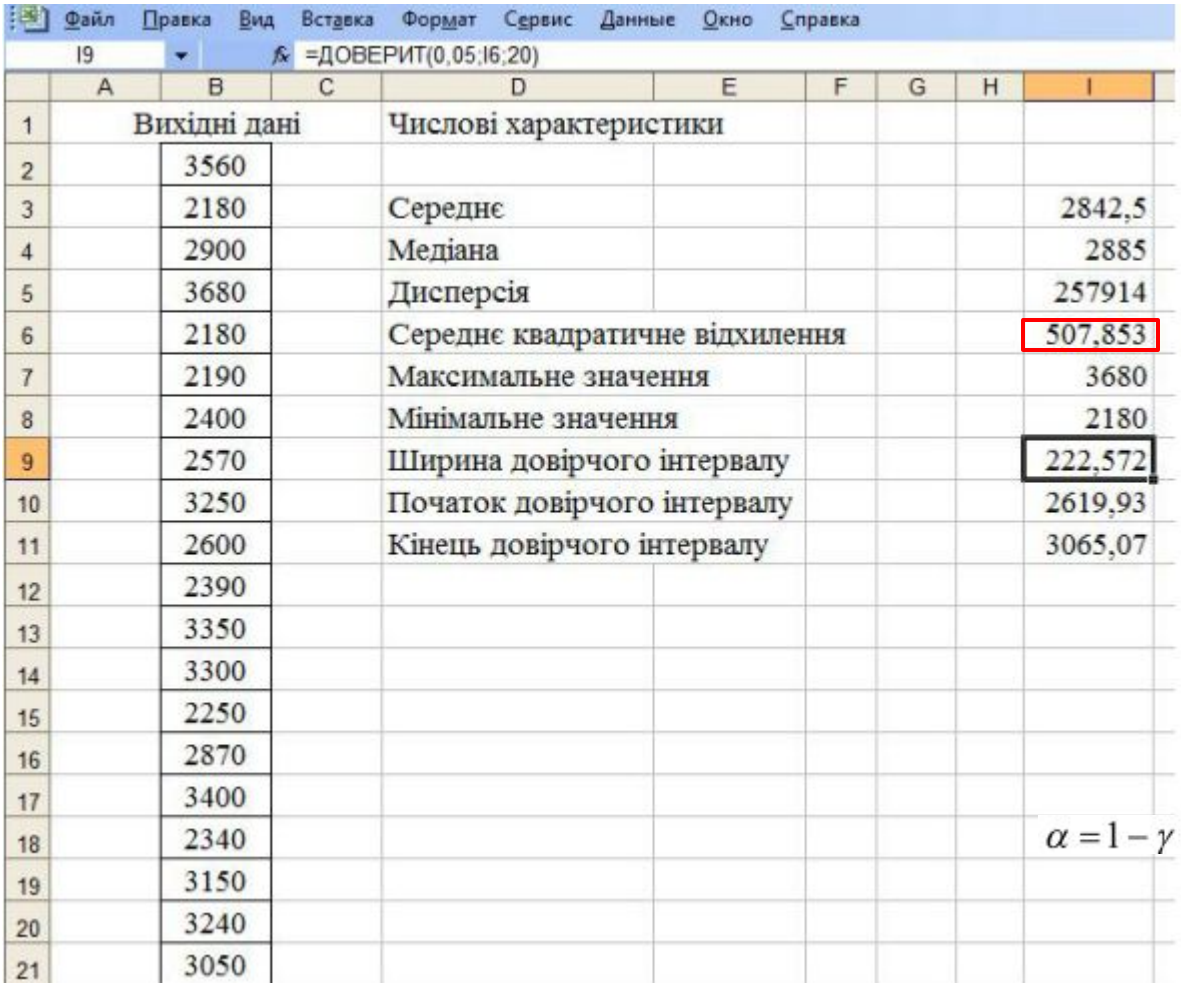
$$s(t,k) = B_k \left(1 + \frac{t^2}{k-1}\right)^{-\frac{k}{2}},$$

СТЬЮДРАСПОБР(вероятность; степени свободы)

НОРМСТОБР (x)

Формула Лапласа

# Приклад застосування Excel




| | 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 | | | | | | | |

=СРЗНАЧ(B2:B21)

=МЕДИАНА(B2:B21)

=ДИСП(B2:B21)

=СТАНДОТКЛОН(B2:B21)

=МАКС(B2:B21)

=МИН(B2:B21)

$$\alpha = 1 - \gamma = 1 - 0,95 = 0,05$$

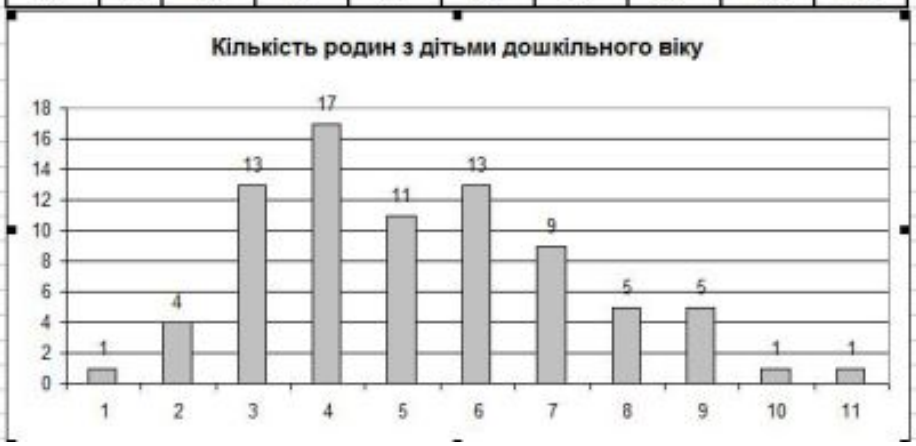
# Приклад побудови гістограми засобами Excel

За даними вибіркового дослідження відома кількість родин з дітьми дошкільного віку у селах деякої області. Побудувати гістограму за даними таблиці.


| | 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 |

# Приклад побудови гістограми засобами Excel

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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 |




Вставка – Диаграмма – Гистограмма, задамо діапазон даних, тобто розраховані частоти і вкажемо групування за стовпцями

# Довірчий інтервал для генерального середнього при відомій генеральної дисперсії

Нехай $X$ – генеральна сукупність, що підкоряється нормальному закону розподілу; $\sigma^2$ – відома генеральна дисперсія; $\{x_1, x_2, ..., x_n\}$ – вибірка з генеральної сукупності об'ємом $n$; $\bar{x}$ – вибіркове середнє. Потрібно знайти довірчий інтервал для генерального середнього $a$ із заданим рівнем надійності $\gamma$.

Шуканий довірчий інтервал знаходиться за формулою:

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

значення $z_{\frac{1-\gamma}{2}}$ ⇨ НОРМСТОБР($\gamma$)

| $\gamma$ | 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 1.4

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

**Розв'язок.** Оскільки $\gamma = 0{,}95$, то $\frac{1-\gamma}{2} = \frac{1-0{,}95}{2} = 0{,}025$.

знайдемо $z_{\frac{1-\gamma}{2}} = z_{0{,}025} = 1{,}96$.

Тоді ширина довірчого інтервалу: $z_{\frac{1-\gamma}{2}} \cdot \frac{\sigma}{\sqrt{n}} = 1{,}96 \cdot \frac{5}{\sqrt{30}} \approx 1{,}79$

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

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 пачкам.

# Довірчий інтервал для генерального середнього при невідомій генеральної дисперсії

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

Шуканий довірчий інтервал знаходиться за формулою:

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

де значення $t_{\frac{1-\gamma}{2},n-1}$ знаходиться з таблиці розподілу Стьюдента, яка є у статистичних довідниках, або за допомогою вбудованої функції Excel СТЬЮДРАСПОБР ($\frac{1-\gamma}{2}$, $n$ - 1). Величина $t_{\frac{1-\gamma}{2},n-1} \cdot \frac{S}{\sqrt{n-1}}$ є шириною довірчого інтервалу.

## Приклад 1.4

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

**Розв'язок.** Оскільки $\gamma = 0{,}95$, то $\frac{1-\gamma}{2} = \frac{1-0{,}95}{2} = 0{,}025$; $n = 30$, тоді $n - 1 = 29$. За допомогою Excel знайдемо $t_{\frac{1-\gamma}{2},n-1} = t_{0{,}025;29}$. Натиснемо $f_x$ у командному рядку, виберемо категорію ***Статистические*** і функцію СТЬЮДРАСПОБР; задамо параметри 0,025 і 29

Отримаємо $t_{0{,}025;29} = 2{,}3638$.

Тоді ширина довірчого інтервала $t_{\frac{1-\gamma}{2},n-1} \cdot \frac{S}{\sqrt{n-1}} = 2{,}3638 \cdot \frac{4}{\sqrt{29}} \approx 1{,}75$ і довірчий інтервал за формулою (1.27):

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

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

# 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 $\gamma$.

Шуканий довірчий інтервал знаходиться за формулою:

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

де значення $z_{\frac{1-\gamma}{2}}$ знаходиться з табл. 1.18 або за допомогою вбудованої функції Excel НОРМСТОБР($\gamma$). Величина $z_{\frac{1-\gamma}{2}} \cdot \sqrt{\frac{w(1-w)}{n}}$ є шириною довірчого інтервалу.

## Приклад 1.5

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

**Розв'язок.** Оскільки $\gamma = 0{,}95$, то $\frac{1-\gamma}{2} = \frac{1-0{,}95}{2} = 0{,}025$. З табл. 1.18 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 інтервала:

$$z_{\frac{1-\gamma}{2}} \cdot \sqrt{\frac{w(1-w)}{n}} = 1{,}96\sqrt{\frac{0{,}075(1-0{,}075)}{2000}} \approx 0{,}012,$$

за формулою (1.28) отримаємо довірчий інтервал:

$$w - 0{,}012 < W < w + 0{,}012;$$
$$0{,}075 - 0{,}012 < W < 0{,}075 + 0{,}012;$$
$$0{,}063 < W < 0{,}087.$$

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 обсягу продукції.

Лекція 2

# Статистичні гіпотези

1. Поняття про статистичні гіпотези
2. Перевірка гіпотези про вид закону розподілу досліджуваної величини
3. Перевірка гіпотез про генеральні середні і дисперсії
4. Перевірка статистичних гіпотез із використанням Microsoft Excel

# Поняття про статистичні гіпотези

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

Типи статистичних гіпотез:

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 розумінні, що при перевірці бажано було б встановити її справедливість. Основній гіпотезі $H_0$ протиставляються інші гіпотези $H_1$ , $H_2$ , …, які називаються альтернативними.

Прийняття основної або однієї 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.

# Поняття про статистичні гіпотези

Якщо сформульовані гіпотези $H_0$ – основна та $H_1$ альтернативна (конкуруюча) і обраний критерій перевірки справедливості основної гіпотези, то прийняття $H_0$ означає відкидання $H_1$, а відкидання $H_0$ означає 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 $\alpha$.

Величина $1-\alpha$ 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 $\beta$.

Величина $1-\beta$ є ймовірністю відкидання невірної гіпотези і називається потужністю критерію.

Чим менше значення рівня значущості, тим менша ймовірність відкинути вірну гіпотезу. Зазвичай рівень значущості обирається дослідником рівним 0,1; 0,05; 0,01 або 0,001. Якщо, наприклад, обраний рівень значущості $\alpha$ =0,01, то ризик відкинути вірну гіпотезу виникає в одному випадку із ста.

# Поняття про статистичні гіпотези

Перевірка статистичних гіпотез здійснюється за такою послідовністю :

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.

# Перевірка гіпотези про вид закону розподілу досліджуваної величини

Перевірка гіпотези про вид закону розподілу досліджуваної величини має велике значення для прикладних досліджень. Необхідність перевірки виникає при виборі критерію, оскільки для багатьох з них висувається вимога нормального розподілу статистичних даних.

Припустимо, що з деякої генеральної сукупності X, яка розглядається як випадкова величина, обрана вибірка

| $x_i$ | $x_1$ | $x_2$ | … | $x_k$ |
|---|---|---|---|---|
| $n_i$ | $n_1$ | $n_2$ | … | $n_k$ |

Отриманий 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 формування вибірки або залежно від виду графіка емпіричної щільності розподілу (гістограми) у випадку неперервної випадкової величини X і полігону частот, якщо величина X дискретна.

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

# Перевірка гіпотези про вид закону розподілу досліджуваної величини

При здійсненні такої заміни немає впевненості, що закон розподілу обраний правильно. Розроблено процедуру, яка дозволяє оцінити степінь відповідності обраного закону даним вибірки.

Критерії здійснення такої перевірки називаються критеріям найбільш відомим з яких є критерій Пірсона x2 (хі-квадрат) $\chi^2 = \sum_{i=1}^{k} \frac{\left(n_i - n_i'\right)^2}{n_i'}$

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

Якщо відповідні теоретичні та емпіричні частоти співпадають, $x^2 = 0$.

Тобто, чим ближче $x^2$ до нуля, тим краще узгоджуються вибіркові дані та обраний теоретичний закон розподілу.

Розраховане значення критерія x2 порівнюється з його критичним значенням , яке знаходиться за статистичними таблицями

# Перевірка гіпотези про вид закону розподілу досліджуваної величини

Перевірка гіпотези про закон розподілу величини X здійснюється за етапами:

1) З генеральної сукупності X формується 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 x 2.
6) Обирається рівень значущості (або рівень довіри ) і знаходиться критичне значення $\chi 2$ .
7) Порівнюються розраховане і критичне значення критерію x 2 і робиться висновок про справедливість запропонованої гіпотези.

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

## Приклад 2.1

За даним інтервальним статистичним рядом знайти закон розподілу випадкової величини X.

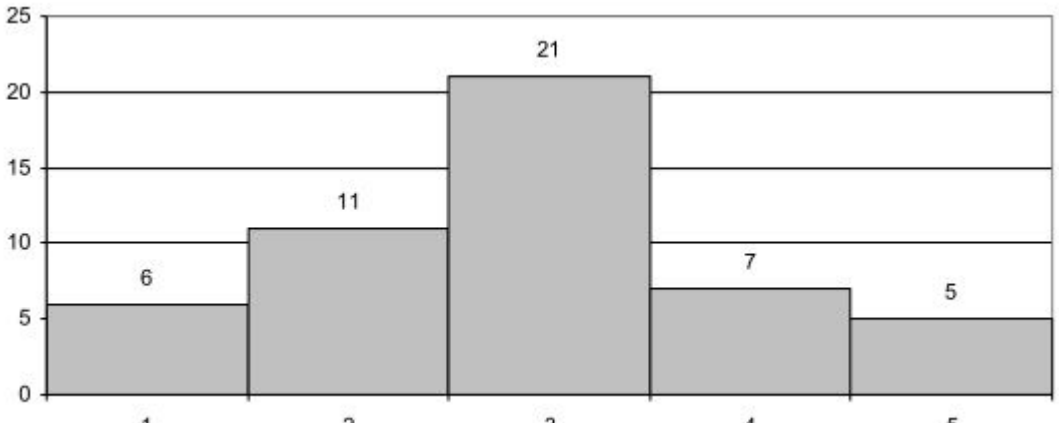
| $[a_i;a_{i+1})$ | [−2;−1,2) | [−1,2;−0,4) | [−0,4;0,4) | [0,4;1,2) | [1,2;2) |
|---|---|---|---|---|---|
| $n_i$ | 6 | 11 | 21 | 7 | 5 |

Розв'язок. Для визначення виду закону розподілу побудуємо гістограму за даними табл.

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.




Щільність розподілу випадкової величини, розподіленої за нормальним законом, має вигляд $f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}$, де $a$ і $\sigma$ – параметри розподілу.

Знайдемо означені параметри, враховуючи, що $\bar{x}=a;\ S^2=\sigma^2$. Розрахунки

## Приклад 2.1

| $[a_i;a_{i+1})$ | $[-2;-1,2)$ | $[-1,2;-0,4)$ | $[-0,4;0,4)$ | $[0,4;1,2)$ | $[1,2;2)$ |
|---|---|---|---|---|---|
| $n_i$ | 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

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

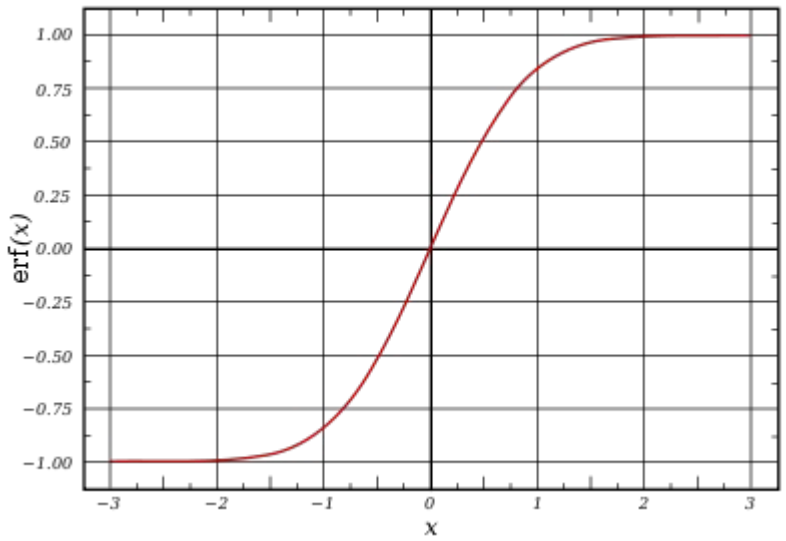
$$S^2=\frac{1}{n}\sum(x_i-\bar{x})^2 n_i=\frac{1}{50}(13,572+5,452+0,194+5,620+14,382)=0,7844;$$

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

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

# Приклад 2.1

Для знаходження значення критерію $\chi^2$ 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 $\Phi$ – функція Лапласа,



## Приклад 2.1

| $[a_i;a_{i+1})$ | [−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 |

$$\chi^2=\sum_{i=1}^{k}\frac{\left(n_i-n_i'\right)^2}{n_i'}\approx 4{,}16$$

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

$$\chi^2<\chi^2_{\alpha,l}$$

гіпотеза $H_0$ про нормальний розподіл приймається, гіпотеза $H_1$ відкидається.

# 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 (коли визначається перевага однієї з технологій виготовлення певної продукції, або наявність підвищення продуктивності праці після внесення змін в процес виробництва, або при перевірці якості продукції).

Здійснення 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:

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

Гіпотеза $H_0$ приймається, якщо розраховане значення $F$ менше критичного значення розподілу Фішера $F_{крит}$, взятого із рівнем значущості $\alpha$ і степенями свободи $l_1$ та $l_2$ для чисельника і знаменника 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 ($\alpha$; $l_1$; $l_2$).

## Приклад 2.2

Відомо 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. Враховуючи, що дані розподілені за нормальним законом, перевірити гіпотезу про рівність дисперсій.

| | 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 табл. є двома вибірками.
Перша – вибірка значень величини $X_1$ – продуктивність праці робітників, що пройшли навчання, Друга – вибірка величини $X_2$ – продуктивність 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).

| $x_i$ | 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 |

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

Зміщена вибіркова дисперсія

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

## Приклад 2.2

| $x_i$ | 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 |

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

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

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

Оскільки $F > F_{\text{крит}}$, то гіпотезу $H_0$ відкидаємо і приймаємо гіпотезу $H_1$ – дисперсії нерівні, тобто вибірки здобуті 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 вимірюються 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 у такій послідовності:

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:

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

де $n_1$, $n_2$ – об'єми вибірок. При цьому $R_1$ – сума рангів меншої за об'ємом вибірки.

# Перевірка гіпотези про рівність генеральних дисперсій. Критерій Зігеля-Тьюкі

Якщо $2R_1 > n_1(n_1 + n_2 + 1) + 1$, $Z$ розраховується за формулою:

$$Z = \frac{2R_1 - n_1(n_1 + n_2 + 1) - 1}{\frac{n_2}{3} \cdot \sqrt{n_1(n_1 + n_2 + 1)}}$$

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

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

6) Обирається рівень значущості $\alpha$.

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

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

## Приклад 2.3

У результаті дослідження надійності приладів двох виробників отримано дані про час (в годинах) безаварійної роботи (табл. 1). Враховуючи, що дані не розподілені за нормальним законом, перевірити гіпотезу про рівність дисперсій.

| Виробник | Час безаварійної роботи | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 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 1 є двома вибірками.

Перша – вибірка значень величини $\mathbf{X_1}$ – час безаварійної роботи приладів виробника 1;

Друга – вибірка величини $\mathbf{X_2}$ – час безаварійної роботи приладів виробника 2.

Сформулюємо гіпотези: $\mathbf{H_0}$ – дисперсії генеральних сукупностей, з яких зроблено вибірки, рівні: $\mathbf{S^1{}_2 = S^2{}_2}$;

$\mathbf{H_1}$ дисперсії не рівні: $\mathbf{S^1{}_2 \neq S^2{}_2}$

Перевіримо справедливість гіпотези $\mathbf{H_0}$ за критерієм Зігеля-Тьюкі.

## Приклад 2.3

Сформуємо об’єднану вибірку, присвоїмо її елементам ранги і знайдемо їх суму. Результати розрахунків оформимо у вигляді таблиці. .підкреслимо елементи першої вибірки.

| 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 |

## Приклад 2.3

Розрахуємо за формулою значення нормальної випадкової величини 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 $\alpha = 0,05$. 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.

# 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:

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

Розраховане значення $t$-критерія порівнюється з критичним значенням $t_{крит}$, де $t_{крит}$ – критичне значення розподілу Стьюдента з параметрами $\frac{\alpha}{2}$ і степенем свободи $l = n_1 + n_2 - 2$, яке надається в статистичних таблицях або знаходиться за допомогою вбудованих функцій SPSS та 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:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Розраховане значення $t$-критерію також порівнюється з критичним значенням $t_{крит}$, але степінь свободи розраховується за формулою:

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

# Перевірка гіпотези про рівність генеральних середніх. Критерій Стьюдента

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

$$t = \frac{\overline{x_d}}{S_d / \sqrt{n-1}}$$

де $\overline{x_d}$ – вибіркове середнє для вибірки різниць, $S_d$ – вибіркове середнє квадратичне відхилення для вибірки різниць, $n$ – об'єм вибірки різниць.

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

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

$$|t| < t_{крит}.$$

## Приклад 2.4

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

Розв'язок. За умовами задачі було зроблено дві вибірки:

Перша – вибірка об'ємом $\mathbf{n_1 = 10}$ значень величини $\mathbf{X_1}$ – часу, потрібного для виготовлення деталей за першою технологією;

Друга – – вибірка об'ємом $\mathbf{n_1 = 16}$ значень величини величини $\mathbf{X_2}$ часу, потрібного для виготовлення деталей за другою технологією.

Відомі вибіркові середні $\bar{x}_1$ = 30 с та $\bar{x}_2$ = 28 с – середній час, необхідний для виготовлення деталей за першою і другою технологіями відповідно. Відомі дисперсії часу для вибірок: 21

$\mathbf{S^2_1 = 1\ с^2}$ **та** $\mathbf{S^2_2 = 2\ с^2}$ **.** 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. Використаємо критерій Фішера. Обчислимо значення **F-критерія:**

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

## Приклад 2.4

Знайдемо критичне значення розподілу Фішера $F_{крит}$: оберемо рівень значущості $\alpha$ =0,05; врахуємо, що степені свободи $l_1 = n_1 - 1 = 9$ та $l_2 = n_2 - 1 = 15$. Тоді FРАСПОБР ($\alpha$; $l_2$; $l_1$) = 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 $\alpha$ =0,05. Тоді

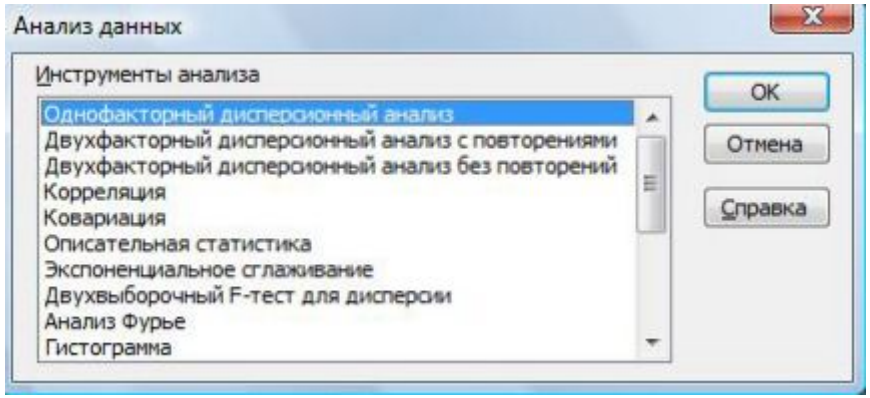
$$t_{крит} = \text{СТЬЮДРАСПОБР}\left(\frac{\alpha}{2}; l\right) = \text{СТЬЮДРАСПОБР}\ (0{,}025;\ 24) = 2{,}39.$$

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

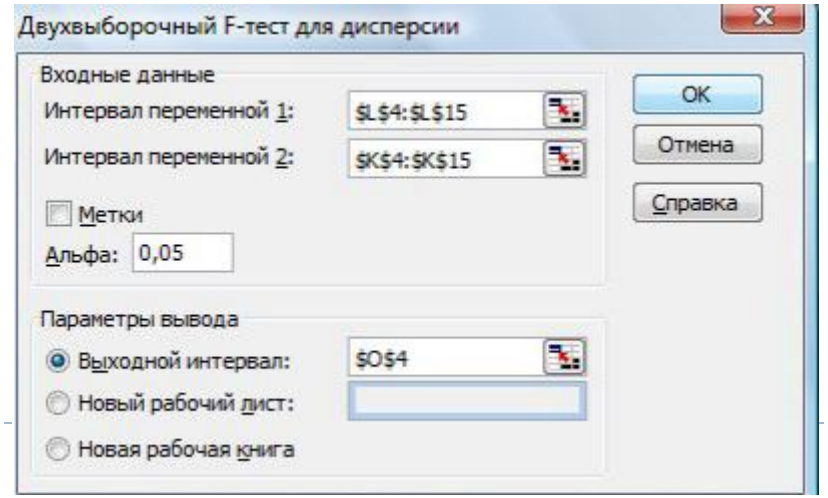
**Висновок:** для вироблення деталей за першою технологією потрібно, у середньому, більше часу.

## Двохвибірковий F-тест для дисперсій

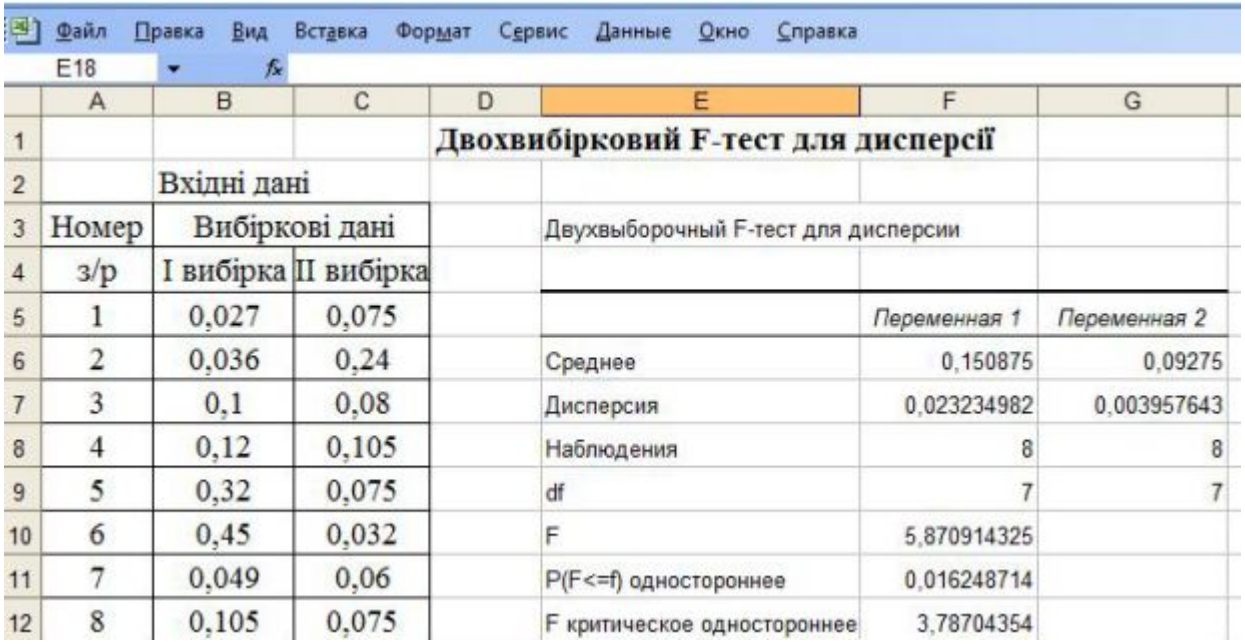
1) Вибрати в меню послідовно пункти Сервис – Анализ данных, після чого з'явиться вікно для вибору інструмента аналізу




2) Вибрати у діалоговому вікні інструмент Двухвыборочный F-тест для дисперсии, після чого з'явиться вікно для вибору параметрів, задати області данних



## Двохвибірковий F-тест для дисперсій



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 | |

# Перевірка статистичних гіпотез із використанням Microsoft Excel

## Двохвибірковий t-тест для середніх

Перевірку гіпотез про рівність генеральних середніх за критерієм Стьюдента можна виконати за допомогою пакета аналізу даних Microsoft Excel. Для здійснення перевірки необхідно у вікні для вибору інструмента аналізу 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.

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 | | | | | | | |

Лекція 3

# Основи кореляційного аналізу

1. Поняття кореляційного зв'язку між досліджуваними величинами. Групування даних для кореляційного аналізу
2. Коефіцієнт кореляції Пірсона
3. Коефіцієнт кореляції Спірмена
4. Множинний та частинний коефіцієнти кореляції
5. Кореляційний аналіз із використанням Microsoft Excel

# Поняття кореляційного зв’язку між досліджуваними величинами

Кореляційний аналіз - математичний апарат для виявлення зв’язків і оцінки
їх сили (тісноти) між ознаками різних об'єктів і явищ.

В багатьох прикладних задачах необхідно виявити залежність між двома властивостями (ознаками) X і Y одного і того ж економічного об’єкта або між
певними ознаками різних об’єктів. Якщо вказані ознаки допускають кількісне
вимірювання, і, з погляду економічної теорії, виходячи з економічної
характеристики об’єкта, ознака Y залежить від ознаки X. Тоді X можна назвати
незалежною змінною або факторною ознакою, а Y – залежною змінною або
результативною ознакою.

Якщо кожному значенню факторної ознаки X відповідає одне і тільки
одне значення результативної ознаки Y, то говорять, що між цими ознаками
існує функціональний зв’язок: $Y = f(X)$

# Поняття кореляційного зв’язку між досліджуваними величинами

Якщо кожному значенню факторної ознаки X відповідає безліч значень результативної ознаки Y, то говорять, що між цими ознаками існує статистичний зв’язок.

Наприклад, якщо $X$ приймає $l$ значень $X=\{x_1, x_2, ..., x_l\}$ і кожному її значенню $x_i$ відповідає множина значень $Y$, тобто:

значенню $x_1$ відповідає множина $\{y_{11}, y_{12}, ..., y_{1m_1}\}$;

значенню $x_2$ відповідає множина $\{y_{21}, y_{22}, ..., y_{2m_2}\}$;

…

значенню $x_l$ відповідає множина $\{y_{l1}, y_{l2}, ..., y_{lm_l}\}$,

то між $X$ та $Y$ існує статистичний зв’язок.

Вивчення статистичного зв'язку є складним і трудомістким процесом, у якому потрібно аналізувати багатовимірні таблиці даних. Тому вивчається не статистичний, а кореляційний зв'язок між X та Y.

## Поняття кореляційного зв'язку між досліджуваними величинами

Якщо кожному значенню факторної ознаки X відповідає певне середнє значення результативної ознаки Y, то говорять, що між цими ознаками існує кореляційний зв'язок. Тобто кореляційний $\overline{Y} = f(X)$ функціональна залежність між зна

Наприклад, якщо $X$ приймає $l$ значень $X = \{x_1, x_2, ..., x_l\}$ і кожному її значенню $x_i$ відповідає середнє множини значень $Y$, тобто:

значенню $x_1$ відповідає $\overline{y}_{x_1} = \frac{y_{11} + y_{12} + ... + y_{1m_1}}{m_1}$;

значенню $x_2$ відповідає $\overline{y}_{x_2} = \frac{y_{21} + y_{22} + ... + y_{1m_2}}{m_2}$;

...

значенню $x_l$ відповідає $\overline{y}_{x_l} = \frac{y_{l1} + y_{l2} + ... + y_{lm_l}}{m_l}$,

то між X та Y існує кореляційний зв'язок.

Основними задачами кореляційного аналізу є:

- вивчення сили зв'язку між двома і більше ознаками досліджуваного об'єкта;
- встановлення факторів, що найбільше впливають на результативну ознаку;
- виявлення невідомих причинно-наслідкових зв'язків між ознаками

# Групування даних для кореляційного аналізу

Вибіркові дані для вивчення кореляційного зв'язку між ознаками $X$ та $Y$ 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, $i=\overline{1,n}$.

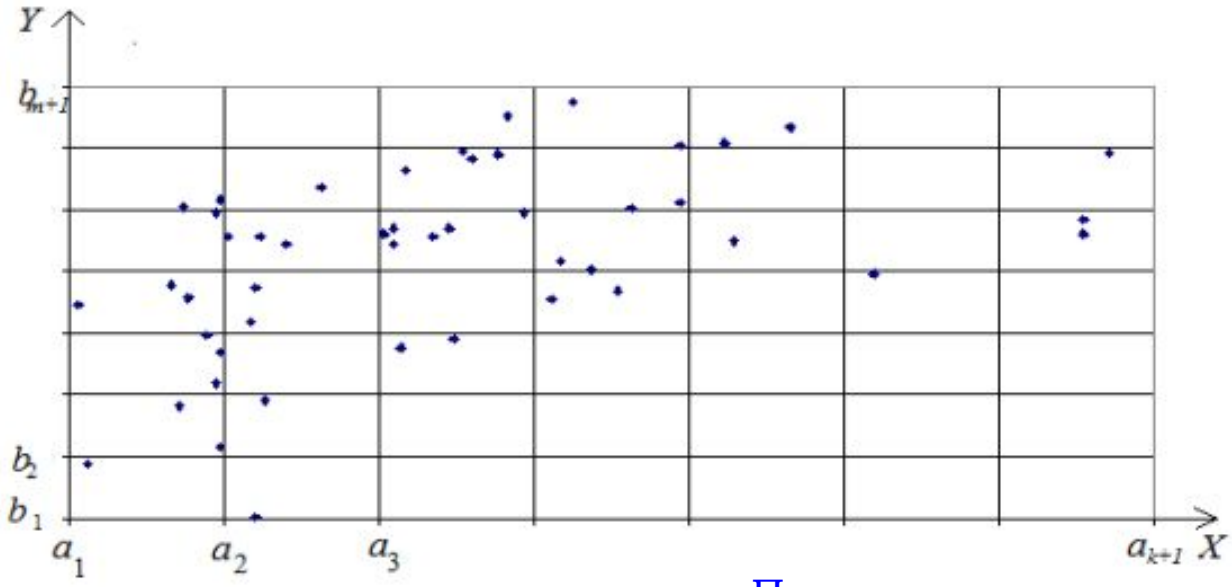
Якщо кількість пар значень достатньо велика (принаймні $n>20$), то для зручності розрахунків дані групуються.

Для групування даних необхідно:

1) Розбити множини значень $X$ та $Y$ на інтервали, використовуючи формулу Стерджеса , кількість інтервалів для $X$ та $Y$ може бути різною (позначення: $k$ – кількість інтервалів для $X$; $m$ – кількість інтервалів для $Y$).

2) Зобразити дані графічно: побудувати на площині точки з координатами $(x_i;y_j)$. В результаті отримується площина, розбита на прямокутники, в кожному з яких може бути множина точок . Вказане графічне зображення вибіркових даних називається **полем кореляції**.

# Групування даних для кореляційного аналізу




Поле кореляції

3) Побудувати кореляційну таблицю . В першому рядку, розбитому на дві частини, записуються інтервали $[a_i;a_{i+1})$ для $X$ та їх середини $x_i$. У першому стовпці, розбитому на дві частини, записуються інтервали $[b_j;b_{j+1})$ для $Y$ та їх середини $y_j$. В центральній частині таблиці записуються частоти $n_{ij}$ – кількість точок, що потратили в прямокутник, обмежений по $X$ інтервалом $[a_i;a_{i+1})$ і по $Y$ інтервалом $[b_j;b_{j+1})$. В останньому рядку таблиці записуються частоти $n_i$ для $X$ – кількості точок, що потратили в прямокутники, які відповідають інтервалу $[a_i;a_{i+1})$, тобто $n_i=\sum_{j=1}^{m} n_{ij}$ – сума частот $n_{ij}$ в стовпці з номером $i$. В останньому стовпці таблиці записуються частоти $n_j$ для $Y$ – кількості точок, що потрапили в прямокутники, які відповідають інтервалу $[b_j;b_{j+1})$, тобто $n_j=\sum_{i=1}^{k} n_{ij}$ – сума частот $n_{ij}$ в рядку з номером $j$.

Кореляційну таблицю можна розглядати як своєрідний подвійний статистичний ряд.

# Групування даних для кореляційного аналізу

| $X$ (інтервали і їх середини) / $Y$ (інтервали і їх середини) | | $[a_1;a_2)$ | $[a_2;a_3)$ | … | $[a_k;a_{k+1})$ | $n_j = \sum_{i=1}^{k} n_{ij}$ |
|---|---|---|---|---|---|---|
| | | $x_1$ | $x_2$ | … | $x_k$ | |
| $[b_1;b_2)$ | $y_1$ | $n_{11}$ | $n_{21}$ | … | $n_{k1}$ | $n_1$ |
| $[b_2;b_3)$ | $y_2$ | $n_{12}$ | $n_{22}$ | … | $n_{k2}$ | $n_2$ |
| … | … | … | … | … | … | … |
| $[b_m;b_{m+1})$ | $y_m$ | $n_{1m}$ | $n_{2m}$ | … | $n_{km}$ | $n_m$ |
| $n_i = \sum_{j=1}^{m} n_{ij}$ | | $n_1$ | $n_2$ | … | $n_k$ | |

4) За даними кореляційної таблиці будується ряд, що відображає залежність середнього значення $Y$ від $X$ . В першому рядку таблиці записуються середини інтервалів $x_i$. В другому – відповідні середні значення $\overline{y}_{x_i}$, що знаходяться за формулами:

$$\overline{y}_{x_1} = \frac{y_1 n_{11} + y_2 n_{12} + \ldots + y_m n_{1m}}{n_1}; \qquad \overline{y}_{x_2} = \frac{y_1 n_{21} + y_2 n_{22} + \ldots + y_m n_{2m}}{n_2}; \qquad \ldots \qquad ;$$

$$\overline{y}_{x_k} = \frac{y_1 n_{k1} + y_2 n_{k2} + \ldots + y_m n_{km}}{n_k}.$$

| $x_i$ | $x_1$ | $x_2$ | … | $x_k$ |
|---|---|---|---|---|
| $\overline{y}_{x_i}$ | $\overline{y}_{x_1}$ | $\overline{y}_{x_2}$ | … | $\overline{y}_{x_k}$ |
| $n_i$ | $n_1$ | $n_2$ | … | $n_k$ |

В результаті отримується статистичний ряд, що містить значення $X$, відповідні середні значення $Y$ та частоти. За даними такого ряду проводиться кореляційний аналіз.

# Коефіцієнт кореляції Пірсона

Для оцінки тісноти (або сили) зв'язку між $X$ та $Y$ існує коефіцієнт кореляції. У випадку, коли між $X$ та $Y$ існує лінійний зв'язок та вибіркові дані розподілені за нормальним законом, використовується **коефіцієнт кореляції Пірсона**, який ще називається параметричним коефіцієнтом кореляції.

Коефіцієнт кореляції Пірсона розраховується за формулою:

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{S_x \cdot S_y}, \qquad (3.1)$$

де $\bar{x}$ – вибіркове середнє величини $X$;

$\bar{y}$ – вибіркове середнє величини $Y$;

$\overline{xy}$ – вибіркове середнє величини $XY$;

$S_x$ – вибіркове середнє квадратичне відхилення величини $X$;

$S_y$ – вибіркове середнє квадратичне відхилення величини $Y$.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i; \quad \bar{y} = \frac{1}{n}\sum_{j=1}^{m} y_j n_j; \quad \overline{xy} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{m} x_i y_j n_{ij};$$

$$S_x = \sqrt{\frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i - \left(\frac{1}{n}\sum_{i=1}^{k} x_i n_i\right)^2}; \quad S_y = \sqrt{\frac{1}{n}\sum_{j=1}^{m} y_j^2 n_j - \left(\frac{1}{n}\sum_{j=1}^{m} y_j n_j\right)^2},$$

# Коефіцієнт кореляції Пірсона

$$r=\frac{n\sum_{i=1}^{k}\sum_{j=1}^{m}x_i y_j n_{ij}-\left(\sum_{i=1}^{k}x_i n_i\right)\left(\sum_{j=1}^{m}y_j n_j\right)}{\sqrt{n\sum_{i=1}^{k}x_i^2 n_i-\left(\sum_{i=1}^{k}x_i n_i\right)^2}\sqrt{n\sum_{j=1}^{m}y_j^2 n_j-\left(\sum_{j=1}^{m}y_j n_j\right)^2}}$$

$$r=\frac{n\sum_{i=1}^{n}x_i y_i-\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{\sqrt{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}\sqrt{n\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2}}$$

**Властивості коефіцієнта кореляції Пірсона**

1) Коефіцієнт кореляції Пірсона приймає значення на проміжку $[-1;1]$, тобто $-1\le r\le 1$.

2) Якщо $0{,}3\le|r|\le 0{,}5$, то зв'язок вважається слабким; якщо $0{,}5<|r|\le 0{,}7$, то зв'язок вважається середнім; $0{,}7<|r|\le 1$, то зв'язок вважається сильним.

3) Якщо $r>0$, то зв'язок називається додатнім, тобто зі збільшенням значень $X$ значення $Y$ також збільшуються. Якщо $r<0$, то зв'язок називається від'ємним, тобто зі збільшенням значень $X$ значення $Y$ зменшуються.

# Коефіцієнт кореляції Пірсона

Оскільки сила зв'язку між $X$ та $Y$ оцінюється за вибірковими даними, то необхідна перевірка її **статистичної значущості**, тобто оцінка можливості розповсюдити отримані результати на всю генеральну сукупність.

Перевірка статистичної значущості коефіцієнта кореляції Пірсона здійснюється за допомогою так званої $t$-статистики, яка розраховується за формулою:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}. \qquad (3.4)$$

Розраховане значення $t$-статистики порівнюється з критичним значенням $t_{крит}$. $t_{крит}$ – табличне значення розподілу Стьюдента, яке також можна знайти за допомогою вбудованої статистичної функції Excel СТЬЮДРАСПОБР ($\alpha$ ; $l$), де $\alpha$ – обраний дослідником рівень значущості, $l$ – степінь свободи, $l = n-2$.

Якщо розраховане значення $t$-статистики більше критичного $|t| > t_{крит}$, то коефіцієнт кореляції вважається значущим на обраному рівні $\alpha$.

**Зауваження.** Слід пам'ятати, що коефіцієнт кореляції Пірсона показує силу лінійного зв'язку. Якщо між $X$ та $Y$ існує сильний нелінійний зв'язок, коефіцієнт кореляції Пірсона може дорівнювати нулю.

## Приклад 3.1

За наявними даними про рівнем оплати праці X і продуктивності праці Y для 14 однотипних підприємств оцінити тісноту зв'язку між X і Y. Визначити можливість розповсюдження результатів розрахунків на всі підприємства такого типу.

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $X$ | 32 | 30 | 36 | 40 | 41 | 47 | 56 | 54 | 60 | 55 | 61 | 67 | 69 | 76 |
| $Y$ | 20 | 24 | 28 | 30 | 31 | 33 | 34 | 37 | 38 | 40 | 41 | 43 | 45 | 48 |

Розв'язок. Дані є вибіркою значень X і відповідних значень Y.
Оскільки кількість даних невелика ($n$ =14), то їх можна не групувати. Для оцінки тісноти зв'язку між X і Y розрахуємо коефіцієнт кореляції Пірсона. Розрахунки для зручності оформимо у вигляді таблиці

## Приклад 3.1

| $x_i$ | $y_i$ | $x_i^2$ | $y_i^2$ | $x_i y_i$ |
|---|---|---|---|---|
| 32 | 20 | 1024 | 400 | 640 |
| 30 | 24 | 900 | 576 | 720 |
| 36 | 28 | 1296 | 784 | 1008 |
| 40 | 30 | 1600 | 900 | 1200 |
| 41 | 31 | 1681 | 961 | 1271 |
| 47 | 33 | 2209 | 1089 | 1551 |
| 56 | 34 | 3136 | 1156 | 1904 |
| 54 | 37 | 2916 | 1369 | 1998 |
| 60 | 38 | 3600 | 1444 | 2280 |
| 55 | 40 | 3025 | 1600 | 2200 |
| 61 | 41 | 3721 | 1681 | 2501 |
| 67 | 43 | 4489 | 1849 | 2881 |
| 69 | 45 | 4791 | 2025 | 3105 |
| 76 | 48 | 5779 | 2304 | 3848 |
| Суми | | | | |
| 724 | 492 | 40134 | 18138 | 26907 |

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{j=1}^{n} y_j\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{j=1}^{n} y_j^2 - \left(\sum_{j=1}^{n} y_j\right)^2}} = \frac{14 \cdot 26907 - 724 \cdot 492}{\sqrt{14 \cdot 40134 - 724^2}\sqrt{14 \cdot 18138 - 492^2}}$$

$$= \frac{20490}{\sqrt{37700}\sqrt{11868}} \approx 0{,}969.$$

За значенням коефіцієнта кореляції можна зробити висновок, що між X і Y існує сильний додатній зв'язок.

## Приклад 3.1

Перевіримо статистичну значущість знайденого коефіцієнта кореляції Пірсона.

Розрахуємо t-статистику за формулою

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0{,}969\sqrt{14-2}}{\sqrt{1-0{,}969^2}} \approx 13{,}59.$$

Знайдемо $t_{крит}$, враховуючи, що $l = n - 2 = 14 - 2 = 12$. Оберемо рівень значущості $\alpha = 0{,}01$. Тоді $t_{крит}$ = СТЬЮДРАСПОБР (0,01; 12) = 3,055.

Оскільки розраховане значення $t$-статистики більше критичного $13{,}59 > 3{,}055$, то коефіцієнт кореляції можна вважати значущим на обраному рівні $\alpha = 0{,}01$.

Висновок.

Між рівнем механізації праці та її продуктивністю на підприємствах, що досліджувалися, існує сильний додатній зв'язок: чим більше рівень механізації праці, тим вище її продуктивність. Висновок дійсний для всіх підприємств такого типу.

# Коефіцієнт кореляції Спірмена

Для оцінки сили зв'язку між $X$ та $Y$ у випадку, коли між $X$ та $Y$ існує нелінійний зв'язок або вибіркові дані не розподілені за нормальним законом, варто використовувати коефіцієнт кореляції Спірмена.

Коефіцієнт кореляції Спірмена розраховується за формулою:

$$r_S(X,Y) = 1 - \frac{6\sum_{i=1}^{n} d_i^2 + T_X + T_Y}{n(n^2-1)},$$

де $n$ – кількість пар вибіркових даних;

$d_i$ – різниці між рангами $i$-го значення $X$ та відповідного значення $Y$;

$T_X, T_Y$ – поправки, пов'язані з однаковими рангами; розраховуються за формулами:

$$T_X = \frac{\sum_{i=1}^{L_X}\left(T_{Xi}^3 - T_{Xi}\right)}{12}; \quad T_Y = \frac{\sum_{i=1}^{L_Y}\left(T_{Yi}^3 - T_{Yi}\right)}{12},$$

де $L_X, L_Y$ – кількість зв'язок (груп однакових рангів);

$T_{Xi}, T_{Yi}$ – розміри $i$-тих зв'язок (кількість елементів в них).

Статистична значущість коефіцієнта кореляції Спірмена перевіряється так, як і коефіцієнта кореляції Пірсона.

## Приклад 3.2

Вивчається залежність між продуктивністю праці робітників X (тис. грн.) та їх емоційним відношенням до своєї професійної діяльності Y (бали). Відповідні дані подано у табл. Оцінити силу зв'язку між досліджуваними факторами за коефіцієнтом кореляції Спірмена. Перевірити його статистичну значущість.

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $X$ | 52 | 37 | 32 | 26 | 53 | 31 | 36 | 32 | 54 | 64 | 47 | 35 | 34 | 28 | 36 |
| $Y$ | 16 | 12 | 5 | 4 | 17 | 6 | 15 | 7 | 13 | 20 | 10 | 10 | 10 | 5 | 19 |

## Приклад 3.2

**Розв'язок.** Дані є вибірковими парами значень $(x_i; y_i)$, $i = \overline{1,n}$; $n$ – кількість пар, $n = 15$. Знайдемо коефіцієнт кореляції Спірмена, необхідні розрахунки оформимо у вигляді таблиці , використовуючи позначення: $d_{x_i}$ – ранг $x_i$, $d_{y_i}$ – ранг $y_i$.

| $x_i$ | 52 | 37 | 32 | 26 | 53 | 31 | 36 | 32 | 54 | 64 | 47 | 35 | 34 | 28 | 36 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $y_i$ | 16 | 12 | 5 | 4 | 17 | 6 | 15 | 7 | 13 | 20 | 10 | 10 | 10 | 5 | 19 |
| $d_{x_i}$ | 12 | 10 | 4,5 | 1 | 13 | 3 | 8,5 | 4,5 | 14 | 15 | 11 | 7 | 6 | 2 | 8,5 |
| $d_{y_i}$ | 12 | 9 | 2,5 | 1 | 13 | 4 | 11 | 5 | 10 | 15 | 7 | 7 | 7 | 2,5 | 14 |
| $d_i$ | 0 | −1 | −2 | 0 | 0 | 1 | 2,5 | 0,5 | −4 | 0 | −4 | 0 | 1 | 0,5 | 5,5 |
| $d_i^2$ | 0 | 1 | 4 | 0 | 0 | 1 | 6,25 | 0,25 | 16 | 0 | 16 | 0 | 1 | 0,25 | 30,25 |

Пояснимо, як заповнюється рядок 3: знаходимо найменше зі значень $x_i$ (це 26) та присвоюємо йому ранг 1; знаходимо наступне найменше (це 28) і присвоюємо йому ранг 2; наступним найменшим є 31, йому присвоюємо ранг 3; наступними найменшими є два значення 32, якщо б вони були різними, то їм би присвоїли ранги 4 і 5, але оскільки вони однакові, то присвоюємо їм середній ранг $\frac{4+5}{2} = 4{,}5$; і т. д.

## Приклад 3.2

Знаходимо суму квадратів різниць рангів: $\sum_{i=1}^{15} d_i^2 = 1 + 4 + 1 + 6{,}25 + 0{,}25 + + 16 + 16 + 1 + 0{,}25 + 30{,}25 = 76.$

Знаходимо поправки, що пов'язані з однаковими рангами. В стрічці рангів $d_{x_i}$ є дві групи однакових рангів, в першій з них 2 елемента, в другій теж два. Отже, $L_X = 2$, $T_{X_1} = 2$, $T_{X_2} = 2$.

В стрічці рангів $d_{y_i}$ є дві групи однакових рангів, в першій з них 2 елемента, в другій – три елемента. Отже, $L_Y = 2$, $T_{Y_1} = 2$, $T_{Y_2} = 3$.

Підставимо отримані дані в формули (3.6) і знайдемо поправки:

$$T_X = \frac{\sum_{i=1}^{L_X}\left(T_{Xi}^3 - T_{Xi}\right)}{12} = \frac{\left(2^3 - 2\right) + \left(2^3 - 2\right)}{12} = 1;$$

$$T_Y = \frac{\sum_{i=1}^{L_Y}\left(T_{Yi}^3 - T_{Yi}\right)}{12} = \frac{\left(2^3 - 2\right) + \left(3^3 - 3\right)}{12} = 2{,}5.$$

$$r_S(X,Y) = 1 - \frac{6\sum_{i=1}^{n} d_i^2 + T_X + T_Y}{n(n^2 - 1)},$$

## Приклад 3.2

Обчислимо коефіцієнт кореляції Спірмена за формулою

$$r_S(X,Y) = 1 - \frac{6\sum_{i=1}^{n} d_i^2 + T_X + T_Y}{n(n^2-1)} = 1 - \frac{6 \cdot 76 + 1 + 2{,}5}{15(15^2-1)} \approx 1 - 0{,}14 = 0{,}86.$$

Згідно значення коефіцієнта кореляції можна зробити висновок, що між $X$ та $Y$ існує сильний додатній зв'язок.

Перевіримо статистичну значущість знайденого коефіцієнта кореляції.

Розрахуємо $t$-статистику за формулою

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0{,}86\sqrt{15-2}}{\sqrt{1-0{,}86^2}} \approx 6{,}17.$$

Знайдемо $t_{крит}$, враховуючи, що $l = n - 2 = 15 - 2 = 13$. Оберемо рівень значущості $\alpha = 0{,}001$. Тоді $t_{крит}$ = СТЬЮДРАСПОБР (0,001; 13) = 4,22.

Оскільки розраховане значення $t$-статистики більше критичного $6{,}17 > 4{,}22$, то коефіцієнт кореляції можна вважати значущим на обраному рівні $\alpha = 0{,}001$.

Висновок. Між продуктивністю праці та емоційним відношенням працівника до професійної діяльності існує сильний додатній зв'язок.

*Висновок дійсний для всієї генеральної сукупності, з якої було зроблено вибірку.*

# Множинний та частинний коефіцієнти кореляції

У випадку, коли досліджуваний об'єкт або явище характеризується більш ніж двома ознаками $X_1, X_2, \dots, X_k$, необхідно вивчати множинні залежності. Для оцінки сили зв'язку між певною ознакою $X_i$ та усіма іншими ознаками використовують **множинний коефіцієнт кореляції**, який позначається $R_i$.

Для розрахунку множинного коефіцієнта кореляції необхідно:

1) Побудувати матрицю парних коефіцієнтів кореляції $r_{ij}$, $i=\overline{1,k}$ між ознаками $X_i$ та $X_j$:

$$A=\begin{pmatrix} 1 & r_{12} & \dots & r_{1k} \\ r_{21} & 1 & \dots & r_{2k} \\ \dots & \dots & \dots & \dots \\ r_{k1} & r_{k2} & \dots & r_{kk} \end{pmatrix}.$$

2) Знайти визначник $|A|$ матриці $A$ та алгебраїчне доповнення $A_{ii}$ елемента $r_{ii}$ цієї матриці.

3) Розрахувати множинний коефіцієнт кореляції за формулою:

$$R_i=\sqrt{1-\frac{|A|}{A_{ii}}}.$$

# Множинний та частинний коефіцієнти кореляції

Перевірка статистичної значущості множинного коефіцієнта кореляції здійснюється за допомогою $t$-статистики, яка розраховується за формулою:

$$t = \frac{R^2 (n-k)}{(1-R^2)(k-1)},$$

де $n$ – кількість взаємопов'язаних значень ознак $X_i$, $i = \overline{1,k}$.

Розраховане значення $t$-статистики порівнюється з критичним значенням $F_{крит}$. $F_{крит}$ – табличне значення розподілу Фішера, яке також можна знайти за допомогою вбудованої статистичної функції Excel FРАСПОБР ($\alpha$; $l_1$; $l_2$), де $\alpha$ – обраний дослідником рівень значущості; $l_1$, $l_2$ – степені свободи: $l_1 = k - 1$, $l_2 = n - k$.

Якщо розраховане значення $t$-статистики більше критичного $|t| > F_{крит}$, то множинний коефіцієнт кореляції вважається значущим на обраному рівні значущості $\alpha$.

У випадку, коли необхідно дослідити кореляційний зв'язок між ознаками $X_i$ та $X_j$, $i = \overline{1,k}$, $j = \overline{1,k}$, із множини ознак $X_1, X_2, \ldots, X_k$ досліджуваного об'єкта або явища, який не залежить від впливу інших ознак, розраховується **частинний коефіцієнт кореляції**, який позначається $R_{ij}$.

# Множинний та частинний коефіцієнти кореляції

Для розрахунку частинного коефіцієнта кореляції необхідно:

1) Побудувати матрицю парних коефіцієнтів кореляції $A$.

2) Знайти алгебраїчні доповнення $A_{ii}, A_{jj}, A_{ij}$ елементів $r_{ii}, r_{jj}, r_{ij}$ відповідно.

3) Розрахувати частинний коефіцієнт кореляції за формулою: $R_{ij} = \frac{-A_{ij}}{\sqrt{A_{ii}A_{jj}}}$.

Перевірка статистичної значущості частинного коефіцієнта кореляції здійснюється за допомогою $t$-статистики, яка розраховується за формулою:

$$t = \frac{R_{ij}\sqrt{n-k+2}}{\sqrt{1-R_{ij}^{\,2}}},$$

де $n$ – кількість взаємопов'язаних значень ознак $X_i, i = \overline{1,k}$.

Розраховане значення $t$-статистики порівнюється з критичним значенням $t_{крит}$. $t_{крит}$ – табличне значення розподілу Стьюдента, яке також можна знайти за допомогою вбудованої статистичної функції Excel СТЬЮДРАСПОБР ($\alpha$ ; $l$), де $\alpha$ – обраний дослідником рівень значущості, $l$ – степінь свободи, $l = n - k + 2$.

Якщо розраховане значення $t$-статистики більше критичного $|t| > t_{крит}$, то частинний коефіцієнт кореляції вважається значущим на обраному рівні значущості $\alpha$.

**Зауваження. 1)** Вважається, що для коректного використання множинного і частинного коефіцієнтів кореляції необхідно, щоб вибіркові дані мали сумісний нормальний розподіл, однак перевірка цієї умови на практиці зазвичай не виконується, оскільки пов'язана зі значними труднощами у розрахунках.

**2)** Замість парного коефіцієнта кореляції Пірсона можна використовувати також парний коефіцієнт кореляції Спірмена.

**3)** Кореляційна матриця завжди симетрична відносно головної діагоналі, оскільки $r_{ij} = r_{ji}, i = \overline{1,k}, j = \overline{1,k}$. Елементи головної діагоналі завжди дорівнюють 1, оскільки вони є коефіцієнтами кореляції $X_i$ та $X_i$.

## Приклад 3.3

Для вивчення залежності отриманого доходу Z від мотивованості працівників X (бали) і величини інвестицій Y було проведено дослідження шести малих підприємств (табл).

Визначити силу зв'язку між Z та X та Y, використовуючи множинний коефіцієнт кореляції. Порівняти силу зв'язку між Z та X, між Z та Y за части

| | | | | | | |
|---|---|---|---|---|---|---|
| $X$ | 26 | 35 | 36 | 40 | 41 | 45 |
| $Y$ | 2,1 | 2,3 | 2,4 | 2,6 | 2,9 | 3 |
| $Z$ | 18 | 21 | 22,1 | 25,3 | 28 | 28,5 |

**Розв'язок.** За умовою задачі, необхідно для об'єкта, що характеризується трьома ознаками $X$, $Y$ та $Z$ ($k$=3), розрахувати множинний коефіцієнт кореляції $R_Z$ і частинні коефіцієнти кореляції $R_{XZ}$ та $R_{YZ}$ на основі шести взаємопов'язаних трійок вибіркових даних $(x_i, y_i, z_i)$, $i = \overline{1,n}$, $n = 6$.

## Приклад 3.3

| Розрахункова таблиця | | | | | | | Суми |
|---|---|---|---|---|---|---|---|
| $x_i$ | 26 | 35 | 36 | 40 | 41 | 45 | 223 |
| $y_i$ | 2,1 | 2,3 | 2,4 | 2,6 | 2,9 | 3 | 15,3 |
| $z_i$ | 18 | 21 | 22,1 | 25,3 | 28 | 28,5 | 142,9 |
| $x_i^2$ | 676 | 1225 | 1296 | 1600 | 1681 | 2025 | 8503 |
| $y_i^2$ | 4,41 | 5,29 | 5,76 | 6,76 | 8,41 | 9 | 39,63 |
| $z_i^2$ | 324 | 441 | 488,41 | 640,09 | 784 | 812,25 | 3489,75 |
| $x_i\, y_i$ | 54,6 | 80,5 | 86,4 | 104 | 118,9 | 135 | 579,4 |
| $x_i\, z_i$ | 468 | 735 | 795,6 | 1012 | 1148 | 1282,5 | 5441,1 |
| $y_i\, z_i$ | 37,8 | 48,3 | 53,04 | 65,78 | 81,2 | 85,5 | 371 |

## Приклад 3.3

$$r_{XY} = r_{YX} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}} = \frac{6 \cdot 579{,}4 - 223 \cdot 15{,}3}{\sqrt{6 \cdot 8503 - 223^2}\sqrt{6 \cdot 39{,}63 - 15{,}3^2}} \approx$$

$$\approx 0{,}935;$$

$$r_{XZ} = r_{ZX} = \frac{n\sum_{i=1}^{n} x_i z_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} z_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} z_i^2 - \left(\sum_{i=1}^{n} z_i\right)^2}} = \frac{6 \cdot 5441{,}1 - 223 \cdot 142{,}9}{\sqrt{6 \cdot 8503 - 223^2}\sqrt{6 \cdot 3489{,}75 - 142{,}9^2}} \approx$$

$$\approx 0{,}954;$$

$$r_{YZ} = r_{ZY} = \frac{n\sum_{i=1}^{n} y_i z_i - \left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} z_i\right)}{\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}\sqrt{n\sum_{i=1}^{n} z_i^2 - \left(\sum_{i=1}^{n} z_i\right)^2}} = \frac{6 \cdot 371{,}62 - 15{,}3 \cdot 142{,}9}{\sqrt{6 \cdot 39{,}63 - 15{,}3^2}\sqrt{6 \cdot 3489{,}75 - 142{,}9^2}} \approx$$

$$\approx 0{,}991;$$

## Приклад 3.3

$$A=\begin{pmatrix} 1 & r_{12} & \dots & r_{1k} \\ r_{21} & 1 & \dots & r_{2k} \\ \dots & \dots & \dots & \dots \\ r_{k1} & r_{k2} & \dots & r_{kk} \end{pmatrix}.$$

Таким чином, кореляційна матриця має вигляд:

$$A=\begin{pmatrix} 1 & 0,935 & 0,954 \\ 0,935 & 1 & 0,991 \\ 0,954 & 0,991 & 1 \end{pmatrix}.$$

Знайдемо визначник $|A|$ матриці $A$ та алгебраїчне доповнення $A_{ZZ}=A_{33}$:

$$|A|=\begin{vmatrix} 1 & 0,935 & 0,954 \\ 0,935 & 1 & 0,991 \\ 0,954 & 0,991 & 1 \end{vmatrix}=1+2\cdot 0,935\cdot 0,991\cdot 0,954-0,954^2-0,991^2-$$

$$-0,935^2\approx 0,0015;$$

$$A_{ZZ}=A_{33}=(-1)^{3+3}\begin{vmatrix} 1 & 0,935 \\ 0,935 & 1 \end{vmatrix}=1-0,935^2\approx 0,1258;$$

тоді $R_Z=R_3=\sqrt{1-\dfrac{|A|}{A_{33}}}=\sqrt{1-\dfrac{0,0015}{0,1258}}\approx 0,994$. Значення множинного коефіцієнта кореляції $R_z$ показує, що величина $Z$ тісно пов'язана з $X$ та $Y$.

$$\Delta = \begin{vmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{vmatrix} \begin{matrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \\ a_{3,1} & a_{3,2} \end{matrix} - \begin{vmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{vmatrix} \begin{matrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \\ a_{3,1} & a_{3,2} \end{matrix}$$

Приклад 3.3+

правило Саррюса

$$A = \begin{pmatrix} 7 & -3 & 5 \\ 1 & 4 & -1 \\ 3 & -4 & -2 \end{pmatrix}.$$

$$\det A = 7 \cdot 4 \cdot (-2) + (-3) \cdot (-1) \cdot 3 + 5 \cdot 1 \cdot (-4) -$$
$$-(5 \cdot 4 \cdot 3 + (-3) \cdot 1 \cdot (-2) + 7 \cdot (-1) \cdot (-4)) =$$
$$= -56 + 9 - 20 - 60 - 6 - 28 = -161.$$

$$|A| = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix} =$$

$$= a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31} - a_{12}a_{21}a_{33} - a_{11}a_{23}a_{32}.$$

## Приклад 3.3

Перевіримо статистичну значущість множинного коефіцієнта кореляції $R_Z$. Знайдемо t-статистику за формулою

$$t = \frac{R^2(n-k)}{(1-R^2)(k-1)} = \frac{0{,}994^2(6-3)}{(1-0{,}994^2)(3-1)} \approx 124{,}09.$$

Знайдемо $F_{крит}$, враховуючи, що $l_1 = k-1 = 3-1 = 2$; $l_2 = n-k = 6-3 = 3$. Оберемо рівень значущості $\alpha = 0{,}01$. Тоді $F_{крит}$=FРАСПОБР (0,01; 2; 3)=30,82. Оскільки $t > F_{крит}$, то множинний коефіцієнт кореляції $R_Z$ є статистично значущим на рівні значущості $\alpha = 0{,}01$.

Для обчислення частинних коефіцієнтів кореляції $R_{XZ} = R_{13}$ та $R_{YZ} = R_{23}$ знайдемо алгебраїчні доповнення:

$$A_{13} = (-1)^{1+3}\begin{vmatrix} 0{,}935 & 1 \\ 0{,}954 & 0{,}991 \end{vmatrix} = 0{,}935 \cdot 0{,}991 - 0{,}954 \approx -0{,}027;$$

$$A_{23} = (-1)^{2+3}\begin{vmatrix} 1 & 0{,}935 \\ 0{,}954 & 0{,}991 \end{vmatrix} = (-1)(0{,}991 - 0{,}935 \cdot 0{,}954) \approx -0{,}099;$$

$$A_{11} = (-1)^{1+1}\begin{vmatrix} 1 & 0{,}991 \\ 0{,}991 & 1 \end{vmatrix} = (1 - 0{,}991^2) \approx 0{,}018;$$

$$A_{22} = (-1)^{2+2}\begin{vmatrix} 1 & 0{,}954 \\ 0{,}954 & 1 \end{vmatrix} = (1 - 0{,}954^2) \approx 0{,}09.$$

$$A = \begin{pmatrix} 1 & 0{,}935 & 0{,}954 \\ 0{,}935 & 1 & 0{,}991 \\ 0{,}954 & 0{,}991 & 1 \end{pmatrix}.$$

## Приклад 3.3

Для обчислення частинних коефіцієнтів кореляції

$$R_{13}=\frac{-A_{13}}{\sqrt{A_{11}A_{33}}}=\frac{-(-0,027)}{\sqrt{0,018\cdot 0,126}}\approx 0,577;\qquad R_{23}=\frac{-A_{23}}{\sqrt{A_{22}A_{33}}}=\frac{-(-0,099)}{\sqrt{0,09\cdot 0,126}}\approx 0,929.$$

Значення частинних коефіцієнтів кореляції показують, що величина $Z$ пов'язана з величиною $Y$ сильніше, ніж з величиною $X$.

Знайдемо t-ста

$$t=\frac{R_{ij}\sqrt{n-k+2}}{\sqrt{1-R_{ij}^{\ 2}}}=\frac{0,577\sqrt{6-3+2}}{\sqrt{1-0,577^2}}\approx 1,581.$$

Знайдемо критичне значення $t_{\text{крит}}$, враховуючи, що $l=n-k+2=6-3+2=5$. Оберемо рівень значущості $\alpha=0,01$. Тоді $t_{\text{крит}}$ = СТЬЮДРАСПОБР (0,01;5) = 4,032. Оскільки розраховане значення $t$-статистики менше критичного $|t|<t_{крит}$, то частинний коефіцієнт кореляції $R_{13}$ не є значущим на рівні значущості $\alpha=0,01$.

## Приклад 3.3

Перевіримо статистичну значущість частинного коефіцієнта кореляції $R_{23}$. Знайдемо $t$-статистику:

$$t = \frac{R_{ij}\sqrt{n-k+2}}{\sqrt{1-R_{ij}^{\ 2}}} = \frac{0{,}929\sqrt{6-3+2}}{\sqrt{1-0{,}929^2}} \approx 5{,}614.$$

Оскільки розраховане значення $t$-статистики більше критичного $|t| > t_{крит}$, то частинний коефіцієнт кореляції $R_{23}$ є значущим на рівні значущості $\alpha = 0{,}01$.

Висновок: отриманий дохід залежить від мотивованості працівників та величини інвестицій. При цьому дохід значно сильніше залежить від величини інвестицій ніж від зарплати працівників. Сила зв'язку між доходом та зарплатою середня і не є статистично значущою.

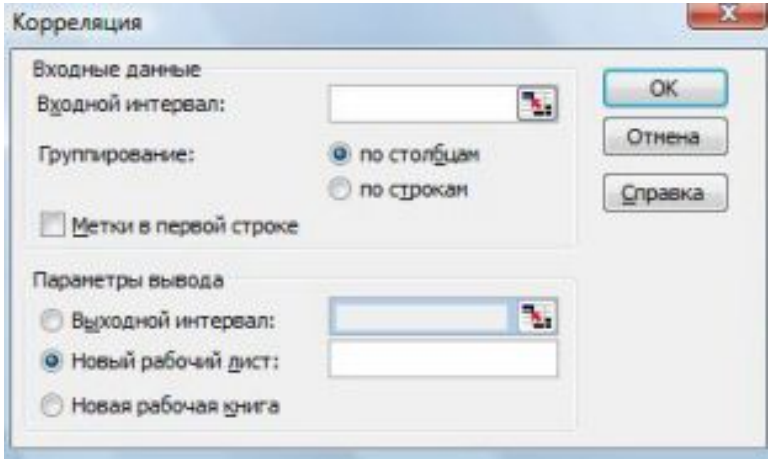
# Кореляційний аналіз із використанням Microsoft Excel

Вбудовані сервісні функції Microsoft Excel дозволяють розраховувати парні коефіцієнти кореляції Пірсона. Для отримання матриці парних коефіцієнтів кореляції необхідно:

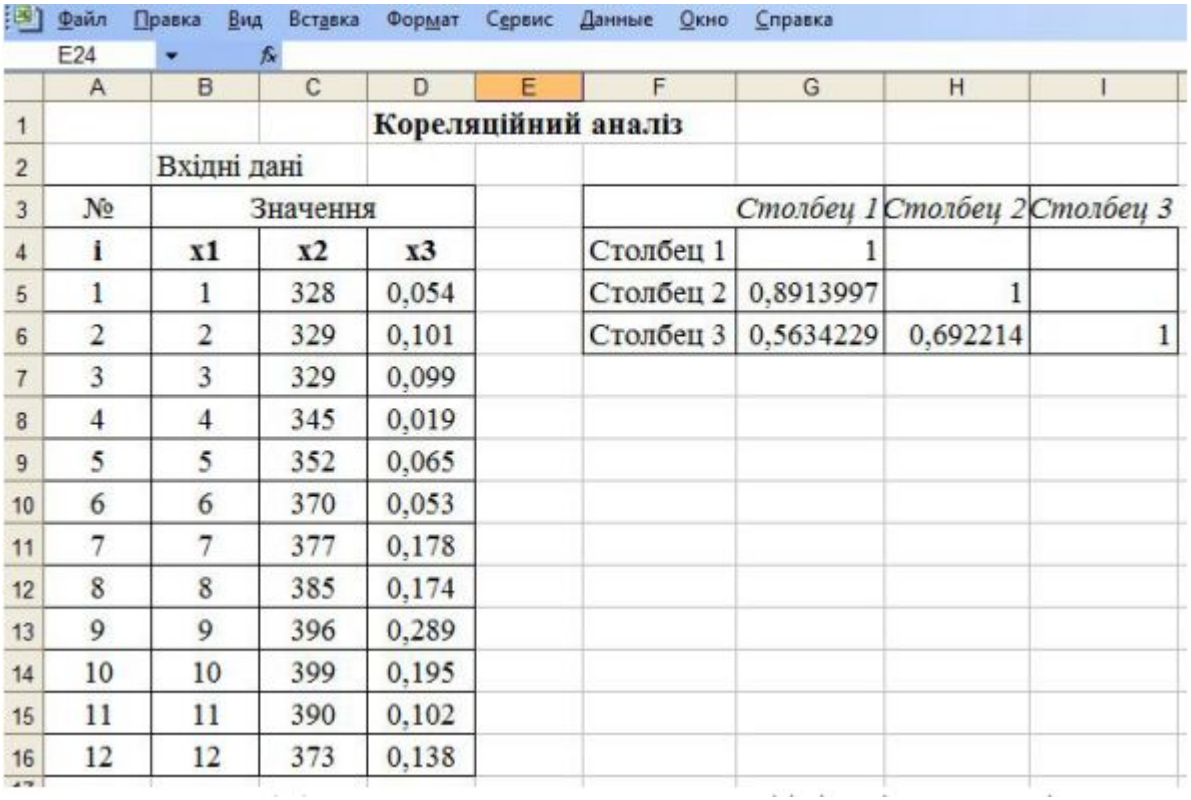
1) Вибрати ***Сервис – Анализ данных.***

2) У діалоговому вікні для вибору інструмента аналізу вибрати інструмент ***Корреляция***. З'явиться вікно для задання параметрів (рис. 3.2).

3) Задати параметри для розрахунку коефіцієнтів кореляції. У графі ***Входной интервал*** вказати масив даних; у графі ***Группирование*** вказати тип групування, наприклад ***По столбцам***, у графі ***Выходной интервал*** вказати ту чарунку, починаючи з якої будуть представлятись вихідні дані – парні коефіцієнти кореляції. Натиснути ***ОК***.



# Кореляційний аналіз із використанням Microsoft Excel



Файл Правка Вид Вставка Формат Сервис Данные Окно Справка

E24

**Кореляційний аналіз**

Вхідні дані

| № | Значення | | |
|---|---|---|---|
| **i** | **x1** | **x2** | **x3** |
| 1 | 1 | 328 | 0,054 |
| 2 | 2 | 329 | 0,101 |
| 3 | 3 | 329 | 0,099 |
| 4 | 4 | 345 | 0,019 |
| 5 | 5 | 352 | 0,065 |
| 6 | 6 | 370 | 0,053 |
| 7 | 7 | 377 | 0,178 |
| 8 | 8 | 385 | 0,174 |
| 9 | 9 | 396 | 0,289 |
| 10 | 10 | 399 | 0,195 |
| 11 | 11 | 390 | 0,102 |
| 12 | 12 | 373 | 0,138 |

| | *Столбец 1* | *Столбец 2* | *Столбец 3* |
|---|---|---|---|
| Столбец 1 | 1 | | |
| Столбец 2 | 0,8913997 | 1 | |
| Столбец 3 | 0,5634229 | 0,692214 | 1 |

Лекція 5

# Побудова регресійних моделей

1. Встановлення виду кореляційної залежності
2. Лінійна регресія
3. Нелінійна регресія
4. Множинна лінійна регресія
5. Регресія у Microsoft Excel

# Регресійний аналіз

При вивченні тісноти зв'язку між різними ознаками економічного чи соціального 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, яка 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

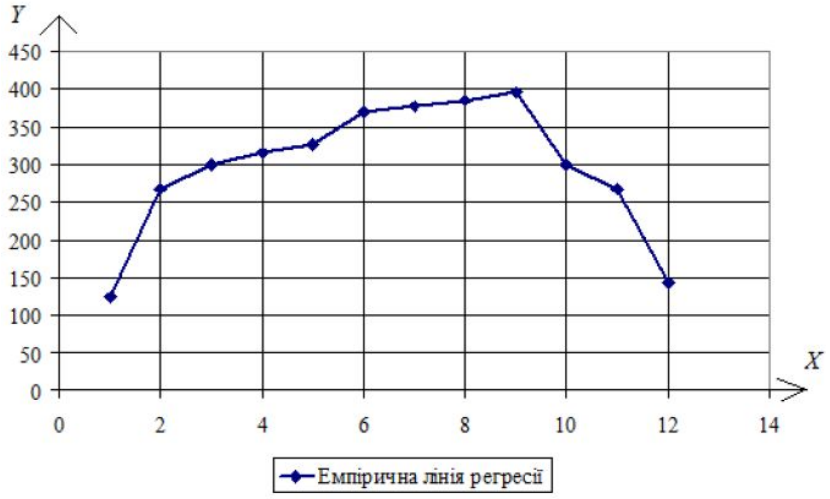
| $x_i$ | $x_1$ | $x_2$ | … | $x_k$ |
|---|---|---|---|---|
| $\overline{y}_{x_i}$ | $\overline{y}_{x_1}$ | $\overline{y}_{x_2}$ | … | $\overline{y}_{x_k}$ |
| $n_i$ | $n_1$ | $n_2$ | … | $n_k$ |

# Лінія регресії

Згруповані дані зображуються графічно, що часто дозволяє визначити вид залежності Y від X.

Ламана лінія, що сполучає точки з координатами ( , ), називається
емпіричною лінією регресії.
Якщо емпірична лінія регресії значно наближається до прямої лінії, то висувається гіпотеза про наявність лінійного зв'язку між досліджуваними ознаками


Гіпотетична лінійна залежність




Гіпотетична нелінійна залежність

# Лінійна регресія

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

$$\overline{y_x} = ax + b$$

де $a, b$ – параметри моделі.

Побудова лінійної регресійної моделі – це знаходження параметрів рівняння . Параметри рівняння регресії можна знайти за методом найменших квадратів.

# Метод найменших квадратів

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

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

Якщо вибіркові дані не згруповані, то система спрощується:

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

Перевірка правильності побудови рівняння регресії здійснюється за основним варіаційним рівнянням:

$$Q = Q_p + Q_o,$$

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: $Q=\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2$; $Q_p=\sum_{i=1}^{n}\left(y_{i,\text{теор}}-\bar{y}\right)^2$; $Q_o=\sum_{i=1}^{n}\left(y_{i,\text{теор}}-y_i\right)^2$; а середнє значення за формулою $\bar{y}=\frac{\sum_{i=1}^{n}y_i}{n}$.

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

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

де $n$ – кількість спостережень, $l$ – 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 $FРАСПОБР\left(\alpha,k_1,k_2\right)$, де $k_1=l-1$; $k_2=n-l$ – степені свободи, $\alpha$ – 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:

$$R^2 = 1 - \frac{Q_o}{Q} = \frac{Q_p}{Q}.$$

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.

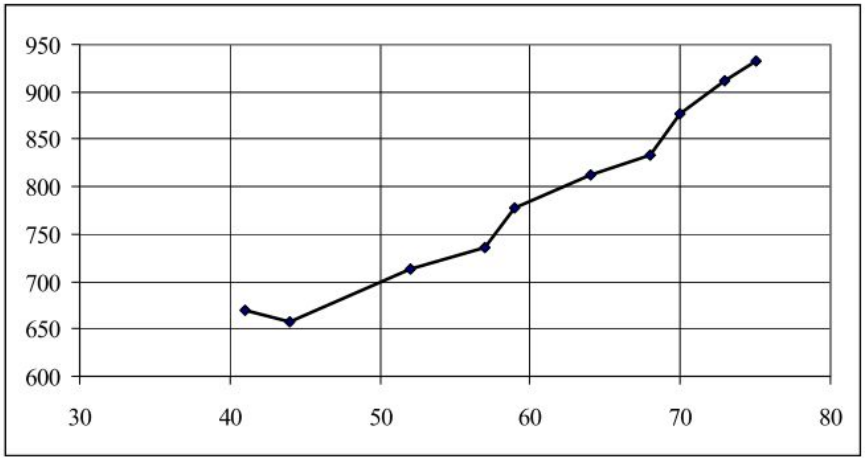
## Приклад 4.1

Побудувати регресійну модель, що описує залежність сумарних виробничих затрат Y (тис. грн.) від об'ємів виробництва X (тис. од.). Відповідні статистичні дані задано у табл.

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

задано вибіркові дані: значення $x_i$, $i=\overline{1,n}$ 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




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 $y = ax + b$.

## Приклад 4.1

Другий етап: знайдемо параметри $a,b$ рівняння регресії, для чого складемо систему для не згрупованих даних. Необхідні розрахунки для зручності оформимо у вигляді таблиці

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

| 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:

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

## Приклад 4.1

$$\begin{cases} 37625a + 603b = 487643 \\ 603a + 10b = 7918 \end{cases}.$$

Знайдемо визначник основної матриці системи, яка складена 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 попереднього 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 $y = 8{,}06x - 305{,}83$.

## Приклад 4.1

Третій етап: перевіримо правильність побудови моделі $Q = Q_p + Q_o$,

її статистичну значущість за $F$-статистикою $F = \dfrac{Q_p (n-l)}{Q_o (l-1)}$,

адекватність вибірковим даним за коефіцієнтом детермінації $R^2 = 1 - \dfrac{Q_o}{Q} = \dfrac{Q_p}{Q}$.

## Приклад 4.1

знайдемо загальну варіацію,

варіації регресії та залишків; необхідні розрахунки оформимо у вигляді таблиці

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

## Приклад 4.1

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 значущим на рівні значущості 0,001 та описує 96% вибіркових даних.

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

**Рівняння параболічної регресії:**

$$\overline{y_x} = ax^2 + bx + c\,.$$

Система для знаходження параметрів:

для згрупованих вибіркових даних:

$$\begin{cases} a\sum_{i=1}^{k} x_i^4 n_i + b\sum_{i=1}^{k} x_i^3 n_i + c\sum_{i=1}^{k} x_i^2 n_i = \sum_{i=1}^{k} x_i^2 n_i \overline{y_{x_i}} \\ a\sum_{i=1}^{k} x_i^3 n_i + b\sum_{i=1}^{k} x_i^2 n_i + c\sum_{i=1}^{k} x_i n_i = \sum_{i=1}^{k} x_i n_i \overline{y_{x_i}} \\ a\sum_{i=1}^{k} x_i^2 n_i + b\sum_{i=1}^{k} x_i n_i + c\sum_{i=1}^{k} n_i = \sum_{i=1}^{k} n_i \overline{y_{x_i}} \end{cases}$$

для незгрупованих вибіркових даних:

$$\begin{cases} a\sum_{i=1}^{n} x_i^4 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i^2 y_i \\ a\sum_{i=1}^{n} x_i^3 + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \\ a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i + cn = \sum_{i=1}^{n} y_i \end{cases}$$

**Рівняння гіперболічної регресії:**

$$\overline{y_x} = \frac{a}{x} + b.$$

Система для знаходження параметрів:

для згрупованих вибіркових даних:

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

для незгрупованих вибіркових даних:

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

**Рівняння показникової регресії:**

$$\overline{y_x} = ba^x.$$

| Система для знаходження параметрів: | |
|---|---|
| для згрупованих вибіркових даних: | для незгрупованих вибіркових даних: |
| $\begin{cases} \lg a\sum_{i=1}^{k} x_i^2 n_i + \lg b\sum_{i=1}^{k} x_i n_i = \sum_{i=1}^{k} x_i n_i \lg \overline{y_{x_i}} \\ \lg a\sum_{i=1}^{k} x_i n_i + \lg b\sum_{i=1}^{k} n_i = \sum_{i=1}^{k} n_i \lg \overline{y_{x_i}} \end{cases}$ | $\begin{cases} \lg a\sum_{i=1}^{n} x_i^2 + \lg b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i \lg y_i \\ \lg a\sum_{i=1}^{n} x_i + n\lg b = \sum_{i=1}^{n} \lg y_i \end{cases}$ |

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 4.2

Дано розподіл однотипних підприємств за об'ємом виробництва X (тис. од.) і собівартістю одиниці продукції Y (грн.) (табл. ). Знайти регресійну модель, що описує залежність собівартості продукції від об'єму виробництва.

| X \ Y | 10 | 15 | 20 | 25 |
|---|---|---|---|---|
| 25 | – | - | 1 | 2 |
| 50 | – | 2 | 2 | – |
| 75 | – | 5 | 3 | 1 |
| 100 | 1 | 3 | – | – |
| 125 | 3 | 1 | 1 | – |

## Приклад 4.2

Розв'язок. Для проведення регресійного аналізу за даними табл. побудуємо кореляційну таблицю

| $y_j$ \ $x_i$ | 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 $X$ для чого знайдемо середні значення $\overline{y}_{x_i}$ для кожного значення $x_i$, $i=\overline{1,5}$

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

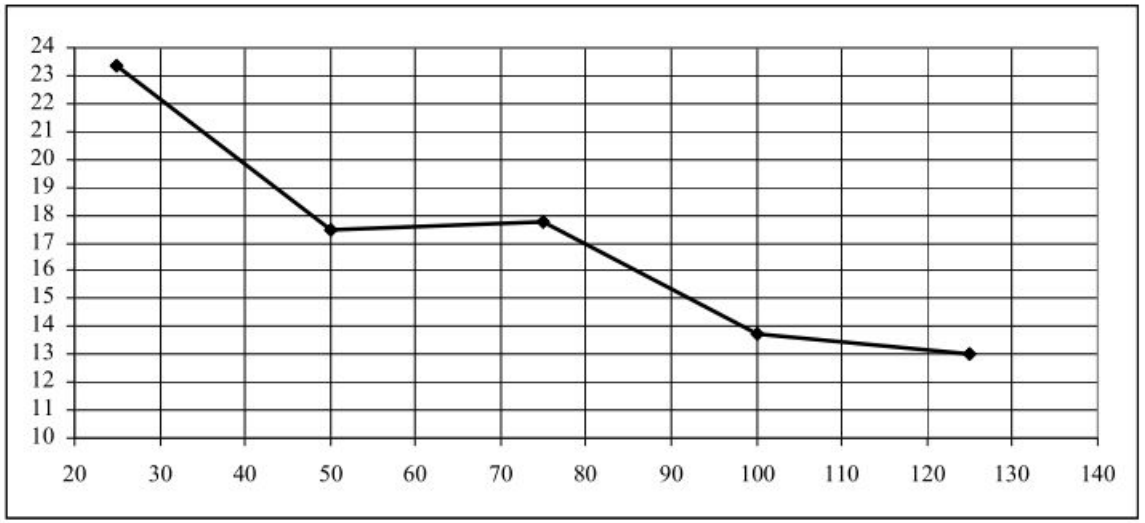
$$\overline{y}_{x_3}=\frac{15\cdot5+20\cdot3+25\cdot1}{9}=17{,}78;\ \overline{y}_{x_4}=\frac{10\cdot1+15\cdot3}{4}=13{,}75;$$

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

## Приклад 4.2

| $x_i$ | 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




Емпірична лінія регресії

## Приклад 4.2

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 , в останньому рядку якої знайдемо відповідні стовпцям суми.

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

для згрупованих вибіркових даних:

$$\begin{cases} a\sum_{i=1}^{k} \frac{1}{x_i^2} n_i + b\sum_{i=1}^{k} \frac{1}{x_i} n_i = \sum_{i=1}^{k} \frac{1}{x_i} \overline{y}_{x_i} n_i \\ a\sum_{i=1}^{k} \frac{1}{x_i} n_i + b\sum_{i=1}^{k} n_i = \sum_{i=1}^{k} \overline{y}_{x_i} n_i \end{cases}$$

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

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

## Приклад 4.2

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

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

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

## Приклад 4.2

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:

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

Знайдемо $F_{кр}$: $F_{кр} = FРАСПОБР(0{,}001; 2-1; 25-2) \approx 14{,}20$. 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 значущим на рівні значущості 0,001 та описує 81% вибіркових даних.

# Множинна лінійна регресія

У процесі аналізу діяльності економічного або соціального об'єкта часто виявляється, що на результативну ознаку цієї діяльності (наприклад, об'єм валової продукції, об'єм продаж, думку респондента відносно певного об'єкта та ін.) впливає декілька факторних ознак: час, вартість сировини і матеріалів, якість обладнання, продуктивність праці, соціальні установки, вплив зовнішніх і внутрішніх факторів та інше. Тоді як модель діяльності об'єкта використовують багатофакторну лінійну регресійну модель, на основі якої розробляються прогнози діяльності, вивчається вплив на діяльність різноманітних показників і виявляються ті показники, покращення яких суттєво збільшує її кінцевий продукт.

Загальний вигляд багатофакторної лінійної регресійної моделі:

$$Y = f(X_1, X_2, \dots, X_m) + \varepsilon,$$

де $Y$ – результативна ознака,
$X_1, X_2, \dots, X_m$ – факторні ознаки,
$m$ – кількість факторних ознак,
$\varepsilon$ – випадкова похибка моделі.

**Зауваження 1.** Задачі побудови багатофакторної регресійної моделі розв'язуються за умов, коли випадкова похибка $\varepsilon$ має нормальний розподіл із нульовим математичним сподіванням, а випадкові похибки кожного вимірювання незалежні та мають однакові дисперсії. Кількість спостережень $n$ повинна перевищувати величину $3(m+1)$.

Крім того, для забезпечення статистичної значущості моделі необхідно дотримуватися основного правила її побудови: **„Факторні ознаки, які включено у модель, повинні бути тісно пов'язані із результативною ознакою і слабо пов'язані (або не мати зв'язку) між собою"**.

Тіснота зв'язку між результативною і факторними ознаками та зв'язку факторних ознак між собою визначається за аналізом парних і частинних коефіцієнтів кореляції (див. п. 3.5). В модель бажано включати тільки ті ознаки, що не мають статистично значущого зв'язку між собою, хоча й вважається, що сильний зв'язок між ними, зазвичай, не впливає на якість прогнозу за моделлю.

Якість моделі визначається за критерієм Фішера, тобто порівнянням статистики $F$ моделі із критичним значенням $F_{\text{кр}}$, де $F_{\text{кр}}(\alpha, k_1, k_2)$ – табличне значення розподілу Фішера, що знаходиться за умов: $\alpha = 0{,}05$; $k_1 = m-1$; $k_2 = n-m$. Якщо $F > F_{кр}$, то модель є достовірною на рівні значущості 0,05 (тобто 95% даних пояснюються побудованою моделлю, 5% – випадкові помилки моделі).

Відносну величину впливу факторних ознак на результативну можна оцінити за формулою: $r_{X_i}^{\ 2} = \frac{t_{X_i}^{\ 2} \cdot R^2}{\sum_{i=1}^{m} t_{X_i}^{\ 2}}$;

де $t_{X_i}^{\ 2}$ – розраховане значення розподілу Стьюдента для ознаки $X_i$;

$R^2$ – загальний коефіцієнт детермінації моделі.

Обчислення, які необхідно провести для побудови багатофакторної регресійної моделі, дуже складні, але застосування засобу Excel ***Регрессия*** пакета ***Анализ данных*** значно полегшує цю роботу.

За допомогою засобу ***Регрессия*** отримують такі результати:

– параметри лінійної регресійної моделі виду $Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_mX_m$, де $b_0$, $b_1$, $b_2$, …, $b_m$ – параметри моделі;

– коефіцієнт детермінації;

– критеріальну статистику для перевірки статистичної значущості моделі;

– теоретичні значення результативної ознаки, отримані за побудованою моделлю та залишки – тобто різниці між теоретичними та емпіричними значеннями цієї ознаки.

**Зауваження 2.** Засобом ***Регрессия*** можна також користуватися при побудові моделі, нелінійної відносно певної факторної ознаки $X_j$, але лінійної відносно коефіцієнта цієї ознаки. Наприклад, якщо необхідно побудувати модель виду $Y = b_0 + b_1X_1 + b_2X_2^3$, то як вхідні дані вказують трійки $(y_i, x_{1i}, x_{2i}^3)$.

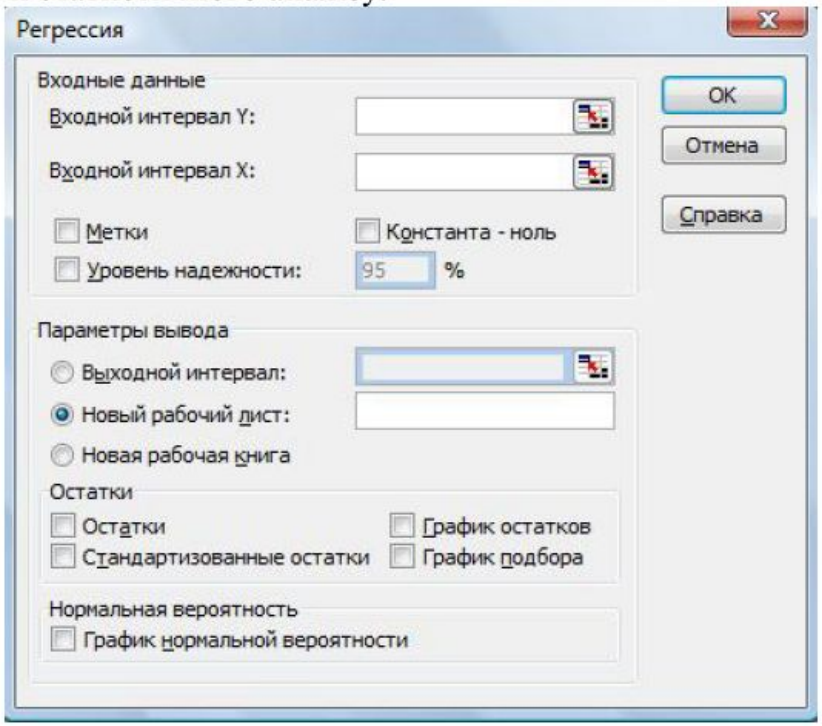
# Регресія у Microsoft Excel

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.

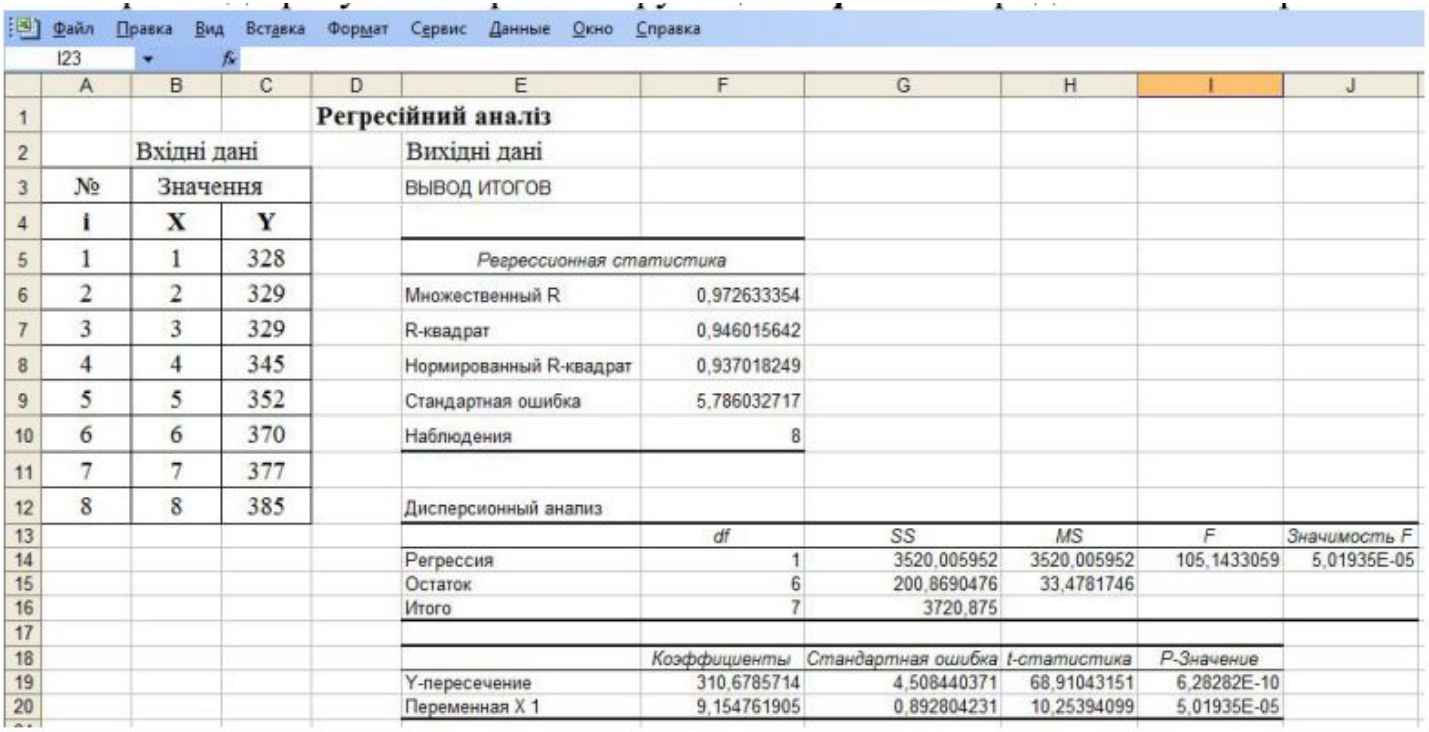
Для побудови лінійної регресійної моделі необхідно:

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 чарунку, починаючи з якої будуть надаватися вихідні дані – параметри рівняння регресії та результати її статистичного аналізу.



# Регресія у Microsoft Excel



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 | |

Результати регресійного аналізу

В таблиці 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. Це значення обчислюється як відношення варіації регресії до варіації залишків (чарунки H14 та H15). В стовпці ***Значимость F*** надано критеріальну статистику. Якщо це значення менше, ніж, наприклад, 0,05, то рівняння регресії є значущим на рівні 0,05. У даному завданні рівняння регресії є значущим на рівні 0,00005.

У графі ***Множественный R-квадрат*** надано значення множинного коефіцієнта кореляції, який показує силу залежності результативної ознаки від факторної (або декілька факторних ознак). У нашому випадку він дорівнює 0,97, що означає сильний зв'язок між $Y$ та $X$.

Коефіцієнт детермінації моделі $R^2$ виводиться у графі ***R-квадрат***, $R^2$= 0,97, тобто 97% даних описується рівнянням регресії.

# Регресія у Microsoft Excel

**Регресійний аналіз**

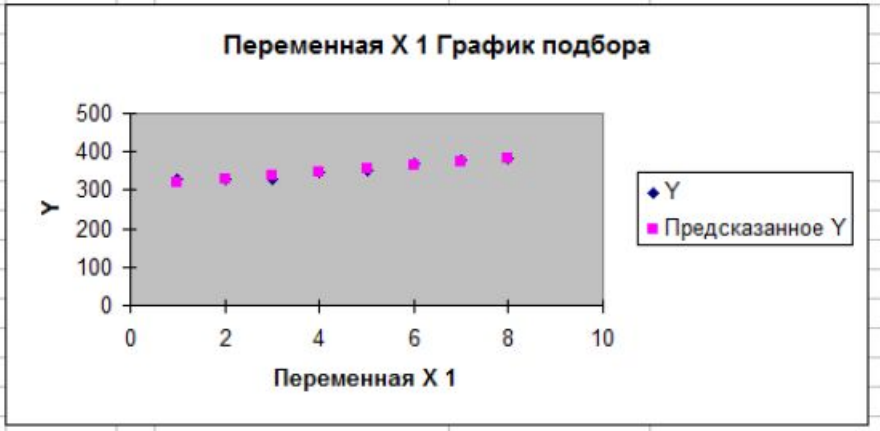
| № | Значення | |
|---|---|---|
| i | X | Y |
| 1 | 1 | 328 |
| 2 | 2 | 329 |
| 3 | 3 | 329 |
| 4 | 4 | 345 |
| 5 | 5 | 352 |
| 6 | 6 | 370 |
| 7 | 7 | 377 |
| 8 | 8 | 385 |

Вхідні дані

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 |




Графік підбору – порівняльна діаграма, що містить емпіричну і теоретичну лінії регресії; таблиця залишків – різниць емпіричних і теоретичних значень Y

## Приклад 4.3

В таблиці вказано дані по заводу за 12 місяців року

| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $Y$ |
|---|---|---|---|---|
| 1 | 328 | 0,054 | 0,3 | 397 |
| 2 | 329 | 0,101 | 0,6 | 670 |
| 3 | 329 | 0,099 | 1,2 | 1209 |
| 4 | 347 | 0,019 | 0,1 | 138 |
| 5 | 352 | 0,065 | 0,3 | 378 |
| 6 | 370 | 0,053 | 0,1 | 79 |
| 7 | 378 | 0,178 | 2,3 | 1883 |
| 8 | 385 | 0,174 | 2,6 | 2124 |
| 9 | 396 | 0,298 | 5,5 | 5069 |
| 10 | 399 | 0,195 | 2,4 | 2618 |
| 11 | 390 | 0,102 | 1,6 | 1265 |
| 12 | 378 | 0,138 | 0,6 | 562 |

Умовні позначення:

$X_1$ – часовий фактор, порядковий номер місяця;

$X_2$ – фонди (тис. грн./робітника);

$X_3$ – фондовіддача (тис. грн. обсягу товарного продукту/тис. грн. основного фонду);

$X_4$ – продуктивність праці (тис. умовних банок/робітника);

$Y$ – валова продукція (тис. умовних банок).

Розробляється проект модернізації заводу, для чого необхідно: побудувати багатофакторну лінійну регресійну модель діяльності заводу; визначити вплив факторних ознак на об'єм валової продукції; виявити найвпливовіші ознаки для визначення напрямків майбутньої модернізації.

## Приклад 4.3

**Розв'язок.** За основним правилом побудови множинної регресійної моделі розв'язок задачі складається з трьох етапів: виявлення факторних ознак, які необхідно включити в модель; побудова моделі; аналіз якості моделі.

*Етап 1.* Виявимо факторні ознаки, що включаються в модель. Для чого:

– розрахуємо парні коефіцієнти кореляції; побудуємо кореляційну матрицю і проведемо її статистичний аналіз;

– на основі результатів статистичного аналізу побудуємо кореляційні плеяди і виявимо ознаки, які необхідно включити в модель;

– у разі необхідності (тобто у випадку існування неявного зв'язку між факторними ознаками) розрахуємо частинні коефіцієнти кореляції та проведемо їх статистичний аналіз.

Парні коефіцієнти кореляції обчислимо за допомогою вбудованих сервісних функцій Excel: перенесемо табл. на сторінку Excel, викличемо ***Сервис – Аналіз данных – Корреляция – ОК***. У графі ***Входной интервал*** вкажемо масив даних табл. у графі ***Группирование*** вкажемо ***По столбцам***, у графі ***Выходной интервал*** вкажемо ту чарунку, починаючи з якої будуть виводитися вихідні дані – парні коефіцієнти кореляції. Отримаємо

## Приклад 4.3

Отримаємо табл., яка є матрицею парних коефіцієнтів кореляції. Чарунки таблиці, розташовані вище головної діагоналі, незаповнені, оскільки таблиця симетрична відносно головної діагоналі.

| | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $Y$ |
|---|---|---|---|---|---|
| $X_1$ | 1,00 | | | | |
| $X_2$ | 0,89 | 1,00 | | | |
| $X_3$ | 0,56 | 0,69 | 1,00 | | |
| $X_4$ | 0,46 | 0,66 | 0,94 | 1,00 | |
| $Y$ | 0,43 | 0,63 | 0,94 | 0,99 | 1,00 |

Розрахуємо критичне значення коефіцієнта кореляції $r_{кр}$ за формулою:

$r_{кр} = \frac{t_{\alpha,k}}{\sqrt{t_{\alpha,k}^{\ 2} + n - 2}}$, $\alpha$ – рівень значущості, $\alpha = 0{,}05$ ; $t_{\alpha,k}$ знайдемо за допомогою вбудованої функції Excel. Викличемо ***Функции*** – ***Статистические*** – ***СТЬЮДРАСПОБР*** – ***Ок***. В графі ***Вероятность*** вкажемо 0,05 (рівень значущості); в графі ***Степени свободы*** вкажемо значення $n - 2 = 12 - 2 = 10$. Отримаємо $t_{\alpha\ k} = 2{,}228$. Тоді:

$$r_{кр} = \frac{t_{\alpha,k}}{\sqrt{t_{\alpha,k}^{\ 2} + n - 2}} = \frac{2{,}228}{\sqrt{2{,}228^2 + 12 - 2}} \approx 0{,}57598.$$

## Приклад 4.3

Виділимо в таблиці елементи, які більші за $r_{кр}$ (це означає, що відповідні ознаки тісно пов'язані між собою).

| | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $Y$ |
|---|---|---|---|---|---|
| $X_1$ | 1,00 | **0,89** | 0,56 | 0,46 | 0,43 |
| $X_2$ | **0,89** | 1,00 | **0,69** | **0,66** | **0,63** |
| $X_3$ | 0,56 | **0,69** | 1,00 | **0,94** | **0,94** |
| $X_4$ | 0,46 | **0,66** | **0,94** | 1,00 | **0,99** |
| $Y$ | 0,43 | **0,63** | **0,94** | **0,99** | 1,00 |

Отже, тісно пов'язані між собою такі факторні ознаки:

$X_1$ та $X_2$ оскільки $r(X_1, X_2) = 0,89 > 0,57598$;

$X_2$ та $X_3$ оскільки $r(X_2, X3) = 0,69 > 0,57598$;

$X_2$ та $X_4$ оскільки $r(X_2, X_4) = 0,66 > 0,57598$;

$X_3$ та $X_4$ оскільки $r(X_3, X_4) = 0,94 > 0,57598$.

З факторних ознак тісно пов'язані із результативною ознакою (із $Y$):

$X_2$ оскільки $r(X_2, Y) = 0,63 > 0,57598$;

$X_3$ оскільки $r(X_3, Y) = 0,94 > 0,57598$;

$X_4$ оскільки $r(X_4, Y) = 0,99 > 0,57598$.

Умовні позначення:

$X_1$ – часовий фактор, порядковий номер місяця;

$X_2$ – фонди (тис. грн./робітника);

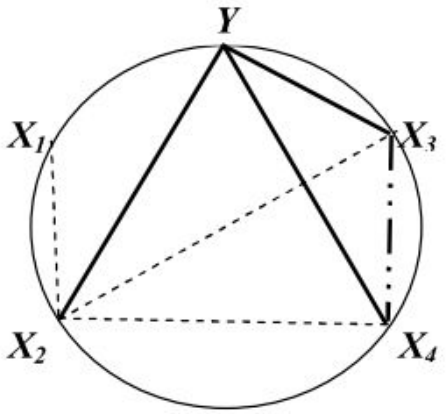
$X_3$ – фондовіддача (тис. грн. обсягу товарного продукту/тис. грн. основного фонду);

$X_4$ – продуктивність праці (тис. умовних банок/робітника);

$Y$ – валова продукція (тис. умовних банок).

## Приклад 4.3

За результатами аналізу кореляційної матриці побудуємо кореляційні плеяди, тобто зобразимо достовірний зв'язок між факторними ознаками графічно




Кореляційний зв'язок між факторними ознаками

Перша кореляційна плеяда: $Y$, $X_2$ , $X_3$ , $X_4$ вказує, що в модель необхідно включити ознаки $X_2$ , $X_3$ , $X_4$, оскільки вони мають зв'язок (тобто впливають) на результативну ознаку $Y$. Тобто із всіх факторних ознак в модель не потрібно включати ознаку $X_1$.

Друга кореляційна плеяда: $X_2$, $X_1$ , $X_3$ , $X_4$ вказує, що в модель можна включити тільки одну з ознак $X_2$, $X_1$ , $X_3$ , $X_4$, оскільки вони пов'язані між собою. Однак наявність дуже сильного (0,94) зв'язку між $X_3$ та $X_4$ свідчить про те, що зв'язок може існувати між $X_2$ та $X_3$ , а між $X_2$ та $X_4$ він може бути тільки наслідком зв'язку $X_3 - X_4$. Або навпаки, зв'язок може існувати між $X_2$ та $X_4$ , а між $X_2$ та $X_3$ він може бути тільки наслідком зв'язку $X_3 - X_4$. Тому, можливо, в модель потрібно включати $X_2$ та одну із ознак $X_3$ та $X_4$. Для того, щоб виявнити це, скористуємось частинними коефіцієнтами кореляції.

## Приклад 4.3

Розрахуємо частинні коефіцієнти кореляції між факторними ознаками $X_2$ і $X_3$, та між $X_2$ і $X_4$. Величина цих частинних коефіцієнтів кореляції дозволить визначити „чистий" зв'язок між вказаними ознаками, тобто зв'язок, що не залежить від впливу всіх останніх факторних ознак.

Частинний коефіцієнт кореляції між факторними ознаками $X_2$ і $X_3$ розраховуємо за формулою:

$$R_{23} = \frac{-A_{23}}{\sqrt{A_{22} A_{33}}},$$

де $A_{23}$ – алгебраїчне доповнення елемента $r_{23}$,
$A_{22}$ – алгебраїчне доповнення елемента $r_{22}$,
$A_{33}$ – алгебраїчне доповнення елемента $r_{33}$.

Алгебраїчне доповнення $A_{23}$ – це визначник матриці, отриманої із матриці $A$ викреслюванням 2-го рядка і 3-го стовпця, помножений на $(-1)^{2+3}$. Аналогічно $A_{22}$ – це визначник матриці, отриманої із матриці $A$ викреслюванням 2-го рядка і 2-го стовпця, помножений на $(-1)^{2+2}$; $A_{33}$ – це визначник матриці, отриманої із матриці $A$ викреслюванням 3-го рядка і 3-го стовпця, помножений на $(-1)^{3+3}$. Визначники обчислюємо за допомогою ***МОПРЕД,*** у графі ***Массив*** вказуємо матрицю, визначник якої потрібно знайти.

Отримаємо:

$$A_{23}=\begin{vmatrix} 1 & 0,89 & 0,46 & 0,43 \\ 0,56 & 0,69 & 0,94 & 0,94 \\ 0,46 & 0,66 & 1 & 0,99 \\ 0,43 & 0,63 & 0,99 & 1 \end{vmatrix}=0,00028; \quad A_{22}=\begin{vmatrix} 1 & 0,56 & 0,46 & 0,43 \\ 0,56 & 1 & 0,94 & 0,94 \\ 0,46 & 0,94 & 1 & 0,99 \\ 0,43 & 0,94 & 0,99 & 1 \end{vmatrix}=0,001005;$$

$$A_{33}=\begin{vmatrix} 1 & 0,89 & 0,46 & 0,43 \\ 0,89 & 1 & 0,66 & 0,63 \\ 0,46 & 0,66 & 1 & 0,99 \\ 0,43 & 0,63 & 0,99 & 1 \end{vmatrix}=0,001442; \quad R_{23}=\frac{-A_{23}}{\sqrt{A_{22}A_{33}}}=\frac{-0,0028}{\sqrt{0,001005\cdot 0,001442}}\approx -0,24;$$

$$A_{24}=\begin{vmatrix} 1 & 0,89 & 0,56 & 0,43 \\ 0,56 & 0,69 & 1 & 0,94 \\ 0,46 & 0,66 & 0,94 & 0,99 \\ 0,43 & 0,63 & 0,94 & 1 \end{vmatrix}=-0,00041; \quad A_{44}=\begin{vmatrix} 1 & 0,89 & 0,56 & 0,43 \\ 0,89 & 1 & 0,69 & 0,63 \\ 0,56 & 0,69 & 1 & 0,94 \\ 0,43 & 0,63 & 0,94 & 1 \end{vmatrix}=0,008578;$$

$$R_{24}=\frac{-A_{24}}{\sqrt{A_{22}A_{44}}}=\frac{0,00041}{\sqrt{0,001005\cdot 0,008578}}\approx 0,1411.$$

Оскільки $|R_{23}| > R_{24}$, то в модель необхідно включати факторну ознаку $X_4$.

**Висновок з етапу 1:** шукана багатофакторна лінійна регресійна модель має вигляд: $Y = b_0 + b_1 X_2 + b_2 X_4$.

*Етап 2.* Побудуємо вказану модель. Для знаходження $b_0, b_1, b_2$ викликаємо ***Сервис – Анализ данных – Регрессия – ОК***. У графі ***Входной интервал Y*** вкажемо відповідний стовпчик даних табл. 4.11; у графі ***Входной интервал X*** вкажемо стовпчики $X_2$ та $X_4$ табл 4.11; у графі ***Выходной интервал*** вкажемо ту чарунку, починаючи з якої будуть виводитися вихідні дані – рівняння регресії. Отримаємо таблицю з результатами регресійного аналізу

ВЫВОД ИТОГОВ

| *Регрессионная статистика* | |
|---|---|
| Множественный | 0,992676928 |
| R-квадрат | 0,985407483 |
| Нормированный | 0,982164702 |
| Стандартная ош | 191,3107107 |
| Наблюдения | 12 |

Дисперсионный анализ

| | *df* | *SS* | *MS* | *F* | *Значимость F* |
|---|---|---|---|---|---|
| Регрессия | 2 | 22243684,82 | 11121842,41 | 303,8772358 | 5,47754E-09 |
| Остаток | 9 | 329398,0921 | 36599,78801 | | |
| Итого | 11 | 22573082,92 | | | |

| | *Коэффициенты* | *Стандартная ошибка* | *t-статистика* | *P-Значение* | *Нижние 95%* | *Верхние 95%* | *Нижние 95,0%* | *Верхние 95,0%* |
|---|---|---|---|---|---|---|---|---|
| Y-пересечение | 657,6665422 | 998,224854 | 0,658836072 | 0,526495682 | -1600,474958 | 2915,808042 | -1600,474958 | 2915,808042 |
| Переменная X 1 | -1,760422463 | 2,859983 | -0,615535881 | 0,553445728 | -8,230154606 | 4,709309679 | -8,230154606 | 4,709309679 |
| Переменная X 2 | 927,0473673 | 48,915520 | 18,9520088 | 1,4588E-08 | 816,3927733 | 1037,701961 | 816,3927733 | 1037,701961 |

В таблиці у графі ***Коэффициенты*** вказані значення параметрів моделі $b_0, b_1, b_2$: $b_0$ – в графі ***Y-пересечение***, $b_1$ – в графі **Переменная *X1***, $b_2$ – в графі **Переменная *X2***. Отже, $b_0$=657,67; $b_1$= −1,76; $b_2$=927,05; багатофакторна лінійна регресійна модель має вигляд:

$$Y = 657,67 - 1,76X_2 + 927,05X_4.$$

Результати регресійного

*Етап 3.* Перевіримо якість побудованої моделі. Скористуємось результатами регресійного аналізу (рис. 4.9).

Перевіримо статистичну значущість моделі. Значення $F$-статистики моделі подано в таблиці у графі $F$: $F$ = 303,877. Критичне значення $F_{кр}$ знайдемо за допомогою статистичної функції Excel $FРАСПОБР(\alpha, k_1, k_2)$, де

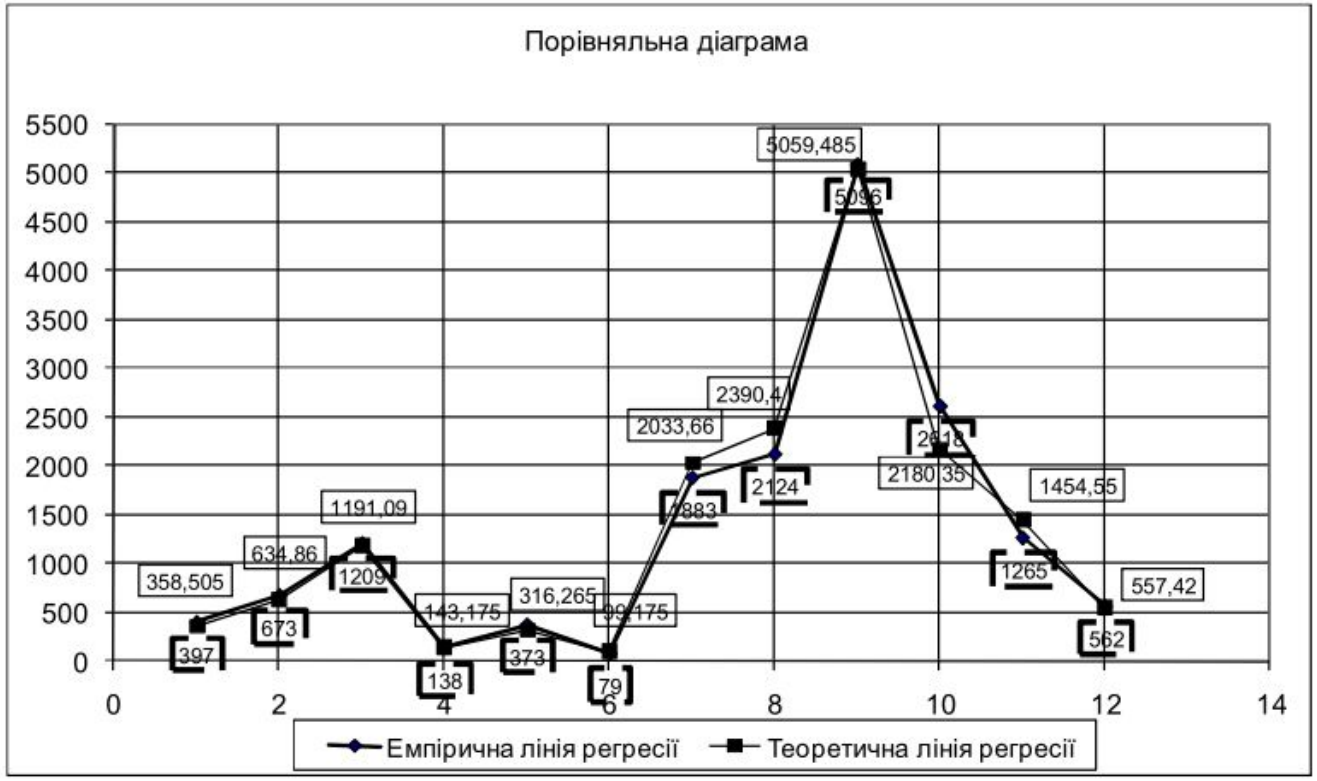
$\alpha = 0{,}05; \quad k_1 = m - 1 = 2 - 1 = 1; \quad k_2 = n - m = 12 - 2 = 10.$

Отже, $F_{кр}$=4,96; $F > F_{кр} \Rightarrow$ рівняння регресії є значущим, модель є достовірною на рівні значущості 0,05. Крім того, в стовпчику ***Значимость F*** є критеріальна статистика, яка показує, що рівняння регресії є значущим на рівні 0,000000005.

У графі ***Множественный R-квадрат*** подано значення множинного коефіцієнта кореляції – 0,99, який показує сильну залежність результативної ознаки від обраних факторних ознак. У графі ***R-квадрат*** бачимо коефіцієнт детермінації моделі $R^2$=0,985, тобто 98,5% даних описуються рівнянням регресії.

Для наочності висновків зобразимо емпіричну і теоретичну лінії регресії

## Приклад 4.3



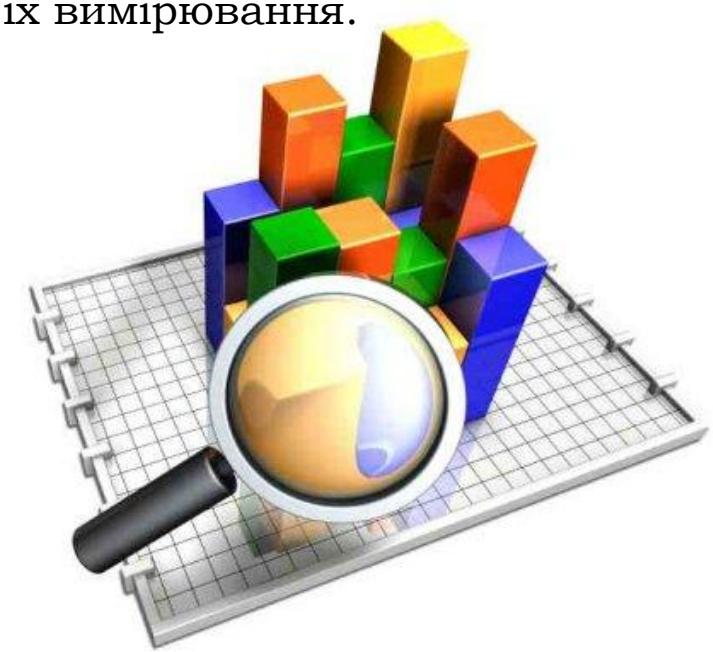

Порівняльна діаграма за результатами регресійного аналізу

**Висновок:** на об'єм валової продукції $Y$ значно впливають факторні ознаки $X_2$ – фонди (тис. грн./робітника) та $X_4$ – продуктивність праці (тис. умовних банок/робітника), що й визначає напрями модернізації заводу.

Лекція 6

# Ряди динаміки. аналіз інтенсивності та тенденцій розвитку

1. Суть та складові елементи ряду динаміки. Види динамічних рядів.
2. Основні характеристики рядів динаміки.
3. Середні показники динаміки.
4. Виявлення тенденцій розвитку явищ.
5. Характеристика сезонних коливань, методи їх вимірювання.



## Суть та складові елементи ряду динаміки

В статистичній практиці доводиться мати справу з великою кількістю даних, що характеризують розвиток явищ в часі.
Для кращого розуміння і аналізу досліджуваних статистичних даних, їх потрібно систематизувати, побудувавши хронологічні ряди, які називаються **рядами динаміки.**
**Ряди динаміки** в статистиці - ряди чисел, що характеризують закономірності і особливості зміни економічних чи суспільних явищ і процесів в часі.

Кожний ряд динаміки складається з двох елементів:
1) періодів або моментів часу, до яких відносяться рівні ряду (t);
2) статистичних показників, які характеризують рівні ряду (y).

При формуванні динамічних рядів для дослідження розвитку економычних чи суспільних явищ в часі потрібно дотримуватись вимоги порівняльності всіх рівнів ряду між собою. Показники ряду динаміки повинні бути порівняльні за територією, колом охоплюваних об'єктів, способами розрахунків, періодами часу, одиницями виміру.

## Види динамічних рядів

В залежності від характеру рівнів ряду розрізняють види рядів динаміки: моментні і інтервальні (періодичні).

**Моментним** називається ряд динаміки, величини якого характеризують стан явищ на певний момент часу.

**Інтервальним** називається такий ряд динаміки, величини якого чисельно характеризують економіко-суспільні явища за певні періоди часу (день, місяць, квартал і т.д.).

Сума рівнів інтервального ряду динаміки характеризує рівень даних явища за більш тривалий проміжок часу.

*Ряди динаміки одномірні і багатомірні.*

**Одномірні** ряди динаміки характеризують зміну одного показника (валовий продукт).

**Багатомірні** ряди динаміки характеризують зміну двох, трьох і більше показників.

Багатомірні динамічні ряди поділяються на паралельні ряди і ряди взаємозв'язаних показників.

**Паралельні** ряди динаміки відображають зміну або одного і того самого показника щодо різних об'єктів, або різних показників щодо одного і того самого об'єкта.

**Ряди взаємозв'язаних показників** характеризують залежність одного явища від іншого (залежність заробітної плати робітників від їхнього тарифного розряду).

*За повнотою часу динамічні ряди поділяються на повні і неповні.*

В **повних** динамічних рядах дати або періоди ідуть один за одним з рівними інтервалами.

В **неповних** динамічних рядах в послідовності часу спостерігаються нерівні інтервали.

За способом вираження рівнів динамічного ряду вони поділяються на ряди

- **абсолютних, середніх і відносних** величин.

## Вимоги до формування динамічних рядів

Статистичні дані, які необхідні для побудови ряду динаміки повинні бути порівняльні за **колом охоплюваних об'єктів**. Непорівняльність може виникнути внаслідок переходу деяких об'єктів із одного підпорядкування в інше.

Порівняльність за колом охоплюваних об'єктів забезпечується **зімкненням динамічних рядів** шляхом заміни абсолютних рівнів відносними.

В моментних рядах динаміки виникає непорівняльність **за критичним моментом реєстрації** рівнів явищ, які піддаються сезонним коливанням.

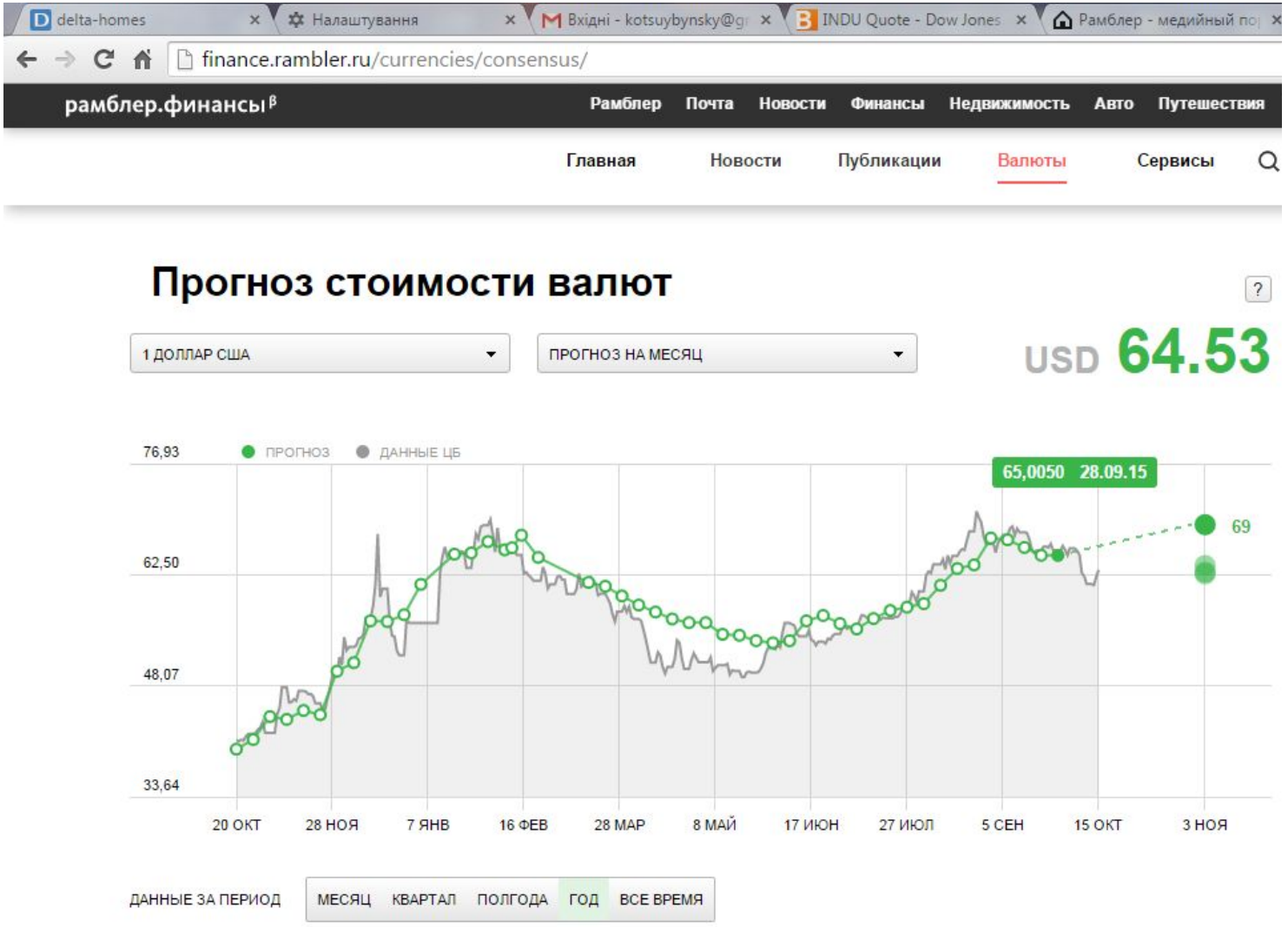
Рівні динамічного ряду повинні бути порівняльні за **методикою їх розрахунку** (Наприклад, за попередні роки чисельність робітників заводу була визначена на початок кожного місяця, тобто на певну дату, а в наступні роки – як середньомісячна чисельність).

Статистичні дані динамічного ряду можуть бути непорівняльними за **різними періодами або тривалістю часу**. Інтервали часу, за які наведені дані динамічного ряду, повинні бути рівні (місяць, квартал, півріччя і т.д.). Можлива непорівняльність **через різні одиниці виміру**.

Непорівняльність статистичних показників динаміки може бути зумовлена також **різною структурою сукупності** за ряд років.

Для приведення даних таких рядів до порівняльного виду використовують **стандартизацію структури** (стандартизовані коефіцієнти доходності, приросту виробництва і т.д.)

# Приклад


delta-homes
Налаштування
Вхідні - kotsuybynsky@g
INDU Quote - Dow Jones
Рамблер - медийный по
finance.rambler.ru/currencies/consensus/
рамблер.финансы β
Рамблер Почта Новости Финансы Недвижимость Авто Путешествия
Главная Новости Публикации Валюты Сервисы
Прогноз стоимости валют
1 ДОЛЛАР США
ПРОГНОЗ НА МЕСЯЦ
USD 64.53
76,93
62,50
48,07
33,64
ПРОГНОЗ
ДАННЫЕ ЦБ
65,0050 28.09.15
69
20 ОКТ
28 НОЯ
7 ЯНВ
16 ФЕВ
28 МАР
8 МАЙ
17 ИЮН
27 ИЮЛ
5 СЕН
15 ОКТ
3 НОЯ
ДАННЫЕ ЗА ПЕРИОД
МЕСЯЦ КВАРТАЛ ПОЛГОДА ГОД ВСЕ ВРЕМЯ

## Основні показники рядів динаміки

Завдання - шляхом аналізу рядів динаміки розкрити і охарактеризувати закономірності, що проявляються на різних етапах розвитку того чи іншого явища, виявити тенденції розвитку та їх особливості.

В процесі аналізу динаміки розраховують і використовують наступні аналітичні показники динаміки:

- абсолютний приріст,
- темп росту,
- темп приросту
- абсолютне значення одного відсотка приросту.

Розрахунок показників ґрунтується на абсолютному або відносному порівнянні між собою рівнів ряду динаміки. При цьому порівнюваний рівень називається **поточним**, а рівень, з яким роблять порівняння - **базисним**. За базу порівняння часто приймають або попередній рівень, або початковий (перший) рівень ряду динаміки.

Якщо кожний рівень порівнюється з попереднім, то отримують **ланцюгові показники динаміки**, а якщо кожний рівень порівнюють з одним і тим же рівнем, взятим за базу порівняння, то такі показники називаються **базисними.**

**Абсолютний приріст** (Δ) обчислюється як різниця між поточним та базисним рівнями і показує, на скільки одиниць підвищився або зменшився рівень порівняно з базисним, за певний період часу:

$$\Delta_{БАЗ} = Y_i - Y_1$$ **базисний приріст**

$$\Delta_{ЛАНЦ} = Y_i - Y_{i-1}$$ **ланцюговий приріст**

$Y_i$ – поточний рівень ряду динаміки;
$Y_1$ – початковий (перший) рівень ряду динаміки;
$Y_{i-1}$ – попередній рівень ряду динаміки
*Знак "+", "–" свідчить про напрям динаміки.*

## Приклад

| Роки | Валовий збір картоплі, млн. т. (у) | Абсолютний приріст, млн. грн. | | Темп росту | | Темп приросту | | Абсолютне значення 1% приросту, млн.т. (А) |
|---|---|---|---|---|---|---|---|---|
| | | Базисний $(\Delta_y^б)$ | Ланцюговий $(\Delta_y^л)$ | Базисний $(T_P^б)$ | Ланцюговий $(T_P^л)$ | Базисний $(T_{ПР}^б)$ | Ланцюговий $(T_{ПР}^л)$ | |
| 2000 | 20 | - | - | 1,000 | - | - | - | - |
| 2001 | 23 | 3 | 3 | 1,150 | 1,150 | 0,150 | 0,150 | 0,20 |
| 2002 | 27 | 7 | 4 | 1,350 | 1,174 | 0,350 | 0,174 | 0,23 |
| 2003 | 28 | 8 | 1 | 1,400 | 1,037 | 0,400 | 0,037 | 0,27 |
| 2004 | 30 | 10 | 2 | 1,500 | 1,074 | 0,500 | 0,074 | 0,28 |
| 2005 | 33 | 13 | 3 | 1,650 | 1,100 | 0,650 | 0,100 | 0,30 |

## Абсолютний приріст

$$\Delta_y^б = y_i - y_1, \quad або \quad \Delta_y^л = y_i - y_{i-1},$$

де $\Delta_y$ - абсолютний приріст;

$y_i$ - поточний рівень ряду динаміки;

$y_1$ - початковий (перший) рівень ряду динаміки;

$y_{i-1}$ - попередній рівень ряду динаміки.

**Коефіцієнт зростання** (Кр) вираховується як відношення порівнюваного рівня до базисного і показує, в скільки разів (відсотків) порівнюваний рівень більший або менший за базисний.

$$Кр_{БАЗ} = \frac{Y_i}{Y_1}$$ **базисний коефіцієнт зростання**

$$Кр_{ЛАНЦ} = \frac{Y_i}{Y_{i-1}}$$ **ланцюговий коефіцієнт зростання**

Між ланцюговими і базисними коефіцієнтами зростання існує взаємозв'язок – *добуток кількох послідовних ланцюгових коефіцієнтів зростання дорівнює базисному коефіцієнту зростання за відповідний період* і, навпаки, *поділивши наступний базисний коефіцієнт зростання на попередній, отримаємо відповідний ланцюговий коефіцієнт зростання.*

$Y_i$ – поточний рівень ряду динаміки;
$Y_1$ – початковий (перший) рівень ряду динаміки;
$Y_{i-1}$ – попередній рівень ряду динаміки
*Знак "+", "–" свідчить про напрям динаміки.*

## Приклад

| Роки | Валовий збір картоплі, млн. т. (у) | Абсолютний приріст, млн. грн. | | Темп росту | | Темп приросту | | Абсолютне значення 1% приросту, млн.т. (А) |
|---|---|---|---|---|---|---|---|---|
| | | Базисний $(\Delta_y^б)$ | Ланцюговий $(\Delta_y^л)$ | Базисний $(T_P^б)$ | Ланцюговий $(T_P^л)$ | Базисний $(T_{ПР}^б)$ | Ланцюговий $(T_{ПР}^л)$ | |
| 2000 | 20 | - | - | 1,000 | - | - | - | - |
| 2001 | 23 | 3 | 3 | 1,150 | 1,150 | 0,150 | 0,150 | 0,20 |
| 2002 | 27 | 7 | 4 | 1,350 | 1,174 | 0,350 | 0,174 | 0,23 |
| 2003 | 28 | 8 | 1 | 1,400 | 1,037 | 0,400 | 0,037 | 0,27 |
| 2004 | 30 | 10 | 2 | 1,500 | 1,074 | 0,500 | 0,074 | 0,28 |
| 2005 | 33 | 13 | 3 | 1,650 | 1,100 | 0,650 | 0,100 | 0,30 |

**Темп росту** $(T_P)$ вираховується як відношення порівнюваного рівня до базисного і показує, в скільки разів (відсотків) порівнюваний рівень більший або менший за базисний.

$$T_P^б = \frac{y_i}{y_1}, \qquad або \qquad T_P^л = \frac{y_i}{y_{i-1}}.$$

**Темп приросту** (Тпр) визначається як відношення абсолютного приросту до абсолютного попереднього або початкового рівня і показує на скільки відсотків порівнюваний рівень більший або менший рівня, прийнятого за базу порівняння.

$$Тпр_{БАЗ} = \frac{\Delta_{БАЗ}}{Y_1}$$

$$Тпр_{БАЗ} = (Кр_{БАЗ} - 1) \cdot 100$$

**базисний темп приросту**

$$Тпр_{ЛАНЦ} = \frac{\Delta_{ЛАНЦ}}{Y_{i-1}}$$

$$Тпр_{ЛАНЦ} = (Кр_{ЛАНЦ} - 1) \cdot 100$$

**ланцюговий темп приросту**

## Приклад

| Роки | Валовий збір картоплі, млн. т. (у) | Абсолютний приріст, млн. грн. | | Темп росту | | Темп приросту | | Абсолютне значення 1% приросту, млн.т. (А) |
|---|---|---|---|---|---|---|---|---|
| | | Базисний $(\Delta_y^б)$ | Ланцюговий $(\Delta_y^л)$ | Базисний $(T_P^б)$ | Ланцюговий $(T_P^л)$ | Базисний $(T_{ПР}^б)$ | Ланцюговий $(T_{ПР}^л)$ | |
| 2000 | 20 | - | - | 1,000 | - | - | - | - |
| 2001 | 23 | 3 | 3 | 1,150 | 1,150 | 0,150 | 0,150 | 0,20 |
| 2002 | 27 | 7 | 4 | 1,350 | 1,174 | 0,350 | 0,174 | 0,23 |
| 2003 | 28 | 8 | 1 | 1,400 | 1,037 | 0,400 | 0,037 | 0,27 |
| 2004 | 30 | 10 | 2 | 1,500 | 1,074 | 0,500 | 0,074 | 0,28 |
| 2005 | 33 | 13 | 3 | 1,650 | 1,100 | 0,650 | 0,100 | 0,30 |

**Темп приросту** $(T_{ПР})$ визначається як відношення абсолютного приросту до абсолютного попереднього або початкового рівня і показує на скільки відсотків порівнюваний рівень більший або менший рівня, прийнятого за базу порівняння.

$$T_{ПР}^б = \frac{\Delta_y^б}{y_1}, \quad \textit{або} \quad T_{ПР}^л = \frac{\Delta_y^л}{y_{i-1}}.$$

## Абсолютне значення одного відсотка приросту.

**Абсолютне значення одного відсотка приросту** (А) визначається шляхом ділення абсолютного приросту на темп приросту за один і той самий період. Абсолютне значення одного відсотка приросту можна вирахувати технічно більш легким шляхом, діленням початкового рівня на 100:

$$A_i = \frac{Y_i - Y_{i-1}}{\left(\frac{Y_i - Y_{i-1}}{Y_{i-1}}\right) \cdot 100} = \frac{Y_{i-1}}{100}$$

## Приклад

| Роки | Валовий збір картоплі, млн. т. (у) | Абсолютний приріст, млн. грн. | | Темп росту | | Темп приросту | | Абсолютне значення 1% приросту, млн.т. (А) |
|---|---|---|---|---|---|---|---|---|
| | | Базисний $(\Delta_y^б)$ | Ланцюговий $(\Delta_y^л)$ | Базисний $(T_P^б)$ | Ланцюговий $(T_P^л)$ | Базисний $(T_{ПР}^б)$ | Ланцюговий $(T_{ПР}^л)$ | |
| 2000 | 20 | - | - | 1,000 | - | - | - | - |
| 2001 | 23 | 3 | 3 | 1,150 | 1,150 | 0,150 | 0,150 | 0,20 |
| 2002 | 27 | 7 | 4 | 1,350 | 1,174 | 0,350 | 0,174 | 0,23 |
| 2003 | 28 | 8 | 1 | 1,400 | 1,037 | 0,400 | 0,037 | 0,27 |
| 2004 | 30 | 10 | 2 | 1,500 | 1,074 | 0,500 | 0,074 | 0,28 |
| 2005 | 33 | 13 | 3 | 1,650 | 1,100 | 0,650 | 0,100 | 0,30 |

**Абсолютне значення одного відсотка приросту** (А) визначається шляхом ділення абсолютного приросту на темп приросту за один і той самий період. Абсолютне значення одного відсотка приросту можна вирахувати технічно більш легким шляхом, діленням початкового рівня на 100.

$$A = \frac{\Delta_y}{T_{ПР}(\%)}, \qquad або \qquad A = \frac{y_б}{100}.$$

## Взаємозвязки. Абсолютне та відносне прискорення

Ланцюгові й базисні характеристики динаміки взаємопов'язані:

1) сума ланцюгових абсолютних приростів дорівнює кінцевому базисному:

$$\sum \Delta_{ЛАНЦ} = \Delta_{БАЗ\ К}$$

2) добуток ланцюгових коефіцієнтів зростання дорівнює кінцевому базисному:

$$\prod Кр_{ЛАНЦ} = Кр_{БАЗ\ К}$$

Якщо швидкість розвитку в межах періоду, що вивчається, неоднакова, порівнянням однойменних характеристик швидкості вимірюється прискорення чи уповільнення динаміки. На базі абсолютних приростів оцінюються **абсолютне та відносне прискорення**.

Абсолютне прискорення –різниця між абсолютними приростами: $\delta = \Delta_t - \Delta_{t-1}$

Відношення коефіцієнтів зростання - **коефіцієнт випередження,** дозволяє порівняти відносну швидкість динамічних рядів однакового змісту по різних об' єктах або різного змісту по одному об'єкту.

# Середні показники динаміки

Динамічні ряди складаються з багатьох варіаційних рівнів тому потребують узагальнюючих характеристик.

Середні показники:

- середні рівні ряду,
- середні абсолютні прирости,
- середні темпи росту і приросту.

В інтервальному ряду з рівними інтервалами середній рівень ряду вираховується за формулою середнього арифметичного :

$$\overline{Y} = \frac{\sum Y}{n}$$

$\sum Y$ сума показників по кожному з інтервалів

n число інтервалів

Якщо окремі періоди інтервального ряду динаміки мають різну довжину, то для визначення середнього рівня використовують рівняння для середнього зваженого :

$$\overline{Y} = \frac{\sum Y \cdot t}{\sum t}$$

Для визначення середнього рівня в моментному динамічному ряду з рівними інтервалами між сусідніми датами застосовують формулу середнього хронологічного:

$$\overline{Y} = \frac{\frac{Y_1}{2} + Y_2 + Y_3 + \ldots + Y_{n-1} + \frac{Y_n}{2}}{n-1}$$

**Середній абсолютний приріст** визначається як середня арифметичне з ланцюгових абсолютних приростів за певні періоди і показує на скільки одиниць в середньому змінився рівень у порівнянні з попереднім:

$$\overline{\Delta} = \frac{\sum \Delta_{ЛАНЦ}}{n-1} = \frac{\Delta_{БАЗ\ К}}{n-1}$$

**Середній коефіцієнт зростання** вираховується за формулою середнього геометричного:

$$\overline{Кр} = \sqrt[n-1]{\prod Кр_{ЛАНЦ}} = \sqrt[n-1]{Кр_{БАЗ\ К}}$$

**Середній темп приросту**

$$\overline{Тпр} = \left(\overline{Кр} - 1\right) \cdot 100$$

При інтерпретації середньої абсолютної чи відносної швидкості динаміки необхідно вказувати часовий інтервал, до якого належать середні, та часову одиницю вимірювання (рік, квартал, місяць, доба тощо).

# Приклад

Кількість працівників ПНУ в січні 2015 р

| На 0,01 | На 6,01 | На 15,01 | На 21,01 | На 29,01 | На 1,02 |
|---|---|---|---|---|---|
| 1210 | 1243 | 1236 | 1248 | 1238 | 1238 |

Середньоспискова чисельність працівників підприємства в січні місяці становитиме:

$$\bar{y} = \frac{\sum yt}{\sum t} = \frac{1210 \cdot 5 + 1243 \cdot 9 + 1236 \cdot 6 + 1248 \cdot 8 + 1238 \cdot 3}{31} = \frac{38351}{31} = 1237 \text{ чол.}$$

Динаміка зміни кількість тракторів в парку агрофірми за 2014 рік

| Дані на початок місяця | 1.01 | 1.02 | 1.03 | 1.04 | 1.05 | 1.06 | 1.07 | 1.08 | 1.09 | 1.10 | 1.11 | 1.12 | 1.01 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Число тракторів, шт. | 622 | 640 | 643 | 640 | 664 | 670 | 682 | 733 | 753 | 768 | 800 | 826 | 888 |

Приклад

Визначимо середнє число тракторів за кожний квартал, перше і друге півріччя і за рік в цілому.

$$\bar{y} = \frac{\frac{1}{2}y_1 + y_2 + y_3 \ldots + y_{n-1} + \frac{1}{2}y_n}{n-1}$$

середнє хронологічне

$$\bar{y}_{Iк.} = \frac{\frac{622}{2} + 640 + 643 + \frac{640}{2}}{3} = \frac{1914}{3} = 638\,шт.;$$

$$\bar{y}_{IIк.} = \frac{\frac{640}{2} + 664 + 670 + \frac{668}{2}}{3} = \frac{1995}{3} = 665\,шт.;$$

$$\bar{y}_{IIIк.} = \frac{\frac{682}{2}2 + 733 + 753 + \frac{768}{2}}{3} = \frac{2211}{3} = 737\,шт.;$$

$$\bar{y}_{IVк.} = \frac{\frac{768}{2} + 800 + 826 + \frac{888}{2}}{3} = \frac{2454}{3} = 188\,шт.;$$

$$\bar{y}_{I\,п.} = \frac{\bar{y}_{I\,к.} + \bar{y}_{II\,к.}}{2} = \frac{638 + 665}{2} = 651{,}5\,шт;$$

$$\bar{y}_{II\,п.} = \frac{\bar{y}_{III\,к.} + \bar{y}_{IV\,к.}}{2} = \frac{737 + 818}{2} = 777{,}5\,шт;$$

$$\bar{y}_{p} = \frac{\bar{y}_{I\,п.} + \bar{y}_{II\,п.}}{2} = \frac{651{,}5 + 777{,}5}{2} = 714{,}5\,шт.$$

# Виявлення тенденцій розвитку явищ

Виявлення основної тенденції (тренду) ряду, є одним з головних методів аналізу і узагальнення динамічних рядів.

Зображена на графіку лінія тренду динамічного ряду вказує зміну досліджуваного явища в часі, звільнену від короткочасних відхилень, викликаних випадковими причинами.

В статистичній практиці виявлення основної тенденції розвитку явищ в часі проводиться методами *укрупнення інтервалів*, *рухомої середньої* і *аналітичним вирівнюванням.*

Укрупнення інтервалів (періодів) часу

Суть цього методу - дані динамічного ряду об'єднуються в групи по періодах і розраховується середній показник на період - триріччя, п'ятиріччя і т.д.

Згладжування за допомогою рухомої середньої. - укрупнення періодів, яке проводиться шляхом послідовних зміщень на одну дату при збереженні постійного інтервалу періоду.

Аналітичне вирівнювання - апроксимація формулами, які виражають тенденцію розвитку (тренд) явищ (пряма, гіпербола, парабола другого порядку, показникова функція, ряди Фур'є, логістична функція, експонента ).

## Приклад

*Укрупнення інтервалів (періодів) часу*

| Роки | Урожайність озимої пшениці, ц/га | Сумарна врожайність, ц/га (за триріччя) | Середня врожайність, ц/га (за триріччя) |
|---|---|---|---|
| 1991 | 15,6 | 50,5 | 16,8 |
| 1992 | 16,0 | | |
| 1993 | 18,9 | | |
| 1994 | 15,7 | 55,3 | 18,4 |
| 1995 | 20,0 | | |
| 1996 | 19,6 | | |
| 1997 | 19,8 | 61,3 | 20,4 |
| 1998 | 21,5 | | |
| 1999 | 20,0 | | |
| 2000 | 27,3 | 79,9 | 26,6 |
| 2001 | 24,4 | | |
| 2002 | 28,2 | | |
| 2003 | 27,9 | 93,7 | 31,2 |
| 2004 | 33,1 | | |
| 2005 | 32,7 | | |

# Приклад

*Згладжування за допомогою рухомої середньої*

| Роки | Урожайність озимої пшениці, ц/га | Сумарна врожайність, ц/га (за триріччя) | Середня врожайність, ц/га (за триріччя) |
|---|---|---|---|
| 1991 | 15,6 | - | – |
| 1992 | 16,0 | 50,5 | 16,8 |
| 1993 | 18,9 | 50,6 | 16,9 |
| 1994 | 15,7 | 54,6 | 18,2 |
| 1995 | 20,0 | 55,3 | 18,4 |
| 1996 | 19,6 | 59,4 | 19,8 |
| 1997 | 19,8 | 60,9 | 20,3 |
| 1998 | 21,5 | 61,3 | 20,4 |
| 1999 | 20,0 | 68,8 | 22,9 |
| 2000 | 27,3 | 71,7 | 23,9 |
| 2001 | 24,4 | 79,9 | 26,6 |
| 2002 | 28,2 | 80,5 | 26,8 |
| 2003 | 27,9 | 89,2 | 29,7 |
| 2004 | 33,1 | 93,7 | 31,2 |
| 2005 | 32,7 | – | – |

**Вирівнювання за прямою** використовується в тих випадках, коли абсолютні прирости приблизно постійні, тобто коли рівні динамічного ряду змінюються в арифметичній прогресії, або близькі до неї.

Рівняння прямої має вигляд: $\hat{Y}_t = a_0 + a_1 t$

де $a_0$, $a_1$ – параметри прямої;

t –позначення часу

Вирівнювання радів динаміки використовують також для знаходження відсутніх членів раду за допомогою інтерполяції і екстраполяції.

**Інтерполяцією** називається в статистиці знаходження відсутнього показника усередині ряду.

**Екстраполяцією** в статистиці називається знаходження невідомих рівнів в кінці або на початку динамічного ряду.

Приклад

| Роки | Урожайність озимої пшениці, ц/га ($y$) | Умовне позначення часу ($t$) |
|---|---|---|
| 1991 | 15,6 | -7 |
| 1992 | 16,0 | -6 |
| 1993 | 18,9 | -5 |
| 1994 | 15,7 | -4 |
| 1995 | 20,0 | -3 |
| 1996 | 19,6 | -2 |
| 1997 | 19,8 | -1 |
| 1998 | 21,5 | 0 |
| 1999 | 20,0 | 1 |
| 2000 | 27,3 | 2 |
| 2001 | 24,4 | 3 |
| 2002 | 28,2 | 4 |
| 2003 | 27,9 | 5 |
| 2004 | 33,1 | 6 |
| 2005 | 32,7 | 7 |
| $n$=15 | 340,7 | 0 |

$$\begin{cases} \sum y = na_0 + a_1 \sum t; \\ \sum ty = a_0 \sum t + a_1 \sum t^2 \end{cases}$$

де $y$ - фактичні рівні динамічного ряду;

$n$ - число членів ряду динаміки.

При відліку часу від середини ряду коли $\sum t = 0$, тоді система рівнянь для знаходження параметрів «$a_0$» і «$a_1$» матиме вигляд: $\begin{cases} \sum y = na_0; \\ \sum ty = a_0 \sum t^2, \end{cases}$

звідки: $a_0 = \frac{\sum y}{n}$, $a_1 = \frac{\sum yt}{\sum t^2}$.

Приклад

| Роки | Урожайність озимої пшениці, ц/га ($y$) | Умовне позначення часу ($t$) | $t^2$ | $Yt$ | Вирівняна урожайність $\hat{y}_t = a_0 + a_1 t$ |
|---|---|---|---|---|---|
| 1991 | 15,6 | -7 | 49 | -109,2 | 14,1 |
| 1992 | 16,0 | -6 | 36 | -96,0 | 15,3 |
| 1993 | 18,9 | -5 | 25 | -94,5 | 16,5 |
| 1994 | 15,7 | -4 | 16 | -62,8 | 17,8 |
| 1995 | 20,0 | -3 | 9 | -60,0 | 19,0 |
| 1996 | 19,6 | -2 | 4 | -39,2 | 20,2 |
| 1997 | 19,8 | -1 | 1 | -19,8 | 21,5 |
| 1998 | 21,5 | 0 | 0 | 0 | 22,7 |
| 1999 | 20,0 | 1 | 1 | 20,0 | 23,9 |
| 2000 | 27,3 | 2 | 4 | 54,6 | 25,2 |
| 2001 | 24,4 | 3 | 9 | 73,2 | 26,4 |
| 2002 | 28,2 | 4 | 16 | 112,8 | 27,7 |
| 2003 | 27,9 | 5 | 25 | 139,5 | 28,9 |
| 2004 | 33,1 | 6 | 36 | 198,6 | 30,1 |
| 2005 | 32,7 | 7 | 49 | 228,9 | 31,4 |
| $n$=15 | 340,7 | 0 | 280 | 346,1 | 340,7 |

$$a_0 = \frac{\sum y}{n} = \frac{340,7}{15} = 22,713; \quad a_1 = \frac{\sum yt}{\sum t^2} = \frac{346,1}{280,0} = 1,236.$$

Звідси рівняння прямої буде мати наступний вигляд: $\hat{y}_t = 22,713 + 1,236t$.

Коефіцієнт регресії ($a_1$=1,236) характеризує середній приріст урожайності озимої пшениці за рік. Величина 22,713 буде показувати теоретичну врожайність 1998 р., для якого ми взяли «0» номер року. Підставляючи у рівняння $\hat{y}_t = 22,713 + 1,236t$ послідовно значення t (-7, -6, -5 і т.д.), отримаємо вирівняний (теоретичний) ряд динаміки урожайності озимої пшениці

## Характеристика сезонних коливань, методи їх вимірювання

**Сезонними коливаннями** називаються стійкі внутрішньорічні коливання в рядах динаміки, обумовлені специфічними умовами виробництва чи споживання певного виду продукції. Для дослідження внутрішньорічних коливань можна використати цілий ряд методів (середнього, Пірсонса, рухомої середньої, аналітичного вирівнювання, рядів Фур'є), які забезпечують їх оцінку з різною точністю, надійністю і трудоємкістю.

Сезонні коливання характеризуються індексом сезонності (Is).

В сукупності індекси сезонності утворюють *сезону хвилю*.

**Індекс сезонності** –процентне відношення однойменних місячних (квартальних) фактичних рівнів динамічних рядів до їх середньорічних або вирівняних рівнів.

Індекс сезонності (сезонну хвилю) розраховують **методом простих середніх**:

$$I_s = \frac{\overline{Yi}}{\overline{Yз}}$$

де Is - індекс сезонності;

Yi – середні місячні або квартальні рівні;

Yз – загальна середня (по всіх місячна або квартальна).

Приклад

Кількість проданих мережею автосалонів автомобілів певної марки впродовж трьох років

| Квартал | Роки | | | Разом | В середньому ($\bar{y}_i$) | Сезонна хвиля, % $I_s = \frac{\bar{y}_i}{\bar{y}_з} \cdot 100$ |
|---|---|---|---|---|---|---|
| | 2012 | 2013 | 2014 | | | |
| I | 1942 | 2126 | 2505 | 6573 | 2191,00 | 82,1 |
| II | 2957 | 2704 | 3704 | 9365 | 3121,67 | 117,0 |
| III | 2504 | 3291 | 3834 | 9629 | 3209,67 | 120,3 |
| IV | 2194 | 1745 | 2513 | 6452 | 2150,67 | 80,6 |
| Разом | 9597 | 9866 | 12556 | 32019 | $\bar{y}_з$=2668,25 | 400,0 |

Індекс сезонності за методом простої середньої визначається за формулою:

$$I_s = \frac{\bar{y}_i}{\bar{y}_з} \cdot 100,$$

$I_s$ - індекс сезонності;

$\bar{y}_i$ – середні місячні або квартальні рівні;

$\bar{y}_з$ – загальна середня (місячна або квартальна).

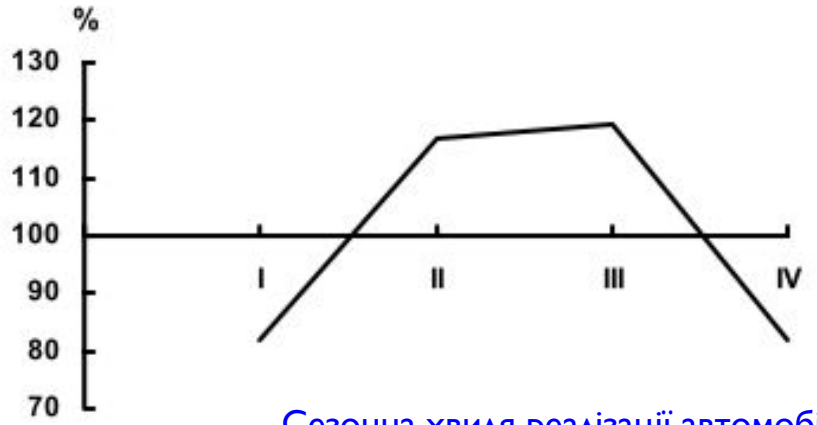
$$I_s^I = \frac{\bar{y}_1}{\bar{y}_з} \cdot 100 = \frac{2191,00}{2668,25} \cdot 100 = 82,1\%;$$

$$I_s^{II} = \frac{\bar{y}_2}{\bar{y}_з} \cdot 100 = \frac{3121,67}{2668,25} \cdot 100 = 117,0\%.$$

Приклад

Кількість проданих мережею автосалонів автомобілів певної марки впродовж трьох років

| Квартал | Роки | | | Разом | В середньому ($\bar{y}_i$) | Сезонна хвиля, % $I_s = \frac{\bar{y}_i}{\bar{y}_з} \cdot 100$ |
|---|---|---|---|---|---|---|
| | 2012 | 2013 | 2014 | | | |
| I | 1942 | 2126 | 2505 | 6573 | 2191,00 | 82,1 |
| II | 2957 | 2704 | 3704 | 9365 | 3121,67 | 117,0 |
| III | 2504 | 3291 | 3834 | 9629 | 3209,67 | 120,3 |
| IV | 2194 | 1745 | 2513 | 6452 | 2150,67 | 80,6 |
| Разом | 9597 | 9866 | 12556 | 32019 | $\bar{y}_з$=2668,25 | 400,0 |




Сезонна хвиля реалізації автомобілів

# Лекція 7

## Індекси

1. Суть та функції індексів у статистичному дослідженні.
2. Види індексів.
3. Методологічні принципи побудови агрегатних індексів.
4. Середньозважені індекси, приведення їх до агрегатної форми.
5. Індекси середніх величин: змінного складу; фіксованого складу і структурних зрушень; їх взаємозв’язок.
6. Характеристики індексів.

## Суть та функції індексів у статистичному дослідженні

Для характеристики соціально-економічних явищ і процесів статистика використовує узагальнюючі показники у вигляді середніх, відносних величин та коефіцієнтів.

*Одним з таких узагальнюючих показників є індекси.*

Індексом у статистиці називається відносний показник, що характеризує зміну рівня соціально-економічного явища в часі або в просторі порівняно з планом, базисним періодом .

Виділяють три сфери застосування економічних індексів.

До *першої сфери* застосування індексів відносять порівняльну характеристику сукупностей параметрів в часі.

Сюди входять індекси динаміки, виконання плану і територіальні індекси.

**Індекси динаміки** - показують зміну якого-небудь складного явища в звітному періоді порівняно з базисним.

**Індекс виконання плану** - використовують для порівняння досягнутого рівня з плановими завданнями.

**Територіальні індекси** застосовують для просторового порівняння рівнів урожайності, цін, продуктивності праці і т.п., в різних регіонах.

# Суть та функції індексів у статистичному дослідженні

Друга сфера застосування індексів полягає у їх використанні для *факторного аналізу складного явища через систему взаємозв'язаних індексів*.

До таких складних явищ можуть бути віднесені
вартість виробленої чи реалізованої продукції, фонд заробітної плати, валовий збір зерна та ін.

*Приклади* :

вартість виробленої продукції дорівнює добутку цін на кількість продукції,

валовий збір зерна – добутку урожайності на посівну площу,

фонд заробітної плати – добутку заробітної плати одного працівника на їх чисельність

За допомогою третьої сфери застосування індексів проводять *аналіз динаміки середніх величин, зміна яких піддається впливу структурних зрушень в середині досліджуваної сукупності.*

Велике значення має вивчення впливу структурних зрушень на динаміку середніх показників через застосування системи взаємозв'язаних індексів змінного складу, постійного (фіксованого) складу і структурних зрушень.

# Класифікація індексів

Всі економічні індекси статистика класифікує за *трьома основними ознаками*:

а) за характером досліджуваних об’єктів;
б) за ступенем охоплення елементів сукупності;
в) за методикою розрахунку загальних індексів.

За характером досліджуваних об'єктів індекси ділять на індекси *об'ємних* (*кількісних*) і
*якісних* показників.

До *першої групи* відносяться індекси фізичного обсягу продукції промисловості, сільського господарства, будівництва та ін.
До другої групи якісних показників відносять індексів цін, собівартості, урожайності і ряд інших.

# Класифікація індексів

За ступеня охоплення елементів сукупності індекси ділять на:

а) індивідуальні;

б) загальні;

в) групові.

**Індивідуальні індекси** характеризують зміну окремих елементів складного явища. В теорії індексів показник, зміну якого характеризує індекс, називається **індексованою величиною**.

Індивідуальні індекси позначають малою латинською буквою «*i*», продукцію в натуральному виразі – через «*q*», ціну одиниці товару – через «*p*», собівартість одиниці продукції – через «*z*» і т.д. Індивідуальні індекси цих ознак визначаються за формулами:

а) фізичного обсягу: $i_q = \frac{q_1}{q_0}$;

б) ціни одиниці товару: $i_p = \frac{p_1}{p_0}$;

в) собівартості одиниці продукції: $i_z = \frac{z_1}{z_0}$;

де $i_q, i_p, i_z$ – індивідуальні індекси фізичного обсягу, ціни і собівартості одиниці продукції;

$q_1$, $q_0$; $p_1$, $p_0$; $z_1$, $z_0$ – фізичний обсяг, ціна, собівартість у звітному і базисному періодах.

# Класифікація індексів

**Загальні індекси** характеризують зміну сукупності в цілому і являють собою відносні числа, що визначають зміни в часі порівняно з плановим, базисним періодами або в просторі складного явища, яке складається з несумірних елементів.

**Груповими** або **субіндексами** називаються такі індекси, які охоплюють не всі елементи сукупності, а тільки яку-небудь частину або їх групу.

**В залежності від методології обчислення,** загальні і групові індекси діляться на агрегатні і середні з індивідуальних індексів.

**Агрегатні індекси** є основною формою економічних індексів, а середні із індивідуальних індексів – похідними, отриманими в результаті перетворення агрегатних індексів.

**Базисні і ланцюгові індекси** обчислюють в тих випадках, коли доводиться вивчати яке-небудь явище суспільного життя за ряд послідовних років.

## Агрегатні індекси як вихідна форма індексів

Агрегатним індексом в статистиці називається загальний індекс, який є відношенням сум добутків індексованих (зіставлюваних) величин порівнюваних періодів на їх відносні ваги

При побудові формул агрегатних індексів використовують наступне правило: «*якщо індексована величина – якісний показник, який знаходять шляхом ділення (ціна, собівартість, урожайність і т.д.) ваги беруться звітного періоду, а якщо індексована величина – кількісний показник, який можна підсумувати (фізичний обсяг продукції, чисельність працівників, посівна площа) ваги беруться базисного періоду*».

**Загальний індекс цін** визначається за формулою:

$$I_p = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1};$$

де *Ip* – загальний індекс цін;

*індекс показує, як змінилися ціни на всі досліджувані товари в звітному періоді порівняно з базисним*

Сума економії або перевитрат від зміни цін визначається як різниця між чисельником і знаменником загального індекса цін:

$$\Delta pq(p) = \Sigma\, p_1 q_1 - \Sigma\, p_0 q_1.$$

# Агрегатні індекси як вихідна форма індексів

**Загальний індекс фізичного обсягу** визначається за формулою:

$$I_q = \frac{\sum q_1 p_0}{\sum q_0 p_0},$$

де $I_q$ – загальний індекс фізичного обсягу продукції;

*індекс показує зміну кількості виробленої або реалізованої продукції в звітному періоді порівняно з базисним*

**Загальний індекс обсягу товарообороту**

$$I_{pq} = \frac{\sum p_1 q_1}{\sum p_0 q_0}.$$

*індекс показує зміну виробництва або реалізації продукції в звітному періоді порівняно з базисним у фактичних цінах*

*індекси взаємозв'язані*

$$I_p \cdot I_q = I_{pq} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \cdot \frac{\sum q_1 p_0}{\sum q_0 p_0} = \frac{\sum p_1 q_1}{\sum p_0 q_0}, \qquad I_p = \frac{I_{pq}}{I_q}, \qquad I_q = \frac{I_{pq}}{I_p}.$$

$q_1$, $q_0$; $p_1$, $p_0$; $z_1$, $z_0$ – фізичний обсяг, ціна, собівартість у звітному і базисному періодах.

Приклад

| Назва товару | Продано товарів, тис. кг. | | Середня ціна одного кілограма, грн. | |
|---|---|---|---|---|
| | 2004р. | 2005р. | 2004р. | 2005р. |
| Морква, кг | 15,3 | 16,2 | 0,22 | 0,20 |
| Яблука, кг | $q_0$ 49,8 | $q_1$ 51,6 | $p_0$ 0,90 | 0,85 $p_1$ |

Обчислимо індивідуальні індекси цін і фізичного обсягу:

$$i_p = \frac{p_1}{p_0} = \frac{0{,}20}{0{,}22} = 0{,}909; \qquad i_p = \frac{p_1}{p_0} = \frac{0{,}85}{0{,}90} = 0{,}944;$$

$$i_q = \frac{q_1}{q_0} = \frac{16{,}2}{15{,}3} = 1{,}059; \qquad i_q = \frac{q_1}{q_0} = \frac{51{,}6}{49{,}8} = 1{,}036.$$

Ціна кілограма моркви в січні 2005р. порівняно з січнем 2004р. знизилась на 9,1 % (100 – 90,9), яблук – на 5,6 % (100 – 94,4), а кількість реалізованої моркви за цей же період збільшилась в 1,059 рази, або на 5,9 % (105,9 – 100), яблук – в 1,036 рази або на 3,6 %.

Визначимо загальний індекс цін на дані продукти:

$$I_p = \frac{\sum p_1 q_1}{\sum p_0 q_1} = \frac{0{,}20 \cdot 16{,}2 + 0{,}85 \cdot 51{,}6}{0{,}22 \cdot 16{,}2 + 0{,}90 \cdot 51{,}6} = \frac{47{,}100}{50{,}004} = 0{,}942.$$

Абсолютна сума виграшу населення від зниження цін:

$$\Delta pq(p) = \sum p_1 q_1 - \sum p_0 q_1 = 47{,}100 - 50{,}004 = -2{,}904 \text{ тис. грн.}$$

## Приклад

| Назва товару | Продано товарів, тис. кг. | | Середня ціна одного кілограма, грн. | |
|---|---|---|---|---|
| | 2004р. | 2005р. | 2004р. | 2005р. |
| Морква, кг | 15,3 | 16,2 | 0,22 | 0,20 |
| Яблука, кг | $q_0$ 49,8 | $q_1$ 51,6 | $p_0$ 0,90 | 0,85 $p_1$ |

Індекси фізичного обсягу продуктів дорівнює:

$$I_q = \frac{\sum q_1 p_0}{\sum q_0 p_0} = \frac{16{,}2 \cdot 0{,}22 + 51{,}6 \cdot 0{,}90}{15{,}3 \cdot 0{,}22 + 49{,}8 \cdot 0{,}90} = \frac{50{,}004}{48{,}196} = 1{,}037.$$

Загальний індекс обсягу товарообороту у фактичних цінах:

$$I_{pq} = \frac{\sum p_1 q_1}{\sum p_0 q_0} = \frac{47{,}100}{48{,}196} = 0{,}977.$$

Абсолютна зміна товарообороту у фактичних цінах становить:

$$\Delta pq = \sum p_1 q_1 - \sum p_0 q_0 = 47{,}100 - 48{,}196 = -1{,}096 \text{ тис. грн.},$$

в тому числі за рахунок зміни кількості проданих товарів:

$$\Delta pq(q) = \sum q_1 p_0 - \sum q_0 p_0 = 50{,}004 - 48{,}196 = 1{,}0808 \text{ тис. грн.}$$

Перевіримо правильність наших розрахунків:

$$I_{pq} = I_p \cdot I_q = 0{,}977 = 0{,}942 \cdot 1{,}037;$$

$$\Delta pq = \Delta pq(p) + \Delta pq(q) = -1{,}096 = -2{,}904 + 1{,}808.$$

Висновок:

Ціни на продукти на ринку в січні 2005р. порівняно з січнем 2004р. знизились на 5,8 %, внаслідок чого населення зекономило 2,904 тис. грн., кількість реалізованих продуктів за цей же період збільшилась в 1,037 рази або на 3,7 %, а товарооборот у фактичних цінах – зменшився на 2,3 % або на 1,096 тис. грн., в тому числі за рахунок зниження цін на 2,904 тис. грн., а за рахунок збільшення кількості проданих товарів він зріс на 1,808 тис. грн.

# Індекси із собівартості і кількості виготовленої продукції

Агрегатний індекс собівартості продукції: $I_z = \frac{\sum z_1 q_1}{\sum z_0 q_1}$.

Загальний індекс фізичного обсягу продукції: $I_q = \frac{\sum q_1 z_0}{\sum q_0 z_0}$.

Загальний індекс обсягу затрат на виробництво продукції: $I_{zq} = \frac{\sum z_1 q_1}{\sum z_0 q_0}$.

Індекси взаємозв'язані

$$I_z \cdot I_q = I_{zq} = \frac{\sum z_1 q_1}{\sum z_0 q_1} \cdot \frac{\sum q_1 z_0}{\sum q_0 z_0} = \frac{\sum z_1 q_1}{\sum z_0 q_0}, \quad I_z = \frac{I_{zq}}{I_q}, I_q = \frac{I_{zq}}{I_z}.$$

Загальна зміна затрат на виробництво продукції в звітному періоді порівняно з базисним дорівнює: $\Delta zq = \Delta zq(z) + \Delta zq(q) = \sum z_1 q_1 - \sum z_0 q_0$, в тому числі за рахунок:

а) зміни собівартості одиниці продукції: $\Delta zq(z) = \sum z_1 q_1 - \sum z_0 q_1$;

б) зміни кількості виробленої продукції: $\Delta zq(q) = \sum q_1 z_0 - \sum q_0 z_0$.

$q_1$, $q_0$; $p_1$, $p_0$; $z_1$, $z_0$ – фізичний обсяг, ціна, собівартість у звітному і базисному періодах.

# Середньозважені індекси

Перетворюють агрегатний індекс в середній з індивідуальних індексів, підставляючи у його чисельник або знаменник замість індексованого показника його вираз, виведений з формули індивідуального індекса. Якщо таку заміну роблять у чисельнику, то агрегатний індекс перетворюється у середній арифметичний, а якщо у знаменнику – в середній гармонічний.

Перетворимо агрегатний індекс фізичного обсягу в середній арифметичний.

$$I_q = \frac{\sum q_1 p_0}{\sum q_0 p_0}; \qquad i_q = \frac{q_1}{q_0}, \qquad \textit{звідси} \qquad q_1 = i_q \cdot q_0.$$

Замінивши в формулі агрегатного індекса фізичного обсягу продукції індексовану величину «$q_1$» на «$i_q \cdot q_0$», отримаємо формулу **середнього арифметичного індекса** фізичного обсягу продукції:

$$\bar{I}_q = \frac{\sum i_q q_0 p_0}{\sum q_0 p_0}.$$

$q_1$, $q_0$; $p_1$, $p_0$; $z_1$, $z_0$ – фізичний обсяг, ціна, собівартість у звітному і базисному періодах.

## Приклад

| Товарні групи | Продано в 2004р., тис. грн. $(q_0 p_0)$ | Індекси кількості проданих товарів $(i_q)$ |
|---|---|---|
| Трикотажні вироби | 150 | 0,98 |
| Тканини | 200 | 1,05 |
| Галантерея | 30 | 1,20 |

Середній арифметичний індекс фізичного обсягу реалізованих товарів дорівнює: $\bar{I}_q = \frac{\sum i_q q_0 p_0}{\sum q_0 p_0} = \frac{0{,}98 \cdot 150 + 1{,}05 \cdot 200 + 1{,}20 \cdot 30}{150 + 200 + 30} = \frac{393}{380} = 1{,}034.$

Перетворимо агрегатний індекс цін у **середній гармонічний.**

$$I_p = \frac{\sum p_1 q_1}{\sum p_0 q_1}; i_p = \frac{p_1}{p_0}, \text{ звідси } p_0 = \frac{p_1}{i_p}, \quad \text{а } \bar{I}_p = \frac{\sum p_1 q_1}{\sum \frac{p_1 q_1}{i_p}}.$$

Приклад

| Товари | Товарооборот в діючих цінах, тис. грн. | | Зміна середніх цін в II кварталі порівняно з I кварталом, % |
|---|---|---|---|
| | I квартал | II квартал | |
| Овочі | 60 | 64 | - 20 |
| М'ясо | 42 | 44 | + 10 |
| Зерно | 35 | 38 | без змін |

Спочатку визначимо індивідуальні індекси цін:

$$i_p = \frac{p_1}{p_0} = \frac{100-20}{100} = 0{,}8; \quad i_p = \frac{p_1}{p_0} = \frac{100+10}{100} = 1{,}1; \quad i_p = \frac{p_1}{p_0} = \frac{100}{100} = 1{,}0.$$

Обчислимо середній гармонічний індекс цін:

$$\overline{I}_p = \frac{\sum p_1 q_1}{\sum \frac{p_1 q_1}{i_p}} = \frac{64+44+38}{\frac{64}{0{,}8}+\frac{44}{1{,}1}+\frac{38}{1{,}0}} = \frac{146}{158} = 0{,}924.$$

Аналогічно перетворюють агрегатний індекс собівартості в середній гармонічний індекс:

$$I_z = \frac{\sum z_1 q_1}{\sum z_0 q_1}; \qquad i_z = \frac{z_1}{z_0}; \qquad z_0 = \frac{z_1}{i_z}; \qquad \overline{I}_z = \frac{\sum z_1 q_1}{\sum \frac{z_1 q_1}{i_z}}.$$

Середні арифметичні і гармонічні індекси повинні співпадати за своєю величиною з відповідними агрегатними індексами.

Вибір форми індекса залежить від поставленого завдання дослідження і від наявності даних, необхідних для обчислення того чи іншого індекса.

# Базисні і ланцюгові індекси з постійними і змінними вагами

В ряді випадків доводиться аналізувати явища суспільного життя не за два, а за три і більше послідовних періодів. В такому разі, в залежності від бази порівняння, обчислюють індекси з постійною базою порівняння (базисні) і змінною базою порівняння (ланцюгові).

**Базисними** називаються індекси, які вираховуються шляхом порівняння даних кожного періоду з даними будь-якого одного періоду, прийнятого за базу порівняння, Наприклад:

$$i_{q1} = \frac{q_1}{q_0}; \quad i_{q2} = \frac{q_2}{q_0}; \quad \ldots \; i_{qn} = \frac{q_n}{q_0}.$$

**Ланцюговими** називаються індекси, обчислені шляхом порівняння даних кожного періоду з даними попереднього періоду, Наприклад:

$$i_{q1} = \frac{q_1}{q_0}; \quad i_{q2} = \frac{q_2}{q_1}; \quad i_{q3} = \frac{q_3}{q_2}; \quad \ldots \; i_{qn} = \frac{q_n}{q_{n-1}}.$$

# Базисні і ланцюгові індекси з постійними і змінними вагами

Між базисними і ланцюговими індексами існує взаємозв'язок, що дозволяє переходити від одного виду індексів до іншого.

Базисні індекси можна визначити через ланцюгові, послідовно перемноживши останні:

$$i_q = \frac{q_1}{q_0} \cdot \frac{q_2}{q_1} \cdot \frac{q_3}{q_2} = \frac{q_3}{q_0}.$$

Ланцюгові індекси визначають через базисні шляхом ділення відповідного базисного індекса до попереднього базисного індекса:

$$i_q \frac{q_2}{q_0} : \frac{q_1}{q_0} = \frac{q_2 q_0}{q_0 q_1} = \frac{q_2}{q_1}.$$

Базисні і ланцюгові індекси є індивідуальні і загальні.

Якщо порівнюваних періодів три і більше, то загальні (базисні і ланцюгові) індекси обчислюють з постійними і змінними вагами.

# Базисні і ланцюгові індекси з постійними і змінними вагами

Якщо для всього індексованого ряду беруть ваги якогось одного періоду, то отримують базисні і ланцюгові індекси з **постійними вагами**. Якщо ваги змінюються від одного індекса до іншого, то матимемо базисні і ланцюгові індекси із **змінними вагами**.

Базисні індекси цін з **постійними вагами**:

$$I_{p1} = \frac{\sum p_1 q_0}{\sum p_0 q_0}; \qquad I_{p2} = \frac{\sum p_2 q_0}{\sum p_0 q_0}; \qquad I_{p3} = \frac{\sum p_3 q_0}{\sum p_0 q_0} \text{ і т.д.}$$

Ланцюгові індекси з **постійними вагами**:

$$I_{p1} = \frac{\sum p_1 q_0}{\sum p_0 q_0}; \qquad I_{p2} = \frac{\sum p_2 q_0}{\sum p_1 q_0}; \qquad I_{p3} = \frac{\sum p_3 q_0}{\sum p_2 q_0} \text{ і т.д.}$$

Базисні індекси із **змінними вагами**:

$$I_{p1} = \frac{\sum p_1 q_1}{\sum p_0 q_1}; \qquad I_{p2} = \frac{\sum p_2 q_2}{\sum p_0 q_2}; \qquad I_{p3} = \frac{\sum p_3 q_3}{\sum p_0 q_3} \text{ і т.д.}$$

Ланцюгові індекси із **змінними вагами**:

$$I_{p1} = \frac{\sum p_1 q_1}{\sum p_0 q_1}; \qquad I_{p2} = \frac{\sum p_2 q_2}{\sum p_1 q_2}; \qquad I_{p3} = \frac{\sum p_3 q_3}{\sum p_2 q_3} \text{ і т.д.}$$

# Індекси змінного, постійного складу і структурних зрушень

Для якісних показників, таких як середня ціна, собівартість, урожайність та інших по однойменній продукції, але віднесеної до різних об'єктів, обчислюють загальні індекси змінного, постійного (фіксованого) складу і структурних зрушень.

Індекси, який характеризує спільний вплив обох чинників, називається **індексом змінного складу** і визначається за формулою:

$$I_{\bar{z}} = \frac{\bar{Z}_1}{\bar{Z}_0} = \frac{\sum z_1 q_1}{\sum q_1} : \frac{\sum z_0 q_0}{\sum q_0},$$

де $I_{\bar{z}}$ – загальний індекс собівартості продукції змінного складу;

$z_1, z_0$ – собівартість одиниці продукції в звітному і базисному періодах;

$q_1, q_0$ – кількість виробленої продукції в натуральному виразі в звітному і базисному періодах;

$\bar{z}_1, \bar{z}_0$ – осереднені ознаки.

# Індекси змінного, постійного складу і структурних зрушень

На величину індекса собівартості змінного складу впливають зміни рівнів собівартості і зміни в структурі (її складі). Щоб виявити роль кожного чинника в загальній динаміці середньої, потрібно індекс змінного складу розкласти на два індекси-співмножники, кожний з яких відображає вплив тільки одного чинника.

Перший індекс, який характеризує вплив тільки індексованої величини (в якому змінюється лише собівартість), називається **індексом постійного (фіксованого) складу**. Він обчислюється за формулою:

$$I_z = \frac{\sum z_1 q_1}{\sum q_1} : \frac{\sum z_0 q_1}{\sum q_1} = \frac{\sum z_1 q_1}{\sum z_0 q_1}.$$

Другий індекс показує, як змінюється середній рівень (середня собіварттість) тільки за рахунок зміни структури явища (структури продукції). Він називається **індексом структурних зрушень** і визначається за формулою:

$$I_{zdq} = \frac{\sum z_0 q_1}{\sum q_1} : \frac{\sum z_0 q_0}{\sum q_0} = \frac{\sum z_0 q_1}{\bar{z}_0 \sum q_1}.$$

$$I_{\bar{z}} = I_{zdq} \cdot I_z; \qquad I_z = \frac{I_{\bar{z}}}{I_{zdq}}; \quad I_{zdq} = \frac{I_{\bar{z}}}{I_z}.$$

Дані про середню собівартість продукції «А» на двох заводах.

Приклад

| Номер заводу | Базисний період | | | Звітний період | | |
|---|---|---|---|---|---|---|
| | Вироблено продукції, тис.шт. ($q_0$) | Собівартість одиниці, грн. ($z_0$) | Частка продукції, ($d_0$) | Вироблено продукції, тис.шт. ($q_1$) | Собівартість одиниці, грн. ($z_1$) | Частка продукції, ($d_1$) |
| 1 | 70 | 25 | 50 | 80 | 22 | 40 |
| 2 | 70 | 23 | 50 | 120 | 20 | 60 |
| Разом: | 140 | x | 100 | 200 | x | 100 |

Обчислимо індекс собівартості продукції змінного складу:

$$I_{\bar{z}} = \frac{\bar{Z}_1}{\bar{Z}_0} = \frac{\sum z_1 q_1}{\sum q_1} : \frac{\sum z_0 q_0}{\sum q_0} = \frac{22 \cdot 80 + 20 \cdot 120}{200} : \frac{25 \cdot 70 + 23 \cdot 70}{140} =$$

$$= \frac{4160}{200} : \frac{3360}{140} = \frac{20{,}8}{24{,}0} = 0{,}867, \textit{ або } 86{,}7\%.$$

Отже, середня собівартість знизилась на 13,3 % (100,0 – 86,7 %).

Економія на одиницю продукції становить – 3,2 грн. (20,8 – 24,0), а на весь обсяг продукції звітного періоду – [-640 тис. грн. ($-3{,}2 \cdot 200$)].

Зниження середньої собівартості одиниці продукції зумовлене зміною собівартості продукції на кожному заводі і зміною структури продукції.

## Приклад

Обчислимо індекси собівартості: а) постійного (фіксованого) складу:

$$I_{\bar{z}} = \frac{\sum z_1 q_1}{\sum q_1} : \frac{\sum z_0 q_1}{\sum q_1} = \frac{22 \cdot 80 + 20 \cdot 120}{200} : \frac{25 \cdot 80 + 23 \cdot 120}{200} = \frac{4160}{200} : \frac{4760}{200} =$$

$$= \frac{20{,}8}{23{,}8} = 0{,}874, \qquad \textit{або} \qquad 87{,}4\,\%.$$

Собівартість продукції по двох заводах разом в середньому знизилась на 12,6 % (100,0 – 87,4). Економія затрат на виробництво продукції в звітному періоді складає 600 тис. грн. [(20,8 – 23,8)· 200];

б) структурних зрушень:

$$I_{zdq} = \frac{\sum z_0 q_1}{\sum q_1} : \frac{\sum z_0 q_0}{\sum q_0} = \frac{23{,}8}{24{,}0} = 0{,}992, \qquad \textit{або} \qquad 99{,}2\,\%.$$

Це означає, що середня собівартість вибору «А» в звітному періоді додатково знизилась на 0,8 % (100,0 – 99,2) за рахунок зміни структури, тобто за рахунок збільшення частки продукції другого заводу з 50 % до 60 %, на якому рівень собівартості був дещо нижчим в порівнянні з першим заводом. За рахунок цієї зміни економія затрат виробництва досягла в звітному періоді 40 тис. грн. [(23,8 – 24,0)· 200].

## Приклад

Проведемо перевірку наших розрахунків:

$$I_{\bar{z}} = I_z \cdot I_{zdq} = 0{,}867 = 0{,}874 \cdot 0{,}992;$$
$$\Delta_{zq} = \Delta_{zq}(z) + \Delta_{zq}(dq) = 640 = 600 + 40.$$

Отже, всі індекси обчислені правильно.

---

Вираховані нами індекси можна обчислити іншим способом, за частками продукції заводів, вираженими в коефіцієнтах:

а) Індекс собівартості змінного складу:

$$I_{\bar{z}} = \frac{\sum z_1 d_1}{\sum z_0 d_0} = \frac{22 \cdot 0{,}4 + 20 \cdot 0{,}6}{25 \cdot 0{,}5 + 23 \cdot 0{,}5} = \frac{20{,}8}{24{,}0} = 0{,}867, \quad \textit{або} \quad 86{,}7\,\%;$$

де $d_1, d_0$ – коефіцієнти частки продукції заводів в звітному і базисному періодах;

б) Індекс собівартості продукції постійного складу:

$$I_z = \frac{\sum z_1 d_1}{\sum z_0 d_1} = \frac{22 \cdot 0{,}4 + 20 \cdot 0{,}6}{25 \cdot 0{,}4 + 23 \cdot 0{,}6} = \frac{20{,}8}{23{,}8} = 0{,}874, \quad \textit{або} \quad 87{,}4\,\%;$$

в) Індекс струтурних зрушень:

$$I_{zdq} = \frac{\sum z_0 d_1}{\sum z_0 d_0} = \frac{25 \cdot 0{,}4 + 23 \cdot 0{,}6}{25 \cdot 0{,}5 + 23 \cdot 0{,}5} = \frac{23{,}8}{24{,}0} = 0{,}992, \quad \textit{або} \quad 99{,}2\,\%,$$

$$\textit{або} \quad I_{zdq} = \frac{I_{\bar{z}}}{I_z} = \frac{0{,}867}{0{,}874} = 0{,}992.$$

## Територіальні індекси

В практиці статистичних досліджень часто виникає потреба зпівставлення рівнів економічних явищ в просторі, для чого використовують територіальні індекси. Територіальні індекси –узагальнюючі відносні величини, що дають порівняльну характеристику в розрізі територій або об’єктів.

При побудові територіальних індексів якісних показників вагами можуть вступати:

- кількісний (екстенсивний) показник тієї території, на якій якісний (інтенсивний) показник економічно кращий;
- кількісний показник однієї з двох порівнюваних територій (об'єктів);
- середній кількісний показник з багатьох порівнюваних територій (об'єктів);
- об'ємний кількісний показник (сума екстенсивних показників декількох територій або об'єктів);
- кількісний показник, прийнятий за стандарт.

# Приклад

| Культура | Середня урожайність, ц/га ($y$) | | Посівна площа, га ($П$) | | | |
|---|---|---|---|---|---|---|
| | по району «А» | по району «В» | по району «А» | по району «В» | по області | |
| | | | | | в га | в % |
| Пшениця | 36 | 42 | 190 | 210 | 2850 | 50 |
| Жито | 19 | 23 | 80 | 100 | 1200 | 21 |
| Ячмінь | 25 | 30 | 110 | 150 | 1650 | 29 |

Обчислимо територіальні індекси урожайності зернових:

а) з вагами району «В»:

$$I_y = \frac{\sum y_A П_B}{\sum y_B П_B} = \frac{36 \cdot 210 + 19 \cdot 100 + 25 \cdot 150}{42 \cdot 210 + 23 \cdot 100 + 30 \cdot 150} = \frac{13210}{15620} = 0{,}846;$$

б) з вагами району «А»:

$$I_y = \frac{\sum y_A П_A}{\sum y_B П_A} = \frac{36 \cdot 190 + 19 \cdot 80 + 25 \cdot 110}{42 \cdot 190 + 23 \cdot 80 + 30 \cdot 110} = \frac{11110}{13120} = 0{,}847;$$

в) з вагами бази порівняння:

$$I_y = \frac{\sum y_B П_A}{\sum y_A П_A} = \frac{42 \cdot 190 + 23 \cdot 80 + 30 \cdot 110}{36 \cdot 190 + 19 \cdot 80 + 25 \cdot 110} = \frac{13120}{11110} = 0{,}181;$$

г) з вагами району «В»:

$$I_y = \frac{\sum y_B П_B}{\sum y_A П_B} = \frac{42 \cdot 210 + 23 \cdot 100 + 30 \cdot 150}{36 \cdot 210 + 19 \cdot 100 + 25 \cdot 150} = \frac{15620}{13210} = 0{,}182.$$

## Приклад

Визначимо територіальні індекси урожайності зернових з обємними вагами:

а) для району «А» порівняно з районом «В»:

$$I_y = \frac{\sum У_A (П_A + П_B)}{\sum У_B (П_A + П_B)} = \frac{36 \cdot (190 + 210) + 19 \cdot (80 + 100) + 25 \cdot (110 + 150)}{42 \cdot (190 + 210) + 23 \cdot (80 + 100) + 30 \cdot (110 + 150)} = 0{,}846;$$

б) для району «В» порівняно з районом «А»:

$$I_y = \frac{\sum У_B (П_A + П_B)}{\sum У_A (П_A + П_B)} = \frac{28740}{24320} = 1{,}182.$$

Аналогічні результати отримаємо, використавши в якості ваг середні посівні площі для окремих культур зернових:

а) для району «А» порівняно з районом «В»:

$$I_y = \frac{\sum У_A \cdot \overline{П}_{(A+B)}}{\sum У_B \cdot \overline{П}_{(A+B)}} = \frac{36 \cdot 200{,}8 + 19 \cdot 90{,}9 + 25 \cdot 131{,}8}{42 \cdot 200{,}8 + 23 \cdot 90{,}9 + 30 \cdot 131{,}8} = \frac{12250{,}9}{14478{,}3} 0{,}846;$$

б) для району «В» порівняно з районом «А»:

$$I_y = \frac{\sum У_B \cdot \overline{П}_{(A+B)}}{\sum У_A \cdot \overline{П}_{(A+B)}} = \frac{14478{,}3}{12250{,}9} 1{,}182.$$

# Приклад

Розрахуємо територіальні індекси урожайності зернових із стандартними вагами:

а) для району «А» порівняно з районом «В»:

$$I_y = \frac{\sum У_A \cdot П_{ст.}}{\sum У_B \cdot П_{ст.}} = \frac{36 \cdot 50 + 19 \cdot 21 + 25 \cdot 29}{42 \cdot 50 + 23 \cdot 21 + 30 \cdot 29} = \frac{2924}{3453} = 0{,}847;$$

а) для району «В» порівняно з районом «А»:

$$I_y = \frac{\sum У_B \cdot П_{ст.}}{\sum У_A \cdot П_{ст.}} = \frac{3453}{2924} = 1{,}181.$$

Отже, урожайність зернових в районі «А» нижча ніж у районі «В» на 15,3 % (100,0 – 84,7), а в районі «В» порівняно з районом «А» вона вища в 1,181 рази, або на 18,1 %.

# Використання системи взаємозв'язаних індексів в аналізі чинників динаміки

Соціально-економічні явища і процеси взаємозв'язані між собою, що виражається у взаємозв'язку відповідних показників. Одна з форм взаємозв'язку між економічними показниками заключається в тому, що їх можна виразити як добуток кількох інших показників. Так, фонд заробітної плати може бути поданий у вигляді добутку заробітної плати одного працівника на загальне число працівників.

$$F = f \cdot T,$$

де $F$ – фонд заробітної плати;

$T$ – число працівників;

$f$ – заробітна плата одного працівника $\left(f = \frac{F}{T}\right)$.

Товарооборот у фактичних цінах можна виразити як добуто цін на кількість проданих товарів: $p \cdot q = pq$;

Показники-співмножники виступають тут як чинники, від величини яких залежить результативна ознака. При економічному аналізі динаміки потрібно виявити і оцінити роль кожного окремого чинника в змінні результативного показника.

Індекс результативного показника завжди виступає як добуток індекса якісного показника на індекс об'ємного показника.

## Приклад

Так, наприклад, загальний індекс товарообороту може бути виражений як добуток індекса цін на індекс фізичного обсягу продукції:

$$I_{pq} = \frac{\sum p_1 q_1}{\sum p_0 q_0} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \cdot \frac{\sum q_1 p_0}{\sum q_0 p_0} = I_p \cdot I_q.$$

Аналогічно можна виразити взаємозв'язок між індексами:

а) собівартості одиниці продукції, обсягом продукції і затратами на її виробництво:

$$I_{zq} = \frac{\sum z_1 q_1}{\sum z_0 q_0} = \frac{\sum z_1 q_1}{\sum z_0 q_1} \cdot \frac{\sum q_1 z_0}{\sum q_0 z_0} = I_z \cdot I_q;$$

б) агрегатним індексом продуктивності праці, затрат часу і фізичним обсягом продукції:

$$I_{\frac{1}{w}} \cdot I_T = I_q = \left( \frac{\sum q_1 p_0}{\sum T_1} : \frac{\sum q_0 p_0}{\sum T_0} \right) \cdot \frac{\sum T_1}{\sum T_0} = \frac{\sum q_1 p_0}{\sum q_0 p_0};$$

в) урожайності, посівної площі і валового збору:

$$I_{\frac{1}{y}} \cdot I_П = I_{yП} = \left( \frac{\sum y_1 П_1}{\sum П_1} : \frac{\sum y_0 П_0}{\sum П_0} \right) \cdot \frac{\sum П_1}{\sum П_0} = \frac{\sum y_1 П_1}{\sum y_0 П_0}, \text{ і т.д.}$$

## Приклад

Дані про реалізацію продуктових товарів на ринку в базисному і звітному періодах.

| Назва продуктів | Базисний період | | | Звітний період | | | Вартість проданого товару у звітному році за цінами базисного року, тис.грн. ($p_0q_1$) |
|---|---|---|---|---|---|---|---|
| | ціна одиниці продукції, грн. ($p_0$) | кількість проданого товару, тис.од. ($q_0$) | вартість проданого товару, тис.грн. ($p_0q_0$) | ціна одиниці продукції, грн. ($p_1$) | кількість проданого товару, тис.од. ($q_1$) | вартість проданого товару, тис.грн. ($p_1q_1$) | |
| Картопля, кг | 0,35 | 45 | 15,75 | 0,30 | 53 | 15,90 | 18,55 |
| Молоко, л | 0,45 | 13 | 5,85 | 0,40 | 20 | 8,00 | 9,00 |
| М'ясо, кг | 4,50 | 10 | 45,00 | 4,00 | 14 | 56,00 | 63,00 |
| Разом | x | x | 66,60 | x | x | 79,90 | 90,55 |

Визначимо індекс товарообороту у фактичних цінах:

$$I_{pq} = \frac{\sum p_1 q_1}{\sum p_0 q_0} = \frac{79,90}{66,60} = 1,2.$$

Різниця між чисельником і знаменником даного індекса дасть нам абсолютний приріст товарообороту в звітному році порівняно з базисним за рахунок зміни двох чинників – цін одиниці товару і обсягу проданих товарів кожного виду:

$$\Delta pq = \sum p_1 q_1 - \sum p_0 q_0 = 79,90 - 66,60 = 13,3 \, тис. грн.$$

## Приклад

Загальний абсолютний приріст товарообороту розкладемо за чинниками у відносних і абсолютних показниках за рахунок зміни:

а) цін: $I_p = \frac{\sum p_1 q_1}{\sum p_0 q_1} = \frac{79{,}90}{90{,}55} = 0{,}882;$

$\Delta pq(p) = \sum p_1 q_1 - \sum p_0 q_1 = 79{,}90 - 90{,}55 = -10{,}65\ тис грн.;$

б) фізичного обсягу продукції: $I_q = \frac{\sum q_1 p_0}{\sum q_0 p_0} = \frac{90{,}55}{66{,}60} = 1{,}360;$

$\Delta pq(q) = \sum q_1 p_0 - \sum q_0 p_0 = 90{,}55 - 66{,}60 = 23{,}95\ тис грн.$

Перевіримо через взаємозв'язок індексів проведені розрахунки:

$$I_{pq} = I_p \cdot I_q = 0{,}882 \cdot 1{,}360 = 1{,}200;$$

$$\Delta pq = \Delta pq(p) + \Delta pq(q) = -10{,}65 + 23{,}95 = 13{,}3\ тис.грн.$$

Таким чином, в цілому по всіх видах товарів загальний обсяг товарообороту за рахунок зниження цін на 11,8 % (100,0 – 88,2) зменшився на 10,65 тис. грн., а за рахунок збільшення обсягу реалізації віх видів товарів в 1,36 рази, або на 36 % (136,0 – 100,0) він зріс на 23,95 тис. грн.

# Фондовий індекс

Фондовий індекс –комплексний показник на основі цін певної групи цінних паперів - «індексного кошика».
При розрахунку індексу його базове значення - сума цін або довільне число (100 або 1000). Для забезпечення порівнянності ціни множать на додаткові коефіцієнти. *Тому абсолютні значення індексів не важливі*.
Значення має динаміка зміни індексу, що дозволяє судити про загальний напрям руху цін в індексному кошику, незважаючи на те, що ціни акцій всередині «індексного кошика» можуть змінюватися різноспрямовано.
Залежно від принципу покладеного в основу вибору цінних паперів для індексу, він може відображати цінову динаміку групи цінних паперів, об'єднаних за якоюсь ознакою (наприклад висока, середня, мала капіталізація акцій) обраного сектора ринку (наприклад, телекомунікації), або широкого ринку акцій в цілому.

Фондові індекси є основою фінансових інструментів (*індексних ф'ючерсів або опціонів*), які використовуються для інвестиційних і спекулятивних цілей або для хеджування ризиків.

*Індекси акцій необхідні для:*

- вивчення факторів, які визначають курси акцій,
- для складання біржових прогнозів;
- прийняття індивідуальних рішень для інвестицій та вибору портфеля інвестицій;
- порівняльного аналізу рентабельності різних форм інвестицій;
- оцінки економічного стану

# Фондовий індекс

Зміни у величині акціонерного капіталу зумовлюють потребу в періодичному оцінюванні.

Індекси акцій розраховують щодня.

Зважування здійснюється за допомогою величини капіталу у вигляді акцій компаній. Коливання в сукупності (злиття, ліквідація та переміщення економічної діяльності) утруднюють використання незмінної схеми зважування.

Індекси акцій відрізняються від індексу цін на величину коефіцієнта поправки:

$$I_a = \frac{\sum p_1 \, q_1}{(\sum p_0 \, q_1) \cdot K_t}$$

$$K_t = \prod_{i=1}^{i=t} A_i$$

де q — номінальний капітал; p — курс.

Ai — компенсаційний фактор, розраховується щодня у формі ланцюгового індексу, забезпечує можливість порівняння індексу з вартісним рівнем портфеля акцій попереднього дня .

**Загальний індекс цін** визначається за формулою:

$$I_p = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1};$$

де *Ip* – загальний індекс цін;

# Промисловий індекс Доу-Джонса

Промисловий індекс Доу-Джонса (Dow Jones Industrial, Dow 30, Dow Jones, The Dow) біржовий індекс цінних паперів (акцій) 30 найбільших американських підприємств. Створений у 1896 році редактором The Wall Street Journal Чарльзом Доу.

Індекс створений для узагальнення інформації про акції індустріальних підприємств.

Залишається найстарішим індексом цінних паперів у США.

В основі індексу Доу-Джонса — теорія Доу про аналіз ринку за допомогою середніх показників котирувань промислових і транспортних акцій.

*Тенденція ринку позитивна, якщо один з середніх показників піднімається вище попереднього локального піку, за яким настає аналогічне зростання іншого показника.*

*Коли ж обидва показники знижуються нижче попереднього локального мінімуму, це означає загальну тенденцію спаду.*

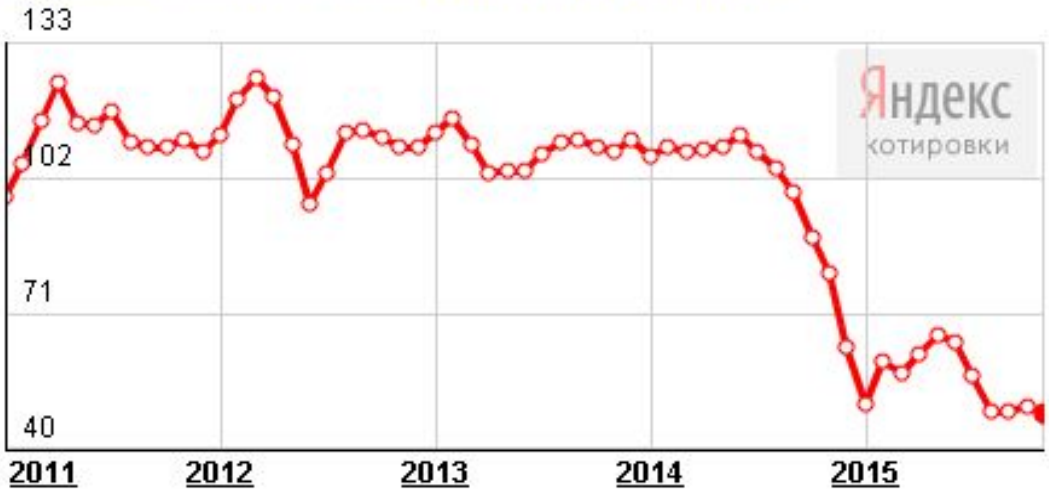
Теорію покладено в основу прогнозування

Індекс Доу-Джонса (промисловий) = (Сумарна ціна акцій 30 корпорацій) / K

K — коригуючий коефіцієнт, який змінюється зі зміною списку корпорацій і в разі кількості акцій.

# Промисловий індекс Доу-Джонса


# Промисловий індекс Доу-Джонса




# Індекс ПФТС

Індекс ПФТС —розраховується щодня за результатами торгів ПФТС на основі середньозваженої ціни за угодами.
У «індексний кошик» входять найліквідніші акції, за якими відбувається найбільше число угод.
Дата 1 жовтня 1997 року є базовим періодом, з якого починається розрахунок індексу.
Суть індексу — відсоток росту середньозважених цін акцій «індексного кошика» відносно базового періоду.
Для розрахунку беруть лише ті акції, що є у вільному обігу на фондовому ринку. Не враховуються акції, що перебувають у власності держави, емітента, стратегічних інвесторів, менеджменту і трудового колективу, а також у перехресному володінні.
Така методика розрахунку підвищує вплив на індекс цінних паперів підприємств, приватизація яких завершена.


Алчевський металургійний комбінат
Авдіївський коксохімічний завод
Азовсталь
Дніпроенерго
Донбасенерго
Західенерго
Єнакієвський металургійний завод
Крюківський вагонобудівний завод
Мотор Січ
Полтавський гірничо-збагачувальний комбінат
Райффайзен банк Аваль
Стахановський вагонобудівний завод
Стирол
Укрнафта
Укрсоцбанк
Укртелеком
Харцизький трубний завод
Центренерго
Південний ГЗК
Ясинівський коксохімічний завод

Лекція 8

# Вибіркове спостереження

1. Поняття про вибіркове спостереження та його основні завдання.
2. Основні умови наукової організації вибіркового спостереження.
3. Методи і способи відбору одиниць у вибіркову сукупність.
4. Знаходження середньої і граничної помилок та необхідної чисельності для різних видів вибірок.
5. Способи поширення даних вибіркового спостереження на генеральну сукупність. .

# Поняття про вибіркове спостереження

Із всіх видів несуцільного спостереження в практиці статистичних досліджень найбільше визначення і застосування отримало вибіркове спостереження

**Вибірковим спостереженням** називається такий вид несуцільного спостереження, за характеристикою відібраної частини одиниць якого судять про всю сукупність.

Розрізняють генеральну і вибіркову сукупності:

**Генеральною сукупністю** називається така маса одиниць, з якої проводиться відбір для дослідження.

**Вибірковою сукупністю** називається частина генеральної сукупності відібрана для обстеження.

Узагальнюючими показниками генеральної сукупності є: середній розмір ознаки «$\bar{x}$», частка «$p$», генеральна дисперсія «$\sigma^2$».

Узагальнюючими показниками вибіркової сукупності є: середня вибіркова «$\tilde{x}$», вибіркова частка «$w$», дисперсія «$\sigma_в^2$».

# Поняття про вибіркове спостереження

До вибіркового спостереження статистика вдається у випадках, коли потрібно зекономити сили і засоби при проведенні дослідження, тобто, коли недоцільно або неможливо проводити суцільне спостереження.

Вибіркове спостереження застосовують також у поєднанні із суцільним для поглиблення дослідження, або уточнення і контролю результатів суцільного спостереження.

Вибіркове спостереження складається з таких етапів:

1) постановка мети спостереження;

2) складання програми спостереження і розробка відповідних даних;

3) вирішення організаційних питань проведення спостереження;

4) визначення відсотка і способу відбору одиниць;

5) проведення відбору;

6) реєстрація відповідних ознак у відібраних для дослідження одиниць;

7) узагальнення даних спостереження та розрахунок їх вибіркових характеристик;

8) знаходження помилок вибірки;

# Основні завдання вибіркового спостереження

Вибіркове спостереження проводиться для вирішення наступних основних завдань:

1) визначення середнього розміру досліджуваної ознаки;

2) визначення питомої ваги (частки);

3) визначення середньої і граничної помилки вибірки;

4) знаходження меж для середньої і частки при повторному і безповторному відборі;

5) визначення потрібної чисельності вибірки;

6) поширення даних вибіркового спостереження на всю сукупність.

# Основні методи формування вибірки

При формуванні вибірки необхідно визначити:
– хто (що) є елементом або *одиницею вибірки* виходячи від сутності дослідження;
– *контур вибірки* - список усіх одиниць генеральної сукупності, з якої формується вибірка;
– *об'єм вибір*ки – кількість елементів у ній.

Приклад,

Фірма – виробник мобільних телефонів бажає вивчити потенційний ринок продукції

*Одиницями вибірки* будуть особи, які приймають рішення про купівлю.

*Контуром вибірки* можуть бути списки:

- Громадян селектовані за віком, родом занять , місцем проживання
- Організацій (корпоративний сектор)

Оскільки вибірка є лише частиною генеральної сукупності, то отримані на основі її вивчення результати не будуть точно відповідати результатам, які можна було б отримати при вивченні всієї генеральної сукупності.

Різниця між результатами дослідження вибірки та генеральної сукупності називається помилкою вибірки.

Помилки вибірки обумовлюються як методами її формування, так і її об' ємом.

# Основні умови наукової організації вибіркового спостереження

Особливістю вибіркового спостереження в порівнянні з іншими видами несуцільного спостереження є те, що при відборі одиниць у вибіркову сукупність забезпечується рівна можливість попадання кожної одиниці у вибірку. Досягається шляхом неупередженого строгого випадкового відбору за схемою, узгодженою з математичною статистикою.

Відповідь на питання про те, яка за розміром різниця між генеральними і вибірковими узагальнюючими показниками, з якою ймовірністю можна судити про цю різницю, дає теорія вибіркового методу, на основі закону великих чисел.

*Розв'язують два завдання:*

1) Розрахунок із заданою ймовірністю межі можливих відхилень вибіркового від відповідного показника в генеральній сукупності;
2) Визначення ймовірності того, що розмір можливих відхилень вибіркового показника від генерального не перевищить встановленої межі.

Закон великих чисел в теорії імовірностей стверджує, що *емпіричне середнє (арифметичне середнє) скінченної вибірки із фіксованого розподілу близьке до теоретичного середнього (математичного сподівання) цього розподілу*.

В залежності від виду збіжності розрізняють слабкий закон великих чисел, коли має місце збіжність за ймовірністю, і посилений закон великих чисел, коли має місце збіжність майже скрізь.

*Завжди знайдеться така кількість випробувань, при якій з будь-якою заданою наперед імовірністю частота появи деякої події буде як завгодно мало відрізнятися від її імовірності.*

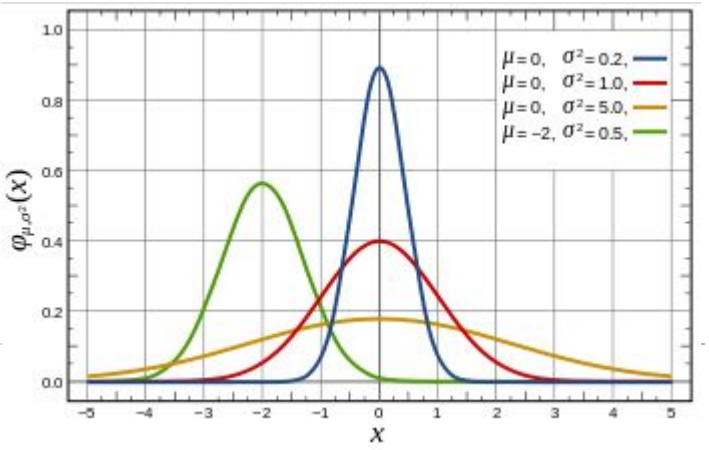
# Основні умови наукової організації вибіркового спостереження

При масовому спостереженні, розподіл емпіричних частот більшості явищ підпорядковується закону нормального розподілу.
за нормальним розподілом більша частина величин зосереджена навколо генерального середнього.
68,3 % чисельності вибірки буде знаходитись в межах ± σ генеральної середньої;
95,4 % чисельності вибірки знаходиться в межах ± 2σ
99,7 % – не вийде за межі ± 3σ.

Принцип строгої випадковості, покладений в основу вибірки, забезпечує об'єктивність, дозволяє встановити межі можливих помилок і отримати достовірні дані для характеристики всієї сукупності явищ.
В цьому випадку вибіркова сукупність називається представницькою (*репрезентативною*).
В її склад входять представники всіх груп, з яких складається генеральна сукупність.
Точність результатів вибіркового спостереження залежить від

- способу відбору одиниць,
- ступеня коливання ознаки в сукупності
- числа одиниць, що їх спостерігатимуть.

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



# Методи і способи відбору одиниць у вибіркову сукупність

Способом відбору називається система організації відбору одиниць з генеральної сукупності.

Розрізняють два методи відбору одиниць у вибіркову сукупність: повторний і безповторний.

Повторним називається такий метод відбору, при якому кожна раніше відібрана одиниця повертається в генеральну сукупність і може знову брати участь у вибірці.

Безповторним називається такий метод відбору, при якому кожна раніше відібрана одиниця не повертається в генеральну сукупність і в подальшій вибірці участі не бере.
Безповторний відбір охоплює постійно нові одиниці сукупності, повторний – одну і ту ж сукупність.

Безповторний відбір дає більш точні результати.

Повторний і безповторний методи відбору, в залежності від характеру одиниці відбору, застосовується в поєднанні з іншими видами відбору.

# Методи і способи відбору одиниць у вибіркову сукупність

На практиці статистичного дослідження використовуються три види відбору:
1) індивідуальний – відбір окремих одиниць сукупності;
2) груповий (серійний) – відбір груп (серій) одиниць;
3) комбінований – комбінація індивідуального і групового.

За способом відбору одиниць для обстеження розрізняють види спостереження:
1) випадкова вибірка;
2) механічна вибірка;
3) типова (районована) вибірка;
4) серійна (гніздова) вибірка;
5) комбінована вибірка;
6) одноступінчаста і багатоступінчаста вибірка;
7) однофазна і багатофазна вибірка;
8) інші види вибірки.

Випадковою називається вибірка, при якій відбір одиниць з генеральної сукупності є випадковим (застосовують жеребкування або таблицю випадкових чисел).
Механічна вибірка – послідовний відбір одиниць через рівні проміжки в порядку їх положення в генеральній сукупності, або в переліку. Інтервали відбору визначаються у відповідності з часткою відбору одиниць (кожна п’ята, десята, сота).

# Методи і способи відбору одиниць у вибіркову сукупність

При типовому відборі генеральну сукупність поділяють на однорідні групи за певною ознакою, райони, зони. З кожної групи випадковим або механічним способом відбирають певну кількість одиниць, пропорційно частці групи в загальній сукупності.

При серійній (кластерній) вибірці відбір одиниць проводять цілими групами (серіями, кластерами) в межах яких обстежують всі одиниці без винятку. Серії для спостереження відбирають випадково, частіше безповторним способом механічної вибірки.

Комбінованою називається вибірка, коли комбінують два або кілька видів вибірок.
Перш за все, комбінують суцільне і вибіркове спостереження. В даному випадку, за основною програмою обстежується генеральна сукупність, а за додатковою – вибіркова.

Одноступінчастою називається вибірка, у випадку коли із сукупності відразу відбираються одиниці або серії одиниць для безпосереднього обстеження.

Багатоступінчаста вибірка передбачає поступове вилучення із генеральної сукупності спочатку укрупнення груп одиниць, потім груп менших за обсягом, і так до тих пір, поки не відберуть відповідні групи або одиниці, які будуть досліджуватись.
Вибірка може бути двох-, трьох і більше ступінчастою.

# Методи і способи відбору одиниць у вибіркову сукупність

Якщо необхідні дані можна отримати на основі вивчення всіх первинно відібраних одиниць, застосовують однофазну вибірку, а якщо тільки на основі деякої її частини, відібраної так, що вона складає підвибірку із початково проведеної вибірки – багатофазну.

Багатофазною називається вибірка, для якої відомості збираються від всіх одиниць відбору, потім відбираються ще деякі одиниці і обстежуються за більш широкою програмою. При багатофазній вибірці на кожній фазі зберігається одна і таж одиниця відбору.
Розрахунок помилок репрезентативності багатоступінчастої і багатофазної вибірок проводиться для кожної ступені і фази окремо.

---

Взаємопроникаючою називається така вибірка, коли із однієї генеральної сукупності проводять одним і тим же способом декілька незалежних вибірок.
Взаємопроникаючі вибірки завжди проводять різні, незалежні один від одного дослідники, що дозволяє порівнювати підсумки по всіх частинах і забезпечити взаємну перевірку їх роботи.
Взаємопроникаючі вибірки дають незалежні одна від одної оцінки значень досліджуваної сукупності, і, якщо результати різних вибірок близькі між собою, то такі оцінки дуже переконливі.

# Методи і способи відбору одиниць у вибіркову сукупність

Направлений відбір використовують тоді, коли за відомим середнім значенням ознаки в генеральній сукупності вибіркова сукупність повинна характеризувати її структуру за іншими ознаками.

Направлений відбір передбачає проведення відбору таким чином, щоб *середній розмір (характеристика) відібраних одиниць дорівнював середньому розміру одиниць всієї сукупності*.

В тому випадку, коли заміна однієї одиниці іншою призводить до наближеної рівності середніх генеральної і вибіркової сукупностей, вибірку вважають врівноваженою і репрезентативною за всіма іншими ознаками сукупності.

*Направленим відбором називається врівноваження за однією ознакою для вибіркового дослідження інших ознак*.

Помилку вибірки направленого відбору визначають в залежності від способу проведення відбору одиниць до врівноваження.

---

Малою вибіркою називається вибіркова сукупність, яка складається з порівняно невеликої кількості одиниць (*десятки*).

При малих вибірках характеристики вибіркової сукупності можна поширити на генеральну сукупність.

# Помилки репрезентативності

Помилки репрезентативності становлять різницю між середніми і відносними показниками вибіркової сукупності та відповідними показниками генеральної сукупності.

Помилки репрезентативності поділяються на систематичні та випадкові.

Систематичні помилки репрезентативності зумовлені внаслідок порушення принципів проведення вибіркового спостереження.

Випадкові помилки репрезентативності зумовлені тим, що вибіркова сукупність не відображає точно середні і відносні показники генеральної сукупності.

# Знаходження середньої помилки для різних видів вибірок

Формули для визначення середньої помилки репрезентативності випадкової і механічної вибірки для повторного і безповторного відбору.

| Спосіб відбору | При визначенні середньої | При визначенні частки |
|---|---|---|
| Повторний | $\mu = \sqrt{\frac{\sigma^2}{n}}$ | $\mu = \sqrt{\frac{w(1-w)}{n}}$ |
| Безповторний | $\mu = \sqrt{\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)}$ | $\mu = \sqrt{\frac{w(1-w)}{n}\left(1-\frac{n}{N}\right)}$ |

$\mu$ – середня помилка репрезентативності;

$\sigma^2$ – середній квадрат відхилень у вибірці;

$n$ – чисельність вибіркової сукупності;

$N$ – чисельність генеральної сукупності;

$\left(1-\frac{n}{N}\right)$ – необстежена частка генеральної сукупності;

$\frac{n}{N}$ – частка обстеженої частини вибіркової сукупності;

$w$ – частка даної ознаки у вибірці;

$(1-w)$ – частка протилежної ознаки у вибірці.

# Знаходження граничної помилки для різних видів вибірок

При вибірковому спостереженні розмір граничної помилки репрезентативності «Δ» може бути *більший* (*дорівнювати* або *менший*) від середньої помилки репрезентативності «μ». Тому величину граничної помилки репрезентативності обчислюють з певною ймовірністю «р», якій відповідає t-разове значення «μ».

З введенням показника кратності помилки «t» формула граничної помилки репрезентативності :

$$\Delta = t\mu; \qquad t = \frac{\Delta}{\mu},$$

μ – середня помилка вибірки;
t – довірчий коефіцієнт

Ймовірність відхилень вибіркового середнього від генерального середнього при достатньо великому обсязі вибірки і обмеженій дисперсії генеральної сукупності підпорядковується закону нормального розподілу.

Ймовірність цих відхилень при різних значеннях «t» визначається за формулою:

$$F(t) = \frac{2}{\sqrt{2\pi}} \int_0^t e^{-\frac{t^2}{2}} dt.$$

для $t = 1$ $p(\Delta \le \mu) = 0{,}683$; для $t = 2$ $p(\Delta \le \mu) = 0{,}954$;

для $t = 3$ $p(\Delta \le \mu) = 0{,}997$; для $t = 4$ $p(\Delta \le \mu) = 0{,}999$.

# Знаходження граничної помилки для різних видів вибірок

Із теореми Чебишева знаходять, що:

$$\bar{x} - \tilde{x} = \pm\Delta_x \quad і \quad \tilde{x} - \Delta_x \le \bar{x} \le \tilde{x} + \Delta_x.$$

Додаючи граничну помилку вибірки до вибіркової частки і віднімаючи її від неї, знаходять межі генеральної частки:

$$p - w = \pm\Delta_p \quad і \quad w - \Delta_p \le p \le w + \Delta_p.$$

На основі формул граничної помилки вибірки розв´язують завдання:

1. визначають довірчі межі генеральної середньої і частки з прийнятою ймовірністю;
2. визначають ймовірність того, що відхилення між вибірковими і генеральними характеристиками не перевищать визначену величину;
3. визначають необхідну чисельність вибірки, яка із заданою ймовірністю забезпечить прийняту точність вибіркових показників.

$w$ – частка даної ознаки у вибірці;

$(1 - w)$ – частка протилежної ознаки у вибірці.

Теоерма Чебишева

Для будь-якого натурального n ≥ 2 знайдеться просте число p в інтервалі n <p <2n.

2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, 61, 67, 71, 73, 79, 83, 89, 97, 101, 103, 107, 109, 113 , 127, 131, 137, 139

## Приклад

При 2% випадковому відборі у відібраних для обстеження 100 деталей встановлено, що середня вага однієї деталі 2500 г; дисперсія 900, із 100 деталей 10 виявились бракованими.
З ймовірністю 0,954 встановити межі середньої ваги однієї деталі в генеральній сукупності, а з ймовірністю 0,997 – межі частки якісних деталей в генеральній сукупності.

Граничну помилку визначаємо $\Delta_x = t\mu$ за формулою безповторного відбору, так як чисельність генеральної сукупності можна знайти:

$$N = \frac{100 \cdot 100}{2} = 5000 \text{ шт.}$$

В спеціальній таблиці знаходимо, що для ймовірності 0,954 $t = 2$, а для ймовірності 0,997 $t = 3$. Таким чином отримаємо:

$$\Delta_x = t\sqrt{\frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right)} = 2\sqrt{\frac{900}{100}\left(1 - \frac{100}{5000}\right)} = 2\sqrt{9 \cdot 0{,}98} = 2 \cdot 3 = 6\,г.$$

Звідси довірчі межі генеральної середньої будуть наступні:

$$\tilde{x} - \Delta_x \le \bar{x} \le \tilde{x} + \Delta_x;$$
$$2500 - 6 \le \bar{x} \le 2500 + 6;$$
$$2494 \le \bar{x} \le 2506.$$

Тобто, з ймовірністю 0,954 можна стверджувати, що середня вага однієї деталі в генеральній сукупності знаходиться в межах від 2494 г. до 2506 г.

Приклад

Вибіркова частка якісних деталей:

$$w = \frac{m}{n} = \frac{90}{100} = 0{,}9,$$

де $m$ – кількість якісних деталей у вибірковій сукупності;

$n$ – кількість відібраних деталей.

$$\Delta_p = t\sqrt{\frac{w-(1-w)}{n}\left(1-\frac{n}{N}\right)} = 3\sqrt{\frac{0{,}9(1-0{,}9)}{100}\left(1-\frac{100}{5000}\right)} = 3\cdot 0{,}03 = 0{,}09,$$

$$w - \Delta_p \le p \le w + \Delta_p,$$

$$0{,}9 - 0{,}09 \le p \le 0{,}9 + 0{,}09,$$

$$0{,}81 \le p \le 0{,}99.$$

Таким чином, з ймовірністю 0,997 можна стверджувати, що частка якісних деталей в генеральній сукупності знаходиться в межах від 81% до 99%.

$$t = \frac{\Delta_x}{\mu} = \frac{6}{3} = 2.$$

Для даного значення ($t = 2$) відповідає ймовірність ($p = 0{,}954$). Це дає право стверджувати, що при визначенні за вибірковими даними середньої ваги деталей ($\tilde{x} = 2500\,г$) допущена помилка, яка не перевищує 6 г.

ймовірність допуску помилки для частки:

$$t = \frac{\Delta_p}{\mu} = \frac{0{,}09}{0{,}03} = 3.$$

Для цього значення ($t = 3$) відповідає ймовірність ($p = 0{,}997$). Таким чином, майже достовірно можна стверджувати, що при визначенні за вибірковими даними частки якісних деталей ($w = 0{,}9$) допущена помилка, яка не перевищує 9 %.

# Чисельність вибірки

Чисельність вибірки залежить від наступних чинників:

1) від варіації досліджуваної ознаки - більша варіація - більшою повинна бути вибірка і навпаки;
2) від розміру можливої граничної помилки вибірки –необхідний менший розмір помилки - більшою повинна бути чисельність вибірки, якщо помилку потрібно зменшити в три рази, то чисельність вибірки збільшують в дев'ять раз;
3) від розміру ймовірності, з якою гарантуватимуть результати вибірки – більша ймовірність- більша повинна бути чисельність вибірки;
4) від способу відбору одиниць у вибіркову сукупність для обстеження.

Формули для знаходження необхідної чисельності вибірки для випадкової і механічної вибірки.

| Способи відбору | Чисельність вибірки | |
|---|---|---|
| | при визначенні середньої | при визначенні частки |
| Повторний | $n = \frac{t^2\sigma^2}{\Delta_x^2}$ | $n = \frac{t^2 w(1-w)}{\Delta_p^2}$ |
| Безповторний | $n = \frac{t^2\sigma^2 N}{\Delta_x^2 N + t^2\sigma^2}$ | $n = \frac{t^2 w(1-w)N}{\Delta_p^2 N + t^2 w(1-w)}$ |

## Приклад

Для району, в якому є 8000 підприємців платників ПДВ , необхідно організувати вибіркове спостереження з метою встановлення усереднених характеристик обороту коштів. Якою повинна бути чисельність вибірки?

При граничній помилці в 30 тис. грн. з ймовірністю ( p=954 ,0 ) і при середньому квадратичному відхиленні 300, визначеному за результатами аналогічних обстежень, необхідна чисельність вибірки повинна бути:

Повторний відбір

$$n = \frac{t^2\sigma^2}{\Delta_x^2} = \frac{2^2 \cdot 300^2}{30^2} = \frac{4 \cdot 90000}{90} = 400$$

Безповторний відбір

$$n = \frac{t^2\sigma^2 N}{\Delta_x^2 N + t^2\sigma^2} = \frac{4 \cdot 90000 \cdot 8000}{900 \cdot 8000 + 4 \cdot 90000} = 380$$

$\mu$ – середня помилка репрезентативності;

$\sigma^2$ – середній квадрат відхилень у вибірці;

$n$ – чисельність вибіркової сукупності;

$N$ – чисельність генеральної сукупності;

$\left(1 - \frac{n}{N}\right)$ – необстежена частка генеральної сукупності;

$\frac{n}{N}$ – частка обстеженої частини вибіркової сукупності;

$w$ – частка даної ознаки у вибірці;

$(1 - w)$ – частка протилежної ознаки у вибірці.

при інших рівних умовах, обсяг вибірки при безповторному відборі завжди буде менший, ніж при повторному.

# Способи поширення даних вибіркового спостереження на генеральну сукупність

Кінцевою практичною метою вибіркового спостереження є поширення його характеристик на генеральну сукупність.

Два способи розповсюдження даних вибіркового спостереження:

1) спосіб прямого перерахунку;
2) спосіб коефіцієнтів.

Спосіб прямого перерахунку застосовують якщо на основі вибірки розраховують об'ємні показники генеральної сукупності, використовуючи для цього вибіркові середню або частку. В першому випадку середній розмір ознаки, визначений в результаті вибіркового спостереження, множиться на кількість одиниць генеральної сукупності.

Спосіб поправочних коефіцієнтів застосовується коли вибіркове спостереження проводиться з метою перевірки і уточнення результатів суцільного спостереження.
В даному випадку, співставляючи дані вибіркового спостереження із суцільним, вираховують поправочний коефіцієнт, який використовують для внесення поправок в матеріали суцільних спостережень.

# Експертне оцінювання

1. Обробка результатів експертного оцінювання
2. Коефіцієнт конкордації
3. Коефіцієнт компетенції .

# Обробка результатів експертного оцінювання

Важливим етапом у підведенні результатів дослідження є прогнозування, яке передбачає визначення значень економічних показників у майбутньому.

*Прогнозування* - початковий етап планування, включає попередній і кінцевий (формальний) прогнози для яких розробляється один або декілька сценаріїв майбутніх подій.

Методи експертних оцінок використовуються для прогнозування майбутніх подій, якщо статистичних даних недостатньо.
Експерт формує своє судження на аналізі групи факторів, оцінюючи ймовірності їх реалізації та впливу на результативну ознаку об'єкта вивчення.
При цьому отримані висновки та оцінки пов'язані з особистістю експерта, тому інший експерт, використовуючи ту саму інформацію, може дійти інших висновків.

При розв'язанні проблем в умовах невизначеності думка групи експертів дає більш надійні результати, ніж думка одного експерта.

Після отримання експертних оцінок проводиться їх обробка та оцінюється достовірність. Обробку результатів експертного оцінювання проводять за *коефіцієнтом конкордації*, який показує степінь згоди думок експертів.

# Коефіцієнт конкордації

Коефіцієнт конкордації $W$ розраховується за формулою:

$$W = \frac{12}{m^2\left(n^3 - n\right)} \sum_{j=1}^{n} \left( \sum_{i=1}^{m} \left( R_{ij} - \frac{n+1}{2} \right) \right)^2,$$

де $n$ – кількість об'єктів оцінювання;
$m$ – кількість експертів;
$R_{ij}$ – ранг $j$-го об'єкта, представленого $i$-м експертом.

Якщо об'єкти оцінювання мають однакові ранги, то коефіцієнт конкордації розраховується за формулою:

$$W = \frac{12}{\frac{1}{12} m^2\left(n^3 - n\right) - m\sum_{j=1}^{m} T_j} \cdot \sum_{j=1}^{n} \left( \sum_{i=1}^{m} \left( R_{ij} - \frac{n+1}{2} \right) \right)^2.$$

$$T_j = \frac{\sum_{i=1}^{L_i} \left( n_i^2 - n_i \right)}{12},$$

де $L_i$ – кількість груп однакових рангів,
$n_i$ – кількість елементів $i$-тої групи для $j$-го експерта.

# Коефіцієнт конкордації

Статистична значущість коефіцієнта конкордації перевіряється порівнянням величини $n(m-1)\cdot W$ з табличним значенням розподілу $\chi^2$ при рівні значущості $\alpha = 0,001$ та $n-1$ степенях свободи.

Якщо коефіцієнт конкордації виявляється не значущим, то використовується методика виведення експерта, думка якого не узгоджується з думкою інших експертів. Для цього будується матриця коефіцієнтів кореляції Пірсона ($r(k,i)$) або рангових коефіцієнтів кореляції Спірмена ($r_S(k,i)$) та виявляється експерт, оцінка якого підкоряється умові:

$$r_j(k,i) = \min_{i=1,\dots,m}\{r(k,i)\},$$

що означає, що думка цього експерта найменше узгоджується з думкою інших експертів. Бали, подані таким експертом, у подальших розрахунках не враховуються.

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

## Приклад

Група експертів з 3 осіб оцінювала час, що необхідний для виконання робіт певного проекту. Результати оцінювання подано у табл. Перевірити степінь узгодженості думок експертів.

| Експерти | Час, необхідний для робіт | | | |
|---|---|---|---|---|
| | Робота 1 | Робота 2 | Робота 3 | Робота 4 |
| 1-й | 6 | 5 | 2 | 4 |
| 2-й | 4 | 7 | 3 | 9 |
| 3-й | 5 | 7 | 3 | 6 |

Здійснимо перевірку за коефіцієнтом конкордації, для чого знайдемо ранги робіт проекту окремо за оцінками кожного з експертів

| Експерти | Ранги робіт | | | |
|---|---|---|---|---|
| | Робота 1 | Робота 2 | Робота 3 | Робота 4 |
| 1-й | 4 | 3 | 1 | 2 |
| 2-й | 2 | 3 | 1 | 4 |
| 3-й | 2 | 4 | 1 | 3 |

У групах рангів оцінок, наданих окремими експертами, немає однакових, тому коефіцієнт конкордації розраховуємо за формулою

Приклад

$$W = \frac{12}{m^2\left(n^3 - n\right)} \sum_{j=1}^{n} \left( \sum_{i=1}^{m} \left( R_{ij} - \frac{n+1}{2} \right) \right)^2,$$

де $n$ – кількість об'єктів оцінювання;

$m$ – кількість експертів;

$R_{ij}$ – ранг $j$-го об'єкта, представленого $i$-м експертом.

Обчислимо величини, що не залежать від індексів сум, враховуючи, що:

$n$ – кількість робіт, $n = 4$; $m$ – кількість експертів, $m = 3$.

Отримаємо:

$$\frac{n+1}{2} = \frac{4+1}{2} = 2{,}5; \quad \frac{12}{m^2\left(n^3 - n\right)} = \frac{12}{3^2\left(4^3 - 4\right)} = \frac{12}{540} \approx 0{,}022.$$

| Розрахункові формули | Результати розрахунків | | | |
|---|---|---|---|---|
| | Робота 1 | Робота 2 | Робота 3 | Робота 4 |
| $R_{ij} - \frac{n+1}{2}$ | 1,5 | 0,5 | −1,5 | −0,5 |
| | −0,5 | 0,5 | −1,5 | 1,5 |
| | −0,5 | 1,5 | −1,5 | 0,5 |
| $\sum_{i=1}^{m} \left( R_{ij} - \frac{n+1}{2} \right)$ | 0,5 | 2,5 | −4,5 | 1,5 |
| $\left( \sum_{i=1}^{m} \left( R_{ij} - \frac{n+1}{2} \right) \right)^2$ | 0,25 | 6,25 | 20,25 | 2,25 |
| $\sum_{j=1}^{n} \left( \sum_{i=1}^{m} \left( R_{ij} - \frac{n+1}{2} \right) \right)^2$ | 29 | | | |

## Приклад

Отже, коефіцієнт конкордації:

$$W = \frac{12}{m^2\left(n^3 - n\right)} \sum_{j=1}^{n} \left( \sum_{i=1}^{m} \left( R_{ij} - \frac{n+1}{2} \right) \right)^2 = 0{,}022 \cdot 29 = 0{,}638$$

Перевіримо його значущість: $n(m-1) \cdot W = 4(3-1) \cdot 0{,}638 = 5{,}104$;

критичне значення $\chi^2$: ХИ2ОБР (0,001; 4 − 1) = 16,27. Оскільки величина $n(m-1) \cdot W$ менша критичного значення $\chi^2$, то коефіцієнт конкордації не є значущим та думки експертів не узгоджені.

Виокремимо експерта, оцінки якого є найбільш неузгодженими. Для цього побудуємо матрицю парних коефіцієнтів кореляції Пірсона

| Експерти | 1-й | 2-й | 3-й |
|---|---|---|---|
| 1-й | 1 | | |
| 2-й | **0,23035** | 1 | |
| 3-й | 0,657143 | 0,797366 | 1 |

Найменшим є значення коефіцієнта кореляції, який показує узгодженість думок першого та другого експертів, тому одного з них необхідно вивести з експертизи.

Доцільно вивести першого експерта, тому що його оцінки є менш узгодженими з оцінками третього експерта.

Розрахуємо коефіцієнт конкордації, враховуючи відсутність оцінок першого експерта.

Приклад

$$\frac{n+1}{2}=\frac{4+1}{2}=2,5\ ;\quad \frac{12}{m^2\left(n^3-n\right)}=\frac{12}{2^2\left(4^3-4\right)}=\frac{12}{240}=0,05$$

| Розрахункові формули | Результати розрахунків | | | |
|---|---|---|---|---|
| | Робота 1 | Робота 2 | Робота 3 | Робота 4 |
| $R_{ij}-\frac{n+1}{2}$ | −0,5 | 0,5 | −1,5 | 1,5 |
| | −0,5 | 1,5 | −1,5 | 0,5 |
| $\sum_{i=1}^{m}\left(R_{ij}-\frac{n+1}{2}\right)$ | −1 | 2 | −3 | 2 |
| $\left(\sum_{i=1}^{m}\left(R_{ij}-\frac{n+1}{2}\right)\right)^2$ | 1 | 4 | 9 | 4 |
| $\sum_{j=1}^{n}\left(\sum_{i=1}^{m}\left(R_{ij}-\frac{n+1}{2}\right)\right)^2$ | 18 | | | |

Коефіцієнт конкордації:

$$W=\frac{12}{m^2\left(n^3-n\right)}\sum_{j=1}^{n}\left(\sum_{i=1}^{m}\left(R_{ij}-\frac{n+1}{2}\right)\right)=0,05\cdot 18=0,9$$

Значення коефіцієнта конкордації після виведення першого експерта збільшилося. Але воно теж не є значущим -пов'язано не з тим, що думки експертів не узгоджуються, а з тим, що кількість експертів надто мала.

Час, необхідний для виконання робіт проекту, розраховується як середнє арифметичне експертних оцінок.

# Коефіцієнт компетенції

Використання коефіцієнта конкордації засновано на припущенні-
*чим більш узгоджені думки експертів, тим достовірнішими є їх оцінки.*
Але
практика показує, що це не завжди вірно, і експерт, який не згоден з думками більшості, може дати найточніші оцінки (*вотум сепаратум*).
Якщо необхідно врахувати думки всіх експертів, то обробку результатів експертного оцінювання слід виконувати за *коефіцієнтом компетентності* експерта.

Метод базується на використанні попередньої оцінки компетентності експертів, які приймають участь у дослідженні.
Оцінка експертів проводиться за критеріями компетентності, серед яких можуть бути:

1. рівень освіти;
2. загальний стаж роботи;
3. стаж роботи за проблемою дослідження;
4. посада.

Важливий критерій - оцінка надійності експерта- розраховується як відношення його правильних оцінок до всіх проведених експертиз.
*Правильними вважаються ті оцінки, які з часом підтвердилися практикою.*
При розрахунку коефіцієнтів компетентності експертів необхідно використовувати єдину для всіх критеріїв шкалу оцінювання. У противному випадку оцінки потрібно буде нормалізувати, тобто привести до однієї шкали.

# Коефіцієнт компетенції

Коефіцієнт компетентності розраховується за формулою:

$$KK_i = \frac{\sum_{j=1}^{m} k_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m} k_{ij}},$$

де $n$ – кількість експертів;

$m$ – кількість критеріїв оцінювання експертів;

$k_{ij}$ – бал, отриманий $i$-м експертом за $j$-м критерієм.

## Приклад

За вхідними даними знайти час, необхідний для виконання проекту, з урахуванням коефіцієнта компетентності експертів. Бали, отримані експертами, подано у табл.

Оцінювання проводилося за трибальною шкалою.

| Експерти | Час, необхідний для робіт | | | |
|---|---|---|---|---|
| | Робота 1 | Робота 2 | Робота 3 | Робота 4 |
| 1-й | 6 | 5 | 2 | 4 |
| 2-й | 4 | 7 | 3 | 9 |
| 3-й | 5 | 7 | 3 | 6 |

| Експерти | Бали, отримані експертами | | |
|---|---|---|---|
| | Критерій 1. Стаж роботи | Критерій 2. Професіоналіз | Критерій 3. Надійність |
| 1-й | 1 | 2 | 2 |
| 2-й | 2 | 3 | 3 |
| 3-й | 2 | 1 | 1 |

Знайдемо коефіцієнти компетентності експертів

| Експерти | Бали, отримані експертами | | | Сума балів кожного експерта |
|---|---|---|---|---|
| | Критерій 1 | Критерій 2 | Критерій 3 | |
| 1-й | 1 | 2 | 2 | 5 |
| 2-й | 2 | 3 | 3 | 8 |
| 3-й | 2 | 1 | 1 | 4 |
| Загальна сума балів | | | | 17 |

для першого експерта $KK_1 = \dfrac{\sum_{j=1}^{m} k_{1j}}{\sum_{i=1}^{n}\sum_{j=1}^{m} k_{ij}} = \dfrac{5}{17} \approx 0,2941;$

для другого і третього $KK_2 = \dfrac{8}{17} \approx 0,4706; \quad KK_3 = \dfrac{4}{17} = 0,2353.$

# Приклад

Розрахуємо час, необхідний для виконання робіт проекту, з урахуванням коефіцієнта компетентності експертів за формулою: $t_j = \sum_{i=1}^{n} KK_i \cdot t_{ij}$; $j = \overline{1,m}$; де $t_i$ – час для $i$-тої роботи; $t_{ij}$ – оцінка часу $i$-тої роботи $j$-м експертом. Результати

| Експерти | $KK_i$ | Час, необхідний для робіт | | | |
|---|---|---|---|---|---|
| | | Робота 1 | Робота 2 | Робота 3 | Робота 4 |
| 1-й | 0,2941 | 6 | 5 | 2 | 4 |
| 2-й | 0,4706 | 4 | 7 | 3 | 9 |
| 3-й | 0,2353 | 5 | 7 | 3 | 6 |
| Час з урахуванням $KK_i$ | | **4,82** | **6,41** | **2,71** | **6,82** |

# Статистичні функції