

26.5.2020

# Лекція №9

Функціональна, статистична та  
кореляційна залежності.

Воробйова А.І доцент, кандидат фіз-мат наук  
КАФЕДРА ІНТЕЛЕКТУАЛЬНИХ ІНФОРМАЦІЙНИХ СИСТЕМ  
СЕКЦІЯ ПРИКЛАДНОЇ ТА ВИЩОЇ МАТЕМАТИКИ

## Зміст

Тема лекції.....	2
План лекції.....	2
Питання для самоконтролю.....	2
Завдання для дистанційного опанування.....	2
I. Елементи теорії кореляції та регресії.....	3
Стохастичний зв'язок між величинами.....	3
II. Коефіцієнт кореляції. Позитивна та негативна кореляція. Сильний і слабкий кореляційний зв'язок. Діаграма розсіювання.....	3
Коефіцієнт кореляції.....	3
Кореляційна залежність.....	6
III. Поняття регресії. Загальна постановка задачі. Рівняння регресії.....	9
Задачі регресійного аналізу.....	9
Рівняння регресії.....	9
Лінійна кореляція.....	11
IV. Приклади «Зв'язок між величинами X та Y».....	12
<i>Приклад 1.</i> ....	12
<i>Приклад 2.</i> ....	13
<i>Приклад 3.</i> ....	16
V. Метод найменших квадратів.....	18
Додаткові джерела.....	21
Статистический анализ в MS Excel.....	21
Линейный коэффициент корреляции Пирсона.....	21
Коэффициент корреляции Пирсона в Excel.....	21
Результаты расчета коэффициента корреляции Пирсона в SPSS.....	21
Коэффициент корреляции Пирсона, 2 способа вычисления.....	21
Висновки.....	22
* Основні поняття, методи і формули статистики.....	22

**ТЕМА ЛЕКЦІЇ.**

**Схема Бернуллі.**

**ПЛАН ЛЕКЦІЇ**

- I. Залежні й незалежні випадкові величини.
- II. Коваріація і кореляція випадкових величин.
- III. Коефіцієнт кореляції та його властивості.
- IV. Лінійна регресія
- V. Метод найменших квадратів.

**ПИТАННЯ ДЛЯ САМОКОНТРОЛЮ**

1. Графічний метод аналізу статистичного взаємозв'язку.
2. Форма, тіснота й спрямованість статистичного взаємозв'язку
3. Лінійна регресія.
4. Визначення коефіцієнтів рівняння лінійної регресії. .
5. Оцінка достовірності існування статистичного взаємозв'язку

**ЗАВДАННЯ ДЛЯ ДІСТАНЦІЙНОГО ОПАНУВАННЯ**

Опрацювати наданий викладачів теоретичний матеріал.

Підготувати *опорний конспект* (основні означення, теореми, властивості, формули).

**Надати відповіді на питання самоконтролю.**

---

## **I. ЕЛЕМЕНТИ ТЕОРІЇ КОРЕЛЯЦІЇ ТА РЕГРЕСІЇ**

### **Стохастичний зв'язок між величинами**

**Статистичні дослідження** мають комплексний характер. Наприклад, при обстеженні пацієнта визначають цілу низку фізіологічних, біохімічних показників, психологічних станів та інших параметрів. Виникає питання про **взаємозв'язок окремих ознак**.

Залежності такого типу називають **стохастичними**. Для визначення ступеня стохастичного зв'язку використовують **кореляційний аналіз**.

## **II. КОЕФІЦІЄНТ КОРЕЛЯЦІЇ. ПОЗИТИВНА ТА НЕГАТИВНА КОРЕЛЯЦІЯ. СИЛЬНИЙ І СЛАБКИЙ КОРЕЛЯЦІЙНИЙ ЗВ'ЯЗОК. ДІАГРАМА РОЗСПІЮВАННЯ**

**Теорія кореляції** дає можливість **виміряти міцність зв'язку між ознаками**, явищами. Таку оцінку здійснюють шляхом розрахунку **коефіцієнта кореляції  $r$** .

У деяких випадках результатом спостережень може бути значення не однієї випадкової величини, а двох (у загальному випадку — декількох випадкових величин). Такий розподіл називають **двовимірним** (у загальному випадку — **багатовимірним**), наприклад, зв'язок між віком дитини та масою її тіла. Кожне спостереження зображують точкою на площині, координати якої є значеннями випадкових величин, що спостерігаються.

### **Коефіцієнт кореляції**

**Результати спостережень** можна записати у вигляді таблиці. Такі таблиці називаються **кореляційними таблицями**. Використовуючи кореляційні таблиці, можна підрахувати **коефіцієнт кореляції  $r$**  між двома випадковими величинами:

$$r = \frac{1}{n} \sum_{i=1}^n \frac{x_i - \bar{x}}{S_x} \cdot \frac{y_i - \bar{y}}{S_y},$$

де  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  — випадкові величини, що спостерігаються попарно,

$\bar{x}$  — середнє значення за вибіркою  $\{x_1, x_2, \dots, x_n\}$ ,

$S_x^2$  — вибіркова дисперсія за вибіркою  $\{x_1, x_2, \dots, x_n\}$ ,

$\bar{y}$  — середнє значення за вибіркою  $\{y_1, y_2, \dots, y_n\}$ ,

$S_y^2$  — вибіркова дисперсія за вибіркою  $\{y_1, y_2, \dots, y_n\}$ .

Згадаємо основні властивості математичного очікування

- $M(C) = C$ , де  $C$  — стала величина;
- $M(k\xi) = kM\xi$ ;
- $M(\xi + \eta) = M\xi + M\eta$  для будь-яких  $\xi, \eta$ ;
- $M(\xi \cdot \eta) = M\xi \cdot M\eta$ ; якщо  $\xi$  та  $\eta$  — незалежні.
- $D\xi = M(\xi^2) - (M\xi)^2 = M(\xi^2) - a^2$ .

Розглянемо нерівність Коші-Буняковського

$$|M(X \cdot Y)| \leq \sqrt{M((X)^2)} \cdot \sqrt{M((Y)^2)}$$

яке застосуємо до випадкових величин  $x - M(x)$  та  $y - M(y)$ , отримаємо

$$\begin{aligned} |M((x - M(x)) \cdot (y - M(y)))| &\leq \sqrt{M((x - M(x))^2)} \cdot \sqrt{M((y - M(y))^2)} = \\ &= \sqrt{D(x)} \cdot \sqrt{D(y)} \end{aligned}$$

З іншого боку,

$$M((x - M(x)) \cdot (y - M(y))) = M(xy - xM(x) - yM(y) + M(x) \cdot M(y))$$

Тоді з останніх двох тверджень випливає

$$M(xy - xM(y) - yM(x) + M(x) \cdot M(y)) \leq \sqrt{D(x)} \cdot \sqrt{D(y)}$$

$$M(xy) - M(x)M(y) - M(y)M(x) + M(x) \cdot M(y) \leq \sqrt{D(x)} \cdot \sqrt{D(y)}$$

$$M(xy) - M(x)M(y) \leq \sqrt{D(x)} \cdot \sqrt{D(y)}$$

Звідси випливає, що коефіцієнт кореляції  $r_{xy} = r$

$$r = \frac{M(xy) - M(x)M(y)}{\sqrt{D(x)} \cdot \sqrt{D(y)}}$$

буде за модулем не перевищувати одиницю  $|r| \leq 1$ .

Зауважимо, що математичне очікування в статистиці замінюється на вибіркове середнє:  $M(x) \sim \bar{x}$  ;  $M(y) \sim \bar{y}$  .

А середнє квадратичне відхилення, тобто корінь від дисперсії замінюється коренем з вибіркової дисперсії за вибіркою:

$$\sigma(x) = \sqrt{D(x)} \sim \sqrt{S_x^2} = S_x \quad \sigma(y) = \sqrt{D(y)} \sim \sqrt{S_y^2} = S_y$$

**Коефіцієнт кореляції** — це число, знак та величина якого характеризують напрям і силу зв'язку. Значення коефіцієнта кореляції може змінюватися від -1 до +1 (включаючи 0,0).

Силу кореляційного зв'язку оцінюють за числовим значенням коефіцієнта кореляції.

*Таблиця. Визначення ступеня кореляційного зв'язку від значення коефіцієнта кореляції*

Значення коефіцієнту кореляції	Кореляційний зв'язок
$ r  < 0,3$	Слабкий
$0,3 <  r  < 0,5$	Помірний
$0,5 <  r  < 0,7$	Значний
$0,7 <  r  < 0,9$	Сильний
$0,9 <  r  < 1$	Дуже сильний

**Знак** коефіцієнта кореляції **вказує на напрям** — прямий чи зворотний

взаємозв'язок між двома змінними.

Абсолютне значення коефіцієнта кореляції характеризує силу та щільність взаємозв'язку, що розглядається.

Якщо коефіцієнт кореляції  $r = 0$ , то величини  $x$  і  $y$  називають некорельованими, такими є незалежні випадкові величини, оскільки для незалежних змінних математичне очікування добутку дорівнює добутку математичних очікувань співмножників.

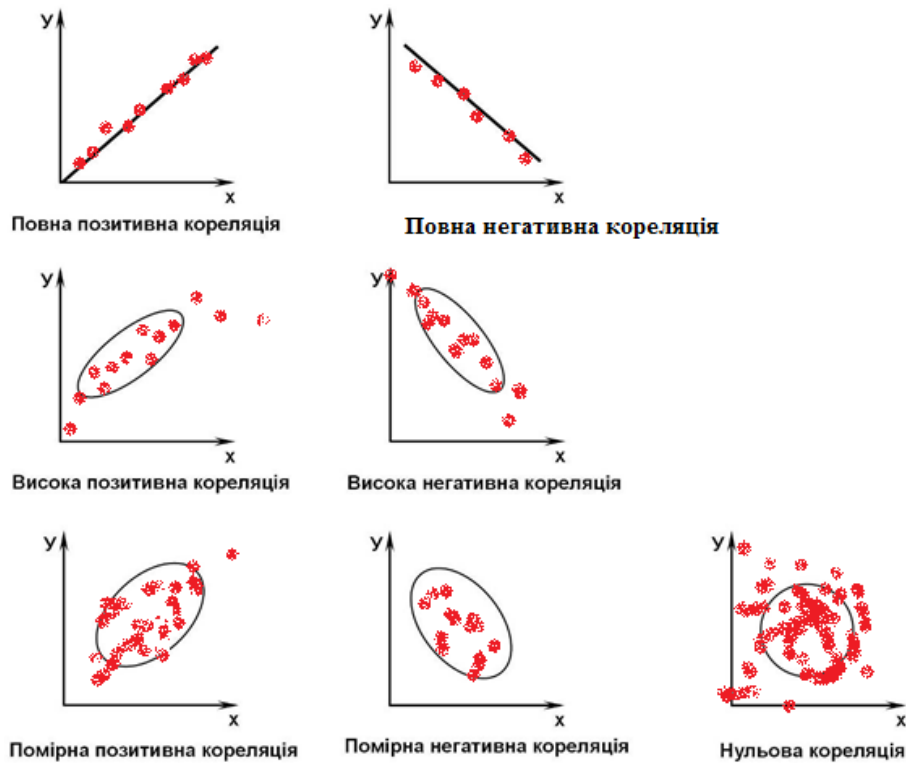
### Кореляційна залежність.

Кореляційна залежність називається позитивною, якщо при збільшенні однієї ознаки збільшується і друга, і негативною, якщо при збільшенні першої ознаки друга зменшується.

Зміст концепції кореляції можна з'ясувати за допомогою так званої *діаграми розсіювання*. При побудові діаграми розсіювання на осях координат відкладають значення відповідних випадкових величин.

Можна вважати, що експериментальні дані потрапляють у геометричну фігуру, котра має форму еліпса: що менша мала вісь еліпса при одній і тій самій великій осі, тим більшим є значення коефіцієнта кореляції; якщо еліпс перетворюється на коло, це означає, що стохастичний зв'язок між змінними відсутній (коефіцієнт кореляції дорівнює нулю). Якщо мала вісь еліпса спрямована до нуля (втягнутий еліпс, що переходить у пряму), відзначають повну позитивну або негативну стохастичну залежність (коефіцієнт кореляції дорівнює  $\pm 1$ ).

Схематичне представлення кореляційної залежності випадкових величин коефіцієнта кореляції зображено на мал.



Мал. Схематичне представлення кореляційної залежності між випадковими величинами

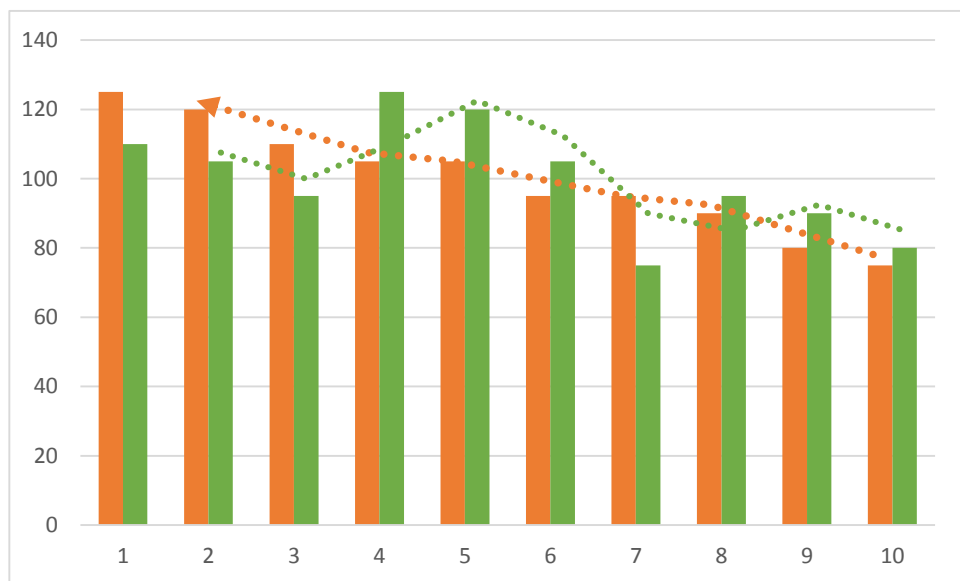
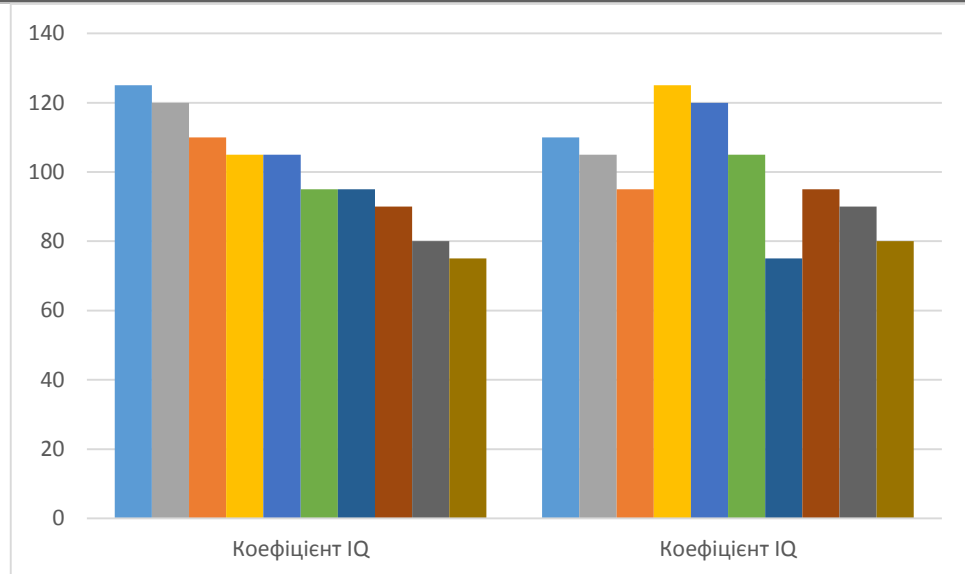
Приклад. Розрахувати коефіцієнт кореляції між середніми значеннями коефіцієнта розумового розвитку (IQ) батьків та їхніх дітей.

Таблиця. Розрахунок коефіцієнта кореляції

Середні значення IQ для обох батьків		Діти		Добуток Z-оцінок
Коефіцієнт IQ	$Z_x = \frac{x - \bar{x}}{s_x}$	Коефіцієнт IQ	$Z_y = \frac{y - \bar{y}}{s_y}$	
125	+1,63	110	+0,65	1,06
120	+1,30	105	+0,33	0,43
110	+0,65	95	-0,33	-0,21
105	+0,33	125	+1,63	0,54
105	+0,33	120	+1,30	0,43
95	-0,33	105	+0,33	-0,11
95	-0,33	75	-1,63	0,54
90	-0,65	95	-0,33	0,21
80	-1,30	90	-0,65	0,85
75	-1,63	80	-1,3	2,12



$\bar{x} = 100$	0,00	$\bar{y} = 100$	0,00	$r = 0,59$
$S_x = 15,33$	1,00	$S_y = 15,33$	1,00	$r = \frac{1}{n} \sum S_x S_y$



Висновок: за значенням  $r = 0,59 > 0$  можна зробити висновок про **помірну позитивну** залежність коефіцієнта IQ батьків та їхніх дітей.

### **III. ПОНЯТТЯ РЕГРЕСІЇ. ЗАГАЛЬНА ПОСТАНОВКА ЗАДАЧІ. РІВНЯННЯ РЕГРЕСІЇ**

**Регресійний аналіз** є одним із найширше використовуваних статистичних методів. Він охоплює велику кількість інших статистичних процедур (*гіпотези про середні і дисперсії, кореляційний і дисперсійний аналіз, планування експерименту* тощо) і розділи інших наук (наприклад лінійна алгебра).

#### **Задачі регресійного аналізу**

***Задачі регресійного аналізу*** є отримання за експериментальними даними ***математичного рівняння (моделі)***, що описує поведінку деякої величини  $y$  ***залежно від***  $x$ .

Знаючи коефіцієнт кореляції, можна за величиною однієї з корелювальних між собою змінних передбачити відповідне значення другої змінної.

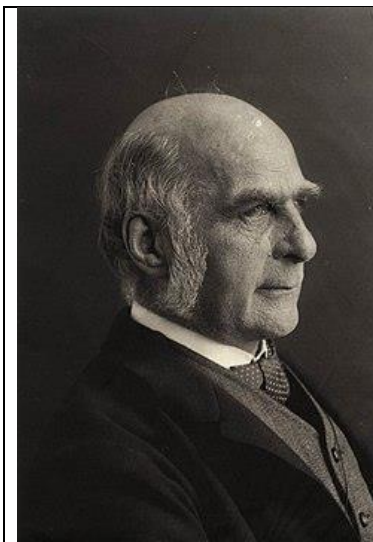
#### **Рівняння регресії**

Рівняння регресії для  $Y$  за  $X$  має вигляд:

$$Z_y = rZ_x,$$

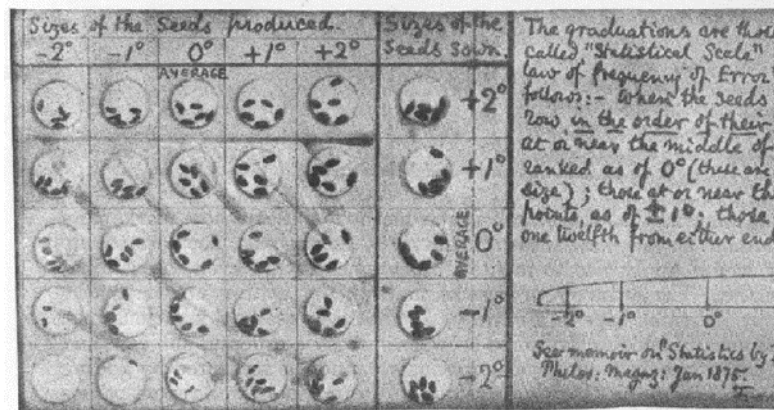
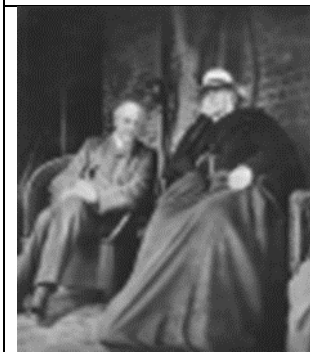
$$\text{де } Z_x = \frac{x_i - \bar{x}}{S_x}, Z_y = \frac{y_i - \bar{y}}{S_y}, S_x = \sqrt{\frac{1}{n} \sum (x_i - \bar{x})^2}, S_y = \sqrt{\frac{1}{n} \sum (y_i - \bar{y})^2}.$$

Термін “регресія” понад 100 років тому було застосовано англійським статистиком *Ф. Голтеном* при вивченні спадкових ознак.



**Сір Френсіс Голтен** (англ. *Francis Galton*; /'gɔ:ltən/, нар. 16 лютого 1822 - пом. 17 січня 1911) – англійський антрополог, дослідник, географ, статистик, полимат, соціолог і психолог.

У 1892 монографія про відбитки пальців *Finger prints* підводить підсумок дослідженням у цій галузі і закладає основні принципи дерматогліфіки. Займається біологічної статистикою, першим запропонував те, як обчислити коефіцієнт кореляції. Запропонував закон регресії спадкових ознак. В останні роки життя займався розробкою основних положень науки евгеніки про створення ідеальної з усякого погляду людини.



Френсіс Галт он у віці 87 років разом із Карлом Пірсоном, його біографом та співробітником <http://galton.org/statistician.html>

Через кілька років він сформулював статистичний коефіцієнт кореляції іншим непрямым маршрутом, кропітко графікуючи і перекроюючи свої дані про біваріантні нормальні розподіли, поки не зрозумів, що формули для еліптичних кривих (тема, популярна в математиці 19 століття, але майже повністю відмирає сьогодні) міг би надати йому метод узагальнення за числом графічного співвідношення, яке він бачив. Це число потім може бути використане для обґрунтування відносин та формування основи для порівнянь. За пропозицією друга, він підійшов до математика в Кембриджі, щоб розробити для нього деталі - підхід, який він зміг дуже плідно домогтися в наступні роки з Пірсоном.

Зміст поняття регресія (повернення до середнього значення) виражав характер зв'язку між зростом батьків та їхніх дітей — тенденції до середнього зросту.

Лінійна кореляція

Якщо кореляція лінійна, то  
рівняння регресії можна записати  
наступним чином:

$$y - b = r \frac{S_2}{S_1} (x - a),$$

$$x - a = \frac{1}{r} \cdot \frac{S_1}{S_2} (y - b),$$

$$\text{де } r \frac{S_2}{S_1} = \operatorname{tg} \alpha,$$

$$\frac{1}{r} \cdot \frac{S_1}{S_2} = \operatorname{tg} \beta$$

— кутові коефіцієнти регресії

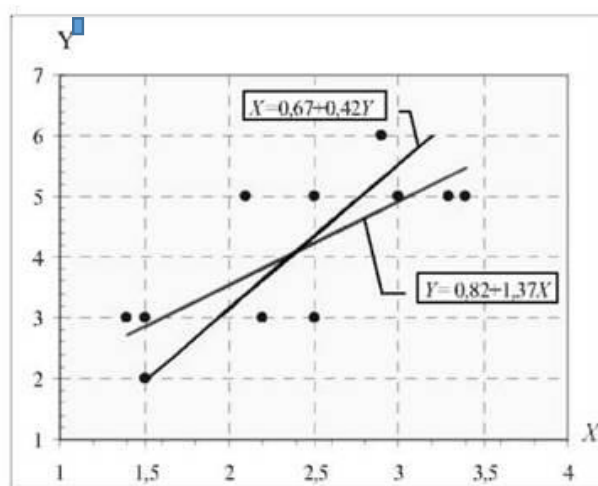
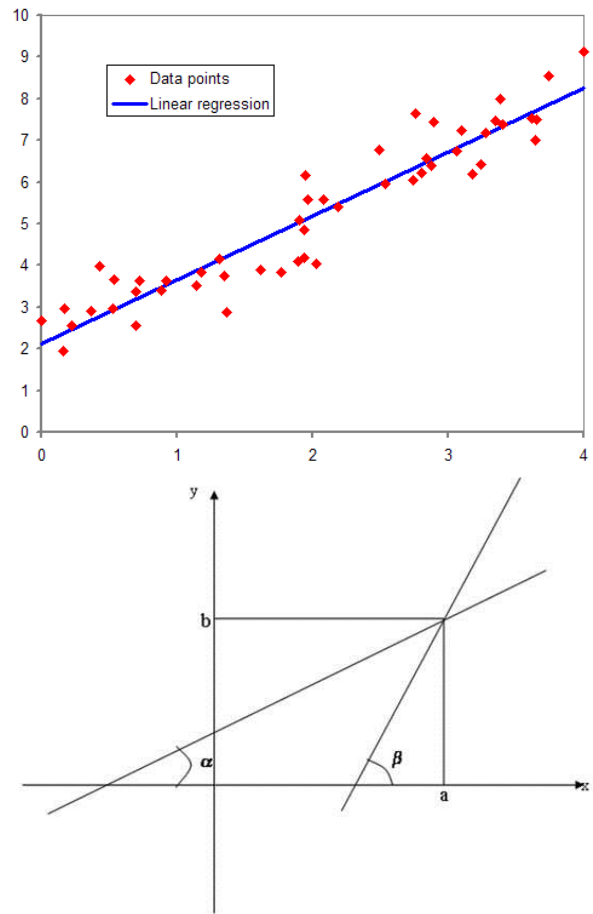


Рис. Рівняння регресії для лінійної кореляції

#### IV.ПРИКЛАДИ «ЗВ'ЯЗОК МІЖ ВЕЛИЧИНАМИ X ТА Y».

##### Приклад 1.

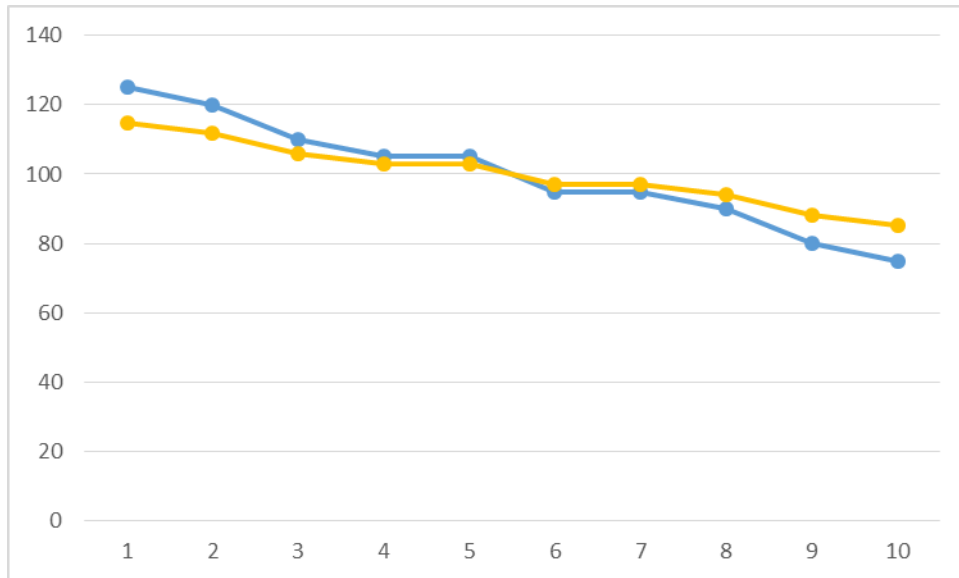
Отримати прогноз значень коефіцієнта IQ у дітей за середніми значеннями коефіцієнта IQ у батьків (коефіцієнти IQ у батьків і дітей є нормально розподіленими незалежними випадковими величинами; табл.).

*Таблиця. Прогноз значень коефіцієнта IQ*

Середні значення IQ для обох батьків		Прогноз $Z_y$	Прогноз коефіцієнта IQ у дітей
Коефіцієнт IQ	$Z_x = \frac{x - \bar{x}}{S_x}$		
125	+1,63	+0,96	114,72
120	+1,30	+0,77	111,80
110	+0,65	+0,38	105,83
105	+0,33	+0,20	103,07
105	+0,33	+0,20	103,07
95	-0,33	-0,20	96,93
95	-0,33	-0,20	96,93
90	-0,65	-0,38	94,17
80	-1,30	-0,77	88,20
75	-1,63	-0,96	85,28
$\bar{x} = 100$	0,00	0,00	100
$S_x^2 = 235$	1,00	0,35	81,75
$S_x = 15,33$	1,00	$S_y = 0,59$	9,04

Результати прогнозу (див. попередню табл., стовпчик 4) ілюструють явище, яке носить назву “регресія до середнього”. У стовпчику 3 стандартне відхилення  $S_y = 0,59$ , тобто воно дорівнює величині коефіцієнта кореляції між прогнозованими значеннями Z-оцінок:  $S_y = r$ .

Дисперсія  $S_y^2 = 0,35$ , тобто  $S_y^2 = r^2$  має особливий зміст: характеризує частину дисперсії значень  $Y$ , яку можна пояснити наявністю кореляції між  $X$  і  $Y$ .

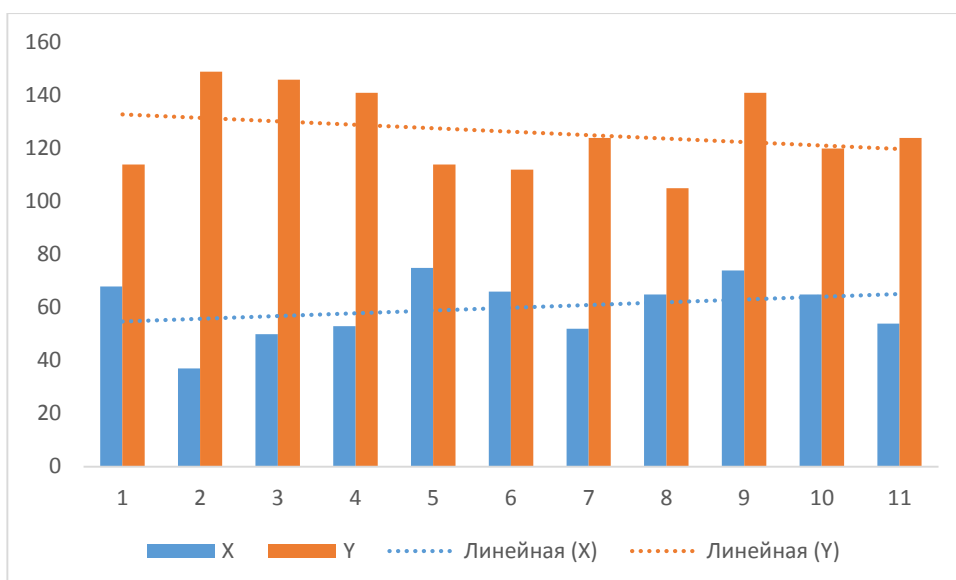


### Приклад 2.

Дослідимо залежність між двома величинами  $Y$  і  $X$ .

Результати спостережень занесемо в таблицю у вигляді двовимірної вибірки:

X	68	37	50	53	75	66	52	65	74	65	54
Y	114	149	146	141	114	112	124	105	141	120	124



Задача:

- 1) обчислити коефіцієнт кореляції;
- 2) дослідити кореляційні зв'язки ;
- 3) записати рівняння лінійної регресії Y на X.

▼ **Розв'язання.**

Використовуючи формули:

$$r_B = \frac{\overline{XY} - \overline{X}\overline{Y}}{\sigma_X \sigma_Y}$$

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{\overline{X^2} - (\overline{X})^2}, \sigma_Y = \sqrt{\sigma_Y^2} = \sqrt{\overline{Y^2} - (\overline{Y})^2}, \overline{Y^2} = \frac{1}{n} \sum_{j=1}^k n_{y_j} y_j^2$$

$$\overline{X^2} = \frac{1}{n} \sum_{i=1}^{11} x_i^2 = \frac{68^2 + 37^2 + 50^2 + 53^2 + 75^2 + 66^2 + 52^2 + 65^2 + 74^2 + 65^2 + 54^2}{11} =$$

$$= \frac{4624 + 1369 + 2500 + 2809 + 5625 + 4356 + 2704 + 4225 + 5476 + 4225 + 2916}{11} = \frac{40829}{11} = 3711,73$$

$$\overline{Y^2} = \frac{1}{n} \sum_{i=1}^{11} y_i^2 = \frac{177992}{11} = 16181,09, \quad \overline{X} = \frac{1}{n} \sum_{i=1}^{11} x_i = \frac{659}{11} = 59,91, \quad \overline{X^2} = 3589,21$$

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{11} y_i = \frac{1390}{11} = 126,36, \quad \overline{Y^2} = 15966,85,$$

$$\overline{XY} = \frac{1}{n} \sum_{i=1}^{11} x_i y_i = \frac{82183}{11} = 7471,2$$

$$r_B = \frac{7471,2 - 59,91 \cdot 126,36}{\sqrt{(3711,73 - 3589,21)(16181,09 - 15966,85)}} = \frac{-99,03}{\sqrt{122,52 \cdot 214,24}} = -\frac{99,03}{162,01} = -0,61$$

$$r_B = \frac{\overline{XY} - \overline{X}\overline{Y}}{\sigma_X \sigma_Y} \quad \sigma_X = \sqrt{\sigma_X^2} = \sqrt{\overline{X^2} - (\overline{X})^2}, \quad \sigma_Y = \sqrt{\sigma_Y^2} = \sqrt{\overline{Y^2} - (\overline{Y})^2}, \quad \overline{Y^2} = \frac{1}{n} \sum_{j=1}^k n_{y_j} y_j^2$$

Отже кореляційний зв'язок між величинами X і Y є значним зворотнім.

$0,3 <  r  < 0,5$	Помірний
$0,5 <  r  < 0,7$	Значний
$0,7 <  r  < 0,9$	Сильний

Знайдемо рівняння лінійної регресії Y на X:

$$Y - 126,36 = -0,61 \cdot \frac{\sqrt{214,24}}{\sqrt{122,52}} (X - 59,91) = -0,61 \cdot \frac{14,63}{11,09} (X - 59,91) = -0,80(X - 59,91)$$

$$Y = 126,36 - 0,80X + 48,53$$

$$Y = 174,59 - 0,80X$$



**Приклад 3.**

Внаслідок проведення п'яти незалежних випробувань дістали п'ять пар значень випадкових величин  $X$  і  $Y$ , заданих у вигляді таблиці:

$X$	2	2,2	2,3	2,5	2,8
$Y$	5	4,8	4,6	4,3	4,2

Обчислити основні числові характеристики випадкових величин  $X$  і  $Y$ .  
Визначити коефіцієнт кореляції цих величин та записати рівняння прямих регресій  $Y$  на  $X$  та  $X$  на  $Y$ .

▼ Розв'язок. Очевидно, що в результаті кожного з п'яти випробувань дістаємо єдину пару значень  $(x_i, y_j)$ , тому для варіант  $x_i$  та  $y_j$  випадкових величин  $X$  і  $Y$  частоти  $n_i = n_j = 1$ . Обчислимо статистичне середнє, дисперсію та середнє квадратичне відхилення для випадкових величин  $X$  і  $Y$ :

$$\bar{x} = \frac{1}{5} \sum_{i=1}^5 x_i n_i = \frac{1}{5} (2 + 2,2 + 2,3 + 2,5 + 2,8) = \frac{11,8}{5} = 2,36$$

$$\bar{y} = \frac{1}{5} \sum_{j=1}^5 y_j n_j = \frac{1}{5} (5 + 4,8 + 4,6 + 4,3 + 4,2) = \frac{22,9}{5} = 4,58$$

$$D_x = \frac{1}{5} (2^2 + 2,2^2 + 2,3^2 + 2,5^2 + 2,8^2) - (2,36)^2 = \frac{28,22}{5} - 5,5696 = 5,644 - 5,5696 = 0,0744;$$

$$D_y = \frac{1}{5} (5^2 + 4,8^2 + 4,6^2 + 4,3^2 + 4,2^2) - (4,58)^2 = \frac{105,33}{5} - 20,9764 = 21,066 - 20,9764 = 0,0896;$$

$$\sigma_x = \sqrt{D_x} = \sqrt{0,0744} \approx 0,273$$

$$\sigma_y = \sqrt{D_y} = \sqrt{0,0896} \approx 0,299$$

Для вивчення залежності між величинами  $X$  і  $Y$  обчислимо вибіркового коефіцієнт кореляції, який визначається за формулою:

$$r_{xy} = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y}.$$

Оскільки  $\overline{xy} = \frac{1}{5}(2 \cdot 5 + 2,2 \cdot 4,8 + 2,3 \cdot 4,6 + 2,5 \cdot 4,3 + 2,8 \cdot 4,2) = \frac{53,65}{5} = 10,73$ , то

$$r_{xy} = \frac{10,73 - 2,36 \cdot 4,58}{0,273 \cdot 0,299} = \frac{0,08}{0,0816} = -0,98.$$

Отже, коефіцієнт кореляції  $|r_{xy}| \approx 1$ , тому залежність між величинами  $X$  і  $Y$  можна вважати лінійною, причому кореляція є від'ємною (значення  $Y$  спадають при зростанні значень  $X$ ).

У цьому випадку лінії регресії є прямими.

Запишемо рівняння прямих регресії  $Y$  на  $X$ :

$$y - \bar{y} = r_{xy} \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

та  $X$  на  $Y$ :

$$x - \bar{x} = r_{xy} \frac{\sigma_x}{\sigma_y} (y - \bar{y}).$$

Тоді рівняння прямих регресії мають вигляд:

$Y$  на  $X$ :

$$y - 4,58 = -0,98 \cdot \frac{0,299}{0,273} (x - 2,36) \Rightarrow y - 4,58 = -1,073(x - 2,36) \Rightarrow y = -1,073x + 7,112$$

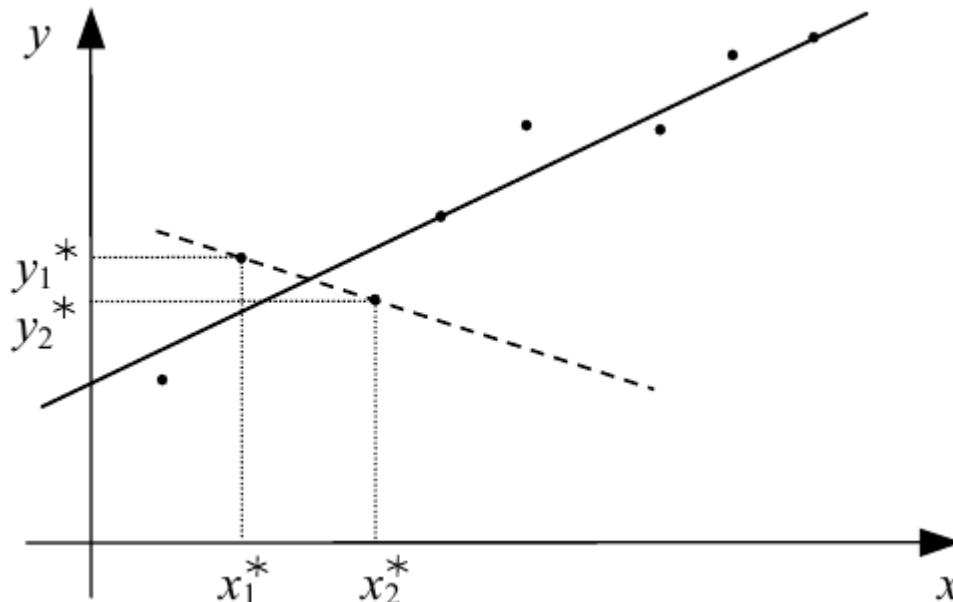
та  $X$  на  $Y$ :

$$x - 2,36 = -0,98 \cdot \frac{0,273}{0,299} (y - 4,58) \Rightarrow x - 2,36 = -0,895(y - 4,58) \Rightarrow x = -0,895y + 6,459.$$

**V.МЕТОД НАЙМЕНШИХ КВАДРАТІВ**

Метод найменших квадратів (МНК), ідея якого належить Гауссу.

Нехай відомо, що вихідний параметр процесу, який вивчається, позначимо його  $y$ , лінійно залежить від вхідного параметра  $x$  (суцільна пряма лінія на рис).



Рисунок— Графічна інтерпретація причин, які обумовлюють необхідність використання МНК

Тобто припустимо, що статична характеристика цього процесу може бути подана у вигляді

$$y = ax + b,$$

де  $a$  і  $b$  — коефіцієнти, для визначення числових значень яких необхідно, як мінімум, задати два значення  $x_1, x_2$  вхідній величині  $x$  і заміряти відповідні їм значення  $y_1, y_2$  вихідної величини  $y$ , оскільки лише під час виконання цих умов для моделі можна скласти систему двох алгебраїчних рівнянь із двома невідомими  $a$  і  $b$

$$\begin{cases} y_1 = ax_1 + b, \\ y_2 = ax_2 + b. \end{cases}$$

Але результати будь-яких експериментальних вимірювань несуть у собі похибки, обумовлені класом точності вимірювальних засобів, дією різноманітних завад, неточністю зчитування показів приладів, округленням

під час приведення даних до однакових умов обробки інформації — список умов виникнення похибок можна продовжити, але для обґрунтування МНК цього досить.

Тож через наявність цих похибок в експериментальних значеннях  $x_1, x_2, y_1, y_2$  безпосередній розв'язок системи рівнянь відносно  $a$  та  $b$  може нести в собі похибку в 10, 100, 1000 і більше відсотків.

Наприклад, якщо використати лише значення  $x_1^*, y_1^*; x_2^*, y_2^*$  для розв'язання системи рівнянь, то похибка буде вже не у відсотках, а у характері функціональної залежності (пунктирна лінія на рис.).

У свій час Гаусс запропонував інший спосіб визначення коефіцієнтів  $a, b$  моделі. Він запропонував сформулювати суму квадратів різниць  $\sum^N$  між теоретично заданими за допомогою рівняння значеннями вихідної координати  $y$  при значеннях аргументу  $x_i, i = \overline{1, N}$  та її експериментальними значеннями  $y_i$ :

$$\Sigma^N = \sum_{i=1}^N (y(x_i) - y_i)^2,$$

а потім знайти такі значення коефіцієнтів  $a, b$  рівняння, котрі мінімізують даний вираз.

Від цієї процедури і назва методу — метод найменших квадратів.

З курсу математичного аналізу відомо, що для знаходження мінімуму якоїсь функції необхідно взяти від неї похідну, прирівняти цю похідну до нуля і розв'язати отримане рівняння — його корінь задає значення аргументу, за якого функція досягає мінімуму, а само значення функції у цій точці, якщо вона опукла донизу, задає її мінімальне значення.

Згідно з цією ідеєю, підставимо у вираз замість  $y(x_i)$  його значення візьмемо від отриманого виразу частинні похідні за  $b$  та  $a$ , які прирівняємо до нуля, тобто

$$\Sigma^N = \sum_{i=1}^N (ax_i + b - y_i)^2,$$

$$\begin{cases} \frac{\partial \Sigma^N}{\partial b} = \sum_{i=1}^N 2 (ax_i + b - y_i)(1) = 0, \\ \frac{\partial \Sigma^N}{\partial a} = \sum_{i=1}^N 2 (ax_i + b - y_i)(x_i) = 0. \end{cases}$$

$$\begin{cases} b \cdot N + a \sum_{i=1}^N x_i = \sum_{i=1}^N y_i, \\ b \sum_{i=1}^N x_i + a \sum_{i=1}^N x_i^2 = \sum_{i=1}^N y_i x_i. \end{cases}$$

Розв'язавши останню систему рівнянь відносно  $b$  і  $a$ , отримаємо такі їх значення, які мінімізують суму квадратів відхилень експериментально виміряних значень величин  $x_i$ ,  $y_i$  від теоретично заданих згідно з вибраною функціональною залежністю.

Рівняння, що входять у означену систему, називають *нормальними рівняннями Гаусса*. Коефіцієнтами у них є суми, які «згладжують» дію похибок вимірювань величин  $x$ ,  $y$  і зменшують їх вплив на оцінки параметрів  $b$ ,  $a$ . Завдяки цьому підвищується точність їх визначення.

## **ДОДАТКОВІ ДЖЕРЕЛА**

### **Статистический анализ в MS Excel**



### **Линейный коэффициент корреляции Пирсона**

Корреляция и регрессия

Див.

<https://statanaliz.info/statistica/korrelyaciya-i-regressiya/linejnyj-koefficient-korrelyacii-pirsona/>

### **Коэффициент корреляции Пирсона в Excel**

В видео показан расчет коэффициента корреляции Пирсона с доверительными интервалами, ранговый коэффициент корреляции Спирмена.

<https://youtu.be/G6Oyg6rOnoY>

### **Результаты расчета коэффициента корреляции Пирсона в SPSS**

<https://www.youtube.com/watch?v=j622ZpLLbJY>

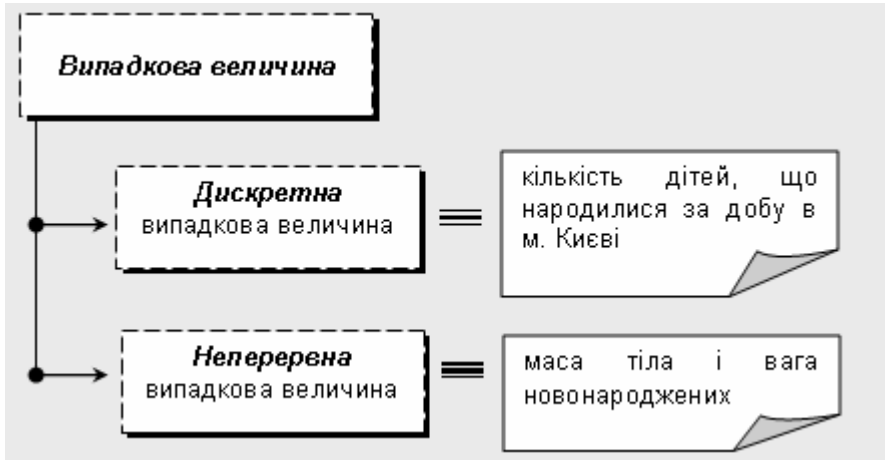
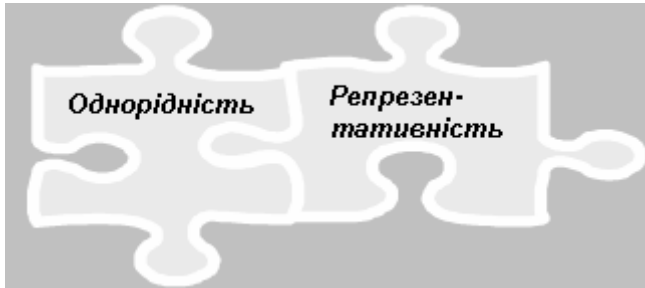
### **Коэффициент корреляции Пирсона, 2 способа вычисления**

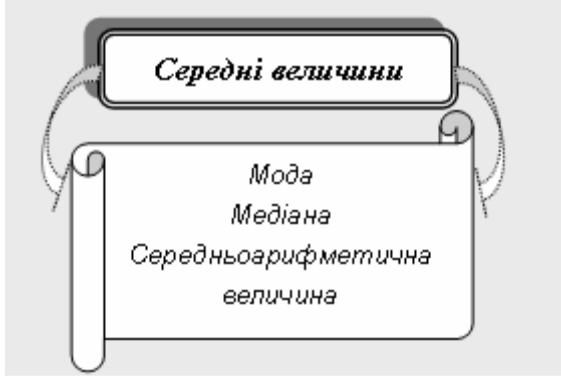

[https://www.youtube.com/watch?v=RXf9i\\_RB7X4](https://www.youtube.com/watch?v=RXf9i_RB7X4)

## **ВИСНОВКИ.**

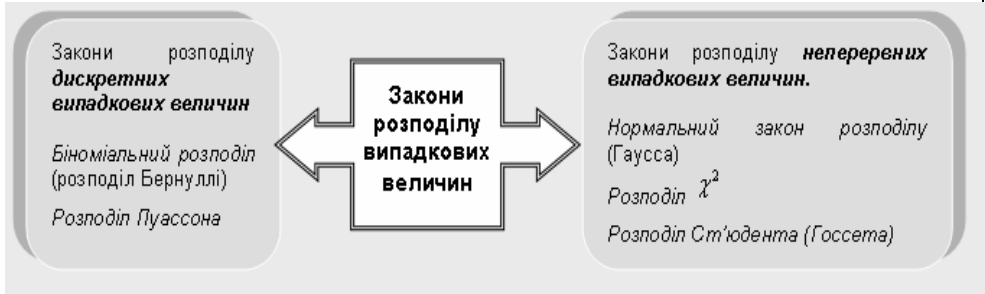
### **\* ОСНОВНІ ПОНЯТТЯ, МЕТОДИ І ФОРМУЛИ СТАТИСТИКИ**

Основні поняття, методи і формули на прикладі біостатистики

Параметр	Властивості, що піддаються оцінці в будь-якій формі (якісній або кількісній)
Статистична сукупність	Група, що складається з великої кількості відносно однорідних елементів (об'єктів), узятих разом у певних межах часу або простору
Випадкова величина	<p>Величина, яка в результаті експерименту, що може бути повторений за незмінних умов велику кількість разів, може набути значень <math>x_1, x_2, \dots, x_n</math>.</p> <p>Види випадкових величин:</p> 
Дискретна випадкова величина	Величина, яка може набувати скінченну кількість значень
Неперервна випадкова величина	Величина, яка може набувати будь-яких числових значень у даному інтервалі значень
Генеральна сукупність	Сукупність, що складається з усіх одиниць спостереження, що можуть бути до неї віднесені відповідно до мети дослідження
Вибірка (вибіркова сукупність)	<p>Частина генеральної сукупності, за властивостями якої судять про генеральну сукупність.</p> <p>Вимоги до вибіркової сукупності:</p> 

Варіаційний ряд	Сукупність значень вивченого в певному експерименті або спостереженні параметра, проранжованих за величинами (зростання або спадання)
Варіанта	Числове значення досліджуваної ознаки; складова варіаційного ряду
Середня величина	<p>Узагальнююча числова характеристика якісно однорідних величин, яка характеризує одним числом усю статистичну сукупність за однією ознакою</p> 
Мода	Значення, найпоширеніше в серії спостережень
Медіана	Значення, що поділяє розподіл на дві рівні частини, центральне або середнє значення серії спостережень, упорядкованих за зростанням або спаданням
Середньоарифметична величина	<p>Середня величина, яка розраховується за формулою:</p> $M = \frac{\sum_{i=1}^s x_i}{n} \quad (1)$
Частота (p)	<p>Абсолютна чисельність окремих варіант у сукупності, що вказує на поширеність цієї варіанти у варіаційному ряду.</p> <p>Види варіаційного ряду відповідно до значення частоти:</p> 
Середнє квадратичне відхилення (σ)	<p>Величина, яка характеризує ступінь розсіювання варіаційного ряду навколо середньої величини:</p> $\sigma = \pm \sqrt{\frac{\sum_{i=1}^n (x_i - x_{\text{сеп}})^2}{n - 1}} \quad (2)$



Коефіцієнт варіації $C_v$	Величина, необхідна для порівняння ступеня розмаїтості ознак, виражених у різноманітних одиницях виміру. Обраховується за формулою: $C_v = \frac{\sigma}{M} \times 100 \quad (3)$
Помилка репрезентативності	Найважливіша статистична величина, необхідна для оцінки достовірності результатів дослідження: $m = \frac{\sigma}{\sqrt{n}} \quad (4)$
Закон розподілу випадкових величин	Функціональна залежність між значеннями випадкових величин та ймовірностями, з якими вони набувають цих значень. Закони розподілу випадкових величин: 
Біноміальний розподіл (розподіл Бернуллі)	Дискретна випадкова величина $x$ , яка може набувати тільки цілих невід'ємних значень з імовірностями $P_n(X = m) = C_n^m p^m q^n$ , $m=0, 1, \dots, n$ , де $p$ – імовірність появи події в кожному випробуванні, $m$ – кількість сприятливих подій, $n$ – загальна кількість випробувань, $q=1-p$ , $C_n^m = \frac{n!}{m!(n-m)!}$ , називається розподіленою за біноміальним законом з математичним сподіванням $np$ та дисперсією $npq$ . Закон Бернуллі використовують тоді, коли необхідно знайти імовірність появи випадкової події, яка реалізується рівно $m$ з серії $n$ випробувань. Біноміальному закону розподілу підпорядковуються випадкові події, такі, як кількість викликів швидкої допомоги за певний проміжок часу, черги до лікаря в поліклініці, епідемії тощо
Розподіл Пуассона	Дискретна випадкова величина $X$ , яка може набувати тільки цілих невід'ємних значень з імовірностями $P_n(X = m) = \frac{\lambda^m e^{-\lambda}}{m!}, m = 0, 1, \dots, \lambda > 0$ , називається розподіленою за законом Пуассона з математичним сподіванням $\lambda$ і дисперсією $\lambda$ , де $\lambda = np$ . Розподіл Пуассона як граничний біноміальний використовується при розв'язуванні задач надійності медичного обладнання та апаратури, поширення епідемії, викликів до хворого дільничних лікарів та інших задач масового обслуговування

Нормальний закон розподілу (Гаусса)	<p>У біології та медицині найчастіше розглядають випадкові величини, які мають нормальний закон розподілу: частоту дихання, частоту серцевих скорочень, динаміку росту популяції тощо. Стандартним нормальним розподілом називають розподіл з нульовим математичним сподіванням і одиничною дисперсією, щільність розподілу якого має наступний вигляд:</p> $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$
Розподіл $\chi^2$	<p>Нехай незалежні випадкові величини <math>x_1, x_2, \dots, x_n</math> розподілені за нормальним законом з <math>m=0</math> та <math>\sigma^2=1</math>. Закон розподілу випадкової величини <math>\chi^2 = \sum_{i=1}^n x_i^2</math> називається «хі-квадрат» розподілом з <math>n</math> ступенями вільності (кількість незалежних координат). Зі збільшенням ступенів вільності розподіл <math>\chi^2</math> наближається до нормального</p>
Розподіл Ст'юдента (Госсета)	<p>Нехай <math>x, y</math> – незалежні випадкові величини, причому <math>x</math> розподілено за нормальним законом з параметрами <math>(0;1)</math>, <math>y</math> – за законом <math>\chi^2</math> з <math>n</math> ступенями вільності. Тоді розподіл випадкової величини <math>t = \frac{x}{\sqrt{y}}</math> називається законом Ст'юдента з <math>n</math> ступенями вільності або <math>t</math>-розподілом.</p> <p>При збільшенні ступенів вільності розподіл Ст'юдента наближається до нормального</p>