Лекція 7. ДВОВИМІРНИЙ СТАТИСТИЧНИЙ РОЗПОДІЛ ВИБІРКИ ТА ЙОГО ЧИСЛОВІ ХАРАКТЕРИСТИКИ

7.1. Двовимірна вибірка. Статистичний розподіл вибірки

Нехай над системою випадкових величин (X,Y) в однакових умовах проведено n незалежних випробувань. Вибіркою обсягом n є послідовність $(x_1,y_1);(x_2,y_2);\ldots(x_n,y_n)$ пар значень, яких набувають складові X та Y системи в цих випробуваннях.

Таблична форма двовимірного розподілу має такий вигляд:

x_i	x_1	x_2	$\ldots$	x_k
y_i	y_1	y_2	$\ldots$	y_k
n_i	n_1	n_2	$\ldots$	n_k

У даному статистичному розподілу x_i та y_i - перелік варіант; n_i - відповідні цим парам варіанти частот.

Кореляційна залежність – це залежність між ознаками X та Y, коли при зміні однієї з ознак змінюється середнє значення іншої.

Кореляційне поле ознак X та Y - це графічне представлення результатів досліджень на координатній площині xOy у вигляді точок з координатами $(x_1,y_1);(x_2,y_2);\ldots(x_n,y_n)$. Кореляційне поле ще називають **діаграмою розсіювання**.

Приклад 7.1. Задано двовимірну вибірку

x_i	1	2	3	4	5	6	7	8	9	10
y_i	2	3,7	6,2	7,9	9,9	12	14,1	16,3	17,8	19,9

Побудувати кореляційне поле.

Розв'язок. Відкладемо на площині xOy точки з координатами (1;2), (2;3,7), (3;6,2) та ін. Отримаємо кореляційне поле для значень ознак *X* та *Y*, на якому чітко видно лінійну залежність *Y* від *X* (рис 7.1).

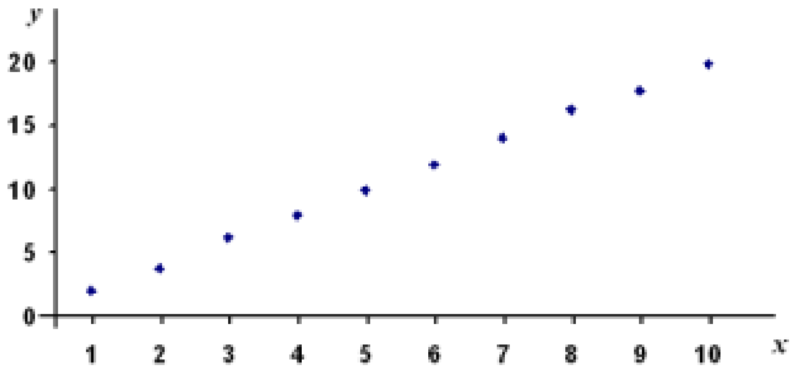


Рис. 7.1.

Кореляційна таблиця двовимірного розподілу:

$Y = y_i$	$X = x_j$					
	x_1	x_2	x_3	$\cdots$	x_m	n_{yi}
y_1	n_{11}	n_{12}	n_{13}	$\cdots$	n_{1m}	n_{y_1}
y_2	n_{21}	n_{22}	n_{23}	$\cdots$	n_{2m}	n_{y_2}
y_3	n_{31}	n_{32}	n_{33}	$\cdots$	n_{3m}	n_{y_3}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
y_k	n_{k1}	n_{k2}	n_{k3}	$\cdots$	n_{km}	n_{y_k}
n_{x_j}	n_{x1}	n_{x2}	n_{x3}	$\cdots$	n_{xm}	

де $x_1, x_2, \ldots, x_m$ - значення варіант ознаки X; $y_1, y_2, \ldots, y_k$ - значення варіант ознаки Y; n_{ij} - частота спільної появи варіант $Y = y_i,\ X = x_j \left(i = \overline{1,k};\ j = \overline{1,m} \right)$; n_{x_j} - частота, з якою зустрічається варіанта x_j; n_{y_i} - частота, з якою зустрічається варіанта y_i; об'єм вибірки за ознакою X, об'єм вибірки за ознакою Y та загальний об'єм вибірки відповідно дорівнюють:

$$n_{x_j} = \sum_{i=1}^{k} n_{ij}; \quad n_{y_i} = \sum_{j=1}^{m} n_{ij}, \quad n = \sum_{i=1}^{k} \sum_{j=1}^{m} n_{ij} = \sum_{i=1}^{k} n_{y_i} = \sum_{j=1}^{m} n_{x_j} \ .$$

7.2. Статистичні оцінки параметрів двовимірної системи

Загальні числові характеристики ознаки X:

Загальна середня величина ознаки X:

$$\overline{x}=\frac{\sum_{j=1}^{m}\sum_{i=1}^{k}x_j n_{ij}}{n}=\frac{\sum_{j=1}^{m}x_j n_{x_j}}{n}; \qquad (7.1)$$

Загальна дисперсія ознаки X:

$$D_x=\frac{\sum_{j=1}^{m}\sum_{i=1}^{k}x_j^2 n_{ij}}{n}-\left(\overline{x}\right)^2=\frac{\sum_{j=1}^{m}x_j^2 n_{x_j}}{n}-\left(\overline{x}\right)^2; \qquad (7.2)$$

або за виправленими вибірковими дисперсіями ознаки X:

$$S_x^2=\frac{\sum_{j=1}^{m}\sum_{i=1}^{k}\left(x_j-\overline{x}\right)^2 n_{ij}}{n-1}=\frac{\sum_{j=1}^{m}\left(x_j-\overline{x}\right)^2 n_{x_j}}{n-1}$$

Загальне середнє квадратичне відхилення ознаки X

$$\sigma_x=\sqrt{D_x}=\sqrt{S_x^2} \qquad (7.3)$$

Загальні числові характеристики ознаки Y:

Загальна середня величина ознаки Y:

$$\overline{y}=\frac{\sum_{i=1}^{k}\sum_{j=1}^{m}y_i n_{ij}}{n}=\frac{\sum_{i=1}^{k}y_i n_{y_i}}{n}; \qquad (7.4)$$

Загальна дисперсія ознаки Y:

$$D_y=\frac{\sum_{i=1}^{k}\sum_{j=1}^{m}y_i^2 n_{ij}}{n}-\left(\overline{y}\right)^2=\frac{\sum_{i=1}^{k}y_i^2 n_{y_i}}{n}-\left(\overline{y}\right)^2; \qquad (7.5)$$

або за виправленими вибірковими дисперсіями ознаки Y:

$$S_y^2=\frac{\sum_{i=1}^{k}\sum_{j=1}^{m}\left(y_i-\overline{y}\right)^2 n_{ij}}{n-1}=\frac{\sum_{i=1}^{k}\left(y_i-\overline{y}\right)^2 n_{y_i}}{n-1}.$$

Загальне середнє квадратичне відхилення ознаки Y

$$\sigma_y=\sqrt{D_y}=\sqrt{S_y^2} \qquad (7.6)$$

7.3. Умовні статистичні розподіли та їх числові характеристики

Умовним статистичним розподілом ознаки Y і при фіксованому значенні $X = x_i$ називають перелік варіант ознаки Y та відповідних їм частот, узятих при фіксованому значенні X.

$Y / X = x_j$.

$Y = y_i$	y_1	y_2	y_3	$\dots$	y_k
n_{ij}	n_{1j}	n_{2j}	n_{3j}	$\dots$	n_{ij}

де $\sum_{i=1}^{k} n_{ij} = n_{x_j}$.

Числові характеристики для такого статистичного розподілу називають умовними. До них належать:

умовна середня ознаки Y: $$\overline{y}_{x=x_j} = \frac{\sum_{i=1}^{k} y_i n_{ij}}{\sum_{i=1}^{k} n_{ij}} = \frac{\sum_{i=1}^{k} y_i n_{ij}}{n_{x_j}}; \quad (7.7)$$

умовна дисперсія ознаки Y: $$D\left(Y / X = x_j\right) = \frac{\sum_{i=1}^{k} y_i^2 n_{ij}}{n_{x_j}} - \left(\overline{y}_{x=x_j}\right)^2; \quad (7.8)$$

умовне середнє квадратичне відхилення ознаки Y:

$$\sigma\left(Y / X = x_j\right) = \sqrt{D\left(Y / X = x_j\right)} \quad (7.9)$$

$D\left(Y / X = x_j\right)$, $\sigma\left(Y / X = x_j\right)$ вимірюють розсіювання варіант ознаки Y щодо умовної середньої величини $\overline{y}_{x=x_j}$.

Умовним статистичним розподілом ознаки X і при $Y = y_i$ називають перелік варіант $X = x_j$ та відповідних їм частот, узятих при фіксованому значенні ознаки $Y = y_i$.

$X / Y = y_i$.

$X = x_j$	x_1	x_2	x_3	$\dots$	x_m
n_{ij}	n_{i1}	n_{i2}	n_{i3}	$\dots$	n_{im}

де $\sum_{j=1}^{m} n_{ij} = n_{y_i}$.

Умовні числові характеристики для цього розподілу:

умовна середня величина ознаки X : $\overline{x}_{y=y_i} = \frac{\sum_{j=1}^{m} x_j n_{ij}}{\sum_{j=1}^{m} n_{ij}} = \frac{\sum_{j=1}^{m} x_j n_{ij}}{n_{y_i}}$; (7.10)

умовна дисперсія ознаки X : $D\left(X \;/\; Y = y_i\right) = \frac{\sum_{j=1}^{m} x_j^2 n_{ij}}{n_{y_i}} - \left(\overline{x}_{y=y_i}\right)^2$; (7.11)

умовне середнє квадратичне відхилення ознаки X :

$$\sigma\left(X \;/\; Y = y_i\right) = \sqrt{D\left(X \;/\; Y = y_i\right)} \qquad (7.12)$$

При відомих значеннях умовних середніх $y^*_{x_j}$, $x^*_{y_i}$ загальні середні ознаки X та Y можна обчислити за формулами:

$$\overline{y} = \frac{\sum_{j=1}^{m} y^*_{x_j} n_{x_j}}{n}; \qquad (7.13)$$

$$\overline{x} = \frac{\sum_{i=1}^{k} x^*_{y_i} n_{y_i}}{n}; \qquad (7.14)$$

7.4. Парний статистичний розподіл вибірки та його числові характеристики

Якщо частота спільної появи ознак X і Y $n_{ij} = 1$ для всіх варіант, то в цьому разі двовимірний статистичний розподіл набуває такого вигляду:

$Y = y_i$	y_1	y_2	y_3	$\ldots$	y_n
$X = x_j$	x_1	x_2	x_3	$\ldots$	x_k

Його називають *парним статистичним розподілом вибірки*. Тут кожна пара значень ознак X і Y з'являється лише один раз.

Обсяг вибірки в цьому разі дорівнює кількості пар, тобто n.

Числові характеристики ознаки X:

середня величина ознаки X: $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$; (7.15)

дисперсія ознаки X: $D_x = \dfrac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2$ (7.16)

виправлена вибіркова дисперсія: $S_x^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$

середнє квадратичне відхилення ознаки X: $\sigma_x = \sqrt{D_x}$ (7.17)

Числові характеристики ознаки Y:

середня величина ознаки Y: $\bar{y} = \dfrac{\sum_{i=1}^{n} y_i}{n}$; (7.18)

дисперсія ознаки Y: $D_y = \dfrac{\sum_{i=1}^{n} y_i^2}{n} - \bar{y}^2$ (7.19)

виправлена вибіркова дисперсія: $S_y^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$.

середнє квадратичне відхилення ознаки Y: $\sigma_y = \sqrt{D_y} = \sqrt{S^2}$ (7.20)

7.5. Кореляційний момент, вибірковий коефіцієнт кореляції

Кореляція - це статистична залежність між випадковими величинами, що носить імовірний характер.

Коваріація (кореляційний момент) двох досліджуваних ознак X та Y – це середнє значення добутків відхилень для кожної пари варіант величин X та Y:

$$K_{xy}^{*}=\frac{\sum_{j=1}^{m}\sum_{i=1}^{k}y_i x_j n_{ij}}{n}-\overline{x}\cdot\overline{y}. \text{ або } K_{xy}^{*}=\frac{\sum_{j=1}^{m}\sum_{i=1}^{k}\left(x_j-\overline{x}\right)\cdot\left(y_i-\overline{y}\right)\cdot n_{ij}}{n-1} \qquad (7.21)$$

На практиці обчислення коваріації для незгрупованих даних або для парного статистичного ряду проводять за формулою:

$$K_{xy}^{*}=\frac{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\cdot\left(y_i-\overline{y}\right)}{n-1} \text{ або } K_{xy}^{*}=\frac{\sum_{i=1}^{n}x_i y_i}{n}-\overline{x}\cdot\overline{y} \qquad (7.22)$$

для згрупованих даних - $K_{xy}^{*}=\frac{\sum_{i=1}^{n}y_i x_i n_i}{n}-\overline{x}\cdot\overline{y}$.

Якщо ознаки X та Y незалежні, то коваріація дорівнює нулю. Обернене твердження не буде справедливим.

Вибірковий коефіцієнт кореляції. Для вимірювання тісноти кореляційного зв’язку обчислюється вибірковий коефіцієнт кореляції r_B за формулою:

$$r_B=\frac{K_{xy}^{*}}{\sigma_x\sigma_y} \qquad (7.23)$$

Як і в теорії ймовірності, $\left|r_B\right|\le 1,\ -1\le r_B\le 1$.

Для оцінювання сили зв’язку між корелюючими ознаками використовують шкалу Чеддока: якщо $\left|r_B\right|=0,1\div 0,3$, то лінійний зв’язок дуже слабкий, якщо $\left|r_B\right|=0,3\div 0,5$ - зв’язок слабкий, якщо $\left|r_B\right|=0,5\div 0,7$ – зв’язок середній, якщо $\left|r_B\right|=0,7\div 0,9$ – зв’язок сильний, якщо $\left|r_B\right|>0,9$– зв’язок дуже сильний.

У випадку повної кореляції всі точки $\left(x_i, y_i\right), i=\overline{1,n}$, будуть розміщені на одній прямій. Якщо $r_B=1$, то між вибірковими даними існує прямий лінійний зв’язок: із збільшенням значень однієї вибірки відповідні значення другої вибірки також збільшуються. Якщо $r_B = -1$, то між вибірковими даними є обернений лінійний зв'язок: із збільшенням значень однієї вибірки відповідні значення другої вибірки зменшуються. Якщо $r_B = 0$, то говорять, що дві вибірки є некорельовані, при цьому точки $\left(x_i, y_i\right)$ розміщені на площині хаотично.

Якщо $0 < r_B < 1$, то можна знайти таку пряму, від якої точки $\left(x_i, y_i\right)$ відхиляються найменше у тому сенсі, що сума квадратів відстаней від $\left(x_i, y_i\right)$ точок до цієї прямої буде мінімальною. Вказана пряма називається *прямою вибіркової лінійної регресії* y на x. Вона визначається рівнянням $y = ax + b$, де $a = r_B \dfrac{\sigma_y}{\sigma_x}$, $b = \overline{y} - a\overline{x}$. Кутовий коефіцієнт даної прямої називається *вибірковим коефіцієнтом регресії* y на x, він показує, на скільки одиниць в середньому змінюється змінна y при x збільшенні на одну одиницю.

Невідомі параметри a і b в рівнянні вибіркової лінійної регресії y на x можуть бути знайдені і як розв'язки нормальної системи методу найменших квадратів:

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

Приклад 7.1. За заданим двовимірним статистичним розподілом вибірки ознак X і Y

Y	X				
	10	20	30	40	n_{y_i}
2	-	2	4	4	**10**
4	10	8	6	6	**30**
6	5	10	5	-	**20**
8	15	-	15	10	**40**
n_{xj}	**30**	**20**	**30**	**20**	

Обчислити: K_{xy}^{*}, r_B та побудувати статистичні розподіли $Y / X = 30$, $X / Y = 4$. Обчислити умовні числові характеристики.

Розв'язання: Щоб обчислити K_{xy}^{*}, r_B визначимо $\overline{x}, \sigma_x, \overline{y}, \sigma_y$. Оскільки $n = \sum_{i=1}^{4}\sum_{j=1}^{4} n_{ij} = \sum_{i=1}^{4} n_{y_i} = \sum_{j=1}^{4} n_{x_j} = 100$, то

$$\overline{x} = \frac{\sum_{j=1}^{4} x_j \cdot n_{x_j}}{n} = \frac{10 \cdot 30 + 20 \cdot 20 + 30 \cdot 30 + 40 \cdot 20}{100} = 24.$$

$$D_x = \frac{\sum_{j=1}^{5} x_j^2 \cdot n_{x_j}}{n} - \left(\overline{x}\right)^2 = \frac{10^2 \cdot 30 + 20^2 \cdot 20 + 30^2 \cdot 30 + 40^2 \cdot 20}{100} - 24^2 = 124.$$

$$\sigma_x = \sqrt{D_x} = \sqrt{124} \approx 11{,}14.$$

$$\overline{y} = \frac{\sum_{i=1}^{4} y_i \cdot n_{y_i}}{n} = \frac{2 \cdot 10 + 4 \cdot 30 + 6 \cdot 20 + 8 \cdot 40}{100} = 5{,}8$$

$$D_y = \frac{\sum_{i=1}^{4} y_i^2 \cdot n_{y_i}}{n} - \left(\overline{y}\right)^2 = \frac{2^2 \cdot 10 + 4^2 \cdot 30 + 6^2 \cdot 20 + 8^2 \cdot 40}{100} - 5{,}8^2 = 4{,}36$$

$$\sigma_y = \sqrt{D_y} = \sqrt{4{,}36} \approx 2{,}1.$$

$$K_{xy}^{*} = \frac{\sum_{i=1}^{4}\sum_{j=1}^{4} y_i x_i n_{ij}}{n} - \overline{x} \cdot \overline{y} = \frac{2 \cdot 10 \cdot 0 + 2 \cdot 20 \cdot 2 + 2 \cdot 30 \cdot 4 + 2 \cdot 40 \cdot 4 +}{100}$$

$$\frac{+4 \cdot 10 \cdot 10 + 4 \cdot 20 \cdot 8 + 4 \cdot 30 \cdot 6 + 4 \cdot 40 \cdot 6 + 6 \cdot 10 \cdot 5 + 6 \cdot 20 \cdot 10 + 6 \cdot 30 \cdot 5 +}{100}.$$

$$\frac{+6 \cdot 40 \cdot 0 + 8 \cdot 10 \cdot 15 + 8 \cdot 20 \cdot 0 + 8 \cdot 30 \cdot 15 + 8 \cdot 40 \cdot 10}{100} - 24 \cdot 5{,}8 = -1{,}6.$$

$$r_B = \frac{K_{xy}^{*}}{\sigma_x \sigma_y} = \frac{-1{,}6}{11{,}14 \cdot 2{,}1} \approx -0{,}068.$$

Умовний статистичний розподіл $Y / X = 30$ матиме такий вигляд:

$Y = y_i$	2	4	6	8
$X = n_{i3}$	4	6	5	15

Обчислимо умовні числові характеристики для цього розподілу:

$$\bar{y}_{X=30} = \frac{\sum_{i=1}^{4} y_i n_{i3}}{\sum_{i=1}^{4} n_{i3}} = \frac{2 \cdot 4 + 4 \cdot 6 + 6 \cdot 5 + 8 \cdot 15}{30} = 6,07;$$

Умовна дисперсія та середнє квадратичне відхилення:

$$D_{(Y/X=30)} = \frac{\sum_{i=1}^{4} y_i^2 n_{i3}}{\sum_{i=1}^{4} n_{i3}} - \left(\bar{y}_{X=30}\right)^2 = \frac{2^2 \cdot 4 + 4^2 \cdot 6 + 6^2 \cdot 5 + 8^2 \cdot 15}{30} - \left(6,07\right)^2 \approx 4,89;$$

$$\sigma_{(Y/X=30)} = \sqrt{D_{(Y/X=30)}} = \sqrt{4,89} \approx 2,21.$$

Умовний статистичний розподіл $X / Y = 4$ матиме такий вигляд:

$X = x_j$	10	20	30	40
$Y = n_{2j}$	10	8	6	6

Обчислимо умовні числові характеристики для цього розподілу:

$$\bar{x}_{Y=4} = \frac{\sum_{j=1}^{4} x_i n_{2j}}{\sum_{j=1}^{4} n_{2j}} = \frac{10 \cdot 10 + 20 \cdot 8 + 30 \cdot 6 + 40 \cdot 6}{30} \approx 22,7;$$

Умовна дисперсія та середнє квадратичне відхилення:

$$D_{(X/Y=4)} = \frac{\sum_{j=1}^{4} x_j^2 n_{2j}}{\sum_{j=1}^{4} n_{2j}} - \left(\bar{x}_{Y=4}\right)^2 = \frac{10^2 \cdot 10 + 20^2 \cdot 8 + 30^2 \cdot 6 + 40^2 \cdot 6}{30} - \left(22,7\right)^2 \approx$$

$$\approx 124,71;$$

$$\sigma_{(X/Y=4)} = \sqrt{D_{(X/Y=4)}} = \sqrt{124,71} \approx 11,17.$$

Приклад 7.2. У 20 рейсах при різних погодних умовах здійснювались вимірювання максимальної швидкості і висоти польоту. Відхилення від розрахункових (у м/с і відповідно в м) наведено в таблиці:

i	1	2	3	4	5	6	7	8	9	10
X	-10	-2	4	10	-1	-16	-8	-2	6	8
Y	-8	-10	22	55	2	-30	-15	5	10	18
Продовження табл.										
i	11	12	13	14	15	16	17	18	19	20
X	-1	4	12	20	-11	2	14	6	-12	1
Y	3	-2	28	62	-10	-8	22	3	-32	8

Скласти інтервальну кореляційну таблицю двовимірного розподілу взявши орієнтовну кількість $m=5$ частинних інтервалів в інтервальному статистичному розподілі системи (X, Y). Знайти точкові оцінки математичного сподівання, дисперсії, кореляційного моменту та коефіцієнта кореляції.

Розв'язання:

Випишемо різні значення варіант, які потрапили у вибірку, у порядку їх зростання. Дістанемо дискретний варіаційний ряд:

X	-16	-12	-11	-10	-8	-2	-1	1	2	4	6	8	10	12	14	20
Y	-32	-30	-15	-10	-8	-2	2	3	5	8	10	18	22	28	55	62

Частоти варіантів за ознакою X: -2; -1; 4 та 6 повторюються по 2 рази.

Частоти варіантів за ознакою Y: -10; -8; 3 та 22 повторюються по 2 рази.

Визначаємо за обсягом вибірки $n=20$ орієнтовну кількість $m=5$ частинних інтервалів в інтервальному статистичному розподілі. За формулами $h_x=\left(x_{\max}-x_{\min}\right)/m$ та $h_y=\left(y_{\max}-y_{\min}\right)/m$

обчислюємо крок інтервалів: $h_x=\left(20+16\right)/5=7,2$ та $h_y=\left(62+32\right)/5=18,8$.

Підсумуємо частоти варіант, які потрапили в кожний із частинних інтервалів, при цьому частоти варіант, які збіглися з межами інтервалів, поділимо порівну між суміжними інтервалами.

Тоді інтервальний статистичний розподіл вибірки можна подати у вигляді таблиці:

i	1	2	3	4	5
(x_{i-1}, x_i)	[-16;-8,8]	[-8,8;-1,6]	[-1,6;5,6]	[5,6;12,8]	[12,8;20]
n_i	4	3	6	5	2
(y_{i-1}, y_i)	[-32;13,2]	[-13,2;5,6]	[5,6;24,4]	[24,4;43,2]	[43,2;62]
n_i	3	9	5	1	2

Кореляційна таблиця двовимірного розподілу:

X \ Y	[-32;-13,2]	[-13,2;5,6]	[5,6;24,4]	[24,4;43,2]	[43,2;62]	$\sum_{j=1}^{5} n_{ij} = n_{i0}$
[-16;-8,8]	2	2	-	-	-	**4**
[-8,8;-1,6]	1	2	-	-	-	**3**
[-1,6;5,6]	-	4	2	-	-	**6**
[5,6;12,8]	-	1	2	1	1	**5**
[12,8;20]	-	-	1	-	1	**2**
$\sum_{j=1}^{5} n_{ij} = n_{0j}$	**3**	**9**	**5**	**1**	**2**	$\sum_{i=1}^{5}\sum_{j=1}^{5} n_{ij} = 20$

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} =$$

$$= \frac{1}{20} \cdot \left(-16 - 12 - 11 - 10 - 8 - 2 \cdot 2 - 1 \cdot 2 + 1 + 2 + 4 \cdot 2 + 6 \cdot 2 + 8 + 10 + 12 + 14 + 20\right) =$$

$$= 1,2.$$

$$\overline{y}=\frac{\sum_{i=1}^{n} y_i}{n}=$$

$$=\left(-32-30-15-10\cdot 2-8\cdot 2-2+2+3\cdot 2+5+8+10+18+22\cdot 2+28+55+62\right)\times$$

$$\times\frac{1}{20}=6{,}15.$$

$$S_x^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2=$$

$$=[\left(-16-1{,}2\right)^2+\left(-12-1{,}2\right)^2+\left(-11-1{,}2\right)^2+\left(-10-1{,}2\right)^2+\left(-8-1{,}2\right)^2+$$

$$+\left(-2-1{,}2\right)^2+\left(-1-1{,}2\right)^2\cdot 3+\left(1-1{,}2\right)^2+\left(2-1{,}2\right)^2+\left(4-1{,}2\right)^2\cdot 2+$$

$$+\left(6-1{,}2\right)^2\cdot 2+\left(8-1{,}2\right)^2+\left(10-1{,}2\right)^2+\left(12-1{,}2\right)^2+\left(14-1{,}2\right)^2+$$

$$+\left(20-1{,}2\right)^2]\cdot 1/19=88{,}38.$$

$$S_y^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2=$$

$$=[\left(-32-6{,}15\right)^2+\left(-30-6{,}15\right)^2+\left(-15-6{,}15\right)^2+\left(-10-6{,}15\right)^2\cdot 2+$$

$$+\left(-2-6{,}15\right)^2+\left(2-6{,}15\right)^2+\left(3-6{,}15\right)^2\cdot 2+\left(5-6{,}15\right)^2+\left(8-6{,}15\right)^2+$$

$$+\left(10-6{,}15\right)^2+\left(18-6{,}15\right)^2+\left(22-6{,}15\right)^2\cdot 2+\left(28-6{,}15\right)^2+\left(-8-6{,}15\right)^2\cdot 2+$$

$$+\left(55-6{,}15\right)^2+\left(62-6{,}15\right)^2]\cdot 1/19=572{,}66.$$

$$K_{xy}^*=\frac{\sum_{j=1}^{m}\sum_{i=1}^{k}\left(x_j-\overline{x}\right)\cdot\left(y_i-\overline{y}\right)}{n-1}=$$

$$=\frac{1}{19}[\left(-10-1{,}2\right)\left(-8-6{,}15\right)+\left(-2-1{,}2\right)\left(-10-6{,}15\right)+\left(4-1{,}2\right)\left(22-6{,}15\right)+$$

$$+\left(10-1{,}2\right)\left(55-6{,}15\right)+\left(-1-1{,}2\right)\left(2-6{,}15\right)+\left(-16-1{,}2\right)\left(-30-6{,}15\right)+$$

$$+\left(-8-1{,}2\right)\left(-15-6{,}15\right)+\left(-2-1{,}2\right)\left(5-6{,}15\right)+\left(6-1{,}2\right)\left(10-6{,}15\right)+$$

$$+\left(8-1{,}2\right)\left(18-6{,}15\right)+\left(-1-1{,}2\right)\left(3-6{,}15\right)+\left(4-1{,}2\right)\left(-2-6{,}15\right)+$$

$$+\left(12-1{,}2\right)\left(28-6{,}15\right)+\left(20-1{,}2\right)\left(62-6{,}15\right)+\left(-11-1{,}2\right)\left(-10-6{,}15\right)+$$

$$+\left(2-1{,}2\right)\left(-8-6{,}15\right)+\left(14-1{,}2\right)\left(22-6{,}15\right)+\left(6-1{,}2\right)\left(3-6{,}15\right)+$$

$$+\left(-12-1{,}2\right)\left(-32-6{,}15\right)+\left(1-1{,}2\right)\left(8-6{,}15\right)]=197{,}86.$$

Для обчислення коефіцієнта кореляції застосуємо формулу:

$$r_B = \frac{K_{xy}^*}{\sqrt{S_x^2}\cdot\sqrt{S_y^2}} = \frac{197,86}{\sqrt{88,38}\cdot\sqrt{572,66}} = 0,88\,.$$

Із значення r_B робимо висновок, що ознаки *X* і *Y* скорельовані і мають майже лінійну залежність.

Приклад 7.3. Залежність кількості масла y_i, що його споживає певна особа за місяць, від її прибутку в гривнях x_i наведена у таблиці:

y_i, *грн.*	10,5	15,8	17,8	19,5	20,4	21,5	22,2	24,3	25,3	26,5
x_i, *грн.*	70	75	82	89	95	100	105	110	115	120
										Продовження табл.
y_i, *грн.*	28,1	30,1	35,2	36,4	37	38,5	39,5	40,5	41	42,5
x_i, *грн.*	125	130	135	140	145	150	155	160	165	170

Обчислити K_{xy}^*, r_B.

Розв'язання: Оскільки обсяг вибірки $n = 20$, то маємо:

$$\overline{x} = \frac{\sum_{i=1}^{20} x_i}{n} = \frac{70+75+82+89+95+100+105+110+115+120+125+}{20}$$

$$\frac{+130+135+140+145+150+155+160+165+170}{20} = 121,8.$$

$$D_x = \frac{\sum_{i=1}^{20} x_i^2}{n} - \left(\overline{x}\right)^2 = \frac{70^2+75^2+82^2+89^2+95^2+100^2+105^2+110^2+115^2+}{20}$$

$$\frac{+120^2+125^2+130^2+135^2+140^2+145^2+150^2+155^2+160^2+165^2+170^2}{20} -$$

$$-\left(121,8\right)^2 = 893,26.$$

$$\sigma_x = \sqrt{D_x} = \sqrt{893,26} = 29,89\,.$$

$$\overline{y} = \frac{\sum_{i=1}^{20} y_i}{n} = \frac{10,5 + 15,8 + 17,8 + 19,5 + 20,4 + 21,5 + 22,2 + 24,3 + 25,3 +}{20}$$

$$\frac{+26,5 + 28,1 + 30,1 + 35,2 + 36,4 + 37 + 38,5 + 39,5 + 40,5 + 41 + 42,5}{20} = 28,63.$$

$$D_y = \frac{\sum_{i=1}^{20} y_i^2}{n} - \left(\overline{y}\right)^2 = \frac{10,5^2 + 15,8^2 + 17,8^2 + 19,5^2 + 20,4^2 + 21,5^2 + 22,2^2 +}{20}$$

$$\frac{+24,3^2 + 25,3^2 + 26,5^2 + 28,1^2 + 30,1^2 + 35,2^2 + 36,4^2 + 37^2 + 38,5^2 + 39,5^2 +}{20}$$

$$\frac{+40,5^2 + 41^2 + 42,5^2}{20} - \left(28,63\right)^2 = 88,3.$$

$$\sigma_y = \sqrt{D_y} = \sqrt{88,3} \approx 9,4.$$

$$K_{xy}^* = \frac{\sum_{i=1}^{20} y_i x_i}{n} - \overline{x} \cdot \overline{y} = \frac{10,5 \cdot 70 + 15,8 \cdot 75 + 17,8 \cdot 82 + 19,5 \cdot 89 + 20,4 \cdot 95 +}{20}$$

$$\frac{+21,5 \cdot 100 + 22,2 \cdot 105 + 24,3 \cdot 110 + 25,3 \cdot 115 + 26,5 \cdot 120 + 28,1 \cdot 125 + 30,1 \cdot 130 +}{20}$$

$$\frac{+35,2 \cdot 135 + 36,4 \cdot 140 + 37 \cdot 145 + 38,5 \cdot 150 + 39,5 \cdot 155 + 40,5 \cdot 160 + 41 \cdot 165 +}{20}$$

$$\frac{+42,5 \cdot 170}{20} - 121,8 \cdot 28,63 = 278.$$

$$r_B = \frac{K_{xy}^*}{\sigma_x \sigma_y} = \frac{278}{29,89 \cdot 9,4} \approx 0,989.$$

Оскільки значення r_B близьке до одиниці, то звідси випливає, що залежність між кількістю масла, споживаного певною особою, та її місячним прибутком майже функціональна.

Приклад 7.4. Обчислити вибірковий коефіцієнт кореляції, знайти рівняння вибіркової лінійної регресії Y на X, та побудувати діаграму розсіювання за вибірковими даними:

x_i	7	8	5	3	7
y_i	1	2	3	1	3

Розв'язання: 1-ший спосіб. Знаходимо числові характеристики:

$$\overline{x} = \frac{1}{5}\left(7+8+5+3+7\right) = 6; \quad \overline{y} = \frac{1}{5}\left(1+2+3+1+3\right) = 2;$$

$$D_x = \frac{1}{5}\left(7^2+8^2+5^2+3^2+7^2\right) - 6^2 = 3,2; \quad D_y = \frac{1}{5}\left(1^2+2^2+3^2+1^2+3^2\right) - 2^2 = 0,8;$$

$$K_{xy}^* = \frac{1}{5}\left(7+16+15+3+21\right) - 12 = 0,4; \quad r_B = \frac{0,4}{\sqrt{3,2}\cdot\sqrt{0,8}} = 0,25;$$

$$a = 0,25\cdot\frac{\sqrt{0,8}}{\sqrt{3,2}} = 0,125; \quad b = 2 - 0,125\cdot 6 = 1,25.$$

Отже, $y = 0,125\cdot x + 1,25$ - рівняння вибіркової лінійної регресії Y на X .

2-гий спосіб. Запишемо нормальну систему методу найменших квадратів. Для цього знайдемо суми:

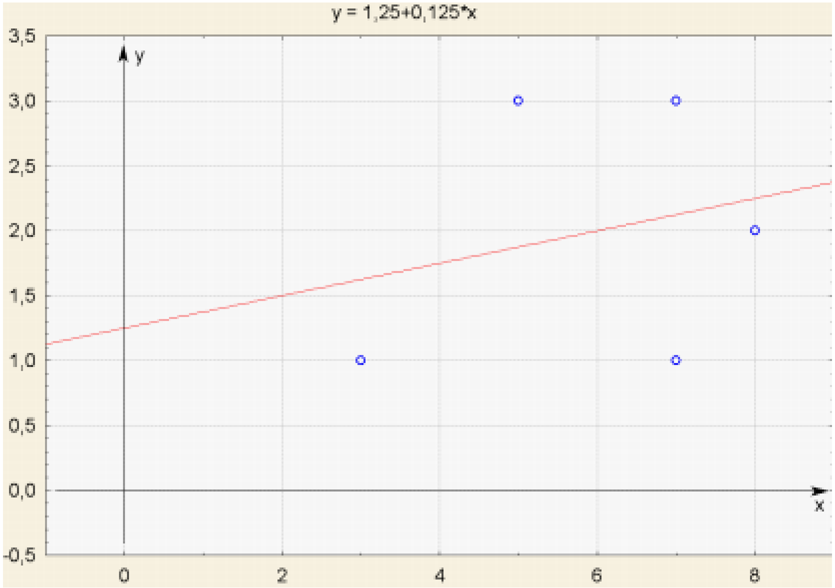
$$\sum_{i=1}^{n} x_i = 30; \quad \sum_{i=1}^{n} x_i^2 = 196; \quad \sum_{i=1}^{n} y_i = 10; \quad \sum_{i=1}^{n} x_i y_i = 62.$$

Маємо:

$$\begin{cases} 196a + 30b = 62 \\ 30a + 5b = 10 \end{cases} \Rightarrow \begin{cases} b = 2 - 6a \\ 196a + 60 - 180a = 62 \end{cases} \Rightarrow \begin{cases} b = 2 - 6a \\ 16a = 2 \end{cases} \Rightarrow \begin{cases} a = 0,125 \\ b = 1,25 \end{cases}.$$

Отже, $y = 0,125\cdot x + 1,25$.

Діаграма розсіювання має вигляд:



7.6. Побудова довірчого інтервалу для коефіцієнта кореляції r_{xy} генеральної сукупності із заданою надійністю γ

Точковою незміщеною статистичною оцінкою для теоретичного коефіцієнта кореляції r_{xy} є вибірковий коефіцієнт кореляції r_B з виправленим середнім квадратичним відхиленням $S = \dfrac{1-r_B^2}{\sqrt{n}}$.

Якщо центрувати і нормувати випадкову величину r_B, то отримаємо величину

$$x_\gamma = \frac{r_B - r_{xy}}{\sigma(r_B)} = \frac{r_B - r_{xy}}{\dfrac{1-r_B^2}{\sqrt{n}}}, \qquad (7.24)$$

що має нормований нормальний закон розподілу $N(0;1)$.

Скориставшись (7.24), дістанемо

$$P\left(\left|\frac{r_B - r_{xy}}{\dfrac{1-r_B^2}{\sqrt{n}}}\right| < x_\gamma\right) = P\left(r_B - x_\gamma \frac{1-r_B^2}{\sqrt{n}} < r_{xy} < r_B + x_\gamma \frac{1-r_B^2}{\sqrt{n}}\right) = \gamma = 2\Phi(x_\gamma).$$

Отже, довірчий інтервал для r_{xy} буде таким:

$$r_B - x_\gamma \frac{1-r_B^2}{\sqrt{n}} < r_{xy} < r_B + x_\gamma \frac{1-r_B^2}{\sqrt{n}}, \quad (7.25)$$

де x_γ знаходимо за таблицею значень Лапласа (табл. 2)

$$\Phi\left(x_\gamma\right) = 0,5 \cdot \gamma \quad (7.26)$$

Приклад 7.5. Побудувати довірчий інтервал з надійністю $\gamma = 0,99$ для коефіцієнта кореляції r_{xy} за двовимірним статистичним розподілом вибірки

$Y = y_i$	$X = x_j$				
	10	20	30	40	n_{y_i}
2	-	2	4	4	10
4	10	8	6	6	30
6	5	10	5	-	20
8	15	-	15	10	40
n_{x_j}	30	20	30	20	-

Розв'язання: Для обчислення K_{xy}^*, r_B визначимо $\overline{x}, \sigma_x, \overline{y}, \sigma_y$:

Оскільки $n = \sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij} = 100$, то

$$\overline{x}=\frac{\sum_{j=1}^{m}x_j n_{x_j}}{n}=\frac{10\cdot 30+20\cdot 20+30\cdot 30+40\cdot 20}{100}=\frac{2400}{100}=24;$$

$$D_x=\frac{\sum_{j=1}^{m}x_j^2 n_{x_j}}{n}-\left(\overline{x}\right)^2=\frac{10^2\cdot 30+20^2\cdot 20+30^2\cdot 30+40^2\cdot 20}{100}-24^2=700-576=124;$$

$$\sigma_x=\sqrt{D_x}=\sqrt{124}\approx 11{,}14.$$

$$\overline{y}=\frac{\sum_{i=1}^{k}y_i n_{y_i}}{n}=\frac{2\cdot 10+4\cdot 30+6\cdot 20+8\cdot 40}{100}=\frac{580}{100}=5{,}8;$$

$$D_y=\frac{\sum_{i=1}^{k}y_i^2 n_{y_i}}{n}-\left(\overline{y}\right)^2=\frac{2^2\cdot 10+4^2\cdot 30+6^2\cdot 20+8^2\cdot 40}{100}-5{,}8^2=38-33{,}64=4{,}36;$$

$$\sigma_y=\sqrt{D_y}=\sqrt{4{,}36}\approx 2{,}1.$$

Для визначення кореляційного моменту K_{xy}^* обчислюють

$$\sum_{i=1}^{k}\sum_{j=1}^{m}y_i x_j n_{ij}=2\cdot 10\cdot 0+2\cdot 20\cdot 2+2\cdot 30\cdot 4+2\cdot 40\cdot 4+4\cdot 10\cdot 10+4\cdot 20\cdot 8+4\cdot 30\cdot 6+$$
$$+4\cdot 40\cdot 6+6\cdot 10\cdot 5+6\cdot 20\cdot 10+6\cdot 30\cdot 5+6\cdot 40\cdot 0+8\cdot 10\cdot 15+8\cdot 20\cdot 0+8\cdot 30\cdot 15+$$
$$+8\cdot 40\cdot 10=13760.$$

$$K_{xy}^*=\frac{\sum_{i=1}^{k}\sum_{j=1}^{m}y_i x_j n_{ij}}{n}-\overline{x}\cdot\overline{y}=\frac{13760}{100}-24\cdot 5{,}8=-1{,}6.$$

Це свідчить про те, що між ознаками X і Y існує від'ємний кореляційний зв'язок.

Для вимірювання тісноти цього зв'язку обчислимо вибірковий коефіцієнт кореляції.

$$r_B=\frac{K_{xy}^*}{\sigma_x\sigma_y}=\frac{-1{,}6}{11{,}14\cdot 2{,}1}\approx -0{,}068.$$

Отже, $r_B=-0{,}068$, тобто тіснота кореляційного зв'язку між знаками X і Y є слабкою.

Запишемо рівняння для знаходження x_γ: $\Phi\left(x_\gamma\right)=\frac{\gamma}{2}=\frac{0,99}{2}=0,495$

Тоді, згідно табл. 2 знаходимо $x_\gamma=2,58$. Підставимо r_B, x_γ та $\sqrt{n}=\sqrt{100}=10$

в формулу (7.25) $r_B-x_\gamma\frac{1-r_B^2}{\sqrt{n}}<r_{xy}<r_B+x_\gamma\frac{1-r_B^2}{\sqrt{n}}$:

$$-0,068-2,58\cdot\frac{1-\left(-0,068\right)^2}{10}<r_{xy}<-0,068+2,58\cdot\frac{1-\left(-0,068\right)^2}{10},$$
$$-0,325<r_{xy}<0,189.$$

Отже, з надійністю $\gamma=0,99$ отримали, що $r_{xy}\in\left(-0,325;\,0,189\right)$.