Практична робота N 2

Тема: *„Двовимірний статистичний розподіл вибірки та його числові характеристики. Парний статистичний розподіл . Поняття про парну регресію."*

Основні теоретичні положення

1. Двовимірний статистичний розподіл та його числові характеристики.

Перелік варіант $Y = y_i$, $X = x_j$ та відповідних їм частот n_{ij} спільної їх появи (тобто пара (x_j, y_i) з'являється у вибірці n_{ij} разів) утворюють ***двовимірний статистичний розподіл вибірки***, що реалізована з генеральної сукупності. Елементам цієї вибірки притаманні кількісні ознаки *X* і *Y*. У табличній формі цей розподіл має такий вигляд:

Таблиця 1

$Y = y_i$	$X = x_j$						
	x_1	x_2	…	x_j	…	x_m	n_{y_i}
y_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1m}	n_{y_1}
y_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2m}	n_{y_2}
…	…	…	…	…	…	…	…
y_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{im}	n_{y_i}
…	…	…	…	…	…	…	…
y_k	n_{k1}	n_{k2}	…	n_{kj}	…	n_{km}	n_{y_k}
n_{x_j}	n_{x_1}	n_{x_2}	…	n_{x_j}	…	n_{x_m}	---

У цій таблиці

$$n_{x_j} = \sum_{i=1}^{k} n_{ij} = n_{1j} + n_{2j} + \ldots + n_{kj}, \quad \textbf{(1)}$$

(кожне число у нижньому рядку таблиці 1 є сумою частот відповідного стовпця);

$$n_{y_i} = \sum_{j=1}^{m} n_{ij} = n_{i1} + n_{i2} + \ldots + n_{im} \quad \textbf{(2)}$$

(кожне число в останньому стовпці таблиці 1 є сумою частот відповідного рядка);

$$\sum_{j=1}^{m} n_{x_j} = \sum_{i=1}^{k} n_{y_i} = n, \quad \textbf{(3)}$$

де n - обсяг вибірки.

Для двовимірного статистичного розподілу використовують такі числові характеристики:

Відносно x	*Відносно* y
$\bar{x} = \frac{\sum_{j=1}^{m} x_j n_{x_j}}{n}$ **(4)**	$\bar{y} = \frac{\sum_{i=1}^{k} y_i n_{y_i}}{n}$ **(5)**
$D_x = \frac{\sum_{j=1}^{m} x_j^2 n_{x_j}}{n} - \left(\bar{x}\right)^2$ **(6)**	$D_y = \frac{\sum_{i=1}^{k} y_i^2 n_{y_i}}{n} - \left(\bar{y}\right)^2$ **(7)**
$\sigma_x = \sqrt{D_x}$ **(8)**	$\sigma_y = \sqrt{D_y}$ **(9)**

Формулами (4) та (5) даються середні значення, (6) та (7) – дисперсії, а (8) і (9) – середньоквадратичні відхилення x та y відповідно.

Для характеристики наявності та щільності лінійного зв'язку між ознаками двовимірного статистичного розподілу використовують наступні числові характеристики:

$$K_{xy}^* = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m} y_i x_j n_{ij}}{n} - \bar{x} \cdot \bar{y} \qquad \textbf{(10)}$$

та

$$r_B = \frac{K_{xy}^*}{\sigma_x \sigma_y}, \qquad \textbf{(11)}$$

Числова характеристика K_{xy}^* називається ***кореляційним моментом*** і вказує на наявність ($K_{xy}^* \neq 0$) чи відсутність ($K_{xy}^* = 0$) кореляційного зв'язку між ознаками X та Y.

Числова характеристика r_B називається ***вибірковим коефіцієнтом кореляції***. Вона вказує на щільність лінійного зв'язку між ознаками X та Y. Для r_B справджується нерівність

$$-1 \le r_B \le 1, \qquad \textbf{(12)}$$

Якщо r_B=0, то між X та Y відсутній лінійний зв'язок (але може бути інша форма залежності (див рис.1)).

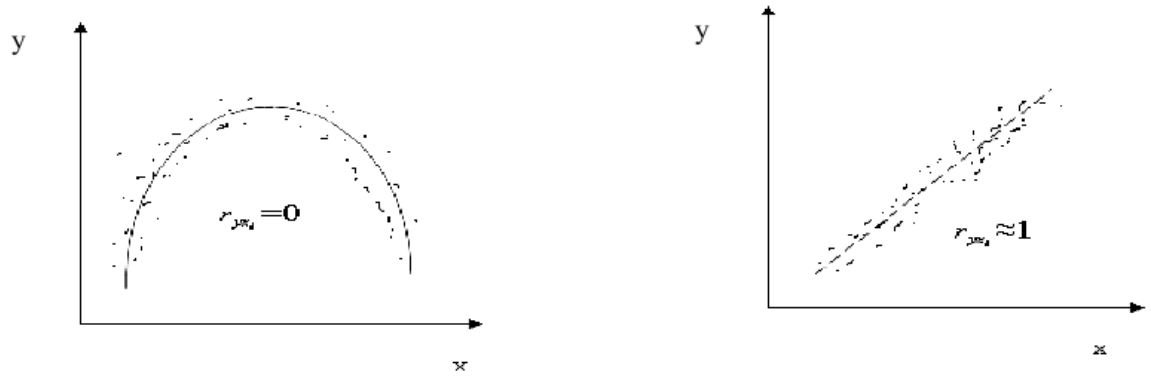


Рис. 1. *Залежність коефіцієнта кореляції від форми зв'язку*

Якщо $r_B > 0$, то зв'язок ознаками X та Y прямий, тобто зі зростанням однієї ознаки збільшується й значення іншої ($x \uparrow\uparrow y$), а в разі $r_B < 0$ лінійний зв'язок ознаками X та Y обернений (зворотній), тобто зі зростанням однієї ознаки значення іншої зменшується ($x \uparrow\downarrow y$). Сила (щільність) лінійного зв'язку встановлюється за шкалою Чеддока (табл. 2), що ранжує модуль r_B.

Таблиця 2

$\lvert r_B \rvert$	Якісна оцінка сили зв'язку
0	**Відсутній**
0,1 - 0,3	**Слабкий**
0,3 - 0,5	**Помірний**
0,5 - 0,7	**Помітний**
0,7 - 0,9	**Високий**
0,9 - 0,99	**Дуже високий (майже функціональний)**
1	**Функціональний**

Умовним статистичним розподілом $Y / X = x_j$ ознаки Y при фіксованому значенні $X = x_j$ називається перелік варіант ознаки $Y = y_i$ та відповідних їм частот n_{ij} $(1 \le i \le k)$, взятих при фіксованому значенні $X = x_j$.

$Y = y_i$	$X = x_j$						
	x_1	x_2	…	x_j	…	x_m	n_{y_i}
y_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1m}	n_{y_1}
y_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2m}	n_{y_2}
…	…	…	…	…	…	…	…
y_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{im}	n_{y_i}
…	…	…	…	…	…	…	…
y_k	n_{k1}	n_{k2}	…	n_{kj}	…	n_{km}	n_{y_k}
n_{x_j}	n_{x_1}	n_{x_2}	…	n_{x_j}	…	n_{x_m}	—

Для табличного зображення умовного статистичного розподілу $Y / X = x_j$ ми маємо з таблиці 1 взяти 1-й та j-й стовпці (для зручності запишемо їх у такій формі, як записувався дискретний розподіл):

Таблиця 3.1

$Y = y_i$	y_1	y_2	…	y_i	…	y_k	$\sum$
n_{ij}	n_{1j}	n_{2j}	…	n_{ij}	…	n_{kj}	n_{x_j}

Умовним статистичним розподілом $X / Y = y_i$ ознаки X при фіксованому значенні $Y = y_i$ називається перелік варіант ознаки $X = x_j$ та відповідних їм частот n_{ij} $(1 \le j \le m)$, взятих при фіксованому значенні $Y = y_i$.

$Y = y_i$	$X = x_j$						
	x_1	x_2	…	x_j	…	x_m	n_{y_i}
y_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1m}	n_{y_1}
y_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2m}	n_{y_2}
…	…	…	…	…	…	…	…
y_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{im}	n_{y_i}
…	…	…	…	…	…	…	…
y_k	n_{k1}	n_{k2}	…	n_{kj}	…	n_{km}	n_{y_k}
n_{x_j}	n_{x_1}	n_{x_2}	…	n_{x_j}	…	n_{x_m}	—

Табличний вигляд умовного статистичного розподілу $X / Y = y_i$ наведений нижче (табл. 3.2). Для його побудови з таблиці 1 взято рядок значень ознаки X (2-й рядок) та i-й рядок (що відповідає значенню $Y = y_i$):

Таблиця 3.2

$X = x_j$	x_1	x_2	…	x_j	…	x_m	$\sum$
n_{ij}	n_{i1}	n_{i2}	…	n_{ij}	…	n_{im}	n_{y_i}

Числові характеристики умовних розподілів

$Y / X = x_j$	$X / Y = y_i$
$\bar{y}_{X=x_j} = \dfrac{\sum_{i=1}^{k} y_i n_{ij}}{n_{x_j}}$ **(13)**	$\bar{x}_{Y=y_i} = \dfrac{\sum_{j=1}^{m} x_j n_{ij}}{n_{y_i}}$ **(14)**
$D(Y / X = x_j) = \dfrac{\sum_{i=1}^{k} y_i^2 n_{ij}}{n_{x_j}} - \left(\bar{y}_{X=x_j}\right)^2$ **(15)**	$D(X / Y = y_i) = \dfrac{\sum_{j=1}^{m} x_j^2 n_{ij}}{n_{y_i}} - \left(\bar{x}_{Y=y_i}\right)^2$ **(16)**
$\sigma(Y / X = x_j) = \sqrt{D(Y / X = x_j)}$ **(17)**	$\sigma(X / Y = y_i) = \sqrt{D(X / Y = y_i)}$ **(18)**

Відповідні середні, дисперсії та середньоквадратичні відхилення називаються ***умовними***.

2. Парний статистичний розподіл та його числові характеристики.

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

Таблиця 4

$X = x_j$	x_1	x_2	…	x_j	…	x_n
$Y = y_i$	y_1	y_2	…	y_j	…	y_n

(або рядок Y пишеться над рядком X чи записуються дані у стовпчик). 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 парного розподілу отримуються з відповідних характеристик двовимірного розподілу:

Відносно x	*Відносно* y
$\bar{x} = \dfrac{\sum_{j=1}^{m} x_j}{n}$ **(19)**	$\bar{y} = \dfrac{\sum_{i=1}^{k} y_i}{n}$ **(20)**
$D_x = \dfrac{\sum_{j=1}^{m} x_j^2}{n} - \left(\bar{x}\right)^2 = \overline{x^2} - \left(\bar{x}\right)^2$ **(21)**	$D_y = \dfrac{\sum_{i=1}^{k} y_i^2}{n} - \left(\bar{y}\right)^2 = \overline{y^2} - \left(\bar{y}\right)^2$ **(22)**
$\sigma_x = \sqrt{D_x}$ **(23)**	$\sigma_y = \sqrt{D_y}$ **(24)**

Аналогічно формула для кореляційного моменту парного розподілу набуває вигляду

$$K_{xy}^{*}=\frac{\sum_{i=1}^{k}\sum_{j=1}^{m}y_i x_j}{n}-\bar{x}\cdot\bar{y}, \qquad \textbf{(25)}$$

а формула для вибіркового коефіцієнта кореляції залишається без змін.

3. Поняття про парну регресію.

Нагадаємо, що ***статистичною*** називають залежність, коли зі зміною однієї випадкової величини змінюється закон розподілу ймовірностей іншої, тобто певному значенню $X=x_i$ відповідає не одне значення змінної Y, а певний статистичний розподіл цієї змінної. Зокрема, статистична залежність виявляється в тому, що зі зміною однієї величини змінюється середнє значення іншої. Така залежність називається ***кореляційною***. Отже, кореляційною залежністю ознаки Y по X називається функціональна залежність середнього значення ознаки Y від X:

$$\bar{y}=\alpha(x).$$

Залежність середнього значення від іншої випадкової величини зображується за допомогою умовного математичного сподівання (умовного середнього). Таку залежність можна виразити співвідношенням

$$M(Y/X)=f(X) \qquad \textbf{(26)}$$

де $M(Y/X)$ — умовне математичне сподівання (умовне середнє).

Функція $f(x)$ називається ***функцією регресії*** Y на X. При цьому X називається ***незалежною (пояснюючою)*** змінною (***регресором, фактором***), Y — ***залежною (пояснюваною)*** змінною (***регресантом, показником***).

Термін "регресія" (рух назад, повернення до попереднього стану) увів Френсіс Галтон наприкінці XIX ст., проаналізувавши залежність між зростом батьків і зростом дітей. Він помітив, що зріст дітей у дуже високих батьків у середньому менший, ніж середній зріст батьків.

У дуже низьких батьків, навпаки, середній зріст дітей вищий. В обох випадках середній зріст дітей прямує (повертається) до середнього зросту людей у даному регіоні. Звідси й вибір терміна, що відбиває таку залежність.

Отже, під терміном «регресія» розуміється функціональна залежність між умовним математичним сподіванням (умовним середнім) випадкової величини Y від значень пояснювальної змінної X.

Проте реальні значення залежної змінної не завжди збігаються з її умовним математичним сподіванням (умовним середнім) і при одному і тому ж значенні пояснювальної змінної значення залежної змінної Y можуть бути різними внаслідок впливу випадкових факторів. Тому аналітична залежність (у вигляді функції $Y=f(X)$) має бути доповнена випадковою складовою ε, що відображає вплив на результативний показник всіх неврахованих факторів. Тоді з їх врахуванням залежність (26) необхідно записати так:

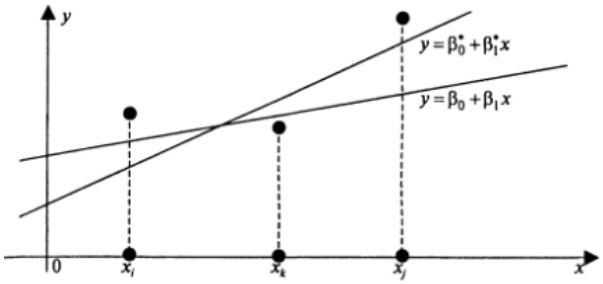
$$Y=M(Y/X)+\varepsilon=f(X)+\varepsilon, \qquad \textbf{(27)}$$

Співвідношення (27) називають ***регресійною моделлю***. Випадкову величину ε назвемо ***збуренням (залишком, відхиленням)***. Її значення можуть змінюватися від одного спостереження до іншого. Наприклад, при вивченні залежності національного доходу від капітальних вкладень збурююча змінна включала би в себе вплив на національний дохід таких факторів, як число працюючих у сфері виробництва, продуктивність праці, використання основних фондів і т. д., а також інші випадкові чинники.

Розглянемо найпростіший випадок, коли $f(X)=\beta_0+\beta_1 X$ - лінійна функція. Тоді (27) набуде вигляду

$$Y=M(Y/X)=\beta_0+\beta_1 X+\varepsilon \qquad \textbf{(28)}$$

Співвідношення (28) називають ***теоретичною лінійною регресійною моделлю***, а β_0, β_1 – ***теоретичними параметрами (коефіцієнтами)*** регресії**.** Щоб визначити значення теоретичних коефіцієнтів регресії, необхідно знати й використовувати всі значення змінних X і Y генеральної сукупності, що практично неможливо. Отже, постає задача, щоб за наявності статистичних даних (x_i, y_i), $i = \overline{1,n}$, одержаних шляхом реалізації вибірки обсягом $n \ll N$ із генеральної сукупності, визначити найкращі статистичні оцінки β_0^*, β_1^* для невідомих теоретичних параметрів (коефіцієнтів) β_0, β_1. Отже, нам необхідно побудувати так зване ***емпіричне рівняння*** на базі інформації, одержаної із вибірки.



Емпіричне рівняння регресії має вигляд

$$\hat{Y} = \beta_0^* + \beta_1^* X + e_i \qquad \textbf{(29)}$$

де β_0^*, β_1^* — оцінки невідомих параметрів β_0, β_1, e_i – статистична оцінка ε_i. Для спрощення сприйняття доданок у e_i емпіричному рівнянні регресії зазвичай не записують.

Через розбіжність статистичної бази для генеральної сукупності та вибірки оцінки β_0^*, β_1^* практично завжди відрізняються від дійсних значень коефіцієнтів β_0, β_1, що призводить до розбіжності емпіричної та теоретичної ліній регресії.

Побудова регресії – це найпростіший метод отримати математичну модель, прослідкувати за тенденцією та зробити прогноз для різних економічних та соціальних процесів.

Оцінки β_0^*, β_1^* моделі (29) знаходяться ***методом найменших квадратів (***мінімізуються квадрати відхилень $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \to \min$***)***, який дає наступні формули для їх обчислення:

$$\beta_1^* = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \left(\overline{x}\right)^2} = \frac{K_{xy}^*}{D_x} \qquad \textbf{(30)}$$

$$\beta_0^* = \overline{y} - \beta_1^* \overline{x} \qquad \textbf{(31)}$$

Лінія регресії проходить через точку $(\overline{X}, \overline{Y})$. *Коефіцієнт* β_1^* *є кутовим коефіцієнтом лінії регресії. Він показує, на скільки одиниць в середньому змінюється показник* Y *при збільшенні фактора* X *на одну одиницю.*

Стала β_0^* *дає прогнозне значення залежної змінної при* $x = 0$. *Але вона не завжди має конкретний зміст, оскільки "прогнозування назад" не завжди спрацьовує.*

Для оцінки якості регресії використовується ***коефіцієнт детермінації*** R^2. Для спрощення сприйняття ми не вдаватимемось у його глибинну суть і не даватимемо його строгого означення, а лише скористаємось його зв'язком з коефіцієнтом вибіркової кореляції

$$R^2 = r_B^2 \qquad \textbf{(32)}$$

Коефіцієнт детермінації задовольняє нерівність

$$0 \le R^2 \le 1$$

і, помножений на 100, *показує, на скільки відсотків зміна показника* Y *пояснюється зміною фактора* X. Зрозуміло, що регресія вважається тим кращою, чим ближчий R^2 до 1. Якщо $R^2 < 0,75$, то модель вважається непридатною для прогнозування і слід шукати або іншу форму залежності, або включати в модель інші фактории.

Зразок розв'язання типового варіанту

Задача 1. *Двовимірний статистичний закон розподілу задано таблицею (де k - порядковий номер студента у списку академгрупи ($1 \le k \le 16$)) :*

Таблиця 5

$Y = y_i$	$X = x_j$			n_{y_i}
	k+1	***k+2***	***k+4***	
k	0	0	k+5	
k+5	0	18- k	2	
k+10	k+1	0	16-k	
k+15	4	2	2	
n_{x_j}				

Завдання:

1) Записати двовимірний статистичний закон розподілу, що відповідає Вашому варіанту (тобто записати таблицю 5, підставивши конкретне значення k і заповнити останній рядок n_{x_j} та останній стовпець n_{y_i}).

2) Знайти числові характеристики отриманого двовимірного розподілу:

$$\overline{x},\ D_x,\ \sigma_x,\ \overline{y},\ D_y,\ \sigma_y\ ,\ K^*_{xy},\ r_B\ .$$

3) Скласти та записати умовні закони розподілу

$$Y / X = k + 4 \text{ та } X / Y = k + 10$$

і знайти їхні числові характеристики:

$$\overline{y}_{X=k+4},\ D(Y / X = k + 4),\ \sigma(Y / X = k + 4),$$
$$\overline{x}_{Y=k+10},\ D(X / Y = k + 10),\ \sigma(X / Y = k + 10).$$

Розв'язання. Розв'яжемо дане завдання при $k = 0$.

1) *Запишемо двовимірний статистичний закон розподілу, що відповідає значенню $k = 0$, тобто запишемо таблицю 5, підставивши $k = 0$ і заповнимо останній рядок n_{x_j} та останній стовпець n_{y_i} :*

Таблиця 6

$Y = y_i$	$X = x_j$			n_{y_i}
	1	***2***	***4***	
0	0	0	5	**5**
5	0	18	2	**20**
10	1	0	16	**17**
15	4	2	2	**8**
n_{x_j}	**5**	**20**	**25**	**50**

Для заповнення останнього рядка n_{x_j} додаємо усі елементи відповідного стовпця:

$$n_{x_1} = 0 + 0 + 1 + 4 = 5,$$
$$n_{x_2} = 0 + 18 + 0 + 2 = 20,$$
$$n_{x_3} = 5 + 2 + 16 + 2 = 25,$$
$$n_{x_1} + n_{x_2} + n_{x_3} = 5 + 20 + 25 = 50$$

Аналогічно, для заповнення останнього стовпця n_{y_i} додаємо усі елементи відповідного рядка

$$n_{y_1} = 0+0+5 = 5,$$
$$n_{y_2} = 0+18+2 = 20,$$
$$n_{y_3} = 1+0+16 = 17,$$
$$n_{y_4} = 4+2+2 = 8,$$
$$n_{y_1} + n_{y_2} + n_{y_3} + n_{y_4} = 5+20+17+8 = 50.$$

Таким чином,

$$n_{x_1} + n_{x_2} + n_{x_3} = n_{y_1} + n_{y_2} + n_{y_3} + n_{y_4} = 50 = n,$$

тобто обсяг вибірки $n = 50$.

2) Знайдемо числові характеристики отриманого двовимірного розподілу (табл. 6):

$$\bar{x},\ D_x,\ \sigma_x,\ \bar{y},\ D_y,\ \sigma_y.$$

Для зручності знаходження $\bar{x}$ випишемо окрему таблицю

$Y = y_i$	$X = x_j$			n_{y_i}
	1	2	4	
0	0	0	5	5
5	0	18	2	20
10	1	0	16	17
15	4	2	2	8
n_{x_j}	5	20	25	50

(другий та останній рядки таблиці 6):

Таблиця 6.1

$X = x_j$	1	2	4
n_{x_j}	5	20	25

Середнє $\bar{x}$ знаходимо за формулою (4). Це буде сума добутків відповідних елементів таблиці 6.1, поділена на обсяг вибірки $n = 50$:

$$\bar{x} = \frac{\sum_{j=1}^{m} x_j n_{x_j}}{n} = \frac{1\cdot 5 + 2\cdot 20 + 4\cdot 25}{50} = \frac{145}{50} = 2{,}9.$$

Дисперсію D_x знаходимо за формулою (6):

$$D_x = \frac{\sum_{j=1}^{m} x_j^2 n_{x_j}}{n} - \left(\bar{x}\right)^2 = \frac{1^2\cdot 5 + 2^2\cdot 20 + 4^2\cdot 25}{50} - (2{,}9)^2 = \frac{485}{50} - \left(\frac{29}{10}\right)^2 =$$
$$= \frac{97}{10} - \frac{841}{100} = \frac{970-841}{100} = 1{,}29.$$

За формулою (8) знаходимо середньоквадратичне відхилення σ_x:

$$\sigma_x = \sqrt{D_x} = \sqrt{1{,}29} \approx 1{,}14.$$

Далі знаходимо відповідні числові характеристики по змінній y. Запишемо таблицю

$Y=y_i$	$X=x_j$			n_{y_i}
	1	2	4	
0	0	0	5	5
5	0	18	2	20
10	1	0	16	17
15	4	2	2	8
n_{x_j}	5	20	25	50

(перший та останній стовпець таблиці 6, розміщені рядками):

Таблиця 6.2

$Y=y_i$	0	5	10	15
n_{y_i}	5	20	17	8

Середнє $\overline{y}$ знаходимо за формулою (5):

$$\overline{y}=\frac{\sum_{i=1}^{k} y_i n_{y_i}}{n}=\frac{0\cdot 5+5\cdot 20+10\cdot 17+15\cdot 8}{50}=\frac{390}{50}=7{,}8$$

Дисперсію D_y знаходимо за формулою (7):

$$D_y=\frac{\sum_{i=1}^{k} y_i^2 n_{y_i}}{n}-\left(\overline{y}\right)^2=\frac{0^2\cdot 5+5^2\cdot 20+10^2\cdot 17+15^2\cdot 8}{50}-(7{,}8)^2=$$

$$=\frac{4000}{50}-(7{,}8)^2=80-60{,}84=19{,}16$$

і середньоквадратичне відхилення σ_y знаходимо за формулою (9):

$$\sigma_y=\sqrt{D_y}=\sqrt{19{,}16}\approx 4{,}38.$$

Знайдемо кореляційний момент K_{xy}^* за формулою (10)

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

$$=\frac{0\cdot 1\cdot 0+0\cdot 2\cdot 0+0\cdot 4\cdot 5+5\cdot 1\cdot 0+5\cdot 2\cdot 18+5\cdot 4\cdot 2+10\cdot 1\cdot 1+10\cdot 2\cdot 0+10\cdot 4\cdot 16+15\cdot 1\cdot 4+15\cdot 2\cdot 2+15\cdot 4\cdot 2}{50}-$$

$$-2{,}9\cdot 7{,}8=\frac{1110}{50}-22{,}62=22{,}2-22{,}62=-0{,}42\neq 0$$

Оскільки $K_{xy}^*\neq 0$, то між ознаками X та Y наявний лінійний зв'язок. Щоб охарактеризувати щільність (тісноту) цього зв'язку, обчислимо вибірковий коефіцієнт кореляції r_B за формулою (11), врахувавши, що $K_{xy}^*=-0{,}42$, $\sigma_x\approx 1{,}14$, $\sigma_y\approx 4{,}38$ (обчислені вище):

$$r_B=\frac{K_{xy}^*}{\sigma_x\sigma_y}=\frac{-0{,}42}{1{,}14\cdot 4{,}38}=-\frac{0{,}42}{4{,}9932}\approx -0{,}08.$$

Інтерпретація результату: *оскільки $r_B<0$, то між X та Y зв'язок зворотній, тобто при збільшенні кількісного значення ознаки X, значення Y зменшується. Крім того,*

$$|r_B|=|-0{,}08|=0{,}08<0{,}1,$$

отже, згідно зі шкалою Чеддока (табл. 2) лінійний зв'язок майже відсутній (дуже слабкий).

3) Скласти та записати умовні закони розподілу

$$Y / X = k + 4 \text{ та } X / Y = k + 10$$

і знайти їхні числові характеристики:

$$\overline{y}_{X=k+4},\ D(Y / X = k + 4),\ \sigma(Y / X = k + 4),$$
$$\overline{x}_{Y=k+10},\ D(X / Y = k + 10),\ \sigma(X / Y = k + 10).$$

При $k = 0$ маємо умовні розподіли:

$$Y / X = k + 4 \overset{k=0}{\Rightarrow} Y / X = 4$$

і

$$X / Y = k + 10 \overset{k=0}{\Rightarrow} X / Y = 10.$$

Випишемо розподіл $Y / X = 4$ у табличному вигляді та знайдемо його числові характеристики. Для цього

$Y = y_i$	$X = x_j$			n_{y_i}
	1	2	4	
0	0	0	5	5
5	0	18	2	20
10	1	0	16	17
15	4	2	2	8
n_{x_j}	5	20	25	50

з таблиці 6 вибираємо перший стовпець (значень Y) та останній стовпець частот, що відповідають значенню $X = 4$. Для зручності запишемо отриману таблицю рядками

Таблиця 6.3

$Y = y_i$	0	5	10	15	n_{x_3}
n_{i3} $(j = 3)$	5	2	16	2	25

Далі обчислюємо числові характеристики умовного розподілу $Y / X = 4$, заданого таблицею 6.3, як це робиться звичайно. Умовне середнє $\overline{y}_{X=4}$ обчисляємо за формулою (13). При цьому $X = 4$ при $j = 3$ і $n_{x_3} = 25$. Маємо:

$$\overline{y}_{X=4} = \frac{\sum_{i=1}^{k} y_i n_{i3}}{n_{x_3}} = \frac{0 \cdot 5 + 5 \cdot 2 + 10 \cdot 16 + 15 \cdot 2}{25} = \frac{200}{25} = 8$$

Умовну дисперсію $D(Y / X = 4)$ знайдемо за формулою (15)

$$D(Y / X = 4) = \frac{\sum_{i=1}^{k} y_i^2 n_{i3}}{n_{x_3}} - \left(\overline{y}_{X=4}\right)^2 = \frac{0^2 \cdot 5 + 5^2 \cdot 2 + 10^2 \cdot 16 + 15^2 \cdot 2}{25} - 8^2 =$$

$$= \frac{2100}{25} - 64 = 84 - 64 = 20$$

і умовне середньоквадратичне відхилення $\sigma(Y / X = 4)$ - за формулою (17)

$$\sigma(Y / X = 4) = \sqrt{D(Y / X = 4)} = \sqrt{20} \approx 4{,}47.$$

Випишемо розподіл $X / Y = 10$ у табличному вигляді та знайдемо його числові характеристики. Для цього

$Y=y_i$	$X=x_j$			n_{y_i}
	1	2	4	
0	0	0	5	5
5	0	18	2	20
10	1	0	16	17
15	4	2	2	8
n_{x_j}	5	20	25	50

з таблиці 6 вибираємо рядок значень X та рядок частот, що відповідають значенню $Y=10$:

Таблиця 6.4

$X=x_j$	1	2	4	n_{y_3}
n_{3j} $(i=3)$	1	0	16	17

Умовне середнє $\overline{X}_{Y=10}$ обчисляємо за формулою (14). При цьому $Y=10$ при $i=3$ і $n_{y_3}=17$. Маємо:

$$\overline{x}_{Y=10}=\frac{\sum_{j=1}^{m}x_j n_{3j}}{n_{y_3}}=\frac{1\cdot 1+2\cdot 0+4\cdot 16}{17}=\frac{65}{17}\approx 3{,}82$$

Умовну дисперсію $D(X/Y=10)$ - за формулою (16), враховуючи, що $\overline{X}_{Y=10}=\frac{65}{17}$:

$$D(X/Y=10)=\frac{\sum_{j=1}^{m}x_j^2 n_{3j}}{n_{y_3}}-\left(\overline{x}_{Y=10}\right)^2=\frac{1^2\cdot 1+2^2\cdot 0+4^2\cdot 16}{17}-\left(\frac{65}{17}\right)^2=$$

$$=\frac{257}{17}-\frac{4225}{17^2}=\frac{4369-4225}{289}=\frac{144}{289}\approx 0{,}498$$

і умовне середньоквадратичне відхилення $\sigma(X/Y=10)$ - за формулою (18)

$$\sigma(X/Y=10)=\sqrt{D(X/Y=10)}=\sqrt{\frac{144}{289}}=\frac{12}{17}\approx 0{,}71. \blacksquare$$

Задача 2. *Результати проведеного аналізу залежності кількості проданих пар чоловічого взуття y_i від його розміру x_i наведено у таблиці:*

Таблиця 7

y_i	25	38	65	95	120	140	152	160	165	175	180	185	190	200
x_i	45	43	42	41	40	39	38,5	38	37,5	37	36,5	36	35,5	35

1) Знайти оцінки параметрів β_0^*, β_1^* емпіричної моделі парної лінійної регресії

$$\hat{y}=\beta_0^*+\beta_1^* x.$$

2) Оцінити наявність та тісноту (щільність) лінійного зв'язку з допомогою вибіркового коефіцієнта кореляції.

3) Оцінити якість регресії з допомогою коефіцієнта детермінації R^2.

4) Зобразити кореляційне поле (систему координат xOy із зображеній у ній спостережуваними точками (x_i, y_i), $1 \le i \le n$) та пряму регресії $\hat{y} = \beta_0^* + \beta_1^* x$.

5) З допомогою побудованої регресії зробити точковий прогноз для $x_{\max} + 0{,}5$.

Інтерпретувати отримані результати.

Розв'язання.

1) *Знайдемо оцінки параметрів β_0^*, β_1^* емпіричної моделі парної лінійної регресії*

$$\hat{y} = \beta_0^* + \beta_1^* x.$$

Аналізуючи формули (30) та (31)

$$\beta_1^* = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \left(\overline{x}\right)^2} = \frac{K_{xy}^*}{D_x},\ \beta_0^* = \overline{y} - \beta_1^* \overline{x}$$

для знаходження β_0^*, β_1^*, бачимо, що для спрощення обчислень зручно заповнити наступну таблицю,

Таблиця 8

№	x_i	y_i	$x_i \cdot y_i$	x_i^2	y_i^2
1	45	25	1125	2025	625
2	43	38	1634	1849	1444
3	42	65	2730	1764	4225
4	41	95	3895	1681	9025
5	40	120	4800	1600	14400
6	39	140	5460	1521	19600
7	38,5	152	5852	1482,25	23104
8	38	160	6080	1444	25600
9	37,5	165	6187,5	1406,25	27225
10	37	175	6475	1369	30625
11	36,5	180	6570	1332,25	32400
12	36	185	6660	1296	34225
13	35,5	190	6745	1260,25	36100
14	35	200	7000	1225	40000
Сума	***544***	***1890***	***71213,5***	***21255***	***298598***
Середнє	***38,85714***	***135***	***5086,679***	***1518,214***	***21328,43***

де у першому стовпці нумеруються спостереження (отже, останній номер співпадає з обсягом вибірки, тобто $n = 14$), другий та третій стовпці – це вхідні дані, четвертий стовпець – добутки відповідних значень x_i та y_i, п'ятий – квадрати значень ознаки X, а останній - квадрати значень ознаки Y. У передостанньому рядку «Сума» просумовані усі елементи відповідних стовпців, а у останньому рядку – елементи рядка «Сума» поділені на $n = 14$. Тоді вже готові значення складових для обчислення β_1^* у формулу (30) підставляються з останнього рядка таблиці 8:

$$\beta_1^* = \frac{K_{xy}^*}{D_x} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \left(\overline{x}\right)^2} = \frac{5086{,}679 - 38{,}857 \cdot 135}{1518{,}214 - \left(38{,}857\right)^2} = \frac{-159{,}036}{8{,}337} \approx -19{,}08.$$

Далі знаходимо β_0^* за формулою (31), підставляючи у неї $\overline{x}$ та $\overline{y}$ з останнього рядка таблиці 8, а $\beta_1^* = -19{,}08$:

$$\beta_0^* = \overline{y} - \beta_1^* \overline{x} = 135 - (-19{,}08) \cdot 38{,}857 \approx 876{,}26.$$

Таким чином, емпіричне рівняння парної лінійної регресії має вигляд

$$\hat{y} = 876,26 - 19,08x \qquad \textbf{(33)}$$

Інтерпретація результату: *оскільки* $\beta_1^* < 0$, *то між фактором* X *та показником* Y *зв'язок зворотній* ($x \uparrow\downarrow y$), *тобто при збільшенні* X *значення* Y *зменшується і навпаки. Крім того, значення коефіцієнта регресії*

$$\beta_1^* = -19,08 < 0,$$

дає змогу стверджувати, що при збільшенні фактору X *на 1 одиницю свого виміру показник* Y *зменшиться в середньому на 19,08 одиниць свого виміру, тобто у даному прикладі при збільшенні розміру чоловічого взуття на 1 кількість проданих пар в середньому зменшиться на* $19,08 \approx 19$.

2) Оцінимо наявність та тісноту (щільність) лінійного зв'язку з допомогою вибіркового коефіцієнта кореляції.

Оскільки при обчисленні β_1^* вже знайдено $K_{xy}^* = -159,036$ та $D_x = 8,337$, а $\sigma_x = \sqrt{D_x}$, то для знаходження r_B залишається знайти ще σ_y. Обчислимо

$$\sigma_y = \sqrt{D_y} = \sqrt{\overline{y^2} - \left(\overline{y}\right)^2} = \sqrt{21328,43 - 135^2} = \sqrt{3103} \approx 55,71.$$

Тоді за формулою (11) маємо:

$$r_B = \frac{K_{xy}^*}{\sigma_x \sigma_y} = \frac{-159,036}{\sqrt{8,337} \cdot 55,71} \approx -0,989.$$

Інтерпретація результату: *оскільки* $r_B < 0$, *то між* X *та* Y *зв'язок зворотній, тобто при збільшенні кількісного значення ознаки* X, *значення* Y *зменшується. Крім того,*

$$|r_B| = |-0,989| = 0,989 \approx 0,99,$$

отже, згідно зі шкалою Чеддока (табл. 2) лінійний зв'язок дуже високий (майже функціональний).

3) Оцінимо якість регресії з допомогою коефіцієнта детермінації R^2.

Обчислимо коефіцієнт детермінації R^2, скориставшись формулою (32), що виражає залежність між R^2 та вибірковим коефіцієнтом кореляції

$$R^2 = r_B^2 = 0,989^2 \approx 0,978.$$

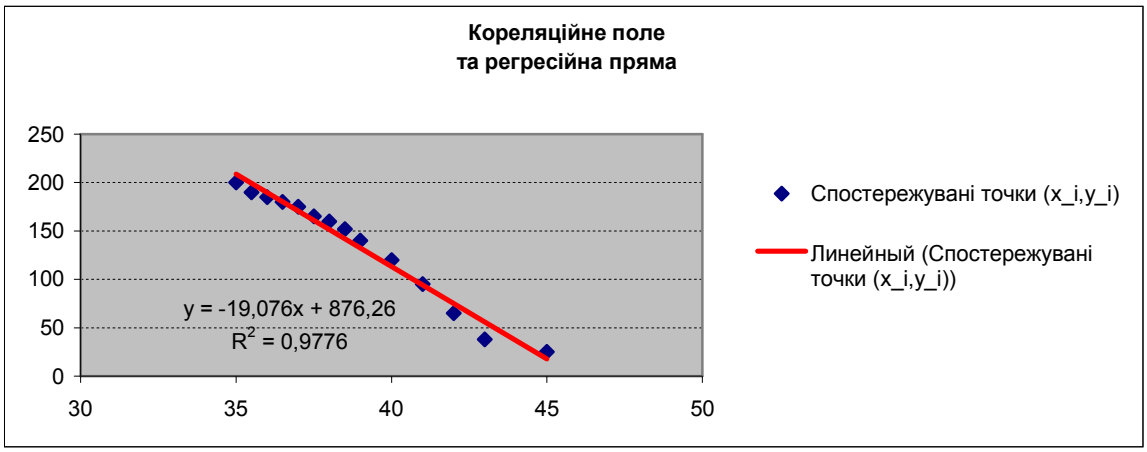
Інтерпретація результату: $R^2 \cdot 100\% \approx 97,8\%$, *отже, зміна показника* Y *на 97,8% пояснюється зміною показника* X, *тобто зміна кількості проданих пар* (y_i) *чоловічого взуття на 97,8% пояснюється зміною його розміру* (x_i) *і лише на 2,2% - іншими факторами та випадковими чинниками.*

4) Зобразимо кореляційне поле та пряму регресії $\hat{y} = 876,26 - 19,08x$.

Для цього в системі координат xOy зобразимо спостережувані точки (x_i, y_i), $1 \le i \le 14$:

(45;25), (43;38), (42;65), (41;95), (40;120), (39;140), (38,5;152), (38;160), (37,5;165), (37;175), (36,5;180), (36;185), (35,5;190), (35;200).

Ці точки в системі координат і утворюють ***кореляційне поле***. Аналіз вигляду кореляційного поля дає перше уявлення про форму залежності між X та Y.



Як видно з рисунку, регресійна пряма досить добре наближає експериментальні (спостережувані) дані.

5) З допомогою побудованої регресії зробимо точковий прогноз для $x_{\max}+0,5$.

Для цього треба підставити значення $x_{\max}+0,5$ у рівняння регресії. У нашому прикладі $x_{\max}=45$, отже, $x_{\max}+0,5=45,5$. Тоді

$$\hat{y}_{прогн}=\hat{y}(45,5)=876,26-19,08\cdot 45,5=876,26-868,14=8,12\approx 8.$$

Таким чином, можна очікувати, що буде продано в середньому 8 пар чоловічого взуття 45,5 розміру. ■

Завдання для самостійного розв'язання

Задача 1. *Двовимірний статистичний закон розподілу задано таблицею (де* k *- порядковий номер студента у списку академгрупи (*$1\le k\le 16$*)) :*

Таблиця 5

$Y=y_i$	$X=x_j$			n_{y_i}
	k+1	**k+2**	**k+4**	
k	0	0	k+5	
k+5	0	18- k	2	
k+10	k+1	0	16-k	
k+15	4	2	2	
n_{x_j}				

Завдання:

1) Записати двовимірний статистичний закон розподілу, що відповідає Вашому варіанту (тобто записати таблицю 5, підставивши конкретне значення k і заповнити останній рядок n_{x_j} та останній стовпець n_{y_i}).

2) Знайти числові характеристики отриманого двовимірного розподілу:

$$\bar{x},\ D_x,\ \sigma_x,\ \bar{y},\ D_y,\ \sigma_y,\ K^*_{xy},\ r_B\ .$$

3) Скласти та записати умовні закони розподілу

$$Y/X=k+4 \text{ та } X/Y=k+10$$

і знайти їхні числові характеристики:

$$\bar{y}_{X=k+4},\ D(Y/X=k+4),\ \sigma(Y/X=k+4),$$
$$\bar{x}_{Y=k+10},\ D(X/Y=k+10),\ \sigma(X/Y=k+10).$$

Задача 2. *За вхідними даними Вашого варіанту (заданим парним статистичним розподілом):*

1) Знайти оцінки параметрів β_0^* , β_1^* емпіричної моделі парної лінійної регресії

$$\hat{y} = \beta_0^* + \beta_1^* x .$$

2) Оцінити наявність та тісноту (щільність) лінійного зв'язку з допомогою вибіркового коефіцієнта кореляції.

3) Оцінити якість регресії з допомогою коефіцієнта детермінації R^2 .

4) Зобразити кореляційне поле (систему координат xOy із зображеній у ній спостережуваними точками (x_i, y_i), $1 \le i \le n$) та пряму регресії $\hat{y} = \beta_0^* + \beta_1^* x$.

5) З допомогою побудованої регресії зробити точковий прогноз, збільшивши максимальне значення фактора на 10% ($x_{\max} + 10\% = x_{\max} + 0{,}1x_{\max} = 1{,}1x_{\max}$).

Інтерпретувати отримані результати.

Вхідні дані для задачі 2.

Варіант 1 (Байрамов Алі)

Залежність кров'яного тиску ***Y*** людини (в умовних одиницях) від довжини руки ***X*** наведена в таблиці:

y_i	115	117	120	122	124	125	127	129
x_i, см	62,1	61,0	59,0	58,0	56,5	56	55	54,5

Варіант 2 (Беленчук Олексій)

Залежність між продуктивністю праці ***Y*** та фондозабезпеченістю ***X*** на підприємствах однієї галузі наведено в таблиці:

y_i, тис.грн.	14,85	11,94	8,03	7,11	9,50	11,60	8,14	7,34
x_i, тис.грн.	60	48	39	28	45	58	27	38

Варіант 3 (Березний Ігор)

Залежність урожайності цукрових буряків Y від кількості внесених у грунт поживних речовин X наведена в таблиці:

y_i, ц/га	369	380	370	395	420	412	436	420
x_i, кг/га	83	92	112	132	144	154	162	189

Варіант 4 (Бужак Андрій)

Залежність маси монети ***Y*** від часу її обігу в роках ***X*** наведена в таблиці:

y_i, мг	9,18	9,10	9,05	8,98	8,94	8,88	8,78	8,75	8,65
x_i, років	5,5	6,8	8,5	12,0	15,9	28,5	36,8	40,0	50,0

Варіант 5 (Бурле Павло)

Залежність між собівартістю ***Y*** та кількістю виготовлених виробів ***X*** наведена в таблиці:

y_i, тис.грн.	2,2	3,5	3,7	3,8	4,5	5,7
x_i, тис.шт.	1,5	1,4	1,2	1,1	0,9	0,8

Варіант 6 (Василевич Павло)

Залежність величини зносу різця ***Y*** від тривалості роботи ***X*** показана в таблиці:

y_i, мм	26,8	26,5	26,3	26,1	25,7	25,3	24,3	24,1	24,0
x_i, год	15	16	17	18	19	20	21	22	23

Варіант 7 (Волощук Назарій)

Залежність денного споживання масла ***Y*** певної особи від розміру її заробітної плати за добу ***X*** наведена в таблиці:

y_i, г	12,5	15,8	17,8	19,5	20,4	21,5	22,2	24,3	26,5
x_i, грн	70	75	82	89	95	100	105	110	120

Варіант 8 (Георгіян Євген)

Залежність числа гризунів ***Y***, які загинули від наявності отрути в їжі при концентрації ***X***, наведена в таблиці:

y_i	32	38	46	49	59	68	73	81	92
x_i, %	3	4	5	6	7	8	9	10	11

Варіант 9 (Гончаров Олександр)

Залежність кількості проданих пар чоловічого взуття ***Y*** від його розміру ***X*** наведена в таблиці:

y_i, шт	10	25	68	136	152	162	170	180
x_i	44	43	42	41	40	39	38	37

Варіант 10 (Григорчук В'ячеслав)

Залежність між собівартістю ***Y*** та кількістю виготовлених виробів ***X*** наведена в таблиці:

y_i, тис.грн.	4,2	5,5	5,7	5,9	6,5	7,8
x_i, тис.шт.	2,5	2,4	2,2	2,1	1,9	1,8

Варіант 11 (Денис Денис)

Залежність урожайності пшениці ***Y*** від глибини зволоження ґрунту ***X*** наведена в таблиці:

y_i, ц/га	10	14	20	26	30	36	40	44	48
x_i, см	0	8	14	20	24	30	34	38	42

Варіант 12 (Дісар Іван)

Залежність пружності ***Y*** сталевих болтів від вмісту в них нікелю ***X*** наведена в таблиці:

y_i, %	39,1	40,5	42,4	43,8	45,6	46,9	48,5	50,0
x_i, %	2,95	2,99	3,00	3,11	3,21	3,29	3,34	3,50

Варіант 13 (Дручук Роман)

Зі старшого класу навмання обраної середньої школи було відібрано групу учнів. Дані про їх середньорічні оцінки з математики ***Y*** та середню оцінку решти дисциплін ***X*** в балах наведені в таблиці:

y_i	45	48	54	59	72	76	82	85	90
x_i	30	31	41	50	60	65	78	71	80

Варіант 14 (Дубець Василь)

Конденсатор було заряджено до повної напруги в певний момент часу, після чого він почав розряджатися. Залежність напруги ***Y*** від часу розрядження ***X*** наведена в таблиці:

y_i	100	85	70	60	45	35	25	22	20
x_i	0	1	2	4	7	9	11	12	13

Варіант 15 (Дуплава Олександр)

Залежність урожайності озимої пшениці ***Y*** від кількості внесених добрів ***X*** наведена в таблиці:

y_i, ц/га	16	19	22	25	26	27	32	33	34
x_i, кг/га	60	70	80	90	100	110	120	130	140

Варіант 16 (Жупник Евеліна)

Показники товарообігу ***Y*** та суми витрат ***X*** , які досліджувалися в 10 магазинах, наведені в таблиці:

y_i, грн.	4800	5100	5300	5550	5700	5850	5960	6050	6250	6400
x_i, грн.	300	250	310	280	400	540	600	640	780	830