Розділ 3. ОСНОВИ КОРЕЛЯЦІЙНОГО АНАЛІЗУ

Будь-який соціо-економічний об'єкт або явище зазвичай характеризується декількома ознаками, тобто різними властивостями. Ці ознаки взаємозв'язані і впливають одна на одну. Крім того, може існувати зв'язок між ознаками різних об'єктів і явищ. Тому в математичній статистиці розроблений апарат для виявлення таких зв'язків і оцінки їх сили (тісноти). Цей математичний апарат називається кореляційним аналізом.

3.1. Поняття кореляційного зв'язку між досліджуваними величинами

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**: $Y = f(X)$.

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 вважається дуже складним і трудомістким процесом, у якому потрібно аналізувати багатовимірні таблиці даних. Тому, зазвичай, вивчається не статистичний, а кореляційний зв'язок між X та Y.

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

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

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

…

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

то між X та Y існує кореляційний зв'язок.

Наприклад, відомо, що з однакових за площею ділянок землі при рівних кількостях внесеного добрива отримують різний урожай. Тому, якщо Y – урожайність зерна, а X – кількість внесеного добрива, то функціонального зв'язку між X та Y немає. Це пояснюється впливом таких випадкових факторів, як температура повітря, кількість опадів і т. ін. Однак досвід показує, що середній урожай є функцією від кількості добрива, тобто 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 найбільш суттєво впливають на результативну ознаку;

− виявлення невідомих причинно-наслідкових зв'язків між ознаками об'єкта.

3.2. 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 (форм. 1.2), 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 (рис. 3.1). Вказане графічне зображення вибіркових даних називається **полем кореляції**.

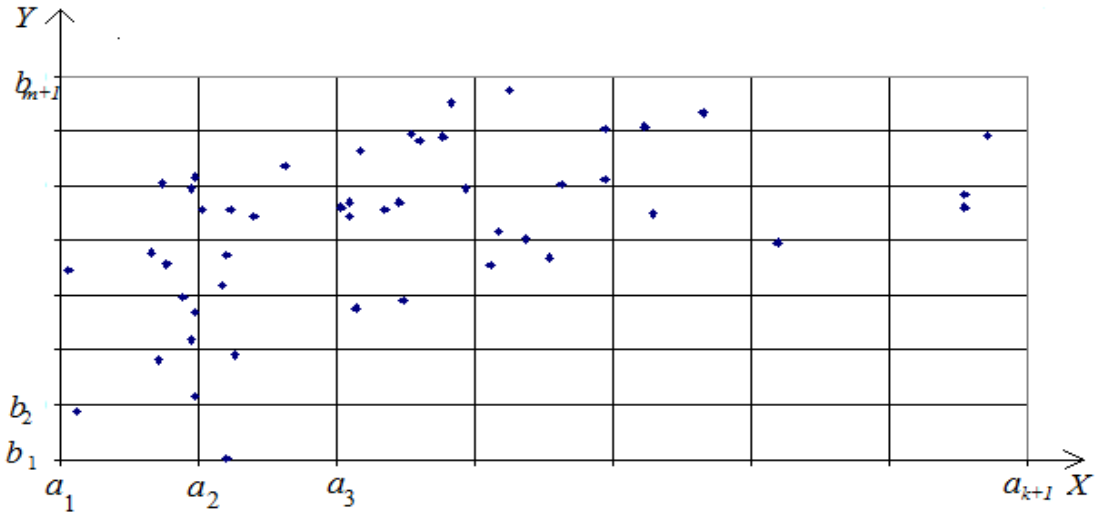

Рисунок 3.1. Поле кореляції

3) Побудувати кореляційну таблицю (табл. 3.1). 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.

Таблиця 3.1

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 (табл. 3.2). В першому рядку таблиці записуються середини інтервалів x_i. В другому – відповідні середні значення $\overline{y}_{x_i}$, що знаходяться за формулами:

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

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

Таблиця 3.2

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.

3.3. 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:

$$r = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{S_x \cdot S_y}, \tag{3.1}$$

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

$\overline{y}$ – 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.

Враховуючи формули для знаходження вибіркових середніх і середніх квадратичних відхилень, а саме:

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

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

отримують більш зручну для розрахунків формулу:

$$r = \frac{n\sum_{i=1}^{k}\sum_{j=1}^{m} x_i y_j n_{ij} - \left(\sum_{i=1}^{k} x_i n_i\right)\left(\sum_{j=1}^{m} y_j n_j\right)}{\sqrt{n\sum_{i=1}^{k} x_i^2 n_i - \left(\sum_{i=1}^{k} x_i n_i\right)^2}\sqrt{n\sum_{j=1}^{m} y_j^2 n_j - \left(\sum_{j=1}^{m} y_j n_j\right)^2}}. \quad (3.2)$$

У випадку незгрупованих даних розрахункова формула суттєво спрощується:

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}. \quad (3.3)$$

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.

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.

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:

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

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 (α ; *l*), де α – 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 α.

Приклад 3.1. За наявними даними про рівень механізації праці X (%) і продуктивності праці Y (од. продукції/год.) для 14 однотипних підприємств (табл. 3.3) оцінити тісноту зв'язку між X і Y. Визначити можливість розповсюдження результатів розрахунків на всі підприємства такого типу.

Таблиця 3.3

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 табл. 3.3 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 за формулою (3.3.) для незгрупованих даних. Розрахунки для зручності оформимо у вигляді таблиці (табл. 3.4).

Таблиця 3.4

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

Отже,

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

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

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

Перевіримо статистичну значущість знайденого коефіцієнта кореляції

Пірсона. Розрахуємо t-статистику за формулою (3.4): $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0{,}969\sqrt{14-2}}{\sqrt{1-0{,}969^2}} \approx 13{,}59$. 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.

Приклад 3.2. За наявними даними про річний об'єм виробництва Y (тис. од. продукції) та основні фонди X (тис. у. од.) для 20 однотипних підприємств (табл. 3.5) оцінити тісноту зв'язку між X і Y. Визначити можливість розповсюдження результатів розрахунків на всі підприємства такого типу.

Таблиця 3.5

Y \ X	12,5	17,5	22,5	27,5
20,5	1	–	–	–
21,5	–	2	–	–
22,5	–	1	2	–
23,5	–	–	3	3
24,5	–	–	–	8

Розв'язок. Згруповані вибіркові дані (табл. 3.5) запишемо у вигляді кореляційної таблиці (табл. 3.6).

Таблиця 3.6

y_j \ x_i	12,5	17,5	22,5	27,5	n_j
20,5	1	0	0	0	1
21,5	0	2	0	0	2
22,5	0	1	2	0	3
23,5	0	0	3	3	6
24,5	–	–	–	8	8
n_i	1	3	5	11	

Для розрахунку коефіцієнта кореляції Пірсона скористаємося формулою (3.2). Розрахунки для зручності оформимо у вигляді таблиці (табл. 3.7).

Таблиця 3.7

x_i	n_i	y_j	n_j	$x_i n_i$	$y_j n_j$	x_i^2	$x_i^2 n_i$	y_j^2	$y_j^2 n_j$
12,5	1	20,5	1	12,5	20,5	156,25	156,25	420,25	420,25
17,5	3	21,5	2	52,5	43	306,25	918,75	462,25	924,5
22,5	5	22,5	3	112,5	67,5	506,25	2531,3	506,25	1518,8
27,5	11	23,5	6	302,5	141	756,25	8318,8	552,25	3313,5
		24,5	8		196			600,25	4802
Суми									
				480	468		11925		10979

Окремо розрахуємо $\sum_{i=1}^{k}\sum_{j=1}^{m} x_i y_j n_{ij}$:

$$\sum_{i=1}^{4}\sum_{j=1}^{5} x_i y_j n_{ij} = 20,5\cdot 12,5 + 21,5\cdot 17,5\cdot 2 + 22,5\cdot 17,5 + 22,5\cdot 22,5 + 23,5\cdot 22,5\cdot 3 +$$

$$+23,5\cdot 27,5\cdot 3 + 24,5\cdot 27,5\cdot 8 = 11330.$$

Підставимо знайдені суми у формулу (3.2):

$$r = \frac{n\sum_{i=1}^{k}\sum_{j=1}^{m} x_i y_j n_{ij} - \left(\sum_{i=1}^{k} x_i n_i\right)\left(\sum_{j=1}^{m} y_j n_j\right)}{\sqrt{n\sum_{i=1}^{k} x_i^2 n_i - \left(\sum_{i=1}^{k} x_i n_i\right)^2}\sqrt{n\sum_{j=1}^{m} y_j^2 n_j - \left(\sum_{j=1}^{m} y_j n_j\right)^2}} = \frac{20\cdot 11330 - 480\cdot 468}{\sqrt{20\cdot 11925 - 480^2}\sqrt{20\cdot 10979 - 468^2}} =$$

$$= \frac{1960}{90\cdot 23,58} \approx 0,924.$$

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

Перевіримо статистичну значущість знайденого коефіцієнта кореляції Пірсона. Розрахуємо t-статистику за формулою (3.4): $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0,924\sqrt{20-2}}{\sqrt{1-0,924^2}} = \frac{3,92}{\sqrt{0,146}} \approx 10,26$. Знайдемо $t_{\text{крит}}$, враховуючи, що $l = n-2 = 20-2 = 18$. Оберемо рівень значущості $\alpha = 0,01$. Тоді $t_{\text{крит}}$ = СТЬЮДРАСПОБР (0,01; 18) = 2,88.

Оскільки розраховане значення t-статистики більше критичного $10,26 > 2,88$, то коефіцієнт кореляції можна вважати значущим на обраному рівні $\alpha = 0,01$.

Висновок. Між річним об'ємом виробництва та основними фондами на підприємствах, що досліджувалися, існує сильний додатній зв'язок. Висновок дійсний для всіх підприємств такого типу.

3.4. 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:

$$r_S(X,Y) = 1 - \frac{6\sum_{i=1}^{n} d_i^2 + T_X + T_Y}{n(n^2-1)}, \tag{3.5}$$

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:

$$T_X = \frac{\sum_{i=1}^{L_X}\left(T_{Xi}^3 - T_{Xi}\right)}{12}; \qquad T_Y = \frac{\sum_{i=1}^{L_Y}\left(T_{Yi}^3 - T_{Yi}\right)}{12}, \tag{3.6}$$

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

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

Зауваження 1. Ранги присвоюються вибірковим даним звичайним способом (див. п. 2.3.4).

Зауваження 2. Статистична значущість коефіцієнта кореляції Спірмена перевіряється так, як і коефіцієнта кореляції Пірсона.

Приклад 3.3. Вивчається залежність між продуктивністю праці робітників X (тис. грн.) та їх емоційним відношенням до своєї професійної діяльності Y (бали). Відповідні дані подано у табл. 3.8. Оцінити силу зв'язку між досліджуваними факторами за коефіцієнтом кореляції Спірмена. Перевірити його статистичну значущість.

Таблиця 3.8

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

Розв'язок. Дані табл. 3.8 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 (табл. 3.9), використовуючи позначення: d_{x_i} – ранг x_i, d_{y_i} – ранг y_i.

Таблиця 3.9

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.

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:

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

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

Обчислимо коефіцієнт кореляції Спірмена за формулою (3.6):

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

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

Знайдемо $t_{\text{крит}}$, враховуючи, що $l = n - 2 = 15 - 2 = 13$. Оберемо рівень значущості $\alpha = 0{,}001$. Тоді $t_{\text{крит}}$ = 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.

3.5. 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:

$$A=\begin{pmatrix} 1 & r_{12} & \dots & r_{1k} \\ r_{21} & 1 & \dots & r_{2k} \\ \dots & \dots & \dots & \dots \\ r_{k1} & r_{k2} & \dots & r_{kk} \end{pmatrix}. \tag{3.7}$$

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

3) Розрахувати множинний коефіцієнт кореляції за формулою:

$$R_i=\sqrt{1-\frac{|A|}{A_{ii}}}\,. \tag{3.8}$$

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

$$t=\frac{R^2\left(n-k\right)}{\left(1-R^2\right)\left(k-1\right)}, \tag{3.9}$$

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

Розраховане значення *t*-статистики порівнюється з критичним значенням $F_{\text{крит}}$. $F_{\text{крит}}$ – табличне значення розподілу Фішера, яке також можна знайти за допомогою вбудованої статистичної функції Excel FРАСПОБР (α; l_1; l_2), де α – 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 α.

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 $X_1, X_2, \dots, X_k$ досліджуваного об'єкта або явища, який не залежить від впливу інших ознак, розраховується **частинний коефіцієнт кореляції**, який позначається R_{ij}.

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:

$$R_{ij} = \frac{-A_{ij}}{\sqrt{A_{ii} A_{jj}}}. \qquad (3.10)$$

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

$$t = \frac{R_{ij}\sqrt{n-k+2}}{\sqrt{1-R_{ij}^{\,2}}}, \qquad (3.11)$$

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

Розраховане значення *t*-статистики порівнюється з критичним значенням $t_{\text{крит}}$. $t_{\text{крит}}$ – табличне значення розподілу Стьюдента, яке також можна знайти за допомогою вбудованої статистичної функції Excel СТЬЮДРАСПОБР (α; l), де α – 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 α.

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 3.4. Для вивчення залежності урожайності зернових культур *Z* (ц/га) від якості пашні *X* (бали) і кількості внесеного добрива *Y* (кг/га) було проведено дослідження шести фермерських господарств, результати якого представлено у табл. 3.10. 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 частинними коефіцієнтами кореляції.

Таблиця 3.10

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), $i=\overline{1,n}$, $n=6$.

Побудуємо матрицю парних коефіцієнтів кореляції, які обчислимо за формулою (3.3). Розрахунки для зручності оформимо у вигляді таблиці (табл. 3.11).

Таблиця 3.11

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

Отже, за формулою (3.3) маємо:

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

$$\approx 0,935;$$

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

$$\approx 0,954;$$

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

$$\approx 0,991;$$

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.

Перевіримо статистичну значущість множинного коефіцієнта кореляції R_Z. Знайдемо t-статистику за формулою (3.9):

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

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:

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

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

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

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

Тоді за формулою (3.10) маємо:

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

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

Перевіримо статистичну значущість частинного коефіцієнта кореляції R_{13}. Знайдемо t-статистику за формулою (3.11):

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

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 $\alpha = 0{,}01$.

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 α =0,01.

Висновок: Урожайність зернових культур сильно пов'язана з якістю пашні і кількістю внесеного добрива. При цьому урожайність значно сильніше залежить від кількості добрива, ніж від якості пашні. Сила зв'язку між урожайністю та якістю пашні середня і не є статистично значущою.

3.6. 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***.

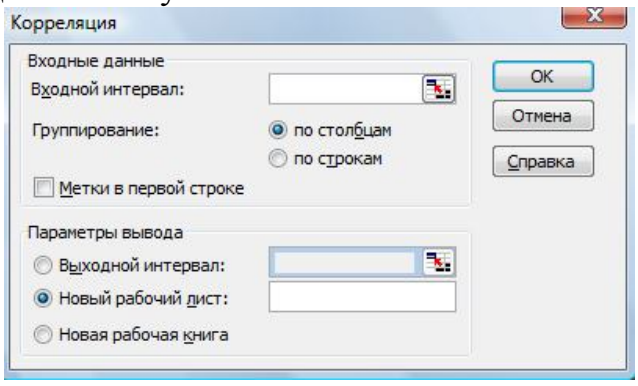

Рисунок 3.2. Вікно надання параметрів кореляційного аналізу

Приклад і результати розрахунків парних коефіцієнтів кореляції продемонстровано на рис. 3.3.

Файл Правка Вид Вставка Формат Сервис Данные Окно Справка

E24

	A	B	C	D	E	F	G	H	I
1				**Кореляційний аналіз**					
2		Вхідні дані							
3	№	Значення					*Столбец 1*	*Столбец 2*	*Столбец 3*
4	**i**	**x1**	**x2**	**x3**		Столбец 1	1		
5	1	1	328	0,054		Столбец 2	0,8913997	1	
6	2	2	329	0,101		Столбец 3	0,5634229	0,692214	1
7	3	3	329	0,099					
8	4	4	345	0,019					
9	5	5	352	0,065					
10	6	6	370	0,053					
11	7	7	377	0,178					
12	8	8	385	0,174					
13	9	9	396	0,289					
14	10	10	399	0,195					
15	11	11	390	0,102					
16	12	12	373	0,138					

Рисунок 3.3. Результати розрахунку коефіцієнтів кореляції

Зауваження. 1) В результаті роботи інструмента аналізу даних ***Корреляция*** розраховується матриця парних коефіцієнтів кореляції Пірсона навіть у випадку встановлення зв'язку між двома величинами.

2) Чарунки матриці, що розташовані вище головної діагоналі, зазвичай залишаються незаповненими, оскільки матриця симетрична відносно головної діагоналі.

3) Засобами Microsoft Excel неможливо розрахувати парні або множинні коефіцієнти кореляції, однак можна значно спростити розрахунки, використовуючи вбудовану математичну функцію МОПРЕД, яка дозволяє обчислити визначник заданої матриці.

Приклад і результати обчислення визначника матриці показано на рис. 3.4.

Файл Правка Вид Вставка Формат Сервис Данные Окно Справка

E6 =МОПРЕД(A5:C7)

	A	B	C	D	E	F
1			**Обчислення визначника**			
2						
3	Вхідні дані: матриця			Визначник заданої матриці		
4						
5	345	671	134			
6	102	204	133		-11322297	
7	147	654	334			
8						

Рисунок 3.4. Обчислення визначника заданої матриці

3.7. Можливості SPSS у дослідженні кореляції

У SPSS вибір методів обчислення коефіцієнтів кореляції залежить від виду шкали, до якої належить змінна. А саме, для інтервальних та номінальних величин − це коефіцієнт кореляції Пірсона; якщо хоча б одна із змінних належить до порядкової шкали або не підпорядковується нормальному розподілу − коефіцієнт рангової кореляції Спірмена або τ Кендала.

Кореляційний аналіз можна здійснювати безпосередньо в процесі побудови таблиць зв'язності для двох змінних за допомогою команд меню ***Анализ − Описательные статистики − Таблицы сопряженности***, вибравши у пункті ***Статистики*** необхідний коефіцієнт, або за допомогою окремих команд меню.

3.7.1. Коефіцієнт кореляції Пірсона

Вивчимо тісноту зв'язку між річним об'ємом виробництва *Y* (тис. од. продукції) та основними фондами *X* за даними табл. 3.5 із прикладу 3.2.

Для цього необхідно:

1) Ввести дані та вибрати в меню послідовно ***Анализ − Корреляции − Парные***, з'явиться діалогове вікно ***Парные корреляции*** (рис. 3.5);

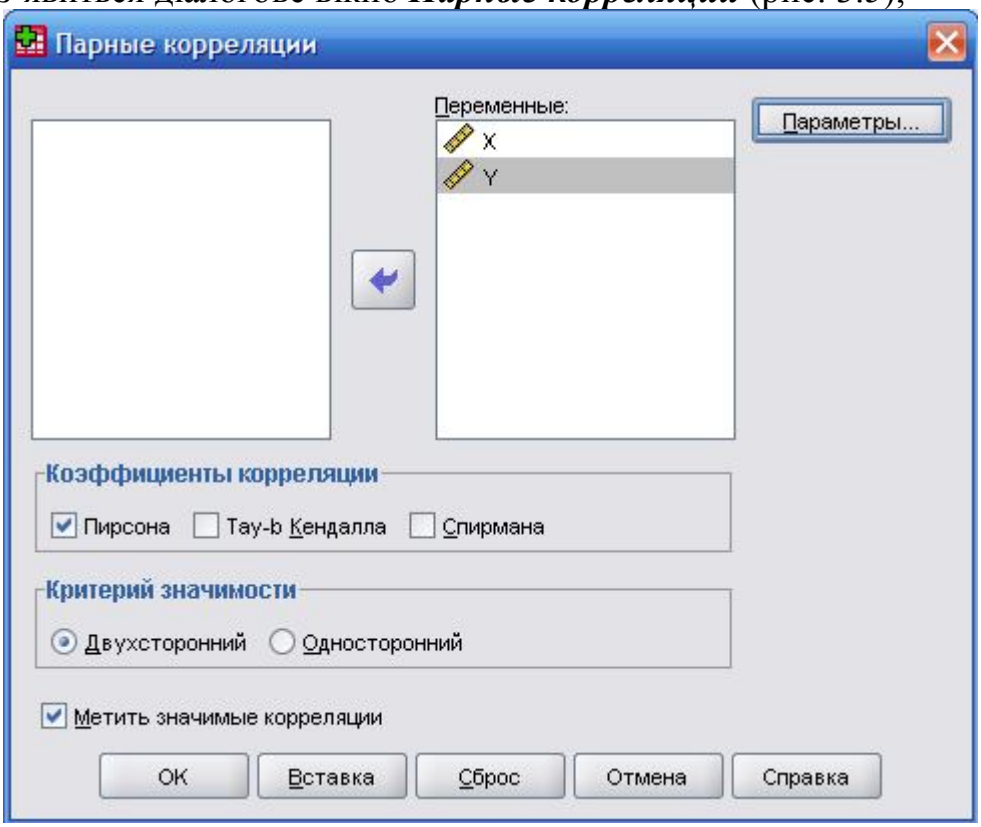

Рисунок 3.5. Діалогове вікно вибору коефіцієнта кореляції

2) Перенести змінні *X*, *Y* у поле ***Переменные***, серед коефіцієнтів кореляції залишити відмітку на ***Пирсона*** та проаналізувати результати у вікні перегляду (рис. 3.6).

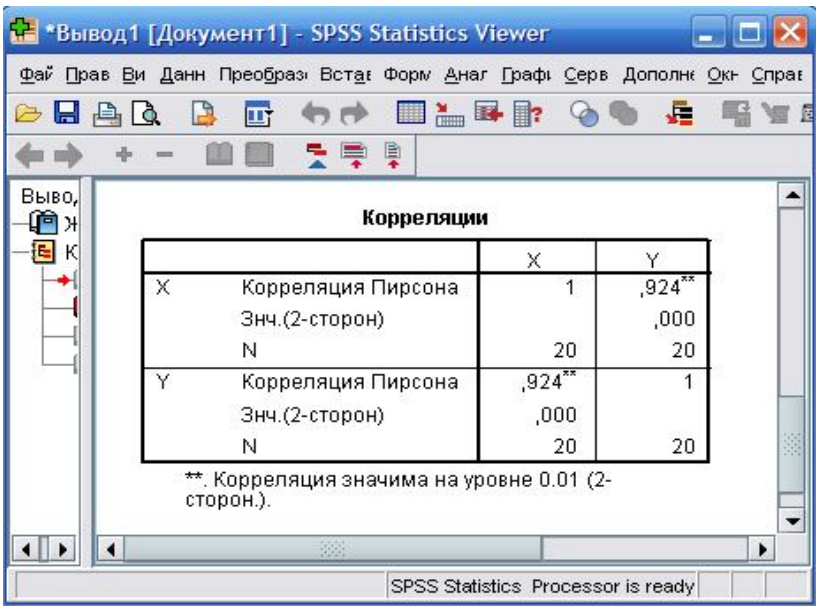

Рисунок 3.6. Результат тісного зв’язку між змінними *X* і *Y*

Так як коефіцієнт кореляції − 0,924, то між річним об'ємом виробництва та основними фондами існує сильний додатній зв’язок. Причому значення коефіцієнта кореляції є значущим на рівні 0,01.

3.7.2. Коефіцієнт кореляції Спірмена

Вивчимо тісноту зв’язку між продуктивністю праці робітників *X* (тис. грн.) та їх емоційним відношенням до своєї професійної діяльності *Y* (бали) за даними, представленими у таблиці 3.8 із прикладу 3.3.

Для цього необхідно:

1) Ввести дані, вибрати в меню послідовно ***Анализ*** − ***Корреляции*** − ***Парные***, у діалоговому вікні ***Парные корреляции*** (рис. 3.5) перенести змінні *X*, *Y* у поле ***Переменные***, серед коефіцієнтів кореляції вибрати коефіцієнт ***Спирмена***, та проаналізувати результат у вікні перегляду (рис. 3.7).

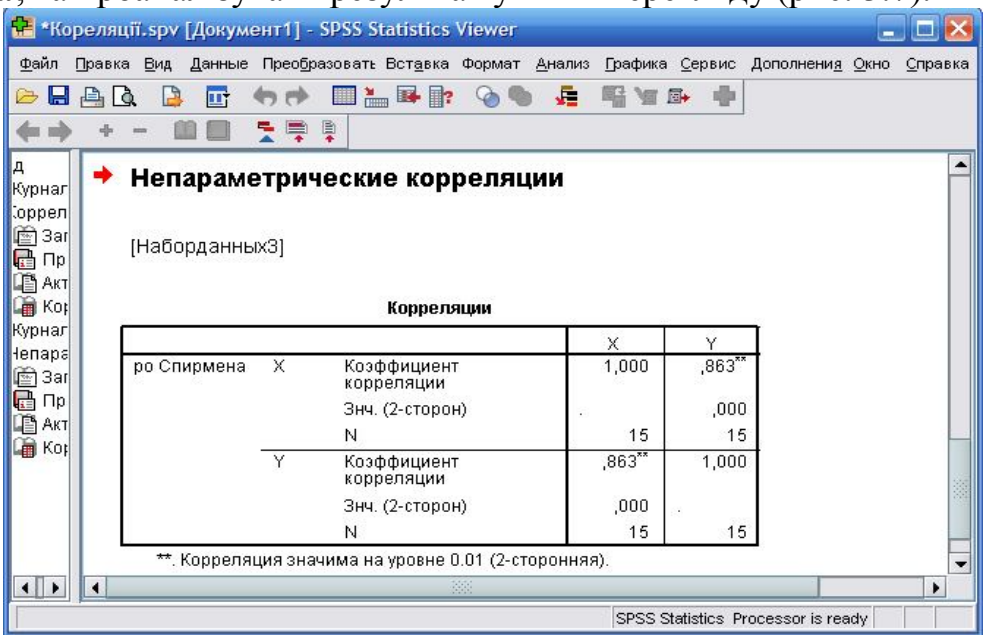

Рисунок 3.7. Міра зв’язку між змінними *X* і *Y* за коефіцієнтом кореляції Спірмена

Коефіцієнт кореляції − 0,863 свідчить про наявність тісного прямого зв'язку між продуктивністю праці та емоційним відношенням працівника до професійної діяльності. Значення коефіцієнта кореляції є значущим на рівні 0,01.

3.7.3. Частинний коефіцієнт кореляції

Проаналізуємо залежність урожайності зернових культур *Z* (ц/га) від якості пашні *X* (бали) і кількості внесеного добрива *Y* (кг/га) за даними табл. 3.10 із прикладу 3.4.

Виконаємо:

1) Введемо дані, виберемо в меню послідовно ***Анализ − Корреляции − Частные***, у діалоговому вікні ***Частные корреляции*** (рис. 3.8) перенесемо змінні *Z*, *X* у поле ***Переменные***, а змінну *Y* у поле ***Исключаемые***. Натиснемо ***ОК***;

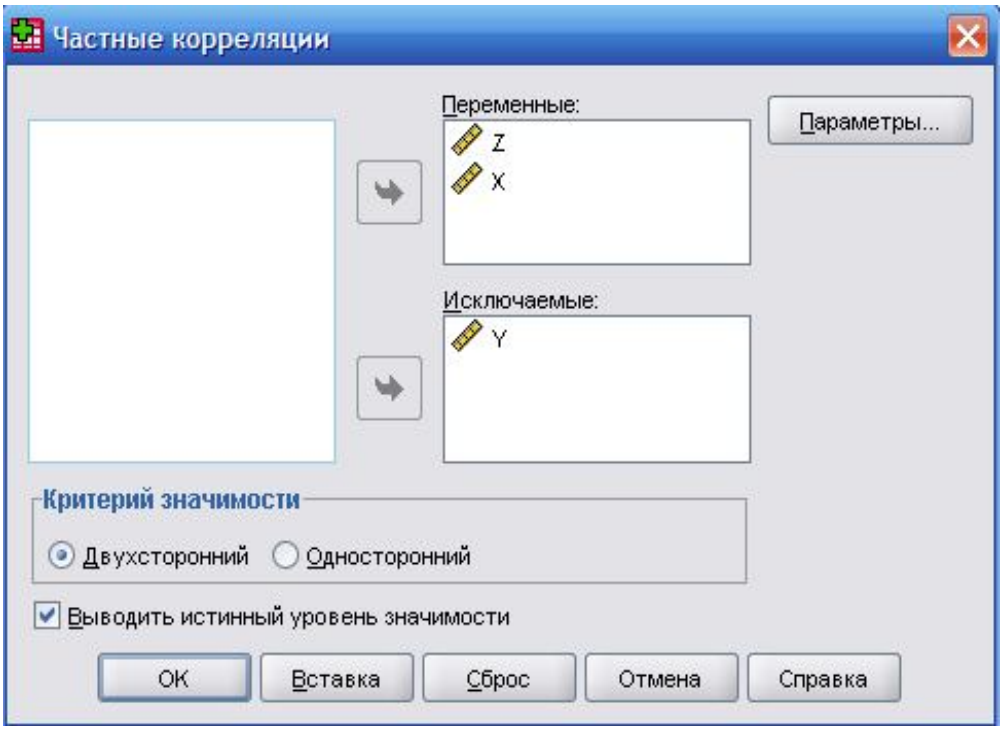

Рисунок 3.8. Діалогове вікно вибору змінних для частинної кореляції

2) Повторимо етапи пункту 1) з поправкою: у діалоговому вікні ***Частные корреляции*** (рис. 3.8) перенесемо змінні *Z, Y* у поле ***Переменные***, а змінну *X* у поле ***Исключаемые***. Отримаємо результати частинного кореляційного зв'язку (рис. 3.9).

Згідно даних кореляційних таблиць (рис. 3.9), урожайність зернових культур (*Z*) тісно пов'язана із кількістю внесеного добрива (*Y*), про що свідчить значення коефіцієнта кореляції − 0,935, яке є значущим на рівні 0,02. Між урожайністю (*Z*) та якістю пашні (*X*) існує помірний зв'язок, який характеризується коефіцієнтом 0,587 і не є статистично значущим.

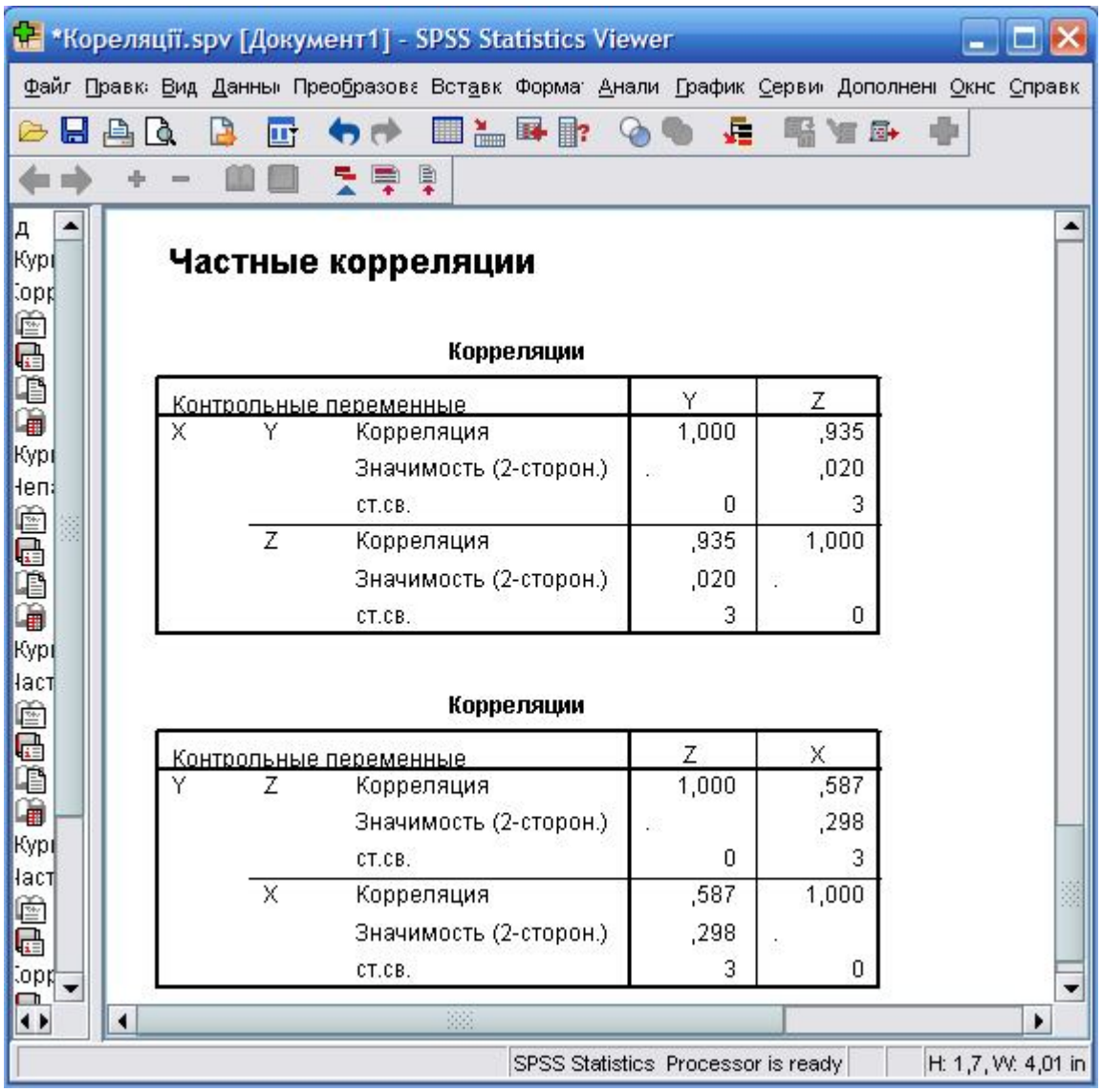

Рисунок 3.9. Частинні коефіцієнти кореляції

Завдання для самостійного виконання

3.1. Визначити силу зв'язку між вагою рослини X (г) і вагою його насіння Y (г) за даними табл. 3.12.

Таблиця 3.12

X	40	50	60	70	80	90	100
Y	20	25	28	30	35	40	45

3.2. В табл. 3.13 наведені дані про роздрібний товарообіг Z (млрд грн.), середню кількість населення X (млн. осіб) та середній дохід Y (млн грн.). Проаналізувати зв'язок між Z та X і Y за частинним і множинним коефіцієнтами кореляції.

Таблиця 3.13

Z	1,2	1,3	2,5	1,4	1,2	0,2	2,4	4,1	1,1
X	1,4	1,4	2,5	1,5	1,3	0,3	2,6	4,2	1,1
Y	1,3	1,3	1,4	1,8	1,5	1,6	1,8	1,9	1,6

3.3. Для дослідження впливу капіталовкладень X (млн грн.) на отриманий річний прибуток Y (млн грн.) було зібрано статистичні дані по 20 великих підприємствах (табл. 3.14). Визначити тісноту зв'язку між означеними факторами.

Таблиця 3.14

Y \ X	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
1,5 – 2,5	1	–	–	–	–
2,5 – 3,5	2	5	2	–	–
3,5 – 4,5	–	3	3	2	–
4,5 – 5,5	–	–	–	2	–

3.4. В таблиці 3.15 наведено дані про щомісячний прибуток Z (тис. у. од.), витрати на рекламу X (тис. у. од.) та вкладення капіталу в цінні папери Y (тис. у. од.). Проаналізувати зв'язок між Z та X і Y за частинним і множинним коефіцієнтами кореляції.

Таблиця 3.15

Z	10	12	12	14	16	17	18
X	0,2	0,5	0,3	0,5	0,5	0,6	0,8
Y	0,8	0,2	1	1,2	0,9	1	1,1

3.5. В табл. 3.16 наведено дані про рівень витрат X (%) та річний дохід Y (млн грн.) 50-ти великих магазинів Визначити тісноту зв'язку між означеними факторами.

Таблиця 3.16

Y \ X	4 – 6	6 – 8	8 – 10	10 – 12	12 – 14
0,5 – 2,0	–	–	2	3	1
2,0 – 3,5	–	4	5	1	–
3,5 – 5,0	–	8	5	5	–
5,0 – 6,5	3	8	2	–	–
6,5 – 8,0	2	1	–	–	–

3.6 – 3.15. За даними табл. 3.17 перевірити гіпотезу про наявність лінійного зв'язку.

Таблиця 3.17

№	X					Y					α
3.6	1	5	3	4	7	1	5	5	2	8	0,05
3.7	3	6	7	8	7	1	3	5	5	4	0,01
3.8	4	7	5	4	5	3	1	2	2	1	0,05
3.9	9	8	3	4	1	0	1	4	3	5	0,01
3.10	1	0	3	3	0	2	3	5	6	4	0,05
3.11	0	4	7	8	5	2	6	8	7	5	0,01
3.12	4	2	3	4	3	8	6	8	7	6	0,05
3.13	7	5	1	0	3	8	6	4	2	4	0,01
3.14	3	5	7	2	5	1	3	5	0	1	0,05
3.15	4	4	8	9	5	6	2	9	9	4	0,01

Питання для самоконтролю

1. Що називається кореляційним аналізом? Яка мета кореляційного аналізу?
2. Що називається кореляційним зв’язком? Статистичним зв’язком?
3. Що називається коефіцієнтом кореляції? Як він використовується у статистичному моделюванні?
4. Як згрупувати вхідні дані при кореляційному аналізі?
5. Що називається полем кореляції? Кореляційною таблицею?
6. Як побудувати кореляційну таблицю?
7. Як за вибірковими даними визначити вид зв’язку?
8. Чим відрізняються коефіцієнти кореляції Пірсона та Спірмена? Які загальні риси мають коефіцієнти кореляції Пірсона і Спірмена?
9. Як мають задаватись вхідні дані для кореляційного аналізу у випадку лінійного зв'язку? У випадку нелінійного зв’язку?
10. Який зв’язок між факторами вважається сильним? Середнім? Слабким?
11. Чи показує коефіцієнт кореляції спрямованість зв’язку?
12. Який висновок робиться, якщо коефіцієнт кореляції є додатнім? Від’ємним?
13. Як присвоюються ранги, якщо вибіркові дані повторюються?
14. Що означає поняття «статистична значущість»?
15. Як перевірити зв’язок між декількома факторами?
16. Які коефіцієнти кореляції обчислюються при множинному кореляційному аналізі?
17. Що таке множинний кореляційний зв’язок?
18. Що таке чистий кореляційний зв’язок?
19. Для чого служить частинний коефіцієнт кореляції? Множинний коефіцієнт кореляції?
20. Що називається кореляційною матрицею? Для чого будується кореляційна матриця?
21. Які властивості кореляційної матриці?
22. Для чого перевіряється статистична значущість коефіцієнта кореляції?
23. Як побудувати кореляційну матрицю засобами MS Excel?
24. Чи можливо знайти множинний та частинний коефіцієнти кореляції засобами Microsoft Excel? Засобами SPSS?

Розділ 4. ПОБУДОВА РЕГРЕСІЙНИХ МОДЕЛЕЙ

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**.

4.1. 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 (див. п. 3.2) і будується статистичний ряд, що містить значення X, відповідні середні значення Y та частоти (табл. 4.1).

Таблиця 4.1

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

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

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

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