Розділ 2. ПЕРЕВІРКА СТАТИСТИЧНИХ ГІПОТЕЗ

2.1. Поняття про статистичні гіпотези

При застосуванні певних статистичних методів обробки даних вибірки часто ставляться вимоги до розподілу даних або до числових характеристик.

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

За змістом статистичні гіпотези можна віднести до таких типів:

1) Гіпотези про вид закону розподілу досліджуваної величини.
2) Гіпотези про числові характеристики досліджуваної величини.
3) Гіпотези про рівність числових характеристик досліджуваних величин.
4) Гіпотези про належність досліджуваних величин до одній генеральної сукупності.
5) Гіпотези про вид моделі, що описує взаємозв'язок між досліджуваними величинами.
6) Гіпотези про належність досліджуваних величин до одного класу.

Статистичні гіпотези позначаються латинськими буквами H_0, H_1, і т.д. Гіпотеза H_0 формулюється як основна в тому розумінні, що при перевірці бажано було б встановити її справедливість. Основній гіпотезі H_0 протиставляються інші гіпотези H_1, H_2, …, які називаються альтернативними.

Прийняття основної або однієї з альтернативних гіпотез здійснюється на основі дослідження статистичних даних. Дослідження проводиться за певним **критерієм**, який обирається відповідно до змісту гіпотези і виду наявних статистичних даних.

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

Оскільки прийняття гіпотези здійснюється на основі статистичних даних, то завжди існує ймовірність помилки.

Ймовірність відкидання гіпотези H_0, якщо вона справедлива, називається ймовірністю помилки першого роду або **рівнем значущості** і позначається α. Величина $1-\alpha$ є ймовірністю прийняття справедливої гіпотези і називається **рівнем довіри**. Ймовірність прийняття гіпотези H_0, якщо вона не вірна, називається ймовірністю помилки другого роду і позначається β. Величина $1-\beta$ є ймовірністю відкидання невірної гіпотези і називається **потужністю критерію**.

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

Зауваження. Перевірка статистичної гіпотези не надає точного висновку щодо її вірності або невірності. Прийняття гіпотези означає, що на прийнятому рівні значущості вона не суперечить статистичним даним.

Перевірка статистичних гіпотез здійснюється за такими **етапами**:

1) Висунення припущень про вид розподілу досліджуваної величини (величин) або про її числові характеристики.

2) Формулювання статистичних гіпотез.

3) Вибір критерію перевірки відповідно до змісту гіпотез і статистичних даних.

4) Вибір рівня значущості залежно від вимог до точності результатів дослідження.

5) Розрахунок значення обраного критерію за статистичними даними.

6) Порівняння розрахованого значення критерію з його критичним значенням і прийняття або відкидання основної гіпотези.

2.2. Перевірка гіпотези про вид закону розподілу досліджуваної величини

Перевірка гіпотези про вид закону розподілу досліджуваної величини має велике значення для прикладних досліджень. Необхідність такої перевірки виникає при виборі критерію, оскільки для багатьох з них висувається вимога нормального розподілу статистичних даних. Означені гіпотези перевіряються при проектуванні систем масового обслуговування, перевірки якості продукції або праці і т. ін.

Припустимо, що з деякої генеральної сукупності X, яка розглядається як випадкова величина, обрана вибірка $\{x_1, x_2, \ldots x_n\}$. За даними вибірки побудовано статистичний ряд (табл. 2.1), що містить варіанти x_i та відповідні частоти n_i, $i = \overline{1,k}$, де k – кількість варіант у випадку дискретного ряду. У випадку інтервального ряду x_i – середини інтервалів, k – кількість інтервалів.

Таблиця 2.1

x_i	x_1	x_2	…	x_k
n_i	n_1	n_2	…	n_k

Отриманий на основі вибіркових даних статистичний ряд називається **емпіричним законом розподілу** величини X.

За даними статистичного ряду можна знайти числові характеристики, які є вибірковими параметрами закону розподілу X. Вид закону розподілу визначається відповідно до умов формування вибірки або залежно від виду графіка емпіричної щільності розподілу (гістограми) у випадку неперервної випадкової величини X і полігону частот, якщо величина X дискретна. Параметри обраного закону розподілу змінюються відповідними вибірковими параметрами.

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

При здійсненні такої заміни немає впевненості, що закон розподілу обраний правильно. Тому розроблено процедуру, яка дозволяє оцінити степінь відповідності обраного закону даним вибірки. Критерії здійснення такої перевірки називаються **критеріями згоди**, найбільш відомим з яких є **критерій Пірсона** χ^2 (хі-квадрат).

Критерій Пірсона χ^2 обчислюється за формулою:

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(n_i - n_i'\right)^2}{n_i'}, \qquad (2.1)$$

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

З формули (2.1) видно, що у випадку, коли відповідні теоретичні та емпіричні частоти співпадають, $\chi^2 = 0$. Тобто, чим ближче χ^2 до нуля, тим краще узгоджуються вибіркові дані та обраний теоретичний закон розподілу.

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

Зауваження. У деяких статистичних таблицях критичне значення χ^2 надається залежно від рівня довіри γ, а $\gamma = 1 - \alpha$.

Отже, перевірка гіпотези про закон розподілу величини X здійснюється за такими **етапами**:

1) З генеральної сукупності X формується вибірка і будується статистичний ряд.
2) Висувається гіпотеза про закон розподілу випадкової величини X.
3) Знаходяться вибіркові параметри обраного закону розподілу.
4) Розраховуються теоретичні частоти.
5) Розраховується критерій χ^2 за формулою (2.1).
6) Обирається рівень значущості α (або рівень довіри γ) і знаходиться критичне значення $\chi^2_{\alpha,l}$ (або $\chi^2_{\gamma,l}$).
7) Порівнюються розраховане і критичне значення критерію χ^2 і робиться висновок про справедливість запропонованої гіпотези.

Приклад 2.1. За даним інтервальним статистичним рядом (табл. 2.2) знайти закон розподілу випадкової величини X.

Таблиця 2.2

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

Розв'язок. Для визначення виду закону розподілу побудуємо гістограму за даними табл. 2.2 (рис. 2.1). За видом гістограми висуваємо гіпотезу про нормальний закон розподілу даної випадкової величини:

H_0 – випадкова величина X розподілена за нормальним законом;

H_1 – випадкова величина X не розподілена за нормальним законом.

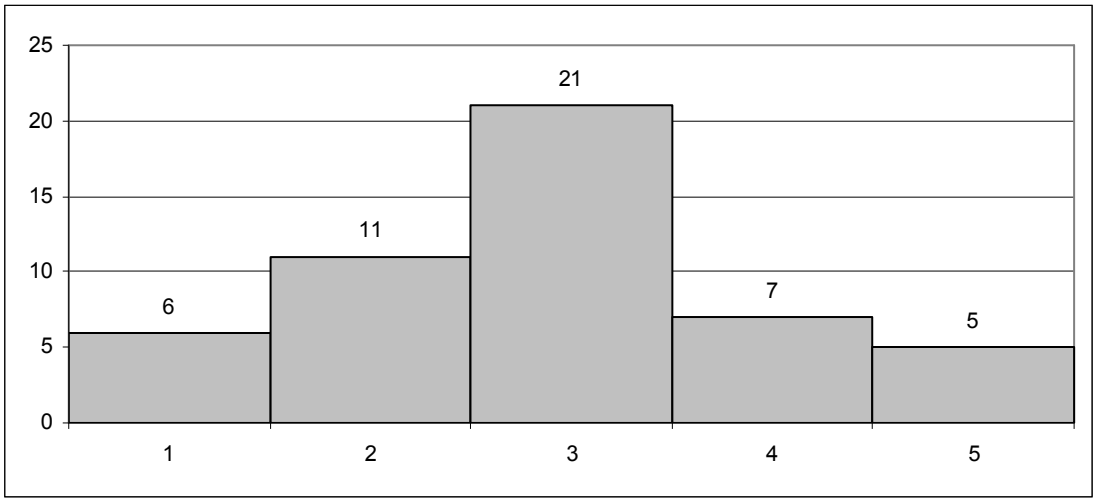


Рисунок 2.1. Гістограма за даними табл. 2.2

Щільність розподілу випадкової величини, розподіленої за нормальним законом, має вигляд $f(x) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}$, де a і σ – параметри розподілу. Знайдемо означені параметри, враховуючи, що $\bar{x} = a$; $S^2 = \sigma^2$. Розрахунки оформимо у вигляді таблиці (табл. 2.3).

Таблиця 2.3

$[a_i; a_{i+1})$	$[-2;-1,2)$	$[-1,2;-0,4)$	$[-0,4;0,4)$	$[0,4;1,2)$	$[1,2;2)$
n_i	6	11	21	7	5
x_i	−1,6	0,8	0	0,8	1,6
$x_i n_i$	−9,6	8,8	0	5,6	8
$\left(x_i - \bar{x}\right)^2 n_i$	13,572	5,452	0,194	5,620	14,382

Знайдемо вибіркове середнє, вибіркову дисперсію і вибіркове середнє квадратичне відхилення за формулами (1.6), (1.13) та (1.17) відповідно:

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

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

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

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

Для знаходження значення критерію χ^2 розрахуємо теоретичні частоти n_i'. Теоретичні частоти можна знайти за формулою $n_i' = np_i$, де p_i – ймовірності попадання випадкової величини в певний інтервал. Для нормального закону розподілу означені ймовірності знаходяться за формулою $P(a_i < X < a_{i+1}) = p_i = \frac{1}{2}\left[\Phi\left(\frac{a_{i+1} - \bar{x}}{S}\right) - \Phi\left(\frac{a_i - \bar{x}}{S}\right)\right]$, де Φ – функція Лапласа, значення якої представлені у статистичних таблицях. Для зручності обчислень побудуємо таблицю (табл. 2.4).

За формулою (2.1) маємо: $\chi^2 = \sum_{i=1}^{k} \frac{\left(n_i - n_i'\right)^2}{n_i'} \approx 4{,}16$. Знайдемо критичне значення $\chi^2_{\alpha,l}$, враховуючи, що $l = k - r - 1 = 5-2-1 = 2$. Рівень значущості α оберемо рівним 0,1. За допомогою Excel знаходимо ХИ2ОБР(0,1; 2) = 4,6.

Отже, оскільки $\chi^2 < \chi^2_{\alpha,l}$, гіпотеза H_0 про нормальний розподіл приймається, гіпотеза H_1 відкидається.

Таблиця 2.4

$[a_i; a_{i+1})$	[−2;−1,2)	[−1,2;−0,4)	[−0,4;0,4)	[0,4;1,2)	[1,2;2)
n_i	6	11	21	7	5
x_i	−1,6	0,8	0	0,8	1,6
$x_i n_i$	−9,6	8,8	0	5,6	8
$\left(x_i - \bar{x}\right)^2 n_i$	13,572	5,452	0,194	5,620	14,382
$\frac{a_i - \bar{x}}{S}$	−2,1498	−1,2465	−0,3433	0,56	1,4633
$\Phi(\frac{a_i - \bar{x}}{S})$	−0,958	−0,785	−0,266	0,425	0,856
p_i	0,0856	0,2595	0,3455	0,2155	0,063
$n_i' = np_i$	4,325	12,975	17,275	10,775	3,15
$\frac{\left(n_i - n_i'\right)^2}{n_i'}$	0,649	0,301	0,803	1,323	1,087

Приклад 2.2. На одній з міських АТС фіксувалася кількість телефонних дзвінків в годину. Спостереження велися на протязі 100 годин, їх результати представлені в табл. 2.5. Чи можна вважати навантаження на АТС стандартним?

Таблиця 2.5

Кількість викликів в годину	0	1	2	3	4	5	6	7
Кількість спостережень	6	27	26	20	10	5	5	1

Розв'язок. Навантаження на АТС можна вважати стандартним, якщо випадкова величина X – кількість телефонних дзвінків, що поступили, підкоряється закону розподілу Пуассона. Сформулюємо гіпотези:

H_0 – випадкова величина X підкоряється закону розподілу Пуассона;

H_1 – випадкова величина X не підкоряється закону розподілу Пуассона.

Закон Пуассона має вигляд: $p_k = \dfrac{e^{-\lambda}\lambda^k}{k!}$, $k = 0,1,\ldots; \lambda > 0$, де λ – параметр розподілу. Крім того, відомо, що $\bar{x} = \lambda;\ S^2 = \lambda$. Отже, для встановлення параметра λ потрібно знайти $\bar{x}$ або S^2.

Випадкова величина X – кількість викликів в годину; тоді кількість спостережень – це відповідні значенням X частоти n_i, а табл. 2.5 є статистичним рядом і емпіричним законом розподілу величини X. Знайдемо $\bar{x}$ і S^2. Для зручності обчислення оформимо у вигляді таблиці (табл. 2.6).

Таблиця 2.6

x_i	0	1	2	3	4	5	6	7	Суми
n_i	6	27	26	20	10	5	5	1	100
$x_i\, n_i$	0	27	52	60	40	25	30	7	241
$\left(x_i - \bar{x}\right)^2 n_i$	34,85	53,68	4,37	6,96	25,28	33,54	64,44	21,07	244,19

Отже, $\bar{x} = \dfrac{1}{n}\sum_{i=1} x_i n_i = \dfrac{1}{100}\cdot 241 = 2,41$; $S^2 = \dfrac{1}{n}\sum_{i=1}(x_i - \bar{x})^2 n_i = \dfrac{1}{100}\cdot 244,19 \approx 2,44$.

Оскільки повинна виконуватися рівність $\bar{x} = \lambda; S^2 = \lambda$, то як параметр можна вибрати або $\bar{x}$, або S^2, або їх середнє арифметичне. Виберемо $\lambda = \dfrac{\bar{x} + S^2}{2} \approx 2,426$.

Таким чином, гіпотеза H_0 – це припущення, що величина X розподілена згідно із законом Пуассона: $p_k = \dfrac{e^{-2,426}\,2,426^k}{k!}$, $k = 0,1,\ldots$

Перевіримо правильність гіпотези за допомогою критерія Пірсона. Знайдемо теоретичні частоти, використовуючи формулу: $p_k = \dfrac{e^{-2,426}\,2,426^k}{k!}$, $k = 0,1,\ldots$ Через k позначимо значення X, тобто x_i.

Відмітимо, що p_i – ймовірності того, що X прийме значення x_i, тобто статистично вони є відносними частотами, теоретичні ж частоти знаходитимемо за формулою: $n_i' = np_i$.

Для зручності при обчисленні теоретичних частот продовжимо таблицю, складену на основі статистичного ряду (табл. 2.7).

Отже, за результатами розрахунків $\chi^2 = \sum_{i=0}^{7} \dfrac{\left(n_i - n_i'\right)^2}{n_i'} = 5,739$.

Для даного завдання $l = k - r - 1$=8−1−1=6. Виберемо рівень значущості α=0,01 і знайдемо за допомогою таблиць або функції ХИ2ОБР табличного процесора Excel значення $\chi^2_{\alpha,l}$: $\chi^2_{0,01;6}$= 16,812. Оскільки для такого рівня довіри (0,99) $\chi^2 < \chi^2_{\alpha,l}$, то гіпотезу H_0 про розподіл Пуассона можна прийняти.

Таблиця 2.7

x_i	0	1	2	3	4	5	6	7	Суми
n_i	6	27	26	20	10	5	5	1	100
$p_i = \frac{e^{-2,426}2,426^{x_i}}{x_1!}$	0,09	0,21	0,26	0,21	0,13	0,06	0,03	0,01	
$n_i' = np_i$	8,84	21,44	26,01	21,03	12,76	6,19	2,50	0,87	
$\frac{\left(n_i - n_i'\right)^2}{n_i'}$	0,91	1,44	4,65E-06	0,05	0,59	0,23	2,49	0,02	5,739

Висновок: навантаження на АТС можна вважати стандартним.

Приклад 2.3. З метою впорядкування роботи міського суспільного транспорту фіксувався час очікування в хвилинах пасажирами тролейбусів на декількох маршрутах. Було проведено 200 вимірювань, їх результати представлені в табл. 2.8. Чи можна вважати, що перевезення по перевірених маршрутах забезпечені раціонально?

Таблиця 2.8

Час очікування	1 – 3	3 – 5	5 – 7	7 – 9	9 – 11	11 – 13
Кількість спостережень	25	30	48	35	42	20

Розв'язок. Можна вважати, що перевезення по перевірених маршрутах забезпечені раціонально, якщо випадкова величина X – час очікування пасажирами транспорту – підкоряється рівномірному закону розподілу. Тобто задача зводиться до перевірки гіпотези про закон розподілу випадкової величини. Сформулюємо гіпотези:

H_0 – випадкова величина X розподілена рівномірно;

H_1 – випадкова величина X не розподілена рівномірно.

Щільність розподілу випадкової величини, що підкоряється рівномірному закону: $f(x) = \begin{cases} \frac{1}{b-a}, a \le x \le b \\ 0, \quad x < a, x > b \end{cases}$, де a; b − параметри розподілу. Крім того, відомо, що $\bar{x} = \frac{a+b}{2}$; $S^2 = \frac{(b-a)^2}{12}$. Тобто для встановлення параметрів a і b потрібно знайти $\bar{x}$ і S^2, після чого розв'язати систему: $\begin{cases} \bar{x} = \frac{a+b}{2} \\ S^2 = \frac{(b-a)^2}{12} \end{cases}$.

Оскільки за умов задачі випадкова X – час очікування транспорту, то кількість спостережень – це відповідні значенням X частоти n_i, а табл. 2.8 – це інтервальний статистичний ряд і емпіричний закон розподілу X. Знайдемо $\bar{x}$ і S^2. Через x_i візьмемо середини відповідних інтервалів. Для зручності обчислення оформимо у вигляді таблиці (табл. 2.9).

Таблиця 2.9

$[a_i; a_{i+1})$	1 – 3	3 – 5	5 – 7	7 – 9	9 – 11	11 – 13	Суми
x_i	2	4	6	8	10	12	
n_i	25	30	48	35	42	20	200
$x_i\, n_i$	50	120	288	280	420	240	1398
$\left(x_i - \bar{x}\right)^2 n_i$	622,5	268,2	47,045	35,704	380,52	502	1856

Отже, $\bar{x} = \frac{1}{n}\sum_{i=1}^{6} x_i n_i = \frac{1}{200}\cdot 1398 = 6{,}99$; $S^2 = \frac{1}{n}\sum_{i=1}^{6}(x_i - \bar{x})^2 n_i = \frac{1}{200}\cdot 1856 \approx 9{,}2799$.

Складемо систему для визначення параметрів рівномірного розподілу і розв'яжемо її:

$$\begin{cases} 6{,}99 = \dfrac{a+b}{2} \\ 9{,}2799 = \dfrac{(b-a)^2}{12} \end{cases} \Rightarrow \begin{cases} a+b = 13{,}98 \\ (b-a)^2 = 111{,}3588 \end{cases} \Rightarrow \begin{cases} a+b = 13{,}98 \\ b-a \approx 10{,}55 \end{cases} \Rightarrow \begin{cases} a \approx 1{,}715 \\ b \approx 12{,}265 \end{cases}.$$

Таким чином, гіпотеза H_0 – це припущення, що X розподілена за рівномірним законом із щільністю розподілу:

$$f(x) = \begin{cases} \dfrac{1}{12{,}265 - 1{,}715}, & 1{,}715 \le x \le 12{,}265 \\ 0, & x < 1{,}715, \quad x > 12{,}265 \end{cases}.$$

Перевіримо справедливість гіпотези H_0 за допомогою критерію Пірсона. Для знаходження теоретичних частот використаємо формулу $n_i' = np_i$, а ймовірності попадання в інтервали p_i знайдемо за формулою: $P(\alpha < X < \beta) = \dfrac{\beta - \alpha}{b - a}$.

Для зручності обчислень теоретичних частот складемо таблицю (табл. 2.10). Врахуємо, що $\beta - \alpha = 2$; $b - a = 12{,}265 - 1{,}715 = 10{,}55$ для всіх інтервалів, окрім першого і останнього. Для першого інтервалу $\beta - \alpha = \beta - a = 3 - 1{,}715 = 1{,}285$; для останнього інтервалу $\beta - \alpha = b - \alpha = 12{,}265 - 11 = 1{,}265$.

Таблиця 2.10

$[a_i; a_{i+1})$	1 – 3	3 – 5	5 – 7	7 – 9	9 – 11	11 – 13	Суми
n_i	25	30	48	35	42	20	200
$\beta - \alpha$	1,285	2	2	2	2	1,265	
$p_i = \frac{\beta - \alpha}{10,55}$	0,122	0,19	0,19	0,19	0,19	0,12	
$n_i' = np_i$	24,360	37,915	37,915	37,915	37,915	23,981	
$\frac{\left(n_i - n_i'\right)^2}{n_i'}$	0,017	1,6522	2,6827	0,2241	0,4402	0,6609	5,677

Отже, $\chi^2 = \sum_{i=0}^{7} \frac{\left(n_i - n_i'\right)^2}{n_i'} = 5,677$.

Для даного завдання $l = k - r - 1$=6−2−1=3. Виберемо рівень значущості $\alpha = 0,01$ і знайдемо за допомогою таблиць або функції ХИ2ОБР табличного процесора Excel значення $\chi^2_{\alpha,l}$: $\chi^2_{0,01;3}$=11,34. Оскільки для такого рівня довіри (0,99) $\chi^2 < \chi^2_{\alpha,l}$, гіпотезу про рівномірний розподіл приймаємо.

Висновок: перевезення по перевірених маршрутах організовані раціонально.

2.3. Перевірка гіпотез про генеральні середні і дисперсії

В прикладних задачах часто виникає необхідність перевірки рівності середніх значень та дисперсій за даними двох або більше вибірок. Наприклад, коли визначається перевага однієї з технологій виготовлення певної продукції, або наявність підвищення продуктивності праці після внесення змін в процес виробництва, або при перевірці якості продукції. Здійснення означеної перевірки виконується за критеріями, що обираються залежно від виду розподілу вибіркових даних і мети дослідження. Для деяких критеріїв перевірки рівності середніх значень висувається додаткова вимога − про рівність генеральних дисперсій.

2.3.1. Перевірка гіпотези про рівність генеральних дисперсій. *F*-критерій (Фішера)

Перевірка гіпотези про рівність генеральних дисперсій здійснюється за *F*-критерієм (Фішера) тільки тоді, коли статистичні дані незалежні і розподілені за нормальним законом. Формулюються гіпотези:

H_0 − дисперсії двох нормально розподілених генеральних сукупностей рівні, тобто $S_1^2 = S_2^2$;

H_1 − дисперсії двох нормально розподілених генеральних сукупностей не рівні, тобто $S_1^2 \neq S_2^2$.

F-критерій (Фішера) розраховується за формулою:

$$F = \frac{S_1^2}{S_2^2}, \quad S_1^2 > S_2^2. \tag{2.2}$$

Гіпотеза H_0 приймається, якщо розраховане значення F менше критичного значення розподілу Фішера $F_{крит}$, взятого із рівнем значущості α і степенями свободи l_1 та l_2 для чисельника і знаменника відповідно: $l_1 = n_1 - 1$, $l_2 = n_2 - 1$, де n_1, n_2 – об'єми вибірок. $F_{крит}$ можна знайти за допомогою вбудованої статистичної функції Excel FРАСПОБР (α; l_1; l_2).

Зауваження. Дисперсія у чисельнику дробу в формулі (2.2) повинна бути більшою дисперсії у знаменнику, тобто значення F-критерія повинно бути більше одиниці.

Приклад 2.4. Відомо дані про продуктивність праці (одиниць продукції за зміну) двох груп працівників: група 1 складається з працівників, що пройшли спеціальний навчальний курс; група 2 – із працівників, що не пройшли курсу (табл. 2.11). Враховуючи, що дані розподілені за нормальним законом, перевірити гіпотезу про рівність дисперсій.

Таблиця 2.11

	Група 1					Група 2				
Продуктивність праці	34	85	96	102	103	63	69	83	89	106
Кількість працівників	5	2	11	8	4	2	6	8	3	1

Розв'язок. Дані табл. 2.11 є двома вибірками. Перша – вибірка значень величини X_1 – продуктивність праці робітників, що пройшли навчання, друга – вибірка величини X_2 – продуктивність праці робітників, що не пройшли навчання.

Сформулюємо гіпотези: H_0 – дисперсії генеральних сукупностей, з яких зроблено вибірки, рівні, $S_1^2 = S_2^2$; H_1 − дисперсії не рівні, $S_1^2 \neq S_2^2$. Перевіримо справедливість гіпотези H_0 за F-критерієм (Фішера).

Знайдемо за вибірковими даними оцінку дисперсії X_1 за формулою (1.14). Розрахунки оформимо у вигляді таблиці (табл. 2.12).

Таблиця 2.12

x_i	34	85	96	102	103	Суми
n_i	5	2	11	8	4	30
$x_i\, n_i$	170	170	1056	816	412	2624
$(x_i - \overline{x})^2 n_i$	14293,42	12,17	801,00	1689,74	965,14	17761,47

Отже, $\overline{x_1} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i = \frac{1}{30} \cdot 2624 \approx 87,47$;

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

Аналогічно знайдемо оцінку дисперсії X_2 (табл. 2.13).

Таблиця 2.13

x_i	63	69	83	89	106	Суми
n_i	2	6	8	3	1	20
$x_i\, n_i$	126	414	664	267	106	1577
$(x_i - \bar{x})^2 n_i$	502,45	582,13	137,78	309,07	737,12	2268,55

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

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

Знайдемо значення *F*-критерію за формулою (2.2). Оскільки $S_1^2 > S_2^2$, то $F = \frac{S_1^2}{S_2^2} = \frac{612{,}46}{119{,}40} \approx 5{,}13$. Знайдемо $F_{\text{крит}}$, враховуючи, що $l_1 = n_1 - 1 = 30 - 1 = 29$; $l_2 = n_2 - 1 = 20 - 1 = 19$. Рівень значущості оберемо $\alpha = 0{,}05$. Тоді $F_{\text{крит}}$= FРАСПОБР (0,05; 29; 19) = 2,077.

Оскільки $F > F_{\text{крит}}$, то гіпотезу H_0 відкидаємо і приймаємо гіпотезу H_1 – дисперсії нерівні, тобто вибірки здобуті з різних генеральних сукупностей.

Висновок: навчальний курс суттєво впливає на продуктивність праці робітників.

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

Якщо статистичні дані не розподілені за нормальним законом або вимірюються з використанням порядкової шкали, то перевірка гіпотези про рівність генеральних дисперсій здійснюється за критерієм Зігеля-Тьюкі. Формулюються гіпотези:

H_0 – дисперсії двох генеральних сукупностей рівні, тобто $S_1^2 = S_2^2$;

H_1 − дисперсії двох генеральних сукупностей не рівні, тобто $S_1^2 \neq S_2^2$.

Перевірка виконується за даними двох вибірок у такій послідовності:

1) Формується об’єднана вибірка.

2) Даним об’єднаної вибірки присвоюються ранги (порядкові номери) за правилом: найменшому значенню присвоюється ранг 1, двом найбільшим – ранги 2 і 3; наступним двом найменшим – ранги 4 і 5; наступним найбільшим – ранги 6 і 7 і т. д. При цьому, якщо кількість елементів вибірки непарна, то її центральний елемент (тобто медіана) не отримує ніякого рангу.

3) Розраховуються суми рангів елементів вихідних вибірок R_1 *і* R_2.

4) Розраховується нормальна випадкова величина *Z* за формулою:

$$Z = \frac{2R_1 - n_1\left(n_1 + n_2 + 1\right) + 1}{\frac{n_2}{3} \cdot \sqrt{n_1\left(n_1 + n_2 + 1\right)}}, \qquad (2.3)$$

де n_1, n_2 – об’єми вибірок. При цьому R_1 – сума рангів меншої за об’ємом вибірки. Якщо $2R_1 > n_1\left(n_1 + n_2 + 1\right) + 1$, *Z* розраховується за формулою:

$$Z = \frac{2R_1 - n_1\left(n_1 + n_2 + 1\right) - 1}{\frac{n_2}{3} \cdot \sqrt{n_1\left(n_1 + n_2 + 1\right)}}. \qquad (2.4)$$

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

$$Z' = Z + \left(\frac{1}{10n_1} - \frac{1}{10n_2}\right)\left(Z^3 - 3Z\right). \qquad (2.5)$$

6) Обирається рівень значущості α.

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

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

Зауваження. Для перевірки правильності присвоєння рангів можна скористатися формулами: $R_1 + R_2 = \frac{\left(n_1 + n_2\right)\left(n_1 + n_2 + 1\right)}{2}$ у випадку парної кількості елементів об'єднаної вибірки; $R_1 + R_2 = \left(n_1 + n_2\right)\left(\frac{n_1 + n_2 + 1}{2} - 1\right)$ у випадку непарної кількості цих елементів.

Приклад 2.5. У результаті дослідження надійності станків двох виробників отримано дані про час (в годинах) безаварійної роботи (табл. 2.14). Враховуючи, що дані не розподілені за нормальним законом, перевірити гіпотезу про рівність дисперсій.

Таблиця 2.14

Виробник	Час безаварійної роботи									
1	280	230	112	176	90	175	216	110	205	115
2	200	126	225	210	260	194	156	240	170	232

Розв'язок. Дані таблиці 2.14 є двома вибірками. Перша – вибірка значень величини X_1 – час безаварійної роботи станків виробника 1; друга – вибірка величини X_2 – час безаварійної роботи станків виробника 2.

Сформулюємо гіпотези: H_0 – дисперсії генеральних сукупностей, з яких зроблено вибірки, рівні: $S_1^2 = S_2^2$; H_1 − дисперсії не рівні: $S_1^2 \neq S_2^2$. Перевіримо справедливість гіпотези H_0 за критерієм Зігеля-Тьюкі.

Сформуємо об'єднану вибірку, присвоїмо її елементам ранги і знайдемо їх суму. Результати розрахунків оформимо у вигляді таблиці (табл. 2.15). Для зручності підкреслимо елементи першої вибірки.

Розрахуємо за формулою (2.3) значення Z, враховуючи, що $n_1 = n_2 = 10$:

$$Z = \frac{2R_1 - n_1\left(n_1 + n_2 + 1\right) + 1}{\frac{n_2}{3} \cdot \sqrt{n_1\left(n_1 + n_2 + 1\right)}} = \frac{2 \cdot 95 - 10(10 + 10 + 1) + 1}{\frac{10}{3} \cdot \sqrt{10(10 + 10 + 1)}} \approx \frac{-19}{48{,}26} \approx -0{,}394.$$

Оберемо рівень значущості $\alpha = 0{,}05$. За допомогою вбудованої функції Excel НОРМРАСП знаходиться ймовірність $P(Z)$:

$P(Z)$ = НОРМРАСП (−0,394; 0; 1; ИСТИНА) = 0,3469.

Оскільки $2P(Z) = 2 \cdot 0{,}3469 = 0{,}6938 > \alpha = 0{,}05$, то гіпотеза H_0 про рівність генеральних дисперсій приймається.

Висновок: дисперсії надійності станків двох виробників однакові.

Таблиця 2.15

Елементи об'єднаної вибірки	Сортована об'єднана вибірка	Ранги елементів об'єднаної вибірки	Ранги елементів першої вибірки	Ранги елементів другої вибірки
280	90	1	1	
230	110	4	4	
112	112	5	5	
176	115	8	8	
90	126	9		9
175	156	12		12
216	170	13		13
110	175	16	16	
205	176	17	17	
115	194	20		20
200	200	19		19
126	205	18	18	
225	210	15		15
210	216	14	14	
260	225	11		11
194	230	10	10	
156	232	7		7
240	240	6		6
170	260	3		3
232	280	2	2	
		Суми	95	115

2.3.3. Перевірка гіпотези про рівність генеральних середніх. Критерій Стьюдента

Критерій Стьюдента використовується для перевірки гіпотез про рівність генеральних середніх, якщо статистичні дані розподілені за нормальним законом. Формулюються гіпотези:

H_0 – середні двох генеральних сукупностей рівні, тобто $\overline{x_1} = \overline{x_2}$;

H_1 − середні двох генеральних сукупностей не рівні, тобто $\overline{x_1} \neq \overline{x_2}$.

Перевірка виконується за даними двох вибірок об'ємом n_1 та n_2. При цьому можливі такі випадки.

Випадок 1. Генеральні дисперсії рівні ($S_1^2 = S_2^2$). Тоді *t*-критерій Стьюдента обчислюється за формулою:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}. \quad (2.6)$$

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

Випадок 2. Генеральні дисперсії не рівні ($S_1^2 \neq S_2^2$). Тоді *t*-критерій Стьюдента обчислюється за формулою:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}. \quad (2.7)$$

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

$$l = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 + 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 + 1}} + 2. \quad (2.8)$$

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

$$t = \frac{\overline{x_d}}{S_d / \sqrt{n-1}}. \quad (2.9)$$

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

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

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

$$|t| < t_{крит}.$$

ПРИКЛАД 2.6. Для виробництва кожної з 10 деталей за першою технологією було витрачено, у середньому, 30 с. Дисперсія часу складала 1 с^2. Для виробництва кожної з 16 деталей за другою технологією було витрачено, у середньому, 28 с із дисперсією часу 2 с^2. Чи можна вважати, що у середньому, для виробництва деталей за першою технологією потрібно більше часу?

Розв'язок. За умовами задачі було зроблено дві вибірки: перша – вибірка об'єму $n_1 = 10$ значень величини X_1 – часу, потрібного для виготовлення деталей за першою технологією; друга – вибірка об'єму $n_2 = 16$ значень величини X_2 – часу, потрібного для виготовлення деталей за другою технологією. Відомі вибіркові середні $\overline{x_1} = 30$ с та $\overline{x_2} = 28$ с – середній час, необхідний для виготовлення деталей за першою і другою технологіями відповідно. Відомі дисперсії часу для вибірок: $S_1^2 = 1$ с^2 та $S_2^2 = 2$ с^2. Потрібно перевірити гіпотезу про рівність генеральних середніх.

Сформулюємо гіпотези:

H_0 – середні двох генеральних сукупностей рівні, тобто $\overline{x_1} = \overline{x_2}$;

H_1 − середні двох генеральних сукупностей не рівні, тобто $\overline{x_1} \neq \overline{x_2}$.

Перед вибором критерію для перевірки потрібно встановити, чи рівні генеральні дисперсії. Використаємо критерій Фішера. За формулою (2.2) обчислимо значення F-критерія:

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

Знайдемо критичне значення розподілу Фішера $F_{крит}$: оберемо рівень значущості α=0,05; врахуємо, що степені свободи $l_1 = n_1 - 1 = 9$ та $l_2 = n_2 - 1 = 15$. Тоді FРАСПОБР (α; l_2; l_1) = FРАСПОБР (0,05; 15; 9) = 3,006. Отже, $F < F_{крит}$, тому генеральні дисперсії можна вважати рівними.

Оскільки генеральні дисперсії рівні (випадок 1), то t-критерій Стьюдента розраховуємо за формулою (2.6):

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{30 - 28}{\sqrt{\frac{10 \cdot 1 + 16 \cdot 2}{10 + 16 - 2}\left(\frac{1}{10} + \frac{1}{16}\right)}} \approx 7{,}032.$$

Знайдемо критичне значення розподілу Стьюдента $t_{крит}$, враховуючи, що

$$l = n_1 + n_2 - 2 = 10 + 16 - 2 = 24.$$

Оберемо значення α=0,05. Тоді

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

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

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

Приклад 2.7. Для виробництва кожної з 51 деталі за першою технологією було витрачено, у середньому, 30 с. Дисперсія часу складала 6 с^2. Для виробництва кожної з 41 деталей за другою технологією було витрачено, у середньому, 25 с із дисперсією часу 3 с^2. Чи можна вважати, що у середньому для виробництва деталей за першою технологією потрібно більше часу?

Розв'язок. За умовами задачі було сформовано дві вибірки: перша – вибірка об'єму $n_1 = 51$ значень величини X_1 – часу, потрібного для виготовлення деталей за першою технологією; друга – вибірка об'єму $n_2 = 41$ значень величини X_2 – часу, потрібного для виготовлення деталей за другою технологією. Відомі вибіркові середні $\overline{x_1} = 30$ с та $\overline{x_2} = 25$ с – середній час, необхідний для виготовлення деталей за першою і другою технологіями відповідно. Відомі дисперсії часу для вибірок: $S_1^2 = 6$ с^2 та $S_1^2 = 3$ с^2. Потрібно перевірити гіпотезу про рівність генеральних середніх.

Сформулюємо гіпотези:

H_0 – середні двох генеральних сукупностей рівні, тобто $\overline{x_1} = \overline{x_2}$;

H_1 − середні двох генеральних сукупностей не рівні, тобто $\overline{x_1} \neq \overline{x_2}$.

Перед вибором критерію для перевірки потрібно встановити, чи рівні генеральні дисперсії. Скористаємось критерієм Фішера. За формулою (2.2) обчислимо значення F-критерію: оскільки $S_1^2 > S_2^2$, то

$$F = \frac{S_1^2}{S_2^2} = \frac{6}{3} = 2.$$

Знайдемо критичне значення розподілу Фішера $F_{крит}$: оберемо рівень значущості $\alpha = 0{,}05$; врахуємо, що степені свободи $l_1 = n_1 - 1 = 50$ та $l_2 = n_2 - 1 = 40$. Тоді FРАСПОБР (α; l_1; l_2) = FРАСПОБР (0,05; 50; 40) = 1,66.

Отже, $F > F_{крит}$, тому генеральні дисперсії можна вважати різними.

Оскільки генеральні дисперсії різні (випадок 2), то t-критерій Стьюдента розраховуємо за формулою (2.7):

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{30 - 25}{\sqrt{\frac{6}{51} + \frac{3}{41}}} = \frac{5}{\sqrt{0{,}19077}} \approx 11{,}45.$$

Знайдемо критичне значення розподілу Стьюдента $t_{крит}$. Оберемо рівень значущості $\alpha = 0{,}05$. Обчислимо степені свободи за формулою (2.8):

$$l = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 + 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 + 1}} + 2 = \frac{\left(\frac{6}{51} + \frac{3}{41}\right)^2}{\frac{\left(\frac{6}{51}\right)^2}{51 + 1} + \frac{\left(\frac{3}{41}\right)^2}{41 + 1}} + 2 \approx 93.$$

Тоді

$t_{\text{крит}}$ = СТЬЮДРАСПОБР ($\frac{\alpha}{2}$; l) = СТЬЮДРАСПОБР (0,025;93) = 2,27.

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

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

Приклад 2.8. Проводилося дослідження продуктивності праці двох механіків з регулювання станків-автоматів. Вимірювався час роботи станків до першої зупинки для регулювання (табл. 2.16). Чи можна вважати, що механіки працюються однаково продуктивно?

Таблиця 2.16

Номер станка	1	2	3	4	5
Час роботи після регулювання першим механіком (год.)	60,2	62,3	61,3	60,7	63,4
Час роботи після регулювання другим механіком (год.)	59,4	58,3	62,1	63,4	60,8

Розв'язок: За умовами задачі було зроблено дві вибірки: перша – вибірка об'єму $n_1 = 5$ значень величини X_1 – часу роботи станків після регулювання першим механіком; друга – вибірка об'єму $n_2 = 5$ значень величини X_2 – часу роботи станків після регулювання другим механіком. Потрібно перевірити гіпотезу про рівність генеральних середніх.

Сформулюємо гіпотези:

H_0 – середні двох генеральних сукупностей рівні, тобто $\overline{x_1} = \overline{x_2}$;

H_1 − середні двох генеральних сукупностей не рівні, тобто $\overline{x_1} \neq \overline{x_2}$.

Оскільки вибірки не є незалежними, то необхідно сформувати парну вибірку (випадок 3), знайти для кожної пари елементів різницю їх значень d, обчислити для парної вибірки $\overline{x_d}$ і S_d та розрахувати t-критерій Стьюдента за формулою (2.9). Розрахунки для зручності оформимо у вигляді таблиці (табл. 2.17).

Таблиця 2.17

$\overline{x_{1i}}$	60,2	62,3	61,3	60,7	63,4	Суми
$\overline{x_{2i}}$	59,4	58,3	62,1	63,4	60,8	
$d = \overline{x_{1i}} - \overline{x_{2i}}$	0,8	4	−0,8	−2,7	2,6	3,9
$(d - \overline{x_d})^2$	0,0004	10,3684	2,4964	12,1104	3,3124	28,288

Знайдемо $\overline{x_d}$ за формулою (1.7): $\overline{x_d} = \frac{1}{n}\sum_{i=1}^{n} d = \frac{3,9}{5} = 0,78$. Знайдемо $S_d{}^2$ за формулою (1.15): $S_d{}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(d - \overline{x_d}\right)^2 = \frac{28,288}{5} = 5,6576$. Тоді $S_d = \sqrt{S_d{}^2} \approx 2,379$.

Знайдемо t-критерій Стьюдента за формулою (2.9):

$$t = \frac{\overline{x_d}}{S_d / \sqrt{n-1}} = \frac{0,78}{2,379 / \sqrt{5-1}} \approx 0,656.$$

Знайдемо критичне значення $t_{крит}$ розподілу Стьюдента. Оберемо рівень значущості $\alpha = 0,05$, врахуємо, що степінь свободи $l = n - 1 = 4$. Тоді $t_{крит}$ = СТЬЮДРАСПОБР ($\frac{\alpha}{2}$; l) = СТЬЮДРАСПОБР (0,025; 4) = 3,495.

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

Висновок: продуктивність праці двох механіків однакова.

2.3.4. Перевірка гіпотези про рівність генеральних середніх. Критерій Уілкоксона

Критерій Уілкоксона використовується для перевірки гіпотези про рівність генеральних середніх за статистичними даними двох вибірок тоді, коли дані не підкоряються нормальному закону розподілу.

Формулюються гіпотези:

H_0 – середні двох генеральних сукупностей рівні, тобто $\overline{x_1} = \overline{x_2}$;

H_1 − середні двох генеральних сукупностей не рівні, тобто $\overline{x_1} \neq \overline{x_2}$.

Для здійснення перевірки необхідно, щоб об'єм першої вибірки n_1 був менший об'єму другої вибірки n_2. Якщо ця умова не виконується, то слід змінити місця вибірок, тобто першу вважати другою. Перевірка виконується за такими етапами:

1) Формується об'єднана вибірка, об'єм якої $n = n_1 + n_2$.

2) Даним об'єднаної вибірки присвоюються ранги (порядкові номери) за правилом: найменшому значенню присвоюється ранг 1, наступному найменшому – ранг 2 і т. д. При цьому, якщо деякі елементи вибірки співпадають, то їм присвоюються середні ранги. Для цього додаються ранги, які мали б ці елементи, якщо були б різні; розраховується їх середнє арифметичне; кожному із однакових елементів присвоюється ранг, що дорівнює розрахованому середньому арифметичному.

3) Розраховуються суми рангів елементів вихідних вибірок R_1 *і* R_2.

4) Обирається величина W: якщо вибірки рівні за об'ємом, то $W = R_1$ або $W = R_2$; якщо вибірки різні за об'ємом, то $W = R_1$, де R_1 – сума рангів меншої за об'ємом вибірки.

5) Розраховується значення W^* критерію Уілкоксона за формулою:

$$W^* = \frac{2W - n_1(n_1 + n_2 + 1) + 1}{\sqrt{\frac{n_1 n_2}{3} \cdot (n_1 + n_2 + 1)}}. \qquad (2.10)$$

6) Якщо у вибірці є дані з однаковими рангами – зв'язки, то критерій Уілкоксона W^* розраховується за формулою:

$$W^* = \frac{2W - n_1\left(n_1 + n_2 + 1\right) + 1}{\sqrt{\frac{n_1 n_2}{12} \cdot \left(n_1 + n_2 + 1\right) - \frac{\sum_{i=1}^{m} t_i\left(t_i^2 - 1\right)}{\left(n_1 + n_2\right)\left(n_1 + n_2 - 1\right)}}}, \qquad (2.11)$$

де m – кількість груп однакових рангів, що містять дані обох вибірок (кількість загальних зв'язок), t_i – кількість елементів в i-тій групі (розмір зв'язок).

6) Обирається рівень значущості α.

7) Розраховується найменший рівень значущості $p_{гр}$ за формулою:

$$p_{гр} = 2 \cdot \left(1 - \Phi\left(\left|W^*\right|\right)\right), \qquad (2.12)$$

де $\Phi(u)$ – функція нормального стандартного розподілу, значення якої можна знайти в статистичних таблицях або за допомогою вбудованої функції Excel НОРМРАСП: $p_{гр}$ = НОРМРАСП (W^*; 0; 1; ИСТИНА).

8) Порівнюються рівень значущості α і величина $p_{гр}$. Якщо $p_{гр} > \alpha$, то гіпотеза H_0 про рівність генеральних дисперсій приймається.

Приклад 2.9. У результаті дослідження надійності станків двох виробників отримано дані про час (в годинах) безаварійної роботи (табл. 2.18). Враховуючи, що дані не розподілені за нормальним законом, перевірити гіпотезу про рівність середніх.

Таблиця 2.18

Виробник	Час безаварійної роботи									
1	280	230	112	176	90	175	216	110	205	115
2	200	126	225	210	260	194	156	240	170	232

Розв'язок. Дані табл. 2.18 є двома вибірками. Перша – вибірка значень величини X_1 – часу безаварійної роботи станків виробника 1; друга – вибірка величини X_2 – часу безаварійної роботи станків виробника 2.

Сформулюємо гіпотези:

H_0 – середні генеральних сукупностей, з яких зроблено вибірки, рівні $\overline{x_1} = \overline{x_2}$;

H_1 – середні не рівні $\overline{x_1} \neq \overline{x_2}$.

Перевіримо справедливість гіпотези H_0 за критерієм Уілкоксона.

Сформуємо об'єднану вибірку, присвоїмо її елементам ранги і знайдемо їх суму. Результати розрахунків оформимо у вигляді таблиці (табл. 2.19). Для зручності підкреслимо елементи першої вибірки.

Оскільки вибірки рівні за об'ємом, то $W = R_1$ або $W = R_2$.

Розрахуємо значення W^* критерія Уілкоксона за формулою (2.10), якщо $W = R_1 = 89$:

$$W^* = \frac{2W - n_1\left(n_1 + n_2 + 1\right) + 1}{\sqrt{\frac{n_1 n_2}{3} \cdot \left(n_1 + n_2 + 1\right)}} = \frac{2 \cdot 89 - 10(10 + 10 + 1) + 1}{\sqrt{\frac{10 \cdot 10}{3} \cdot \left(10 + 10 + 1\right)}} \approx -1{,}476.$$

Таблиця 2.19

Елементи об'єднаної вибірки	Сортована об'єднана вибірка	Ранги елементів об'єднаної вибірки	Ранги елементів першої вибірки	Ранги елементів другої вибірки
280	90	1	1	
230	110	2	2	
112	112	3	3	
176	115	4	4	
90	126	5		5
175	156	6		6
216	170	7		7
110	175	8	8	
205	176	9	9	
115	194	10		10
200	200	11		11
126	205	12	12	
225	210	13		13
210	216	14	14	
260	225	15		15
194	230	16	16	
156	232	17		17
240	240	18		18
170	260	19		19
232	280	20	20	
		Суми	89	121

Оберемо рівень значущості $\alpha = 0{,}05$.

Розрахуємо найменший рівень значущості $p_{гр}$ за формулою (2.12):

якщо $W^* = -1{,}476$, то $\Phi(|W^*|)$ = НОРМРАСП (1,476; 0; 1; ИСТИНА) = 0,93,

тоді $p_{гр} = 2 \cdot \left(1 - \Phi\left(\left|W^*\right|\right)\right) = 2(1-0{,}93) = 0{,}14$.

Отже, оскільки $p_{гр} > \alpha$, то гіпотеза H_0 про рівність середніх приймається на рівні значущості $\alpha = 0{,}05$.

Розрахуємо значення W^* критерія Уілкоксона, якщо $W = R_2 = 121$:

$$W^* = \frac{2W - n_1(n_1 + n_2 + 1) + 1}{\sqrt{\frac{n_1 n_2}{3} \cdot (n_1 + n_2 + 1)}} = \frac{2 \cdot 121 - 10(10 + 10 + 1) + 1}{\sqrt{\frac{10 \cdot 10}{3} \cdot (10 + 10 + 1)}} \approx 1{,}245.$$

Розрахуємо найменший рівень значущості $p_{гр}$:

якщо $W^* = 1{,}245$, то $\Phi(W^*)$ = НОРМРАСП (1,245; 0; 1; ИСТИНА) = 0,89, тоді $p_{гр} = 2 \cdot \left(1 - \Phi\left(\left|W^*\right|\right)\right) = 2(1-0{,}89) = 0{,}22$.

Отже, оскільки $p_{гр} > \alpha$, то гіпотеза H_0 про рівність середніх приймається на рівні значущості $\alpha = 0{,}05$.

Висновок: надійність роботи станків двох виробників однакова.

2.4. Перевірка статистичних гіпотез із використанням Microsoft Excel

2.4.1. Двохвибірковий *F*-тест для дисперсій

Перевірку гіпотези про рівність генеральних дисперсій за *F*-критерієм (Фішера) можна виконати за допомогою пакета аналізу Microsoft Excel. Для використання пакета необхідно:

1) Вибрати в меню послідовно пункти ***Сервис – Анализ данных***, після чого з'явиться вікно для вибору інструмента аналізу (рис. 2.2).

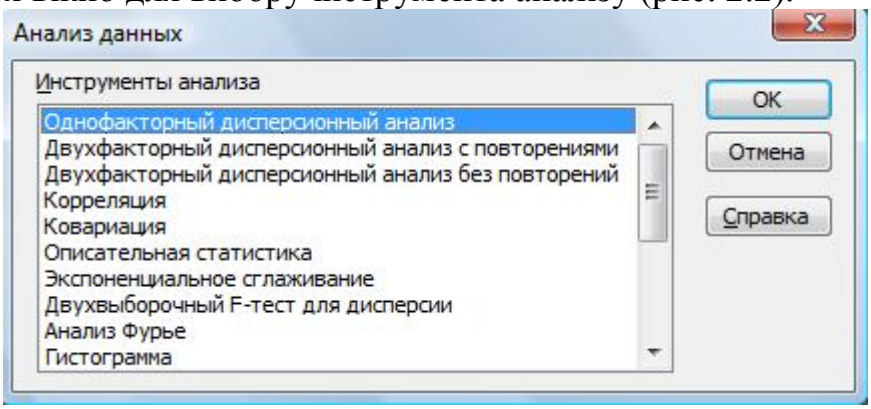


Рисунок 2.2. Діалогове вікно аналізу даних

2) Вибрати у діалоговому вікні інструмент ***Двухвыборочный F-тест для дисперсии***, після чого з'явиться вікно для вибору параметрів (рис. 2.3).

3) Задати всі необхідні параметри і натиснути ОК.

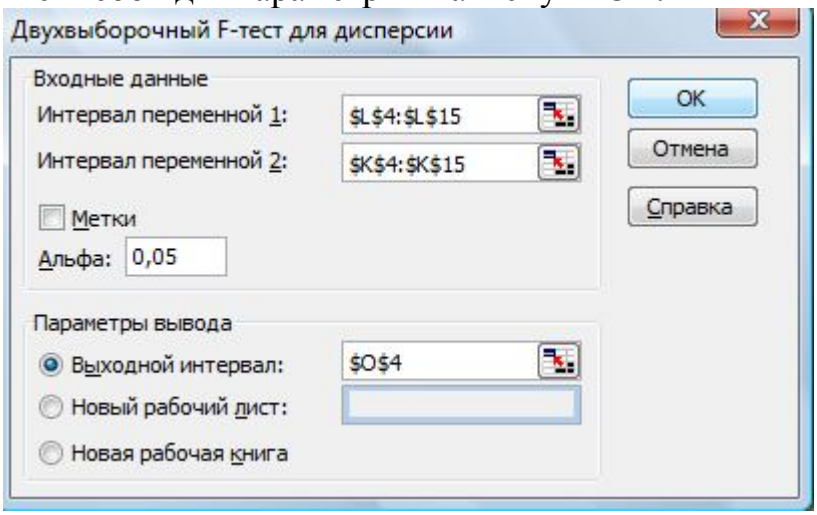
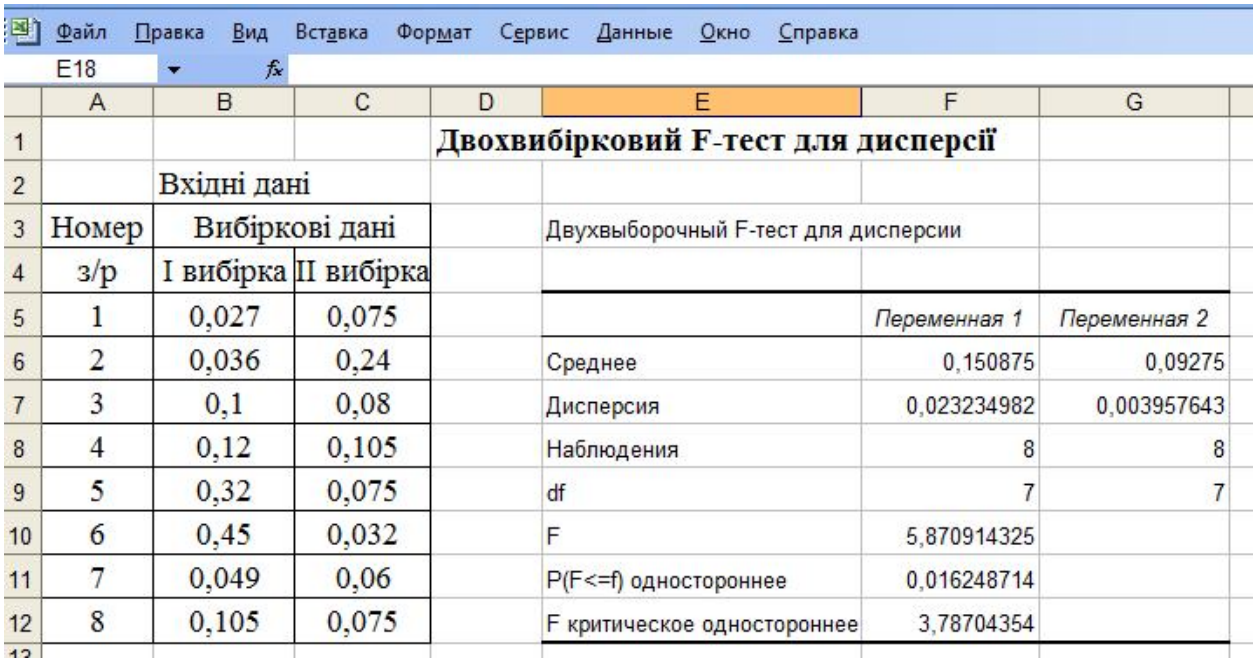


Рисунок 2.3. Вікно надання параметрів

Приклад і результат роботи тесту продемонстровано на рис. 2.4.



Файл Правка Вид Вставка Формат Сервис Данные Окно Справка

E18

	A	B	C	D	E	F	G
1				**Двохвибірковий F-тест для дисперсії**			
2		Вхідні дані					
3	Номер	Вибіркові дані			Двухвыборочный F-тест для дисперсии		
4	з/р	I вибірка	II вибірка				
5	1	0,027	0,075			*Переменная 1*	*Переменная 2*
6	2	0,036	0,24		Среднее	0,150875	0,09275
7	3	0,1	0,08		Дисперсия	0,023234982	0,003957643
8	4	0,12	0,105		Наблюдения	8	8
9	5	0,32	0,075		df	7	7
10	6	0,45	0,032		F	5,870914325	
11	7	0,049	0,06		P(F<=f) одностороннее	0,016248714	
12	8	0,105	0,075		F критическое одностороннее	3,78704354	

Рисунок 2.4. Перевірка рівності генеральних дисперсій

2.4.2. Двохвибірковий *t*-тест для середніх

Перевірку гіпотез про рівність генеральних середніх за критерієм Стьюдента можна виконати за допомогою пакета аналізу даних Microsoft Excel. Для здійснення перевірки необхідно у вікні для вибору інструмента аналізу (рис. 2.2) виконати:

1) У випадку рівних генеральних дисперсій – інструмент ***Двухвыборочный t-тест с одинаковыми дисперсиями***;

2) У випадку різних генеральних дисперсій – інструмент ***Двухвыборочный t-тест с разными дисперсиями***;

3) У випадку залежних вибірок – ***Парный двухвыборочный t-тест для средних***.

Після вибору інструмента аналізу з’явиться вікно для вибору параметрів аналізу, аналогічне зображеному на рис. 2.3, в якому необхідно задати масиви чарунок із вхідними вибірковими даними і рівень значущості (стандартний рівень – 0,05). Приклад і результат роботи тесту наведено на рисунку 2.5.

Файл Правка Вид Вставка Формат Сервис Данные Окно Справка

C22

	A	B	C	D	E	F	G
1				**Двохвибірковий t-тест для середніх**			
2		Вхідні дані					
3	Номер	Вибіркові дані			Двухвыборочный t-тест с различными дисперсиями		
4	з/р	I вибірка	II вибірка				
5	1	0,027	0,075			*Переменная 1*	*Переменная 2*
6	2	0,036	0,24		Среднее	0,150875	0,09275
7	3	0,1	0,08		Дисперсия	0,023234982	0,003957643
8	4	0,12	0,105		Наблюдения	8	8
9	5	0,32	0,075		Гипотетическая разность средних	0	
10	6	0,45	0,032		df	9	
11	7	0,049	0,06		t-статистика	0,996970694	
12	8	0,105	0,075		P(T<=t) одностороннее	0,172413357	
13					t критическое одностороннее	1,833112923	
14					P(T<=t) двухстороннее	0,344826713	
15					t критическое двухстороннее	2,262157158	
16							

Рисунок 2.5. Перевірка рівності генеральних середніх

2.5. Перевірка статистичних гіпотез із використанням SPSS

За допомогою пакета програм SPSS можна виконувати перевірку статистичних гіпотез різноманітними методами. Якщо при порівнянні генеральних середніх та дисперсій передбачається, що вибірки підпорядковуються нормальному закону, то використовуються методи, що базуються на критеріях: *t*-критерії, критерій Стьюдента, Лівіня, Шефе, *F*-тест, тести Тьюкі, Дункана, Габріеля, Темхена і т.д…

Непараметричні критерії перевірки статистичних гіпотез використовуються тоді, коли вибірка не підпорядковується нормальному закону. Серед таких критеріїв: тести хі-квадрат, критерій Колмогорова-Смірнова, *U*-тест Мана-Уітні, тест Мозеса, Уалда-Вольфовіца, Уілкоксона, Фрідмана, *W* Кендала, знаковий тест та ін.

Ці та багато інших методів доступні для використання за допомогою описаної програми. Розглянемо деякі з них.

2.5.1. Перевірка гіпотези про вид закону розподілу досліджуваної величини

Щоб визначити вид закону розподілу, його ототожнюють із одним уже відомим (нормальним, рівномірним, Пуассона, експоненціальним) і висувають гіпотези:

а) нульова гіпотеза H_0 − закони співпадають,

б) конкуруюча H_1 − закони не співпадають.

При рівні значущості (Асимпт. знч.) $p > 0{,}05$ нульова гіпотеза приймається, при p <= 0,05 − приймається конкуруюча.

Для перевірки гіпотези про вид закону розподілу необхідно:

1) Вибрати в меню послідовно ***Анализ − Непараметрические критерии − Одновыборочный Колмогорова-Смирнова***, після чого з'явиться діалогове вікно ***Одновыборочный критерий Колмогорова-Смирнова*** (рис. 2.6);

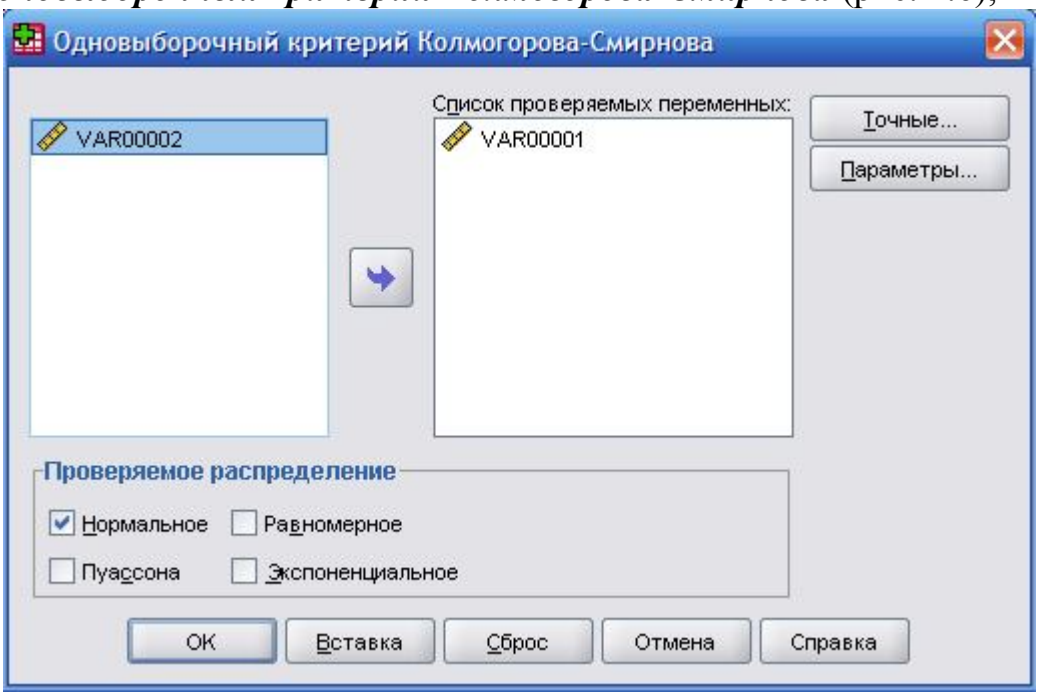


Рисунок 2.6. Критерій перевірки виду ряду розподілу

2) Перенести змінну, яку необхідно перевірити у поле ***Список проверяемых переменных***, у переліку ***Проверяемое распределение*** вибрати вид розподілу, із яким будемо порівнювати задану змінну;

3) Якщо у вікні перегляду результатів (рис. 2.7) рівень значущості (Асимпт. знч. (двухсторонняя) − остання стрічка таблиці, зображеної на рис. 2.7) $p > 0{,}05$, то нульова гіпотеза про вид розподілу приймається.

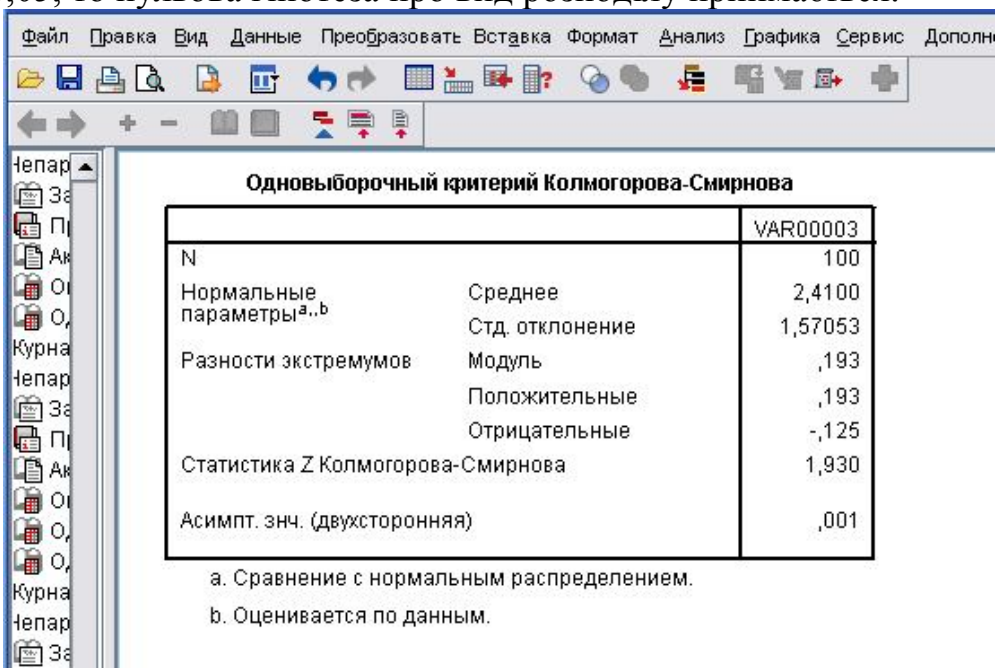


Рисунок 2.7. Приклад, коли ряд розподілу не співпадає з нормальним ($p = 0{,}01$)

2.5.2. Перевірка гіпотези про рівність генеральних дисперсій для незалежних вибірок: *T*- критерій.

Для перевірки рівності генеральних дисперсій нормально розподілених двох незалежних вибірок в SPSS, використовують критерій Лівіня. Висуваються гіпотези: нульова гіпотеза H_0 − дисперсії рівні, конкуруюча H_1 − дисперсії не рівні. При рівні значущості (Знч.) $p > 0{,}05$ нульова гіпотеза приймається, при p <= 0,05 − приймається конкуруюча.

Розглянемо можливості використання пакету програм SPSS для перевірки достовірності прийняття нульової гіпотези згідно даних із вище наведеного прикладу 2.4 (табл. 2.20).

Таблиця 2.20

	Група 1					Група 2				
Продуктивність праці	34	85	96	102	103	63	69	83	89	106
Кількість працівників	5	2	11	8	4	2	6	8	3	1

Для перевірка гіпотези про рівність генеральних дисперсій необхідно:

1) Для зручності дослідження ввести дані дослідження у два стовпчики вкладки *Набор данных* таким чином: у перший стовпчик − продуктивність праці працівників (ПродПрац), враховуючи кількість, у другий − позначення групи (НомГрупи), до якої належить працівник (рис. 2.8);

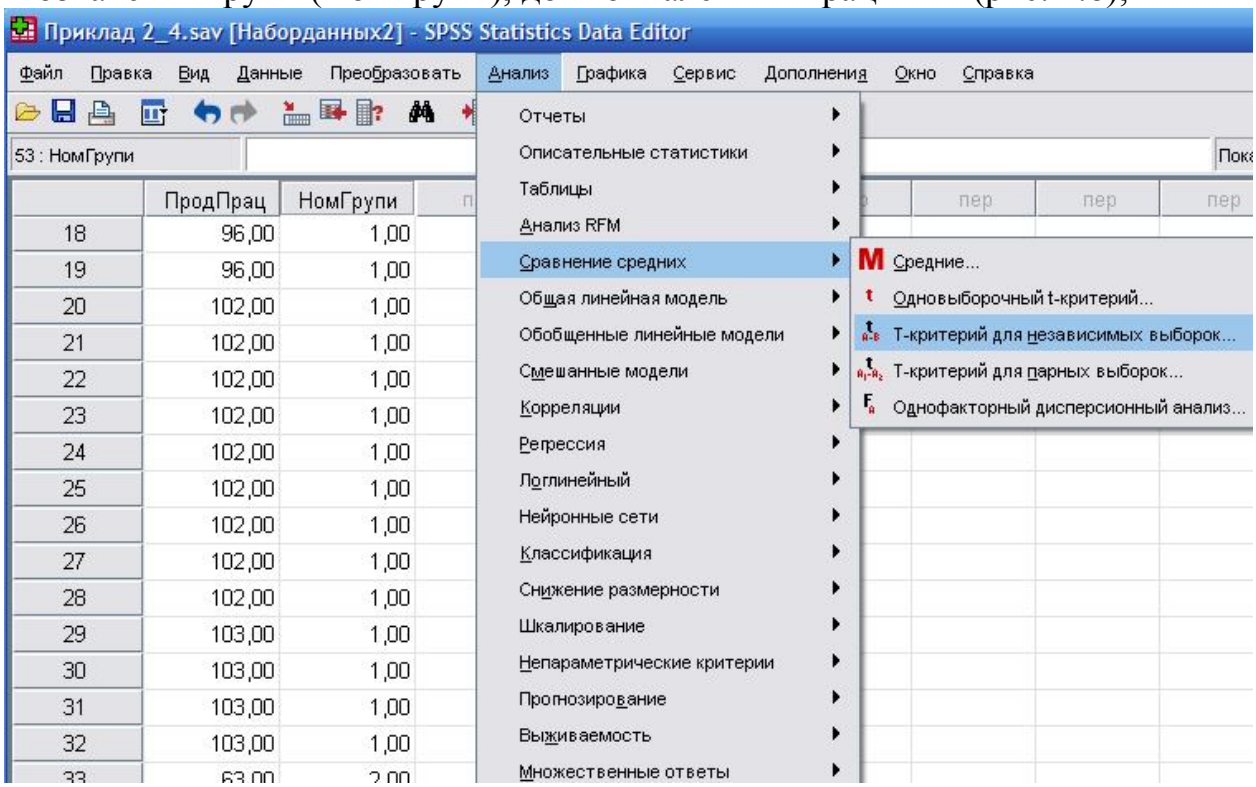


Рисунок 2.8. Вибір критерія перевірки рівності дисперсій незалежних вибірок

2) Вибрати в меню послідовно ***Анализ − Сравнение средних − Т-критерий для независимых выборок***, після чого з'явиться діалогове вікно ***Т-критерий для независимых выборок***, у якому: змінну *ПродПрац* перенести у поле ***Проверять переменные***, а змінну *НомГрупи* − у поле ***Группировать по***.

Активувати діалогове вікно ***Задать группы*** (рис. 2.9), у якому ввести значення 1 та 2 відповідно у поле Групи 1 та Групи 2;

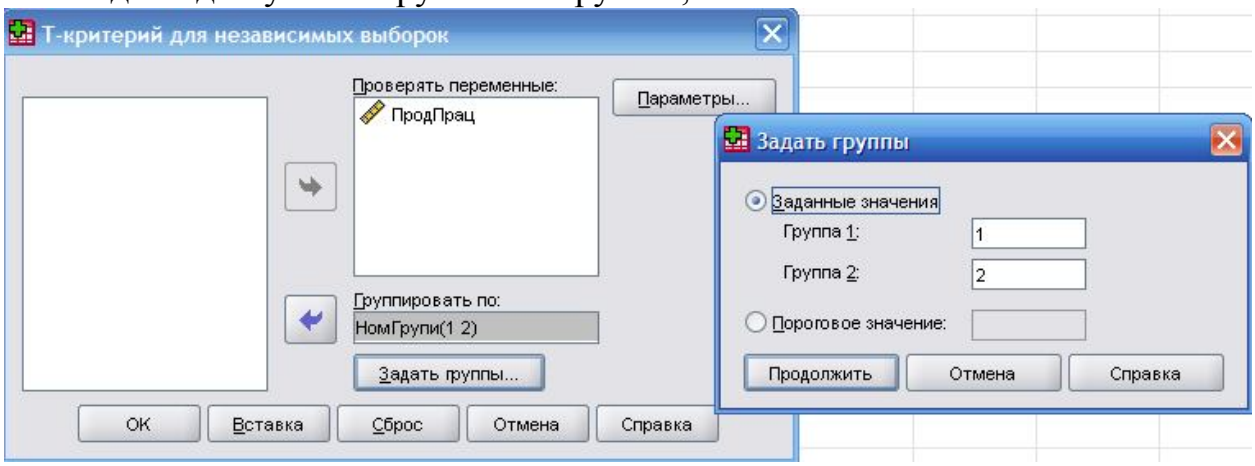


Рисунок 2.9. Етап завдання даних для перевірки критерія про рівність дисперсій

3) Викликати вікно результатів (рис. 2.10) та зробити відповідні висновки згідно із отриманим значенням рівня значущості (Знч.). Для даного прикладу згідно критерію Лівіня $p = 0{,}036 < 0{,}05$, отже нульову гіпотезу про рівність дисперсій приймати не слід.

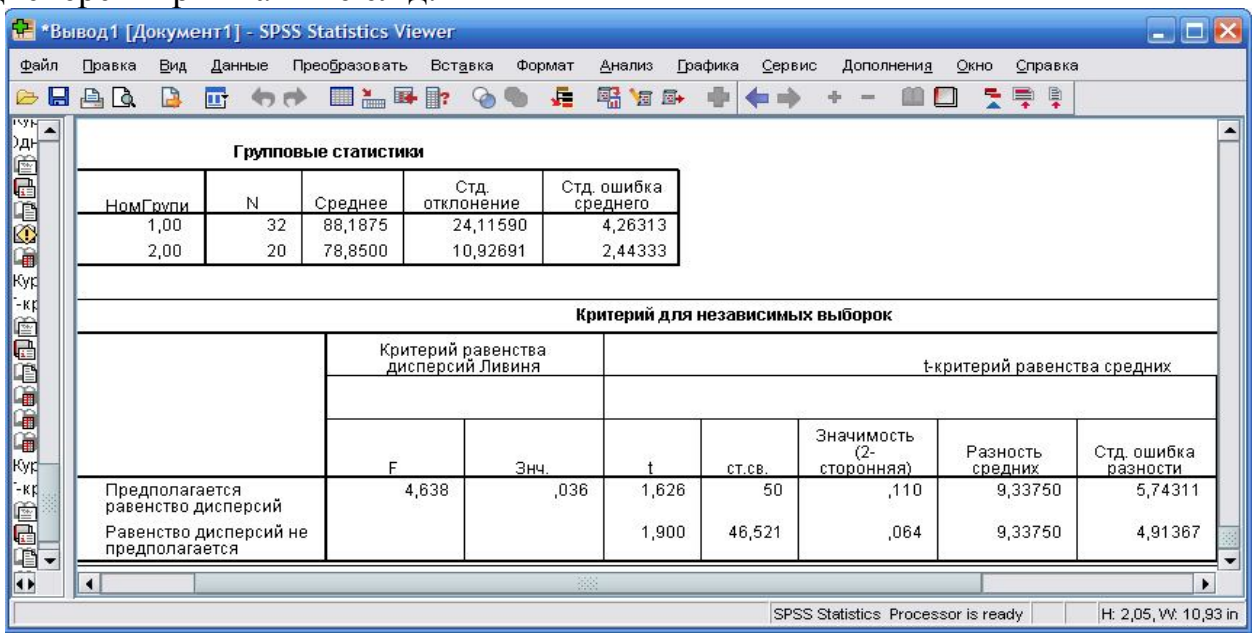


Рисунок 2.10. Вікно виведення результатів критерію для незалежних вибірок

2.5.3. Перевірка гіпотези про рівність генеральних середніх двох незалежних вибірок за допомогою *t*-критерію. Критерій Стьюдента

Для перевірки рівності генеральних середніх нормально розподілених двох незалежних вибірок в SPSS висуваються гіпотези: нульова гіпотеза H_0 − середні рівні, конкуруюча H_1 − генеральні середні нерівні. При рівні значущості (Значимость (2-сторонняя)) $p > 0{,}05$ приймається нульова гіпотеза, при p <= 0,05 − приймається конкуруюча.

Перевіримо рівність генеральних середніх для уже згадуваного набору даних із прикладу 2.4 (табл. 2.20).

Для перевірки рівності генеральних середніх методами SPSS необхідно:

1) Виконати дії пп. 1)-3) із пункту **2.5.2** та розглянути інші дані, представлені у таблиці вікна результатів (рис. 2.10). Зробити висновки, враховуючи, що гіпотезу про рівність дисперсій відкинуто. Тому до уваги приймаємо останню стрічку таблиці: значення $t = 1{,}9$, степені свободи $df = 46{,}521$, рівень значущості $p = 0{,}064 > 0{,}05$.

Отже нульову гіпотезу про рівність генеральних середніх (середню продуктивність праці у двох групах) варто прийняти.

2.5.4. Перевірка гіпотези про рівність генеральних середніх. Критерій Уілкоксона

Для перевірки рівності генеральних середніх двох залежних вибірок критерієм Уілкоксона в SPSS висуваються гіпотези: нульова гіпотеза H_0 − генеральні середні рівні, конкуруюча H_1 − генеральні середні нерівні. Якщо рівень значущості (Асимпт. знч. (двухсторонняя)) $p > 0{,}05$, то приймається нульова гіпотеза, якщо ж $p <= 0{,}05$ − приймається конкуруюча.

Розглянемо, як використовується метод Уілкоксона у пакеті програм SPSS для даних, описаних у прикладі 2.9.

Для перевірки рівності генеральних середніх необхідно:

1) Ввести стрічкові дані табл. 2.18 у два стовпчики вкладки *Набор данных*. Вибрати в меню послідовно ***Анализ − Непараметрические критерии − Для двух связанных выборок***. У діалоговому вікні ***Критерии для связанных выборок*** перенести дві змінні, що відображають надійність праці станків різних виробників у поле ***Тестовые пары***. Вибрати серед заданих критерій Уілкоксона (рис. 2.11) та перейти у вікно виведення результатів;

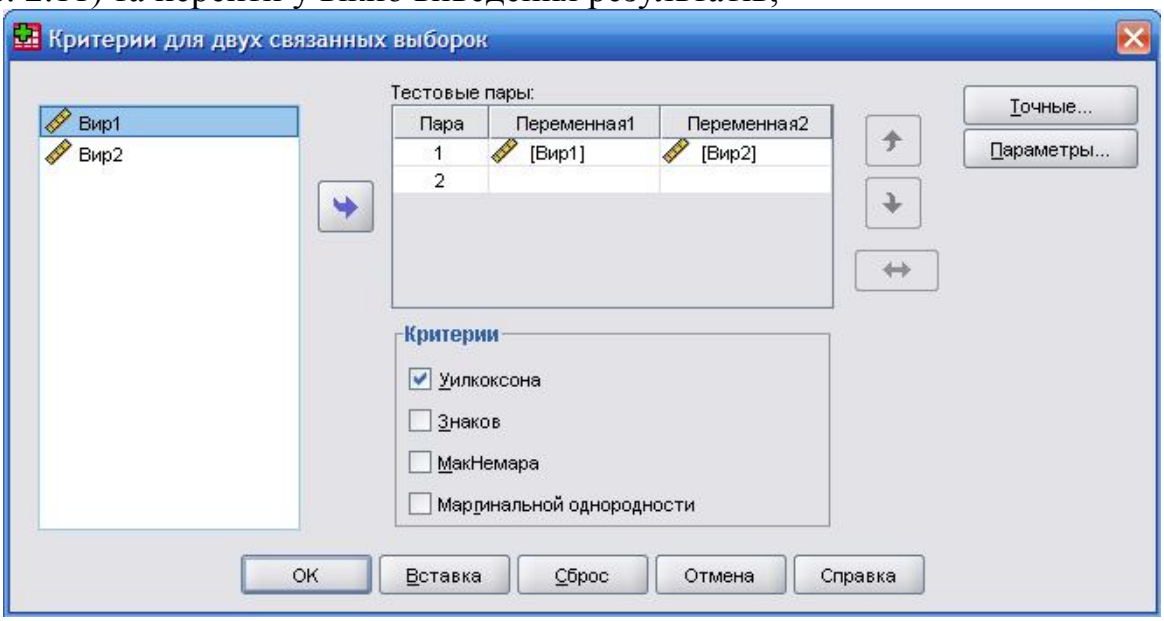


Рисунок 2.11. Вибір критерія перевірки гіпотези про рівність середніх

2) Враховуючи, що рівень значущості $p = 0{,}333 > 0{,}05$ (рис. 2.12), можна зробити висновок про однакову надійність роботи станків обох виробників.

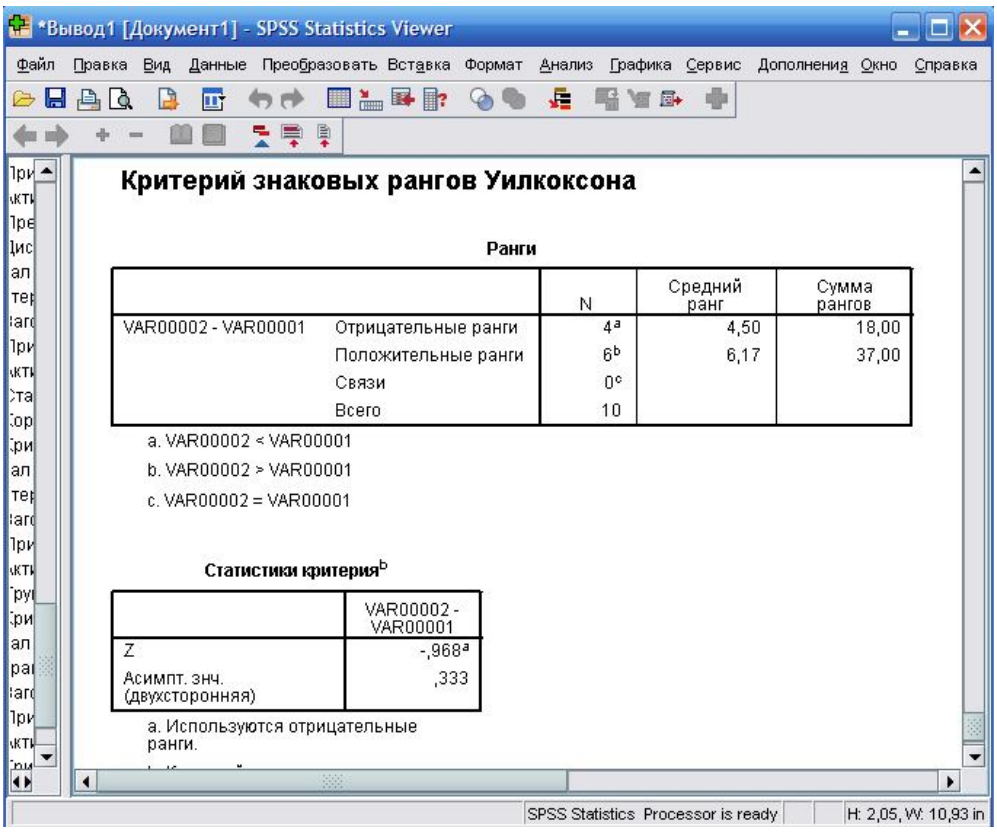


Рисунок 2.12. Результати статистики критерія Уілкоксона

Завдання для самостійного виконання

2.1. При дослідженні якості продукції, яку випускають дві фірми, було перевірено 50 одиниць продукції першої фірми і 63 одиниці другої. Середня оцінка продукції першої фірми склала 36 балів, другої – 32 бали. Дисперсія оцінок виявилася рівною 10. Визначити фірму, що виробляє більш якісну продукцію, на рівні значущості 0,01.

2.2. Середній річний об'єм виробництва 10 однотипних компаній в регіоні *A* склав 5900 одиниць продукції; 8 компаній того ж типу в регіоні *B* – 5690 одиниць. Вибіркова дисперсія об'єму виробництва в регіоні *A* склала 1000, в регіоні *B* – 1500. Перевірити гіпотезу про рівність середніх значень на рівні значущості 0,05.

2.3. Середня урожайність пшениці у фермерському господарстві складала 75 ц/га. Після внесення нового добрива середня урожайність підвищилася на 5 ц/га. Вибіркова дисперсія склала 2,5 ц/га. Перевірити на рівні значущості 0,01 гіпотезу про доцільність нового добрива.

2.4. Середня продуктивність праці на підприємстві складала 30 одиниць продукції за зміну з дисперсією 4. Після внесення змін в організацію праці середня продуктивність склала 34 з дисперсією 7. Перевірити на рівні значущості 0,01 гіпотезу про доцільність внесення змін в організацію праці.

2.5 – 2.14. Для виробництва кожної з $n_1 = 53$ деталей за першою технологією було витрачено, у середньому $\overline{x_1}$ секунд часу з дисперсією S_1^2. Для виробництва кожної з $n_2 = 43$ деталей за другою технологією було витрачено, у середньому $\overline{x_2}$ секунд часу з дисперсією S_2^2 (табл. 2.21). Чи можна зробити висновок, що для виробництва деталей за першою технологією потрібно, у середньому, більше часу, ніж за другою. Гіпотезу перевірити на рівні значущості α.

Таблиця 2.21

№	$\overline{x_1}$	S_1^2	$\overline{x_2}$	S_2^2	α
2.5	38	4	31	2	0,05
2.6	39	5	32	3	0,01
2.7	33	7	31	8	0,05
2.8	37	8	34	7	0,01
2.9	35	4	32	5	0,05
2.10	37	5	36	4	0,01
2.11	37	7	35	7	0,05
2.12	38	8	33	8	0,01
2.13	42	3	40	5	0,05
2.14	40	2	34	4	0,01

2.16. Знайти закон розподілу величини X – кількості відвідувачів ресторану швидкого харчування за даними табл. 2.22.

Таблиця 2.22

Час роботи	9.00 – 11.00	11.00 – 13.00	13.00 – 15.00	15.00 – 17.00	17.00 – 19.00	19.00 – 21.00
Кількість відвідувачів	25	53	68	73	52	39

Питання для самоконтролю

1. Що називається статистичною гіпотезою?
2. Чим відрізняється рівень значущості від рівня довіри?
3. За якими етапами здійснюється перевірка статистичних гіпотез?
4. Що означає прийняття гіпотези?
5. Що називається теоретичним законом розподілу випадкової величини? Емпіричним законом розподілу?
6. Що називається критерієм згоди?
7. Як перевірити гіпотезу про вид закону розподілу випадкової величини?
8. Яка мета перевірки гіпотез про рівність середніх значень та дисперсій?
9. Які критерії перевірки гіпотез про рівність генеральних середніх?
10. Які умови застосування критеріїв про рівність генеральних середніх?

11. Які критерії перевірки гіпотез про рівність генеральних дисперсій ви знаєте?
12. Які умови застосування критеріїв про рівність генеральних середніх?
13. Що називається рангом? Яка процедура ранжування?
14. Які засоби MS Excel призначені для перевірки статистичних гіпотез?
15. Як перевіряються статистичні гіпотези засобами SPSS?

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**.

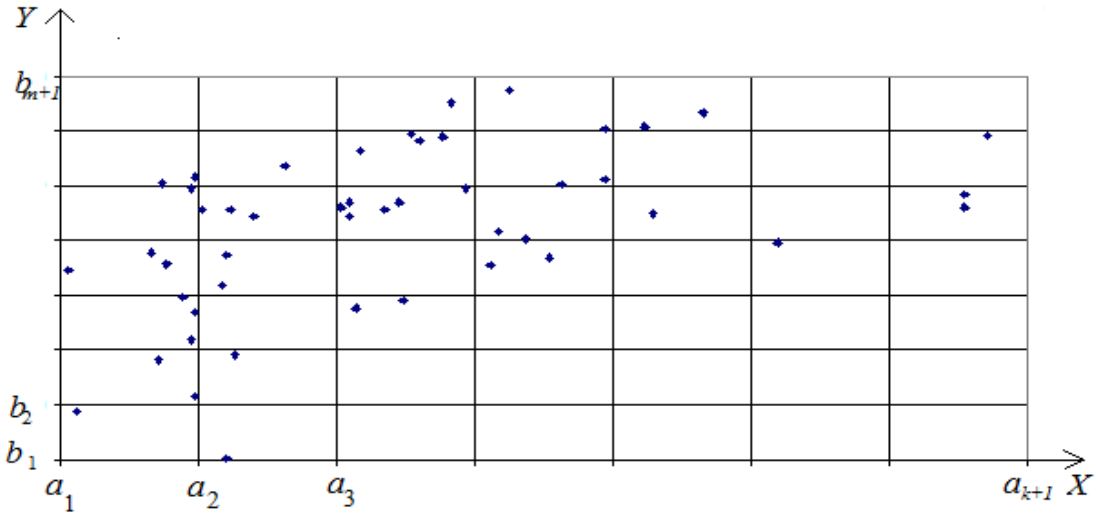


Рисунок 3.1. Поле кореляції

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***.

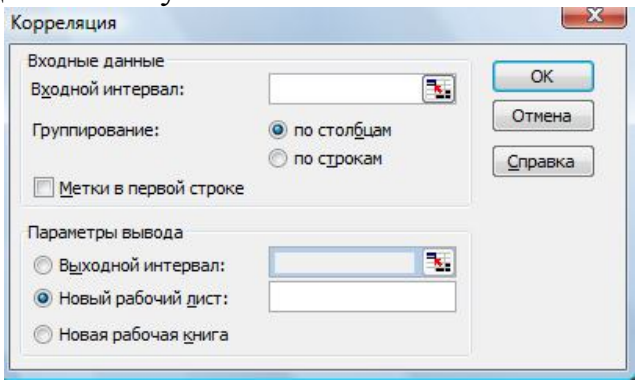


Рисунок 3.2. Вікно надання параметрів кореляційного аналізу

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);

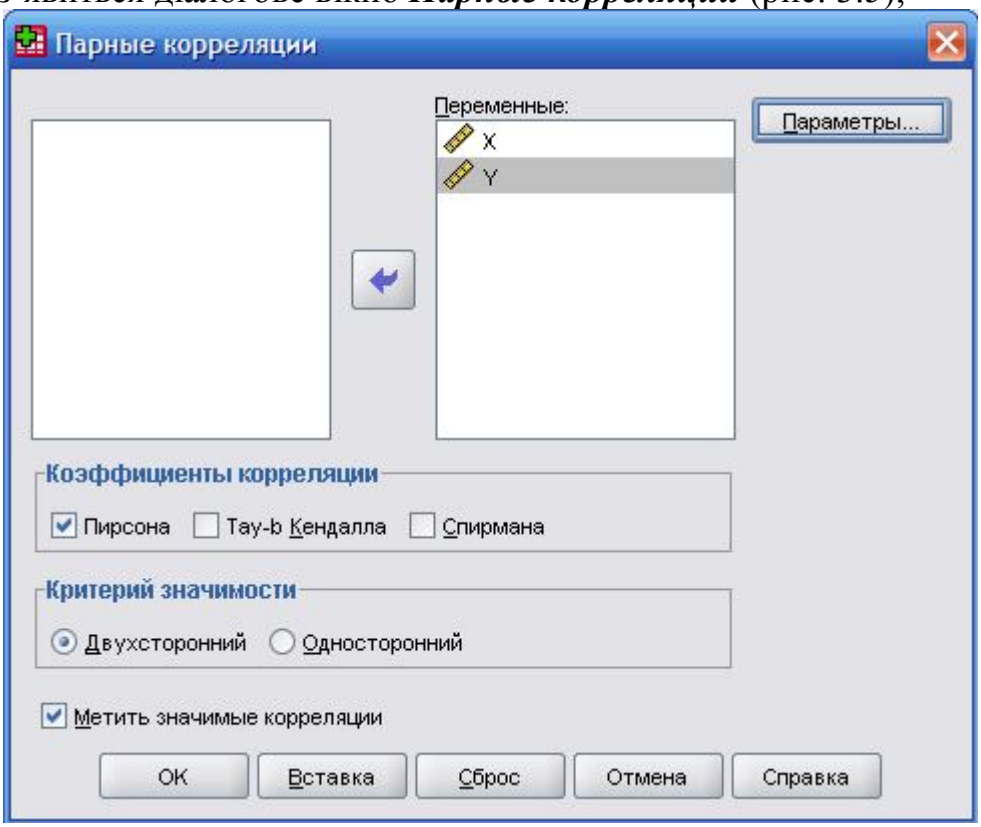


Рисунок 3.5. Діалогове вікно вибору коефіцієнта кореляції

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

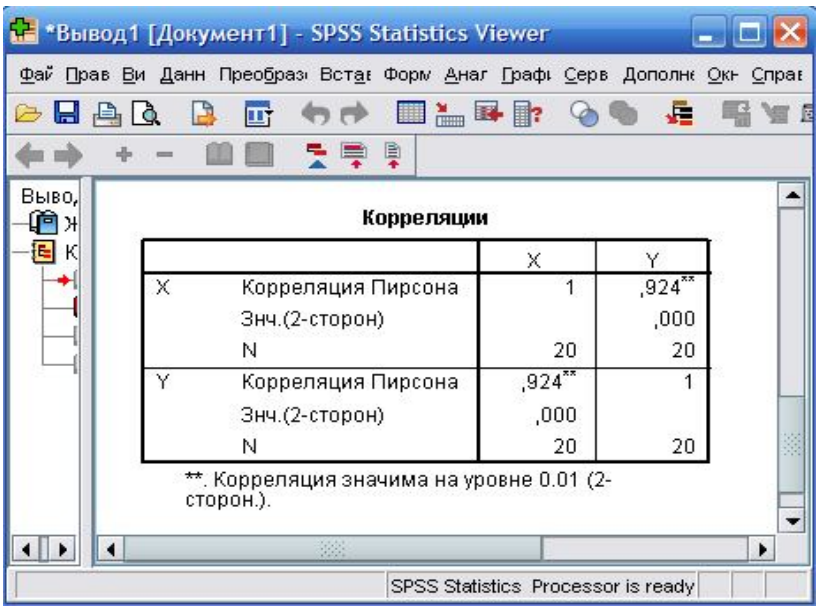


Рисунок 3.6. Результат тісного зв’язку між змінними *X* і *Y*

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).

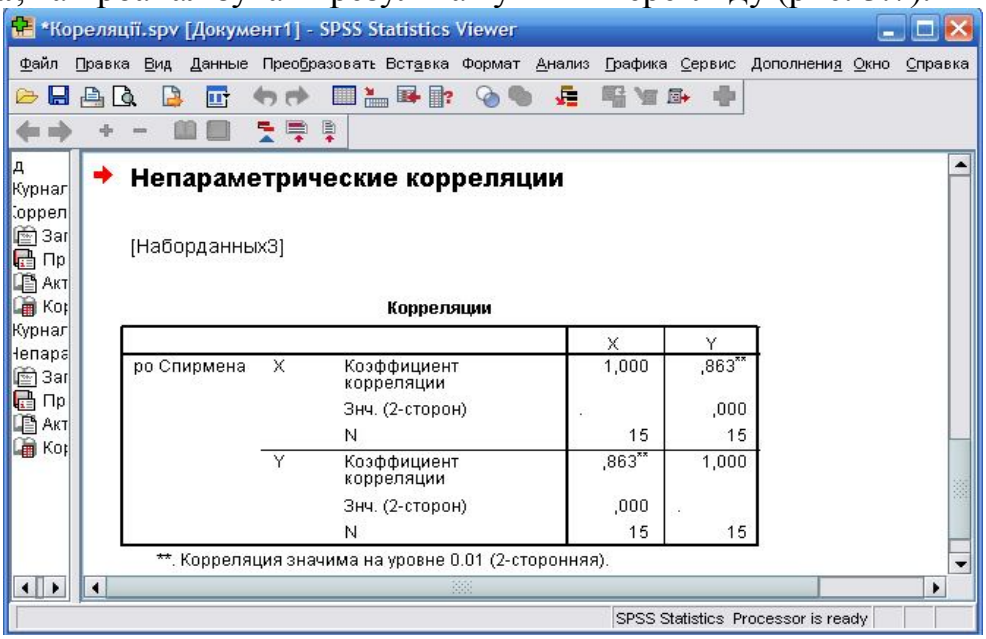


Рисунок 3.7. Міра зв’язку між змінними *X* і *Y* за коефіцієнтом кореляції Спірмена

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***;

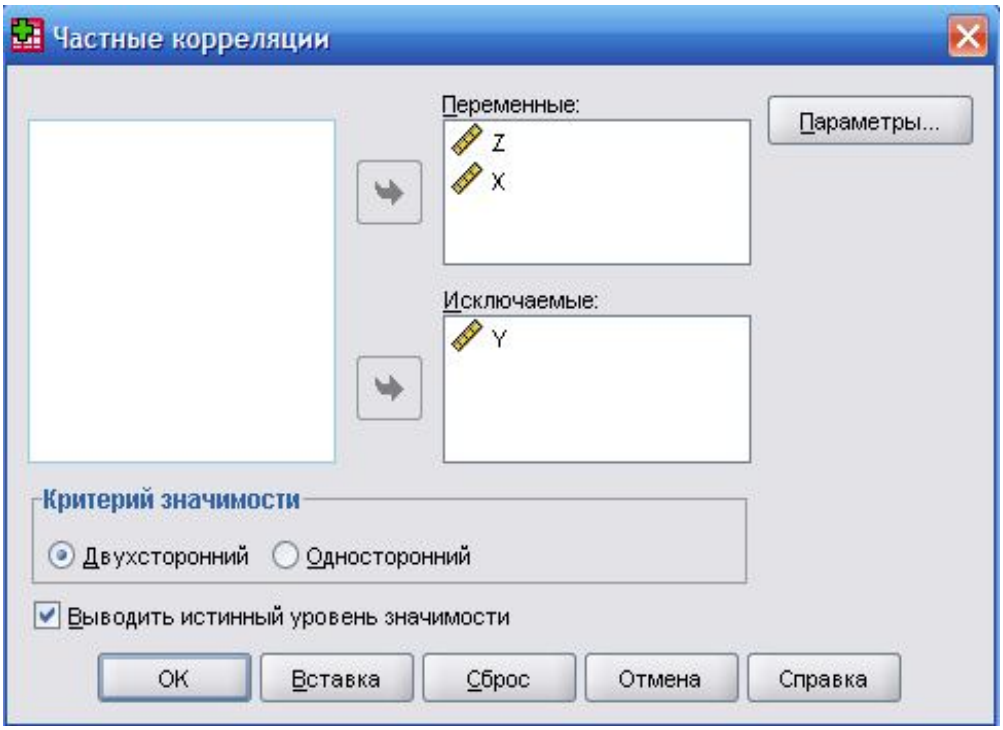


Рисунок 3.8. Діалогове вікно вибору змінних для частинної кореляції

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.

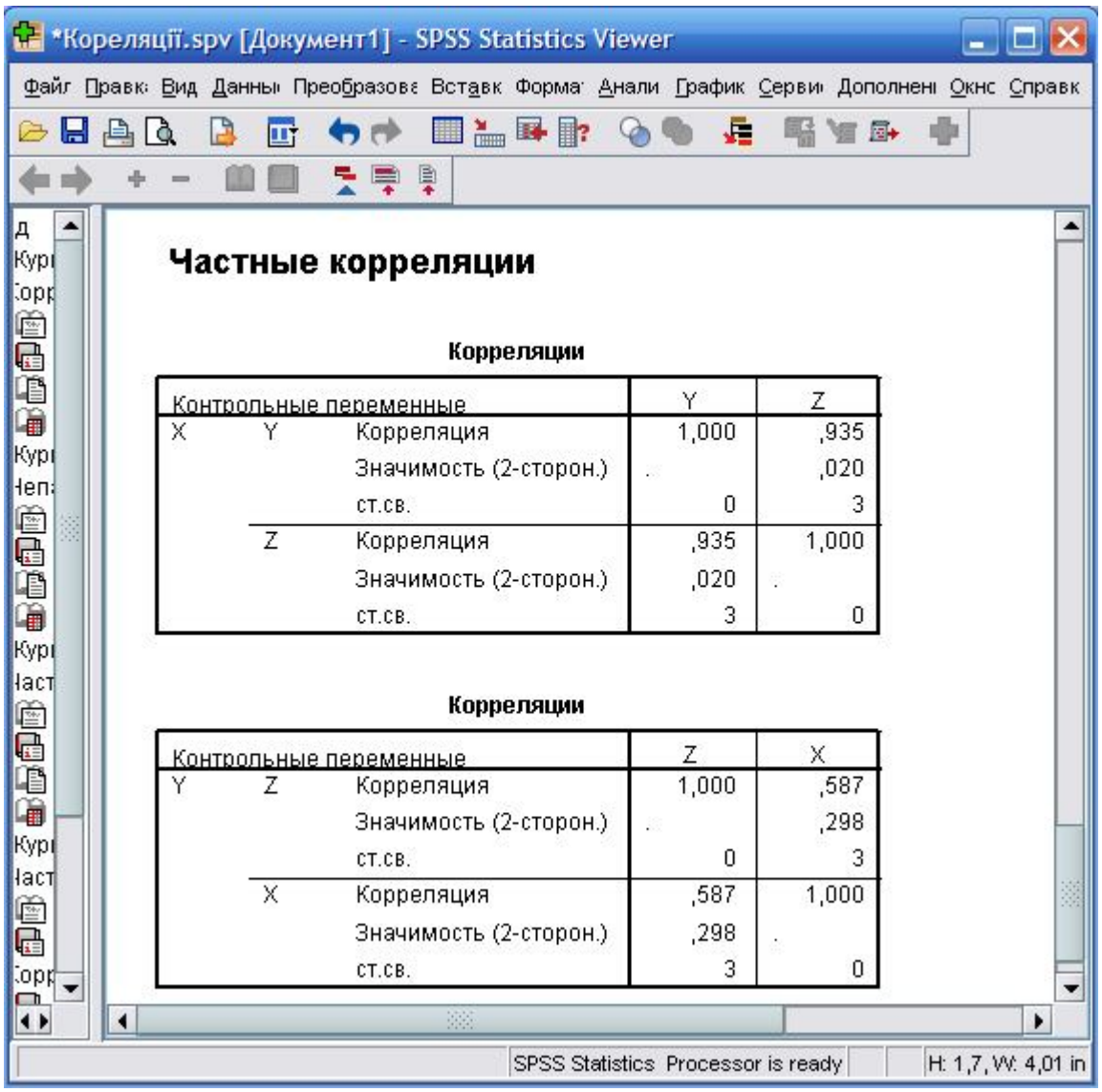


Рисунок 3.9. Частинні коефіцієнти кореляції

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).

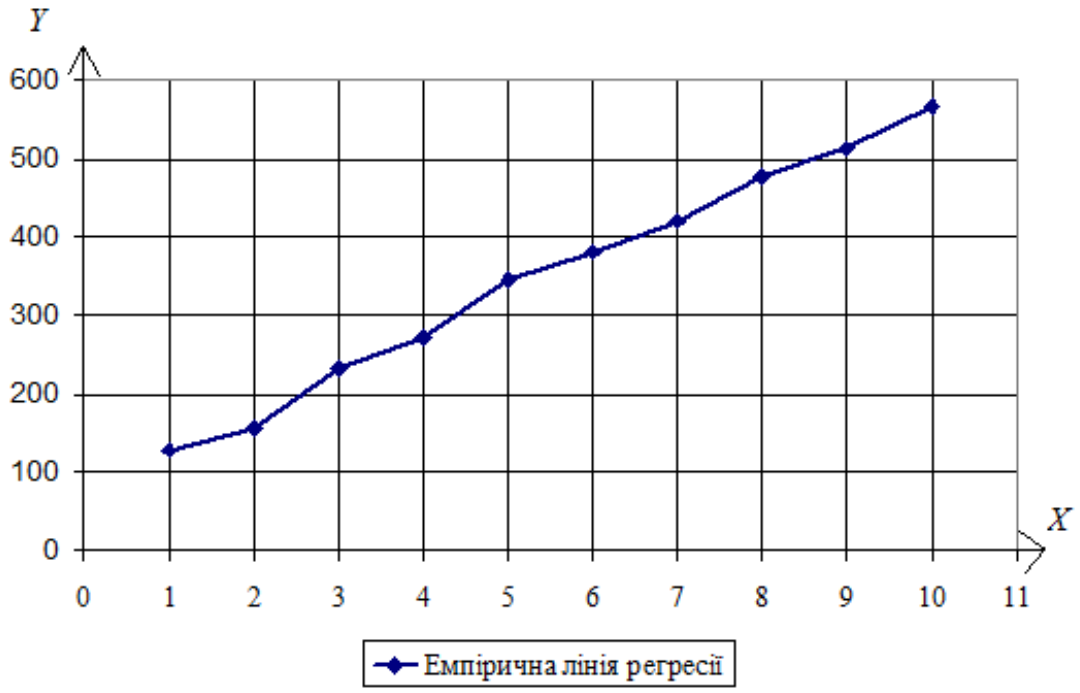


Рисунок 4.1. Гіпотетична лінійна залежність

В іншому випадку висувається гіпотеза про наявність нелінійного зв'язку (рис. 4.2).

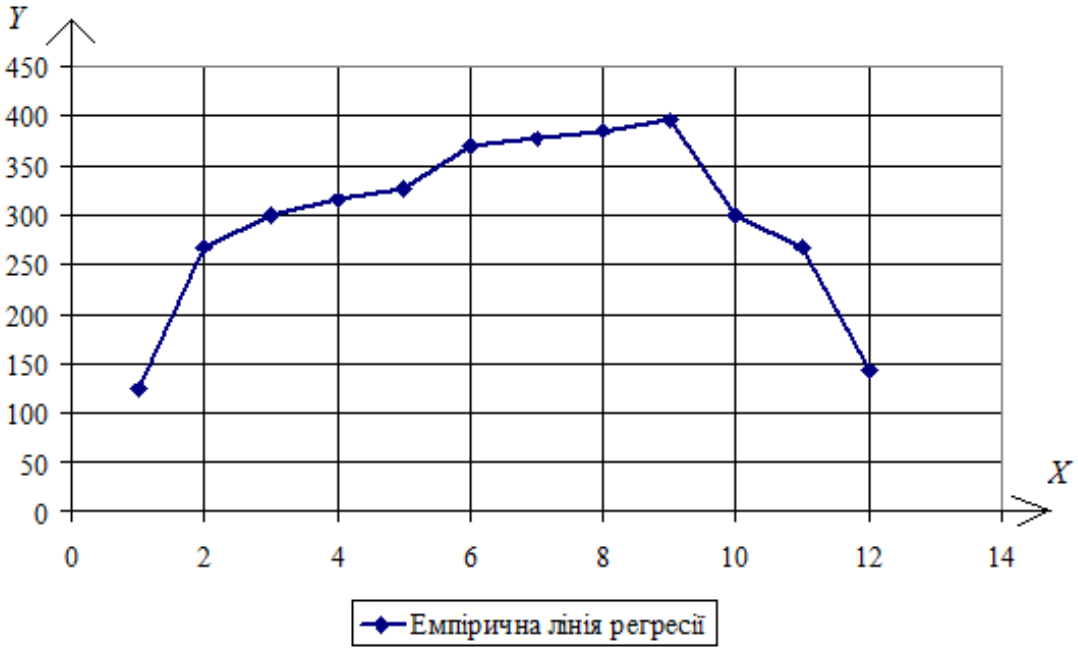


Рисунок 4.2. Гіпотетична нелінійна залежність

4.2. Лінійна регресія

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

$$\overline{y_x} = ax + b\,, \qquad (4.1)$$

де a, b − параметри моделі.

Побудова лінійної регресійної моделі – це знаходження параметрів рівняння (4.1). Параметри рівняння регресії можна знайти за **методом найменших квадратів**.

Ідея методу найменших квадратів

Нехай при вивченні залежності *Y* від *X* було отримано вибіркові дані: $x_1, x_2, ..., x_n$ – значення величини *X*, $y_1, y_2, ..., y_n$ – відповідні значення *Y*. За вибірковими даними було побудовано рівняння регресії $y = ax + b$. Якщо в рівняння підставити замість *x* значення $x_1, x_2, ..., x_n$, то будуть отримані теоретичні значення *Y*: $y_{1,теор}, y_{2,теор}, ..., y_{n,теор}$, які відрізняються від $y_1, y_2, ..., y_n$. Різниця значень $y_{i,теор} - y_i$ називається помилкою регресійної моделі і позначається e_i. Якщо параметри рівняння підбираються так, щоб сума квадратів помилок була мінімальною, то говорять, що вони отримані за методом найменших квадратів.

У випадку лінійної регресії параметри рівняння регресії за методом найменших квадратів знаходяться з системи лінійних алгебраїчних рівнянь:

$$\begin{cases} a\sum\limits_{i=1}^{k} x_i^2 n_i + b\sum\limits_{i=1}^{k} x_i n_i = \sum\limits_{i=1}^{k} x_i n_i \overline{y_{x_i}} \\ a\sum\limits_{i=1}^{k} x_i n_i + b\sum\limits_{i=1}^{k} n_i = \sum\limits_{i=1}^{k} n_i \overline{y_{x_i}} \end{cases}. \qquad (4.2)$$

Якщо вибіркові дані не згруповані, то система (4.1) значно спрощується:

$$\begin{cases} a\sum\limits_{i=1}^{n} x_i^2 + b\sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} x_i y_i \\ a\sum\limits_{i=1}^{n} x_i + bn = \sum\limits_{i=1}^{n} y_i \end{cases}. \qquad (4.3)$$

Перевірка правильності побудови рівняння регресії здійснюється за основним варіаційним рівнянням:

$$Q = Q_p + Q_o, \qquad (4.4)$$

де $Q = \sum\limits_{i=1}^{k} \left(\overline{y_{x_i}} - \overline{y} \right)^2 n_i$ − загальна варіація, тобто сума квадратів відхилень емпіричних значень *Y* від середнього $\overline{y} = \dfrac{\sum\limits_{i=1}^{k} \overline{y_{x_i}} n_i}{n}$;

$Q_p = \sum\limits_{i=1}^{k} \left(y_{i,теор} - \overline{y} \right)^2 n_i$ − варіація регресії, тобто сума квадратів відхилень теоретичних значень *Y* від середнього, що обумовлена регресією;

$Q_o = \sum_{i=1}^{k} \left(y_{i,теор} - \overline{y}_{x_i} \right)^2 n_i$ – варіація залишків, тобто сума квадратів відхилень теоретичних значень Y від емпіричних.

У випадку незгрупованих даних загальна варіація, варіації регресії і залишків знаходяться за формулами: $Q = \sum_{i=1}^{n} \left(y_i - \overline{y} \right)^2$; $Q_p = \sum_{i=1}^{n} \left(y_{i,\text{теор}} - \overline{y} \right)^2$; $Q_o = \sum_{i=1}^{n} \left(y_{i,\text{теор}} - y_i \right)^2$; а середнє значення за формулою $\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$.

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

$$F = \frac{Q_p (n-l)}{Q_o (l-1)}, \tag{4.5}$$

де n – кількість спостережень, l – кількість груп у кореляційній таблиці або кількість параметрів моделі у випадку незгрупованих даних. Розраховане значення F-статистики порівнюється з критичним значенням $F_{кр}$ розподілу Фішера, яке можна знайти за статистичними таблицями або за допомогою вбудованої функції Excel $FРАСПОБР(\alpha, k_1, k_2)$, де $k_1 = l - 1$; $k_2 = n - l$ – степені свободи, α – рівень значущості.

Адекватність моделі вибірковим даним можна оцінити за коефіцієнтом детермінації R^2, що показує частину варіації значень результативної ознаки Y, що пояснюється рівнянням регресії. Коефіцієнт детермінації розраховується за формулою:

$$R^2 = 1 - \frac{Q_o}{Q} = \frac{Q_p}{Q}. \tag{4.6}$$

Значення коефіцієнта детермінації знаходяться в інтервалі $[0;1]$, тобто $0 \le R^2 \le 1$. Чим ближче R^2 до 1, тим краще отримане рівняння регресії пояснює поведінку результативної ознаки. Наприклад, якщо $R^2 = 0{,}98$, то 98% варіації результативної ознаки Y пояснюється рівнянням регресії.

Приклад 4.1. Побудувати регресійну модель, що описує залежність сумарних виробничих затрат Y (тис. грн.) від об'ємів виробництва X (тис. од.). Відповідні статистичні дані задано у табл. 4.2.

Таблиця 4.2

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

Розв'язок. В табл. 4.2 задано вибіркові дані: значення x_i, $i = \overline{1,n}$ величини X та відповідні значення y_i, $i = \overline{1,n}$; кількість пар – $n = 10$ невелика, тому для проведення регресійного аналізу їх можна не групувати.

Перший етап аналізу: визначимо вид залежності Y від X. Побудуємо емпіричну лінію регресії (рис. 4.3).

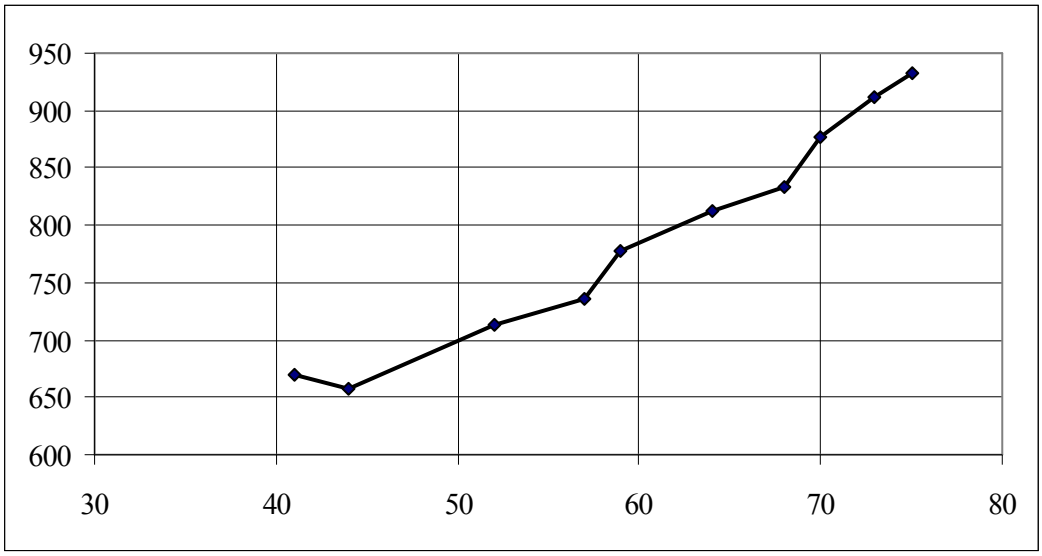


Рисунок 4.3. Емпірична лінія регресії

Оскільки емпірична лінія регресії наближається до прямої лінії, то висуваємо гіпотезу про лінійну залежність Y від X, тобто рівняння регресії будемо шукати у вигляді $y = ax + b$.

Другий етап: знайдемо параметри a, b рівняння регресії, для чого складемо систему (4.3) для не згрупованих даних. Необхідні розрахунки для зручності оформимо у вигляді таблиці (табл. 4.3).

Таблиця 4.3

Розрахункова таблиця											Суми
x_i	41	44	52	57	59	64	68	70	73	75	603
y_i	670	657	713	736	778	812	833	876	911	932	7918
x_i^2	1681	1936	2704	3249	3481	4096	4624	4900	5329	5625	37625
$x_i y_i$	27470	28908	37076	41952	45902	51968	56644	61320	66503	69900	487643

Отже, складемо систему для знаходження параметрів рівняння регресії та розв'яжемо її за правилом Крамера:

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

Знайдемо визначник основної матриці системи, яка складена із коефіцієнтів перед невідомими: $\Delta = \begin{vmatrix} 37625 & 603 \\ 603 & 10 \end{vmatrix} = 37625 \cdot 10 - 603^2 = 12641$.

Знайдемо допоміжні визначники, що отримуються із попереднього заміною відповідного стовпця коефіцієнтів на стовпець вільних членів:

$$\Delta a = \begin{vmatrix} 487643 & 603 \\ 7918 & 10 \end{vmatrix} = 487643 \cdot 10 - 603 \cdot 7918 = 101876;$$

$$\Delta b = \begin{vmatrix} 37626 & 487643 \\ 603 & 7918 \end{vmatrix} = 37626 \cdot 7918 - 48743 \cdot 603 = 3866021.$$

Знайдемо невідомі за формулами Крамера:

$$a = \frac{\Delta a}{\Delta} = \frac{101876}{12641} \approx 8,06; \quad b = \frac{\Delta b}{\Delta} = \frac{3866021}{12641} \approx 305,83.$$

Отже, шукане рівняння регресії має вигляд $y = 8,06x - 305,83$.

Третій етап: перевіримо правильність побудови моделі за рівнянням (4.4), її статистичну значущість за *F*-статистикою (4.5) і адекватність вибірковим даним за коефіцієнтом детермінації (4.6). Для чого знайдемо загальну варіацію, варіації регресії та залишків; необхідні розрахунки оформимо у вигляді таблиці (табл. 4.4).

Передусім знайдемо $\bar{y}$: $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \approx 791,8$.

Таблиця 4.4

x_i	y_i	$y_{i,\text{теор}}$	$\left(y_i - \bar{y}\right)^2$	$\left(y_{i,\text{теор}} - \bar{y}\right)^2$	$\left(y_{i,\text{теор}} - y_i\right)^2$
41	670	636,26	14835,24	24193,32	1138,52
44	657	660,44	18171,04	17256,64	11,80
52	713	724,91	6209,44	4474,42	141,82
57	736	765,20	3113,64	707,31	852,92
59	778	781,32	190,44	109,77	11,04
64	812	821,62	408,04	889,17	92,52
68	833	853,86	1697,44	3850,90	434,96
70	876	869,97	7089,64	6111,17	36,31
73	911	894,15	14208,64	10475,83	283,87
75	932	910,27	19656,04	14035,10	472,20
		Суми	85579,6	82103,63	3475,974

Отже, $Q = 85579,6$; $Q_p = 82103,63$; $Q_o = 3475,974$; тоді основне варіаційне рівняння $Q = Q_p + Q_o$ для побудованої моделі має вигляд: $85579,6 = 82103,63 + 3475,974$ і є тотожністю, тому рівняння регресії побудовано правильно.

Для перевірки статистичної значущості рівняння регресії знайдемо *F*-статистику, враховуючи, що $n = 10$, $l = 2$ – оскільки шукали рівняння з двома параметрами:

$$F = \frac{Q_p\left(n-l\right)}{Q_o\left(l-1\right)} = \frac{82103\left(10-2\right)}{3475,974\left(2-1\right)} \approx 188,96.$$

Знайдемо $F_{\text{кр}}$: $F_{\text{кр}} = FРАСПОБР\left(0,001; 2-1; 10-2\right) \approx 25,41$. Розраховане значення *F*-статистики більше критичного, тому регресійна модель є статистично значущою на рівні 0,001.

Знайдемо коефіцієнт детермінації R^2: $R^2 = \frac{Q_p}{Q} = \frac{82103,63}{85579,6} \approx 0,96$. Значення коефіцієнта детермінації свідчить, що 96% варіації результативної ознаки Y пояснюються рівнянням регресії.

Висновок: Сумарні виробничі затрати Y (тис. грн.) лінійно залежать від об'єму виробництва X (тис. од.). Залежність описується рівнянням $y = 8,06x - 305,83$, яке є статистично значущим на рівні значущості 0,001 та описує 96% вибіркових даних.

4.3. Нелінійна регресія

Якщо висунуто гіпотезу про наявність нелінійної залежності результативної ознаки (Y) від факторної (X), то регресійний аналіз проводиться за тими ж етапами, як і у випадку лінійної залежності. Вид рівнянь регресії і системи для знаходження їх параметрів для нелінійних залежностей, що найчастіше зустрічаються, надано у табл. 4.5.

Таблиця 4.5

Рівняння параболічної регресії:

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

Система для знаходження параметрів:

для згрупованих вибіркових даних:

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

для незгрупованих вибіркових даних:

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

Рівняння гіперболічної регресії:

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

Система для знаходження параметрів:

для згрупованих вибіркових даних:

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

для незгрупованих вибіркових даних:

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

Рівняння показникової регресії:

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

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

Перевірка статистичної значущості нелінійної регресійної моделі також здійснюється за *F*-статистикою. При цьому для параболічної регресії кількість параметрів $l = 3$, для гіперболічної і показникової – $l = 2$.

Приклад 4.2. Дано розподіл однотипних підприємств за об'ємом виробництва *X* (тис. од.) і собівартістю одиниці продукції *Y* (грн.) (табл. 4.6). Знайти регресійну модель, що описує залежність собівартості продукції від об'єму виробництва.

Таблиця 4.6

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

Розв'язок. Для проведення регресійного аналізу за даними табл. 4.6 побудуємо кореляційну таблицю (табл. 4.7).

Таблиця 4.7

y_j \ x_i	25	50	75	100	125	n_j
10	0	0	0	1	3	4
15	0	2	5	3	1	11
20	1	2	3	0	1	7
25	2	0	1	0	0	3
n_i	3	4	9	4	5	$n = 25$

За даними кореляційної таблиці побудуємо ряд, що відображає залежність середнього значення *Y* від *X* (табл. 3.2), для чого знайдемо середні значення $\overline{y}_{x_i}$ для кожного значення x_i, $i = \overline{1,5}$ і заповнимо табл. 4.8:

$$\overline{y}_{x_1} = \frac{y_1 n_{11} + y_2 n_{12} + y_3 n_{13} + y_4 n_{14}}{n_1} = \frac{20 \cdot 1 + 25 \cdot 2}{3} \approx 23{,}33;\ \overline{y}_{x_2} = \frac{15 \cdot 2 + 20 \cdot 2}{4} = 17{,}5;$$

$$\overline{y}_{x_3} = \frac{15 \cdot 5 + 20 \cdot 3 + 25 \cdot 1}{9} = 17{,}78;\ \overline{y}_{x_4} = \frac{10 \cdot 1 + 15 \cdot 3}{4} = 13{,}75;$$

$$\overline{y}_{x_5} = \frac{10 \cdot 3 + 15 \cdot 1 + 20 \cdot 1}{5} = 13.$$

Таблиця 4.8

x_i	25	50	75	100	125
$\overline{y}_{x_i}$	23,33	17,5	17,78	13,75	13
n_i	3	4	9	4	5

Перший етап аналізу: визначимо вид залежності Y від X. Побудуємо емпіричну лінію регресії (рис. 4.4).

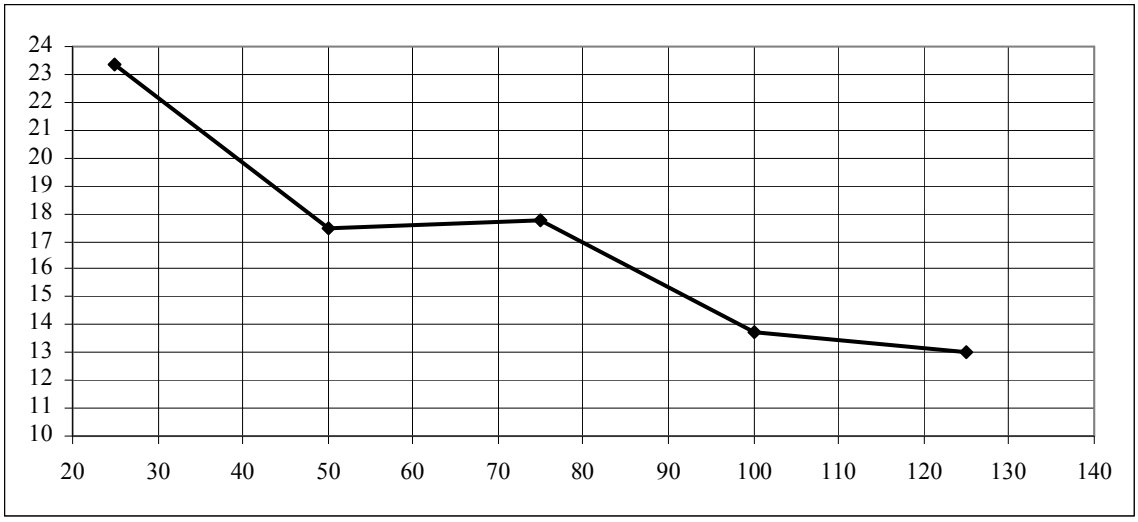


Рисунок 4.4. Емпірична лінія регресії

Оскільки емпірична лінія регресії наближається до гіперболи, то висуваємо гіпотезу про гіперболічну залежність Y від X, тобто рівняння регресії будемо шукати у вигляді $\overline{y}_x = \frac{a}{x} + b$.

Другий етап: знайдемо параметри a, b рівняння регресії, для чого складемо систему для згрупованих даних. Необхідні розрахунки для зручності оформимо у вигляді таблиці (табл. 4.9), в останньому рядку якої знайдемо відповідні стовпцям суми.

Таблиця 4.9

x_i	$\overline{y}_{x_i}$	n_i	$\frac{1}{x_i}$	$\frac{1}{x_i} n_i$	$\frac{1}{x_i^2} n_i$	$\overline{y}_{x_i} n_i$	$\frac{1}{x_i} \overline{y}_{x_i} n_i$
25	23,33	3	0,04	0,12	0,0048	69,99	2,7996
50	17,5	4	0,02	0,08	0,0016	70	1,4
75	17,78	9	0,0133	0,12	0,0016	160,02	2,1336
100	13,75	4	0,01	0,04	0,0004	55	0,55
125	13	5	0,008	0,04	0,0003	65	0,52
			Суми	0,4	0,0087	420,01	7,4032

Отже, складемо систему для знаходження параметрів рівняння регресії та розв’яжемо її за правилом Крамера:

$$\begin{cases} a\sum\limits_{i=1}^{k}\frac{1}{x_i^2}n_i + b\sum\limits_{i=1}^{k}\frac{1}{x_i}n_i = \sum\limits_{i=1}^{k}\frac{1}{x_i}\overline{y_{x_i}}n_i \\ a\sum\limits_{i=1}^{k}\frac{1}{x_i}n_i + b\sum\limits_{i=1}^{k}n_i = \sum\limits_{i=1}^{k}\overline{y_{x_i}}n_i \end{cases} \Rightarrow \begin{cases} 0,0087a + 0,4b = 7,4032 \\ 0,4a + 25b = 420,01 \end{cases}.$$

Головний визначник системи: $\Delta = \begin{vmatrix} 0,00872 & 0,4 \\ 0,4 & 25 \end{vmatrix} = 0,00872 \cdot 25 - 0,4^2 = 0,058$.

Допоміжні визначники: $\Delta a = \begin{vmatrix} 7,4032 & 0,4 \\ 420 & 25 \end{vmatrix} = 7,4032 \cdot 25 - 0,4 \cdot 420 = 17,076$;

$$\Delta b = \begin{vmatrix} 0,00872 & 7,4032 \\ 0,4 & 420 \end{vmatrix} = 0,00872 \cdot 420 - 7,4032 \cdot 0,4 = 0,701207.$$

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

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

Третій етап: перевіримо правильність побудови моделі за рівнянням (4.4), її статистичну значущість за *F*-статистикою (4.5) і адекватність вибірковим даним за коефіцієнтом детермінації (4.6). Для цього знайдемо загальну варіацію, варіації регресії та залишків; необхідні розрахунки оформимо у вигляді таблиці (табл. 4.10).

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

Таблиця 4.10

x_i	y_i	n_i	$y_{i,\text{теор}}$	$\left(y_i - \overline{y}\right)^2 n_i$	$\left(y_{i,\text{теор}} - \overline{y}\right)^2 n_i$	$\left(y_{i,\text{теор}} - y_i\right)^2 n_i$
25	23,33	3	23,866	127,907	149,782	0,863
50	17,5	4	17,978	1,958	5,547	0,914
75	17,78	9	16,015	8,637	5,547	28,028
100	13,75	4	15,034	37,220	12,482	6,594
125	13	5	14,445	72,215	27,737	10,441
			Суми	247,936	201,096	46,840

Отже, $Q = 247,936$; $Q_p = 201,096$; $Q_o = 46,840$; тоді основне варіаційне рівняння $Q = Q_p + Q_o$ для побудованої моделі має вигляд: $247,936 = 201,096 + 46,840$ і є тотожністю, тому рівняння регресії побудовано правильно.

Для перевірки статистичної значущості рівняння регресії знайдемо *F*-статистику, враховуючи, що $n = 25$, $l = 2$ – оскільки шукали рівняння з двома параметрами:

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

Знайдемо $F_{кр}$: $F_{кр} = FРАСПОБР(0{,}001; 2-1; 25-2) \approx 14{,}20$. Розраховане значення F-статистики більше критичного, тому регресійна модель є статистично значущою на рівні 0,001.

Знайдемо коефіцієнт детермінації R^2: $R^2 = \frac{Q_p}{Q} = \frac{201{,}096}{247{,}936} \approx 0{,}81$. Значення коефіцієнта детермінації свідчить, що 81% варіації результативної ознаки Y пояснюється рівнянням регресії.

Висновок: Залежність собівартості одиниці продукції Y (грн.) від об'єму виробництва X (тис. од.) описується рівнянням $\overline{y_x} = \frac{294{,}41}{x} + 12{,}09$, яке є статистично значущим на рівні значущості 0,001 та описує 81% вибіркових даних.

4.4. Множинна лінійна регресія

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

Загальний вигляд багатофакторної лінійної регресійної моделі:

$$Y = f(X_1, X_2, \ldots, X_m) + \varepsilon, \qquad (4.7)$$

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

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

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

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

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

Відносну величину впливу факторних ознак на результативну можна оцінити за формулою:

$$r_{X_i}^{\,2} = \frac{t_{X_i}^{\,2} \cdot R^2}{\sum_{i=1}^{m} t_{X_i}^{\,2}}; \tag{4.8}$$

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

R^2 – загальний коефіцієнт детермінації моделі.

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

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

− параметри лінійної регресійної моделі виду $Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_m X_m$, де b_0, b_1, b_2, …, b_m – параметри моделі;

− коефіцієнт детермінації;

− критеріальну статистику для перевірки статистичної значущості моделі;

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

Зауваження 2. Засобом ***Регрессия*** можна також користуватися при побудові моделі, нелінійної відносно певної факторної ознаки X_j, але лінійної відносно коефіцієнта цієї ознаки. Наприклад, якщо необхідно побудувати модель виду $Y = b_0 + b_1 X_1 + b_2 X_2^3$, то як вхідні дані вказують трійки (y_i, x_{1i}, x_{2i}^3).

4.5. Регресія у Microsoft Excel

Пакет аналізу даних Microsoft Excel надає можливість будувати регресійні моделі, але тільки у випадку лінійної залежності результативної ознаки Y від факторної ознаки X і тільки для незгрупованих вибіркових даних.

Для побудови лінійної регресійної моделі необхідно:

1) Викликати ***Сервис – Анализ данных – Регрессия – ОК***. З'явиться вікно для надання вхідних даних (рис. 4.5).

2) У графі ***Входной интервал Y*** та ***Входной интервал X*** вказати відповідні стовпці даних; у графі ***Выходной интервал*** вказати ту чарунку, починаючи з якої будуть надаватися вихідні дані – параметри рівняння регресії та результати її статистичного аналізу.

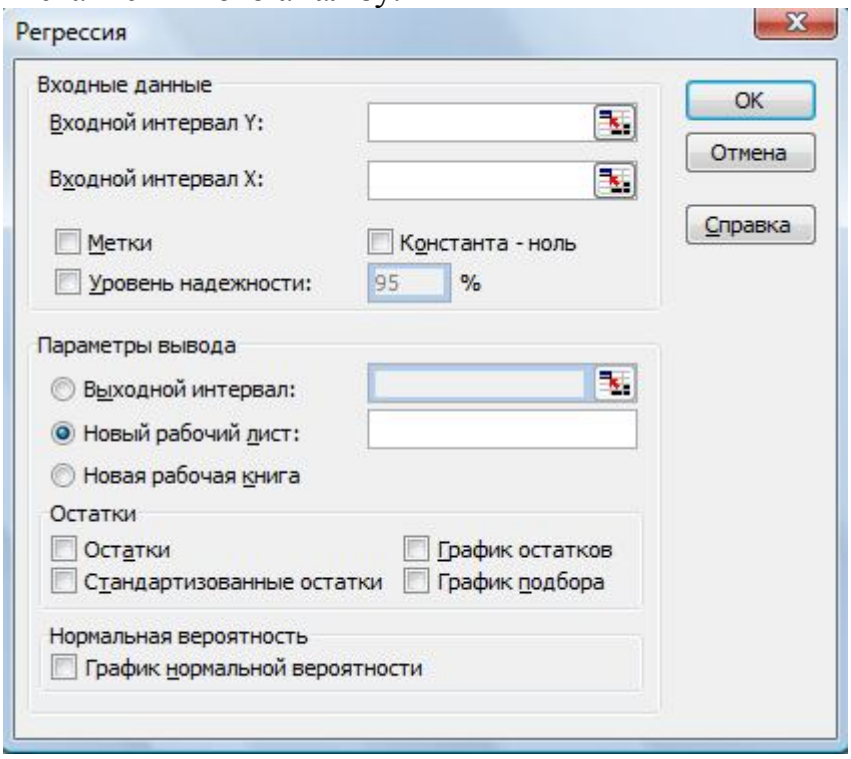


Рисунок 4.5. Діалогове вікно функції ***Регрессия***

Приклад і результати роботи функції ***Регрессия*** представлено на рис. 4.6.

Файл Правка Вид Вставка Формат Сервис Данные Окно Справка

I23

	A	B	C	D	E	F	G	H	I	J
1					**Регресійний аналіз**					
2		Вхідні дані			Вихідні дані					
3	№	Значення			ВЫВОД ИТОГОВ					
4	**i**	**X**	**Y**							
5	1	1	328		*Регрессионная статистика*					
6	2	2	329		Множественный R	0,972633354				
7	3	3	329		R-квадрат	0,946015642				
8	4	4	345		Нормированный R-квадрат	0,937018249				
9	5	5	352		Стандартная ошибка	5,786032717				
10	6	6	370		Наблюдения	8				
11	7	7	377							
12	8	8	385		Дисперсионный анализ					
13						*df*	*SS*	*MS*	*F*	*Значимость F*
14					Регрессия	1	3520,005952	3520,005952	105,1433059	5,01935E-05
15					Остаток	6	200,8690476	33,4781746		
16					Итого	7	3720,875			
17										
18						*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	
19					Y-пересечение	310,6785714	4,508440371	68,91043151	6,28282E-10	
20					Переменная X 1	9,154761905	0,892804231	10,25394099	5,01935E-05	

Рисунок 4.6. Результати регресійного аналізу

В таблиці (рис. 4.6) у графі ***Коэффициенты*** вказані значення параметрів моделі a та b: b − в графі ***Y-пересечение***, a − в графі ***Переменная X1***. Отже, побудована лінійна регресійна модель має вигляд:

$$y = 69{,}15x + 310{,}68\,.$$

Для перевірки статистичної значущості моделі надається значення F-статистики у графі F: F = 105,14. Це значення обчислюється як відношення варіації регресії до варіації залишків (чарунки H14 та H15). В стовпці ***Значимость F*** надано критеріальну статистику. Якщо це значення менше, ніж, наприклад, 0,05, то рівняння регресії є значущим на рівні 0,05. У даному завданні рівняння регресії є значущим на рівні 0,00005.

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

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

Крім того, можна задати: графік підбору – порівняльна діаграма, що містить емпіричну і теоретичну лінії регресії; таблиця залишків – різниць емпіричних і теоретичних значень Y (рис. 4.7).

Регресійний аналіз

Вхідні дані

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

Вихідні дані

ВЫВОД ОСТАТКА

Наблюдение	Предсказанное Y	Остатки
1	319,8333333	8,166666667
2	328,9880952	0,011904762
3	338,1428571	-9,142857143
4	347,297619	-2,297619048
5	356,452381	-4,452380952
6	365,6071429	4,392857143
7	374,7619048	2,238095238
8	383,9166667	1,083333333

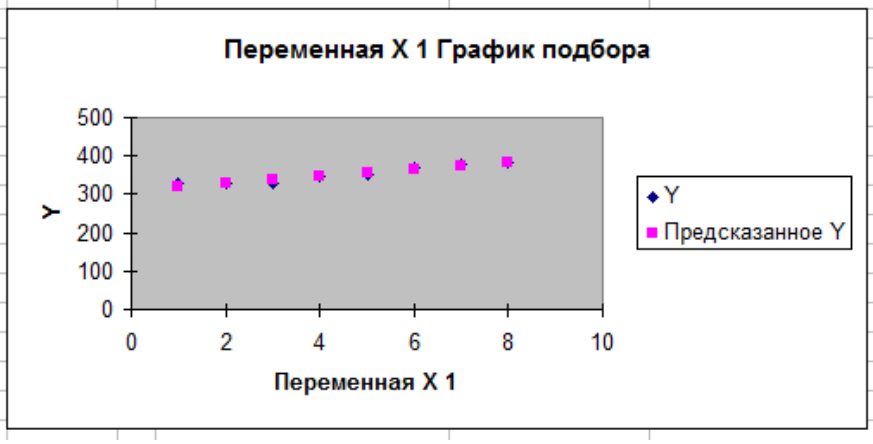


Рисунок 4.7. Додаткові результати регресійного аналізу

Розглянемо детальніше методику побудови множинної лінійної регресійної моделі засобами Microsoft Excel.

Приклад 4.3. В таблиці 4.11 вказано дані по консервному заводу с. Виноградне Одеської області за 12 місяців минулого року (табл. 4.11).

Умовні позначення:

X_1 – часовий фактор, порядковий номер місяця;

X_2 – фонди (тис. грн./робітника);

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

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

Y – валова продукція (тис. умовних банок).

Таблиця 4.11

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

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

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

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

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

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

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

Парні коефіцієнти кореляції обчислимо за допомогою вбудованих сервісних функцій Excel: перенесемо табл. 4.11 на сторінку Excel, викличемо ***Сервис – Аналіз данных – Корреляция – ОК***. У графі ***Входной интервал***

вкажемо масив даних табл. 4.11; у графі ***Группирование*** вкажемо ***По столбцам***, у графі ***Выходной интервал*** вкажемо ту чарунку, починаючи з якої будуть виводитися вихідні дані – парні коефіцієнти кореляції. Отримаємо табл. 4.12, яка є матрицею парних коефіцієнтів кореляції. Чарунки таблиці, розташовані вище головної діагоналі, незаповнені, оскільки таблиця симетрична відносно головної діагоналі.

Таблиця 4.12

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

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

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

Доповнимо табл. 4.12. Виділимо в ній елементи, які більші за $r_{кр}$ (це означає, що відповідні ознаки тісно пов'язані між собою). Отримаємо табл. 4.13.

Таблиця 4.13

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

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

X₁ та ***X₂*** оскільки $r(X_1, X_2) = 0,89 > 0,57598$;

X₂ та ***X₃*** оскільки $r(X_2, X3) = 0,69 > 0,57598$;

X₂ та ***X₄*** оскільки $r(X_2, X_4) = 0,66 > 0,57598$;

X₃ та ***X₄*** оскільки $r(X_3, X_4) = 0,94 > 0,57598$.

З факторних ознак тісно пов'язані із результативною ознакою (із ***У***):

X₂ оскільки $r(X_2, Y) = 0,63 > 0,57598$;

X₃ оскільки $r(X_3, Y) = 0,94 > 0,57598$;

X₄ оскільки $r(X_4, Y) = 0,99 > 0,57598$.

За результатами аналізу кореляційної матриці побудуємо кореляційні плеяди, тобто зобразимо достовірний зв'язок між факторними ознаками графічно (рис. 4.8).

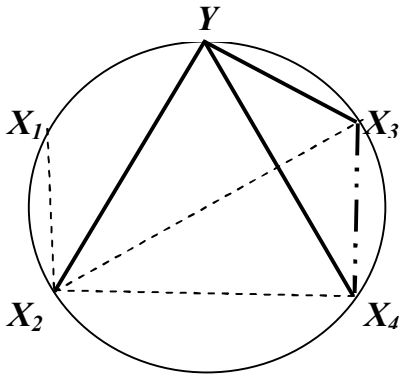


Рис.4.8. Кореляційний зв'язок між факторними ознаками

Перша кореляційна плеяда: ***Y, X₂ , X₃ , X₄*** вказує, що в модель необхідно включити ознаки ***X₂ , X₃ , X₄***, оскільки вони мають зв'язок (тобто впливають) на результативну ознаку ***Y***. Тобто із всіх факторних ознак в модель не потрібно включати ознаку ***X₁***.

Друга кореляційна плеяда: ***X₂, X₁ , X₃ , X₄*** вказує, що в модель можна включити тільки одну з ознак ***X₂, X₁ , X₃ , X₄***, оскільки вони пов'язані між собою. Однак наявність дуже сильного (0,94) зв'язку між ***X₃*** та ***X₄*** свідчить про те, що зв'язок може існувати між ***X₂*** та ***X₃*** , а між ***X₂*** та ***X₄*** він може бути тільки наслідком зв'язку ***X₃ − X₄***. Або навпаки, зв'язок може існувати між ***X₂*** та ***X₄*** , а між ***X₂*** та ***X₃*** він може бути тільки наслідком зв'язку ***X₃ − X₄***. Тому, можливо, в модель потрібно включати ***X₂*** та одну із ознак ***X₃*** та ***X₄***. Для того, щоб виявнити це, скористуємось частинними коефіцієнтами кореляції.

Розрахуємо частинні коефіцієнти кореляції між факторними ознаками ***X₂*** і ***X₃***, та між ***X₂*** і ***X₄***. Величина цих частинних коефіцієнтів кореляції дозволить визначити „чистий" зв'язок між вказаними ознаками, тобто зв'язок, що не залежить від впливу всіх останніх факторних ознак.

Частинний коефіцієнт кореляції між факторними ознаками ***X₂*** і ***X₃*** розраховуємо за формулою:

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

де A_{23} – алгебраїчне доповнення елемента r_{23},

A_{22} – алгебраїчне доповнення елемента r_{22},

A_{33} – алгебраїчне доповнення елемента r_{33}.

Алгебраїчне доповнення A_{23} – це визначник матриці, отриманої із матриці A викреслюванням 2-го рядка і 3-го стовпця, помножений на $(-1)^{2+3}$. Аналогічно A_{22} – це визначник матриці, отриманої із матриці A викреслюванням 2-го рядка і 2-го стовпця, помножений на $(-1)^{2+2}$; A_{33} – це визначник матриці, отриманої із матриці A викреслюванням 3-го рядка і 3-го стовпця, помножений на $(-1)^{3+3}$. Визначники обчислюємо за допомогою вбудованих функцій Excel: викликаємо ***Функции – Математические –***

МОПРЕД, у графі ***Массив*** вказуємо матрицю, визначник якої потрібно знайти. Отримаємо:

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

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

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

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

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

Висновок з етапу 1: шукана багатофакторна лінійна регресійна модель має вигляд: $Y=b_0+b_1X_2+b_2X_4$.

Етап 2. Побудуємо вказану модель. Для знаходження b_0, b_1, b_2 викликаємо ***Сервис – Анализ данных – Регрессия – ОК***. У графі ***Входной интервал Y*** вкажемо відповідний стовпчик даних табл. 4.11; у графі ***Входной интервал X*** вкажемо стовпчики ***X_2*** та ***X_4*** табл 4.11; у графі ***Выходной интервал*** вкажемо ту чарунку, починаючи з якої будуть виводитися вихідні дані – рівняння регресії. Отримаємо таблицю з результатами регресійного аналізу (рис. 4.9).

ВЫВОД ИТОГОВ

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

Дисперсионный анализ

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

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

Рисунок 4.9. Результати регресійного аналізу

В таблиці (рис. 4.9) у графі ***Коэффициенты*** вказані значення параметрів моделі b_0, b_1, b_2: b_0 − в графі ***Y*-пересечение**, b_1 − в графі **Переменная *X1***, b_2 − в графі **Переменная *X2*.** Отже, b_0=657,67; b_1= −1,76; b_2=927,05; багатофакторна лінійна регресійна модель має вигляд:

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

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

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

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

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

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

Для наочності висновків зобразимо емпіричну і теоретичну лінії регресії (рис. 4.10).

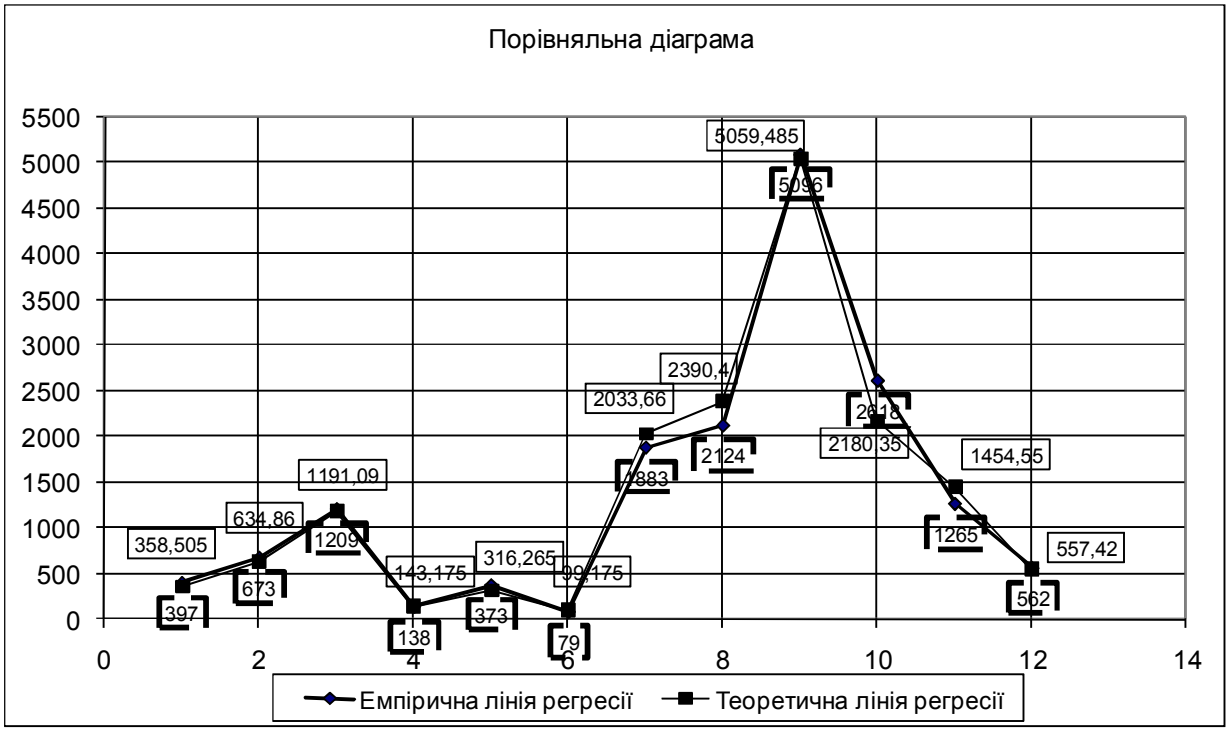


Рисунок 4.10. Порівняльна діаграма за результатами регресійного аналізу

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

4.6. Регресійний аналіз засобами SPSS

Щоб знайти та дослідити рівняння лінії регресії, варто побудувати емпіричну лінію регресії та визначити її вид. За допомогою SPSS можна вивчати кілька видів регресійного аналізу, які зображені на рис. 4.11.

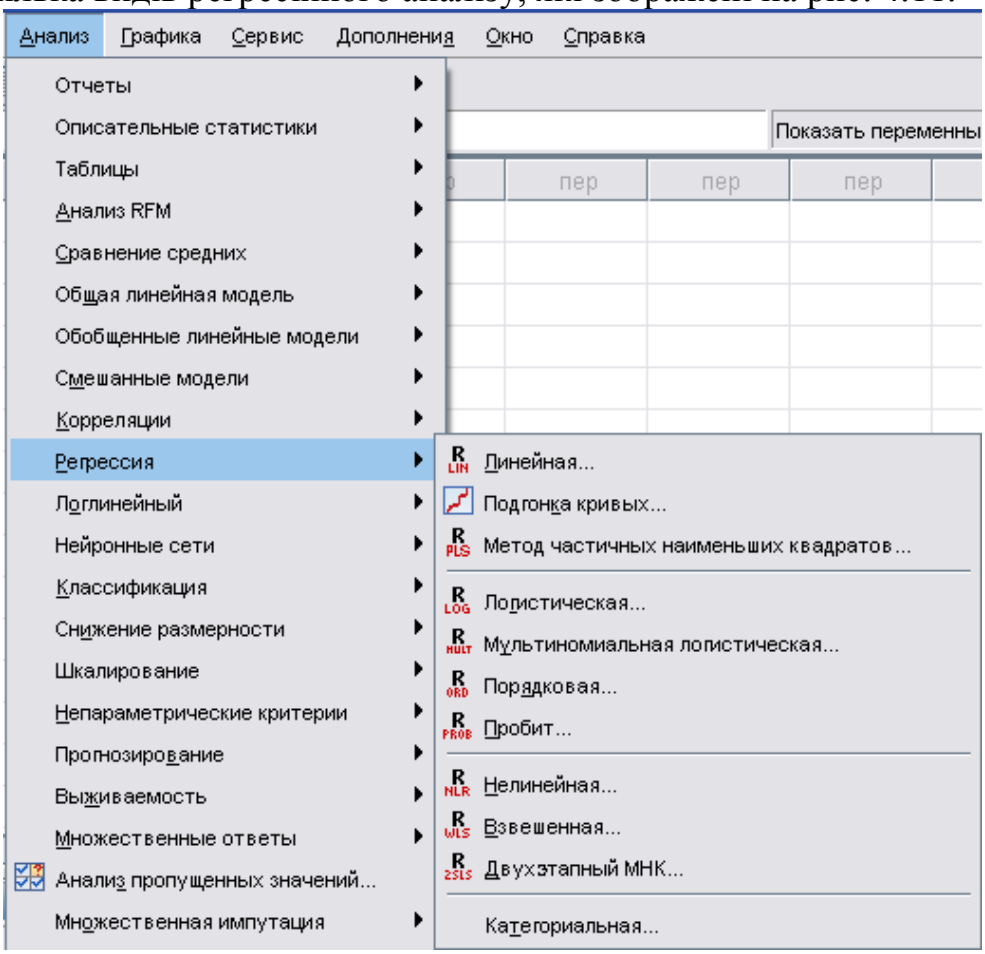


Рисунок 4.11. Види регресій у SPSS

4.6.1. Лінійна регресія

Знайдемо рівняння лінії регресії та побудуємо регресійну пряму, які характеризують залежність сумарних виробничих затрат Y (тис. грн.) від обсягів виробництва X (тис. од.) за статистичними даними, представленими у табл. 4.2 із прикладу 4.1.

1) Введемо стрічкові дані табл. 4.2 у два стовпчики вкладки *Набор данных* і побудуємо регресійну пряму. Виберемо в меню ***Графика*** – ***Рассеяния/точки***, відкриється діалогове вікно у якому виберемо вид ***Простая диаграмма рассеяния*** і натиснемо ***Задать***. На вісь OX перенесемо змінну X, на OY – змінну Y.

Отримаємо графік (рис. 4.12), який дає підстави спрогнозувати лінійну залежність Y від X. Отже, вивчаємо лінійну регресію.

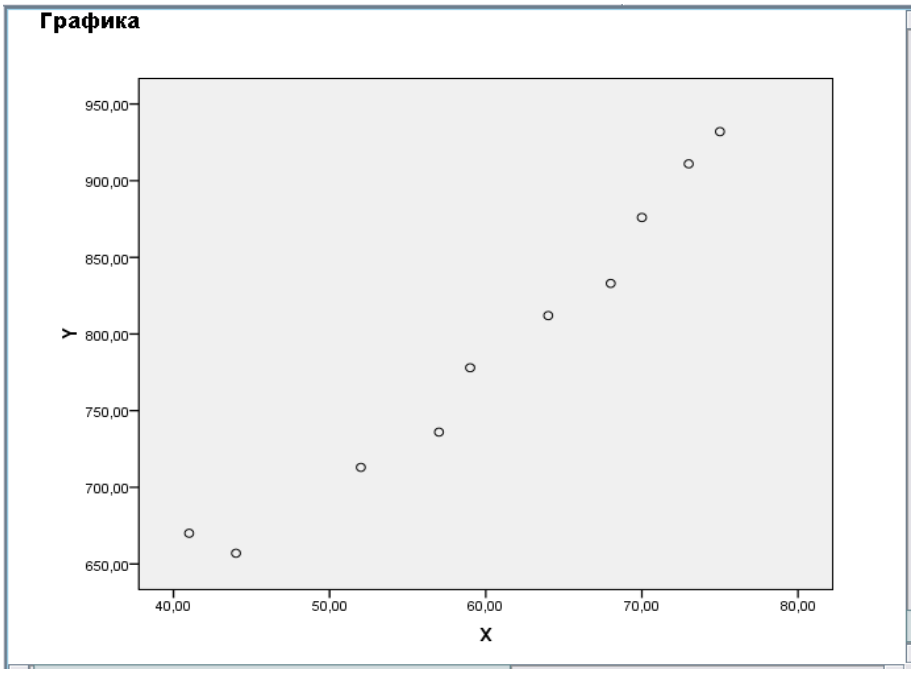


Рисунок 4.12. Графік емпіричної лінії регресії засобами SPSS

2) Виберемо в меню послідовно ***Анализ − Регрессия − Линейная***. У діалоговому вікні ***Линейная регрессия*** перенесемо змінну X у поле ***Независимые переменные***, а змінну Y − в поле ***Зависимые переменные*** (рис. 4.13) та перейдемо у вікно виведення результатів;

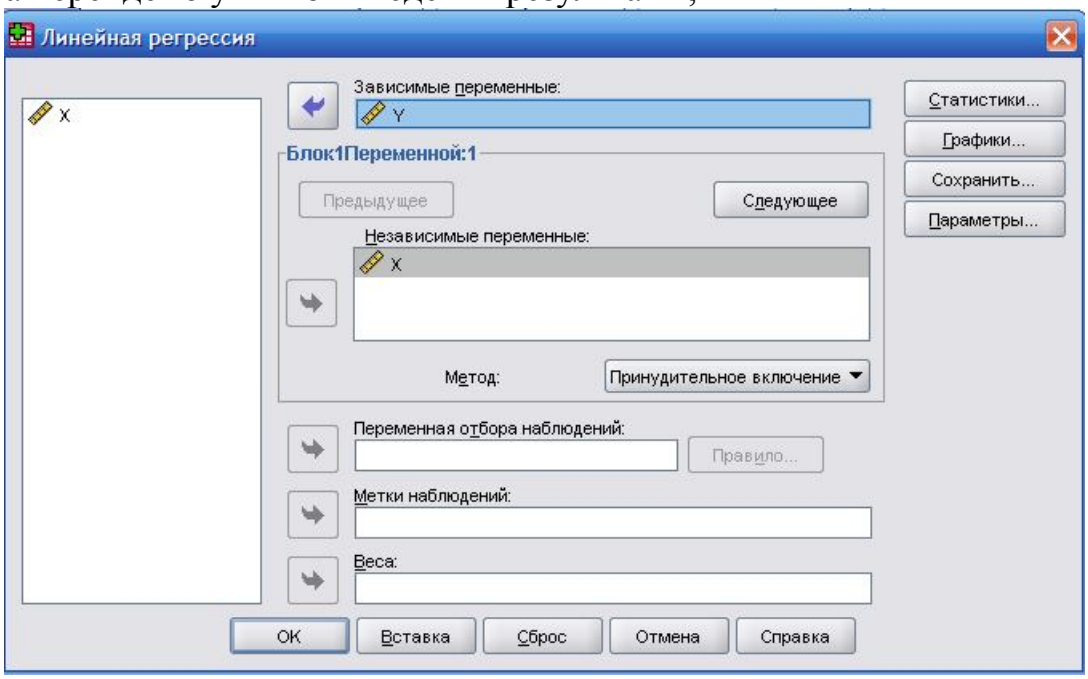


Рисунок 4.13. Діалогове вікно вибору параметрів лінійної регресії

3) В таблиці ***Сводка для модели*** (рис. 4.14) під назвою ***R-квадрат*** зберігається значення коефіцієнта детермінації 0,959, який свідчить про те, що рівняння регресії описує 95,9% вибіркових даних. У таблиці ***Коэффициенты*** в першому стовпчику знаходяться коефіцієнти рівняння регресії, а саме: $y = 8,059x - 305,832$.

Сводка для модели

Модель	H	R-квадрат	Скорректированный R-квадрат	Стд. ошибка оценки
1	,979[a]	,959	,954	20,84459

a. Предикторы: (конст) X

Дисперсионный анализ[b]

Модель		Сумма квадратов	ст.св.	Средний квадрат	Щ	Знч.
1	Регрессия	82103,626	1	82103,626	188,963	,000[a]
	Остаток	3475,974	8	434,497		
	Всего	85579,600	9			

a. Предикторы: (конст) X

b. Зависимая переменная: Y

Коэффициенты[a]

Модель		Нестандартизованные коэффициенты		Стандартизованные коэффициенты	t	Знч.
		B	Стд. Ошибка	Бета		
1	(Константа)	305,832	35,962		8,504	,000
	X	8,059	,586	,979	13,746	,000

a. Зависимая переменная: Y

Рисунок 4.14. Результати розрахунку рівняння лінійної регресії

4.6.2. Нелінійна регресія

Знайдемо рівняння лінії регресії та побудуємо регресійну пряму, які характеризують залежність собівартості одиниці продукції Y (грн.) від обсягів виробництва X (тис. од.) за статистичними даними, представленими у табл. 4.6 із прикладу 4.2.

Для зручності, скористаємось елементами вище проведеного дослідження, де, згідно графіка, припускається, що емпірична лінія регресії наближається до гіперболи і висувається гіпотеза про гіперболічну залежність Y від X:

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

Введемо стрічкові дані таблиці 4.8 у стовпчики вкладки ***Данные*** редактора ***Набор данных***.

1) Виберемо в меню послідовно ***Анализ*** − ***Регрессия*** − ***Нелинейная***. У діалоговому вікні ***Нелинейная регрессия*** (рис. 4.15) перенесемо змінну Y у поле ***Зависимые переменные***, у полі ***Выражение, задающее модель*** задаємо вираз: $a/X+b$ та активуємо кнопку ***Параметры***, перейшовши у вікно ***Нелинейная регрессия***: ***Параметры*** (рис. 4.15);

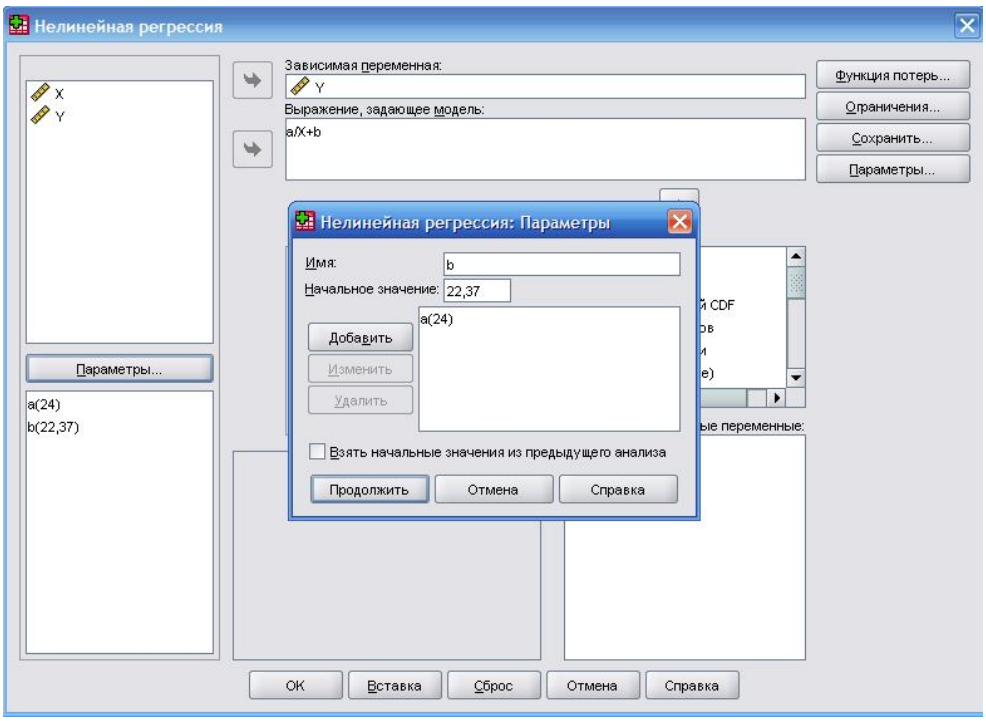


Рисунок 4.15. Діалогове вікно вибору параметрів нелінійної регресії

2) В полі ***Имя*** (діалогове вікно на першому плані рис. 4.15) введемо a, його початкове значення має «покривати» максимальне значення y_i із табл. 4.8. Тому вибираємо значення − 24. Натискаємо ***Добавить*** і задаємо значення b − 22,37 (так як $x_1 = 25,\ a = 24,\ y_1 = 23,33$ і $b = y_1 - \frac{a}{x_1}$). Переходимо у вікно виведення результатів (рис. 4.16).

Оценки параметра

Параметр	Оценка	Стд. Ошибка	Доверительный интервал 95 %	
			Нижняя граница	Верхняя граница
a	300,739	56,971	119,432	482,045
b	11,579	1,233	7,655	15,502

Корреляции оценок параметров

	a	b
a	1,000	-,844
b	-,844	1,000

Дисперсионный анализ[a]

Источник	Сумма квадратов	ст.св.	Средние квадраты
Регрессия	1518,173	2	759,086
Остаток	6,557	3	2,186
Нескорректированный итог	1524,730	5	
Скорректированный итог	67,464	4	

Зависимая переменная: Y

a. R в квадрате = 1 - (остаточная сумма квадратов) / (скорректированная сумма квадратов) = ,903.

Рисунок 4.16. Результати розрахунку рівняння параболічної регресії

Із першого стовпчика таблиці ***Оценки параметра*** знаходимо коефіцієнти і складаємо рівняння: $y = \frac{300,7}{x} + 11,6$, яке описує 90,3% вибіркових даних, про що свідчить коефіцієнт детермінації.

Якщо гіпотетично неможливо зробити припущення про вид лінії регресії, то варто використати можливості програми щодо підбору виду кривої.

Приклад 4.4. Побудувати регресійну модель, що характеризує залежність обсягу продажу деякої продукції в день *Y* (тис. грн.) від кількості днів рекламної компанії *X* (дні). Дані наведено у табл. 4.14.

Таблиця 4.14

X	11	12	22	23	25	30	34	56	78	90
Y	540	530	505	490	483	465	470	485	484	470

Розв'язок. Для спрощення дослідження про вид та рівняння емпіричної лінії регресії, необхідно:

1) Ввести дані табл. 4.14 у стовпчики вкладки ***Данные*** редактора ***Набор данных***.

2) Вибрати в меню послідовно ***Анализ − Регрессия − Подгонка кривых***. У діалоговому вікні ***Подгонка кривых*** перенести змінну *X* у поле ***Независимая переменная***, а змінну *Y* у поле ***Зависимые*** і активувати усі моделі у полі ***Модели***, відзначивши їх галочками (рис. 4.17);

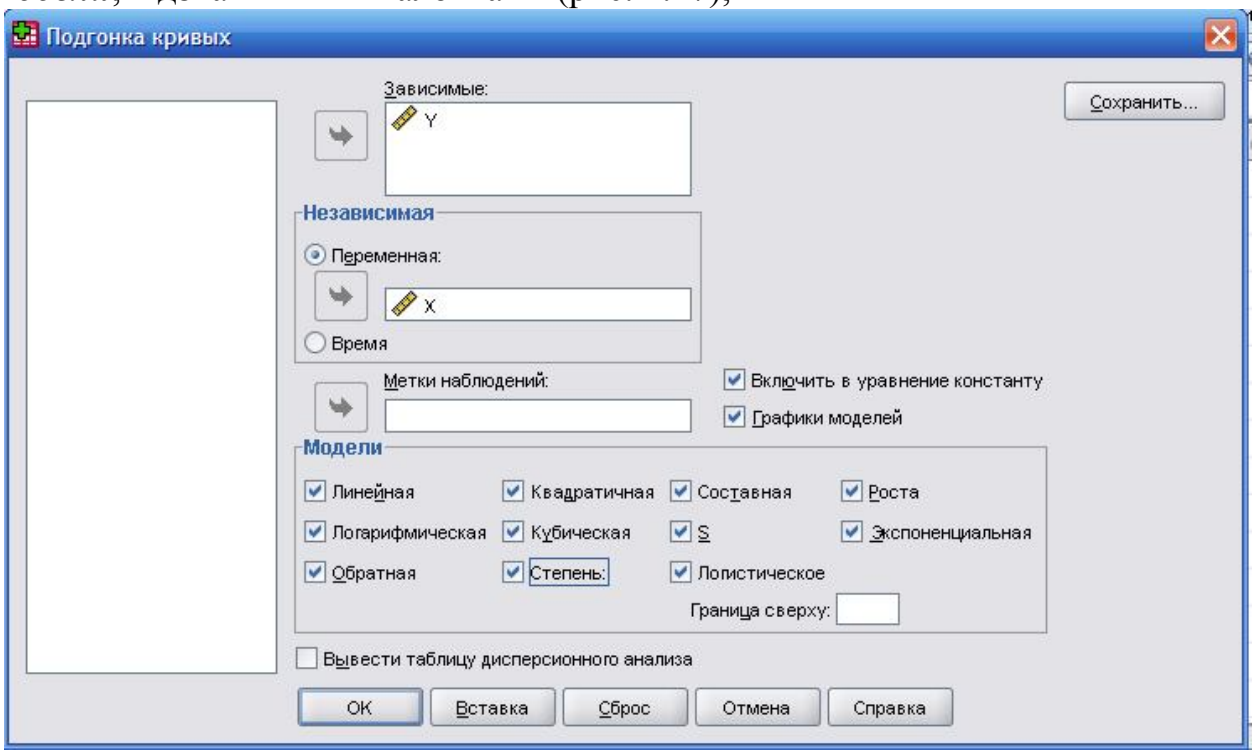


Рисунок 4.17. Діалогове вікно вибору гіпотетичної моделі лінії регресії

3) Проаналізувати дані таблиць вікна виводу результатів (рис. 4.18).

Сводка модели и оценки параметров

Зависимая переменная:Y

Уравнение	Сводка для модели					Оценки параметра			
	R-квадрат	F	ст.св.1	ст.св.2	Знч.	Константа	b1	b2	b3
Линейный	,335	4,039	1	8	,079	512,657	-,537		
Логарифмическая	,577	10,916	1	8	,011	584,680	-27,090		
Обратная	,795	30,985	1	8	,001	458,412	828,724		
Квадратичный	,665	6,939	2	7	,022	557,874	-3,225	,027	
Кубический	,941	31,955	3	6	,000	621,955	-9,391	,183	-,001
Составная	,334	4,014	1	8	,080	512,038	,999		
Степенная	,572	10,711	1	8	,011	590,673	-,054		
S	,786	29,307	1	8	,001	6,131	1,642		
Роста	,334	4,014	1	8	,080	6,238	-,001		
Экспоненциальная	,334	4,014	1	8	,080	512,038	-,001		
Логистическая	,334	4,014	1	8	,080	,002	1,001		

Независимой переменной является X.

Рисунок 4.18. Результати підбору виду регресії

Дані першого стовпчика таблиці ***Сводка модели и оценки параметров*** (рис. 4.18) ***R-квадрат*** є коефіцієнтами детермінації, які показують скільки відсотків вибіркових об'єктів охоплює кожний вид рівняння. Необхідно вибрати максимальне значення: в нашому випадку − 0,941.

Отже, емпіричною лінією регресії є кубічна парабола, яка охоплює 94,1% досліджуваних даних. Це підтверджує також найменше значення рівня значущості серед усіх інших видів рівняння p = 0,000 (у таблиці стовпчик ***Знч***).

Запишемо рівняння: $\overline{y}_x = 621{,}96 - 9{,}391x + 0{,}183x^2 - 0{,}001x^3$, графік якого, поряд з іншими, зображений на рис. 4.19.

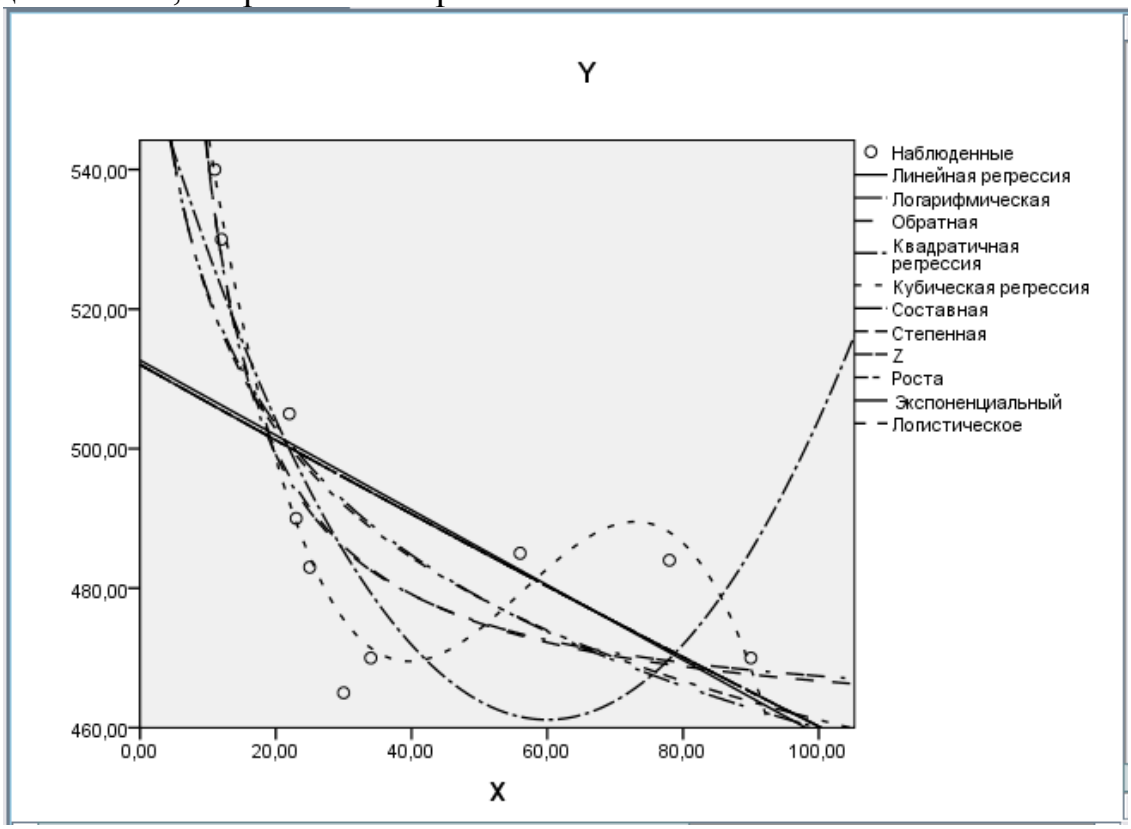


Рисунок 4.19. Графіки підгонки кривих

Висновок: Залежність обсягу продажу деякої продукції в день Y (тис. грн.) від кількості днів рекламної компанії X (дні) описується рівнянням $\bar{y}_x = 621{,}96 - 9{,}391x + 0{,}183x^2 - 0{,}001x^3$, яке характеризує 94,1% варіації результативної ознаки Y з ймовірністю випадковості отриманого результату $p = 0{,}000$.

4.6.3. Множинна лінійна регресія

Множинний регресійний аналіз передбачає вивчення залежності між кількома незалежними ознаками. Ознаки можуть належати до інтервальної або порядкової шкал. Якщо ж ознака відноситься до номінальної шкали і може бути дихотомічною, то її можна розписати на кілька дихотомічних змінних. Наприклад, ознаку *Освіта (середня, середня професійна, неповна вища, вища)* можна представити як: *Освіта1 (1 − середня, 0 − не середня), Освіта2 (1 − середня професійна, 0 − не середня професійна); Освіта3 (1 − неповна вища, 0 − не неповна вища)* і т. д.

Множинний аналіз лінійної регресії в SPSS можна провести з допомогою кількох методів. Автоматично активований метод *Принудительное включение*, який не варто використовувати для множинного аналізу. Даний метод передбачає одночасну обробку усіх незалежних ознак, тому об'єктивними можуть бути лише результати аналізу рівняння регресії з однією ознакою (змінною). Для множинного аналізу варто вибрати один із покрокових методів. Використовуючи прямий метод, незалежні змінні, які мають найбільші значення коефіцієнтів частинної кореляції з залежною змінною, поетапно вносяться у рівняння регресії. Обернений метод передбачає вилучення незалежних змінних з найменшими значеннями частинних коефіцієнтів кореляції із гіпотетичного рівняння лінії регресії, яке містить усі змінні. Процес продовжується до того часу, поки відповідний регресійний коефіцієнт не виявиться незначущим (у даному випадку рівень значущості дорівнює 0,1).

Приклад 4.5. У 10 компаніях вивчається взаємозв'язок між середньорічними цінами на рекламу X_1 (млн. грн.), рівнем затрат на проведення реклами X_2 (% до вартості реалізованої продукції) та вартістю реалізованої рекламної продукції Y (млн. грн.). Дані наведено у таблиці 4.15

Таблиця 4.15

№ компанії	X_1	X_2	Y
1	3	4	20
2	3	3	25
3	5	3	20
4	6	5	30
5	7	10	32
6	6	12	25
7	8	12	29
8	9	11	37
9	9	15	36
10	10	15	40

Вважаючи, що між показниками існує лінійна залежність, визначити параметри рівняння регресії та оцінити адекватність обраної моделі.

Розв'язок. Для знаходження коефіцієнтів рівняння емпіричної лінії регресії, необхідно:

1) Ввести дані табл. 4.15 у стовпчики вкладки ***Данные*** редактора ***Набор данных***.

2) Вибрати в меню послідовно ***Анализ − Регрессия − Линейная…*** У діалоговому вікні ***Линейная регрессия*** перенести змінні X_1, X_2 у поле ***Независимые переменные***, а змінну Y у поле ***Зависимые***. Вибрати один із обернених покрокових методів: ***Удалить*** (рис. 4.20).

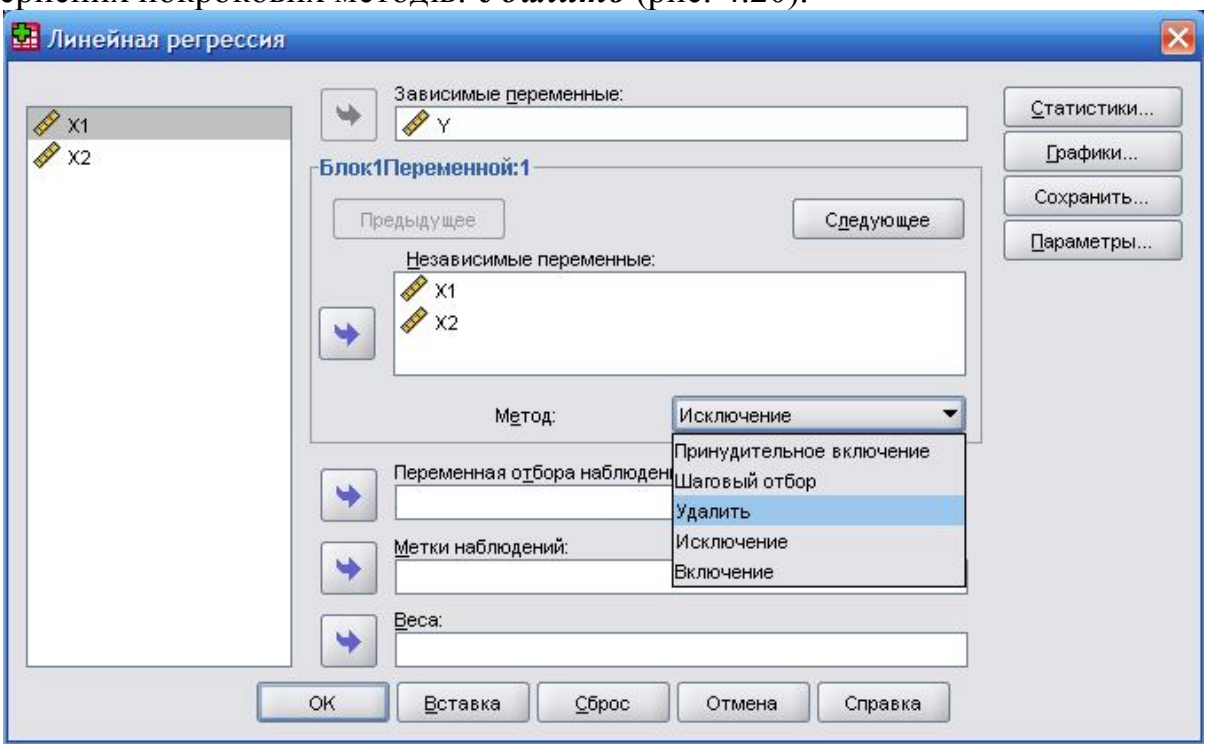


Рисунок 4.20. Діалогове вікно множинної лінійної регресії

3) Проаналізувати інформацію вікна виведення результатів (рис. 4.21) та зробити відповідні висновки.

Дані стовпчика ***R-квадрат*** таблиці ***Сводка для модели*** є коефіцієнтами детермінації, які вказують на степінь відповідності між регресійними моделями і вхідними даними. Коефіцієнт детермінації для моделі 1 дорівнює 0,798. Отже, 80% досліджуваних об'єктів описує перша модель емпіричної лінії регресії.

У стовпчику ***Нестандартизованные коэффициенты: B*** таблиці ***Коэффициенты*** подано значення коефіцієнтів рівняння регресії, а саме:

$$Y_{X_1X_2} = 12{,}51 + 2{,}67X_1 - 0{,}083X_2 .$$

Даними стовпчика ***Стандартизованные коэффициенты Бета*** таблиці ***Коэффициенты*** є регресійні коефіцієнти, які вказують важливість незалежних ознак, використаних в рівнянні лінії регресії. Значення 0,943 показує, наскільки важливі показники середньорічних цін на рекламу (X_1) для визначення вартості реалізованої рекламної продукції (Y).

У таблиці ***Исключенные переменные*** дані стовпчика ***Частная корреляция*** показують тісний зв'язок між незалежними і залежною змінними ($r_{X_1Y} = 0,89$, $r_{X_2Y} = 0,78$).

Сводка для модели

Модель	H	R-квадрат	Скорректированный R-квадрат	Стд. ошибка оценки
1	,893[a]	,798	,741	3,546
2	,000[b]	,000	,000	6,963

a. Предикторы: (конст) X2, X1

b. Предиктор (константа)

Коэффициенты[a]

Модель		Нестандартизованные коэффициенты		Стандартизованные коэффициенты	t	Знч.
		B	Стд. Ошибка	Бета		
1	(Константа)	12,507	3,595		3,479	,010
	X1	2,672	1,027	,943	2,601	,035
	X2	-,083	,525	-,057	-,157	,879
2	(Константа)	29,400	2,202		13,351	,000

a. Зависимая переменная: Y

Исключенные переменные[b]

Модель		Бета включения	t	Знч.	Частная корреляция	Статистики коллинеарности
						Толерантность
2	X1	,893[a]	5,614	,001	,893	1,000
	X2	,777[a]	3,488	,008	,777	1,000

a. Предиктор (константа)

b. Зависимая переменная: Y

Рисунок 4.21. Результати розрахунку коефіцієнтів рівняння множинної регресії

Висновок: на вартість реалізованої рекламної продукції (Y) значно впливають середньорічні ціни на рекламу (X_1) та рівень затрат на проведення реклами (X_2). Залежність від згаданих факторів можна описати наступним рівнянням регресії:

$$Y_{X_1X_2} = 12,51 + 2,67X_1 - 0,083X_2 .$$

Завдання для самостійного виконання

4.1. Відомо дані про обсяг виробництва сільськогосподарської продукції (грн.) на 1 особу АР Крим (табл. 4.11). Побудувати регресійну модель за даними таблиці, оцінити її статистичну значущість та адекватність.

Таблиця 4.11

Рік	2001	2002	2003	2004	2005	2006	2007
Обсяг виробництва с/г продукції	1000	995	949	926	1049	1112	1596

4.2. Відомо дані про чисельність наукових та науково-технічних працівників, що припадають на 1000 осіб (табл. 4.12). Побудувати регресійну модель за даними таблиці, оцінити її статистичну значущість та адекватність.

Таблиця 4.12

Рік	2001	2002	2003	2004	2005	2006	2007
Чисельність наукових та науково-технічних працівників	1,1	1,1	1,0	1,0	1,0	1,0	0,9

4.3. Відомо дані про інвестиції в основний капітал в розрахунку на 1 особу (табл. 4.13). Побудувати регресійну модель за даними таблиці, оцінити її статистичну значущість та адекватність.

Таблиця 4.13

Рік	2001	2002	2003	2004	2005	2006	2007
Інвестиції в основний капітал	376,8	600,2	735,7	955,2	1376,2	1704,1	2375,6

4.4. Відомо дані про обсяг інноваційної продукції в розрахунку на 1 особу (табл. 4.14). Побудувати регресійну модель за даними таблиці, оцінити її статистичну значущість та адекватність.

Таблиця 4.14

Рік	2001	2002	2003	2004	2005	2006	2007
Обсяг інноваційної продукції	38,4	139,0	264,5	172,3	313,1	469,9	282,2

4.5. Відомо дані про обсяг експорту товарів в розрахунку на 1 особу (табл. 4.15). Побудувати регресійну модель за даними таблиці, оцінити її статистичну значущість та адекватність.

Таблиця 4.15

Рік	2001	2002	2003	2004	2005	2006	2007
Обсяг експорту товарів	84,6	107,3	109,2	158,1	137,1	178,1	201,7

4.6 – 4.15. Знайти рівняння ліній регресії, які описують залежність Y від X за даними кореляційних табл. 4.16–4.17, оцінити їх статистичну значущість та адекватність.

Таблиця 4.16

Y	X					
	10	20	30	40	50	60
5	*a*	*b*				
10		*c*	*d*			
15			*e*	*f*	*g*	
20			*h*	*k*	*m*	
25				*n*	*p*	*q*

Таблиця 4.17

	4.6	4.7	4.8	4.9	4.10	4.11	4.12	4.13	4.14	4.15
a	2	2	4	1	4	2	2	2	3	4
b	3	6	2	5	2	4	4	4	3	2
c	7	4	6	5	6	3	6	6	5	5
d	3	4	4	3	2	7	2	3	4	3
e	2	7	6	9	5	5	3	6	20	5
f	50	35	45	40	40	30	50	45	22	45
g	2	8	2	2	5	10	2	4	8	5
h	1	2	2	4	2	7	1	2	5	2
k	10	10	8	11	8	10	10	8	10	8
m	6	8	6	6	7	8	6	6	6	7
n	4	5	4	4	4	5	4	4	4	4
p	7	6	7	7	7	6	7	7	7	7
q	3	3	4	3	8	3	3	3	3	3

Питання для самоконтролю

1. Що називається регресійним аналізом?
2. Що називається регресійною моделлю?
3. Що називається факторними ознаками? Результативною ознакою?
4. Які види регресійних моделей Ви знаєте?
5. Як повинні бути задані вхідні дані для регресійного аналізу?
6. Як сформулювати гіпотезу про вид регресійної моделі?
7. Що таке теоретичні та емпіричні значення результативної ознаки?
8. Що називається емпіричною лінією регресії? Теоретичною лінією регресії?
9. Як побудувати емпіричну лінію регресії? Теоретичну лінію регресії?
10. Як знайти параметри регресійної моделі?
11. Як будується розрахункова таблиця у регресійному аналізі?
12. Як перевірити правильність побудованої регресійної моделі?
13. Як перевірити адекватність побудованої регресійної моделі вхідним даним?
14. Що називається коефіцієнтом детермінації і як він використовується у статистичному моделюванні?
15. Як перевірити статистичну значущість побудованої регресійної моделі?
16. Що називається багатофакторною лінійною регресією?
17. Які етапи побудови багатофакторної лінійної регресійної моделі?
18. Як обґрунтовується вибір факторних ознак для побудови моделі?
19. Що називається кореляційними плеядами?
20. Як оцінюється вплив факторних ознак на результативну?
21. Як здійснюється прогноз за багатофакторною лінійною регресійною моделлю?
22. Як визначити вид нелінійної регресійної моделі?

23. Як перевірити статистичну значущість нелінійної регресійної моделі?
24. Як знайти критичне значення критерія Фішера?
25. Як здійснюється прогноз за нелінійною регресійною моделлю?
26. Для чого перевіряється статистична значущість регресійної моделі?
27. Чому при наявному нелінійному зв’язку будують декілька регресійних моделей?
28. Як з декількох регресійних моделей обрати найбільш адекватну?
29. Як побудувати порівняльну діаграму у регресійному аналізі?
30. Як побудувати регресійну модель засобами MS Excel? SPSS?

Розділ 5. Проблемні питання прикладних досліджень

5.1. Формування вибірки

5.1.1. Основні методи формування вибірки

Формування вибірки – це базовий етап будь-якого прикладного дослідження (основні вимоги до вибірки викладено у п. 1.1.). При формуванні вибірки досліднику необхідно визначити:

− хто (що) є елементом або ***одиницею вибірки*** виходячи від сутності дослідження;

− ***контур вибірки***, тобто список усіх одиниць генеральної сукупності, з якої формується вибірка;

− ***об'єм вибірки*** − кількість елементів у ній.

Наприклад, якщо фірма – виробник мобільних телефонів бажає вивчити потенціальний ринок своєї продукції, то одиницями вибірки будуть особи, які приймають рішення про вибір комунікаційного обладнання: керівники організацій або голови родин. Контуром вибірки можуть бути списки: організацій, фірм, домовласників і т.п.

Оскільки вибірка є лише частиною генеральної сукупності, то отримані на основі її вивчення результати не будуть точно відповідати результатам, які можна було б отримати при вивченні всієї генеральної сукупності. Різниця між результатами дослідження вибірки та генеральної сукупності називається ***помилкою вибірки***. Помилки вибірки обумовлюються як методами її формування, так і її об'ємом.

Ймовірнісні методи формування вибірки

При формуванні вибірки часто використовують ймовірнісні методи, при яких всі елементи генеральної сукупності мають однакову ймовірність бути включеними в вибірку. Такими методами є простий випадковий відбір, систематичний відбір, кластерний відбір та стратифікований відбір.

При ***простому випадковому відборі*** передбачається, що для всіх елементів генеральної сукупності ймовірність бути обраним в вибірку відома і однакова. Така ймовірність визначається як відношення об'єму вибірки до об'єму генеральної сукупності.

Простий випадковий відбір може здійснюватись за допомогою таблиць випадкових чисел або генератора випадкових чисел MS Excel. При цьому: кожному елементу генеральної сукупності присвоюється порядковий номер; генерується необхідна кількість випадкових чисел; обираються ті елементи генеральної сукупності, номери яких співпадають з випадковими числами. Так, наприклад, при проведенні телефонного інтерв'ю комп'ютер випадково генерує телефонні номери.

Недоліком цього методу є необхідність визначення кожного елемента генеральної сукупності, що часто просто неможливо.

При ***систематичному відборі*** необхідно: визначити всі елементи генеральної сукупності; визначити інтервал відбору – відношення об'єму генеральної сукупності до об'єму вибірки; відібрати в вибірку елементи генеральної сукупності з урахуванням інтервалу.

Наприклад, якщо як генеральна сукупність використовується телефонний довідник, в якому 5500 номерів, а необхідний об'єм вибірки дорівнює 500, то інтервал відбору визначається як $5500 / 500 = 11$. Це означає, що кожний одинадцятий телефонний номер потрапляє у вибірку.

Кластерний відбір базується на розподілі елементів генеральної сукупності на групи (кластери), кожна з яких репрезентативна всій сукупності. У подальшому випадково обирається один із кластерів, елементи якого вважаються генеральною сукупністю. На ньому проводиться певне дослідження, результати якого розповсюджуються на всі інші кластери і на всю генеральну сукупність.

Наприклад, якщо планується дослідження думки населення певної області, то кластерами можуть вважатися райони цієї області. При цьому дослідник припускає, що населення кожного району ідентичне населенню області у цілому (що не завжди правильно).

У випадку, коли дослідження проводиться на певній території, можна розбити цю територію на кластери за допомогою вибіркової решітки, яка накладається на карту території і визначає кластери.

Недоліком усіх описаних методів є необхідність припущення про симетричний розподіл характеристик генеральної сукупності, що практично зустрічається дуже рідко. Наприклад, населення різних районів Одеської області значно відрізняється за рівнем доходів, основними видами діяльності і т. ін., тому її райони не можуть вважатися ідентичними як один одному, так і всій області в цілому.

Якщо розподіл характеристик генеральної сукупності не є симетричним, використовується ***стратифікований відбір***. В цьому випадку генеральна сукупність розподіляється на групи (страти) залежно від певної характеристики. Наприклад, населення області розподіляється на групи залежно від рівня доходу. Кожна страта виконує роль генеральної сукупності, з якої формується вибірка.

При цьому необхідно враховувати пропорційність стратифікування. Якщо об'єм вибірки для певної страти пропорційний розміру страти у відношенні до всієї генеральної сукупності, то вибірка називається пропорційно стратифікованою. Якщо ця умова не виконується, то слід застосовувати вагові коефіцієнти, які урівноважать розміри страт.

Неймовірнісні методи формування вибірки

Використання ймовірнісних методів формування вибірки є найбільш правильним з точки зору статистики, але вони трудомісткі і, відповідно, матеріально невигідні. Тому в прикладних дослідженнях часто використовують неймовірнісні методи: відбір за критерієм, відбір згідно спостереження, відбір за квотами та відбір за принципом зручності.

Метод ***відбору за принципом зручності*** полягає в тому, що формування вибірки здійснюється самим зручним з боку дослідника способом за мінімум часу та грошових витрат. Наприклад, опитування покупців проводиться в певному магазині. При цьому вибірка часто не є репрезентативною, але існують способи оцінки її похибки.

При ***формуванні вибірки за критеріями*** використовується думка спеціалістів-експертів щодо її складу. Так часто формуються фокус-групи.

Метод ***відбору за спостереженням*** використовується тоді, коли контури вибірки дуже обмежені. Наприклад, при проведенні дослідження попиту на продукцію, що не має широкого кола споживачів.

За цим методом дослідник формує початкову вибірку, об'єм якої значно менший необхідного. У процесі дослідження вибірка розширюється за рахунок пропозицій і рекомендацій респондентів, які вже прийняли участь в опитуванні.

Метод ***відбору за квотами*** використовується за умови наявності чітко визначених характеристик респондентів, думку яких доцільно вивчити при дослідженні. У цьому випадку в вибірку відбирають ті елементи генеральної сукупності, що мають певні характеристики. Об'єм вибірки (квота) також визначається на початку дослідження. Відбір елементів закінчується при заповненні квоти.

Зауваження. При формуванні вибірки дослідник повинен знайти баланс між затратами на збір даних і об'ємом вибірки. Методи формування вибірки повинні відповідати цілям дослідження.

5.1.2. Визначення об'єму вибірки

В практичних дослідженнях використовується кілька методів розрахунку об'єму вибірки. Але незалежно від результатів розрахунків слід пам'ятати, що:

− об'єм вибірки залежить передусім від вартості дослідження;

− точність отриманих результатів залежить не стільки від об'єму вибірки, скільки від методу її формування.

Об'єм вибірки може бути просто процентом від об'єму генеральної сукупності. Наприклад, припускається, що для отримання досить точних результатів достатньо 5% елементів генеральної сукупності. Однак перевірити таке припущення неможливо, тому рівень точності результатів невідомий.

Якщо певне дослідження проводиться регулярно, то доцільно використовувати вибірки однакового об'єму. Крім того, якщо відомо, що дослідження певного типу вже проводилося, можна обрати той же об'єм вибірки. Але при цьому не враховуються можливі зміни умов проведення дослідження.

При проведенні соціологічних опитувань вважається, що для пробних (пілотних) опитувань достатня вибірка об'ємом 100 – 250 осіб. При масових опитуваннях (генеральна сукупність становить 5000 осіб і більше) об'єм вибіркової сукупності повинен становити 10% генеральної сукупності, але не більше 2 – 2,5 тис. осіб. Це гарантує достатньо достовірні результати дослідження. Помилки вибірки, які інколи при цьому трапляються, бувають наслідком невірних вихідних статистичних даних про параметри контрольних ознак генеральної сукупності; недостатнього об'єму вибіркової сукупності, неправильного застосування способу відбору одиниць аналізу (наприклад, відбір із неправильно складеного списку, невдалий вибір місця, часу проведення опитування тощо).

Найбільш коректним є статистичний метод розрахунку об'єму вибірки, заснований на визначенні мінімально необхідного об'єму вибірки залежно від вимог до точності результатів дослідження.

У статистичному методі використовуються такі позначення:

n – об'єм вибірки;

N – об'єм генеральної сукупності;

z − нормативне відхилення оцінки. Обирається залежно від рівня значущості (табл. 5.1). Зазвичай потрібен рівень значущості $\alpha = 0,05$, тоді z=1,96;

S^2 – дисперсія вибірки;

S – середнє квадратичне відхилення вибірки;

p – варіація вибірки. Наприклад, якщо відомо, що 70% населення не вживають йогурт, то варіація вибірки p = 70. Якщо варіація невідома, то приймається p = 50;

$q = 100 - p$;

e – допустима помилка, яка обирається дослідником.

Таблиця 5.1

α	0,4	0,3	0,2	0,15	0,1	0,05	0,03	0,01	0,003
z	0,84	1,03	1,29	1,44	1,65	1,96	2,18	2,58	3,0

Зауваження. Для визначення варіації певної генеральної сукупності доцільно провести попереднє (пілотне) дослідження. Крім того, слід враховувати, що максимальною є варіація p = 50. Тобто така варіація відповідає найгіршому випадку.

Наведемо розрахункові формули для визначення об'єму вибірки у випадку, коли вибірка становить менше 5% від генеральної сукупності і вважається великою:

$$n = \frac{z^2 pq}{e^2} \ . \tag{5.1}$$

Формула (5.1) використовується, якщо потрібно розрахувати кількість респондентів соціологічного опитування за умов, що на питання існує два варіанти відповіді «Так» або «Ні».

$$n = \frac{z^2 S^2}{e^2} \ . \tag{5.2}$$

Формула (5.2) використовується, якщо з попередніх досліджень відома дисперсія або середнє квадратичне відхилення.

Якщо такі дослідження не проводилися і вибірка формується вперше, то використовується формула, в якій помилка e пов'язана із середнім квадратичним відхиленням:

$$n = \frac{z^2}{e_1^{\,2}}, \text{ де } \ e_1^2 = \frac{e}{S} \ . \tag{5.3}$$

Якщо об'єм вибірки більше 5% генеральної сукупності, то вибірка вважається маленькою і в формули 5.1 – 5.3 вводиться уточнюючий коефіцієнт:

$$k = \sqrt{\frac{N-n}{N-1}} \ . \qquad (5.4)$$

Тоді об'єм вибірки n_k розраховується за формулою:

$$n_k = n \cdot k = n \cdot \sqrt{\frac{N-n}{N-1}} \ . \qquad (5.5)$$

Приклад 5.1. Необхідно визначити кількість респондентів для опитування при дослідженні ринку копченої риби.

Розв'язок. Якщо кількість споживачів невідома, приймаємо варіацію вибірки $p = 50$. Рівень значущості $\alpha = 0,05$, тоді z=1,96, припустима помилка 4%. Тоді за формулою 5.1 маємо:

$$n = \frac{z^2 pq}{e^2} = \frac{1,96^2 \cdot 50 \cdot 50}{4^2} = 600 \text{ осіб.}$$

Якщо потрібна менша помилка, наприклад 3%, то за тією ж формулою маємо:

$$n = \frac{z^2 pq}{e^2} = \frac{1,96^2 \cdot 50 \cdot 50}{3^2} = 1067 \text{ осіб.}$$

Якщо проводилося попереднє дослідження і відомо, що 70% респондентів є споживачами, то об'єм вибірки дорівнює:

$$n = \frac{z^2 pq}{e^2} = \frac{1,96^2 \cdot 70 \cdot 30}{4^2} = 504 \text{ особи.}$$

Приклад 5.2. Проводиться дослідження якості послуг, які надаються перукарнями міста. Необхідно визначити кількість респондентів для опитування, якщо відомо, що у місті 500 перукарень.

Розв'язок. Приймаємо варіацію вибірки $p = 50$. Рівень значущості $\alpha = 0,05$, тоді z=1,96, припустима помилка 10%. Тоді за формулою 5.1 маємо:

$$n = \frac{z^2 pq}{e^2} = \frac{1,96^2 \cdot 50 \cdot 50}{10^2} = 96 \text{ перукарень.}$$

У даному випадку об'єм вибірки становить 19% від генеральної сукупності (загальна кількість перукарень – 500 од.), тобто перевищує 5%. Тому розрахуємо об'єм вибірки з урахування уточнюючого коефіцієнта за формулою 5.5:

$$n_k = n \cdot k = 96 \cdot \sqrt{\frac{500-96}{500-1}} = 86 \text{ перукарень.}$$

5.2. Обробка результатів експертного оцінювання

5.2.1. Коефіцієнт конкордації

Дуже важливим етапом у підведенні результатів дослідження є прогнозування. Часто прогнозування передбачає визначення значень економічних або соціологічних показників у майбутньому. Наприклад, прогнозування думки респондентів щодо рекламованого товару, ціни на товари, об'єм їх продаж.

Прогнозування є початковим етапом планування і включає в себе попередній і кінцевий (формальний) прогнози, для яких розробляється один або декілька сценаріїв майбутніх подій.

Методи експертних оцінок використовуються для прогнозування майбутніх подій, якщо відсутні статистичні дані або їх недостатньо. Вони також застосовуються для кількісного вимірювання таких подій, для яких не існує інших способів вимірювання.

Припускається, що експерт формує своє судження на аналізі групи факторів, оцінюючи ймовірності їх реалізації та впливу на результативну ознаку об'єкта вивчення. Але, при цьому отримані висновки та оцінки пов'язані з особистістю експерта, тому інший експерт, використовуючи ту саму інформацію, може дійти інших висновків. Тому вважається, що при розв'язанні проблем в умовах невизначеності, думка групи експертів дає більш надійні результати, ніж думка одного експерта.

Після отримання експертних оцінок проводиться їх обробка та оцінюється достовірність. Обробку результатів експертного оцінювання можна проводити за ***коефіцієнтом конкордації***, який показує степінь згоди думок експертів. Найбільш достовірні оцінки отримуються за умов узгодженості думок експертів.

Коефіцієнт конкордації W розраховується за формулою:

$$W = \frac{12}{m^2\left(n^3 - n\right)} \sum_{j=1}^{n} \left(\sum_{i=1}^{m} \left(R_{ij} - \frac{n+1}{2} \right) \right)^2 , \qquad (5.6)$$

де n – кількість об'єктів оцінювання;

m – кількість експертів;

R_{ij} – ранг j-го об'єкта, представленого i-м експертом.

Якщо об'єкти оцінювання мають однакові ранги, то коефіцієнт конкордації розраховується за формулою:

$$W = \frac{12}{\frac{1}{12} m^2\left(n^3 - n\right) - m\sum_{j=1}^{m} T_j} \cdot \sum_{j=1}^{n} \left(\sum_{i=1}^{m} \left(R_{ij} - \frac{n+1}{2} \right) \right)^2 . \qquad (5.7)$$

При цьому T_j обчислюється за формулою:

$$T_j = \frac{\sum_{i=1}^{L_i} \left(n_i^2 - n_i \right)}{12}, \qquad (5.8)$$

де L_i − кількість груп однакових рангів,

n_i − кількість елементів *i*-тої групи для *j*-го експерта.

Статистична значущість коефіцієнта конкордації перевіряється порівнянням величини $n(m-1)\cdot W$ з табличним значенням розподілу χ^2 при рівні значущості $\alpha = 0,001$ та $n-1$ степенях свободи.

Якщо коефіцієнт конкордації виявляється не значущим, то використовується методика виведення експерта, думка якого не узгоджується з думкою інших експертів. Для цього будується матриця коефіцієнтів кореляції Пірсона ($r(k,i)$) або рангових коефіцієнтів кореляції Спірмена ($r_S(k,i)$) та виявляється експерт, оцінка якого підкоряється умові:

$$r_j(k,i) = \min_{i=1,\dots,m}\{r(k,i)\}, \qquad (5.9)$$

що означає, що думка цього експерта найменше узгоджується з думкою інших експертів. Бали, подані таким експертом, у подальших розрахунках не враховуються.

Алгоритм повторюється, поки коефіцієнт конкордації не стає значущим.

Приклад 5.3. Група експертів з 3 осіб оцінювала час, що необхідний для виконання робіт певного проекту. Результати оцінювання подано у табл. 5.2. Перевірити степінь узгодженості думок експертів.

Таблиця 5.2

Експерти	Час, необхідний для робіт			
	Робота 1	Робота 2	Робота 3	Робота 4
1-й	6	5	2	4
2-й	4	7	3	9
3-й	5	7	3	6

Розв'язок. Здійснимо перевірку за коефіцієнтом конкордації, для чого знайдемо ранги робіт проекту окремо за оцінками кожного з експертів (табл. 5.3).

Таблиця 5.3

Експерти	Ранги робіт			
	Робота 1	Робота 2	Робота 3	Робота 4
1-й	6	5	2	4
2-й	4	7	3	9
3-й	5	7	3	6

У групах рангів оцінок, наданих окремими експертами, немає однакових, тому коефіцієнт конкордації розраховуємо за формулою 5.6.

Обчислимо величини, що не залежать від індексів сум, враховуючи, що: *n* – кількість робіт, *n* = 4; *m* – кількість експертів, *m* = 3.

Отримаємо:

$$\frac{n+1}{2} = \frac{4+1}{2} = 2,5; \quad \frac{12}{m^2(n^3-n)} = \frac{12}{3^2(4^3-4)} = \frac{12}{540} \approx 0,022.$$

Подальші обчислення для зручності представимо у вигляді таблиці (табл. 5.4).

Таблиця 5.4

Розрахункові формули	Результати розрахунків			
	Робота 1	Робота 2	Робота 3	Робота 4
$R_{ij} - \frac{n+1}{2}$	1,5	0,5	−1,5	−0,5
	−0,5	0,5	−1,5	1,5
	−0,5	1,5	−1,5	0,5
$\sum_{i=1}^{m}\left(R_{ij} - \frac{n+1}{2}\right)$	0,5	2,5	−4,5	1,5
$\left(\sum_{i=1}^{m}\left(R_{ij} - \frac{n+1}{2}\right)\right)^2$	0,25	6,25	20,25	2,25
$\sum_{j=1}^{n}\left(\sum_{i=1}^{m}\left(R_{ij} - \frac{n+1}{2}\right)\right)^2$	29			

Отже, коефіцієнт конкордації:

$$W = \frac{12}{m^2\left(n^3 - n\right)}\sum_{j=1}^{n}\left(\sum_{i=1}^{m}\left(R_{ij} - \frac{n+1}{2}\right)\right)^2 = 0,022 \cdot 29 = 0,638$$

Перевіримо його значущість: $n\left(m-1\right)\cdot W = 4(3-1)\cdot 0,638 = 5,104;$

критичне значення χ^2: ХИ2ОБР (0,001; 4 − 1) = 16,27. Оскільки величина $n\left(m-1\right)\cdot W$ менша критичного значення χ^2, то коефіцієнт конкордації не є значущим та думки експертів не узгоджені.

Виокремимо експерта, оцінки якого є найбільш неузгодженими. Для цього побудуємо матрицю парних коефіцієнтів кореляції Пірсона (табл. 5.5).

Таблиця 5.5

Експерти	1-й	2-й	3-й
1-й	1		
2-й	**0,23035**	1	
3-й	0,657143	0,797366	1

З таблиці 5.5 видно, що найменшим є значення коефіцієнта кореляції, який показує узгодженість думок першого та другого експертів, тому одного з них необхідно вивести з експертизи. Доцільно вивести першого експерта, тому що його оцінки є менш узгодженими з оцінками третього експерта.

Розрахуємо коефіцієнт конкордації (табл. 5.6), враховуючи відсутність оцінок першого експерта.

Отже, $\frac{n+1}{2} = \frac{4+1}{2} = 2,5$; $\frac{12}{m^2\left(n^3-n\right)} = \frac{12}{2^2\left(4^3-4\right)} = \frac{12}{240} = 0,05$

Таблиця 5.6

Розрахункові формули	Результати розрахунків			
	Робота 1	Робота 2	Робота 3	Робота 4
$R_{ij} - \frac{n+1}{2}$	−0,5	0,5	−1,5	1,5
	−0,5	1,5	−1,5	0,5
$\sum_{i=1}^{m}\left(R_{ij} - \frac{n+1}{2}\right)$	−1	2	−3	2
$\left(\sum_{i=1}^{m}\left(R_{ij} - \frac{n+1}{2}\right)\right)^2$	1	4	9	4
$\sum_{j=1}^{n}\left(\sum_{i=1}^{m}\left(R_{ij} - \frac{n+1}{2}\right)\right)^2$	18			

Коефіцієнт конкордації:

$$W = \frac{12}{m^2\left(n^3 - n\right)}\sum_{j=1}^{n}\left(\sum_{i=1}^{m}\left(R_{ij} - \frac{n+1}{2}\right)\right)^2 = 0,05 \cdot 18 = 0,9$$

Отже, значення коефіцієнта конкордації після виведення першого експерта значно збільшилося. Але воно теж не є значущим. Однак це пов'язано не з тим, що думки експертів не погоджуються, а з тим, що кількість експертів надто мала.

Час, необхідний для виконання робіт проекту, розраховується як середнє арифметичне експертних оцінок.

5.2.2. Коефіцієнт компетенції

Використання коефіцієнта конкордації засновано на припущенні, що чим більш узгоджені думки експертів, тим достовірнішими є їх оцінки. Але практика показує, що це не завжди вірно, і експерт, який не згоден з думками більшості, може дати найточніші оцінки.

Якщо дослідник бажає врахувати думки всіх експертів, то обробку результатів експертного оцінювання слід виконувати за ***коефіцієнтом компетентності*** експерта.

Цей метод базується на використанні попередньої оцінки компетентності експертів, які приймають участь у дослідженні. Оцінка експертів проводиться за критеріями компетентності, серед яких можуть бути: рівень освіти; загальний стаж роботи; стаж роботи за проблемою дослідження; посада тощо. Крім того, важливим критерієм є оцінка надійності експерта, яка розраховується як відношення його правильних оцінок до всіх проведених експертиз. Правильними вважаються ті оцінки, які з часом підтвердилися практикою.

При розрахунку коефіцієнтів компетентності експертів необхідно використовувати єдину для всіх критеріїв шкалу оцінювання. У противному випадку оцінки потрібно буде нормалізувати, тобто привести до однієї шкали.

Коефіцієнт компетентності розраховується за формулою:

$$KK_i = \frac{\sum_{j=1}^{m} k_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m} k_{ij}}, \qquad (5.9)$$

де n – кількість експертів;

m – кількість критеріїв оцінювання експертів;

k_{ij} – бал, отриманий i-м експертом за j-м критерієм.

Приклад 5.4. За вхідними даними прикладу 5.3. знайти час, необхідний для виконання робіт проекту, з урахуванням коефіцієнта компетентності експертів. Бали, отримані експертами, подано у табл. 5.7. Оцінювання проводилося за трьохбальною шкалою.

Таблиця 5.7

Експерти	Бали, отримані експертами		
	Критерій 1. Стаж роботи	Критерій 2. Професіоналіз	Критерій 3. Надійність
1-й	1	2	2
2-й	2	3	3
3-й	2	1	1

Розв'язок. Знайдемо коефіцієнти компетентності експертів за формулою 5.9. Необхідні розрахунки внесемо у табл. 5.8.

Таблиця 5.8

Експерти	Бали, отримані експертами			Сума балів кожного експерта
	Критерій 1	Критерій 2	Критерій 3	
1-й	1	2	2	5
2-й	2	3	3	8
3-й	2	1	1	4
Загальна сума балів				17

Отже:

для першого експерта $KK_1 = \frac{\sum_{j=1}^{m} k_{1j}}{\sum_{i=1}^{n}\sum_{j=1}^{m} k_{ij}} = \frac{5}{17} \approx 0,2941;$

аналогічно для другого і третього $KK_2 = \frac{8}{17} \approx 0,4706; \quad KK_3 = \frac{4}{17} = 0,2353.$

Розрахуємо час, необхідний для виконання робіт проекту, з урахуванням коефіцієнта компетентності експертів за формулою: $t_j = \sum_{i=1}^{n} KK_i \cdot t_{ij};\ j = \overline{1,m};$ де t_i – час для i-тої роботи; t_{ij} – оцінка часу i-тої роботи j-м експертом. Результати представлено у табл. 5.9.

Таблиця 5.9

Експерти	$КК_i$	Час, необхідний для робіт			
		Робота 1	Робота 2	Робота 3	Робота 4
1-й	0,2941	6	5	2	4
2-й	0,4706	4	7	3	9
3-й	0,2353	5	7	3	6
Час з урахуванням $КК_i$		**4,82**	**6,41**	**2,71**	**6,82**

5.3. Аналіз часових рядів із сезонною варіацією

Множина даних, отриманих у результаті спостережень, що проводилися регулярно через рівні інтервали часу, називається ***часовим рядом***. У часових рядах час є факторною ознакою. Зміна в часі результативної ознаки називається ***трендом***.

В процесі господарської діяльності окремі галузі промисловості, торгівля і сфера послуг стикаються з циклічними коливаннями, які викликані сезонним характером виробництва та споживання товарів і послуг. Повторення даних через певний проміжок часу називається ***сезонною варіацією***.

Для аналізу тенденції зміни результативної ознаки на основі часового ряду сезонну варіацію даних необхідно виключити (провести десезоналізацію даних). Після цього за допомогою моделі лінійної регресії можна знайти рівняння тренда.

За допомогою рівняння тренда розробляються прогнози на наступні часові періоди. Кожен прогноз містить похибки, які бувають систематичними і випадковими. ***Систематичні похибки*** виникають внаслідок невірної моделі тренда, порушення сезонної варіації у неналежний бік і т. ін. ***Випадкові похибки*** – ті, що не можна пояснити моделлю тренда. Похибки обчислюються за рівнянням тренда і фактичними даними за формулами:

середнє абсолютне відхилення $$MAD = \sum_{i=1}^{n} \frac{|e_i|}{n}; \qquad (5.10)$$

середньоквадратична похибка $$MSE = \sum_{i=1}^{n} \frac{e_i^2}{n}, \qquad (5.11)$$

де e_i – різниця фактичного і трендового значень.

Як трендові найчастіше використовуються: адитивна модель, мультиплікативна модель та модель експоненційного згладжування.

Для адитивної моделі:

фактичне значення A = трендове значення T + сезонна варіація S + похибка E

Для мультиплікативної моделі:

фактичне значення A = трендове значення T * сезонна варіація S * похибка E

Для моделі експоненційного згладжування:

новий прогноз = $\alpha \cdot$фактичний результат в останній період + + $(1-\alpha) \cdot$прогноз в останній період

Константу згладжування α вибирають з відрізка [0; 1]. В умовах стабільності α належить відрізку [0,2; 0,4], при швидкій зміні результативної ознаки − α вибирають з відрізка [0,7; 0,9].

Побудову та аналіз моделей розглянемо на прикладі.

Приклад 5.5. В табл. 5.12 вказаний об'єм продажу (тис. грн.) за 11 кварталів. На основі цих даних зробити прогноз на наступні 2 квартали.

Таблиця 5.12

Квартал	1	2	3	4	5	6	7	8	9	10	11
Об'єм продажу	4	6	4	5	10	8	7	9	12	14	15

Розв'язок. Побудуємо адитивну модель за формулою:

фактичне значення A = трендове значення T + сезонна варіація S + похибка E

1. Перш за все виключимо вплив сезонної варіації. Користуємось методом ковзкого середнього. Для зручності обчислень заповнимо табл. 5.13.

Таблиця 5.13

Квартал	Об'єми продажу	Ковзке середнє за 4 квартали	Центроване ковзке середнє	Оцінка сезонної варіації
1	4	–	–	–
2	6	–	–	–
3	4	4,75	5,5	−1,5
4	5	6,25	6,5	−1,5
5	10	6,75	7,125	2,875
6	8	7,5	8	0
7	7	8,5	8,75	−1,75
8	9	9	9,75	−0,75
9	12	10,5	11,5	0,5
10	14	12,5	–	–
11	15	–	–	–

Пояснення до таблиці: 1 рік = 4 квартали, тому знайдемо середні значення об'єму продажу за 4 послідовних квартали. Тобто додаємо 4 послідовних числа із 2-го стовпчика і ділимо їх на 4. Результат записуємо у 3-й стовпчик – навпроти третього доданка). Якщо ковзке середнє обчислюється для непарної кількості періодів часу, то результат записується напроти середнього доданка і отримані середні не потрібно центрувати. У нашому випадку середні значення необхідно центрувати, для чого обчислюємо середнє арифметичне двох сусідніх чисел 3-го стовпчика і записуємо у 4-й. Оцінка сезонної варіації (5-й стовпчик) – це різниця фактичних даних і відповідних центрованих ковзких середніх (2-й стовпчик – 4-й стовпчик).

2. Знайдемо сезонну варіацію для кожного з чотирьох кварталів року. Для зручності результати обчислень оформимо у вигляді табл. 5.14.

Оцінки сезонної варіації запишемо у стовпчиках під відповідним номером кварталу. У кожному стовпчику обчислюємо середнє значення. Обчислюємо суму середніх значень (у даному прикладі вона дорівнює −1). Значення сезонної варіації повинні бути скоректовані так, щоб сума середніх значень дорівнювала 0 (середня варіація за рік). Для цього знаходимо коректувальний

коефіцієнт: суму середніх ділимо на 4 (число кварталів у році) та віднімаємо від усіх середніх даний коефіцієнт.

Таблиця 5.14

	Номер кварталу				
	1	2	3	4	
Оцінки сезонної варіації	–	–	−1,5	−1,5	
	2,875	0	−1,75	−0,75	
	0,5	-	–	–	Сума
Середнє	1,7	0,0	−1,6	−1,1	−1
Скоректована сезонна варіація	2,0	0,2	−1,3	−0,9	0,0

3. Виключимо сезонну варіацію із фактичних даних (табл. 5.15).

Таблиця 5.15

Квартал	Об'єми продажу A	Сезонна варіація S	Десезонолізований об'єм продажу $A-S=T+E$
1	4	2	2
2	6	0,2	5,8
3	4	−1,3	5,3
4	5	−0,9	5,9
5	10	2	8
6	8	0,2	7,8
7	7	−1,3	8,3
8	9	−0,9	9,9
9	12	2	10
10	14	0,2	13,8
11	15	−1,3	16,3

Із фактичних даних (2-й стовпчик) віднімаємо сезонну варіацію (3-й стовпчик) і записуємо результат в 4-й стовпчик.

4. Знайдемо рівняння тренда у вигляді лінійної регресійної моделі: $T = ax + в$. Для знаходження коефіцієнтів скористуємося статистичними функціями ОТРЕЗОК та НАКЛОН майстра функцій з пакету Excel. Отримаємо: $a = 1,1; в = 1,9$.

Отже, трендові значення об'єму продажу =1,9 + 1,1 * номер кварталу.

5. Розрахуємо похибки обчислень. Для зручності результати обчислень оформимо у вигляді табл. 5.15.

Таблиця 5.15

Квартал	Об'єми продажу A	Десезонолізований об'єм продажу $A - S = T + E$	Трендові значення	Похибка e_t	$\lvert e_t \rvert$	e_t^2
1	2	3	4	5	6	7
1	4	2	3	−1	1	1
2	6	5,8	4,1	1,7	1,7	2,89
3	4	5,3	5,2	0,1	0,1	0,01
4	5	5,9	6,3	−0,4	0,4	0,16
5	10	8	7,4	0,6	0,6	0,36

Продовження таблиці 5.15

1	2	3	4	5	6	7
6	8	7,8	8,5	−0,7	0,7	0,49
7	7	8,3	9,6	−1,3	1,3	1,69
8	9	9,9	10,7	−0,8	0,8	0,64
9	12	10	11,8	−1,8	1,8	3,24
10	14	13,8	12,9	0,9	0,9	0,81
11	15	16,3	14	2,3	2,3	5,29
				Сума	11,6	16,58

$$MAD = \sum_{i=1}^{n} \frac{|e_t|}{n} = \frac{11,6}{11} \approx 1,1; \quad MSE = \sum_{i=1}^{n} \frac{e_t^2}{n} = \frac{16,58}{11} \approx 1,5.$$

6. Зробимо прогноз об'ємів продажу на наступні два квартали:
Прогноз об'ємів продажу в 12 кварталі: (1,9+1,1*12)+(−0,9)=14,2 тис. грн.
Прогноз об'ємів продажу в 13 кварталі: (1,9+1,1*13)+2=18,2 тис. грн.

Приклад 5.6. В табл. 5.16 вказаний об'єм продажу (тис. грн.) за 11 кварталів. На основі цих даних зробити прогноз на наступні 2 квартали.

Таблиця 5.16

Квартал	1	2	3	4	5	6	7	8	9	10	11
Об'єм продажу	63	74	79	120	67	79	88	130	69	82	90

Розв'язок. Побудуємо мультиплікативну модель за формулою:

фактичне значення A = трендове значення T * сезонна варіація S * похибка E

1. Перш за все виключимо вплив сезонної варіації. Користуємось методом ковзкого середнього. Для зручності обчислень заповнимо табл. 5.17.

Пояснення до таблиці: центроване ковзке середнє обчислюється аналогічно як і для адитивної моделі. Оцінка сезонної варіації (5-й стовпчик) – це частка від ділення фактичних даних на відповідні центровані ковзкі середні (2-й стовпчик / 4-й стовпчик).

Таблиця 5.17

Квартал	Об'єми продажу	Ковзке середнє за 4 квартали	Центроване ковзке середнє	Оцінка сезонної варіації
1	63	–	–	–
2	74	–	–	–
3	79	84	84,5	0,935
4	120	85	85,625	1,401
5	67	86,25	87,375	0,767
6	79	88,5	89,75	0,880
7	88	91	91,25	0,964
8	130	91,5	91,875	1,415
9	69	92,25	92,5	0,746
10	82	92,75	–	–
11	90	–	–	–

2. Знайдемо сезонну варіацію для кожного з чотирьох кварталів року. Для зручності результати обчислень оформимо у вигляді табл. 5.18.

Таблиця 5.18

	Номер квартала				
	1	2	3	4	
Оцінки сезонної варіації	–	–	0,935	1,401	
	0,767	0,880	0,964	1,415	
	0,746	–	–	–	Сума
Середнє	0,756	0,880	0,950	1,408	3,994
Скоректована сезонна варіація	0,757	0,881	0,952	1,410	4,0

Оцінки сезонної варіації запишемо у стовпчику з відповідним номером квартала. У кожному стовпчику обчислюємо середнє значення. Обчислюємо суму середніх значень (у даному прикладі вона дорівнює 3,994). Значення сезонної варіації повинні бути скоректовані так, щоб сума середніх значень дорівнювала 4 (4 частки від чотирьох кварталів – середня варіація за рік). Для цього знаходимо коректувальний коефіцієнт: 4 ділимо на суму середніх та множимо усі середні на цей коефіцієнт.

3. Виключимо сезонну варіацію із фактичних даних (табл. 5.19).

Таблиця 5.19

Квартал	Об'єми продажу *A*	Сезонна варіація *S*	Десезонолізований об'єм продажу *A / S = T * E*
1	63	0,757	83,176
2	74	0,881	83,953
3	79	0,951	83,074
4	120	1,410	85,096
5	67	0,757	88,457
6	79	0,881	89,626
7	88	0,951	92,538
8	130	1,410	92,188
9	69	0,757	91,098
10	82	0,881	93,029
11	90	0,951	94,641

Фактичні дані (2-й стовпчик) ділимо на сезонну варіацію (3-й стовпчик) і записуємо результат в 4-й стовпчик.

4. Знайдемо рівняння тренда у вигляді лінійної регресійної моделі: $T = ax + в$. Для знаходження коефіцієнтів скористуємося статистичними функціями ОТРЕЗОК та НАКЛОН майстра функцій із пакету Excel. Отримаємо: $a = 1,2$; $в = 81,6$.

Отже, трендові значення об'єму продаж = 81,6 + 1,2 * номер квартала.

5. Розрахуємо похибки обчислень. Для зручності результати обчислень оформимо у вигляді таблиці 5.20.

$MAD = \sum_{i=1}^{n} \frac{|e_t|}{n} = \frac{11,2}{11} \approx 1$; $MSE = \sum_{i=1}^{n} \frac{e_t^2}{n} = \frac{17,1}{11} \approx 1,6$ – похибки складають приблизно 1%.

6. Зробимо прогноз об'ємів продажу на наступні два квартали:

Прогноз об'ємів продажу в 12 кварталі: $(81,6+1,2\cdot 12)\cdot 1,41=135,4$ тис. грн.
Прогноз об'ємів продажу в 13 кварталі: $(81,6+1,2\cdot 13)\cdot 0,757=73,6$ тис. грн.

Таблиця 5.20

Квартал	Об'єми продажу A	Десезонолізований об'єм продажу $A/S=T*E$	Трендові значення	Похибка e_t	$\lvert e_t \rvert$	e_t^2
1	63	83,176	82,8	0,4	0,4	0,16
2	74	83,953	84	0,0	0,0	0,00
3	79	83,074	85,2	−2,2	2,2	4,84
4	120	85,096	86,4	−1,3	1,3	1,69
5	67	88,457	87,6	0,9	0,9	0,81
6	79	89,626	88,8	0,9	0,9	0,81
7	88	92,538	90	2,4	2,4	5,76
8	130	92,188	91,2	1,0	1,0	1,00
9	69	91,098	92,4	−1,3	1,3	1,69
10	82	93,029	93,6	−0,5	0,5	0,25
11	90	94,641	94,8	−0,3	0,3	0,09
				Сума	11,2	17,10

Приклад 5.7. В таблиці 5.21 вказаний об'єм продажу (тис. грн.) за 11 кварталів. На основі цих даних зробити прогноз на наступні 2 квартали.

Таблиця 5.21

Квартал	1	2	3	4	5	6	7	8	9	10	11
Об'єм продажу	4	5	5	6	9	9	8	10	11	13	16

Розв'язок. Побудуємо модель експоненційного згладжування за формулою:

новий прогноз = $\alpha\cdot$фактичний результат в останній період +
+ $(1-\alpha)\cdot$прогноз в останній період

Нехай $\alpha = 0,8$, тоді $1-\alpha = 0,2$. Перший прогноз обираємо рівним першому фактичному значенню, далі користуємось формулою експоненційного згладжування.

Експоненційне згладжування зручно проводити за допомогою сервісних функцій Excel. Необхідно викликати ***Сервис – Аналіз данных – Экспоненциальное сглаживание – ОК***. У графі ***Фактор затухания*** вказати значення $1-\alpha$ (стандартне значення 0,3).

Для зручності результати обчислень оформимо у вигляді табл. 5.22.

Таблиця 5.22

Квартал	Об'єми продажу	Прогноз
1	**2**	**3**
1	4	4
2	6	4
3	4	5,6
4	5	4,32

Продовження таблиці 5.22

1	2	3
5	10	4,864
6	8	8,9728
7	7	8,19456
8	9	7,238912
9	12	8,6477824
10	14	11,32955648
11	15	13,4659113
12		13,77

Похибка прогнозу обчислюється аналогічно як і для адитивної і мультиплікативної моделей.

5.4. Елементи факторного аналізу

5.4.1. Елементи факторного аналізу

Факторний аналіз − сукупність моделей і методів, призначених для стискання інформації, яка міститься в кореляційній матриці. Він допомагає виявити приховані фактори, які пояснюють взаємозв'язки між спостережуваними ознаками досліджуваного об'єкта. Кількість ознак може бути великою і зв'язки між ними надзвичайно складними, однак, спостерігаючи за об'єктом, ми можемо виявити невелику кількість факторів, які впливають на досліджувані ознаки. Факторний аналіз передбачає класифікацію ознак, які мають подібний характер зміни при переході від одного об'єкта спостереження до іншого.

Обґрунтована заміна великої кількості ознак, описаних об'єктами спостережень, меншим числом комплексних характеристик (факторів) складають зміст факторного аналізу. Кожний **фактор** − це група взаємопов'язаних ознак, які визначають змістовну інтерпретацію даного фактора. При цьому в один фактор об'єднуються ознаки, які тісно корелюють між собою. Ознаки з різних факторів характеризуються слабким кореляційним зв'язком.

Основними етапами факторного аналізу є:

1) збір емпіричних даних і підготовка кореляційної (коваріаційної) матриці;

2) виділення початкових факторів і побудова факторної структури (обчислення факторних навантажень): проводиться вибір методу обчислення, визначається кількість факторів на основі змістовних або математичних міркувань;

3) обертання факторної структури: вибір критерію обертання;

4) змістовна інтерпретація результатів факторного аналізу;

5) обчислення факторних значень.

Припустимо, що досліджується деякий масив з n емпіричних ознак. За допомогою методів кореляційного аналізу можна встановити залежності між ними, обчисливши коефіцієнти кореляції. Тоді вся множина з n емпіричних

ознак розіб’ється на окремі групи за величиною коефіцієнтів кореляції. Наприклад, перша ознака тісно пов'язана з четвертою і шостою, а шоста з першою, перша з п’ятою і т. д., причому з іншими ознаками зв'язок виявляється значно слабкішим. Тоді ці взаємопов'язані ознаки утворюють загальну функціональну одиницю, яку ми і називаємо фактором. Наприклад, при аналізі успішності студентів деякий фактор має високу додатну кореляцію з оцінкою з вищої математики, інформатики, розділів фізики і високу від’ємну кореляцію з політології, філософії, історії − фактор характеризує точне мислення.

Створення математичної моделі факторного аналізу базується на припущенні про те, що усі зміни значень ознак обумовлені зміною деяких прихованих властивостей спостережуваних об'єктів. Ці приховані властивості називаються **загальними факторами** і їх кількість має бути меншою від числа ознак, за допомогою яких вони вимірюються. Кожний такий фактор має окреме значення значущості для різних досліджуваних ознак. Рівень значущості кожного фактора називається його **факторним навантаженням**. Він визначає степінь впливу загального фактора на зміну даної ознаки.

На зміну значень спостережуваної ознаки можуть впливати також деякі суб’єктивні, властиві тільки цій ознаці, зміни. Вони можуть бути викликані випадковими помилками, похибками вимірів і т.д.. Причини усіх таких не взаємообумовлених змін об’єднуються в поняття **специфічного фактора.**

Отже, зміни значень спостережуваних ознак залежать від двох складових: загальних факторів і специфічних. Нехай X_i − *i*-та емпірична ознака. Позначимо через U_i − загальну частину цієї ознаки (частина змін, викликана впливом загальних факторів) і ε_i − специфічні фактори, зміни яких не пов'язані один з одним і не залежать від змін інших показників зміни ознаки X_i. Тоді зміна ознаки X_i розкладається на суму загальної частини усіх змін і змін, викликаних впливом специфічних факторів.

$$X_i = U_i + \varepsilon_i .$$

Подальший розвиток ідеї факторного аналізу ґрунтується на тому припущенні, що дані *n* змінних U_i є лінійними комбінаціями меншого числа інших змінних F_j, які називаються факторами, тобто

$$U_i = \omega_{1i}F_1 + \omega_{2i}F_2 + ... + \omega_{ki}F_k , \text{ де } i = \overline{1,n} ,$$

ω_{ji} − факторні навантаження факторів F_j, які характеризують степінь впливу j-го загального фактора на i-у емпіричну ознаку. Об’єднуючи вище зазначені формули, отримаємо, що X_i виражається через загальні і специфічні фактори таким чином:

$$X_i = \omega_{1i}F_1 + \omega_{2i}F_2 + ... + \omega_{ki}F_k + \varepsilon_i , \text{ де } i = \overline{1,n} .$$

Причому:

1. Загальні фактори F_j є або некорельованими випадковими величинами з дисперсією, що дорівнює 1, або невідомими невипадковими параметрами.

2. Специфічні фактори ε_i мають нормальний розподіл, не корелюють між собою і не залежать від загальних факторів.

Тому наступним кроком має бути оцінка значущості загальних і специфічних факторів для зміни значень ознаки X_i. Для цього розглянемо дисперсію ознаки X_i:

$$D(X_i) = \frac{1}{N-1}\sum_{p=1}^{N}(x_{ip} - \bar{x}_i)^2 ,$$

де N – кількість об'єктів спостереження; x_{ip} – значення ознаки X_i для p-го об'єкта спостереження $(p = \overline{1,N})$; x_i – середнє значення i-ї ознаки. Так як U_i і ε_i не корелюють між собою, то:

$$D(X_i)=D(U_i)+D(\varepsilon_i).$$

Чим більше значення $D(U_i)$, тим більша частина змін ознаки X_i залежить від загальних факторів.

Значення дисперсій загальних факторів $D(F_j)$ дозволяють ранжувати їх за степенями впливу на зміну ознак X_i.

Математична модель факторного аналізу є розділом багатомірної статистики і потребує достатньо глибоких знань у таких розділах вищої математики, як матрична алгебра, основи математичної статистики і математичного аналізу. Тому опишемо змістовну структуру факторного аналізу, опускаючи складні математичні і логічні обґрунтування.

Усі методи оцінки зв'язків ознак можна розділити на дві групи: прямі і непрямі. Прямі − це методи, які дозволяють визначити силу зв'язку до проведення факторного аналізу. До непрямих методів відносяться апостеріорні оцінки, коли до проведення факторного аналізу вони не відомі (беруться довільні оцінки), а в якості додаткової умови береться кількість факторів.

Найпоширеніший прямий метод оцінки зв'язку між ознаками базується на припущенні про те, що усі оцінки зв'язків дорівнюють 1. Тобто, вони дорівнюють діагональним елементам кореляційної матриці, а, отже, дисперсії специфічних факторів дорівнюють 0. У цьому полягає зміст однієї з найвідоміших моделей факторного аналізу − методу головних компонент.

Метод головних компонент. Для оцінки факторних навантажень, в якості критерію, використовується мінімум розбіжності між кореляційною матрицею початкових ознак і тією, яка виходить після оцінювання навантажень. Іншими словами, метод головних компонент здійснює перехід до нової системи координат, яка є системою ортонормованих лінійних комбінацій. Лінійні комбінації є власними векторами кореляційної матриці. Перша головна компонента − це лінійна комбінація, яка має найбільшу дисперсію. Друга компонента має найбільшу дисперсію серед усіх інших лінійних комбінацій, які не корелюють з першою головною компонентою.

Метод головних факторів. Це одна з найпоширеніших моделей факторного аналізу. Метод вимагає попередньої оцінки дисперсій і для нього критерієм оптимальної оцінки факторних навантажень є максимальна наближеність початкових кореляцій ознак до тих, які отримані в моделі після оцінювання навантажень. Для визначення кількості факторів використовуються різні статистичні критерії, за допомогою яких перевіряється гіпотеза про незначущість матриці кореляційних лишків.

Метод максимальної правдоподібності (Д. Лоулі) на відміну від попередніх моделей факторного аналізу ґрунтується не на попередній оцінці дисперсій, а на апріорному визначенні кількості загальних факторів. У разі великої вибірки дозволяє отримати статистичний критерій значущості отриманого факторного рішення.

Метод мінімальних залишків (Г. Харман) ґрунтується на мінімізації недіагональних елементів залишкової матриці кореляцій, і так як метод максимальної правдоподібності, вимагає попереднього вибору кількості факторів, що пояснюють спільні зміни ознак.

Перераховані методи відрізняються за способом пошуку розв'язання основного рівняння факторного аналізу. Вибір методу вимагає великого досвіду роботи. Проте деякі дослідники використовують відразу декілька методів, і виділені в усіх методах фактори вважають найбільш стійкими.

Моделі факторного аналізу називають іноді «прямими» або «початковими» в тому розумінні, що отримувані з їх допомогою оцінки навантажень напряму залежать від значень початкових ознак. Але на практиці може виникнути ситуація, коли навантаження при емпіричних ознаках утворюють таку комбінацію знаків і величин, яка важко інтерпретується. Ці випадки змусили дослідників шукати шляхи зміни початкових навантажень, щоб отримати результат, який краще інтерпретується. Один з методів вирішення цієї задачі − обертання факторів, тобто поворот відповідних факторам координатних осей, який проводиться не в просторі початкових ознак, а в просторі знайдених факторів.

Отже, наступним етапом факторного аналізу є обертання факторів, яке базується на принципах простої структури Терстоуна:

− в кожній стрічці факторної структури має бути хоча б один нуль;

− в кожному стовпчику − принаймні *к* нулів (*к* − кількість загальних факторів);

− для кожної пари стовпчиків можна знайти принаймні *к* параметрів (емпіричних ознак), для яких елементи факторної структури дорівнюють нулю в одному з двох стовпчиків і не дорівнюють нулю − в іншому.

На основі цих принципів побудовано велику кількість аналітичних методів, які максимізують деякий критерій. Суть процесу обертання будь-якої пари векторів полягає в знаходженні такого кута між новим і старим напрямом факторів, який давав би найбільший приріст обраного критерію. Обертання факторів в просторі дає можливість кожній ознаці охарактеризуватися переважаючим впливом якогось одного фактора.

У сучасних пакетах статистичної обробки даних найчастіше використовуються методи обертання: варімакс, квартімакс і еквімакс. Обертання методом варімаксу спрощує значення стовпчиків факторної матриці, зводячи їх до 1 або 0. Обертання методом квартімаксу спрощує значення елементів стрічок факторної матриці. І, нарешті, еквімакс займає проміжне положення − при обертанні факторів за цим методом одночасно робиться спроба спростити значення елементів і стовпчиків, і стрічок.

Отримавши після процедури обертання факторне розв’язання (факторну матрицю), можна переходити до інтерпретації і найменування факторів. Цей етап роботи цілком і повністю залежить від інтуїції, рівня обізнаності і практичного досвіду дослідника. Щоб зрозуміти природу конкретного фактора, необхідно проаналізувати зміст ознак, які входять у даний фактор, і спробувати виявити спільні для них риси. Чим більша кількість ознак з великим значенням навантажень у цьому факторі, тим легше розкрити його природу. Для вибору назви фактора немає формалізованих прийомів. В якості попереднього варіанту можна використовувати ім'я ознаки, яка увійшла до фактора з найбільшим навантаженням.

Отже, процес інтерпретації факторних навантажень − це пошук таких загальних властивостей системи спостережень, які б могли бути описані в термінах одночасного збільшення (зменшення) однієї групи ознак на противагу зменшенню (збільшенню) іншої групи або просто збільшенню (зменшенню) якої-небудь частини ознак.

Якщо фактори знайдені і представлені, то на останньому кроці факторного аналізу, додатковим ознакам (додаткові запитання анкети) можна присвоїти значення цих факторів, так звані факторні значення. Тоді для кожного об’єкта спостереження значення великої кількості ознак можна перевести у значення невеликої кількості факторів − нових ознак.

Факторний аналіз є складною процедурою. Як правило, досконале факторне розв’язання (досить просте і таке, що змістовно інтерпретується) вдається отримати щонайменше після декількох циклів його проведення − від відбору ознак до спроби інтерпретації після обертання факторів. Для успішного проведення факторного аналізу необхідно дотримуватись основних вимог:

1) Змінні мають належати до шкали інтервалів (за класифікацією Стівенса). Передбачається, що порядкові змінні підлягають факторному аналізу, якщо їм надати числових значень. Не слід включати дихотомічні змінні в аналіз, якщо завдання не передбачають зменшення кількості ознак.

2) Відбираючи ознаки для факторного аналізу, слід враховувати, що на один фактор має припадати не менше трьох ознак.

3) Не варто включати у факторний аналіз ознаки, які мають дуже слабкі зв’язки з іншими ознаками. Велика ймовірність того, що вони не ввійдуть ні до одного фактора. Якщо в роботі не стоїть завдання сформувати шкалу нового опитування на основі факторного аналізу або якого-небудь аналогічного завдання, то не слід також включати усі ознаки, які мають між собою дуже тісні зв’язки. Швидше за все, вони утворюють один фактор. Чим більше таких ознак включається у факторний аналіз, тим більша ймовірність того, що вони утворюють перший фактор і до нього приєднається більшість інших ознак.

4) Стійкість виявленої факторної структури (її невипадковість) тим менша, чим більше складових її факторів. Вона також нестійка при малій кількості випробовуваних об’єктів.

5.4.2. Факторний аналіз засобами SPSS

Проілюструємо описаний метод на прикладі складеної і апробованої нами анкети. Етап дослідження був проведений для населення м. Одеси віком 18-35 років, в процесі якого вивчалася думка молоді щодо новинок вітчизняного ринку − продуктів з вмістом ГМО. Респондентам пропонували висловити свою думку до наступних положень:

1. Потрібно досконало вивчити вплив ГМО на організм людини.
2. Необхідно виробити стандарт, який би регулював допустиме відсоткове відношення генно-модифікованих організмів у продуктах харчування.
3. В Україні існує велика кількість натуральних продуктів.
4. Кількість населення і так іде на спад, а тут ще й продукти з ГМО.
5. Варто прислухатися до розвитку інноваційних технологій − природні ресурси вичерпні.
6. Кількість продуктів з ГМО на ринку України варто обмежити.
7. Натуральні продукти зникли.
8. Необхідно вивчити результати досліджень міжнародних компаній щодо вмісту неприродних компонентів.
9. Заборона використання продуктів з вмістом ГМО принесе шкоду економіці країни.
10. Варто агітувати населення країни щодо споживання натуральних продуктів.
11. Ми не маємо достовірної інформації про вміст продуктів, які ми вживаємо.
12. Порушуючи права споживачів, виробники компрометують себе.
13. Ми варті якісних продуктів.
14. Не хочу навіть чути про ГМО.
15. Апробація продуктів з ГМО − поступове скорочення кількості населення.

Оцінки ставилися за семибальною шкалою: від повної незгоди (-3) до повної згоди (3). За результатами опитування 110 респондентів проводився факторний аналіз засобами SPSS. Запитання анкети введені під іменами *X1*, *X2*,… , *X15* відповідно. Для виявлення факторів п’ятнадцяти ознак опишемо можливості використання програми.

Для цього в меню необхідно вибрати послідовність команд:

1) ***Анализ → Снижение размерности → Факторный анализ*** ... Відкриється діалогове вікно *Факторный анализ* (рис. 5.1);

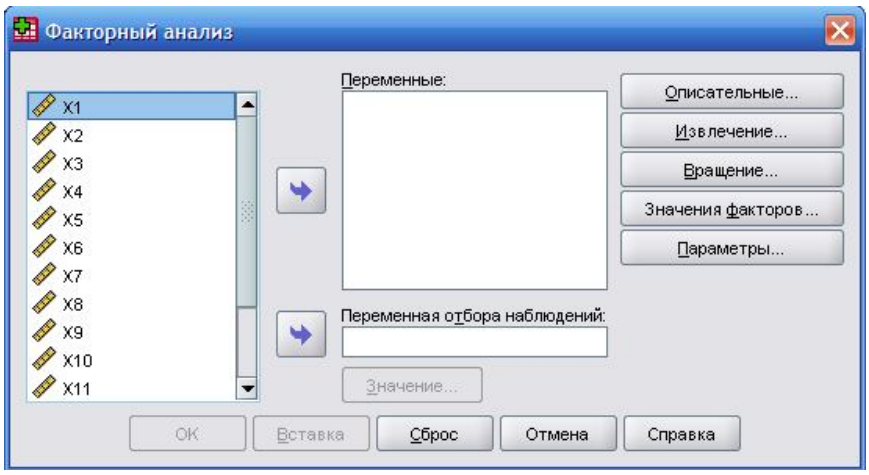


Рисунок 5.1. Діалогове вікно для факторного аналізу

2) Змінні *Xl − Xl5* необхідно перенести в поле *Переменные* і ознайомитись з можливостями різних кнопок даного діалогового меню.

Після натискування на кнопці *Описательные…* відкриється діалогове вікно *Факторный анализ: Описательные*, зображене на рис. 5.2, у якому варто залишити виведення первинних результатів, які включають початкові відносні дисперсії простих факторів, власні значення і процентні частки об'єднаної дисперсії. Часто виникає необхідність у виведенні одновимірних статистик і кореляційних коефіцієнтів, яку можна реалізувати за допомогою функцій та можливостей даного діалогового вікна.

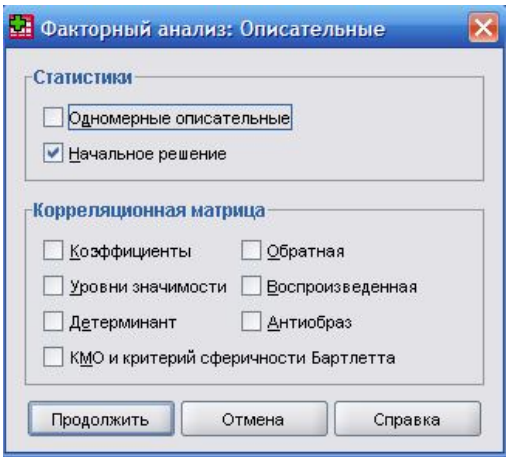


Рисунок 5.2. Діалогове вікно вибору параметрів факторного аналізу

За допомогою кнопки *Извлечение* можна вибрати метод відбору (перелік методів зображено на рис. 5.3) факторів; залишаємо встановлений автоматично аналіз головних компонентів. Кількість вибраних факторів прирівнюється до власних значень, які перевищують одиницю. Величину значення можна відкорегувати відповідною опцією.

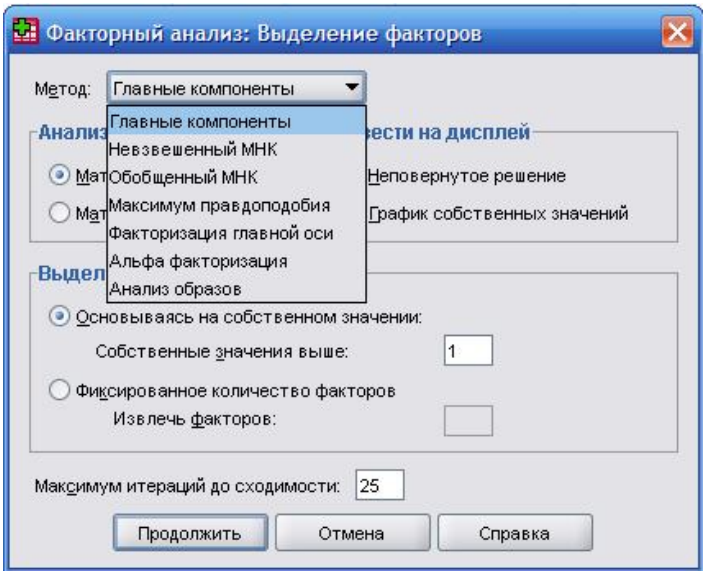


Рисунок 5.3. Методи відбору змінних у фактори

3) У діалоговому вікні *Факторный анализ: Вращение* вибирається метод обертання. Активуємо метод варімаксу і залишаємо активним виведення поверненої матриці факторів. Можна здійснити інтерпретацію факторних навантажень в графічному виді, в якому перші три фактори будуть представлені в тривимірному просторі; у випадку наявності тільки двох факторів − на площині.

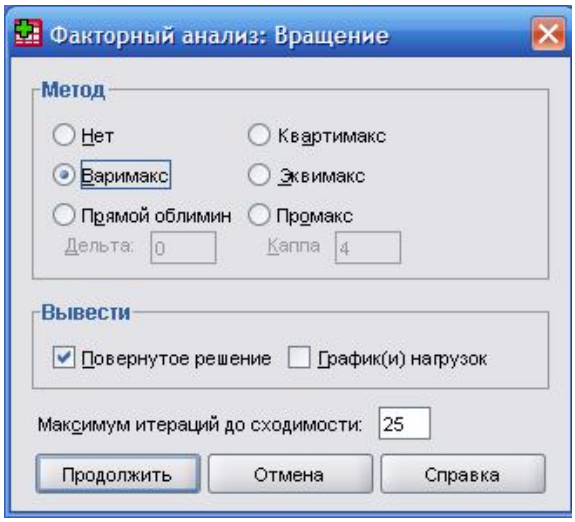


Рисунок 5.4. Діалогове вікно вибору методу обертання

Якщо потрібно знайти значення факторів і зберегти їх у вигляді додаткових змінних, варто натиснути на кнопку *Значения факторов* і відмітити *Сохранить как переменные*. Автоматично встановлений регресійний метод.

Пункт *Параметры* призначений для обробки пропущених значень.

4) Для проведення розрахунків натискаємо на *ОК*. У вікні виводу з'являться результати. Спочатку приводяться первинні статистики, які наведені у табл. 5.23:

Таблиця 5.23

Повна пояснена дисперсія

Компоненти	Початкові власні значення			Суми квадратів навантажень		
	Всього	%	Кумулятивний %	Всього	% дисперсії	Кумулятивний %
1	5,146	34,308	34,308	3,466	23,105	23,105
2	1,945	12,970	47,278	2,536	16,907	40,013
3	1,415	9,433	56,711	2,505	16,698	56,711
4	,990	6,601	63,312			
5	,936	6,238	69,550			
6	,760	5,068	74,617			
7	,693	4,622	79,240			
8	,612	4,083	83,323			
9	,529	3,529	86,852			
10	,473	3,151	90,004			
11	,433	2,889	92,893			
12	,339	2,262	95,1555			
13	,301	2,007	97,161			
14	,245	1,635	98,797			
15	,181	1,203	100,000			

Метод відбору: Аналіз головних компонент

За даними таблиці можна побачити, що значення трьох власних факторів більше за одиницю. Отже, для аналізу відібрано тільки три фактори. Перший фактор пояснює 34,308 % сумарної дисперсії, другий − 12,97 % і третій − 9,433 %. Оскільки ми відмінили виведення неповерненої матриці факторів, то даними табл. 5.24 є значення поверненої матриці.

Таблиця 5.24

Матриця повернутих компонент

	Компонент		
	1	2	3
X1	−,466	,628	−,191
X2	−,141	,657	,215
X3	,327	−,153	,711
X4	,533	−,106	,394
X5	−,362	,783	4,52E-02
X6	−1,2E-02	−3,8E-02	,763
X7	,525	3,58E-02	,543
X8	− 117	719	− 267
X9	2,56E-02	,551	−8,8E-02
X10	,252	−9,5E-02	,685
X11	,125	,392	−,292
X12	,802	−,199	,108
X13	,685	−,110	,465
X14	,837	−,144	−2,5E-02
X15	,725	−4,8E-02	,144

Метод відбору: Аналіз головних компонент
Метод обертання: Варімакс з нормалізацією Кайзера
а. Обертання виконано за 8 ітерацій

Найголовніша частина факторного аналізу − пояснення відібраних факторів. Для цього в кожній стрічці поверненої факторної матриці потрібно відмітити те факторне навантаження, яке має найбільше абсолютне значення.

Ці факторні навантаження слід розуміти як кореляційні коефіцієнти між ознаками і факторами. Так положення *X1* найбільше корелює з фактором 2, величина кореляції складає 0,628; ознака *X2* також найсильніше корелює з фактором 2 згідно значення 0,657; ознака *X3* найтісніше − з фактором 3 (0,711) і т. д. У більшості випадків включення окремої змінної в один фактор, здійснюване на основі коефіцієнтів кореляції, є однозначним. В окремих випадках, наприклад, як в ситуації із ознакою *X7*, вона може відноситися до двох факторів одночасно. Можуть бути також і ознаки, в нашому прикладі *X11*, які не можна включити ні в один із вибраних факторів.

Отже, на основі вище зазначеного, ознаки можна віднести в наступному порядку до трьох факторів:

Фактор 1

Кількість населення і так іде на спад, а тут ще й продукти з ГМО.

Натуральні продукти зникли.

Порушуючи права споживачів, виробники компрометують себе.

Ми варті якісних продуктів.

Не хочу навіть чути про ГМО.

Апробація продуктів з ГМО − поступове скорочення кількості населення.

Фактор 2

Потрібно досконало вивчити вплив ГМО на організм людини.

Необхідно виробити стандарт, який би регулював допустиме відсоткове відношення генно-модифікованих організмів у продуктах харчування.

Варто прислухатися до розвитку інноваційних технологій − природні ресурси вичерпні.

Необхідно вивчити результати досліджень міжнародних компаній щодо вмісту неприродних компонентів.

Заборона використання продуктів з вмістом ГМО принесе шкоду економіці країни.

Ми не маємо достовірної інформації про вміст продуктів, які ми вживаємо.

Фактор 3

В Україні існує велика кількість натуральних продуктів.

Кількість продуктів з ГМО на ринку України варто обмежити.

Натуральні продукти зникли.

Варто агітувати населення країни щодо споживання натуральних продуктів.

Так як положення «Натуральні продукти зникли» має однакові значення навантажень як для фактора 1, так і для фактора 3, то воно включається в обидва фактори.

Останнім і вирішальним кроком факторного аналізу є виявлення і опис змістового зв’язку факторів.

Перший фактор зібрав усі положення, які агресивно налаштовані по відношенню до появи на вітчизняному ринку продуктів з вмістом ГМО.

Другий фактор об'єднує положення, які, в деякій мірі, підтримують обмежене споживання продуктів з вмістом ГМО.

До третього фактора увійшли точки зору, які, не відкидаючи споживання генно-модифікованих продуктів, підтримують споживання натуральних продуктів.

Відповідно до порядку висловлювань ці три фактори можна коротко охарактеризувати за допомогою наступних виразів: 1) негативне відношення, 2) підтримка обмеженого споживання продуктів з вмістом ГМО, 3) нейтральна позиція.

Як показує практика, не завжди фактори можна чітко пояснити. Якщо неможливо здійснити обґрунтоване пояснення факторів, то проведений факторний аналіз можна вважати невдалим.

Оскільки зробили розрахунок значень факторів, то відповідно до трьох відібраних автоматично утворилися три нові змінні, під назвою fac1_l, fac2_l і fac3_l, які містять обчислені значення факторів.

Факторні значення, як правило, лежать в межах від − 3 до 3 і за своєю величиною пояснюють тісноту зв'язку кожної ознаки з відповідним фактором.

5.5. Основні вимоги до аналізу даних та формування звіту

Етап аналізу отриманої інформації − це зіставлення отриманої про вивчений об'єкт інформації з уже відомим об'ємом знань про нього.

Основною метою аналізу даних дослідження є пояснення змісту окремих результатів, об'єднання і виділення узагальнюючих положень, зведення їх в одну теоретичну систему.

Для отримання надійних і достовірних результатів емпіричного аналізу варто дотримуватися ряду вимог:

1) дослідник повинен мати уявлення про логіку використаних математичних методів;

2) застосовувати правильно підібрані математичні методи для аналізу даних досліджень, адже множина математично-статистичних методів буває достатньо різноманітною в залежності від типу дослідження: емпіричного, прикладного або теоретичного.

3) варто попередньо провести пробну обробку на невеликій кількості масиву даних.

У процесі аналізу та узагальнення результатів умовно можна виділити кілька етапів.

Перший − це етап впорядкування, класифікації, групування даних у відповідності з дослідницькими гіпотезами.

Другий етап аналізу та інтерпретації − це узагальнення даних, перевірка значущості і достовірності числових характеристик.

Третій етап − перевірка дослідницьких гіпотез за допомогою отриманих числових характеристик.

Результати дослідження, зокрема соціологічного, завжди відображаються у звіті. Звіт повинен бути основою для подальшої теоретичної роботи і формою впровадження результатів дослідження в практику.

Структура змісту звіту залежить від типу дослідження: теоретичного чи прикладного. У звіті описуються проблеми, цілі і завдання, об'єкт та предмет, інтерпретуються основні поняття, подається стан вивчення та дослідження проблеми у сучасній науці, обґрунтовується вибірка, методи збору інформації, аналізуються результати і визначається степінь розв'язання поставлених завдань.

Основні вимоги до звіту:

1. У звіті повинні бути відображені всі взаємопов'язані групи проблем у відповідності з логікою наукового пошуку.

2. Кожний розділ звіту має складатися із двох частин: у першій − проблеми та результати, у другій − висновки.

3. Звіт формується незалежно від послідовності запитань в анкеті чи опитуванні.

4. Правильне оформлення звіту.

Звіт має:

− містити кількісні параметри вивченого об'єкта;

− описувати основні тенденції і темпи розвитку явищ;

− розкривати взаємозв'язок між ознаками та явищами за допомогою отриманих показників.

Універсального рецепту для правильної інтерпретації кількісних показників у науці немає. Це залежить від професіоналізму, компетентності, культури, ціннісних орієнтацій і установок дослідника, науковця.

До звіту, як правило, додається пояснювальна записка, в якій описуються результати, і додатки з таблицями, графіками.

Варто пам'ятати, що великий об'єм цифрового матеріалу дезорієнтує споживача інформації, тому ефективнішою є графічна інтерпретація результатів дослідження.

5.6. Професійне маніпулювання результатами дослідження

Дослідник-практик перед проведенням спостереження або експерименту періодично вивчає додаткові графіки та діаграми допоміжних матеріалів. Вивчаючи готову інформацію, можна проаналізувати помилки, неточності та непорозуміння, які були допущенні попередніми дослідниками. Це, з однієї сторони, змушує уважніше відноситись до представлення результатів дослідження, а з іншої, набувати досвіду маніпулювання даних для ефективного відображення необхідних висновків. Адже не виключено, що коли-небудь доведеться виконати роботу, результати якої можуть зачепити чиї-небудь інтереси або стануть предметом гострих суперечок.

Правильно відображена вихідна інформація допоможе уникнути звинувачень в упередженості і некомпетентності зі сторони осіб, зацікавлених в протилежних результатах дослідження.

Відомо, що методи графічного відображення даних часто стають об’єктами свідомої або неусвідомленої фальсифікації. Так, яскраво виражені негативні дані або, навпаки, дані, що навіюють нестримний оптимізм, передбачають упущення в графіку (таблиці) якого-небудь фактора, що істотно впливає на результати.

Маневрування статистичними даними часто виникає від незнання прийомів статистики та невміння правильно використовувати відповідні методи. Особливо часто це спостерігається в журналістиці, зокрема, в період передвиборних перегонів.

Річ у тому, що основна проблема полягає не в статистичних даних як таких, а в їх інтерпретації. Одну і ту ж вихідну інформацію можна по-різному тлумачити − від критичної оцінки до схвальної. Такі трактування базуються на прийомах впливу на свідомість публіки. У своїй основі ці прийоми зводяться до декількох способів роботи з матеріалами.

Для простоти та чіткості ми проілюструємо різні способи маніпулювання графічно представленою статистичною інформацією.

Спосіб перший: приховування одиниць вимірювання. Яскравим прикладом використання даного способу можуть бути результати дослідження неуспішності студентів. Статистика кількості незадовільних оцінок з вищої математики у деякому ВУЗі показує жахливі результати (табл. 5.24, рис. 5.5).

Таблиця 5.24

Відомості про незадовільні результати здачі іспитів з вищої математики

	2004	**2005**	**2006**	**2007**	**2008**
Технічні спеціальності	90	93	107	123	165
Економічні спеціальності	85	86	90	97	102
Гуманітарні напрями	67	68	72	89	96
Всього	242	247	269	309	363

Візуальна інтерпретація даних формує лише жахливе уявлення про якість знань студентів. Адже, починаючи з 2005 р. крива графіка стрімко зростає. Навіть у випадку, якщо графічний матеріал супроводжується відповідною таблицею (як це зроблено в даному випадку), більшість читачів орієнтується тільки на графік.

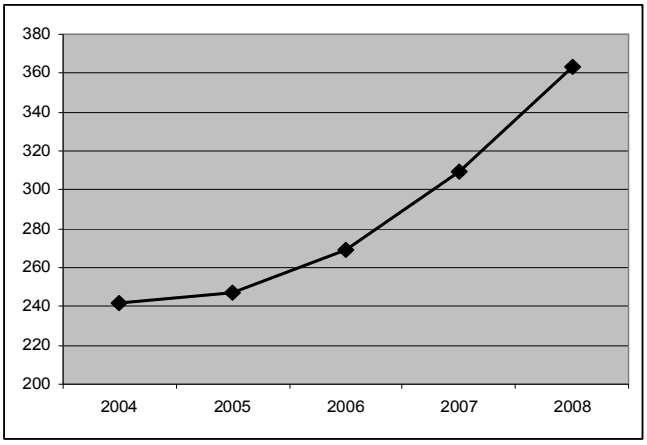


Рисунок 5.5. Результати здачі іспитів з вищої математики: незадовільні оцінки

Причин такого негативного кількісного аналізу успішності з вищої математики може бути кілька: опущено фактори, які впливають на значення показників (загальна кількість студентів ВУЗу зросла); некоректно вибрано одиниці вимірювання.

При представлені результатів необхідно доповнити таблицю хоча б відносними одиницями виміру.

Вводячи в таблицю 1 додатковий параметр − відношення кількості негативних до загальної кількості усіх оцінок з предмету (у відсотках), ми отримаємо реальнішу ситуацію досліджуваної ознаки (табл. 5.25).

Таблиця 5.25

Відомості про незадовільні оцінки з вищої математики

	2004		**2005**		**2006**		**2007**		**2008**	
Технічні спеціальності	90	**8,6%**	93	**9%**	107	**9,5%**	123	**8,8%**	165	**9,2%**
Економічні спеціальності	85	**10%**	86	**9,5%**	90	**9,5%**	97	**9%**	102	**9%**
Гуманітарні напрями	67	**12%**	68	**10%**	72	**11%**	89	**12%**	96	**11%**
Всього	242		247		269		309		363	

Різкий приріст показників неуспішності пояснюється збільшенням кількості студентів у ВУЗі. Обчисливши середнє значення відсотків зі спеціальностей для кожного року, отримаємо такі дані:

$$\frac{8,6+10+12}{3}=10,2;\quad \frac{9+9,5+10}{3}=9,5;\quad \frac{9,5+9,5+11}{3}=10;\quad \frac{8,8+9+12}{3}=9,9;$$

$$\frac{9,2+9+11}{3}=9,7.$$

Вони не лише не ідентичні попереднім, а й показують абсолютно інші тенденції змін неуспішності студентів за роками (рис. 5.6).

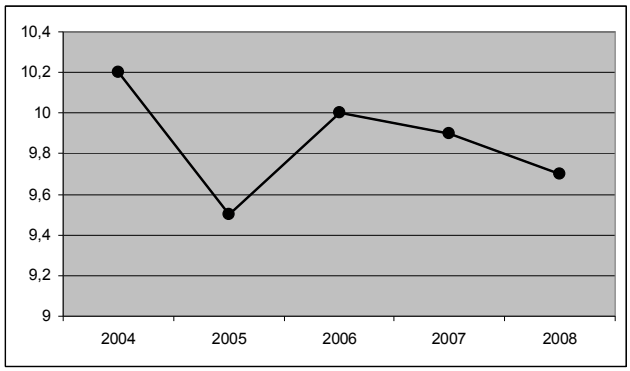
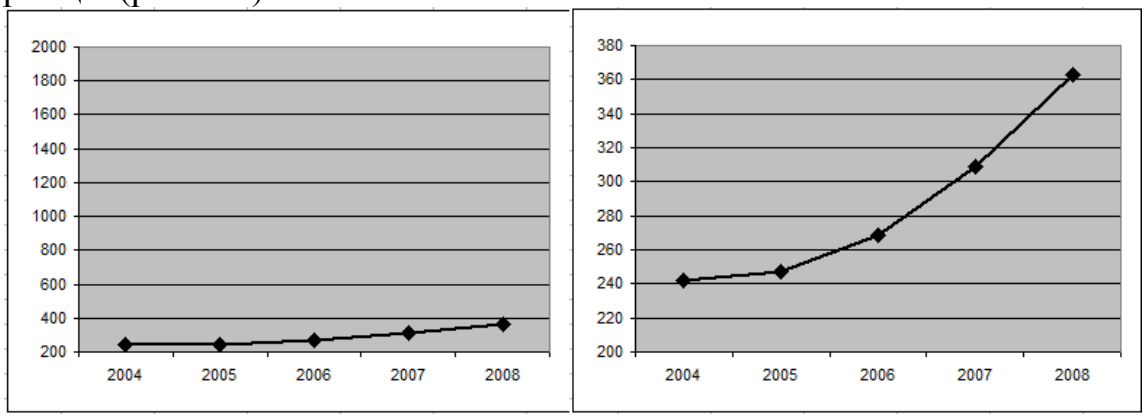


Рисунок 5.6. Відносні частки незадовільних оцінок з вищої математики по роках

Доповнюючи емпіричні дані одним показником, ми змінили характер поведінки вивченої ознаки. Хоча теоретично створити аналітичний матеріал, який усесторонньо схарактеризує об’єкт і предмет дослідження − це мистецтво досвідченого дослідника-практика.

Другий спосіб полягає в умінні маніпулювання осями графіків. Два наступні графіки показують, як по-різному може сприйматися одна і та ж інформація (рис. 5.7).



а) б)

Рисунок 5.7. Результати здачі іспитів з вищої математики: незадовільні оцінки

Графіки побудовані на основі ідентичних даних, а їх візуальне зображення інтерпретується по-різному. На рис. 5.7а) ми спостерігаємо допустиму тенденцію росту кількості незадовільних оцінок протягом п’яти років. На відміну від цього рис. 5.7б) демонструє катастрофічну ситуацію спаду успішності студентів. Такі результати отримані шляхом стискання-розтягування осі ординат відповідного графіка.

Інколи використовується масштабування, яке теж допомагає маніпулювати результатами, гіперболізуючи або мінімізуючи їх. Знявши числові поділки на шкалі можна досягнути ще кращого ефекту. Максимально звужуючи вісь *OX* та розтягуючи *OY*, спостерігатимемо тенденцію до деякого збільшення числового показника − це стане картиною нестримного росту (катастрофічного падіння) характеристики досліджуваної ознаки.

Третій спосіб передбачає усвідомлений або випадковий аналіз частини даних для отримання бажаних для замовника результатів. Тут часто керуються принципом «якщо факти не підходять до теорії − тим гірше для фактів» [1, 178].

За допомогою вищеописаного прикладу графічно продемонструємо спад відносної частки незадовільних оцінок студентів з вищої математики (рис. 5.8). Звідси можна прийти до висновку, що із кожним навчальним роком рівень успішності студентів даного ВУЗу зростає.

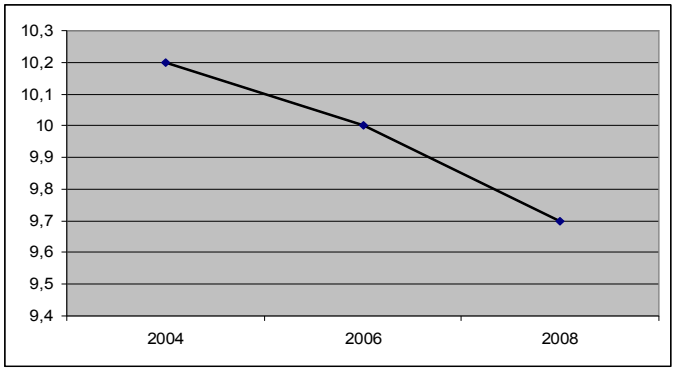


Рисунок 5.8. Відносні частки незадовільних оцінок з вищої математики

Таким чином, дослідник може підвищити репутацію та рейтингову оцінку навчального закладу.

Отже, ми на одному прикладі продемонстрували різні можливості інтерпретації результатів дослідження. Це ні в якому разі не фабрикування результатів, а лише незначне маневрування значеннями і графіками. Адже, раціональність мислення людини часто поступається емоційному сприйманні навколишньої дійсності.

Питання для самоконтролю

1. Що таке тренд?
2. Що таке сезонна варіація?
3. Як розраховується ковзке середнє?
4. Який алгоритм побудови адитивної моделі?
5. Який алгоритм побудови мультиплікативної моделі?
6. Який алгоритм побудови експоненційного згладжування?
7. Як здійснюється прогноз за методом експоненційного згладжування?
8. Яким чином константа згладжування впливає на прогноз?
9. Як розраховується помилка прогнозу?
10. Як обґрунтовується оптимальність моделі для прогнозування?
11. Що таке фактор?
12. Що таке факторне навантаження?
13. Чим відрізняються загальні та спеціальні фактори?
14. Які методи оцінки зв’язків факторного аналізу ви знаєте?
15. Які методи обертання факторів використовую статистичні пакети програм?
16. Що таке інтепретація факторів?
17. Для чого знаходяться факторні значення?
18. Які вимоги до звіту ви знаєте?

ЛІТЕРАТУРА

1. Агабекян Р. Л. Математические методы в социологии. Анализ данных и логика вывода в эмпирическом исследовании: учеб. пособ. для вузов / Р. Л. Агабекян, М. М. Кириченко, С. В. Усатиков. − Ростов н/Д: Феникс, 2005. − 192 с.
2. Бутник О. М. Економіко-математичне моделювання перехідних процесів у соціально-економічних системах: [монографія] / Бутник О. М. – Х.: Видавничий Дім „ІНЖЕК”; СПД Лібуркіна Л. М., 2004. – 304 с.
3. Бююль А. SPSS: искусство обработки информации. Platinum Edition, пер. с нем. / А. Бююль, П . Цёфель.− СПб.: ООО «ДиаСофтЮП», 2005. − 608 с.
4. Валентинов В. А. Эконометрика: практикум / Валентинов В. А. – М.: РДЛ, 2007. – 436 с.
5. Дослідження операцій: Навч. посіб. / М. Г. Медведєв, О. В. Колодінська. – [2-ге вид., перер. і доп.]. – К.: Вид-во Європ. ун-ту, 2006. – 158 с.
6. Дубина А. Г. Excel для экономистов и менеджеров / А. Г. Дубина, С. С. Орлова, И. Ю. Шубина, А. В. Хромов. – СПб.: Питер, 2004. – 295 с.
7. Екимов С. В. Нетрадиционные подходы в экономико-математическом моделировании: [монография] / Екимов С. В. – Днепропетровск: Наука и образование, 2004. – 240 с.
8. Ермолаев О. Ю. Математическая статистика для психологов: [учебник] / Ермолаев О. Ю. – [2-е изд. испр.]. – М.: Московский психолого-социальный институт Флинта, 2003. – 336 с.
9. Карагодова О. О. Дослідження операцій: навч. посіб. / О. О. Карагодова, В. Р. Кігель, В. Д. Рожок. – К.: Центр учбової літератури, 2007. – 256 с.
10. Лапач С. Н. Статистика в науке и бизнесе / С. Н. Лапач, А. В. Чубенко , П. Н. Бабич. – К.: МОРИОН, 2002. – 640 с.
11. Макаренко Т. І. Моделювання та прогнозування у маркетингу: навч. посіб. / Макаренко Т. І. – К.: Центр навчальної літератури, 2005. – 160 с.
12. Минько А. А. Статистический анализ в MS Excel / Минько А. А. – М.: Изд. дом «Вильямс», 2004. – 448 с.
13. Наследов А. Д. SPSS: Компьютерный анализ данных в психологии и социальных науках / Наследов А. Д. − СПб.: Питер, 2005. − 416 с.
14. Невежин В. П. Сборник задач по курсу «Экономико-математическое моделирование» / В. П. Невежин, С. И. Кружилов. – М.: ОАО «Изд. дом „Городец”», 2005. – 320 с.
15. Просветов Г. И. Эконометрика: Задачи и решения: учебно-метод. пособ. – [4-е изд., доп.] / Просветов Г. И. – М.: Издательство РДЛ, 2007. – 192 с.
16. Таха Х. Введение в исследование операций. – [7-е изд.] / Таха Х., пер. с англ. – М.: Изд. дом "Вильямс", 2007. – 912 с.
17. Шимко П. Д. Статистика / П. Д. Шимко, М. П. Власов. − Ростов н/Д: Феникс, 2003. − 448 с. – [Серия «Учебники, учебные пособия»].
18. Экономико-математические методы и прикладные модели: учеб. пособ.; Под ред. В. В. Федосеева. – М.: ЮНИТИ, 1999. – 321 с.

Навчальне видання

ВАСИЛЕНКО Оксана Анатоліївна,
СЕНЧА Ірина Анатоліївна

Математично-статистичні методи аналізу в прикладних дослідженнях

Навчальний посібник

Редактор ***Г. Ю. Греля***

Комп′ютерне верстання ***Є. С. Корнійчук***

Здано в набір 9.04.2012 Підписано до друку 10.05.2012.
Формат 60/88/16 Зам. № 4847.
Тираж 100 прим. Обсяг: 10,5 ум. друк. арк.
Віддруковано на видавничому устаткуванні фірми RISO
у друкарні редакційно-видавничого центру ОНАЗ ім. О.С. Попова
ОНАЗ, 2012