Лекція 5

КОРЕЛЯЦІЙНИЙ АНАЛІЗ

Основні положення кореляційного аналізу

Поняття кореляції з'явилося разом із поняттям регресії в середині XIX ст. завдяки роботам англійських статистиків Ф. Гальтона і К. Пірсона.

В сфері ІТ технологій між змінними величинами існує залежність, коли кожному значенню однієї змінної відповідає множина можливих значень іншої змінної, тобто кожній змінній відповідає умовний розподіл другої змінної. Така залежність отримала назву **статистичної** або **стохастичної.** Прикладами статистичного зв'язку є залежність швидкості доступу від кількості користувачів, ефективність роботи ПЗ в залежності від обраної моделі алгоритма, і т.ін.

Статистична залежність між двома змінними, при якій кожному значенню однієї змінної відповідає певне умовне математичне сподівання іншої, називається **кореляційною** (від латинського «*correlatio»* - співвідношення, взаємозв'язок).

Кореляційна залежність має вигляд:

$$M_x(Y) = \varphi(x)$$

або $$M_y(X) = \psi(y) \ ,$$

де $\varphi(x) \neq const \ i \ \psi(y) \neq const$. Ці рівняння називаються **модельними рівняннями регресії** (від лат. «*regressio»*- рух у зворотньому напрямку) або просто рівняннями регресії відповідно Y по X та X по Y; функції $\varphi(x)$ і $\psi(y)$ - **модельними функціями регресії**, а їх графіки – **лініями регресії.**

В теорії кореляції розв'язуються дві основні задачі:

1) За даними кореляційної таблиці визначають формулу кореляційного зв'язку: або $y_x = f(x)$, або $x_y = \varphi(y)$.

2) Оцінюють силу кореляційного зв'язку, тобто обчислюють коефіцієнт кореляції.

5.1. Лінійна кореляційна залежність та пряма регресії

Для відшукання рівнянь регресії необхідно знати закон розподілу двовимірної випадкової величини (X,Y), що не завжди можливо. На практиці дослідник має лише вибірку пар значень (x_i, y_i). В цьому випадку досліджують найкращу оцінку – **вибіркову лінію регресії *Y* по *X***:

$$y_x = \varphi(x, b_0, b_1, ..., b_p),$$

де y_x - умовне середнє змінної Y при фіксованому значенню змінної $X = x$;

$b_0, b_1, ..., b_p$ - параметри кривої.

Аналогічно визначається **вибіркова лінія регресії *X* по *Y***:

$$x_y = \widehat{\psi}(y, c_0, c_1, ..., c_p),$$

де x_y - умовне середнє змінної X при фіксованому значенню змінної $Y = y$;

$c_0, c_1, ..., c_p$ - параметри кривої.

Ці рівняння називають **вибірковими рівняннями регресії** відповідно Y по X та X по Y. При вдало визначених функціях $\hat{\varphi}(x, b_0, b_1, ..., b_p)$ і $\hat{\psi}(y, c_0, c_1, ..., c_p)$ із збільшенням обсягу вибірки ($n \to \infty$) вони будуть збігатися за ймовірністю до функцій $\varphi(x)$ і $\psi(y)$.

Дані про статистичну залежність зручно задавати у вигляді **кореляційної таблиці**. Розглянемо, як приклад, таблицю залежності між добовим відвідуванням інтернет - магазину Y (тис.чол) та витратами на «агресивну» рекламу X (тис.грн) для сукупності 50 однотипних інтернет - магазинів (табл.5.1).

Таблиця 5.1

Витрати на рекламу, тис.грн (X)	Середини інтервалів	Кількість відвідувань на добу , тис.чол. (Y)					Всього n_i	Групове середнє, тис.чол $(\overline{y}_i)$
		7-11	11-15	15-19	19-23	23-27		
	y_i / x_i	9	13	17	21	25		
20-25	22,5	2	1	-	-	-	3	10,3
25-30	27,5	3	6	4	-	-	13	13,3
30-35	32,5	-	3	11	7	-	21	17,8
35-40	37,5	-	1	2	6	2	11	20,3
40-45	42,5	-	-	-	1	1	2	23,0
Всього n_i		5	11	17	14	3	50	-
Групове середнє тис.грн, $(\overline{x}_j)$		25,5	29,3	37,9	35,4	39,2	-	-

Зобразимо отриману залежність графічно точками координатної площини. Таке зображення статистичної залежності називається **полем кореляції**. Для кожного значення x_i, $(i=1,2,...,l)$, тобто для кожного рядка кореляційної таблиці обчислимо групові середні

$$\overline{y}_i = \frac{\sum_{j=1}^{m} y_i n_{ij}}{n_i},$$

де n_{ij} - частоти пар (x_i, y_j) і $n_i = \sum_{j=1}^{m} n_{ij}$; m - кількість інтервалів за змінною Y.

Обчислені групові середні $\overline{y}_i$ розташуємо в останньому стовпці кореляційної таблиці і зобразимо графічно у вигляді ламаної, що називається **емпіричною лінією регресії Y по X** (рис. 5.1).

Аналогічно, для кожного значення y_j, $(j=1,2,...,m)$ обчислимо групові середні (розміщені у нижньому рядку кореляційної таблиці):

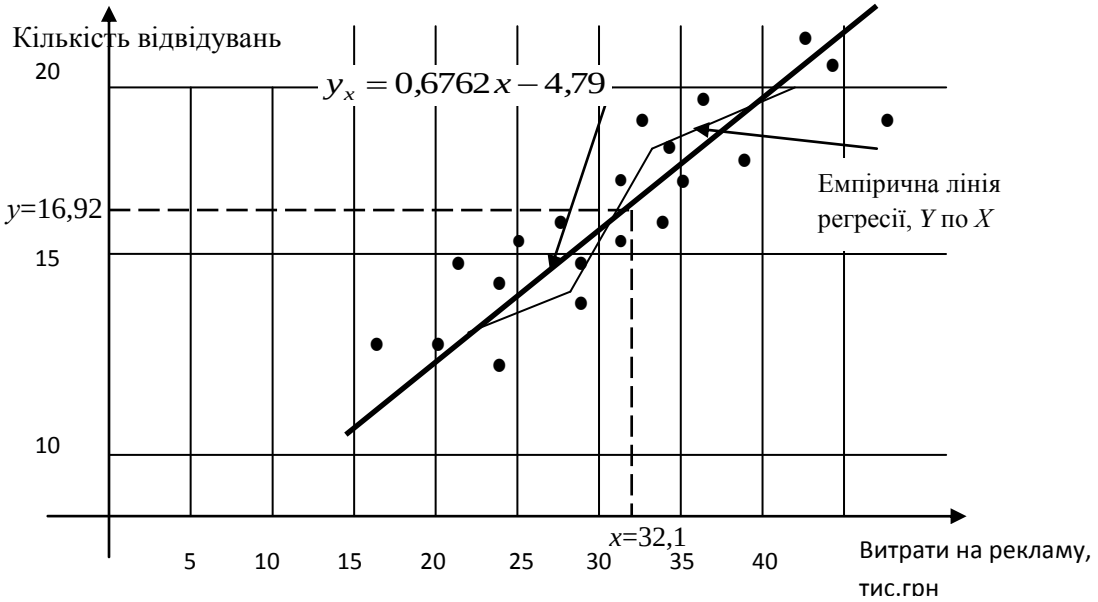

Рис. 5.1. Емпірична лінія регресії Y по X

$$\overline{x}_j = \frac{\sum_{i=1}^{l} x_i n_{ij}}{n_j},$$

де $n_j = \sum_{i=1}^{l} n_{ij}$, l - кількість інтервалів за змінною X. За виглядом ламаної лінії можна припустити наявність лінійної кореляційної залежності Y по X між двома змінними, які розглядаються. Ця залежність графічно буде більш точною, якщо збільшити обсяг вибірки:

$$n = \sum_{i=1}^{l} n_i = \sum_{j=1}^{m} n_j = \sum_{i=1}^{l} \sum_{j=1}^{m} n_{ij} .$$

Тому рівняння регресії будемо шукати у вигляді:

$$y_x = b_0 + b_1 x .$$

Параметри рівняння b_0, b_1 знайдемо за методом найменших квадратів, тобто відшукаємо значення b_0, b_1 мінімізуючи функцію

$$S = \sum_{i=1}^{l} (y_{x_i} - \overline{y}_i)^2 n_i = \sum_{i=1}^{l} (b_0 + b_1 x_i - \overline{y}_i)^2 n_i \to min$$

Необхідні умови екстремуму:

$$\begin{cases} \dfrac{\partial S}{\partial b_0} = 2\sum\limits_{i=1}^{l}(b_0 + b_1 x_i - \bar{y}_i)n_i = 0 \\ \dfrac{\partial S}{\partial b_1} = 2\sum\limits_{i=1}^{l}(b_0 + b_1 x_i - \bar{y}_i)x_i n_i = 0 \end{cases}$$

Після перетворень отримаємо систему нормальних рівнянь для визначення параметрів b_0, b_1:

$$\begin{cases} b_0 \sum\limits_{i=1}^{l} n_i + b_1 \sum\limits_{i=1}^{l} x_i n_i = \sum\limits_{i=1}^{l} \bar{y}_i n_i \\ b_0 \sum\limits_{i=1}^{l} x_i n_i + b_1 \sum\limits_{i=1}^{l} x_i^2 n_i = \sum\limits_{i=1}^{l} x_i \bar{y}_i n_i \end{cases}$$

Перетворимо вирази:

$$\sum_{i=1}^{l} \bar{y}_i n_i = \sum_{i=1}^{l} \left(\frac{\sum\limits_{j=1}^{m} y_j n_{ij}}{n_i} \right) n_i = \sum_{i=1}^{l} \sum_{j=1}^{m} y_j n_{ij} = \sum_{j=1}^{m} y_j \sum_{i=1}^{l} n_{ij} = \sum_{j=1}^{m} y_j n_j,$$

$$\sum_{i=1}^{l} x_i \bar{y}_i n_i = \sum_{i=1}^{l} x_i \left(\frac{\sum\limits_{j=1}^{m} y_j n_{ij}}{n_i} \right) n_i = \sum_{i=1}^{l} \sum_{j=1}^{m} x_i y_j n_{ij}.$$

Поділимо обидві частини нормальних рівнянь на n. Отримаємо систему у вигляді:

$$\begin{cases} b_0 + b_1 \bar{x} = \bar{y} \\ b_0 \bar{x} + b_1 \overline{x^2} = \overline{xy} \end{cases}$$

де

$$\bar{x} = \frac{\sum\limits_{i=1}^{l} x_i n_i}{n}, \quad \bar{y} = \frac{\sum\limits_{j=1}^{m} y_i n_i}{n},$$

$$\overline{xy} = \frac{\sum\limits_{i=1}^{l} \sum\limits_{j=1}^{m} x_i y_j n_i}{n},$$

$$\overline{x^2} = \frac{\sum\limits_{i=1}^{l} x_i^2 n_i}{n}.$$

Підставимо значення b_0 з першого рівняння системи в рівняння регресії:

$$y_x = \bar{y} - b_1\bar{x} + b_1 x \Leftrightarrow y_x - \bar{y} = b_1(x - \bar{x})$$

Коефіцієнт b_1 називається **вибірковим коефіцієнтом регресії *Y* по *X*** ($b_1 = b_{yx}$), отже,

$$y_x - \bar{y} = b_{yx}(x - \bar{x})$$

Коефіцієнт регресії Y по X показує, на скільки одиниць в середньому зміниться Y при збільшенні X на одну одиницю.

Розв'яжемо остаточно нормальну систему і знайдемо b_1:

$$b_1 = b_{yx} = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{s_x^2} = \frac{\mu}{s_x^2}$$

де s_x^2 - вибіркова дисперсія змінної X; μ - вибірковий кореляційний момент, або **вибіркова коваріація**.

Розмірковуючи аналогічно, з рівняння регресії X по Y $x_x - \bar{x} = b_{xy}(y - \bar{y})$ матимемо: $b_{xy} = \frac{\mu}{s_y^2}$ - вибірковий коефіцієнт регресії X по Y, який показує, на скільки одиниць в середньому зміниться X при збільшенні Y на одну одиницю. $s_y^2 = \overline{y^2} - \bar{y}^2$ - вибіркова дисперсія змінної Y. Коефіцієнти регресії b_{xy} і b_{yx} мають однакові знаки, які визначаються знаком μ. Коефіцієнти b_{yx} і $1/b_{xy}$ визначають кутові коефіцієнти відповідних ліній регресії (лінії перетинаються в точці $(\bar{x};\bar{y})$) (див. рис. 5.2).

Приклад 5.1. За даними таблиці 5.1 знайти рівняння регресії Y по X і X по Y та пояснити їх зміст.

Розв'язання. Обчислимо всі необхідні суми:

$$\sum_{i=1}^{5} x_i n_i = 22{,}5 \cdot 3 + 27{,}5 \cdot 13 + 32{,}5 \cdot 21 + 37{,}5 \cdot 11 + 42{,}5 \cdot 2 = 1605$$

$$\sum_{i=1}^{5} x_i^2 n_i = 22{,}5^2 \cdot 3 + 27{,}5^2 \cdot 13 + 32{,}5^2 \cdot 21 + 37{,}5^2 \cdot 11 + 42{,}5^2 \cdot 2 = 52612{,}5$$

$$\sum_{j=1}^{5} y_j n_j = 9 \cdot 5 + 13 \cdot 11 + 17 \cdot 17 + 21 \cdot 14 + 25 \cdot 3 = 846$$

$$\sum_{j=1}^{5} y_j^2 n_j = 9^2 \cdot 5 + 13^2 \cdot 11 + 17^2 \cdot 17 + 21^2 \cdot 14 + 25^2 \cdot 3 = 15226$$

$$\sum_{i=1}^{5}\sum_{j=1}^{5} x_i y_j n_{ij} = 22{,}5 \cdot 9 \cdot 2 + 22{,}5 \cdot 1 \cdot 13 + \ldots 42{,}5 \cdot 1 \cdot 21 + 42{,}5 \cdot 1 \cdot 25 = 27895$$

Знаходимо вибіркові характеристики і параметри рівняння регресії:

$$\bar{x} = \frac{1605}{50} = 32{,}1(\text{тис.грн});\ \bar{y} = \frac{846}{50} = 16{,}92(\text{тис.чол});$$

$$s_x^2 = \frac{52612{,}5}{50} - 32{,}1^2 = 21{,}84;\ s_y^2 = \frac{15226}{50} - 16{,}92^2 = 18{,}2336;$$

$$\mu = \frac{27895}{50} - 32{,}1 \cdot 16{,}92 = 14{,}768;$$

$$b_{yx} = \frac{14{,}768}{21{,}84} = 0{,}6762;\ b_{xy} = \frac{14{,}768}{18{,}2336} = 0{,}8099.$$

Отже, рівняння регресії:

$$y_x - 16{,}92 = 0{,}6762(x - 32{,}1) \Leftrightarrow y_x = 0{,}6762x - 4{,}79$$

$$x_y - 32{,}1 = 0{,}8099(y - 16{,}92) \Leftrightarrow x_y = 0{,}8099x + 18{,}40$$

З першого рівняння регресії Y по X випливає, що при збільшенні витрат на рекламу на 1 тис.грн. кількість відвідувачів на добу Y збільшиться в середньому на 0,6762 тис.чол. Друге рівняння регресії X по Y показує, що для збільшення кількості відвідувачів на добу на 1 тис.чол. необхідно в середньому збільшити витрати на рекламу на 0,8099 тис.грн. Зауважимо, що вільні члени рівнянь не мають реального змісту.►

5.2. Коефіцієнт кореляції

Оцінимо тісноту лінійної кореляційної залежності. Виберемо стандартну систему одиниць виміру, в якій дані за різними характеристиками виявилися б такими, що можуть бути порівняні між собою. Ця система використовує в якості одиниці виміру змінної її середнє квадратичне відхилення s. Представимо рівняння регресії $y_x - \bar{y} = b_{yx}(x - \bar{x})$ у вигляді:

$$\frac{y_x - \bar{y}}{s_x} = \left(b_{yx} \frac{s_x}{s_y} \right) \frac{x - \bar{x}}{s_x}$$

Величина $r = b_{yx} \frac{s_x}{s_y}$ показує на скільки величин s_y зміниться в середньому Y, коли X збільшиться на одне s_x.

Величина $r = b_{yx} \frac{s_x}{s_y}$ є показником тісноти лінійного зв'язку і називається вибірковим **коефіцієнтом кореляції**.

На рис. 1 наведено дві кореляційні залежності змінної Y по X: у випадку *а)* залежність між змінними менш тісна, ніж у випадку *б)*.

Якщо $r > 0$, кореляційний зв'язок називається **прямим**, а якщо $r < 0$ — **оберненим.**

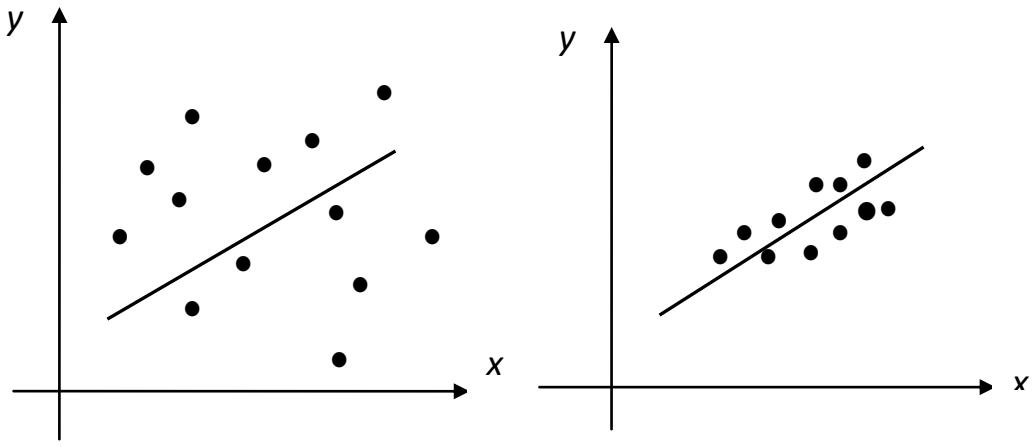

Рис. 5.2. Дві кореляційні залежності змінної Y по X:
у випадку *а)* залежність між змінними менш тісна, ніж у випадку *б)*

Формула для r симетрична відносно змінних Y та X, отже, можна записати $r = b_{xy}\frac{s_y}{s_x}$. Перемножимо обидві формули для r:

$$r^2 = b_{yx} \cdot b_{xy} \Leftrightarrow r = \pm\sqrt{b_{yx} \cdot b_{xy}}$$

Отже, коефіцієнт кореляції r змінних Y та X є середнє геометричне коефіцієнтів регресії і має їх знак.

Приклад 5.2. Обчислити коефіцієнт кореляції між кількістю коштів на рекламу X і кількістю відвідувачів на добу Y за даними табл. 5.1.

Розв'язання. В прикладі 1 були обчислені коефіцієнти регресії $b_{yx} = 0{,}6762;\ b_{xy} = 0{,}8099.$ Отже, $r = +\sqrt{0{,}6762 \cdot 0{,}8099} = 0{,}740$. Таким чином, зв'язок між змінними прямий і достатньо тісний. ►

Основні властивості коефіцієнта кореляції

(обсяг вибірки достатньо великий)

1. Коефіцієнт кореляції приймає значення на відрізку $[-1;1]$.

В залежності від того наскільки $|r|$ наближається до одиниці розрізняють зв'язок слабкий, помірний, відчутний, достатньо тісний, тісний і вельми тісний.

2. Якщо всі значення змінних збільшити (зменшити) на одне і те саме число або в одне і те саме число разів, то величина коефіцієнта кореляції не зміниться.

3. При $r = \pm 1$ кореляційний зв'язок представляє **лінійну функціональну залежність**. При цьому лінії регресії Y по X та X по Y співпадають і всі спостереженні значення розташовані на спільній прямій.

4. При $r = 0$ лінійний кореляційний зв'язок відсутній. При цьому групові середні змінних співпадають з їх загальними середніми, а лінії регресій Y по X та X по Y паралельні осям координат. Рівність $r = 0$ каже лише про відсутність лінійної кореляції, але не про відсутність взагалі кореляційної або статистичної залежності.

Основні положення кореляційного аналізу. Двовимірна модель

Кореляційний аналіз (кореляційна модель) - метод, що застосовується тоді, коли данні спостережень або досліду можна вважати випадковими і такими, що вибрані із сукупності, яка розподілена за багатовимірним нормальним законом.

Основна задача кореляційного аналізу – полягає у виявлені зв'язку між випадковими змінними шляхом точкової та інтервальної оцінок різних коефіцієнтів кореляції. Додаткова задача кореляційного аналізу – полягає в оцінці рівнянь регресії однієї змінної по другій.

Розглянемо двовимірну модель. Щільність спільного нормального розподілу двох змінних X та Y має вигляд:

$$\varphi_N(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}e^{-L(x,y)},$$

де $$L(x,y)=-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-a_x}{\sigma_x}\right)^2-2\rho\frac{x-a_x}{\sigma_x}\cdot\frac{y-a_y}{\sigma_y}+\left(\frac{y-a_y}{\sigma_y}\right)^2\right)$$

a_x, a_y - математичні сподівання змінних X та Y;

σ_x^2, σ_y^2 - дисперсії змінних X та Y;

ρ - коефіцієнт кореляції між змінними X та Y, який визначається через коваріацію K_{xy} за формулою:

$$\rho=\frac{K_{xy}}{\sigma_x\sigma_y}=\frac{M(XY)-a_xa_y}{\sigma_x\sigma_y}.$$

Величина ρ характеризує тісноту зв'язку між випадковими змінними X та Y лише у випадку лінійного зв'язку. Ступінь розсіювання значень змінної відносно лінії регресії визначається двома факторами: дисперсією і коефіцієнтом кореляції і не залежить від значення змінної.

Генеральна сукупність певною мірою аналогічна поняттю випадкової величини і її закону розподілу, тому для параметрів $a_x, a_y, \sigma_x^2, \sigma_y^2, \rho$ використовується

інша термінологія: a_x, a_y (або $\bar{x}_0, \bar{y}_0$) – генеральні середні; σ_x^2, σ_y^2 - генеральні дисперсії; K_{xy} і ρ - генеральні коваріація і коефіцієнт кореляції.

Для оцінки генерального коефіцієнта кореляції ρ і модельних рівнянь регресії за вибіркою необхідно замінити параметри $a_x, a_y, \sigma_x^2, \sigma_y^2, K_{xy}$ їх грунтовними оцінками:

a_x, a_y заміняємо на $\bar{x} = \frac{\sum_{i=1}^{l} x_i n_i}{n}$, $\bar{y} = \frac{\sum_{j=1}^{m} y_i n_i}{n}$;

σ_x^2, σ_y^2 заміняємо на $s_x^2 = \overline{x^2} - \bar{x}^2$, $s_y^2 = \overline{y^2} - \bar{y}^2$;

K_{xy} заміняємо на $\mu = \overline{xy} - \bar{x} \cdot \bar{y}$.

Перевірка значущості та інтервальна оцінка параметрів зв'язку

В практичних дослідженнях про тісноту кореляційної залежності судять по значенню вибіркового коефіцієнта кореляції r. Оцінка r величина випадкова. Нехай $r \neq 0$. В цьому випадку перевіряється гіпотеза H_0 про відсутність лінійного кореляційного зв'язку між змінними в генеральній сукупності. Якщо ця гіпотеза справедлива, то статистика $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ має t - розподіл Стьюдента з $k = n-2$ ступенями свободи. Гіпотеза H_0 відкидається, якщо $|t| = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}} > t_{1-\alpha,k}$, де $t_{1-\alpha,k}$ - табличне значення t - критерію Стьюдента, який визначається на рівні значущості α при кількості ступенів свободи $k = n-2$.

Приклад 5.3. Перевірити на рівні α=0,05 значущість коефіцієнта кореляції між змінними X та Y за даними табл. 1.

Розв'язання. В прикладі 1 обчислено коефіцієнт кореляції $r = 0{,}740$. Статистика критерію $t = \dfrac{0{,}740\sqrt{50-2}}{\sqrt{1-0{,}740^2}} = 7{,}62$. Для рівня значущості α=0,05 і кількості ступенів свободи $k = 50-2 = 48$ знаходимо критичне значення статистики $t_{0{,}95;48} = 2{,}01$. Оскільки $t > t_{0{,}95;48}$, то коефіцієнт кореляції між кількістю відвідувань сайту на добу Y і витратами на рекламу X значно відмінний від нуля.►

Для значущого коефіцієнта кореляції r доцільно знайти довірчий інтервал, який із заданою надійністю $\gamma = 1-\alpha$ накриває невідомий генеральний коефіцієнт кореляції ρ. Для побудови такого інтервалу використовують спеціально підібрані функції від r, які збігаються до добре відомих розподілів. Найчастіше використовують Z **- перетворення Фішера:**

$$z = \frac{1}{2} ln \frac{1+r}{1-r}$$

Розподіл z вже при невеликих n є приблизно нормальним з математичним сподіванням $M(z) = \dfrac{1}{2} ln \dfrac{1+\rho}{1-\rho} + \dfrac{\rho}{2(n-1)}$ і дисперсією $\sigma_z^2 = \dfrac{1}{n-3}$.

Спочатку будується довірчий інтервал для $M(z)$:

$$z - t_{1-\alpha} \frac{1}{\sqrt{n-3}} \le M(z) \le z + t_{1-\alpha} \frac{1}{\sqrt{n-3}},$$

де $t_{1-\alpha}$ - нормоване відхилення z, що визначається за допомогою функції Лапласа: $\Phi(t_{1-\alpha}) = \gamma = 1-\alpha$.

Для визначення границь довірчого інтервалу для ρ існують спеціальні таблиці. За їх відсутності користуються формулою $r = thz = \dfrac{e^z - e^{-z}}{e^z + e^{-z}}$.

Якщо коефіцієнт кореляції значущий, то коефіцієнти регресії b_{yx} і b_{xy} також значно відрізняються від нуля, а інтервальні оцінки для відповідних

генеральних коефіцієнтів регресії β_{yx} і β_{xy} можуть бути отримані за формулами, що спираються на те, що статистики $\left(b_{yx}-\beta_{yx}\right)/s_{b_{yx}}$, $\left(b_{xy}-\beta_{xy}\right)/s_{b_{xy}}$ мають t-розподіл Стьюдента з $\left(n-2\right)$ ступенями свободи:

$$b_{yx}-t_{1-\alpha;n-2}\frac{s_y\sqrt{1-r^2}}{s_x\sqrt{n-2}}\le\beta_{yx}\le b_{yx}+t_{1-\alpha;n-2}\frac{s_y\sqrt{1-r^2}}{s_x\sqrt{n-2}}$$

$$b_{xy}-t_{1-\alpha;n-2}\frac{s_x\sqrt{1-r^2}}{s_y\sqrt{n-2}}\le\beta_{xy}\le b_{xy}+t_{1-\alpha;n-2}\frac{s_x\sqrt{1-r^2}}{s_y\sqrt{n-2}}$$

z - перетворення Фішера може бути застосовано при перевірці різних гіпотез відносно коефіцієнта кореляції. Наприклад, для перевірки значущості розбіжностей двох коефіцієнтів кореляції r_1 і r_2, отриманих за вибірками обсягів n_1 і n_2 для перевірки нульової гіпотези $H_0:\rho_1=\rho_2$

застосовується статистика $t=\dfrac{z(r_1)-z(r_2)}{\sqrt{\dfrac{1}{n_1-3}+\dfrac{1}{n_2-3}}}$.

Приклад 5.4. За даними таблиці 1 знайти з надійністю 0,95 інтервальні оцінки (довірчі інтервали) параметрів зв'язку між кількістю відвідувачів сайту Y і витратами на рекламу X.

Розв'язання . Оскільки коефіцієнт кореляції X по Y значущий (див. приклад 2), то побудуємо довірчий інтервал для генерального коефіцієнта кореляції ρ, застосовуючи z - перетворення Фішера:

$z=\frac{1}{2}ln\frac{1+0{,}740}{1-0{,}740}=0{,}9505$. За таблицею функцій Лапласа і за умови $\Phi(\mathrm{t}_{1-\alpha})=0{,}95$, знаходимо $t_{0{,}05}=1{,}96$. Побудуємо довірчий інтервал для $M(z)$: $0{,}9505-1{,}96\frac{1}{\sqrt{50-2}}\le M(z)\le 0{,}9505+1{,}96\frac{1}{\sqrt{50-2}}$, або $0{,}6646\le M(z)\le 1{,}2364$. Знаходимо границі довірчого інтервалу для ρ, використовуючи спеціальну таблицю чи формулу: $0{,}581\le\rho\le 0{,}844$.

Генеральний коефіцієнт кореляції ρ на рівні значущості 0,05 (з надійністю 0,95) накривається знайденим інтервалом.

Тепер побудуємо довірчі інтервали для генеральних коефіцієнтів регресії β_{yx} і β_{xy}. Спочатку визначимо середнє квадратичне відхилення змінних:

$$s_x = \sqrt{s_x^2} = \sqrt{21{,}84} = 4{,}673;\ s_y = \sqrt{s_y^2} = \sqrt{18{,}2336} = 4{,}270.$$

$$0{,}6762 - 2{,}01\frac{4{,}270\sqrt{1-0{,}74^2}}{4{,}673\sqrt{50-2}} \le \beta_{yx} \le 0{,}6762 + 2{,}01\frac{4{,}270\sqrt{1-0{,}74^2}}{4{,}673\sqrt{50-2}}.$$

Або $0{,}4979 \le \beta_{yx} \le 0{,}8545$.

Аналогічно : $0{,}5963 \le \beta_{xy} \le 1{,}0235$ ►.

При змістовній інтерпретації параметрів ρ, β_{yx}, β_{xy} слід врахувати в першу чергу їх інтервальні (а не тільки точкові) оцінки.

Приклад 5.5. При дослідженні зв'язку між продуктивністю праці ІТ-фахівців і рівнем фінансування їх роботи на підприємствах однієї галузі промисловості, які розташовані в двох різних районах держави, обчислені коефіцієнти кореляції $r_1 = 0{,}95$ і $r_2 = 0{,}88$ за вибірками обсягів відповідно $n_1 = 14$ і $n_2 = 20$. З'ясувати, чи є на рівні значущості $\alpha = 0{,}05$ суттєві розбіжності в тісноті зв'язку між змінними, що розглядаються на підприємствах галузі в цих районах.

Розв'язання. Гіпотеза, що перевіряється $H_0 : \rho_1 = \rho_2$. Альтернативна гіпотеза $H_1 : \rho_1 \ne \rho_2$. Статистика обчислюється за формулою:

$$t = \frac{z(0{,}95) - z(0{,}88)}{\sqrt{\frac{1}{14-3} + \frac{1}{20-3}}} = \frac{1{,}832 - 1{,}376}{\sqrt{0{,}15}} = 1{,}18.\ \ t < t_{0{,}95} = 1{,}96.$$ Отже, гіпотеза $H_0 : \rho_1 = \rho_2$ не відкидається, тобто немає підстав вважати розбіжності суттєвими.►

Кореляційне відношення і індекс кореляції

Введений вище коефіцієнт кореляції є повноцінним показником тісноти зв'язку лише у випадку лінійної залежності між змінними. Але часто виникає необхідність у достовірному показнику інтенсивності зв'язку при будь-якій формі залежності. За правилом додавання дисперсій:

$$s_y^2 = s_{iy}'^2 + \delta_{iy}^2 ,$$

де $s_y^2 = \frac{\sum_{j=1}^{m}\left(y_j - \bar{y}\right)^2 n_i}{n}$ — загальна дисперсія змінної;

$s_{iy}'^2 = \frac{\sum_{i=1}^{l} s_{iy}^2 n_i}{n}$ — середнє групових дисперсій $s_{iy}^2 = \frac{\sum_{j=1}^{m}\left(y_j - \bar{y}_i\right)^2}{n}$ (або залишкова дисперсія);

$\delta_{iy}^2 = \frac{\sum_{i=1}^{l}(\bar{y}_i - \bar{y})^2 n_i}{n}$ — міжгрупова дисперсія.

Залишковою дисперсією вимірюють ту частину варіації Y, яка виникає через змінність неврахованих факторів, незалежних від X. Міжгрупова дисперсія виражає ту частину варіацій Y, яка обумовлена змінністю X. Величина $\eta_{yx} = \sqrt{\frac{\delta_{iy}^2}{s_y^2}}$ дістала назву **емпіричного кореляційного відношення** $\boldsymbol{Y}$ по $\boldsymbol{X}$.

Чим тісніший зв'язок , тим більший вплив на варіацію змінної Y має змінність X у порівнянні із не врахованими факторами, тим вище η_{yx}. Величина η_{yx}^2 – **емпіричний коефіцієнт детермінації**, показує, яка частина загальної варіації Y обумовлена варіацією X. Аналогічно вводиться емпіричне кореляційне відношення X по Y: $\eta_{xy} = \sqrt{\frac{\delta_{ix}^2}{s_x^2}}$.

Основні властивості кореляційних відношень

(при досить великому обсягу вибірки n)

1. Кореляційне відношення – це невід'ємна величина, не більша за 1: $0 \le \eta \le 1$.
2. Якщо $\eta = 0$, то кореляційний зв'язок відсутній.
3. Якщо $\eta = 1$, то між змінними існує функціональна залежність.
4. $\eta_{yx} \neq \eta_{xy}$, тобто на відміну від коефіцієнта кореляції r (для якого $r_{yx} = r_{xy} = r$) при обчисленні кореляційного відношення важливо, яку змінну вважати незалежною, а яку – залежною.

Емпіричне кореляційне відношення η_{yx} є показником розсіювання точок кореляційного поля відносно емпіричної лінії регресії , що відображається ламаною, яка об'єднує значення $\overline{y}_i$. Проте, у зв'язку з тим, що закономірна зміна $\overline{y}_i$ порушується випадковими зигзагами ламаної, які виникають як наслідок остаточної дії неврахованих факторів, η_{yx} перебільшує тісноту зв'язку. Тому, разом із η_{yx} , розглядається показник тісноти зв'язку R_{yx}, що характеризує розсіювання точок кореляційного поля відносно лінії регресії y_x. Показник R_{yx} дістав назву **теоретичного кореляційного відношення або індексу кореляції *Y* по *X*:**

$$R_{yx} = \sqrt{\frac{\delta_y^2}{s_y^2}} = \sqrt{1 - \frac{s_y'^2}{s_y^2}},$$

де δ_y^2 і $s_y'^2$ - дисперсії, в яких групові середні $\overline{y}_i$ замінені умовними середніми $\overline{y}_{x_i}$, обчисленими за рівнянням регресії.

Аналогічно обчислюється **індекс кореляції** *X* по *Y:*

$$R_{xy} = \sqrt{\frac{\delta_x^2}{s_x^2}} = \sqrt{1 - \frac{s_x'^2}{s_x^2}}$$

Перевагою показників η і R є те, що вони можуть бути обчислені при будь-якій формі зв'язку між змінними. Хоча η і підвищує тісноту зв'язку у порівнянні з R, та для його обчислення не потрібно знати рівняння регресії. Кореляційні відношення η і R зв'язані з коефіцієнтом кореляції r наступним чином: $0 \le |r| \le R \le \eta \le 1$. Можна показати, що у випадку лінійної моделі $y_x - \bar{y} = b_{yx}(x - \bar{x})$, індекс кореляції R_{yx} дорівнює за абсолютною величиною коефіцієнту кореляції *r*. Розбіжність між η^2 і R^2 (чи r^2) може бути використана для перевірки лінійності кореляційної залежності. Перевірка значущості кореляційного відношення η

грунтується на тому, що статистика $F = \dfrac{\eta^2(n-m)}{(1-\eta^2)(m-1)}$

(де m - кількість інтервалів групування) має F - розподіл Фішера – Снедекора з $k_1 = m-1$ *і* $k_2 = n-m$ ступенями свободи. Тому η значно відрізняється від нуля, якщо $F > F_{\alpha;k_1;k_2}$, де $F_{\alpha;k_1;k_2}$ — табличне значення F - критерію на рівні значущості α при числі ступенів свободи $k_1 = m-1$ і $k_2 = n-m$.

Індекс кореляції R двох змінних є значущим, якщо величина статистики $F = \dfrac{R^2(n-2)}{1-R^2}$ більше табличного значення $F_{\alpha;k_1;k_2}$, де $k_1 = 1$ *і* $k_2 = n-2$.

Приклад 5.6. За даними табл. 5.1 обчислити кореляційне відношення η_{yx} і індекс кореляції R_{yx} і перевірити їх значущість на рівні $\alpha = 0{,}05$.

Розв'язання. Визначимо η_{yx}. В прикладі 1 обчислені: загальне середнє $\bar{y} = 16{,}92$; дисперсія $s_y^2 \approx 18{,}23$; групові середні в табл. 1. Частоти інтервалів також знаходяться в таблиці. Розрахунки представимо у вигляді таблиці 2.

$\delta_{iy}^2 = 517,8/50 = 10,36$. $\eta_{yx} = \sqrt{\dfrac{10,36}{18,23}} = 0,754$. Значення η_{yx} близьке до значення $r = 0,740$. Тому припущення про лінійний зв'язок є обгрунтованим.

Для обчислення R_{yx} за рівнянням регресії $y_x = 0,6762x - 4,79$ знаходимо значення y_{x_i} (див. табл. 6.2), $\delta_y^2 = 502,8/50 = 10,04$ і

$$R_{yx} = \sqrt{\frac{10,04}{18,23}} = 0,742.$$

Таблиця 5.2

x_i	n_i	$\bar{y}_i$	$(\bar{y}_i - \bar{y})^2 n_i$	y_{x_i}	$(\bar{y}_{x_i} - \bar{y})^2 n_i$
22,5	3	10,3	131,5	10,4	127,5
27,5	13	13,3	170,4	13,8	126,5
32,5	21	17,8	16,3	17,2	1,6
37,5	11	20,3	125,7	20,6	149,0
42,5	2	23,0	73,9	23,9	97,4
Σ			517,8	-	502,0

Бачимо, що $R_{yx} = r$ (розбіжності викликані правилами округлення при обчисленнях). Тому, у випадку лінійного зв'язку, достатньо обчислити лише r. Величина коефіцієнта детермінації $R_{yx}^2 = 0,551$ показує, що варіація залежної змінної Y на 55,1% пояснюється варіацією незалежної змінної X.

Для перевірки значущості η_{yx} (кількість інтервалів групування $m = 5$) знайдемо $F = \dfrac{0,754^2(50-5)}{(1-0,754^2)(5-1)} = 14,82$. Табличне значення $F_{0,05;4;45} = 2,57$. Оскільки $F > F_{0,05;4;45}$, то η_{yx} значно відрізняється від нуля. Аналогічно перевіряємо значущість R_{yx}:

$$F = \frac{0,742^2(50-2)}{1-0,742^2} = 58,8 \; F > F_{0,05;1;48} = 4,04.$$

Отже, індекс кореляції є значущим.►