

ТЕМА: КОРЕЛЯЦІЙНА АНАЛІЗА

1.1. Загальні означення

1. Кореляція (пізньолатинське *correlatio* – співвідношення) – взаємозв'язок, взаємозалежність, взаємовідповідність; співвідношення понять, функцій, предметів, підприємств; почуттів, поведінки людей тощо. Вперше (у 1800 – 1805 р. р.) у науковий обіг було введено означення «кореляція» Жоржем Кув'є.

2. Існує цілий клас завдань, пов'язаних із аналізою зв'язку між двома ($x \sim y$) чи кількома змінними ($x_1, x_2, \dots \sim y$), або $x_1, x_2, \dots \sim y_1, y_2, \dots$.

3. Розрізняють два види математичних залежностей (зв'язків) однієї величини від іншої або кількох інших величин:

а) **детермінована** (функціональна; не випадкова; жорстка; чиста; строга; математична; регулярна залежність тощо), коли кожному значенню змінної (аргументу (x)) або кільком залежним величинам (аргументам $x_1, x_2, \dots, x_j, \dots, x_n$) відповідає певне значення іншої величини (функції (y)) або певним значенням інших величин.

Такий зв'язок:

$$y = f(x) \quad \text{або} \quad y = f(x_1, x_2, \dots, x_n) \quad (1.1)$$

носить назву функціонального, $\dots$,

Так, величини $s = \frac{gt^2}{2}$, де $g = \text{const}$, кожному значенню аргумента (t) відповідає певне детерміноване значення (S) функції;

б) **випадкова**, коли між випадковими величинами такого зв'язку немає. Залежність між випадковими величинами зветься схоластичною (випадковою, ймовірнісною, недетермінованою, нежорсткою, нестрогою, нерегулярною, кореляційною тощо), коли кожному значенню аргумента ($\alpha_x, \sigma_x^2, \sigma_x, \dots$) відповідає певний закон розподілу функції, тобто, коли на зміну однієї (або кількох) випадкових величин інша випадкова величина реагує

змінною параметрів свого розподілу ($a_y, \sigma_y^2, \sigma_y, \dots$). Такий зв'язок носить назву стохастичного, ймовірного, кореляційного, . . .

4. Чому існують стохастичні зв'язки? Окрім впливу основних факторів та їх рівнів, є:

- а) вплив випадкових неконтрольованих факторів та їх рівнів на об'єкт дослідження або природний об'єкт спостереження;
- б) вплив випадкових контрольованих але некерованих факторів;
- в) окрім того неможливо, практично не вдається зафіксувати рівні факторів, вони «плавають» біля певної середньої.

Все це впливає на функцію відгуку, і вона теж має випадкові значення.

5. Стохастичний зв'язок має дві складові:

- 1) власне стохастичну складову St , яка пов'язана із залежністю випадкових величин y і визначається дією залежних факторів;
- 2) власне випадкову складову Ra , яка обумовлена дією індивідуальних випадкових факторів, які впливають на одну із випадкових величин.

6. а) якщо $St \neq 0, Ra = 0$ – функціональний зв'язок;

б) якщо $St = 0, Ra \neq 0$ (чи і $Ra = 0$) – незалежні величини;

в) якщо $St \neq 0, Ra \neq 0$ – стохастичний (кореляційний) зв'язок;

г) якщо одна величина y (змінна a_y і σ_y^2) не реагує на зміну a_x , а реагує лише на зміну σ_x^2 , то такі величини є некеровані.

7. Від співвідношення St / Ra залежить сила щільності зв'язку і це є предметом кореляційної аналізи.

Тому, завдання кореляційної аналізи (разом із регресійною аналізою) (регресія від латинського *regressio* – зведення, звертання, зворочення від латинського *regressus* – зворотній рух, відход) полягає у тому, щоби встановити якісні показники математичної моделі:

- а) на скільки модель краще, добротніше (адекватно) описує залежність $y \sim f(x)$, [чи $y = f(x_1, x_2, \dots, x_i, \dots, x_n)$ тощо]. Лінійна (1) чи нелінійна (2) залежність (рис.1.1).

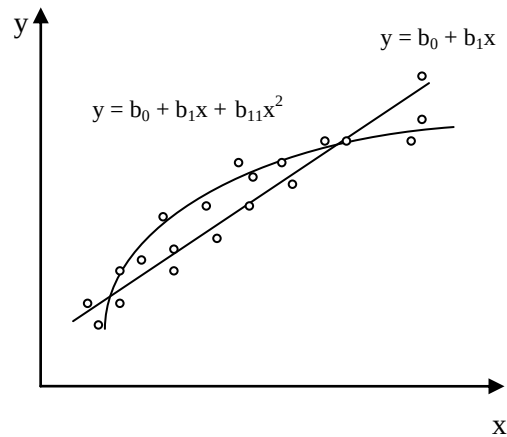


Рис.1.1. Апроксимація експериментальних даних лінійною (1) та нелінійною (2) функціями.

б) Яка щільність зв'язку між x та y . (рис.1.2)

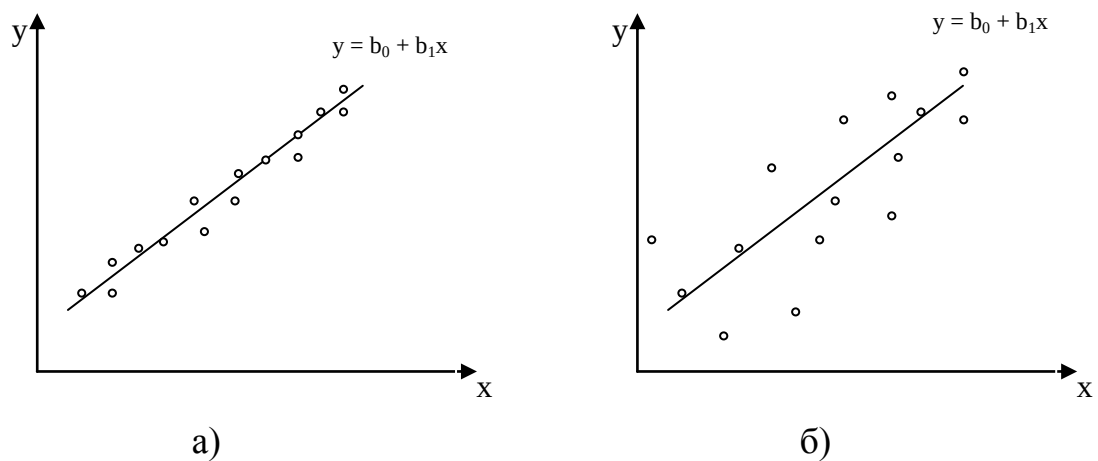


Рис. 1.2. Сильний (а) та слабкий (б) кореляційний лінійний зв'язок $y = f(x)$.

8. Тобто виникає необхідність виразити щільність (силу зв'язку) певним числом:

- а) для лінійного зв'язку – коефіцієнт кореляції;
- б) для нелінійного зв'язку – кореляційне співвідношення (індекс кореляції).

1.2. Лінійний зв'язок, коефіцієнт кореляції

1. Нехай маємо дві генеральні сукупності двох випадкових величин. Обидві величини мають нормальний закон розподілу (табл. 1.1).

Таблиця 1.1

Параметри генеральних сукупностей величин

Сукупність величин		Параметри
$X:$	$x_1, x_2, x_3, \dots, x_i, \dots, x_n$	$N \rightarrow \infty, \mu_x, \sigma_x^2, \sigma_x, \dots$
$Y:$	$y_1, y_2, y_3, \dots, y_i, \dots, y_n$	$N \rightarrow \infty, \mu_y, \sigma_y^2, \sigma_y, \dots$

2. Кількісною оцінкою щільності лінійного зв'язку між величинами X і Y двох генеральних сукупностей буде – генеральний коефіцієнт кореляції:

$$\rho_{x,y} = \rho_{1,2} = \rho = \frac{M_{1/1}}{\sigma_x \cdot \sigma_y}, \quad (1.2)$$

$$\text{де} \quad M_{1/1} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - a_x)(y - a_y) \varphi(x, y) dx dy; \quad (1.3)$$

– генеральний змішаний центральний момент другого порядку;

$$a_x = \int_{-\infty}^{\infty} x \varphi(x) dx \quad - \quad (1.4)$$

– математичне сподівання випадкової величини X (генеральна середня);

$$a_y = \int_{-\infty}^{\infty} y \varphi(y) dy \quad - \quad (1.5)$$

– математичне сподівання випадкової величини Y (генеральна середня);

$\varphi(x), \varphi(y)$ – щільність ймовірності розподілу випадкової величини X, Y відповідно;

$\varphi(x, y)$ – щільність ймовірностей двовимірного сумісного розподілу випадкових величин X, Y ;

$$\sigma_x = \sqrt{\sigma_x^2}, \quad \sigma_y = \sqrt{\sigma_y^2} \quad - \quad (1.6)$$

– генеральні середньо квадратичні відхилення.

Генеральні дисперсії:

$$\sigma_x^2 = \int_{-\infty}^{\infty} (x - a_x)^2 \varphi(x) dx ; \quad (1.7)$$

$$\sigma_y^2 = \int_{-\infty}^{\infty} (y - a_y)^2 \varphi(y) dy . \quad (1.8)$$

Коефіцієнт кореляції змінюється у межах:

$$0 \leq |\rho| \leq 1 , \quad (1.9)$$

$$\text{або } 0\% \leq |\rho| \leq 100\% ; \quad (1.10)$$

$$\rho: -1 \dots 0 \dots +1; \quad (1.11)$$

$$\rho: -100\% \dots 0 \dots +100\% . \quad (1.12)$$

Якщо:

1) $\rho=0$, то x, y незалежні випадкові величини, тому такі величини некорельовані;

2) $\rho \rightarrow 0$, для нелінійних випадкових величин;

3) $|\rho|=1$ (-1, або +1), зв'язок функціональний;

4) $0 < |\rho| < 1$, стохастичний кореляційний зв'язок;

5) якщо $\rho > 0$, то при $x \uparrow \rightarrow y \uparrow$ або при $x \downarrow \rightarrow y \downarrow$;

6) якщо $\rho < 0$, то при $x \uparrow \rightarrow y \downarrow$ або при $x \downarrow \rightarrow y \uparrow$.

3. Розрахунок вибіркового коефіцієнта кореляції.

Позначимо вибіркового коефіцієнта кореляції так: $r = r_{1,2} = r_{x,y}$.

Оцінками генеральних показників є вибіркові:

вибіркова сукупність $\xrightarrow[\text{ОЦІНКА}]{} \text{генеральна сукупність}$

$$\overline{X} \xrightarrow[\text{ОЦІНКА}]{} \mu_x$$

$$\overline{y} \xrightarrow[\text{ОЦІНКА}]{} \mu_y$$

$$S_x^2 \xrightarrow[\text{ОЦІНКА}]{} \sigma_x^2$$

$$S_y^2 \xrightarrow[\text{ОЦІНКА}]{} \sigma_y^2$$

$$S_x \xrightarrow[\text{ОЦІНКА}]{} S_x^2$$

$$S_y \xrightarrow[\text{ОЦІНКА}]{} S_y^2$$

$$\begin{array}{ccc}
 \dots\dots\dots & & \dots\dots\dots \\
 \dots\dots\dots & & \dots\dots\dots \\
 r_{x,y} & \xrightarrow{\text{ОЦІНКА}} & \rho_{x,y}
 \end{array}$$

Розраховують вибірковий коефіцієнт кореляції за формулою:

$$r_{x,y} = \frac{m_{1/1}}{S_x \cdot S_y}, \quad (1.13)$$

$$\begin{aligned}
 \text{де } m_{1/1} &= \frac{1}{N-1} \sum_{i=1}^N [(x_i - \bar{x})(y_i - \bar{y})] = \frac{1}{N-1} \left[\sum_{i=1}^N (x_i \cdot y_i) - \frac{1}{N} \sum_{i=1}^N x_i \sum_{i=1}^N y_i \right] = \\
 &= \frac{1}{N-1} \left[\sum_{i=1}^N (x_i \cdot y_i) - N \bar{x} \bar{y} \right] -
 \end{aligned} \quad (1.14)$$

– вибірковий змішаний центральний момент другого порядку.

Вибіркові середньо квадратичні відхилення:

$$\begin{aligned}
 S_x = \sqrt{S_x^2} &= \sqrt{\frac{1}{N-1} \left[\sum_{i=1}^N (x_i - \bar{x})^2 \right]} = \sqrt{\frac{1}{N-1} \left[\sum_{i=1}^N x_i^2 - \frac{1}{N} \left(\sum_{i=1}^N x_i \right)^2 \right]} = \\
 &= \sqrt{\frac{1}{N-1} \left[\sum_{i=1}^N x_i^2 - N (\bar{x})^2 \right]};
 \end{aligned} \quad (1.15)$$

$$\begin{aligned}
 S_y = \sqrt{S_y^2} &= \sqrt{\frac{1}{N-1} \left[\sum_{i=1}^N (y_i - \bar{y})^2 \right]} = \sqrt{\frac{1}{N-1} \left[\sum_{i=1}^N y_i^2 - \frac{1}{N} \left(\sum_{i=1}^N y_i \right)^2 \right]} = \\
 &= \sqrt{\frac{1}{N-1} \left[\sum_{i=1}^N y_i^2 - N (\bar{y})^2 \right]}.
 \end{aligned} \quad (1.16)$$

Тоді:

$$r_{x,y} = \frac{\sum_{i=1}^N [(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}} = \frac{\sum_{i=1}^N (x_i y_i) - \frac{1}{N} \sum_{i=1}^N x_i \sum_{i=1}^N y_i}{\sqrt{\left[\sum_{i=1}^N x_i^2 - \frac{1}{N} \left(\sum_{i=1}^N x_i \right)^2 \right]} \cdot \sqrt{\left[\sum_{i=1}^N y_i^2 - \frac{1}{N} \left(\sum_{i=1}^N y_i \right)^2 \right]}} =$$

$$= \frac{\sum_{i=1}^N (x_i y_i) - N (\bar{x} \bar{y})}{\sqrt{\left[\sum_{i=1}^N x_i^2 - N (\bar{x})^2 \right] \cdot \left[\sum_{i=1}^N y_i^2 - N (\bar{y})^2 \right]}} \quad (1.17)$$

1.3. Значущість та довірчі інтервали для коефіцієнта кореляції

1. Вибірковий коефіцієнт кореляції $r_{x,y}$ як випадкова величина може приймати різні значення при повторних дослідженнях (або спостереженнях), тобто при «витягуванні» вибірки з генеральної сукупності.

Тоді будемо мати різні оцінки генерального коефіцієнта кореляції:

2. Навіть, коли $\rho_{x,y} = 0$ (немає лінійного кореляційного зв'язку) $r_{x,y} \neq 0$.

Тоді необхідно перевірити нульову гіпотезу :

$$\left. \begin{array}{l} H_0: \rho = 0 \\ \uparrow \text{ оцінка} \\ |r_{x,y}| \neq 0 \end{array} \right\} \quad (\text{немає лінійного кореляційного зв'язку})$$

Альтернативна гіпотеза:

$$\left. \begin{array}{l} H_1: \rho \neq 0 \\ \uparrow \text{ оцінка} \\ |r_{x,y}| \neq 0 \end{array} \right\} \quad (\text{є лінійний кореляційний зв'язок})$$

3. Перевірка значущості коефіцієнта кореляції за критерієм Стьюдента.

При великому обсязі виборки, наприклад $N > 100$ і $|\rho| \neq 1$ (тобто відсутності функціонального зв'язку) розподіл вибіркового коефіцієнта кореляції $r_{x,y}$ підпорядкований нормальному закону розподілу (Н.З.Р.) з параметрами :

$$\left. \begin{array}{l} a_r = \rho_{x,y} \\ \sigma_r = \frac{\sqrt{1 - \rho_{x,y}^2}}{\sqrt{N - 1}} \end{array} \right\} \quad (1.18)$$

Так як при $N \rightarrow \infty$ (або $N \rightarrow \max$) $\sigma_r \approx S_r$, то

$$S_r = \frac{\sqrt{1 - r^2}}{\sqrt{N - 1}} \quad (1.19)$$

Пронормуємо r : введемо нову випадкову змінну Z

$$Z = \frac{r - a_r}{\sigma_r} = \frac{r - \rho}{\sqrt{1 - \rho^2}} \sqrt{N - 1}, \quad (1.20)$$

Тоді величину Z розподілено за нормальним законом розподілу з параметрами:

$$\left. \begin{aligned} a_z &= 0 \\ \sigma_z^2 &= 1, \sigma_z = 1 \end{aligned} \right\} \quad (1.21)$$

Якщо $\sigma_r \approx S_r$, тоді (19) зводиться, що Z підпорядковано розподілу Стьюдента і перевірка $H_0: \rho = 0$ зводиться до розрахунку:

$$t_{\text{ПОЗР}} = \frac{r}{\sqrt{1 - r^2}} \sqrt{N - 2} \quad \text{і порівняння } |t_{\text{ПОЗР}}| \text{ з табличним значенням критерія}$$

Стьюдента: $t_{\text{ТАБ}}\{\alpha; f = N - 2\}$.

Якщо: $|t_{\text{ПОЗР}}| \leq t_{\text{ТАБ}}$, то з ймовірністю $p = 1 - \alpha$ приймаємо $H_0: \rho = 0$ (немає лінійного зв'язку):

Тут оцінка ступеня нелінійності:

$$\xi_2(t) = \frac{t_{\text{ТАБ}}}{|t_{\text{ПОЗР}}|} \quad (1.22)$$

Якщо $|t_{\text{ПОЗР}}| > t_{\text{ТАБ}}$, то з ймовірністю $p = 1 - \alpha$ нульова гіпотеза $H_0: \rho = 0$ відкидається, приймається альтернативна гіпотеза $H_1: \rho \neq 0$ (є щільний лінійний зв'язок):

Тут оцінка ступеня лінійності:

$$\xi_1(t) = \frac{|t_{\text{ПОЗР}}|}{t_{\text{ТАБ}}}. \quad (1.23)$$

Коефіцієнт кореляції значущий.

Для довірчої ймовірності $p = 1 - \alpha$ довірчий інтервал генерального коефіцієнта кореляції становить:

$$p \{ [r - (t_{TAB} S_r)] \leq \rho < [r + (t_{TAB} S_r)] \} = 1 - \alpha ,$$

де $t_{TAB} \{ \alpha; f = N - 2 \}$; рівень значущості : $\alpha = 1 - p$.

4. При малому обсязі вибірки (обмеженому) і порівняно високої лінійної кореляції: ($N < 60 - 100$) розподіл вибіркового коефіцієнта кореляції r суттєво відрізняється від нормального закону розподілу (рис. 1.3)

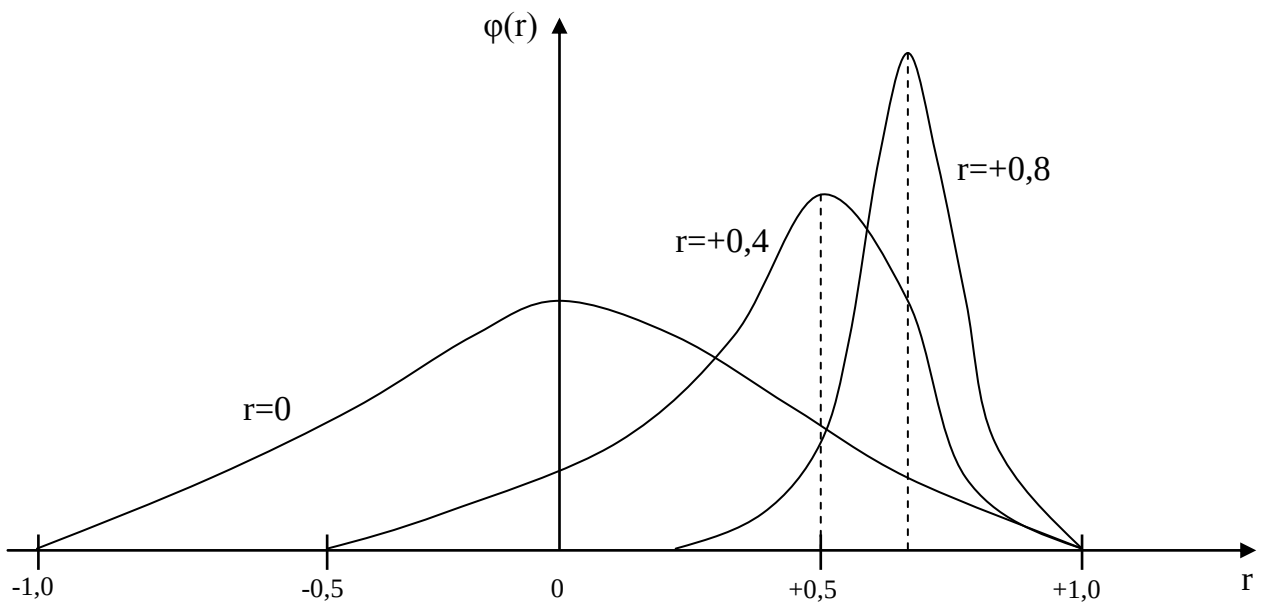


Рис. 1.3. Розподіл вибіркового коефіцієнта кореляції.

Використаємо перетворення Фішера:

$$Z = \frac{1}{2} \ln \frac{1+r}{1-r}, \quad (1.24)$$

де r – вибіровий коефіцієнт кореляції.

Випадкова величина Z буде розподілена за нормальним законом (рис. 1.4) з параметрами:

$$\mu_z = \frac{1}{2} \ln \frac{1+\rho}{1-\rho} + \frac{\rho}{2(N-2)} \quad (1.25)$$

$$\sigma_z = \frac{1}{\sqrt{N-3}} \quad (1.26)$$

Так як $\left[\frac{\rho}{2(N-2)} \ll \sigma_z \right]$, то в (1.3) цим числом можна знехтувати тоді:

$$\mu_z \approx \frac{1}{2} \ln \frac{1+\rho}{1-\rho} \quad (1.27)$$

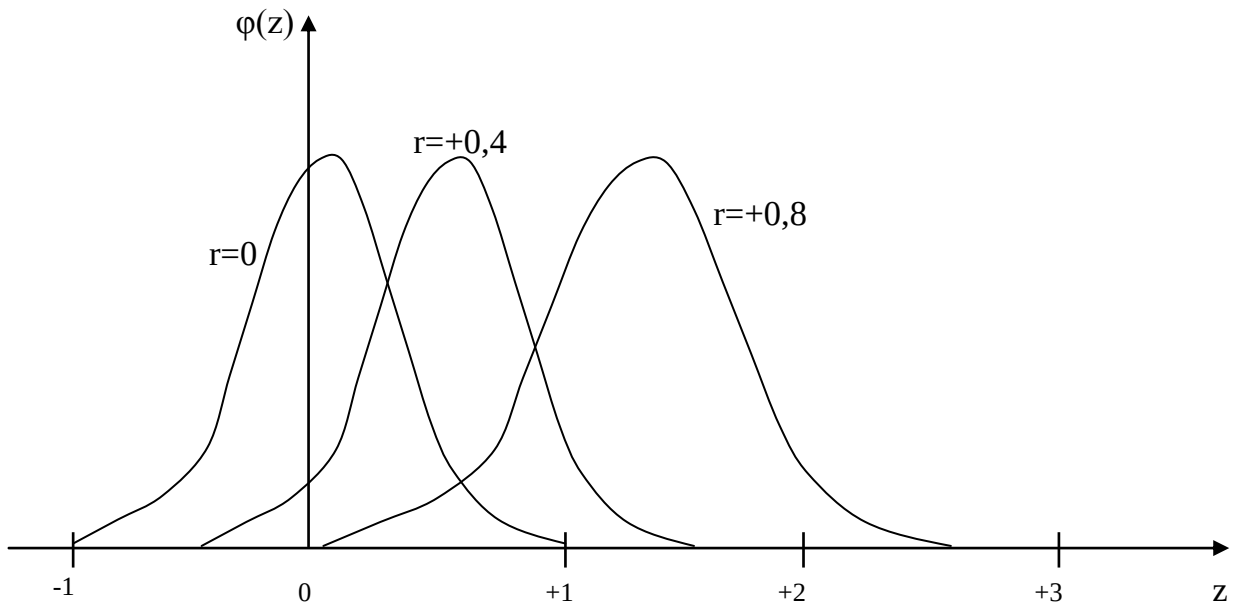


Рис. 1.4. Нормальний розподіл випадкової величини Z.

Перевірка нульової гіпотези $H_0: \rho = 0$ зводиться до розрахунку:

$$Z_{\text{роЗР}} = \frac{1}{2} \ln \frac{1+r}{1-r} \quad (1.28)$$

і порівняння $|Z_{\text{роЗР}}|$ з добутком $(Z_{\text{ТАБ}} \{ p = 1 - \frac{\alpha}{2} \} \cdot \sigma_z)$, де $Z_{\text{ТАБ}}$ – квантиль нормованого нормального розподілу для ймовірності $p = 1 - \frac{\alpha}{2}$.

Якщо $|Z_{\text{роЗР}}| \leq (Z_{\text{ТАБ}} \{ p = 1 - \frac{\alpha}{2} \} \cdot \sigma_z)$, то нульову гіпотезу $H_0: \rho = 0$ (немає лінійного зв'язку) приймаємо.

Тут оцінка ступеня нелінійності зв'язку:

$$\xi_2(Z) = \frac{\left[Z_{\text{ТАБ}} \left\{ p = 1 - \frac{\alpha}{2} \right\} \cdot \sigma_z \right]}{|Z_{\text{РОЗР}}|}; \quad (1.29)$$

Якщо $|Z_{\text{РОЗР}}| > (Z_{\text{ТАБ}} \{ p = 1 - \frac{\alpha}{2} \} \cdot \sigma_z)$, то нульова гіпотеза відкидається, приймається $H_1: \rho \neq 0$, щільний лінійний зв'язок коефіцієнт кореляції значущий.

Тут оцінка ступеня лінійности зв'язку:

$$\xi_1(Z) = \frac{|Z_{\text{РОЗР}}|}{\left[Z_{\text{ТАБ}} \left\{ p = 1 - \frac{\alpha}{2} \right\} \cdot \sigma_z \right]} \quad (1.30)$$

Довірчі інтервали та довірна ймовірність для μ_z :

$$p [(Z_1) < \mu_z < (Z_2)] = 1 - \alpha, \quad (1.31)$$

$$\text{де } Z_1 = Z - (z_{1-\alpha/2} \cdot \sigma_z); \quad (1.32)$$

$$Z_2 = Z + (z_{1-\alpha/2} \cdot \sigma_z); \quad (1.33)$$

Довірчі інтервали та довірна ймовірність для ρ :

$$p [r_1 < \rho < r_2] = 1 - \alpha, \quad (1.34)$$

$$r_1 = \text{th} z_1; \quad (1.35)$$

$$r_2 = \text{th} z_2$$

$$\begin{aligned} \text{th}(z) &= \frac{\sinh(z)}{\cosh(z)} = \frac{\frac{\exp(z) - \exp(-z)}{2}}{\frac{\exp(z) + \exp(-z)}{2}} = \frac{\exp(z) - \exp(-z)}{\exp(z) + \exp(-z)} = \frac{\frac{\exp(z)}{\exp(-z)} - 1}{\frac{\exp(z)}{\exp(-z)} + 1} = \\ &= \frac{\exp(2z) - 1}{\exp(2z) + 1}; \end{aligned} \quad (1.36)$$

$$r_1 = \text{th} z_1 = \frac{\exp(2z_1) - 1}{\exp(2z_1) + 1} \quad (1.37)$$

$$r_2 = \text{th} z_2 = \frac{\exp(2z_2) - 1}{\exp(2z_2) + 1} \quad (1.38)$$

5. Перевірку значущості коефіцієнта кореляції проводять за його нижньою межею довірчої ділянки для абсолютного значення коефіцієнта кореляції:

$$r_{KP} \left\{ p = 1 - \frac{\alpha}{2}; f = N - 2 \right\} \quad (1.39)$$

Якщо $|r| \leq r_{KP}$, то H_0 приймається – немає значущого лінійного зв'язку.

Тут оцінка ступеня нелінійності зв'язку:

$$\xi_2(r) = \frac{r_{KP}}{|r|}; \quad (1.40)$$

Якщо $|r| > r_{KP}$, то H_0 відкидають – є значущий лінійний зв'язок.

Тут оцінка ступеня лінійності зв'язку:

$$\xi_1(r) = \frac{|r|}{r_{KP}}; \quad (1.41)$$

За перетвореннями Фішера змінна

$$t = |r| \sqrt{\frac{N - 2}{1 - r^2}} \quad (1.42)$$

підпорядковується розподілу Стюдента: тому розраховується:

$$t_{PO3P} = |r| \sqrt{\frac{N - 2}{1 - r^2}} \quad (1.43)$$

і порівнюється з $t_{TAB} \{ \alpha, f = N - 2 \}$;

$$\frac{t_{TAB}}{\sqrt{N - 2}} = \frac{|r|}{\sqrt{1 - r^2}}, \quad (1.44)$$

звідки межа, нижче якої вже не відрізняється від $r = 0$.

1.4. Порівняння двох генеральних коефіцієнтів кореляції

1. Іноді виникає необхідність порівняти два вибірових коефіцієнта кореляції з метою перевірки нульової гіпотези при відсутності статистичної різниці між двома генеральними коефіцієнтами кореляції ρ_1 і ρ_2 , які

оцінюються за двома відповідними вибірковими коефіцієнтами кореляції r_1 і r_2 :

$$\begin{array}{ccc}
 H_0: \rho_1 = \rho_2 & & \\
 \uparrow \quad \uparrow \text{ оцінка} & & \\
 r_1 \neq r_2 & &
 \end{array} \quad (1.45)$$

2. Розраховують:

$$Z_1 = \frac{1}{2} \ln \frac{1 + r_1}{1 - r_1}; \quad (1.46)$$

$$Z_2 = \frac{1}{2} \ln \frac{1 + r_2}{1 - r_2}; \quad (1.47)$$

$$Z_{\text{РОЗР}} = \left| \frac{z_1 - z_2}{\sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}} \right|. \quad (1.48)$$

3. Із відповідної таблиці знаходять квантиль нормованого нормального розподілу Гауса для $p = 1 - \alpha$:

$$z_{\text{ТАБ}} \{ p = 1 - \frac{\alpha}{2} \}. \quad (1.49)$$

4. Якщо $|z_{\text{РОЗР}}| \leq z_{\text{ТАБ}} \{ p = 1 - \frac{\alpha}{2} \}$, то нульова гіпотеза H_0 приймається. Цим з рівнем значущості α (для ймовірності $p = 1 - \alpha$) стверджується, що між ρ_1 і ρ_2 немає статистичної різниці, тобто можна вважати, що вибірки взяті із загальної сукупності.

Ступінь рівності двох коефіцієнтів кореляції:

• якщо $|z_{\text{РОЗР}}| > z_{\text{ТАБ}}$, то

$$\xi_1(z) = \frac{z_{\text{ТАБ}}}{|z_{\text{РОЗР}}|}, \text{ то ця різниця статистично існує.} \quad (1.50)$$

Ступінь нерівності двох коефіцієнтів кореляції:

• якщо $|z_{\text{РОЗР}}| \leq z_{\text{ТАБ}}$, то

$$\xi_2(z) = \frac{|z_{\text{PO3P}}|}{z_{\text{TAB}}}.$$
(1.51)

ТЕМА: ЛІНІЙНА МНОЖИННА КОРЕЛЯЦІЯ ТА РЕГРЕСІЯ

2.1. Теоретична частина

1. У випадку k -змінних рівняння множинної регресії у натуральній шкалі має вигляд:

$$\bar{y}_1(2,3,\dots, i,\dots, k) = b_1 + b_2x_2 + b_3x_3 + \dots + b_ix_i + \dots + b_kx_k, \quad (2.1)$$

де $\bar{y}_1(2,3,\dots, i,\dots, k)$ - умовне середнє значення залежної величини y_1 , яке відповідає певним значенням незалежних величин $x_2, x_3, \dots, x_i, \dots, x_k$.

2. **Проведемо процедуру нормування:** перейдемо до нової випадкової нормалізованої змінної t_i :

$$\left. \begin{aligned} t_i &= \frac{y_j - \bar{y}_i}{S_{yi}}; \\ t_j &= \frac{x_j - \bar{x}_i}{S_{xi}}, \end{aligned} \right\} \quad (2.2)$$

тоді, всі величини y_i та всі їх залежності знайдуть вираження у стандартній шкалі.

3. Таким чином, **рівняння регресії (1)** в нормованому вигляді набуде форми:

$$\bar{t}_1(2,3,\dots, i,\dots, k) = \beta_2t_2 + \beta_3t_3 + \dots + \beta_it_i + \dots + \beta_kt_k, \quad (2.3)$$

де $\bar{t}_1(2,3,\dots, i,\dots, k)$ - умовне середнє значення нормованої (стандартної) залежної величини t_1 , яке відповідає певним значенням нормованих незалежних величин $t_2, t_3, \dots, t_i, \dots, t_k$;

$t_2, t_3, \dots, t_i, \dots, t_k$ - значення нормованих (стандартних) незалежних величин $y_2, y_3, \dots, y_i, \dots, y_k$;

$\beta_2, \beta_3, \dots, \beta_i, \dots, \beta_k$ - нормовані (стандартні) коефіцієнти множинної регресії за рівнянням (2.3).

4. **Нормовані коефіцієнти множинної регресії** визначимо за системою лінійних рівнянь:

$$\left. \begin{aligned}
 r_{1,2} &= \beta_2 r_{22} + \beta_3 r_{32} + \beta_4 r_{42} + \dots + \beta_i r_{i2} + \dots + \beta_k r_{k2} \\
 r_{1,3} &= \beta_2 r_{23} + \beta_3 r_{33} + \beta_4 r_{43} + \dots + \beta_i r_{i3} + \dots + \beta_k r_{k3} \\
 r_{1,4} &= \beta_2 r_{24} + \beta_3 r_{34} + \beta_4 r_{44} + \dots + \beta_i r_{i4} + \dots + \beta_k r_{k4} \\
 &\dots\dots\dots \\
 &\dots\dots\dots \\
 r_{1,i} &= \beta_2 r_{2i} + \beta_3 r_{3i} + \beta_4 r_{4i} + \dots + \beta_i r_{ii} + \dots + \beta_k r_{ki} \\
 &\dots\dots\dots \\
 &\dots\dots\dots \\
 r_{1,k} &= \beta_2 r_{2k} + \beta_3 r_{3k} + \beta_4 r_{4k} + \dots + \beta_i r_{ik} + \dots + \beta_k r_{kk}
 \end{aligned} \right\} , \quad (2.4)$$

де $r_{12}, r_{13}, r_{14}, \dots, r_{1i}, \dots, r_{1k}, \dots, r_{22}, r_{23}, r_{2i}, r_{22}, r_{23}, r_{24}, \dots, r_{ik}, \dots, r_{kk}$ – коефіцієнти парної лінійної кореляції між змінними $y_1 \sim x_2; y_1 \sim x_3; x_2 \sim x_3; x_2 \sim x_4; \dots; x_i \sim x_j; x_i \sim x_k; \dots; x_k \sim x_k$.

5. Щільність зв'язку змінної y_1 із сукупністю змінних $x_2, x_3, \dots, x_i, \dots, x_k$ у випадку лінійної множинної кореляції визначається за коефіцієнтом множинної кореляції:

$$r_1(2,3,\dots, i,\dots, k) = \sqrt{\beta_2 r_{12} + \beta_3 r_{13} + \dots + \beta_i r_{1i} + \dots + \beta_k r_{1k}} \quad (2.5)$$

6. Коефіцієнт множинної кореляції після корекції (врахування числа коефіцієнтів рівняння (51) – числа параметрів рівняння регресії в натуральній шкалі дорівнює:

$$\bar{r}_1(2,3,\dots, i,\dots, k) = \sqrt{1 - [1 - r_1^2(2,3,\dots, i,\dots, k)] \cdot \left(\frac{N-1}{N-k} \right)}, \quad (2.6)$$

де N – число спостережень;

k – число параметрів моделі (2.1).

7. Розрахунок коефіцієнтів моделі (2.1) в натуральній шкалі:

$$b_i = \beta_1 \frac{S_{y1}}{S_i}, \text{ де } i=2, 3, \dots, k \quad (2.7)$$

$$b_1 = \bar{y}_1 - [b_2 \bar{x}_2 + b_3 \bar{x}_3 + \dots + b_i \bar{x}_i + \dots + b_k \bar{x}_k], \quad (2.8)$$

де $\bar{y}_1, \bar{x}_2, \bar{x}_3, \dots, \bar{x}_i, \dots, \bar{x}_k$ - середнє значення відповідних величин.

8. Оцінка середньої квадратичної помилки розрахунку величини $\bar{y}_{1(2,3,\dots,i,\dots,k)}$ в рівнянні моделі (2.1) дорівнює:

$$\delta_{\sigma b} = S_{y1} \sqrt{[1 - \bar{r}_1^2(2,3,\dots,i,\dots,k)] \cdot \left(\frac{N-1}{N-2} \right)} \quad (2.9)$$