5.3. Характеристика кореляційного зв'язку між елементами (ознаками)

Після визначення мети і задачі дослідження, а також встановлення доцільності опрацювання матеріалів експериментальних досліджень біологічного або екологічного об'єкту методами кореляційного аналізу проводиться визначення номінальних ознак, що обрані для порівняння, шляхом вимірювання, зважування, встановлення кількості, визначення кольору та інше. Одна з аналізованих ознак приймається як незалежна (аргумент) (x), друга – як сполучена ознака (y) (функція). Звичайно як незалежну обирають ту ознаку або той елемент, який більш доступний для спостереження.

Розміщення залежної ознаки відносно незалежної на осі координат може бути прямолінійним або криволінійним.

Кореляційний зв'язок характеризується коефіцієнтом кореляції (r), який має значення в межах від 0 до +1 і від 0 до –1. При значенні r від 0 до +1 маємо справу з прямою кореляційною залежністю, коли це значення від 0 до –1 – залежність зворотня.

Чим ближче значення коефіцієнта кореляції наближається до 1, тим тісніший (більш щільний) зв'язок між ознаками, що досліджуються.

Коли $r = \pm 1$, маємо справу з функціональним прямолінійним характером зв'язку.

Коли r наближається до 0, то вірогідність наявності прямолінійного зв'язку між ознаками дуже мала, однак тіснота криволінійного зв'язку може бути досить високою.

При наявності кореляції дослідник має справу не з прирощенням (збільшенням або зменшенням) функції, а із взаємосполученою варіацією ознак. Варіації певної кількості ознак (y), яка відповідає відповідному значенню аргументу (x), від середнього її значення характеризується показником, який має назву *коваріації* (Cov):

або

$$Cov = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

X і Y є

$$Cov = \frac{\sum (x_i y_i - n\bar{x}\bar{y})}{n}.$$

величинами різнойменними, тому при кореляційному аналізі проводиться співставлення не саме їх, а перетворених відхилень від середніх у вигляді неіменованих значень нормованих відхилень (t_x):

$$t_x = \frac{x_i - \bar{x}}{\sigma_x}$$

і

$$t_y = \frac{y_i - \bar{y}}{\sigma_y}.$$

Звідси одержується коефіцієнт кореляції (r):

$$r = \frac{\sum t_x \cdot t_y}{n}$$

або

$$r = \frac{Cov}{\sigma_x \sigma_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n\sigma_x \sigma_y}.$$

Отже, одержується неіменоване значення коефіцієнта кореляції, яке має вираз у долях від одиниці.

Якщо в наведеній вище формулі позначити

$$x_i - \bar{x} = \alpha_x \text{ і } \quad y_i - \bar{y} = \alpha_y,$$

то вона може бути представлена у більш скороченому вигляді:

$$\sigma_y^2 = \frac{\sum \alpha_y^2}{n},$$

Якщо врахувати, що

а $$\sum \alpha_x^2 = n\sigma_x^2$$

тобто: і $$\sum \alpha_y^2 = n\sigma_y^2,$$

то наведену вище формулу (*r*) можна представити так:

$$r = \frac{\sum \alpha_x \cdot \alpha_y}{n\sigma_x\sigma_y} = \frac{\sum \alpha_x \cdot \alpha_y}{\sqrt{n\sigma_x^2 \cdot n\sigma_y^2}} = \frac{\sum \alpha_x \cdot \alpha_y}{\sqrt{\sum \alpha_x^2 \cdot \sum \alpha_y^2}}.$$

Таблиця 5. Макет допоміжної таблиці для розрахунків коефіцієнта кореляції

Дані замірів		α_x	α_y	$\alpha_x\alpha_y$	α_x^2	α_y^2
x	y	$\alpha_x=x_i-x$	$\alpha_y=y_i-y$			

Остання формула зветься формулою Пірсона, вона звільняє від необхідності попередніх розрахунків середніх квадратичних відхилень, що

полегшує визначення коефіцієнтів кореляції.

Для розрахунків коефіцієнта кореляції за формулою Пірсона використовується допоміжна таблиця (табл. 5).

Коефіцієнт кореляції, який визначається для відповідної виборки варіант, так саме, як і окремі варіанти, що досліджуються, є величина випадкова. Тому виникає необхідність також визначати ступінь його наближення до показника генеральної сукупності значень *r*. Цей показник позначається літерою (ρ). Для вирішення наведеного завдання також застосовується нульова гіпотеза. Вона полягає в припущенні, що ρ = 0, тобто, що між випадковими величинами *X* і *Y* кореляція відсутня.

Для перевірки цієї нульової гіпотези використовується порівняння показників з критерієм *t*–Стьюдента (t_s). Значення *t* при досить значній кількості спостережень (*n* і 100) є відношенням коефіцієнта кореляції до своєї помилки (m_r), яка визначається за формулою

$$m_r = \frac{1-r^2}{\sqrt{N-1}}.$$

Тоді

$$t = \frac{r}{m_r} = \frac{r\sqrt{N-1}}{1-r^2}.$$

Якщо в досліді кількість спостережень менше 100 ($n < 100$), за критерій для перевірки нульової гіпотези приймається

$$t = \sqrt{\frac{1-r^2}{N-2}}.$$

Якщо $t > t_s$, то нульова гіпотеза відкидається. Це означає, що в генеральній сукупності $\rho \neq 0$, тобто одержаний коефіцієнт кореляції (*r*) достовірно відрізняється від 0, а між *X* і *Y* існує кореляційний зв'язок. При $t < t_s$ зберігається нульова гіпотеза, а одержане відхилення (*r*) від 0 є випадковим.

Приклад. У виборці, в якій n = 50, одержаний коефіцієнт кореляції r = 0,54. Треба визначити достовірність одержаного значення (r). Вираховуємо критерій достовірності (t):

$$t = \frac{0{,}54\sqrt{50-2}}{\sqrt{1-0{,}54^2}} = \frac{6{,}92}{0{,}87} = 8{,}0.$$

За таблицями Стьюдента для $R = 50 - 2 = 48$ і P =0,01 знаходимо: t_s= 2,68. Оскільки $t = 8{,}0 > t_s = 2{,}68$, нульова гіпотеза відкидається, тобто одержане значення є достовірне на 0,05 рівні значущості.

Слід мати на увазі різні методичні підходи до використання нульової гіпотези для перевірки критерію достовірності середнього значення виборки і критерію достовірності коефіцієнта кореляції. При визначенні критерію достовірності середнього значення нульова гіпотеза полягає в припущенні, що між параметрами вибіркової і генеральної сукупності різниця відсутня, тобто $П_в - П_г = 0$ (розд. 5).

При визначенні ступеня наближення коефіцієнта кореляції вибіркової сукупності до гіпотетичного коефіцієнта кореляції генеральної сукупності (ρ) застосування нульової гіпотези полягає в припущенні, що між випадковими величинами X і Y кореляція відсутня, тобто ρ = 0, що означає припущення повної відсутності кореляції між X і Y.

Р.Фішер запропонував для оцінки ступеню кореляції в малих виборках застосувати не коефіцієнт кореляції, а пов'язану з ним допоміжну величину Z (зет):

$$Z = 1{,}15129 \lg \frac{1+r}{1-r}.$$

Z змінює своє значення від $-\infty$ до $+\infty$ і при малих виборках дає більш надійні результати, ніж коефіцієнт кореляції r.

Перетворення коефіцієнта кореляції в показник Z здійснюється за

специфічною таблицею, складеною Фішером (додатки 3, 4).

Критерій достовірності показника Z визначається:

$$\sigma_Z = \frac{1}{\sqrt{n-3}};$$

Цей критерій діє як для малих, так і для великих виборок, коли замість коефіцієнта кореляції застосовується відповідне до нього значення Z.

Послідовність дій при застосуванні показника Z наступна. За показником коефіцієнта кореляції r за таблицю Фішера встановлюють значення Z (зет). Далі:

- визначають величину помилки Z за формулою

$$t_Z = \frac{Z}{\sigma_Z} = Z\sqrt{n-3};$$

- за таблицею знаходять значення критерію t_Z:

$$\Delta_Z = t_{\sigma^2}$$

- значення t_Z порівнюється із стандартом за таблицю Стьюдента для прийнятого рівня значимості (P) і числа ступенів вільності: $k = n - 2$;
- за величиною максимальної помилки –
  – знаходять границі довірчого інтервалу для генерального параметру.

$$\sigma_Z = \frac{1}{\sqrt{28-3}} = \frac{1}{5} = 0{,}20.$$

Приклад. По виборці $n = 28$ одержаний $r = 0{,}52$. За таблицею (додаток 5) знаходимо, що цьому значенню (r) відповідає значення $Z = 0{,}576$. Вираховуємо помилку:

$$\Delta_Z = t_{\sigma^2} = 1{,}96 \cdot 0{,}20 = 0{,}392$$

Звідси

За таблицею критерію *t*-Стьюдента (додаток 2) для $k = 28 - 2 = 26$ і $P = 0{,}05$ знаходимо $t_s = 2{,}006$. Порівняємо: $t_\phi = 2{,}88 > t_s = 2{,}06$; звідси робимо висновок про те, що нульова гіпотеза відхиляється.

Далі за величиною

знаходимо межі довірчого інтервалу для показника *Z* (зет):

- нижня межа = 0,576 – 0,392 = 0,184;
- верхня межа = 0,576 + 0,392 = 0,968.

Користуючись додатком 5 переводимо значення "зет" в величини коефіцієнта кореляції (*r*) і знаходимо його довірчі границі:

- нижня границя – 0,18;
- верхня границя – 0,74.

Знайдені межі довірчого інтервалу свідчать про те, що величина коефіцієнту кореляції в генеральній сукупності знаходяться в межах $0{,}18 < r < 0{,}74$; тобто наведений вище коефіцієнт кореляції: $r = 0{,}52$ визначений з достатньою точністю.

Якщо ставиться завдання визначити коефіцієнт кореляції із заданим показником достовірності, то попередньо визначається мінімальна кількість

$$n = \frac{(1{,}96)^2}{(0{,}2554)^2} + 3 = \frac{3{,}842}{0{,}065} + 3 = 59 + 3 = 62.$$

спостережень, при якій буде забезпечена задана достовірність. Для цього використовується формула

$$n = \frac{t_Z^2}{Z^2} + 3;$$

де n – необхідна мінімальна кількість парних спостережень (об'єм виборки); t_Z – задана по прийнятому порогу довірчої імовірності величина критерію достовірності Z (зет).

Приклад. Для $r = 0{,}25$ і $n = 20$ величина $Z = 0{,}2554$. Звідси $t_Z = 0{,}2554\sqrt{(17)} = 1{,}05$. Для $P = 0{,}05$ і $K = 20 - 2 = 18$ $t_s = 2{,}10$, тобто $t_\phi = 1{,}05 < t_s = 2{,}10$. Це означає неможливість відкидання нульової гіпотези.

Питання: яке число спостережень (n) потрібно провести, щоб із заданою імовірністю $P = 0{,}95$ зробити остаточний висновок про наявність або відсутність кореляції між ознаками X і Y.

Використовуючи останню формулу і виходячи з того, що для імовірності $P = 0{,}95$ відповідає $t = 1{,}96$, знаходимо:

$$t = \frac{r_1 - r_2}{\sqrt{(m_{r1})^2 + (m_{r2})^2}}.$$

Тобто відповідь на поставлене запитання можна одержати, якщо провести не менше 62 спостережень.

У біометричних дослідженнях зустрічаються ситуації, коли необхідно дати оцінку вибіркових сукупностей, що порівнюються, за показниками тісноти зв'язку, тобто за коефіцієнтами кореляції відповідних ознак в цих сукупностях. В таких випадках необхідно оцінити достовірність різниці між відповідними коефіцієнтами кореляції. Для цього визначається критерій t різниці коефіцієнтів кореляції:

$$m_{DZ} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}.$$

Одержане значення t порівнюється із значенням t_s. Різницю можна вважати достовірною, якщо $t > t_s > 3$.

Оцінку достовірності різниці між коефіцієнтами кореляції можна також

здійснити за допомогою критерія Z Фішера. Для цього за таблицею (додаток 4) переводимо значення коефіцієнтів кореляції (r_1 і r_2) в показники критерію Z (Z_1 і Z_2). Позначимо різницю між Z_1 і Z_2 літерою D:

$$D = Z_1 - Z_2.$$

Далі визначається помилка цієї різниці як сума помилок Z_1 і Z_2, тобто

$$t_D = \frac{Z_1 - Z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}.$$

Критерій достовірності оцінки різниці значень Z_1 і Z_2 є її відношення до наведеної помилки:

$$r_{xyz} = \sqrt{\frac{r_{xz}^2 + r_{yz}^2 + 2r_{xy}r_{xz}r_{yz}}{1 - r_{xy}^2}}.$$

Одержане значення t_D за описаним вже принципом порівнюється із табличним значенням критерія t_{St}.