Тема 4. Рангова кореляція

4.1. Коефіцієнт рангової кореляції Спірмена

У багатьох випадках результати спостережень подаються не у вигляді кількісних вимірювань, а у вигляді бальних оцінок (рангів). Наприклад, студенти у групі можуть бути впорядковані по номерам за середнім балом в сесії, країни– за кількістю населення, учасники конкурсу – за зайнятим місцем тощо. При цьому інколи виникає можливість упорядкувати об'єкти дослідження за двома або більше показниками. У зв'язку з цим виникає задача дослідження кореляції цих показників.

Нехай n об'єктів дослідження, розташованих за рівнем якості, характеризуються парами рангів $(x_i, y_i), i = 1,2,...,n$. Потрібно з'ясувати рівень кореляції між двома ознаками, x та y. Для цього використовують *коефіцієнт рангової кореляції Спірмена*. Цей показник розраховують за формулою:

$$R = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)}, \tag{4.1}$$

де $d_i = x_i - y_i$ – різниця рангів для i-го об'єкту спостереження.

Це значення можна використати як наближення коефіцієнта кореляції, він є менш точним у порівнянні зі звичайним коефіцієнтом кореляції, оскільки при його розрахунку не враховуються кількісні значення характеристик об'єктів, а лише їх порядок. Як і для звичайного коефіцієнта кореляції, значення R змінюється від –1 до 1. Чим ближче абсолютне значення коефіцієнта рангової кореляції до одиниці, тим більш щільним є зв'язок між факторами.

Приклад 4.1. У таблиці наведено дані про місця, що займають 8 провідних компаній галузі за собівартістю продукції (фактор *x*) та часткою ринку (фактор *y*). Обчислити коефіцієнт рангової кореляції Спірмена.

Таблиця 4.1. Дані про розподіл компаній галузі за собівартістю продукції та часткою ринку

Підприємство	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
Фактор *x*	8	3	1	4	2	7	5	6
Фактор *y*	3	5	6	7	8	4	1	2
d=x–y	5	–2	–5	–3	–6	3	4	4

Маємо:

$$\sum_{i=1}^{8} d_i^2 = 25+4+25+9+36+9+16+16 = 140.$$

$$R = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2-1\right)} = 1 - \frac{6\cdot 140}{8\cdot 63} = -0,67.$$

Отримане значення коефіцієнта рангової кореляції Спірмена свідчить про наявність взаємозв'язку між собівартістю продукції компанії та часткою ринку. При цьому спостерігається обернена залежність: зі зростанням собівартості продукції компанії її частка ринку зменшується.

4.2 Коефіцієнт рангової кореляції Кендалла

Іншим показником, що характеризує узгодженість двох факторів, є *коефіцієнт рангової кореляції Кендалла.* Для обчислення цього показника ранги N_x значень показника *x* розташовують у порядку зростання, при цьому для кожного значення рангу N_x фіксують ранг N_y відповідного значення показника *y*. Ідеальна кореляція між цими показниками буде спостерігатися у тому випадку, коли послідовності значень N_x та N_y будуть співпадати. Коефіцієнт рангової кореляції Кендалла дозволяє визначити міру відповідності цих послідовностей.

Для кожного значення N_y послідовно визначають кількість розташованих за ним рангів, що перевищують N_y, а також кількість рангів, менших, ніж N_y. Перша група рангів враховується зі знаком «+», їх суму позначимо *P*. Ранги другої групи враховуються зі знаком «–», нехай їх сума дорівнює *Q*.

$1+2+\ldots+n=\frac{n(n+1)}{2}$

Максимальне значення P досягається у випадку, коли ранги N_y значень фактору y співпадають з рангами N_x значень фактору x і кожна послідовність рангів співпадає з послідовністю натуральних чисел від 1 до n, розміщених у порядку зростання. Тоді після першої пари значень $N_x = 1$ та $N_y = 1$ кількість перевищень цих значень рангів буде дорівнювати n–1, після другої пари $N_x = 2$ та $N_y = 2$ ця кількість буде дорівнювати n–2 і так далі. Отже, у випадку, коли ранги x та y співпадають, а кількість рангів дорівнює n, маємо:

$$P_{\max} = (n-1) + (n-2) + \ldots + 3 + 2 + 1 = \frac{n(n-1)}{2}.$$

Якщо послідовність рангів y має обернену тенденцію по відношенню до послідовності рангів x, то Q матиме таке ж максимальне значення по модулю:

$$|Q_{\max}| = \frac{n(n-1)}{2}.$$

Якщо ранги y не співпадають з рангами x, то знаходять суму $S = P + Q$. Відношення цієї суми до $P_{\max} = |Q_{\max}| = \frac{n(n-1)}{2}$ дорівнює коефіцієнту рангової кореляції Кендалла:

$$\tau = \frac{2S}{n(n-1)}. \qquad (4.2)$$

Приклад 4.2. Маємо дані для 10 аграрних компаній про врожайність картоплі *y* (ц/га) та кількість внесених на 1 га мінеральних добрив *x* (кг). Вихідні дані наведені у розрахунковій таблиці 4.2. З допомогою коефіцієнта рангової кореляції Кендалла виміряти щільність взаємозв'язку між цими показниками.

Розглянемо, як відбувається підрахунок балів.

Оскільки ранги *x*, тобто N_x, розташовані у порядку зростання, підрахунок балів здійснюємо, спостерігаючи за змінами N_y. Після першої пари 9 значень N_y більші 1 і жодного значення, меншого за 1. Тому у першому рядку стоїть 9 у стовпчику зі знаком «+» та 0 у стовпчику зі знаком «–». Після другої пари, де значення N_y=3, спостерігається 7 випадків, коли ранги у перевищують 3, і один випадок (N_y=2), коли ранг менший 3. Відповідно, у другому рядку записані цифри 7 у стовпчику зі знаком «+» та 1 у стовпчику зі знаком «–».

Знайшовши суму елементів стовпчика зі знаком «+», отримуємо $P = 38$, підсумовуючи числа у стовпчику зі знаком «–» і враховуючи знак, отримуємо значення $Q = -7$. Тоді $S = P + Q = 38 - 7 = 31$.

Таблиця 4.2. Розрахункова таблиця для визначення коефіцієнта рангової кореляції Кендалла

x	*y*	Ранги		Підрахунок балів	
		N_x	N_y	«+»	«–»
138	218	1	1	9	0
175	240	2	3	7	1
190	232	3	2	7	0
196	280	4	6	4	2
200	260	5	4	5	0
235	310	6	9	1	3
250	290	7	7	2	1
260	278	8	5	2	0
275	300	9	8	1	0
290	320	10	10	–	–
$n = 10$				$P = 38$	$Q = -7$

Звідси знаходимо коефіцієнт кореляції Кендалла:

$$\tau = \frac{2S}{n(n-1)} = \frac{2 \cdot 31}{10 \cdot 9} = 0,69.$$

Отримане значення коефіцієнта рангової кореляції свідчить про достатньо високу (вищу за середню, оскільки $|\tau| > 0{,}5$) щільність зв'язку між факторами *x* та *y*.

Аналогічно виконують розрахунок показника τ для випадку протилежної спрямованості рангів факторів *x* та *y*. Розглянемо цей випадок на прикладі.

Приклад 4.3. У розрахунковій таблиці 4.3 наведено дані щодо погодинної оплати праці на підприємстві *x* (г.о.) та рівня плинності кадрів *y* (кількість працівників, що звільнилися за рік). Розрахувати коефіцієнт рангової кореляції Кендалла для цих даних.

Таблиця 4.3. Розрахункова таблиця для визначення коефіцієнта рангової кореляції Кендалла при протилежній спрямованості рангів

x	*y*	Ранги		Підрахунок балів	
		N_x	N_y	«+»	«–»
3	34	1	7	1	6
4	35	2	8	0	6
5	33	3	6	0	5
6	28	4	5	0	4
7	20	5	3	1	2
8	24	6	4	0	2
9	15	7	2	0	1
10	11	8	1	–	–
n=8				$P = 2$	$Q = -26$

За даними таблиці 4.3 отримуємо $P = 2$, $Q = -26$, $S = P + Q = 2 - 26 = -24$.

Звідси знаходимо коефіцієнт рангової кореляції Кендалла:

$$\tau = \frac{2S}{n(n-1)} = \frac{2 \cdot (-24)}{8 \cdot 7} = -0{,}857.$$

Отримане від'ємне значення коефіцієнта, за абсолютною величиною близьке до 1, свідчить про наявність досить щільного зворотного зв'язку між факторами *x* та *y*.

Зауважимо, що з допомогою коефіцієнтів рангової кореляції Спірмена та Кендалла можна вимірювати щільність взаємозв'язку, не лише між кількісними, але й якісними (атрибутивними) ознаками (стать, професія тощо), впорядкованими певним чином.

4.3. Коефіцієнт конкордації

Якщо досліджується щільність зв’язків між більше, ніж двома факторами, то для її кількісної оцінки можна використати *коефіцієнт конкордації (множинний коефіцієнт рангової кореляції*):

$$W = \frac{12S}{m^2\left(n^3 - n\right)}, \tag{4.3}$$

де, m – кількість факторів (ознак), n – кількість спостережень, S – сума квадратів відхилень суми n рангів від їх середньої величини:

$$S = \sum_{i=1}^{n}\left(\sum_{j=1}^{m} R_{ij}\right)^2 - \frac{\left(\sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij}\right)^2}{n}.$$

Формулу (4.3) застосовують, якщо ранги для кожної ознаки не повторюються. Якщо зустрічаються пов’язані ранги (ранги, що повторюються), то коефіцієнт конкордації розраховують з врахуванням кількості таких пов’язаних рангів по кожному фактору:

$$W = \frac{12S}{m^2\left(n^3 - n\right) - m\sum_{k=1}^{m}\left(k^3 - k\right)}, \tag{4.4}$$

де k – кількість однакових рангів за кожною ознакою.

Розглянемо розрахунок коефіцієнта конкордації для випадку, коли для кожної ознаки ранги, що повторюються відсутні.

Приклад 4.4. Для 4 опитаних сімей отримано дані щодо їх річного доходу, кількості дітей та річних заощаджень. Відповідні дані наведені у розрахунковій таблиці 4.4. Розрахувати коефіцієнт конкордації для оцінки щільності зв'язку між цими показниками.

Занесемо необхідні проміжні розрахунки у таблицю 4.4. Маємо $m = 3$, $n = 4$.

Таблиця 4.4. Розрахункова таблиця для визначення коефіцієнта конкордації

Порядковий номер сім'ї	Річний дохід, тис. г.о., x_1	Кількість дітей у сім'ї, x_2	Річні заощадження, тис. г.о., x_3	Ранги факторів			Сума рангів, $\sum_{i=1}^{m} R_i$	Квадрат суми рангів $\left(\sum_{i=1}^{m} R_i\right)^2$
				R_1	R_2	R_3		
A	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	30	2	2,5	1	2	1	4	16
2	35	1	3,1	2	1	2	5	25
3	38	3	4,2	3	3	4	10	100
4	40	4	3,6	4	4	3	11	121
Σ							30	262

Позначивши R_{ij} ранг i-го фактору у j-ої одиниці спостереження, впорядковуємо кожний з трьох факторів (стовпці 4-6), потім знаходимо суму рангів у кожному рядку і записуємо її у стовпці 7. У стовпець 8 заносимо квадрат цієї суми і знаходимо суму елементів цього стовпця. У нашому прикладі отримуємо:

$$\sum_{i=1}^{n}\left(\sum_{j=1}^{m} R_{ij}\right)^2 = 262\,.$$

Потім суму елементів 7-го стовпця підносимо до квадрату і ділимо на кількість спостережень n. Цю частку віднімаємо від суми елементів 8-го стовпця:

$$S = \sum_{i=1}^{n}\left(\sum_{j=1}^{m} R_{ij}\right)^2 - \frac{\left(\sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij}\right)^2}{n} = 262 - \frac{30^2}{4} = 37\,.$$

Підставивши знайдене значення S у формулу (4.3), отримуємо коефіцієнт конкордації:

$$W = \frac{12S}{m^2\left(n^3 - n\right)} = \frac{12 \cdot 37}{3^2 \cdot \left(4^3 - 4\right)} = 0{,}82 .$$

Коефіцієнт конкордації може набувати значень від 0 до 1. Отримане значення $W = 0{,}82$, близьке до 1, дає змогу зробити висновок про наявність досить щільного зв'язку між факторами x_1, x_2 та x_3. Проте, щоб це твердження не було помилковим, коефіцієнт конкордації слід перевірити на істотність. Для цього використовуємо критерій χ^2. За відсутності пов'язаних рангів його розраховують за формулою:

$$\chi^2 = \frac{12S}{mn\left(n+1\right)}. \tag{4.5}$$

За наявності пов'язаних рангів використовуємо формулу:

$$\chi^2 = \frac{12S}{mn\left(n+1\right) - \frac{\sum_{k=1}^{m}\left(k^3 - k\right)}{n-1}}. \tag{4.6}$$

Тут k – кількість рангів, що повторюються.

Фактичне (розраховане) значення χ^2 порівнюють з табличним значенням. Його знаходять за таблицею χ^2-критерію у відповідності з рівнем значущості α (він дорівнює 0,05 або 0,1) та параметром $v = n - 1$. Якщо $\chi^2 > \chi^2_{табл.}$, то коефіцієнт W є істотним.

У нашому прикладі за відсутності пов'язаних рангів маємо:

$$\chi^2 = \frac{12 \cdot 37}{3 \cdot 4 \cdot \left(4+1\right)} = 7{,}4 ;\ \chi^2{}_{табл.} = 7{,}81 \quad \left(\alpha = 0{,}05; v = 4 - 1 = 3\right).$$

Оскільки $\chi^2 < \chi^2_{табл.}$, то на рівні значущості $\alpha = 0{,}05$ знайдений коефіцієнт конкордації не можна визнати істотним. Це значить, що використана кількість спостережень є недостатньою для гарантії відповіді з ймовірністю $1 - \alpha = 0{,}95$.

Якщо прийняти $\alpha = 0{,}1$, то $\chi^2{}_{табл.} = 6{,}25$, і на цьому рівні істотності, тобто з ймовірністю 1–0,1 = 0,9 коефіцієнт конкордації може вважатися істотним, оскільки $\chi^2 > \chi^2_{табл.}$.

Коефіцієнт конкордації часто використовують у експертних оцінках для того, щоб визначити ступінь узгодженості думок експертів про важливість певного показника або скласти рейтинг окремих одиниць за певною ознакою.

Розглянемо приклад розрахунку коефіцієнту конкордації за наявності рангів, що повторюються.

Приклад 4.5.

Нехай два експерти (m = 2) упорядкували чотири ознаки (n = 4), що впливають на певний результат, за їх важливістю. Відповідні дані наведено у розрахунковій таблиці 4.5. Обчислити коефіцієнт конкордації з метою оцінки узгодженості думок експертів.

Таблиця 4.5. Експертна оцінка ознак

Факторна ознака, x_i	Ранг, встановлений експертом		Сума рангів для кожної ознаки	Квадрат суми рангів
	першим	другим		
1	**2**	**3**	**4**	**5**
x_1	1	1,5	2,5	6,25
x_2	2,5	1,5	4	16
x_3	2,5	4	6,5	42,25
x_4	4	3	7	49
Σ	10	10	20	113,50

Враховуючи наявність пов’язаних факторів, для розрахунку коефіцієнта конкордації використаємо формулу (4.6).

У нашому прикладі $m = 2$, $n = 4$, отже

$$S = 113{,}5 - \frac{20^2}{4} = 13{,}5.$$

Тут і у першого, і у другого експерта два пов’язані ранги, тому маємо:

$$\sum_{k=1}^{2}\left(k^3 - k\right) = \left(2^3 - 2\right) + \left(2^3 - 2\right) = 12.$$

Підставивши всі знайдені значення у формулу (4.4), отримуємо значення коефіцієнта конкордації:

$$W = \frac{12S}{m^2\left(n^3 - n\right) - m\sum_{k=1}^{m}\left(k^3 - k\right)} = \frac{12 \cdot 13{,}5}{2^2\left(4^3 - 4\right) - 2 \cdot 12} = 0{,}75.$$

Значення W є достатньо великим. Перевіримо його на істотність з допомогою критерію χ^2. За формулою (4.6) знаходимо фактичне значення критерію:

$$\chi^2 = \frac{12S}{mn\left(n+1\right) - \frac{\sum_{k=1}^{m}\left(k^3 - k\right)}{n-1}} = \frac{12 \cdot 13{,}5}{2 \cdot 4 \cdot 5 - \frac{12}{3}} = 4{,}5.$$

Для $v = n - 1 = 3, \alpha = 0{,}05$ знаходимо табличне значення $\chi^2_{табл.}$: $\chi^2_{табл.} = 7{,}81$. Оскільки $\chi^2 < \chi^2_{табл.}$, то коефіцієнт конкордації не можна вважати істотним. Можливо, для формування узгодженої думки слід залучити більшу кількість експертів або розглянути додаткові факторні ознаки.

4.4 Інші непараметричні методи дослідження зв'язків між факторами

Розглянемо методику оцінки зв'язку між якісними ознаками з використанням коефіцієнтів спряження.

В основі обчислення щільності зв'язку між атрибутивними (якісними) ознаками знаходиться побудова таблиці взаємного спряження (взаємозалежності) (таблиця 4.6), у якій наводяться комбінаційні розподіли сукупностей за факторною та результативною ознаками.

Таблиця 4.6. Загальний вигляд таблиці взаємного спряження

Групи за ознакою x	Групи за ознакою y						
	Група 1	Група 2	…	Група j	…	Група m_2	Разом
Група 1	f_{11}	f_{12}	…	f_{1j}	…	f_{1m_2}	f_{10}
Група 2	f_{21}	f_{22}	…	f_{2j}	…	f_{2m_2}	f_{20}
…	…	…	…	…	…	…	…
Група i	f_{i1}	f_{i2}		f_{ij}		f_{im_2}	f_{i0}
…	…	…	…	…	…	…	…
Група m_1	$f_{m_1 1}$	$f_{m_1 2}$	…	$f_{m_1 j}$	…	$f_{m_2 m_1}$	$f_{m_1 0}$

Разом	f_{01}	f_{02}	…	f_{0j}	…	$f_{0m_2 1}$	n

Величина f_{ij} – це число спостережень на перетині *i*-го рядка та *j*-го стовпця, тобто частота групи *i* у групі *j,* а f_{i0} та f_{0j} – відповідно підсумкові частоти за ознакою *x* та ознакою *y*. У випадку відсутності стохастичної залежності між ознаками частки умовних розподілів збігаються і дорівнюють часткам безумовного розподілу (часткам розподілу по підсумковому рядку). Розбіжність між фактичною кількістю спостережень у клітинках таблиці 4.6 і теоретично можливою за повної відсутності зв'язку оцінюють за допомогою показника χ^2, який розраховують за формулою:

$$\chi^2 = n\left[\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\frac{f_{ij}^2}{f_{i0}f_{j0}} - 1\right]. \tag{4.7}$$

За відсутності зв'язку між ознаками $\chi^2 = 0$.

Для вимірювання щільності зв'язку між ознаками використовують кілька коефіцієнтів спряження. Найчастіше використовують *коефіцієнт Чупрова*. Він обчислюється за формулою:

$$K_Ч = \sqrt{\frac{\chi^2}{n\sqrt{(m_1-1)(m_2-1)}}}. \tag{4.8}$$

Тут *n* – кількість спостережень.

Якщо кількості виділених груп за кожною ознакою рівні, тобто $m_1 = m_2$ і між ознаками існує функціональний зв'язок, то коефіцієнт Чупрова дорівнює 1. Проте, якщо $m_1 \neq m_2$, то значення коефіцієнта Чупрова відмінне від 1 навіть за наявності функціонального зв'язку між ознаками.

Модифікацією коефіцієнта Чупрова є коефіцієнт Крамера:

$$K_К = \sqrt{\frac{\chi^2}{n(m-1)}}. \tag{4.9}$$

Тут $m = \min(m_1, m_2)$.

Оцінити щільність зв'язку між якісними ознаками можна також за допомогою коефіцієнта Пірсона:

$$K_{П} = \sqrt{\frac{\chi^2}{n + \chi^2}}. \qquad (4.10)$$

Значення коефіцієнтів Чупрова, Крамера та Пірсона коливаються у межах від 0 до 1. Коефіцієнт Чупрова враховує кількість виділених груп за кожною ознакою і дає найбільш обережну оцінку щільності зв'язку. Якщо значення цього коефіцієнта $K_{Ч} \geq 0{,}3$, то можна говорити про помірний або щільний зв'язок між ознаками. Перевірка істотності зв'язку здійснюється на основі χ^2-критерію з $V = (m_1 - 1)(m_2 - 1)$ ступенями вільності.

Приклад 4.6. На основі даних, наведених у таблиці 4.7, дослідити щільність зв'язку між категоріями працівників підприємства та задоволеністю рівня оплати праці.

Розв'язання. Обчислимо значення χ^2. За формулою (4.7) маємо:

$$\chi^2 = 131\left(\frac{625}{71\cdot 40}+\frac{225}{60\cdot 40}+\frac{1600}{71\cdot 80}+\frac{1600}{60\cdot 80}+\frac{36}{71\cdot 11}+\frac{25}{60\cdot 11}-1\right)=$$

$$=1,69.$$

Обчислимо значення коефіцієнта Чупрова:

$$K_ч = \sqrt{\frac{1,69}{131\sqrt{2\cdot 1}}} = 0,1.$$

Таблиця 4.7. Дані щодо задоволеності рівнем оплати праці різних категорій працівників підприємства

Група працівників	Кількість працівників		
	Задоволений оплатою праці	Незадоволений оплатою праці	Разом
Управлінський та інженерно-технічний персонал	25	15	40
Робітники	40	40	80
Допоміжний персонал	6	5	11
Разом	71	60	131

Оскільки значення цього коефіцієнта менше 0,3, то можна говорити, про дуже слабкий зв'язок між ознаками, що розглядаються. Аналогічний висновок отримуємо, використавши χ^2-критерій. Для рівня значимості $\alpha = 0,05$ та $V = (3-1)(2-1) = 2$ ступенів вільності з таблиць розподілу χ^2 отримуємо $\chi^2_{табл.} = 5,99 > 1,69$, тому слід прийняти гіпотезу про відсутність зв'язку між ознаками, що розглядаються.