ЛЕКЦІЯ 13

КОЕФІЦІЄНТИ КОРЕЛЯЦІЇ

13.1. Коефіцієнт вибіркової кореляції (коефіцієнт кореляції Пірсона)

Розглянемо статистичні вибірки об’єму n, які позначимо $X; Y$:

$$X = \{x_1, \dots, x_n\}; Y = \{y_1, \dots, y_n\}.$$

Частоти кожного елемента вибірок дорівнюють 1.

Незміщеними оцінками кореляційного моменту $K(X;Y)$ і дисперсій $\sigma^2(X)$ і $\sigma^2(Y)$ (лекція 12) є відповідно вибірковий кореляційний момент (вибіркова коваріація) $K_{\text{В}}(X;Y)$ і виправлені вибіркові дисперсії, які обчислюються за формулою (13.1):

$$s^2(X) = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x_{\text{В}}})^2 \text{ і } s^2(Y) = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (y_i - \overline{y_{\text{В}}})^2. \quad (13.1)$$

Тоді, якщо відомі (x_1, y_1); $(x_2, y_2), \dots, (x_n, y_n)$ – фактичні значення змінних X і Y, то вибіркова коваріація між ними розраховується наступним чином (13.2):

$$K_{\text{В}}(X;Y) = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left((x_i - \overline{x_{\text{В}}}) \cdot (y_i - \overline{y_{\text{В}}}) \right), \quad (13.2)$$

де $\overline{x_{\text{В}}} = \frac{\sum_{i=1}^{n} x_i}{n}$; $\overline{y_{\text{В}}} = \frac{\sum_{i=1}^{n} y_i}{n}$ – середні арифметичні відповідних величин.

Видно, що у виразі для $K_{\text{В}}(X;Y)$ присутній один зв'язок (добуток середніх), тому для нього, як і для вибіркових дисперсій, число ступенів свободи дорівнює $n - 1$.

Той факт, що для отримання незміщеної оцінки s^2 в знаменнику вибіркової дисперсії довелося n замінити на $n - 1$, безпосередньо пов'язаний з тим, що величина відносно якої беруться відхилення, сама залежить від елементів вибірки. Якщо б у формулі вибіркової дисперсії були дві такі величини, то n потрібно було б замінити на $n - 2$ і т.д.

Кожна величина, яка залежить від елементів вибірки і є у формулі вибіркової дисперсії, називається *зв'язком*. Виявляється, що знаменник вибіркової дисперсії завжди дорівнює різниці між об’ємом вибірки n і числом зв'язків, накладених на цю вибірку. Число $k = n - l$ буде вважатися *числом ступенів свободи*.

Вибірковий коефіцієнт кореляції (*коефіцієнт кореляції Пірсона*) розраховується за формулою (13.3):

$$\rho_{\text{П}}(X,Y) = \rho_{\text{В}}(X,Y) = \frac{K_{\text{В}}(X;Y)}{\sqrt{s^2(X) \cdot s^2(Y)}} = \frac{\sum_{i=1}^{n} (x_i - \overline{x_{\text{В}}}) \cdot (y_i - \overline{y_{\text{В}}})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x_{\text{В}}})^2 \cdot \sum_{i=1}^{n} (y_i - \overline{y_{\text{В}}})^2}}. \quad (13.3)$$

Для зручності обчислень на практиці будемо використовувати формулу (13.4):

$$\rho_{\text{П}}(X,Y)=\frac{\sum_{i=1}^{n}(x_i\cdot y_i)-\frac{(\sum_{i=1}^{n}x_i)\cdot(\sum_{i=1}^{n}y_i)}{n}}{\sqrt{\left(\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}\right)\cdot\left(\sum_{i=1}^{n}y_i^2-\frac{\left(\sum_{i=1}^{n}y_i\right)^2}{n}\right)}}. \quad (13.4)$$

Дану формулу легко отримати з формули (13.3) розкривши дужки та замінивши: $\overline{x_{\text{В}}}=\frac{\sum_{i=1}^{n}x_i}{n}$; $\overline{y_{\text{В}}}=\frac{\sum_{i=1}^{n}y_i}{n}$.

При цьому оцінка дисперсії характеризує ступінь розкиду значень навколо їх середнього або варіабельність та визначається за формулою (13.5):

$$s^2(X)=\frac{1}{n-1}\cdot\sum_{i=1}^{n}(x_i-\overline{x_{\text{В}}})^2. \quad (13.5)$$

В загальному випадку для отримання незміщеної оцінки дисперсії суму квадратів необхідно поділити на число ступенів свободи. Так як вибірка вже використовувалася один раз для визначення вибіркового середнього вибірки X, то число накладених зв'язків в даному випадку дорівнює одиниці, а число ступенів свободи $k=n-1$.

Однак, більш природно вимірювати ступінь розкиду значень змінних в тих же одиницях, в яких вимірюється і сама змінна. Цю задачу вирішує показник, що називається *середньоквадратичним відхиленням* (*стандартним відхиленням*) або *стандартною похибкою* і визначається співвідношенням (13.6):

$$s(X)=\sqrt{s^2(X)}. \quad (13.6)$$

Коефіцієнт кореляції Пірсона намагається встановити лінію, яка найкраще допасовується до набору даних із двох змінних, по суті викладаючи очікувані значення, а отриманий коефіцієнт кореляції Пірсона вказує, наскільки далеким від очікуваних значень є фактичний набір даних. Залежно від знаку нашого коефіцієнта кореляції Пірсона ми можемо отримати як від'ємну, так і додатну кореляцію, якщо якийсь зв'язок між змінними нашого набору даних існує (рис. 13.1).

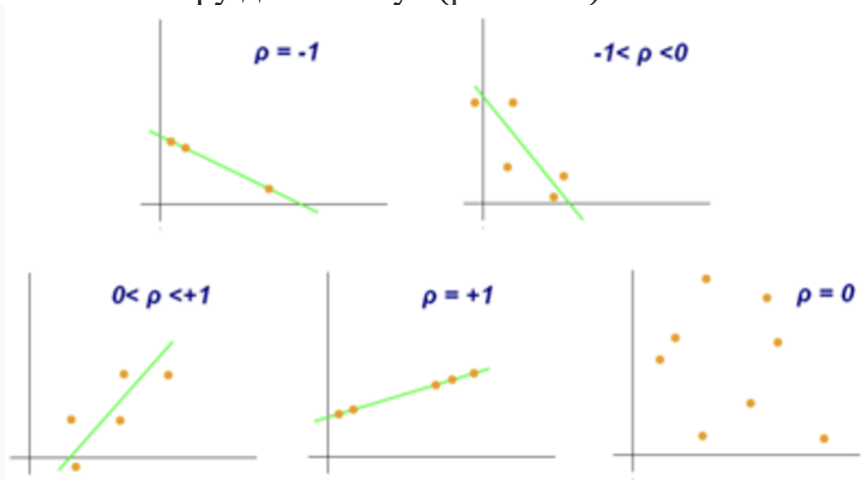

Рис. 13.1. Приклади діаграм розсіювання з різними значеннями коефіцієнту кореляції Пірсона

Під час інтерпретації коефіцієнта кореляції Пірсона необхідно враховувати такі моменти:

1. Коефіцієнт Пірсона можна використовувати для різних шкал (шкала відношень, інтервальна або порядкова) за винятком дихотомічної шкали.

Для якісної оцінки коефіцієнтів кореляції різні шкали, найбільш часто застосовується шкала Чеддока. Згідно неї в залежності від значень модуля коефіцієнта кореляції зв'язок може мати одну з оцінок:

Таблиця 13.1. *Шкала Чеддока виявлення сили зв'язку за значенням парного коефіцієнта кореляції*

Значення коефіцієнта кореляції (за модулем)	$0{,}1 - 0{,}3$	$0{,}3 - 0{,}5$	$0{,}5 - 0{,}7$	$0{,}7 - 0{,}9$	$0{,}9 - 0{,}99$
Характеристика сили зв'язку	слабкий	помірний	помітний	сильний	дуже сильний

2. Кореляційний зв'язок далеко не завжди означає зв'язок причинно-наслідковий. Інакше кажучи, якщо ми знайшли, припустімо, позитивну кореляцію між зростом і вагою в групи випробовуваних, то це зовсім не означає, що зріст залежить від ваги або навпаки (обидві ці ознаки залежать від третьої (зовнішньої) змінної, яка в цьому разі пов'язана з генетичними конституціональними особливостями людини).

3. $\rho_{\text{П}}(X;Y) = 0$ може спостерігатися не тільки за відсутності зв'язку між x і y, а й у разі сильного нелінійного зв'язку (рис. 13.2 в).

У цьому разі від'ємна і позитивна кореляції врівноважуються і в результаті створюється ілюзія відсутності зв'язку.

За наявності прямого (позитивного) зв'язку між змінними (рис. 13.2 а) хмара розсіювання має більш-менш сплощену еліптичну форму, довга вісь якої спрямована вправо і вгору. Іншими словами, у разі зростання значення однієї змінної є тенденція до збільшення іншої змінної.

У разі негативного зв'язку між змінними (рис. 13.2 б) довга вісь хмари розсіювання спрямована праворуч донизу, тобто збільшення значень однієї змінної відповідає закономірному зниженню значень іншої.

Зрештою, якщо хмара розсіювання має округлу форму (рис. 13.2 в), то можна припустити, що кореляція між змінними відсутня або, принаймні, вона дуже незначна.

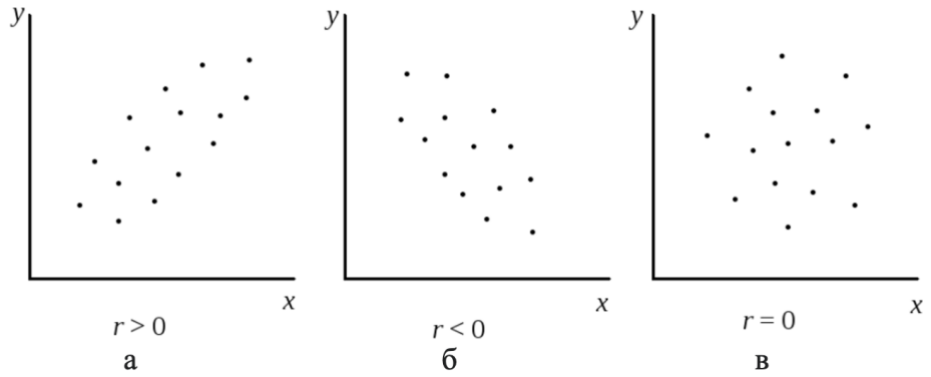

Рис. 13.2. Графічне представлення зв'язку між змінними (хмара розсіювання точок має різну форму залежно від характеру зв'язку

Варто відмітити, що величина коефіцієнта кореляції не є доказом того, що між ознаками, що досліджуються, є причинно-наслідковий зв'язок, а являє собою оцінку ступеня взаємної узгодженості в змінах ознак. Для того, щоб встановити причинно-наслідкову залежність необхідним є аналіз якісної природи явищ.

Оскільки оцінка тісноти зв’язку за допомогою коефіцієнта кореляції проводиться, як правило, на основі більш чи менш обмеженої інформації про явище, що досліджується, то виникає питання: наскільки правомірним є висновок по вибірковим даним про наявність кореляційного зв’язку в тій генеральній сукупності. з якої була отримана вибірка?

Нехай X і Y мають нормальний розподіл. У цьому випадку при досить великому об’ємі вибірки n коефіцієнт $\rho_{\text{П}}(X;Y)$ наближено дорівнює генеральному коефіцієнту $\rho(X;Y)$. Проте оцінити похибку, яка виникає при цьому, дуже важко. Це і не обов’язково, оскільки точне значення $\rho(X;Y)$ в розрахунках практично не використовується, а треба лише як показник наявності кореляції між Y і X. Вибірковий коефіцієнт кореляції $\rho_{\text{П}}(X;Y)$ застосовується в основному для перевірки загальної гіпотези про наявність кореляції між спостережуваними величинами, не вдаючись у детальні оцінки сили цієї кореляції.

Вплив перетворення даних на $\rho_{\text{П}}(X;Y)$:

1). Лінійні перетворення x і y типу $b \cdot x + a$ і $d \cdot y + c$ не змінять величину кореляції між x і y.

2). Лінійні перетворення x і y у разі $b < 0, d > 0$, а також у разі $b > 0$ і $d < 0$ змінюють знак коефіцієнта кореляції, не змінюючи його величини.

Достовірність (або, інакше, статистична значущість) коефіцієнта кореляції Пірсона може бути визначена різними способами:

- За таблицями критичних значень коефіцієнтів кореляції Пірсона. Якщо отримане в розрахунках значення $\rho_{\text{П}}(X;Y)$ перевищує критичне (табличне) значення для даної вибірки, коефіцієнт Пірсона вважається статистично значущим. Число ступенів свободи в цьому випадку відповідає $k = n - 2$, де n – число пар порівнюваних значень (об’єм вибірки).

- За таблицями критичних значень коефіцієнтів кореляції Пірсона. Якщо отримане в розрахунках значення $\rho_{\text{П}}(X;Y)$ перевищує критичне (табличне) значення для даної вибірки, коефіцієнт Пірсона вважається статистично значущим. Число ступенів свободи в цьому випадку відповідає $k = n - 2$, де n – число пар порівнюваних значень (об'єм вибірки).
- За коефіцієнтом Стьюдента, який обчислюється як відношення коефіцієнта кореляції до його похибки.

Похібка коефіцієнта кореляції обчислюється за формулою (13.7):

$$\delta = \sqrt{\frac{1-\rho_{\text{П}}^2(X;Y)}{n-2}}, \qquad (13.7)$$

де δ – помилка коефіцієнта кореляції, $\rho_{\text{П}}(X;Y)$ – коефіцієнт кореляції Пірсона; n - кількість порівнюваних пар.

Приклад 13.1. *Нехай є два набори даних:*

$$X = \{1; 2; 3; 4\};\ Y = \{2; 1; 4; 5\}.$$

Встановити силу кореляційного зв'язку між ними.

Розв'язання. 1. Для обчислення коефіцієнта кореляції Пірсона отримаємо наступну розрахункову таблицю (табл. 13.2):

Таблиця 13.2. *Розрахункова таблиця*

	x_i	y_i	x_i^2	y_i^2	$x_i \cdot y_i$
	1	2	1	4	2
	2	1	4	1	2
	3	4	9	16	12
	4	5	16	25	20
$\sum$	10	12	30	46	36

2. Проводимо обчислення та підставляємо значення в формулу (13.4):

$$\sum x_i = 10; \sum x_i^2 = 30; \sum y_i = 12; \sum y_i^2 = 46; \sum(x_i \cdot y_i) = 36.$$

$$\rho_{\text{П}}(X;Y) = \frac{36 - \frac{10 \cdot 12}{4}}{\sqrt{\left(30 - \frac{10^2}{4}\right) \cdot \left(46 - \frac{12^2}{4}\right)}} = \frac{6}{\sqrt{5 \cdot 10}} \approx 0{,}85.$$

3. Згідно шкали Чеддока між наборами даних є сильний додатний кореляційний зв'язок.

Відповідь. Між наборами даних є сильний додатний кореляційний зв'язок.

Приклад 13.2. 22 *старшокласники були протестовані за двома тестами:* РСК (*рівень суб'єктивного контролю*) *і* МдоУ (*мотивація до успіху*). *Отримано такі результати* (*табл.* 13.3):

Таблиця 13.3. *Дані тестувань старшокласників*

№	РСК (x_i)	МдоУ (y_i)	№	РСК (x_i)	МдоУ (y_i)
1	27	18	12	24	12
2	24	19	13	27	15
3	27	16	14	25	15
4	30	13	15	37	23
5	25	17	16	35	24
6	18	13	17	25	20
7	28	19	18	22	14
8	31	19	19	26	21
9	31	10	20	34	24
10	30	24	21	25	17
11	18	13	22	31	17

Перевірити гіпотезу про те, що для людей із високим рівнем інтернальності (бал РСК*) характерний високий рівень мотивації до успіху (бал* МдоУ*).*

Розв'язання. 1. Для обчислення коефіцієнта кореляції Пірсона отримаємо наступну розрахункову таблицю (табл. 13.4):

Таблиця 13.4. *Розрахункова таблиця*

x_i	y_i	x_i^2	y_i^2	$x_i \cdot y_i$
x_1	y_1	x_1^2	y_1^2	$x_1 \cdot y_1$
x_2	y_2	x_2^2	y_2^2	$x_2 \cdot y_2$
…	…	…	…	…
x_{22}	y_{22}	x_{22}^2	y_{22}^2	$x_{22} \cdot y_{22}$
$\sum x_i$	$\sum y_i$	$\sum x_i^2$	$\sum y_i^2$	$\sum (x_i \cdot y_i)$

2. Проводимо обчислення та підставляємо значення в формулу (13.4):

$\sum x_i = 600; \sum x_i^2 = 16864; \sum y_i = 383; \sum y_i^2 = 7025; \sum (x_i \cdot y_i) = 10689.$

$$\rho_{\text{П}}(X;Y)=\frac{10689-\frac{600\cdot 383}{22}}{\sqrt{\left(16864-\frac{600^2}{22}\right)\cdot\left(7025-\frac{383^2}{22}\right)}}\approx 0{,}58.$$

3. Визначимо статистичну значущість коефіцієнта кореляції Пірсона. Для цього обчислюємо похибку коефіцієнт Стьюдента, як відношення коефіцієнта кореляції Пірсона до похибки.

$$\delta=\sqrt{\frac{1-(0{,}58)^2}{22}}\approx 0{,}18;\ t_{\text{сп}}=\frac{\rho_{\text{П}}(X;Y)}{\delta}\approx 3{,}22.$$

У дод. 6 Гмурмана знаходимо стандартні значення коефіцієнта Стьюдента для 1-го, 2-го і 3-го рівнів значущості ($\alpha=0{,}05;0{,}01;0{,}001$) за числа ступенів свободи $k=n-2=20$: $t_{\text{кр}}=2{,}09;2{,}85;3{,}85$.

Відповідь: Кореляція між показниками тестів РСК і МдоУ є статистично значущою для 1-го і 2-го рівнів значущості.

13.2. Коефіцієнти рангової кореляції

Коефіцієнти рангової кореляції, такі як коефіцієнт рангової кореляції Спірмена та коефіцієнт рангової кореляції Кендалла (τ), вимірюють, до якої міри в разі збільшення однієї змінної інша змінна схильна збільшуватися, не вимагаючи, щоби це збільшення було подано лінійною залежністю. Якщо за збільшення однієї змінної інша зменшується, то коефіцієнти рангової кореляції будуть від'ємними. Ці коефіцієнти рангової кореляції часто розглядають як альтернативи коефіцієнту Пірсона, які використовують або для зменшення кількості обчислень, або для того, щоби зробити коефіцієнт менш чутливим до не нормальності в розподілах. Проте ця точка зору має мало математичних підстав, оскільки коефіцієнти рангової кореляції вимірюють інший тип зв'язку, ніж коефіцієнт кореляції Пірсона, і їх найкраще розглядати як показники іншого типу зв'язку, а не як альтернативну міру генерального коефіцієнту кореляції.

Щоб унаочнити природу рангової кореляції та її відмінність від лінійної кореляції, розгляньмо наступні чотири пари чисел $(x;y)$: $(0;1),(10;100),(101;500),(102;2000)$.

В міру просування від кожної пари до наступної x збільшується, й те саме робить y. Цей взаємозв'язок ідеальний, у тому сенсі, що збільшення в x завжди супроводжується збільшенням в y. Це означає, що ми маємо ідеальну рангову кореляцію, й обидва коефіцієнти кореляції Спірмена та Кендалла дорівнюють 1, тоді як у цьому прикладі коефіцієнт кореляції Пірсона дорівнює $0{,}7544$, вказуючи на те, що точки далеко не лежать на одній прямій. Так само, якщо y завжди зменшується, коли x збільшується, коефіцієнти рангової кореляції становитимуть -1, тоді як коефіцієнт кореляції Пірсона може бути або не бути близьким до -1, залежно від того, наскільки близько до прямої лінії розташовані ці точки. Хоча в граничних випадках ідеальної рангової кореляції ці два коефіцієнти рівні (чи то обидва

$+1$, чи обидва -1), зазвичай це не так, і тому значення цих двох коефіцієнтів неможливо порівнювати змістовно. Наприклад, для трьох пар (1,1) (2,3) (3,2) коефіцієнт Спірмена дорівнює $\frac{1}{2}$, а коефіцієнт Кендалла дорівнює $\frac{1}{3}$.

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 $(x_i; y_i)$ $i = 1,2,\dots,n$. Потрібно з'ясувати рівень кореляції між двома ознаками, x та y. Для цього використовують *коефіцієнт рангової кореляції Спірмена*. Цей показник розраховують за формулою (13.8):

$$\rho_{\text{В}}(X,Y) = \rho_{\text{С}}(X,Y) = 1 - \frac{6\cdot\sum_{i=1}^{n} d_i^2}{n\cdot(n^2-1)}, \qquad (13.8)$$

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 коефіцієнтом кореляції Пірсона, оскільки при його розрахунку не враховуються кількісні значення характеристик об'єктів, а лише їх порядок. Як і для коефіцієнта кореляції $\rho_{\text{П}}(X,Y)$, значення $\rho_{\text{С}}(X,Y)$ змінюється від – 1 до 1. Чим ближче абсолютне значення коефіцієнта рангової кореляції до одиниці, тим більш щільним є зв'язок між факторами.

У більш загальному випадку, якщо у вибірках X, Y присутні зв'язні ранги, тобто ранги елементів вибірки, які повторюються, коефіцієнт рангової кореляції Спірмена обчислюється за формулами (13.9) та (13.10):

$$\rho_{\text{С}}(X,Y) = 1 - \frac{6\cdot\sum_i d_i^2 + T_X + T_Y}{n\cdot(n^2-1)}, \qquad (13.9)$$

де

$$T_X = \frac{N_X^3 - N_X}{12} \text{ та } T_Y = \frac{N_Y^3 - N_Y}{12}, \qquad (13.10)$$

де N_X та N_Y, відповідно, кількість рангів, що повторюються, у вибірках X та Y. Для вибірки, яка не має повторюваних елементів, відповідний коефіцієнт із формули (13.9) дорівнює нулю, тобто отримуємо формулу (13.8).

Кроки для обчислення коефіцієнта кореляції Спірмена

Впорядкування даних: Присвоїти ранги кожному набору даних X і Y. Найменшому значенню присвоюється ранг 1, наступному — 2 і так далі.

Якщо є однакові значення (збіги), їм присвоюється середній ранг (між наступними двома номерами рангів).

Обчислення різниць рангів: Для кожної пари даних $(x_i; y_i)$ знайти різницю між рангами d_i.

Зведення різниць до квадрату: Обчислити квадрат різниць рангів для кожної пари d_i^2.

Обчислення коефіцієнтів T_X; T_Y: В разі наявності зв'язних рангів, дані коефіцієнти обчислюються за формулою (13.9).

Обчислення коефіцієнта кореляції Спірмена: Підставити знайдені значення в формулу (13.7) і обчислити коефіцієнт.

Застосування

- Коефіцієнт рангової кореляції Спірмена використовується тоді, коли:
- Дані не мають нормального розподілу або містять викиди.
- Зв'язок між змінними не є лінійним, але очікується монотонний.
- Дані представлені у вигляді ранжувань.

Приклад 13.3. *Нехай є два набори даних:*

$$X = \{85; 95; 80; 70; 60\};\ Y = \{80; 60; 90; 85; 70\}.$$

Встановити степінь кореляційного зв'язку між ними.

Розв'язання. Розставляємо ранги в обох наборах даних від найменшого до найбільшого.

Ранги для X: $R(X) = \{4; 5; 3; 2; 1\}$.

Ранги для Y: $R(Y) = \{3; 1; 5; 4; 2\}$.

Різниці рангів d_i: $d_i = \{1; 4; -2; -2; -1\}$.

Розставляємо ранги в обох наборах даних від найбільшого до найменшого.

Ранги для X: $R(X) = \{2; 1; 3; 4; 5\}$.

Ранги для Y: $R(Y) = \{3; 5; 1; 2; 4\}$.

Різниці рангів d_i: $d_i = \{-1; -4; 2; 2; 1\}$.

З чого можна зробити висновок, що d_i в обох випадках відрізнятимуться лише знаком, яким можна знехтувати при піднесенні до квадрату. Тобто спосіб ранжування можна обирати на власний розсуд.

Квадрати різниць d_i^2: $d_i^2 = \{1{,}16{,}4{,}4{,}1\}$,

$$\sum d_i^2 = 26.$$

Коефіцієнти T_X; T_Y: Оскільки зв'язних рангів немає, то $T_X = T_Y = 0$.

Обчислення коефіцієнта кореляції Спірмена (13.8):

$$\rho_C(X, Y) = 1 - \frac{6 \cdot 26}{5 \cdot 24} = 1 - 1{,}3 = -0{,}3.$$

Цей результат показує помірну від'ємну кореляцію між змінними X і Y.

Відповідь. Між наборами даних є помірний від'ємний кореляційний зв'язок.

Приклад 13.4. *Нехай є два набори даних:*

$$X = \{85; 95; 85; 70; 60\};\ Y = \{80; 70; 90; 80; 60\}.$$

Встановити степінь кореляційного зв'язку між ними.

Розв'язання. Розставляємо ранги в обох наборах даних від найбільшого до найменшого. Також необхідно врахувати повтори елементів в обох вибірках.

Ранги для X: $R(X) = \{2{,}5; 1; 2{,}5; 4; 5\}$.

Ранги для Y: $R(Y) = \{2{,}5; 4; 1; 2{,}5; 5\}$.

Різниці рангів d_i: $d_i = \{0; -3; 1{,}5; 1{,}5; 0\}$.

Квадрати різниць d_i^2: $d_i^2 = \{0; 9; 2{,}25; 2{,}25; 0\}$,

$$\sum d_i^2 = 13{,}5.$$

Коефіцієнти T_X; T_Y: Оскільки обидві вибірки мають по два елементи, які повторюються, тобто $N_X = N_Y = 2$, то $T_X = T_Y = \frac{2^3-2}{12} = \frac{1}{2}$.

Обчислення коефіцієнта кореляції Спірмена: За формулою (13.9):

$$\rho_C(X,Y) = 1 - \frac{6 \cdot 13{,}5 + 0{,}5 + 0{,}5}{5 \cdot 24} \approx 1 - 0{,}683 \approx 0{,}317.$$

Цей результат показує помірну додатну кореляцію між змінними X і Y.

Відповідь. Між наборами даних є помірний додатний кореляційний зв'язок.

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.

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:

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

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

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

Якщо ранги y не співпадають з рангами x, то знаходять суму $S = P + Q$. Відношення цієї суми до $P_{max} = |Q_{max}|$ дорівнює *коефіцієнту рангової кореляції Кендалла* (13.11):

$$\rho_{\text{B}}(X,Y) = \rho_{\text{K}}(X,Y) = \frac{2\cdot S}{n\cdot(n-1)} = \frac{4\cdot P}{n\cdot(n-1)} - 1 = 1 - \frac{4\cdot Q}{n\cdot(n-1)}. \quad (13.11)$$

Таким чином, для будь-яких двох пар даних (x_i, y_i) можна зробити такі висновки:

Узгоджена пара: Якщо $x_i < x_j, y_i < y_j$, або якщо $x_i > x_j, y_i > y_j$.

Кількість узгоджених пар дорівнює P.

Неузгоджена пара: Якщо $x_i < x_j, y_i > y_j$, або $x_i > x_j, y_i < y_j$.

Кількість неузгоджених пар дорівнює Q (береться зі знаком « − »).

Нейтральна пара: Якщо $x_i = x_j$ або $y_i = y_j$ (такі пари не враховуються у підрахунках).

Отже, коефіцієнт кореляції Кендалла визначається як різниця між кількістю узгоджених пар і неузгоджених пар, нормована до кількості можливих пар.

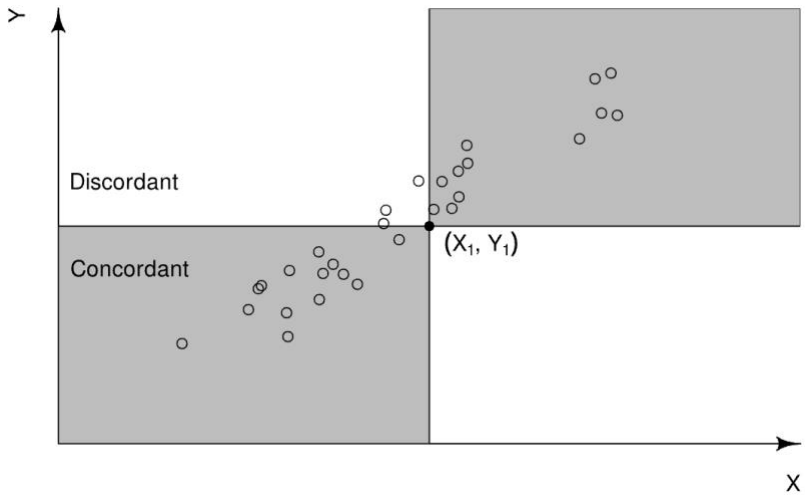

Рис. 13.5. Візуалізація узгоджених та неузгоджених з точкою (X_1, Y_1) пар

На рис. 13.5. усі точки в сірих прямокутниках є узгодженими, а всі точки в білих прямокутниках є неузгодженими з точкою (X_1, Y_1). Загалом на графіку є $n = 30$ точок, які утворюють $C_{30}^2 = 435$ можливих пар, 395 з цих пар є узгодженими, 40 пар — неузгодженими, що дає коефіцієнт кореляції рангу Кендалла 0,816.

Приклад 13.5. *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* (*табл.* 13.5). *З допомогою коефіцієнта рангової кореляції Кендалла виміряти щільність взаємозв'язку між цими показниками.*

Таблиця 13.5. *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	−	−
$n = 10$				$P = 38$	$Q = -7$

Розв'язання. 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 $S = P + Q = 38 - 7 = 31$.

Звідси знаходимо коефіцієнт кореляції Кендалла (13.11):

$$\tau(X, Y) = \frac{2 \cdot 31}{10 \cdot 9} = \frac{62}{90} \approx 0{,}69.$$

Отримане значення коефіцієнта рангової кореляції, згідно шкали Чеддока, свідчить про помітний кореляційний зв'язку між показниками x та y.

Відповідь. Між показниками є помітний додатний кореляційний зв'язок.

Приклад 13.6. *У розрахунковій таблиці* (*табл.* 13.6) *наведено дані щодо погодинної оплати праці на підприємстві* x (*г.о.*) *та рівня плинності кадрів*

у (*кількість працівників, що звільнилися за рік*). *Розрахувати коефіцієнт рангової кореляції Кендалла для цих даних*.

Таблиця 13.6. *Розрахункова таблиця для визначення коефіцієнта рангової кореляції Кендалла*

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	–	–
$n = 8$				$P = 2$	$Q = -26$

Розв'язання. Розрахункова таблиця для визначення коефіцієнта рангової кореляції Кендалла при протилежній спрямованості рангів має вигляд (табл. 13.6):

За даними розрахункової табл. 13.6 отримуємо $P = 2, Q = -26$, $S = P + Q = 2 - 26 = -24$.

Звідси знаходимо коефіцієнт рангової кореляції Кендалла (13.11):

$$\tau(X,Y) = \frac{2\cdot(-24)}{8\cdot7} = \frac{4\cdot2}{8\cdot7} - 1 = 1 - \frac{4\cdot26}{8\cdot7} = -\frac{6}{7} \approx -0{,}86.$$

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

Відповідь. Між показниками є сильний від'ємний кореляційний зв'язок.

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.

13.3. Коефіцієнт кореляції Фехнера

Коефіцієнт кореляції Фехнера, запропонований у 2-й половині XIX століття Г.Т. Фехнером, є найпростішою мірою зв'язку між двома змінними. Він ґрунтується на зіставленні двох психологічних ознак x_i і y_i, виміряних на одній і тій самій вибірці, за зіставленням знаків відхилень індивідуальних значень від середнього: $s(x_i - \overline{x_B})$ і $s(y_i - \overline{y_B})$. Висновок про кореляцію між двома змінними роблять на підставі підрахунку числа збігів і розбіжностей цих знаків.

Коефіцієнт кореляції Фехнера обчислюється за формулою (13.12):

$$\rho_B(X,Y) = \rho_\Phi(X,Y) = \frac{n_a - n_b}{n_a + n_b}, \qquad (13.12)$$

де a — збіги знаків, b — розбіжності знаків; n_a — кількість збігів, n_b — кількість розбіжностей

Приклад 13.7. *Нехай x_i і y_i – дві ознаки, виміряні на одній і тій самій вибірці випробовуваних. За допомогою коефіцієнта кореляції Фехнера встановити рівень кореляційного зв'язку між ними.*

Розв'язання. Для обчислення коефіцієнта Фехнера необхідно обчислити середні значення для кожної ознаки, а також для кожного значення змінної – знак відхилення від середнього (табл. 13.7):

Таблиця 13.7. *Розрахункова таблиця для обчислення коефіцієнта кореляції Фехнера*

№	x_i	y_i	$s(x_i - \overline{x_В})$	$s(y_i - \overline{y_В})$.	Позначення
1	16	20	+	+	a
2	15	17	–	–	a
3	19	16	+	–	b
4	12	22	–	+	b
5	9	18	–	+	b
6	20	12	+	–	b
7	18	15	+	–	b
8	14	18	–	+	b
9	15	16	–	–	a
10	17	18	+	+	a

Згідно таблиці (13.7) обчислюємо вибіркові середні значення $\overline{x_В} = 15{,}5$; $\overline{y_В} = 17{,}2$. Тоді $n_a = 4$, $n_b = 6$. За формулою (13.12):

$$\rho_Ф(X, Y) = \frac{4 - 6}{4 + 6} = -\frac{1}{5} = -0{,}2.$$

Відповідь. Між показниками є слабкий від'ємний кореляційний зв'язок.

13.4. Розробка алгоритму формування портфеля послуг згідно з принципом об'єднання на основі статистичного аналізу інформації про надання послуг

Математична постановка задачі: задано множину послуг $P = \{P_j\}$, $j = \overline{1, m}$. Кожна послуга P_j з множини P представлена набором значень:

$$P_j = (P_{j1}, \dots, P_{ji}, \dots, P_{jn}),$$

де P_{ji} – кількість звернень по послугу P_j $(j = \overline{1, m})$ за день i $(j = \overline{1, m})$, n – кількість днів в періоді, за який здійснюється аналіз, m – кількість послуг.

Необхідно на основі наявної інформації встановити взаємозв'язок між

послугами множини P. На основі аналізу сили встановленого взаємозв'язку сформувати рекомендації щодо зв'язування послуг в один або декілька портфелів послуг.

Теоретичне обгрунтування підходу: кожна послуга P_j з множини P може розглядатись як дискретна випадкова величина, яка представлена своїм рядом розподілу. Тоді взаємозв'язок між цими величинами можна встановити на основі аналізу коефіцієнту парної кореляції Пірсона (13.3).

Сила кореляціційних зв'язків між послугами встановлюється згідно шкали Чеддока (лекція 12). При формуванні рекомендацій щодо об'єднання послуг доцільно враховувати зв'язки, сила яких визначається діапазоном $0{,}7 \le \rho_{\text{П}}(P_k, P_j)) \le 1$.

Слід зазначити, що для отримання достовірних результатів аналізу необхідно встановлювати статистичний зв'язок між послугами, використовуючи тільки статистично значущі значення коефіцієнтів кореляції.

Під *статистично значущім кореляційним зв'язком* між парами послуг мається на увазі ті пари послуг, замовлення однієї з яких добре узгоджується із замовленням іншої, тобто сила зв'язку встановлюється не тільки кількісно (значення коефіцієнта кореляції), а й якісно (що обидві послуги замовлять ті самі люди).

Перевірити статистичну значущість розрахованого коефіцієнта кореляції можна, обчисливши значення t-критерія Стьюдента (13.13):

$$t_{\text{сп } k,j} = \rho_{\text{П}}(P_k, P_j) \cdot \sqrt{\frac{n-2}{1-\rho_{\text{П}}^2(P_k;P_j)}}, \qquad (13.13)$$

де $\rho_{\text{П}}(P_k, P_j)$ – коефіцієнт кореляції випадкових величин P_k, P_j, n – розмір вибірки (кількість днів в періоді), та порівнявши його із критичним табличним значенням цього критерія $t_{\text{кр}}$ (яке береться із таблиці розподілу Стьюдента (дод. 6 Гмурмана) із врахуванням заданого рівня значущості $\alpha = 0{,}05$, що є достатнім рівнем для отримання достовірних результатів, та кількості ступенів свободи $k = n - 2$).

Якщо розраховане значення $|t_{\text{сп } k,j}| > t_{\text{кр}}$, то відповідний коефіцієнт кореляції є статистично значущим.

Проведення статистичного аналізу для сформованого датасету

Для встановлення наявності зв'язку між кожною парою послуг побудовано матрицю парних коефіцієнтів кореляції, яка представлена на рис. 13.6. Враховуючи симетричність матриці відносно головної діагоналі, на рис. 13.6 наведено половину матриці. Значення коефіцієнтів, які відповідають помітним та сильним зв'язкам, виділено червоним кольором. Так, значення коефіцієнта кореляції для пари послуг 34 та 37 дорівнює 0,7, що згідно зі шкалою Чеддока, свідчить про наявність сильного статистичного зв'язку між цими послугами.

В загальному випадку вибір послуг для аналізу рекомендовано

здійснювати за кількістю звернень, що перевищують вибіркове середнє. В прикладі вибрано 24 послуги з 84, кількість звернень за цими послугами складає 2173, або 91% від загальної кількості звернень.

	2	3	5	6	7	13	14	24	26	30	34	37	40	45	46	47	50	56	63	69	71	74	76	84
2		0	0,2	0,5	0	-0	0,2	-0	0,3	-0	-0	-0	0,2	-0	-0	-0	-0	-0	0	0,4	0,2	0	-0	-0
3			0,1	0	0,1	0,2	-0	-0	-0	0,1	-0	-0	-0	0,1	0,1	0,1	0	-0	-0	0,1	-0	0,1	0	0,2
5				0,3	-0	0,2	0,1	0,1	0,2	0,5	0,2	0,3	-0	0	-0	0,1	-0	0	0,4	0,4	0,1	0,2	-0	0,2
6					-0	0,2	0,1	0,1	0,3	-0	0,2	0,4	0,4	0,4	-0	0,2	-0	0,1	0,1	0,5	0,2	0,1	-0	0,4
7						0,1	0,2	0,1	0,1	-0	-0	-0	-0	-0	-0	0,3	-0	0,2	-0	-0	0,2	-0	0,2	0,4
13							0,1	-0	0,3	0,3	-0	0	0,3	0,1	-0	0,1	0,3	0,4	-0	0,1	0,1	0,1	0,1	0,4
14								0,1	0,1	0,1	-0	0,1	0,3	0,2	0	0,4	0,3	-0	0,5	0,3	0,3	0,2	0,5	-0
24									0,3	-0	0	0	-0	0,2	0,4	0,2	-0	0,3	0	0,3	0,4	0	-0	0,4
26										-0	-0	-0	-0	-0	0,2	0,2	-0	0,1	-0	0,4	0	0,1	-0	0,4
30											0,1	0,1	-0	-0	0	0,2	0,1	0,3	-0	0,1	0,2	0,3	0,2	-0
34												0,7	0,2	0,2	0,3	0,1	-0	-0	0,1	-0	0	0	-0	0
37													0,2	0,5	0	0,1	-0	0,1	0,4	0,1	0	-0	0	0,2
40														0,2	0,1	0,2	0,4	-0	0,1	-0	-0	0	0,2	-0
45															0,2	0,1	0,4	0,2	0,3	0,2	-0	-0	0,5	0,1
46																0,3	0	-0	-0	-0	0	0	-0	0,1
47																	-0	-0	-0	0,1	0,1	0	0,2	0,3
50																		0,1	0,2	0,1	-0	0,3	0,6	-0
56																			-0	0,1	0,5	0,2	0	0,3
63																				0,3	-0	0	0,1	-0
69																					0,4	-0	0,2	0
71																						0,2	-0	0,3
74																							-0	0
76																								-0
84																								

Рис. 13.6. Матриця парних коефіцієнтів кореляції

Для отримання достовірних результатів аналізу необхідно перевірити гіпотезу про статистичну значущість отриманих значень коефіцієнтів кореляції з рис. 13.6. Для цього вибирається табличне значення критерія Стьюдента для $n = 24$ (ступінь свободи $k = 22$), $\alpha = 0{,}05$, яке дорівнює $t_{\text{сп}} = 2{,}07$.

На рис. 3 наведено матрицю розрахованих за формулою (13.13) значень t-критерія для перевірки статистичної значущості розрахованих коефіцієнтів кореляції (тут $n = 24$, r – коефіцієнт кореляції, значущість якого перевіряється).

Якщо розраховане значення $|t_{\text{сп}}| > t_{\text{кр}}$, то відповідний коефіцієнт кореляції є статистично значущим (в матриці на рис. 13.7 такі значення t-критерія виділено червоним кольором).

	2	3	5	6	7	13	14	24	26	30	34	37	40	45	46	47	50	56	63	69	71	74	76	84
2		0,1	1	2,4	0,1	0,8	0,9	0,3	1,3	0,8	0,8	0,8	0,7	0,1	1,1	0,9	1,1	0,6	0,1	2,2	0,9	0	0,3	0,2
3			0,5	0	0,5	1,2	0,3	0,5	0,3	0,2	1,6	1,4	0,2	0,3	0,5	0,2	0	0,3	1,5	0,2	0,2	0,3	0,2	0,9
5				1,5	2,1	1	0,3	0,6	1	2,4	0,8	1,3	1	0,2	0,3	0,2	1,2	0,1	2,1	1,8	0,3	1	1,3	0,7
6					0,5	0,9	0,5	0,3	1,6	0,7	1	2,2	2	1,9	0,5	0,9	0,2	0,3	0,4	2,6	0,8	0,4	0,3	1,9
7						0,7	0,9	0,5	0,6	0,9	2,3	1	0,6	0,7	0,9	1,4	0,9	1,2	1,5	0,3	1	1,6	1	2,2
13							0,5	0,1	1,4	1,3	0,7	0,1	1,3	0,4	0,5	0,4	1,5	1,9	0,3	0,5	0,3	0,6	0,3	1,9
14								0,5	0,6	0,4	0,8	0,5	1,3	1,2	0,1	2,3	1,6	0,3	2,4	1,5	1,3	0,9	2,5	0,1
24									1,4	0,3	0,1	0,2	2,3	1,1	1,9	0,7	0,7	1,3	0,1	1,5	2,1	0	0,8	2,1
26										0,1	1,2	0,8	0,7	0,9	1	1,1	0,2	0,3	0,6	2,2	0,1	0,6	1,1	1,8
30											0,6	0,6	0,1	0,3	0,2	1	0,3	1,4	0	0,3	0,7	1,2	1	0,1
34												4,7	0,9	1,1	1,4	0,3	0,4	0,3	0,7	0,2	0	0	0,6	0,1
37													1	2,4	0,2	0,5	0,3	0,5	1,9	0,4	0,1	0,6	0,1	1,1
40														1,1	0,3	0,7	1,9	1,1	0,4	0,1	0,8	0	1,2	1
45															0,8	0,3	1,9	1	1,3	1,2	0,2	0,3	2,5	0,4
46																1,3	0,2	0,6	1,2	0,3	0,1	0,1	0,6	0,4
47																	0,5	0,7	0,1	0,5	0,6	0,1	1,1	1,3
50																		0,5	1	0,3	1,2	1,4	3,3	0,8
56																			1	0,5	2,6	1	0,2	1,6
63																				1,3	1,1	0,2	0,6	1,6
69																					1,8	0,3	1	0,1
71																						1	0	1,3
74																							0,7	0,2
76																								1
84																								

Рис. 13.7. Матриця коефіцієнтів критерія Стьюдента

Згідно з даними, наведеними в матриці на рис. 13.7, можемо побачити, що сильний зв'язок послуг 37 та 34, є статистично значущим. Послуга 76 є помітно пов'язаною з послугою 50. Про це свідчить значення коефіцієнта кореляції 0,6 для цієї пари послуг, що видно із рис. 13.7. Цей зв'язок також є статистично значущим, про що свідчить значення t-критерія для цієї пари послуг на рис. 13.7.

Перелік послуг, пов'язаних із послугами 37 та 76 згідно проведеного аналізу, наведено в табл. 13.8, 13.9 відповідно.

Таблиця 13.8. *Перелік послуг, пов'язаних із послугою* 37 «*Державна реєстрація інших* (*відмінних від права власності*) *речових прав на нерухоме майно*»

№	Коефіцієнт кореляції	Номер послуги	Назва послуги
1	0,71143376	34	Державна реєстрація права власності на нерухоме майно
2	0,45735186	45	Надання відомостей з Державного земельного кадастру у формі копій документів, що створюються під час ведення Державного земельного кадастру.
3	0,429127614	6	Видача довідки до нотаріальної контори про реєстрацію місця проживання громадянина за даною адресою на день смерті (для оформлення спадщини та інше)

Таблиця 13.9. *Перелік послуг, пов'язаних із послугою* 76 «*Призначення грошової компенсації вартості одноразової натуральної допомоги* «*пакунок малюка*»»

№	Коефіцієнт кореляції	Номер послуги	Назва послуги

1	0,576879259	50	Надання державної допомоги при народженні дитини
2	0,470902426	14	Видача довідки про склад зареєстрованих у житловому приміщенні осіб
3	0,46632422	45	Надання відомостей з Державного земельного кадастру у формі копій документів, що створюються під час ведення Державного земельного кадастру

В табл. 13.8, 13.9 наведено значення коефіцієнтів кореляції, які є статистично значущими.

Для формування портфелів послуг на основі проведеного аналізу вибираються послуги, для яких встановлено наявність сильного статистично значущого зв'язку.

Для розглянутого прикладу в результаті формуються рекомендації щодо створення двох портфелів послуг – один включає послуги 34 і 37, а інший – послуги 76 та 50.

13.5. Встановлення рівня залежності та її статистичної значущості між прогнозованими та реальним курсами криптовалюти

Встановлення рівня залежності між прогнозованими та реальним курсами криптовалюти

Постановка задачі: Нехай задано пари статистичних вибірок (X, Y_i), $i = \overline{1;3}$ (X – реальні курси криптовалюти за певний проміжок часу; Y_i – прогнози курсу криптовалюти в аналогічні моменти часу, що зроблені за допомогою алгоритмів АУДСМ, ARIMA та Експоненційного згладжування відповідно).

Необхідно встановити рівень залежності між парами вибірок (X, Y_i), враховуючи порядок слідування елементів у вибірці.

Обгрунтування: Загалом, для встановлення рівня залежності між двома вибірками використовується коефіцієнт кореляції Пірсона (13.3).

Оскільки для даного дослідження істотним є порядок слідування елементів у вибірках (реальний та прогнозований курси криптовалют розглядаються і порівнюються у відповідний момент часу), то рекомендовано використовувати коефіцієнт рангової кореляції, зокрема коефіцієнти Спірмена, яка обчислюється за формулами (13.9) та (13.10).

Слід зазначити, що коефіцієнт рангової кореляції Спірмена (13.9) є менш точними у порівнянні з коефіцієнтом кореляції Пірсона (13.3), оскільки при його розрахунку не враховуються кількісні значення елементів вибірок, а лише їх порядок. Це свідчить про необхідність враховувати коефіцієнт кореляції Пірсона при більш глибокому визначенні рівня залежності вибірок X та Y_i, $i = \overline{1;3}$. Його встановлення відбувається за допомогою шкали Чеддока (лекція 12).

Запропонований підхід дає змогу оцінити наскільки тісний кореляційний зв'язок існує між реальними та прогнозованими курсами обраної криптовалюти за вказаний проміжок часу.

Результат: Для вибірок, результати представляються у наступному вигляді (табл. 13.10).

Таблиця 13.10. *Аналіз кореляційного зв'язку між парами вибірок* (X, Y_i), $i = \overline{1;3}$

Вибірка	Коефіцієнт рангової кореляції Спірмена $\rho_C(X, Y_i)$	Коефіцієнт рангової кореляції Кендалла $\rho_K(X, Y_i)$	Коефіцієнт кореляції Пірсона $\rho_П(X, Y_i)$	Сила зв'язку з вибіркою X
Y_1	0,97828283	0,9111111	0,99829258	Дуже сильний зв'язок як з точки зору подібності елементів вибірок, так і з точки зору врахування їх порядку слідування
Y_2	0,39040404	0,3333333	0,581244474	Помітний зв'язок з точки зору подібності елементів вибірок, та помірний з точки зору врахування їх порядку слідування
Y_3	0,09949495	0,0666667	0,58275805	Помітний зв'язок з точки зору подібності елементів вибірок, але майже відсутній з точки зору врахування їх порядку слідування

В даній таблиці коефіцієнти кореляції обчислюються за формулами (13.3), (13.9), (13.10), (13.11), а сила зв'язку з вибіркою X встановлюється за допомогою табл. 3.

З табл. 13.10 видно, що вибірка Y_1 має найсильніший кореляційний зв'язок з вибіркою X серед усіх розглянутий вибірок (дуже сильний зв'язок). Середній рівень зв'язку з вибіркою X має вибірка Y_2, найгірші показники має вибірка Y_3.

Встановлення статистичної значущості прогнозованих курсів криптовалюти

Постановка задачі: Нехай задана пара вибірок $(X, Y_i), i = \overline{1;3}$. Між ними відомий коефіцієнти рангової кореляції Спірмена, Кендалла або Пірсона.

Потрібно при рівні значущості α перевірити гіпотезу про значущість відповідного коефіцієнта кореляції.

Обґрунтування: Для розв'язання сформульованої задачі, використовувалися наступні правила, які входять до t-критерію Стьюдента.

Правило 1. Для того, щоб при рівні значущості α перевірити гіпотезу про значущість коефіцієнтів рангової кореляції Спірмена необхідно обчислити критичне значення (13.14):

$$T_{\text{кр } i} = t_{\text{кр}} \cdot \sqrt{\frac{1-\rho_C^2(X,Y_i)}{n-2}}, \qquad (13.14)$$

де критичне значення $t_{\text{кр}}$ знаходиться по таблиці розподілу Стьюдента (дод. 6 Гмурмана) за рівнем значущості α та ступенем свободи $k = n - 2$, $\rho_C(X, Y_i)$ – коефіцієнти рангової кореляції Спірмена, $i = \overline{1;3}$, n – об'єм вибірок X та Y_i.

Якщо $|\rho_C(X, Y_i)| \le T_{\text{кр } i}$, то гіпотеза приймається, в противному разі – відхиляється, тобто між вибірками існує значущий кореляційний зв'язок.

Правило 2. Для того, щоб при рівні значущості α перевірити гіпотезу про значущість коефіцієнтів рангової кореляції Кендалла $\rho_{\text{К}}(X, Y_i)$, $i = \overline{1;3}$, необхідно обчислити критичне значення (13.15):

$$T_{\text{кр}\,i} = T_{\text{кр}} = z_{\text{кр}} \cdot \sqrt{\frac{2\cdot(2\cdot n+5)}{9\cdot n\cdot(n-1)}}, \qquad (13.15)$$

де при заданому рівні значущості α необхідно відшукати критичну точку $z_{\text{кр}}$ з рівняння $\Phi\left(z_{\text{кр}}\right) = \frac{1-\alpha}{2}$ (див. дод. 2 Гмурмана).

Якщо $|\rho_{\text{К}}(X, Y_i)| \le T_{\text{кр}}$, то гіпотеза приймається, в противному разі – відхиляється, тобто між вибірками існує значущий кореляційний зв'язок.

Правило 3. Для того, щоб при рівні значущості α перевірити гіпотезу про значущість коефіцієнта кореляції Пірсона необхідно обчислити спостережуване значення (13.16):

$$t_{\text{сп}\,i} = \rho_{\text{П}}(X, Y_i) \cdot \sqrt{\frac{n-2}{1-\rho_{\text{П}}^2(X,Y_i)}}, \qquad (13.16)$$

де n – об'єм вибірок X та Y_i, а $\rho_{\text{П}}(X, Y_i)$ – коефіцієнти кореляції Пірсона, $i = \overline{1;3}$.

Формула (13.16) є аналогічною до формули (13.13). Різниця полягає у позначенні вибірок.

Далі необхідно порівняти його із табличним критичним значенням цього критерія $t_{\text{кр}}\,(\alpha, k)$ (яке береться із таблиці розподілу Стьюдента (дод. 6 Гмурмана) з урахуванням заданого рівня значущості α, що є достатнім рівнем для отримання достовірних результатів, та кількості ступенів свободи $k = n-2$).

Якщо $|t_{\text{сп}\,i}| \le t_{\text{кр}}(\alpha, k)$, то гіпотеза приймається, інакше – відхиляється, тобто між вибірками існує значущий кореляційний зв'язок.

Під *значущим кореляційним зв'язком* мається на увазі, що прогнозовані курси криптовалюти, тобто вибірки Y_i добре узгоджуються з реальними курсами в аналогічні моменти часу, тобто з вибіркою X. Зокрема, якщо розглянути пару вибірок (X, Y_1), в рамках запропонованого підходу можна простежити, на скільки прогноз курсу криптовалюти з урахуванням дописів відомих людей в соціальних мережах корелюється з реальними курсами валют за аналогічний період часу. Це дає змогу оцінити рівень впливу цих дописів на курси криптовалюти, що і є основною метою даного дослідження.

Результат: Для вибірок Y_i з урахуванням коефіцієнтів рангової кореляції, наведених в табл. 13.10, результати представляються у наступному вигляді (табл. 13.11-13.13):

Таблиця 13.11. *Результати перевірки гіпотези про значущість коефіцієнта рангової кореляції Спірмена між парами вибірок* (X, Y_i) *при* $\alpha = 0{,}05$ *і* $t_{\text{кр}}(0{,}05; 8) = 2{,}31$

Вибірка	Коефіцієнт рангової кореляції Спірмена $\rho_{\text{С}}(X, Y_i)$	Значення $T_{\text{кр}\,i}$	Висновки про прийняття або відкидання гіпотези
Y_1	0,97828283	0,16	значущий кореляційний зв'язок
Y_2	0,39040404	0,75	незначущий кореляційний зв'язок
Y_3	0,09949495	0,81	незначущий кореляційний зв'язок

Таблиця 13.12. *Результати перевірки гіпотези про значущість коефіцієнта кореляції Кендалла між парами вибірок* (X, Y_i) *при* $n = 10;\ \alpha = 0{,}05$ і $z_{\text{кр}}(0{,}05) = 1{,}96$

Вибірка	Коефіцієнт кореляції Кендалла $\rho_{\text{К}}(X, Y_i)$	Значення $T_{\text{кр}}$	Висновки про прийняття або відкидання гіпотези
Y_1	0,9111111	0,49	значущий кореляційний зв'язок
Y_2	0,3333333	0,49	незначущий кореляційний зв'язок
Y_3	0,0666667	0,49	незначущий кореляційний зв'язок

Таблиця 13.13. *Результати перевірки гіпотези про значущість коефіцієнта кореляції Пірсона між парами вибірок* (X, Y_i) *при* $\alpha = 0{,}05$ і $t_{\text{кр}}(0{,}05; 8) = 2{,}31$

Вибірка	Коефіцієнт кореляції Пірсона $\rho_{\text{П}}(X, Y_i)$	Значення $t_{\text{сп}\,i}$	Висновки про прийняття або відкидання гіпотези
Y_1	0,99829258	48,34	значущий кореляційний зв'язок
Y_2	0,581244474	2,02	незначущий кореляційний зв'язок
Y_3	0,58275805	2,03	незначущий кореляційний зв'язок

В даних таблицях коефіцієнти кореляції обчислюються за формулами (13.3), (13.9), (13.10) та (13.11), а критичні значення – за формулами (13.14), (13.15) та (13.16) відповідно.

З табл. 13.11-13.13 видно, що в парі вибірок (X, Y_1) переважно існує значущий кореляційний зв'язок, в той час, як в парах вибірок (X, Y_2) та (X, Y_3) кореляційний зв'язок є незначущим.

Більш докладно про статистичні критерії розглядатиметься в лекції 18.