МІНІСТЕРСТВО ВНУТРІШНІХ СПРАВ УКРАЇНИ
ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ВНУТРІШНІХ СПРАВ

Факультет № 6
Кафедра соціології та психології

ТЕКСТ ЛЕКЦІЇ

з навчальної дисципліни **«Математичні методи в психології»**
обов'язкових компонент
освітньої програми першого (бакалаврського) рівня вищої освіти

053 Психологія (практична психологія)

Тема №6. Методи кореляційного аналізу.

Харків 2023

ЗАТВЕРДЖЕНО
Науково-методичною радою
Харківського національного
університету внутрішніх справ
Протокол від 30.08.2023 № 7

СХВАЛЕНО
Вченою радою факультету № 6
Протокол від 25.08.2023 № 7

ПОГОДЖЕНО
Секцією Науково-методичної
ради ХНУВС гуманітарних та
соціально- економічних дисциплін
Протокол від 29.08.2023 № 7

Розглянуто на засіданні кафедри соціології та психології (протокол №8 від 15.08.2023)

Розробник:
Доцент кафедри соціології та психології, кандидат психологічних наук, доцент Твердохвалова Ю.Л.

Рецензенти:
1. Професор кафедри психології Харківського національного педагогічного університету імені Г.С. Сковороди, доктор психологічних наук, професор, Кузнєцов М.А.
2. Доцент кафедри соціології та психології факультету № 6 Харківського національного університету внутрішніх справ, кандидат психологічних наук, доцент Греса Н.В.

План лекції

6.1. Сутність методів встановлення статистичних взаємозв'язків. Основні властивості коефіцієнта кореляції

6.2. Кореляція метричних змінних

6.3. Кореляція рангових змінних

6.4. Кореляція дихотомічних змінних

Рекомендована література (основна, допоміжна), інформаційні ресурси в Інтернеті

Основна:

1. Климчук В.О. Математичні методи у психології. Навчальний посібник для студентів психологічних спеціальностей. Київ : Освіта України. 2009. 288 с.
2. Телейко А.Б. Чорней Р.К. Математико-статистичні методи в соціології та психології : Навч. посібник. Київ : МАУП, 2007. 424 с.
3. Руденко В.М., Руденко Н.М. Математичні методи в психології : підручник. Київ : Академвидав, 2009. 384 с.

Допоміжна:

1. Літнарович Р.М. Основи математичної статистики у психології : Навчальний посібник. Ч.3. Рівне : МЕГУ, 2006. 49 с.
2. Татьянчиков А.О. Методичні рекомендації до виконання лабораторних робіт з курсу «Методи психологічного дослідження: математичні методи в психології». Одеса : Вид-во Університету Ушинського, 2019. 38 с.
3. Климчик В.О. Кластерний аналіз: використання в психологічних дослідженнях// Практична психологія та соціальна робота. 2006. №4. С. 30-36.
4. Циба В.Т. Математичні основи соціологічних досліджень: кваліметричний підхід. - К.: МАУП, 2002. - 248 с.
5. Климчук В.О. Викладання курсу "Математичні методи у психології" в умовах кредитно-модульної системи // Соціальна психологія. 2008. №2 (28). С. 180-189.

Текст лекції

Одним із найважливіших завдань всякого дослідження, зокрема і психолого-педагогічного, є встановлення зв'язку між величинами або факторами, зміна яких визначає сутність процесу, що вивчається. Щоб пізнати яке-небудь явище, треба вивчити не тільки його зв'язки з навколишніми явищами (факторами), але й також взаємозв'язки всіх його сторін, тобто треба встановити закономірності змін взаємопов'язаних явищ і показників, що їх характеризують.

В роботі практичного психолога часто необхідно аналізувати залежність між двома або декількома змінними величинами (ознаками). Якщо дві деякі характеристики отримані для одного і того ж «об'єкта», мають тенденцію змінюватися сумісно так, що створюється можливість передбачити одну з них за значенням іншої, то кажуть, що ці характеристики **корелюють** одна з одною. Відповідно в статистиці кореляція виражає степінь взаємозв'язку між такими характеристиками. Кількісно ця степінь взаємозв'язку виражається за допомогою **коефіцієнта кореляції**.

6.1. Сутність методів встановлення статистичних взаємозв'язків. Основні властивості коефіцієнта кореляції

Міри зв'язку виявляють співвідношенням, як правило, між двома змінними, які виміряні на одній вибірці. Ці зв'язки визначають через обчислення коефіцієнтів кореляції.

Термін «кореляція» вперше застосував французький палеонтолог Ж. Кюв'є, який вивів «закон кореляції частин і органів тварин» (цей закон дозволяє відновлювати за знайденими частинам тіла вигляд всієї тварини). У статистику зазначений термін ввів англійський біолог і статистик Ф. Г альтон.

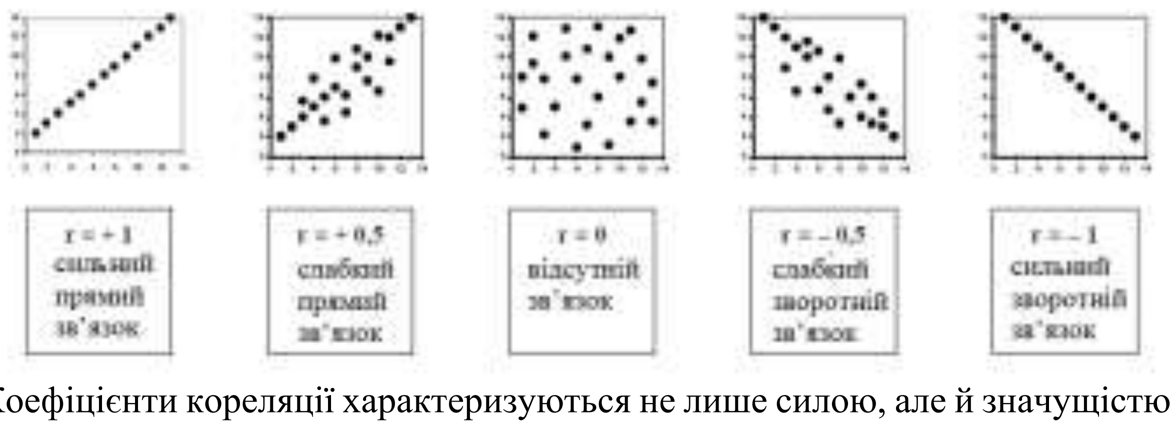
Кореляційний аналіз для двох випадкових величин складається з:

- побудови кореляційного поля і складання кореляційної таблиці;
- обчислення вибіркових коефіцієнтів кореляції і кореляційних відносин;
- перевірки статистичної значущості зв'язку.

Основне призначення кореляційного аналізу - виявлення зв'язку між двома або більше досліджуваними змінними, яка розглядається як спільна узгоджена зміна двох досліджуваних характеристик. Кореляційний зв'язок не можна вважати свідченням причинно-наслідкового зв'язку. Коефіцієнт кореляції характеризується за формою, напрямком і силою.

За формою кореляційний зв'язок може бути лінійним або нелінійним. Більш зручним для виявлення та інтерпретації кореляційного зв'язку є лінійна форма. Для лінійного кореляційного зв'язку можна виділити два основних напрямки: позитивний («прямий зв'язок») і негативний («зворотний зв'язок»).

Лінійну кореляцію можна кількісно виміряти. Степінь зв'язку між ознаками виражається величиною, яка називається коефіцієнтом кореляції. Позначається r. Значення даного коефіцієнта можуть знаходитися в діапазоні від –1 до +1. Можливі варіанти зв'язку, відповідні їм коефіцієнти кореляції та їх інтерпретації зобразимо за допомогою діаграм розсіяння:



Коефіцієнти кореляції характеризуються не лише силою, але й значущістю. Сильна кореляція може виявитися випадковою при малому обсязі вибірки, а слабка кореляція може виявитися високо значущою при великому обсязі вибірки.

Сила зв'язку безпосередньо вказує, наскільки синхронно проявляється спільна мінливість досліджуваних змінних. Наочне уявлення про характер ї зв'язку дає діаграма розсіювання - графік, осі якого відповідають значенням двох змінних, а кожен досліджуваний є точкою.

В якості числової характеристики ймовірнісної зв'язку коефіцієнти кореляції використовують значення в діапазоні від -1 до +1. Після проведення розрахунків дослідник, як правило, відбирає тільки найбільш сильні кореляції, які в подальшому інтерпретуються:

- сильний або тісний зв'язок при $r \geq 0,7$;
- середній при $0,5 \leq r < 0,7$;
- помірний при $0,3 \leq r < 0,5$;
- слабкий при $0,2 \leq r < 0,3$;
- дуже слабкий при $r < 0;2$.

Критерієм для відбору «досить сильних» кореляцій може бути як абсолютне значення самого коефіцієнта кореляції так і відносна величина цього коефіцієнта, яка визначається за рівнем статистичної значущості (від 0,01 до 0,1), який залежить від розміру вибірки. У малих вибірках для подальшої інтерпретації коректніше відбирати сильні кореляції на підставі рівня статистичної значущості. Для досліджень, які проведені на великих вибірках, краще використовувати абсолютні значення коефіцієнтів кореляції.

Таким чином, завдання кореляційного аналізу зводиться до встановлення напрямку (позитивного чи негативного) і форми (лінійної, нелінійної) зв'язку між ознаками, вимірюванню її тісноти, і перевірки рівня статистичної значущості отриманих коефіцієнтів кореляції.

На даний час розроблено безліч різних коефіцієнтів кореляції. Найбільш вживаними є коефіцієнти кореляції r-Пірсона, r-Спірмена і r-Кендалла. Як правило, комп'ютерні статистичні програми в меню «Кореляції» пропонують саме ці три коефіцієнта як основні, а для вирішення інших дослідницьких завдань пропонуються методи порівняння груп.

Вибір методу обчислення коефіцієнта кореляції, в першу чергу, залежить від:

- типу шкали, в якій виміряні змінні (номінальна, рангова, інтервальна, абсолютна);
- виду нормальності розподілу даних.

Вивчення зв'язку між ознаками, які приймають випадкові значення, починається з оцінювання його лінійності.

У психолого-педагогічних дослідженнях використовують чотири різновиди моделей взаємозв’язків, які ґрунтуються на визначенні коефіцієнтів взаємної спряженості та кореляції і тісно пов’язані із вживаними вимірювальними шкалами:

1) коефіцієнти кореляції змінних, виміряних на рівні шкал інтервалів і відношень (коефіцієнт лінійної кореляції Пірсона);

2) коефіцієнти кореляції змінних, виміряних за порядковою шкалою:

- коефіцієнти взаємозв'язку двох змінних (коефіцієнт кореляції Спірмена, міра зв'язку Кендалла);
- коефіцієнт взаємозв'язку декількох змінних (коефіцієнт конкордації W);

3) коефіцієнти кореляції змінних, виміряних на рівні номінальних шкал:

- коефіцієнти чотириклітинної зв'язаності типу 2x2 (коефіцієнт асоціації А, коефіцієнт контингенції Юла Q);
- коефіцієнти багатоклітинної зв'язаності типу *mxn* (коефіцієнт взаємної зв'язаності Пірсона С і коефіцієнт взаємної зв'язаності Чупрова К);

4) коефіцієнти кореляції змінних, виміряних за різними типами шкал (бісеріальний коефіцієнт кореляції, точково-бісеріальний коефіцієнт кореляції і рангово- бісеріальний коефіцієнт кореляції).

Коефіцієнт кореляції як міра зв'язку між випадковими величинами – також випадкова величина, що має ймовірнісний характер і потребує статистичного оцінювання.

6.2. Кореляція метричних змінних

Для вивчення взаємозв'язку двох метричних змінних, виміряних на одній і тій же вибірці, застосовується *коефіцієнт кореляції r-Пірсона.* Сам коефіцієнт характеризує наявність тільки лінійного зв'язку між ознаками. Коефіцієнт лінійної кореляції є параметричним методом і його коректне застосування можливо тільки в тому випадку, якщо дані двох змінних відрізняється від нормального виду в незначній мірі.

При обробці даних «вручну» необхідно обчислити коефіцієнт кореляції, а потім визначити p-рівень значущості (з метою перевірки даних користуються таблицями критичних значень, які складені для цього критерію). Величина коефіцієнта лінійної кореляції Пірсона не може перевищувати +1 і бути менше ніж -1. Ці два числа +1 і -1 є межами для коефіцієнта кореляції. Коли при розрахунку виходить величина, більша або менша +1 -1, це свідчить, що сталася помилка в обчисленнях.

При обчисленнях на комп'ютері статистична програма (SPSS, Statistica) супроводжує обчислений коефіцієнт кореляції більш точним значенням p-рівня.

Для статистичного рішення про прийняття або відхилення H_0 зазвичай встановлюють $p = 0,05$, а для великого обсягу вибірки досліджуваних (100 і більше) $p = 0,01$. Якщо $p < 0,05$, H_0 відхиляється і робиться змістовний висновок, що виявлений статистично достовірний (значущий) зв'язок між досліджуваними змінними (позитивний чи негативний в залежності від знака кореляції). Коли $p > 0,05$, H_0 не відхиляється, змістовний висновок обмежений констатацією, що зв'язок (статистично достовірний) не виявлений.

Якщо зв'язок не виявлено, але є підстави вважати, що зв'язок насправді існує, слід перевірити можливі причини недостовірності зв'язку.

Нелінійність зв'язку- для встановлення цього необхідно проаналізувати графік двовимірного розсіювання. Якщо зв'язок нелінійний, але монотонний необхідно перейти до рангових кореляція. Якщо зв'язок не монотонний, то доречно поділити вибірку досліджуваних на частини, в яких зв'язок монотонний, і обчислити кореляції окремо для кожної частини вибірки, або поділити вибірку на контрастні групи і далі порівнювати їх за рівнем вираженості ознаки.

Наявність викидів і виражена асиметрія розподілу однієї або обох змінних. Для цього необхідно проаналізувати гістограми розподілу частот обох ознак. При наявності викидів або асиметрії виключити викиди або перейти до рангових кореляція.

Неоднорідність вибірки (проаналізувати графік двовимірного розсіювання). Спробувати розділити вибірку на частини, в яких зв'язок може мати різні напрямки.

Якщо ж зв'язок статистично достовірний, то перш ніж робити змістовний висновок, необхідно виключити можливість помилкової кореляції:

- *зв'язок обумовлений викидами.* При наявності викидів перейти до рангових кореляція або виключити викиди;
- *зв'язок обумовлений впливом третьої змінної.* Якщо є таке явище, необхідно обчислити кореляцію не тільки для всієї вибірки, а й для кожної групи окремо. Якщо «третя» змінна метрична - обчислити приватну кореляцію.

КОЕФІЦІЄНТ КОРЕЛЯЦІЇ ПІРСОНА

Для пошуку міри зв'язку між двома змінними розглянемо відхилення кожної змінної від середнього арифметичного: $X_i - X$ та $Y_i - Y$ Візьмемо для прикладу такі ряди даних .

Таблиця

№	Інтелект X	Успішність Y
1	90	5
2	120	6
3	120	8
4	135	9
5	135	9
6	120	10
7	100	7
8	90	6
9	80	5
10	140	6
11	190	11
$\sum$	1320	82
$\overline{X}$	120	7,5

Тепер подивимося, що буде відбуватися з добутками ($X_i - X$)($Y_i - Y$) при їх підрахунку для різних досліджуваних. Візьмемо таких досліджуваних, у яких високий рівень інтелекту (І) та успішності (У), низький рівень І та У, високий рівень І та низький рівень У.

1. Високий рівень І та У (досліджуваний11) – добуток буде додатнім.
2. Низький рівень І та У (досліджуваний 8) – добуток теж буде додатнім.

3. Високий рівень I та низький рівень У (досліджуваний 10) – добуток буде від'ємним.

Таким чином, **якщо** змінні X та Y в основному пов'язані прямо, то більшість добутків $(X_i - X)(Y_i - Y)$ будуть додатні, а значить і сума $\sum(X_i - X)(Y_i - Y)$ теж буде додатною.

Якщо ж більшість змінних пов'язані обернено, то більшість добутків $(X_i - X)(Y_i - Y)$ будуть від'ємними, а значить і сума $\sum(X_i - X)(Y_i - Y)$ теж буде від'ємною.

Таким чином, ми отримали достатньо ефективну міру зв'язку. Однак, вона має недолік, пов'язаний із впливом чисельності вибірки. Тому отриману суму ділять на **(n-1)**.

У результаті отримують величину, яка називається **коваріацією**, і позначається s_{xy}

$$s_{xy} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

Відмітьте, що коваріація змінної сама з собою є дисперсією.

Коваріація є достатньо задовільною мірою зв'язку для багатьох задач, однак, залежить від впливу стандартних відхилень обох груп.

У результаті ділення коваріації на стандартні відхилення цих груп отримується **коефіцієнт кореляції Пірсона**. Позначається він r_{xy}

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

З допомогою цієї формули коефіцієнт кореляції Пірсона означається, однак важко обчислюється. Тому для зручності обчислень з допомогою математичних перетворень виводять таку формулу:

$$r_{xy} = \frac{n \cdot \sum_i (x_i \cdot y_i) - \left(\sum_i x_i\right)\left(\sum_i y_i\right)}{\sqrt{\left(n \cdot \sum_i x_i^2 - \left(\sum_i x_i\right)^2\right) \cdot \left(n \cdot \sum_i y_i^2 - \left(\sum_i y_i\right)^2\right)}}$$

ОБЧИСЛЕННЯ КОЕФІЦІЄНТА КОРЕЛЯЦІЇ ПІРСОНА

Нехай нам потрібно визначити, чи існує зв'язок між рівнем успішності учінв та їх особистісною тривожністю. Після проведення дослідження ми отримуємо два ряди даних. Обчислимо коефіцієнт лінійної кореляції Пірсона для того, щоб встановити характер зв'язку між ними.

ПРОЦЕДУРА ОБЧИСЛЕННЯ

Передусім для спрощення обчислень дані варто згрупувати у таблицю.

Таблиця

Обчислення коефіцієнта кореляції Пірсона

№	Учні	Успішність учнів (x)	Особистісна тривожність (y)	x_i^2	y_i^2	$(x_i \cdot y_i)$
1.	А.О.	10	48	100	2304	480
2.	Л.А.	12	49	144	2401	588
3.	М.П.	10	45	100	2025	450
4.	С.А.	9	38	81	1444	342
5.	Р.К.	8	34	64	1156	272
6.	В.О.	7	25	49	625	175
7.	Д.К.	10	42	100	1764	420
8.	В.К.	9	41	81	1681	369
9.	М.А.	11	43	121	1849	473
10.	Д.К.	10	46	100	2116	460
11.	З.Ц.	12	48	144	2304	576
12.	В.Ж.	8	23	64	529	184
		$\sum_i x_i = 116$	$\sum_i y_i = 482$	$\sum_i x_i^2 = 1148$	$\sum_i y_i^2 = 20198$	$\sum_i (x_i \cdot y_i) = 4789$

Після проведення всіх обчислень значення з нижніх комірок таблиці підставляються у формулу коефіцієнта лінійної кореляції Пірсона, і отримується значення r_{xy} =0,87

6.3. Кореляція рангових змінних

Якщо до кількісних даних неприйнятний коефіцієнт кореляції r- Пірсона, то для перевірки гіпотези про зв'язок двох змінних після попереднього рангування можуть бути застосовані коефіцієнти кореляції r- Спірмена або τ-Кендалла.

Для коректного обчислення обох коефіцієнтів (Спірмена і Кендалла) результати вимірювань повинні бути представлені в шкалі рангів або інтервалів. Принципових відмінностей між цими критеріями не існує, але прийнято вважати, що коефіцієнт Кендалла є більш «змістовним», так як він більш повно і детально аналізує зв'язки між змінними, перебираючи всі можливі відповідності між парами значень. Коефіцієнт Спірмена більш точно враховує саме кількісну ступінь зв'язку між змінними.

Коефіцієнт рангової кореляції Спірмена є непараметрическим аналогом класичного коефіцієнта кореляції Пірсона, але при його розрахунку враховуються не середнє арифметичне і дисперсія змінних, а ранги.

Застосування коефіцієнта кореляції Спірмена для перевірки гіпотез подібне до використання коефіцієнта r-Пірсона. У комп'ютерних програмах (SPSS, Statistica) рівні значущості для однакових коефіцієнтів r-Пірсона і r- Спірмена на завжди збігаються.

Перевага коефіцієнта r-Спірмена в порівнянні з коефіцієнтом r-Пірсона - більшій чутливості до зв'язку. Ми використовуємо його в наступних випадках:

- наявність істотного відхилення розподілу хоча б однієї змінної від нормального вигляду (асиметрія, викиди);
- поява криволінійного (монотонного) зв'язку.

Обмеженням для застосування коефіцієнта r-Спірмена є:

- для кожної змінної не менше 5 спостережень;
- коефіцієнт при великій кількості однакових рангів по змінним дає огрублене значення.

КОЕФІЦІЄНТ РАНГОВОЇ КОРЕЛЯЦІЇ СПІРМЕНА

У випадку, коли обидві змінні виміряні в порядкових шкалах, використовують ранговий коефіцієнт кореляції Спірмена, який легко можна вивести із коефіцієнта кореляції Пірсона, зробивши певні припущення стосовно рангових величин (Х та У приймають значення, рівні 1, 2, 3, ...).

Уявімо, що ми провели дослідження, в якому експерт з допомогою рангування визначав, наскільки студенту подобається його група (Х), а також його куратор (У).

Обчислення коефіцієнта рангової кореляції Спірмена

№	X	У	$(X_i - Y_i)$	$(X_i - Y_i)^2$
1	1	5	-4	16
2	8	1	7	49
3	3	2	1	1
4	2	3	-1	1
5	4	4	0	0
6	5	6	-1	1
7	6	9	-3	9
8	7	10	-3	9
9	10	8	2	4
10	9	7	2	4
$\sum$	55	55		94

Формула Спірмена ґрунтується на припущенні, що $\sum X_i = \sum Y_i$, і має такий вигляд:

$$r_s = 1 - \frac{6\sum (X_i - Y_i)^2}{n(n^2 - 1)}$$

В нашому випадку r_s =0,43

Однак, при обчисленні коефіцієнта кореляції Спірмена може виникнути проблема зв’язаних рангів.

Уявімо, що наш експерт не зміг прорангувати 1, 3 та 4 досліджуваних. В такому випадку їм слід приписати ранги за таким правилом:

Уявно все ж таки проставити цим досліджуваним ранги (1, 2, 3), потім знайти їх середнє арифметичне і реально проставити їм знайдений середній ранг (2). Подальше рангування слід продовжувати вже не з 2, а з того рангу, на якому закінчилося уявне рангування (після 3). Такі ранги називають зв’язаними.

Однак, якщо ми тепер порахуємо суми рангів в обох змінних, то побачимо, що базове для формули Спірмена припущення $\sum X_i = \sum Y_i$ не справджується. У такому випадку див. табл.:

Таблиця

Кількість зв'язаних рангів	Зв'язаних рангів мало	Зв'язаних рангів багато
Проблема	Відмінності між сумами незначні	Відмінності між сумами значні
Вирішення	Ігнорувати відмінності, пам'ятаючи, що формула може дати похибку	Застосувати формулу Пірсона або τ-Кендалла (див. далі) із поправкою на зв'язані рангів

Коефіцієнт рангової кореляції τ-Кендалла є самостійним оригінальним методом, що спирається на обчислення співвідношення пар значень двох вибірок, що мають однакові або різні тенденції (зростання або спадання значень). Цей коефіцієнт називають ще коефіцієнтом конкордації.

Таким чином, основною ідеєю даного методу є те, що про направлення зв'язку можна судити, попарно порівнюючи між собою досліджуваних: якщо у пари досліджуваних зміна X збігається за напрямком зі зміною Y, це свідчить про позитивний зв'язок, якщо не збігається - про негативний зв'язок. У цьому методі одна змінна представляється у вигляді монотонної послідовності в порядку зростання величин, іншій змінній присвоюються відповідні рангові місця. Кількість інверсій (порушень монотонності в порівнянні з першим рядом) використовується у формулі для кореляційних коефіцієнтів. При підрахунку τ-Кендалла «вручну» дані спочатку впорядковуються за змінною X. Потім для кожного досліджуваного підраховується, скільки разів його ранг по Y виявляється меншим, ніж ранг досліджуваних, що знаходяться нижче. Результат записується в стовпець «Збіги». Сума всіх значень стовпчика «Збіг» і є P - загальне число збігів, підставляється в формулу для обчислення коефіцієнта Кендалла, який є більш простим в обчислювальному відношенні, але при зростанні вибірки, на відміну від r-Спірмена, обсяг обчислень зростає не пропорційно, а в геометричній прогресії. При обчисленнях на комп'ютері в статистичної програмі (SPSS, Statistica) коефіцієнт Кендалла обраховується аналогічно коефіцієнтам r-Спірмена і r- Пірсона. Обчислений коефіцієнт кореляції τ-Кендалла характеризується більш точним значенням p-рівня.

Застосування коефіцієнта Кендалла є кращим, якщо у вихідних даних є викиди.

КОЕФІЦІЄНТ τ-КЕНДАЛЛА

Коефіцієнт τ-Кендалла є альтернативним коефіцієнтом для обчислення зв'язку в тих же випадках, що і коефіцієнт рангової кореляції Спірмена, але ґрунтується на дещо інших припущеннях.

Таблиця

№	Х (подобається група)	У (подобається куратор)	Співпадання (Р)	Інверсії (Q)
1	1	5	5	4
4	2	3	6	2
3	3	2	6	1
5	4	4	5	1
6	5	6	4	1
7	6	9	1	3
8	7	10	0	3
2	8	1	2	0
10	9	7	1	0
9	10	8	0	0
		$\sum$	30	15

Для підрахунку коефіцієнта τ-Кендалла використовують такий алгоритм :

1. Впорядковують ранги по одній із змінних.

2. Підраховується кількість "співпадань" (Р): для кожного об'єкта підраховується, скільки разів його ранг по У виявляється менше від рангів об'єктів, які знаходяться нижче нього. 3. Підраховується кількість "інверсій" (Q): для кожного об'єкта підраховується, скільки разів його ранг по У виявляється більше, ніж ранги об'єктів, які знаходяться нижче.

4. Обчислюється коефіцієнт τза формулою:

$$\tau = \frac{P - Q}{n(n-1)/2},$$

де:

P - кількість співпадань,

Q - кількість інверсій,

n - загальна кількість об'єктів.

Цю формулу можна спростити, якщо врахувати, що $n(n-1)$ 2 – це загальна кількість можливих пар n об'єктів, і $n(n-1)$ 2= $P + Q$. Тоді можна отримати дві абсолютно еквівалентні формули:

$$\tau = \frac{4P}{n(n-1)} - 1 \text{ або } \tau = 1 - \frac{4Q}{n(n-1)}$$

В нашому випадку ми отримаємо τ=0,33.

Як можна помітити, τ-Кендалла та r-Спірмена для одних і тих же даних дають **різні числові результати**. Причина – в різних логіках побудови коефіцієнтів. Якщо r-Спірмена можна проінтерпретувати як коефіцієнт Пірсона для рангових величини, то τ-Кендалла відображає різницю ймовірностей між тим, що дані в обох рядах даних мають однаковий порядок і тим, що дані в обох рядах даних мають різний порядок.

6.4. Кореляція дихотомічних змінних

При порівнянні двох змінних, виміряних в дихотомічній шкалі, мірою кореляційної зв'язку служить так званий *коефіцієнт φ,* який представляє собою коефіцієнт кореляції для дихотомічних даних.

Величина коефіцієнта φ лежить в інтервалі між +1 та -1. Він може бути як позитивним, так і негативним, характеризуючи напрямок зв'язку двох дихотомічних ознак. Проте інтерпретація ф може породжувати специфічні проблеми. Неправильно вважати, що інтерпретуються значення r = 0,60 і φ = 0,60 однакові. Коефіцієнт φ можна обчислити методом кодування, а також використовуючи так звані четирьохклітинні таблиці або таблицю спряженості.

Нехай маємо ряди даних: Х – сімейний стан (0-неодружений, 1 – одружений), а У – навчання у вузі (0 – продовжує навчання, 1 - виключений).

№	1	2	3	4	5	6	7	8	9	10	11	12
Х	0	1	0	0	1	1	0	1	0	0	0	1
У	0	1	1	0	1	0	0	1	0	1	0	1

Очевидно, що до таких даних прямо застосувати коефіцієнт кореляції Пірсона неможливо. Однак, в результаті певних математичних міркувань з нього можна отримати формулу для нашого випадку. Вона матиме такий вигляд:

$$\varphi = \frac{p_{xy} - p_x p_y}{\sqrt{p_x q_x p_y q_y}},$$

де:

p_{xy} – відсоток людей, що мають одиницю по Х та по У одночасно, p_x – відсоток людей, що мають одиницю по Х, p_y – відсоток людей, що мають одиницю по У, q_x – відсоток людей, що мають нуль по Х, q_y – відсоток людей, що мають нуль по У. При цьому $q_x = 1 - p_x$, а $q_y = 1 - p_y$ Для отримання цієї формули достатньо формулу коефіцієнта кореляції Пірсона поділити на n та з'ясувати, що середнє арифметичне для дихотомічних шкал є виразом відсотку кількості одиниць.

Якщо нас не цікавлять відсотки одиниць та нулів, ми користуємося **таблицею спряженості ознак**, яка має в загальному такий вигляд

		Змінна Х		Всього
		0	1	
Змінна У	1	a	b	**a+b**
	0	c	d	**c+d**
	Всього	**a+c**	**b+d**	n

Тоді обчислення проводять за такою формулою:

$$\varphi = \frac{bc - ad}{\sqrt{(a+c)(b+d)(a+b)(c+d)}}$$

Для нашого випадку таблиця спряженості ознак матиме такий вигляд

		Змінна Х		Всього
		0	1	
Змінна У	1	2	4	**6**
	0	5	1	**6**
	Всього	**7**	**5**	12

Отже, $\varphi = 0{,}507$

Для застосування коефіцієнта кореляції φ необхідно дотримуватися таких умов:

- порівнювані ознаки повинні бути виміряні в дихотомічній шкалі;
- число ознак в порівнюваних змінних X і Y має бути однаковим.

Даний вид кореляції розраховують в комп'ютерній програмі SPSS на підставі визначення подібності/відмінності. Деякі статистичні процедури, такі як факторний аналіз, кластерний аналіз, багатовимірне шкалювання, побудовані на застосуванні цих принципів, а іноді самі представляють додаткові можливості для обчислення цих характеристик.

У тих випадках коли одна змінна вимірюється в дихотомічній шкалі (змінна X), а інша в шкалі інтервалів або відносин (змінна Y), використовується *бісеріальний коефіцієнт кореляції,* наприклад, при перевірці гіпотез про вплив статі дитини на показник зросту і ваги. Цей коефіцієнт змінюється в діапазоні від -1 до +1, але його знак для інтерпретації результатів не має значення. Для його застосування необхідно дотримуватися таких умов:

- порівнювані ознаки повинні бути виміряні в різних шкалах: одна X - в дихотомічній шкалі; інша Y - в шкалі інтервалів або відносин;
- змінна Y повинна мати нормальний закон розподілу;
- число варіюючих ознак в порівнюваних змінних X і Y має бути однаковим.

Якщо ж змінна X виміряна в дихотомічній шкалі, а змінна Y в ранговій, необхідно використовувати *рангово-бісеріальний коефіцієнт кореляції,* який тісно пов'язаний з τ-Кендалла. Інтерпретація результатів така ж сама.

Завдання на самостійну підготовку

1. Характеристика кореляційного аналізу даних
2. Алгоритми розрахунку на матеріалі психологічних даних:

- лінійної кореляції Пірсона;
- рангової кореляції Спірмена;
- «τ» Кендалла;
- бісереального коефіцієнту кореляції;
- рангово-бісереального коефіцієнту кореляції;
- χ^2 - Пірсона;
- φ-коефіцієнта погодженості;
- коефіцієнта множинної кореляції;
- коефіцієнта приватної кореляції.