УДК 519.233.5

ПОРІВНЯЛЬНИЙ АНАЛІЗ КОЕФІЦІЄНТІВ КОРЕЛЯЦІЇ ПІРСОНА ТА СПІРМЕНА ДЛЯ РІЗНИХ ТИПІВ ДАНИХ

Анастасія Піскун

Миколаївський національний аграрний університет, м. Миколаїв, України

Науковий керівник: О. Є. Богатєнкова,
асистент Миколаївський національний аграрний університет

Кореляційний аналіз застосовують для виявлення взаємозв'язків між ознаками під час опрацювання експериментальних, освітніх та соціально-економічних даних. Проте однакові спостереження можуть давати різні оцінки залежно від обраного коефіцієнта. Це пояснюється тим, що коефіцієнт Пірсона характеризує лінійний зв'язок, а коефіцієнт Спірмена – монотонний зв'язок між рангами. Тому вибір показника має спиратися на шкалу вимірювання, форму залежності та наявність нетипових спостережень.

Мета роботи –порівняти властивості коефіцієнтів кореляції Пірсона та Спірмена для кількісних і порядкових даних та визначити умови їх обґрунтованого застосування. Для цього використано порівняння визначень і припущень методів та обчислювальний експеримент на малих модельних вибірках. Такий підхід дає змогу відокремити вплив форми залежності від впливу окремого викиду.

Коефіцієнт Пірсона визначають як відношення коваріації двох змінних до добутку їхніх стандартних відхилень. Він враховує числові відстані між значеннями та змінюється від −1 до 1. Додатний знак відповідає узгодженому зростанню змінних, від'ємний – їх зміні у протилежних напрямках. Абсолютне значення, близьке до одиниці, свідчить про сильний лінійний зв'язок. Для обчислення показника обидві змінні мають бути неконстантними; нормальність розподілу не є умовою самого розрахунку [1].

Водночас описова оцінка та статистична перевірка гіпотези мають різні передумови. Стандартний параметричний тест для коефіцієнта Пірсона спирається на модель незалежних пар спостережень із двовимірного нормального розподілу за нульової гіпотези. Відхилення від цих умов потребує обережності під час інтерпретації p-значень і довірчих інтервалів. Автоматична відмова від Пірсона лише через ненормальність однієї ознаки не враховує ані мети аналізу, ані форми зв'язку [1].

Коефіцієнт Спірмена обчислюють як кореляцію Пірсона між рангами спостережень. Однаковим значенням надають середні ранги. Метод придатний для порядкових ознак, для яких можна встановити порядок, але відстані між категоріями не обов'язково рівні. Для кількісних даних він відображає узгодженість зростання або спадання змінних без вимоги лінійної форми залежності. Натомість довільне кодування номінальних категорій числами не створює змістовного порядку і не обґрунтовує застосування рангової кореляції [2].

Рангування зменшує вплив величини екстремального значення, оскільки після перетворення зберігається його місце у впорядкованій вибірці. Однак Спірмен не усуває вплив усіх нетипових спостережень: зміна порядку пар також змінює результат. Крім того, велика кількість однакових рангів зменшує деталізацію даних. Отже, ранговий підхід забезпечує інший спосіб вимірювання зв'язку, а не універсальне виправлення недоліків вихідної вибірки [2; 3].

Для обчислювального порівняння сформовано чотири набори по п'ять пар. У першому випадку значення першої ознаки становлять 1, 2, 3, 4, 5, а другої –2, 4, 6, 8, 10. У другому перша ознака незмінна, а друга набуває значень 1, 4, 9, 16, 25. У третьому використано перший набір із заміною останнього значення другої ознаки на 100. У четвертому перша ознака становить −2, −1, 0, 1, 2, а друга –4, 1, 0, 1, 4. Коефіцієнти обчислено функціями pearsonr і spearmanr бібліотеки SciPy; результати округлено до трьох десяткових знаків [1; 2].

Таблиця 1 - Результати обчислення на модельних вибірках

Тип залежності	Пірсон, r	Спірмен, ρ
Лінійна	1,000	1,000
Монотонна нелінійна	0,981	1,000
Лінійна із заміною одного значення	0,743	1,000
Немонотонна симетрична	0,000	0,000

Як показано в таблиці 1, за точної лінійної залежності обидва показники дорівнюють одиниці. За квадратичного зростання коефіцієнт Пірсона становить 0,981, тоді як Спірмена –1,000. Відмінність зумовлена тим, що порядок значень зберігається повністю, хоча числові прирости неоднакові. Це ілюструє різницю між лінійністю та монотонністю, яку важливо враховувати під час інтерпретації результатів.

Заміна одного значення на 100 знижує коефіцієнт Пірсона до 0,743, але залишає коефіцієнт Спірмена рівним 1,000, оскільки порядок спостережень не змінюється. Цей результат стосується конкретної конструкції прикладу: викид, який змінює ранги, може змінити й рангову кореляцію. У симетричному немонотонному випадку обидва коефіцієнти дорівнюють нулю попри точну функціональну залежність. Отже, нульове значення не доводить незалежності змінних.

Наведені модельні результати є ілюстративними та не становлять емпіричного висновку про генеральну сукупність. Вибірки з п’яти пар використано для прозорого відтворення обчислень, тому p-значення не інтерпретуються. Для малих реальних вибірок перевірку рангової кореляції доцільно виконувати перестановковим методом з урахуванням умов обмінності; асимптотичні оцінки значущості можуть бути неточними [2].

Практичний порядок аналізу передбачає перевірку шкал вимірювання, пропусків, повторів і незалежності спостережень, після чого будують діаграму розсіювання. Для кількісних ознак із приблизно лінійним зв’язком інформативним є коефіцієнт Пірсона. Для порядкових даних або монотонної нелінійної залежності обґрунтованим є коефіцієнт Спірмена. Якщо форма зв’язку складніша, слід досліджувати графік і модель залежності, а не обмежуватися одним числом.

Обчислення двох показників для тих самих пар спостережень може допомогти виявити чутливість висновку до ранжування та екстремальних значень. Проте різниця між ними сама по собі не доводить наявності викидів або нелінійності. Видаляти спостереження лише для збільшення кореляції неприпустимо: спочатку потрібно з’ясувати їх походження та, за необхідності, подати аналіз чутливості. Статистичну значущість також слід відрізняти від практичної сили зв’язку [3].

Окремої уваги потребує трактування різних типів даних. Для неперервних кількісних ознак, наприклад тривалості виконання завдання або концентрації речовини, зміст мають і порядок, і відстані між вимірюваннями. У такому разі лінійна кореляція дозволяє оцінити узгодженість відхилень від середніх значень. Для порядкових відповідей, наприклад рівнів задоволеності від низького до високого, рівність відстаней між сусідніми категоріями не гарантована. Тому ранжування краще узгоджується зі способом вимірювання. Вибір коефіцієнта має враховувати зміст ознаки ще до програмного обчислення, інакше математично коректний результат може не мати предметної інтерпретації.

Дискретний характер ознаки сам по собі не визначає вибору методу. Кількість виконаних завдань є числовою величиною з інтерпретованими різницями, тоді як номер категорії може бути лише умовним позначенням. В обох випадках у таблиці даних зберігаються цілі числа, однак їх вимірювальний зміст відрізняється. Для числових балів слід з’ясувати, чи припустимо трактувати різницю між оцінками як однакову на всій шкалі. Якщо таке припущення сумнівне, ранговий аналіз може бути доцільнішим. Водночас велика кількість однакових балів потребує коректного обчислення середніх рангів і врахування повторів під час перевірки значущості [2].

Підготовка вибірки також впливає на зіставність результатів. Обидва коефіцієнти потрібно розраховувати на однаковому наборі повних пар. Якщо для різних показників застосовано різні правила вилучення пропусків, розходження оцінок може бути спричинене зміною складу вибірки. Перед аналізом слід перевірити, чи не позначено пропущені значення нулями або службовими

числами, а також чи не переплутано одиниці вимірювання. Видалення неповних рядків необхідно описати разом із кількістю залишених спостережень. Так забезпечується можливість відтворити обчислення та відокремити властивості статистичного методу від наслідків очищення даних.

Зміна одиниць вимірювання не обов'язково змінює кореляцію. Додавання сталої до всіх значень або множення на додатний сталий коефіцієнт зберігає коефіцієнт Пірсона. Наприклад, переведення тривалості із хвилин у секунди не змінює оцінки лінійного зв'язку. Коефіцієнт Спірмена зберігається за строго зростаючого перетворення, оскільки порядок значень залишається тим самим. Це пояснює його поведінку в модельному прикладі з квадратами додатних чисел. Однак застосування нелінійного перетворення може змінити лінійну кореляцію. Отже, попередні перетворення ознак потрібно зазначати в описі методики та інтерпретувати разом із результатом.

Для наочного зіставлення методів доцільно побудувати дві діаграми: вихідних значень та відповідних рангів. Перша дає змогу оцінити форму залежності, розкид і можливі екстремальні точки. Друга показує узгодженість порядку спостережень та допомагає пояснити результат рангового аналізу. Якщо на першій діаграмі простежується вигнута, але послідовно зростаюча траєкторія, високе значення Спірмена є очікуваним. Якщо ж залежність змінює напрямок, наприклад спочатку спадає, а потім зростає, низькі значення обох коефіцієнтів можуть поєднуватися з виразною структурою. Тому графічна перевірка має передувати остаточному висновку [3].

Під час аналізу кількох груп слід додатково перевіряти, чи не приховує загальний коефіцієнт відмінності між ними. Наприклад, об'єднання спостережень із різними середніми рівнями ознак може створити загальну тенденцію, яка не відтворюється всередині окремих груп. У такій ситуації доцільно порівняти спільну діаграму з діаграмами для кожної змістовно визначеної підгрупи. При цьому поділ має спиратися на предметні підстави, а не на бажання отримати більший коефіцієнт. Така перевірка допомагає точніше сформулювати, до якої сукупності належить висновок і за яких умов його можна застосовувати.

Відтворюваність обчислювального експерименту забезпечується явним поданням вихідних пар і правил обробки. Для практичного аналізу необхідно зберігати вихідний файл, програмний код, версії використаних бібліотек та результати очищення. У підсумковій таблиці варто наводити назви ознак, кількість пар, обидва коефіцієнти й пояснення форми зв'язку. Значення доцільно округлювати лише під час представлення, залишаючи повну точність для проміжних розрахунків. Порівняння результатів, які відрізняються лише в останньому десятковому знаку, без урахування вибіркової невизначеності не є достатньою підставою стверджувати перевагу одного методу.

Для подальшого дослідження доцільно розширити модельний експеримент: змінювати кількість спостережень, інтенсивність випадкового шуму та частку нетипових точок. Важливо окремо розглядати викиди, що зберігають порядок значень, і спостереження, які його порушують. Для кожного сценарію повторні симуляції дозволять оцінити мінливість коефіцієнтів, а не лише одну отриману величину. Це є напрямом подальшої роботи; у наведеному експерименті такі серії не проводилися. Практична цінність поточного порівняння полягає у прозорій демонстрації властивостей двох оцінок та формуванні послідовного правила їх вибору для конкретної аналітичної задачі.

Таким чином, коефіцієнти Пірсона та Спірмена доповнюють один одного, оскільки описують різні властивості взаємозв'язку. Обчислювальний приклад показав збіг оцінок за лінійної залежності, їх розходження за монотонної нелінійності та різну реакцію на зміну величини крайнього спостереження. Обґрунтований вибір методу потребує поєднання числової оцінки з візуальним аналізом, урахуванням типу даних і перевіркою передумов статистичного висновування. Виявлена кореляція не є доказом причинно-наслідкового зв'язку.

Список використаних джерел

1. SciPy. pearsonr. SciPy API reference. URL: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html (дата звернення: 12.10.2026).

2. SciPy. spearmanr. SciPy v1.14.1 Manual. URL: https://docs.scipy.org/doc/scipy-1.14.1/reference/generated/scipy.stats.spearmanr.html (дата звернення: 12.10.2026).

3. Schober P., Boer C., Schwarte L. A. Correlation Coefficients: Appropriate Use and Interpretation. Anesthesia & Analgesia. 2018. Vol. 126, no. 5. P. 1763-1768. DOI: https://doi.org/10.1213/ANE.0000000000002864.

УДК 519.115

ПРО ОДНУ ЧИСЛОВУ ПОСЛІДОВНІСТЬ, ПОВ'ЯЗАНУ З ПЕРЕСТАНОВКАМИ, ТА ЇЇ АНАЛІТИЧНУ ФОРМУЛУ

Микола Філіпчук, Ольга Філіпчук

Чернівецький національний університет імені Юрія Федьковича, м. Чернівці, Україна

Дослідження обмежених перестановок є класичним напрямом дискретної математики та теорії графів, що має важливе прикладне значення в сучасній криптографії при проєктуванні симетричних шифрів [1] і теорії кодування інформації [2]. Особливе місце в таких задачах посідає аналіз структурних заборон на сумісне розташування елементів. Одним із обмежень, що виникають при оцінці завадостійкості блочних кодів і нелінійних вузлів перемішування інформації, є виключення суміжних пар з простими числовими характеристиками.

Розглянемо числову послідовність $a(n)$, $n \geq 0$, члени якої визначаються наступним чином:

$a(n) =$ *загальна кількість* n*-перестановок з елементів* $1, 2, ..., n$*, у яких добуток будь-яких двох сусідніх елементів не є простим числом.*

Перш за все, нагадаємо, що число 1 не є простим.

При $n = 0$ існує одна порожня перестановка, яка тривіально задовольняє потрібну умову, тому $a(0) = 1$.

При $n = 1$ єдина існуюча перестановка (1) теж тривіально задовольняє потрібну умову, тому $a(1) = 1$.

При $n = 2$ обидві існуючі перестановки $(1,2)$ і $(2,1)$ не задовольняють потрібну умову (оскільки $1 \cdot 2 = 2 \cdot 1 = 2$, а 2 є простим числом), тому $a(2) = 0$.

При $n = 3$ жодна із шести існуючих перестановок $(1,2,3)$, $(1,3,2)$, $(2,1,3)$, $(2,3,1)$, $(3,1,2)$, $(3,2,1)$, теж не задовольняє потрібну умову, тому $a(3) = 0$.

Натомість уже при $n = 4$ із 24-х всеможливих перестановок лише перестановки $(1,4,2,3)$, $(1,4,3,2)$, $(2,3,4,1)$, $(3,2,4,1)$ задовольняють потрібну умову, тому $a(4) = 4$.

Нескладним скриптом на мовах програмування Python або C++ можна доволі швидко знайти перших 12 членів розглядуваної послідовності:

$a(0) = 1$,
$a(1) = 1$,
$a(2) = 0$,
$a(3) = 0$,
$a(4) = 4$,
$a(5) = 12$,
$a(6) = 144$,