УДК 519.23

Основные ранговые коэффициенты корреляции в прикладных задачах

Оборовская А.С.
Алтайский государственный университет, г. Барнаул
oborov.anna@yandex.ru

Аннотация

Одной из важнейших задач любого исследования, связанного с многомерным анализом данных, является задача установления наличия и оценки степени связи между исследуемыми переменными. При изучении монотонных связей для их выявления и оценки силы чаще всего используют ранговые коэффициенты Спирмена и Кенделла. В работе исследованы соотношения этих коэффициентов между собой, а также с классическим коэффициентом корреляции Пирсона. Высказано несколько новых гипотез относительно этих коэффициентов, которые подтверждены полным перебором всех вариантов при некоторых небольших объемах выборки n, для осуществления которого написана компьютерная программа на языке Python.

Ключевые слова: статистическая связь и ее сила, ранговая корреляция, соотношение коэффициентов корреляции, тау Кенделла.

Практически в любом статистическом исследовании ставится задача оценить степень связи двух признаков. Допустим, мы изучаем некоторое конечное множество объектов, где у каждого из них известны значения признаков X и Y. Качественная интерпретация понятия силы связи может быть визуализирована путем построения так называемых полей корреляции, в процессе которого каждый из объектов изображается точкой на плоскости, координаты которой равны значениям этих признаков. Если полученные точки лежат на графике некоторой функции $y = f(x)$, то связь между признаками предельно сильная. Связь следует признать тем менее сильной, чем в большей степени разбросаны выборочные точки вокруг наиболее удачно подобранного такого графика.

Поскольку подобные исследования всегда актуальны, методы анализа связей между показателями постоянно совершенствуются, см., например, [1] и библиографию там.

Стандартной характеристикой, по величине которой оценивают степень статистической связи, является коэффициент корреляции, точнее говоря, тот из них, который принято называть коэффициентом Пирсона. Он был, видимо, впервые предложен в [2].

Выборочный вариант этого коэффициента по связанным выборкам X, Y вычисляют по формуле

$$\rho^*(X,Y) = \frac{\overline{XY} - \overline{XY}}{S_X S_Y}.$$

Однако этот коэффициент не всегда адекватно отображает даже функциональные монотонные связи, “не видя” их. Например, рассмотрим данные таблицы 1.

Таблица 1

Пример слабой корреляции в монотонной связи

X	$Y = X^{256}$	X	$Y = X^{256}$	X	$Y = X^{256}$
1	1	5	2.93874E+89	9	1.3901E+122
2	3.40282E+38	6	4.012E+99	10	1E+128
3	1.17902E+61	7	1.4878E+108		
4	1.15792E+77	8	3.9402E+115		

В этом примере коэффициент Пирсона равен 0,522. Но, если заменить реальные значения признаков их порядковыми номерами по возрастанию (рангами), то получим коэффициент Пирсона, равный 1, поскольку ряд Y заменится на последовательные натуральные числа и совпадет с рядом X. Таким образом, ранговый вариант коэффициента сумел обнаружить имеющуюся здесь функциональную связь, а коэффициент Пирсона ее не увидел.

Отсюда ясно, что, после перехода к рангам, этот коэффициент, видимо, начинает замечать все монотонные связи. В случае же, когда наблюдаемая величина имеет нечисловой характер, замена ее категорий их рангами с точки зрения какого-то отношения порядка на множестве этих категорий, пожалуй, является единственным способом применить стандартные статистические методы к её изучению. Для изучения связей между рядами рангов можно ввести, кроме коэффициента Пирсона, и другие специальные коэффициенты, обычно также называемые коэффициентами корреляции (см., например, [3, 4]).

Основной задачей настоящей работы является подробное изучение двух таких коэффициентов, наиболее часто применяемых на практике, а также выяснение соотношений между величинами этих коэффициентов с анализом причин этих различий.

Перейдем к строгим определениям. Будем называть ранговой переменной показатель, измеренный в ранговой, т.е. порядковой шкале. Для перевода наблюдений в ранговую шкалу предполагается, что значения наблюдаемой величины можно упорядочить. Если значения были числовыми, то для них чаще всего используется естественный порядок – от меньших к большим. Если же имеют дело с нечисловыми значениями, то можно упорядочить их, например, по степени полезности или привлекательности для исследователя.

При этом возможны группы связных рангов – это группы наблюдений с одинаковыми ранговыми значениями. Они могут возникать, если в выборке есть повторяющиеся значения переменных. В таком случае, каждому из повторяющихся значений должен быть присвоен одинаковый ранг, и эти наблюдения будут объединены в группу связных рангов.

Как правило, перевод данных наблюдений в ранговую форму происходит следующим образом:

1. Расположим выборочные данные первичного ряда в порядке возрастания или предпочтения. Тем самым, получен вариационный ряд выборки.

2. Пронумеруем элементы вариационного ряда, начиная с 1.

3. Определим ранг каждого из значений, при этом:

   - если значение в вариационном ряду встречается единственный раз, то ранг равен порядковому номеру;
   - если значение в вариационном ряду встречается два или более раз, то ранг вычисляется как среднее из порядковых номеров, которые присвоены этому значению.

Лемма 1. *Если R_i, $i = 1, \ldots, n$ – ранги n объектов, присвоенные по описанному алгоритму, то, независимо от того, сколько имеется групп объектов с одинаковыми рангами,*

получаем:

$$\sum_{i=1}^{n} R_i = \frac{n(n+1)}{2}.$$

Лемма 2. *При преобразовании группы из m различных последовательных рангов в одинаковые сумма их квадратов уменьшится на*

$$\frac{m^3 - m}{12}.$$

1. Качественная интерпретация и связь коэффициентов

Понятие ранговой переменной первоначально использовалось в психологии. Для изучения зависимости таких переменных Чарльз Эдвард Спирмен в работе [5] предложил числовой коэффициент, получивший впоследствии его имя. В литературных источниках можно найти несколько различных формул для его вычисления.

1) Самая простая формула – при отсутствии в обеих выборках одинаковых рангов:

$$R = 1 - \frac{6\sum\limits_{i=1}^{n} d_i^2}{n^3 - n}, \tag{1}$$

где n – количество значений, участвовавших в ранжировании; $\sum\limits_{i=1}^{n} d_i^2$ – сумма квадратов разностей между рангами одинаковых значений в разных выборках.

2) При наличии одинаковых рангов коэффициент может вычисляться по одной из формул (2) или (3):

$$R = 1 - \frac{6\sum\limits_{i=1}^{n} d_i^2}{n^3 - n - \frac{T_X + T_Y}{2}}, \tag{2}$$

где T_X и T_Y – поправки на одинаковые ранги, которые считаются по формулам

$$T_X = \sum_j \left(t_{j,X}^3 - t_{j,X}\right), \quad T_Y = \sum_j \left(t_{j,Y}^3 - t_{j,Y}\right),$$

$t_{j,X}$ – число одинаковых рангов в j-й группе одинаковых рангов ряда X. Сумма вычисляется по всем группам одинаковых рангов выборки Xили Y соответственно.

В литературе встречается и другая формула:

$$R = 1 - \frac{6\sum\limits_{i=1}^{n} d_i^2 + \frac{T_X + T_Y}{2}}{n^3 - n}. \tag{3}$$

На самом деле обе эти формулы в контексте стандартных статистических методов должны рассматриваться лишь как приближенные. Точная же формула имеет вид (см [2, с. 77]):

$$R = 1 - \frac{\frac{1}{6}\left(n^3 - n\right) - \sum\limits_{i=1}^{n} d_i^2 - \frac{T_X}{12} - \frac{T_Y}{12}}{\sqrt{\left(\frac{1}{6}\left(n^3 - n\right) - \frac{T_X}{6}\right)\left(\frac{1}{6}\left(n^3 - n\right) - \frac{T_Y}{6}\right)}}. \tag{4}$$

Высказанное только что утверждение следует подкрепить следующей точной формулировкой, которая означает, что фактически имеем дело все с тем же коэффициентом корреляции Пирсона.

Теорема 1. *При отсутствии повторяющихся рангов коэффициент Пирсона между ранговыми переменными вычисляется по формуле* (1).

Иногда для оценки степени связи между собой двух признаков, заданных в ранговых шкалах, применяют коэффициент тау Кенделла. Этот коэффициент предложен английским статистиком Морисом Джорджем Кенделлом в работе [6]. Приведем формулы для его вычисления.

Если все элементы каждой из выборок различны. Тогда

$$\tau(X,Y) = \frac{2(P-Q)}{n(n-1)} = \frac{4P}{n(n-1)} - 1.$$

где P, Q – суммарные количества положительных (соответственно, отрицательных) соответствий рангов в выборке Y по отношению к каждому конкретному элементу выборки X. Точнее, для каждого значения x_j, начиная с первого, вычисляют количества P_j, элементов выборки Y, имеющих номера $s > j$ и ранги $y_s > y_j$ и Q_j тех, для которых имеет место противоположное неравенство рангов.

В случае, если среди рангов есть повторения, полагают

$$\tau(X,Y) = \frac{P-Q}{\sqrt{(P+Q-K_X)\,(P+Q-K_Y)}}, \tag{5}$$

где K_X, K_Y – поправки, подобные T_X, T_Y в коэффициенте Спирмена. Точнее,

$$K_X = \frac{1}{2}\sum_j \left(t_{j,X}^2 - t_{j,X}\right)$$

и аналогичная формула имеет место для показателя Y.

Различие коэффициентов Спирмена и Кенделла состоит в том, что если первый учитывает только сам факт наличия “неправильного”, противоречащего предположению о монотонной связи показателей ранга среди значений второй выборки, то второй учитывает ещё и величину отклонения положения этого ранга от его “правильного” положения. Здесь возникает вопрос о том, как соотносить данные коэффициенты корреляции.

В книге [7, с. 76] для ситуации, когда все ранги в обоих рядах различны получены следующие две формулы, используя которые, можно попытаться корректно сравнить два изучаемых коэффициента. Пусть элементы первой выборки упорядочены по возрастанию.

Для коэффициента Спирмена

$$R = 1 - \frac{12}{n^3-n}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} H_{i,j}(j-i).$$

Для коэффициента Кенделла

$$\tau = 1 - \frac{4}{n^2-n}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} H_{i,j},$$

где $H_{i,j}$ – индикатор наличия инверсии между позициями i и j во второй выборке. $H_{i,j} = 1$, если $y_i > y_j$ и 0, если наоборот.

Гипотеза 1. *Чем большее количество инверсий происходит “на довольно большое расстояние” (т.е. разность позиций инверсий $(j-i)$ часто больше 1), тем сильнее коэффициент Спирмена отличается от коэффициента Кенделла.*

Вычитая одну из этих формул из другой, получаем

$$\tau - R = \frac{12}{n(n^2-1}\sum_{j>i} H_{i,j}\left(\frac{n+1}{3} - (j-i)\right).$$

Видно, что вычисляется сумма тех величин $t_{i,j} = \left(\frac{n+1}{3} - (j-i)\right)$, где $y_i > y_j$ (наличие инверсии). Результатом вычисления данной разности может быть как положительное, так и отрицательное число. При этом наибольшее по модулю значение этой суммы будет, в частности, получаться тогда, когда как можно большее количество из $t_{i,j}$ имеют один и тот же знак.

Отметим, что, если переставить элементы выборки X произвольным образом, одновременно переставляя соответствующие им по номерам элементы Y, то ни один из ранговых коэффициентов не поменяет величины. Поэтому всегда без ограничения общности можно считать, что $X = (1, 2, \ldots, n)$. Возьмем

$$X_1 = (1,2,3,4) \text{ и } Y_1 = (2,1,4,3) \text{ или } X_2 = (1,2,3,4) \text{ и } Y_2 = (3,4,1,2). \tag{6}$$

В каждой из двух таблиц ниже укажем все существующие в этих выборках инверсии (сначала номер первого элемента инверсии, под ним – номер второго), разности $(j-i)$ и знак $t_{i,j}$.

Таблица 2

Инверсии в выборках

1	2	3	4
2	1	4	3
Инверсии			
1	3		
2	4		
1	1		
Знаки	+	+		

1.666667
$(n+1)/2$

1	2	3	4
3	4	1	2
Инверсии			
1	1	2	2
3	4	3	4
2	3	1	2
Знаки	-	-	+	-

По последней строке каждой из таблиц видно, что для X_1, Y_1 будут суммироваться только положительные числа $t_{i,j}$, а для выборок X_2, Y_2 в сумму входят как положительные, так и отрицательные числа. Тем не менее, оказывается, что модули разности R и τ для этих выборок одинаковы.

Гипотеза 2. *Пусть объем выборки $n = 2k$ – четное число, а элементы выборки X – последовательные ранги от 1 до n. Тогда для построения ранговой выборки Y, не содержащей равных рангов и обладающей в наибольшей степени различающимися коэффициентами Кенделла и Спирмена существует всего два варианта: $Y = (k/2, k/2-1, \ldots, 1, k, k-1, \ldots, k/2+1)$ или $Y = (k/2+1, \ldots, k, 1, \ldots, k/2)$.*

Эти две выборки Y, о которых говорится в гипотезе 2, условимся называть антисинхронными выборками $(1, 2, \ldots, n)$.

Для антисинхронных выборок (6) получаем

$$\tau(X_1, Y_1) = \frac{1}{3},\ R(X_1, Y_1) = \frac{3}{5},\ \tau(X_2, Y_2) = -\frac{3}{5},\ R(X_2, Y_2) = \frac{1}{3}.$$

Полученная разница между коэффициентами как можно убедиться, перебрав все остальные варианты ранжирования четырех элементов выборки без повторяющихся рангов, оказывается максимальной. Она равна

Имея выборки ($n = 6$) $X = (1, 2, 3, 4, 5, 6)$, а $Y = (3, 2, 1, 6, 5, 4)$ или $X = (1, 2, 3, 4, 5, 6)$, а $Y = (4, 5, 6, 1, 2, 3)$ получаем:

$$\tau\,(X_1, Y_1) = 0.2,\ R\,(X_1, Y_1) = \frac{19}{35} \text{ и } \tau\,(X_2, Y_2) = -0.2,\ R\,(X_2, Y_2) = -\frac{19}{35}.$$

Разность между коэффициентами равна $\frac{12}{35}$. Таким образом, для выборок объема 4 и 6 гипотеза 2 подтверждена методом полного перебора.

Гипотеза 3. *Пусть объем выборки $n = 2k - 1$ – нечетное число, а элементы выборки X – последовательные ранги от 1 до n. Тогда для построения ранговой выборки Y, не содержащей равных рангов и обладающей в наибольшей степени различающимися коэффициентами Кенделла и Спирмена существует всего 4 варианта:*

$$Y = (\lfloor k/2 \rfloor, \lfloor k/2 \rfloor - 1, \ldots, 1, k, k-1, \ldots, \lceil k/2 \rceil,)$$

$$Y = (\lceil k/2 \rceil, \lfloor k/2 \rfloor, \ldots, 1, k, k-1, \ldots, \lceil k/2 \rceil + 1)\,,$$

$$Y = (\lceil k/2 \rceil, \ldots, k, 1, \ldots, \lfloor k/2 \rfloor)$$

$$Y = (\lceil k/2 \rceil + 1, \ldots, k, 1, \ldots, \lceil k/2 \rceil)\,.$$

Для $n = 5$ проверка того, что максимальная разница модулей коэффициентов достигается на следующих выборках Y была также проведена методом полного перебора (таблица 3).

Таблица 3

Антисинхронные выборки для $X = (1, 2, 3, 4, 5)$

Y	Коэф. Кенделла	Коэф. Спирмена
(2, 1, 5, 4, 3)	0.2	0.5
(3, 2, 1, 5, 4)	0.2	0.5
(3, 4, 5, 1, 2)	-0.2	-0.5
(4, 5, 1, 2, 3)	-0.2	-0.5

2. Выводы и перспективы

Подтверждения высказанных гипотез при конкретных n были получены путем полного перебора пар выборок для заданного n, где первая выборка всегда оставалась неизменной (числа расположены в порядке возрастания). Для этого была написана компьютерная программа на языке Python, которая перебирает все ранговые выборки заданного объема n и находит выборки с наибольшей разницей коэффициентов Спирмена и Кенделла. Кроме поиска выборок с максимальной разницей коэффициентов при заданном n программа позволяет пользователю ввести свои данные выборок и на основе этих данных выводит значения коэффициентов Спирмена, Кенделла и Пирсона.

Полученные в ходе работы результаты можно применять прежде всего в учебных курсах статистики для построения убедительных примеров, а также в некоторых практических задачах, определяя с помощью теоретически полученных результатов, какой из двух различных по значениям ранговых коэффициентов корреляции дает более адекватный результат.

Список литературы

1. Баврина А.П., Борисов И.Б. Современные правила применения корреляционного анализа // Медицинский альманах. — 2021. — № 3(68). — С. 70–79.

2. Pearson K. On the Theory of Contingency and its Relation to Association and Normal Correlation. — London : Dulau & Co., 1904. — 36 с.

3. Крамер Г. Математические методы статистики. — М. : Мир, 1975. — 648 с.

4. Гаек Я., Шидак З. Теория ранговых критериев. — М. : Наука, 1971. — 376 с.

5. Spearman C. The proof and measurement of association between two things // American Journal of Psychology. — 1904. — Vol. 15(1). — P. 72–101.

6. Kendall M. A New Measure of Rank Correlation // Biometrika. — 1938. — Vol. 30(1-2). — P. 81–89.

7. Дронов С.В. Методы и модели многомерной статистики. — Барнаул : Изд-во АлтГУ, 2015. — 275 с.