Коэффициент ранговой корреляции Кендалла (Kendall rank correlation coefficient)

> Синонимы: Kendall rank-order correlation coefficient, Kendall's τ coefficient, Тау-корреляция
> Разделы: Метрики

Коэффициент ранговой корреляции Кендалла является статистической мерой силы зависимости признаков, представленных в порядковой (ранговой) шкале. Он является альтернативой коэффициента ранговой корреляции Спирмена, которую предпочтительнее использовать в случае малых размеров выборок. Предложен известным английским статистиком Морисом Кендаллом в 1938 году.

Поскольку корреляция Кендалла является ранговой, то для оценки силы зависимости между признаками, используются не их значения, а соответствующие им ранги. Коэффициент инвариантен по отношению к любому монотонному преобразованию шкалы измерения (по возрастанию или убыванию).

Так же, как и другие меры ранговой корреляции, коэффициент Кендалла является непараметрической оценкой, т.е. не требует каких-либо предположений относительно распределения значений набора данных и его параметров. Это существенно упрощает его использование.

Коэффициент корреляции Кендалла использует пары наблюдений и определяет силу связи на основе шаблона согласованности (concordant) и несогласованности (disconcordant) между парами.

Пусть задан набор наблюдений, представленных парами значений $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, где X и Y — признаки, между которыми ищется зависимость, n — число наблюдений.

Введем в рассмотрение понятие конкордантности (согласованности) и дискоркондантности (несогласованности) наблюдений. В статистике согласованными называются пары наблюдений (x_1, y_1) и (x_2, y_2), если для них выполняется правило:

$$sgn(x_2 - x_1) = sgn(y_2 - y_1),$$

где

$$sgn(x) = \begin{cases} -1, x < 0 \\ 0, x = 0 \\ 1, x > 0 \end{cases}$$

Иными словами, в согласованной паре оба элемента одной пары больше, равны или меньше соответствующих элементов другой пары. Напротив, несогласованной называется пара наблюдений для которой выполняется:

$sgn(x_2 - x_1) = -sgn(y_2 - y_1)$,

что имеет место когда одна пара содержит более высокое значение x, то другая пара содержит более высокое значение y.

Количество несогласованных пар в наборе наблюдений называют **тау-расстоянием** Кендалла или **ранговым расстоянием Кендалла**. Оно представляет собой метрику, которая подсчитывает количество попарных расхождений между двумя ранжированными наборами значений признаков. Чем больше это расстояние, тем больше отличаются два признака, и, следовательно, тем меньше зависимость между ними.

Обозначим n_C (concordant) — число согласованных пар, n_D (disconcordant) — число несогласованных пар. Тогда коэффициент ранговой корреляции Кендалла вычисляется следующим образом:

$\tau = \frac{n_C - n_D}{n} = 1 - \frac{4n_D}{n(n-1)}$.

Поскольку знаменатель данного выражения представляет собой общее количество парных комбинаций, коэффициент τ изменяется в диапазоне $-1 \leq \tau \leq 1$. Значение $\tau = -1$ имеет место, если $n_C = 0$, т.е. все пары несогласованы. $\tau = 0$ если $n_C = n_D$, т.е. количество согласованных и несогласованных пар совпадают. И, наконец, $\tau = 1$ в случае, когда $n_D = 0$, т.е. все пары являются согласованными.

Интерпретация коэффициента ранговой корреляции Кендалла выглядит следующим образом:

- если соответствие между обоими рейтингами идеальное, коэффициент имеет значение 1 (т.е. большему рейтингу X) в паре всегда соответствует больший рейтинг Y;
- если соответствие между двумя рейтингами отсутствует, коэффициент имеет значение -1 (т.е. большему рейтингу X) в паре всегда соответствует меньший рейтинг Y;
- если коэффициент равен 0 (т.е. половина пар согласованы, а половина нет), то можно считать, что зависимость между признаками отсутствует.

Общее количество пар N, которые можно построить для набора данных размером n наблюдений, будет:

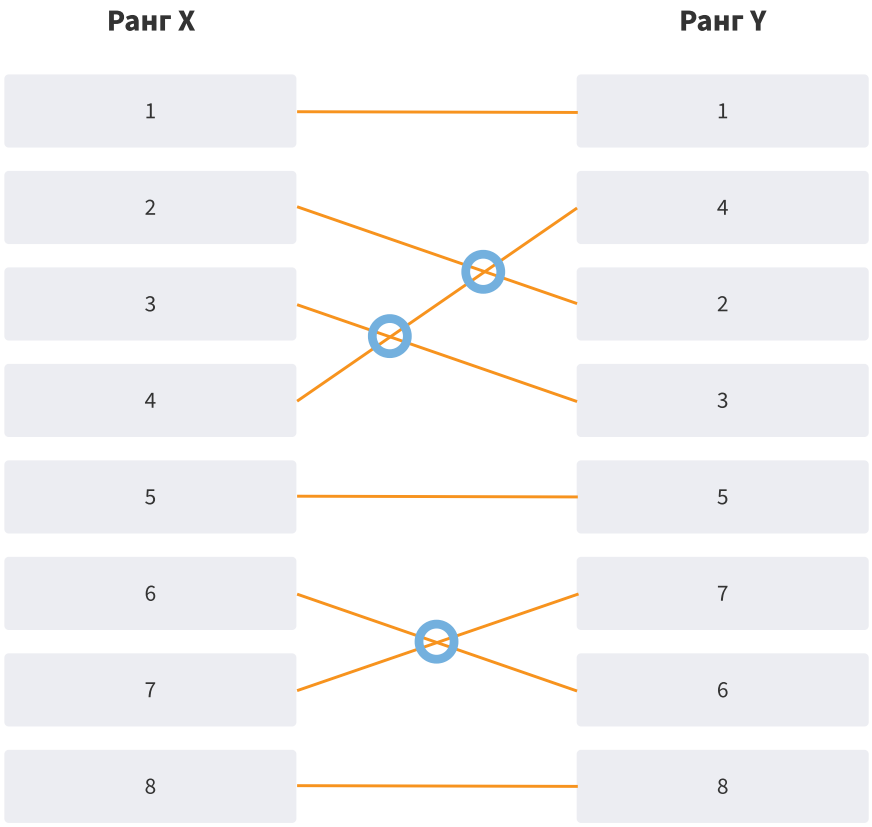
$N = 0.5 \cdot n \cdot (n-1)$.

Существует три типа коэффициентов ранговой корреляции Кендалла τ — $\tau - a$ (tau-a, тау-а), $\tau - b$ (tau-b, тау-би) и $\tau - c$ (tau-c, тау-си)

Коэффициент тау-а — вычисляется только на основании количества согласованных и несогласованных пар:

$\tau_A = \frac{n_C - n_D}{0.5 \cdot n(n-1)}$.

Данный метод в литературе часто называют **методом конкордации**. Для вычисления коэффициента также можно использовать **метод пересечений**. Если в списке рангов для двух признаков соединить ячейки с одинаковыми рангами с помощью линий (как показано на рисунке ниже), то некоторые линии будут пересекаться. Пересечение линий указывает на несогласованность пары. Интуитивно понятно, что число пересечений будет отражать силу зависимости между ранжированными признаками.



Действительно, если пересечений нет, то будет иметь место абсолютная положительная зависимость между признаками X и Y. Если все пары порождают пересечения, то ранги будут разнонаправленными, т.е. имеет место абсолютная отрицательная зависимость, и в этом случае количество пересечений будет равно $0.5 \cdot n \cdot (n-1)$, т.е. максимальному числу несогласованных пар.

Тогда формула для коэффициента ранговой корреляции Кендалла может быть записана в виде:

$$\tau = 1 - \frac{2 \cdot I}{0.5 \cdot n(n-1)},$$

где I — количество пересечений.

В примере на рисунке имеют место 3 пересечения для 8 наблюдений, тогда:

$$\tau = 1 - \frac{2 \cdot 3}{0.5 \cdot 8 \cdot (8-1)} \approx 0.786$$

Коэффициент тау-б — учитывает так называемые **связанные ранги** путем внесения соответствующей поправки. Связанными называются ранги, полученные путем усреднения одинаковых рангов. Количество наблюдений, по которому производится усреднение связанного ранга, называется **длиной связи**.

Формула для вычисления коэффициента тау-би выглядит следующим образом:

$\tau_B = \frac{2(n_C - n_D)}{\sqrt{(n_0 - n_X)(n_0 - n_Y)}}$,(1)

где

- $n_0 = 0.5n(n-1)$ — максимальное число пар;
- $n_X = \sum_i t_{iX}(t_{iX} - 1)$ — поправка на связанные ранги по X, где i — номер группы связей;
- $n_Y = \sum_j t_{jY}(t_{jY} - 1)$ — поправка на связанные ранги по Y, где j — номер группы связей;
- t_{iX} — число связанных значений в i-й группе связей для X (длина связи);
- t_{jY} — число связанных значений в j-й группе связей для Y;

Пары, построенные по наблюдениями со связанными рангами, не считаются ни согласованными, ни несогласованными, поэтому не учитываются при расчете (для них устанавливается 0).

Следует отметить, что иногда в формуле (1) не умножают числитель на 2, а вместо этого используют деление на 2 при вычислении n_X и n_Y.

Рассмотрим пример.

ID	Ранг X	Ранг Y
009	1.5	1
001	1.5	3.5
008	3	3.5
003	4.5	5
012	4.5	6
011	6	2
015	7	8
002	8	7

Для 8 наблюдений может быть построено $0.5 \cdot 8 \cdot (8-1) = 28$ пар. Пример содержит две группы связей по X для наблюдений 009 и 001, и наблюдений 003 и 012, а также одну группу связей по Y для наблюдений 001 и 008. Как можно увидеть, все связи имеют длину

2.

Для пар, в которых присутствует связанный ранг при расчете устанавливается 0. По X таких пар будет 4 — 009, 001, 003 и 012. По Y таких пар будет 2 — 001 и 008. Таким образом, в формуле (1) $t_{jY} = 2$, а $t_{iX} = 2$ — длины связей.

Теперь рассчитаем поправки:

$n_X = (2^2 - 2) + (2^2 - 2) = 4$

$n_Y = (2^2 - 2) = 2.$

Тогда коэффициент тау-би может быть рассчитан следующим образом:

$\tau_B = \frac{2 \cdot (n_C - n_D)}{\sqrt{(n_0 - n_X)(n_0 - n_Y)}} = \tau_B = \frac{2 \cdot (17 - 2)}{\sqrt{(8 \cdot (8-1) - 4)(8 \cdot (8-1) - 2)}} = \frac{30}{53} = 0.566.$

Коэффициент тау-си

$\tau_C = \frac{2(n_C - n_D)}{n^2 (m-1)/m}$. — учитывает число строк и столбцов в таблице. Здесь:

- $m = min(r, s)$;
- r, s — число столбцов и строк в таблице соответственно.

Чаще всего в бизнес-аналитике используется коэффициент тау-би, поэтому когда упоминают ранговый коэффициент корреляции Кендалла обычно имеют в виду именно эту его версию.

В Loginom существует специализированный обработчик Корреляционный анализ, в котором имеется возможность исследовать зависимости между порядковыми признаками с помощью вычисления коэффициента ранговой корреляции τ_B Кендалла.