Ранговая корреляция.

Ранговая корреляция – это метод корреляционного анализа, отражающий отношения переменных, упорядоченных по возрастанию их значения.

Объекты упорядочиваются по возрастанию субъективного признака, затем им присваиваются ранги, которые в дальнейшем сравниваются. Таким образом осуществляется ранговая корреляция.

Коэффициент ранговой корреляции **измеряет степень сходства** между двумя рейтингами, а также может использоваться для оценки **статистической значимости связи** между ними. Применяются статистические критерии значимости, в которых используется ранговая корреляция (t – критерий Стьюдента, U-критерий Манна-Уитни, критерий знакового ранга Уилкоксона).

Ранги - это порядковые номера единиц совокупности в ранжированном ряду. Если **проранжировать совокупность по двум признакам**, связь между которыми изучается, то полное совпадение рангов означает максимально тесную прямую связь, а полная противоположность рангов - максимально тесную обратную связь. Ранжировать оба признака необходимо в одном и том же порядке: либо от меньших значений признака к большим, либо наоборот.

Для практических целей использование ранговой корреляции весьма полезно. Например, если установлена высокая ранговая корреляция между двумя качественными признаками изделий, то достаточно контролировать изделия только по одному из признаков, что удешевляет и ускоряет контроль.

1. Коэффициент ранговой корреляции Спирмена

Коэффициент корреляции рангов, предложенный К. Спирменом, относится к непараметрическим показателям связи между переменными, измеренными в ранговой шкале. При расчете этого коэффициента не требуется никаких предположений о характере распределений признаков в генеральной совокупности. Этот коэффициент определяет степень тесноты связи порядковых признаков, которые в этом случае представляют собой ранги сравниваемых величин.

Величина коэффициента корреляции Спирмена лежит в интервале +1 и -1. Он может быть положительным и отрицательным, характеризуя направленность связи между двумя признаками, измеренными в ранговой шкале.

Ранговый коэффициент корреляции Спирмена ρ подсчитывается по формуле:

$$\rho = 1 - \frac{6\Sigma\, d_i^2}{n^3 - n}$$

d_i - разность между рангами по двум признакам,

n – число сопоставляемых пар.

Расчет коэффициента ранговой корреляции Спирмена.

Первый этап – ранжирование рядов переменных. Процедура ранжирования начинается с расположения переменных по возрастанию их значений. Разным значениям присваиваются ранги, обозначаемые натуральными числами. Если встречается несколько равных по значению переменных, им присваивается усредненный ранг.

Второй этап – непосредственный расчет коэффициента Спирмена: $\rho = 1 - \frac{6\Sigma d_i^2}{n^3 - n}$

Третий этап – статистическая проверка значимости коэффициента Спирмена.

Преимущество коэффициента корреляции рангов Спирмена состоит в том, что ранжировать можно и по таким признакам, которые *нельзя выразить численно*: можно проранжировать кандидатов на занятие определенной должности по профессиональному уровню, по умению руководить коллективом, по личному обаянию и т. п. При экспертных оценках можно ранжировать оценки разных экспертов и найти их корреляции друг с другом, чтобы затем исключить из рассмотрения оценки эксперта, слабо коррелированные с оценками других экспертов. Коэффициент корреляции рангов Спирмена применяется для оценки устойчивости тенденции динамики. Недостатком коэффициента корреляции рангов является то, что одинаковым разностям рангов могут соответствовать совершенно отличные разности значений признаков (в случае количественных признаков). Поэтому для последних следует считать корреляцию рангов приближенной мерой тесноты связи, обладающей меньшей информативностью, чем коэффициент корреляции числовых значений признаков, например, коэффициент Пирсона.

Величины, сравнение которых **не имеет количественного** выражения признака: произведения искусства, литературное творчество, красота, проявление интереса, вкус блюд, маркетинговые предпочтения в потреблении(одежда, гаджеты, еда, развлечений, поездок и т.п. Новинки кино, предпочтения – рейтинги фильмов, телепередач, жанровые литературные предпочтения различных групп населения, индивидуальные предпочтения при осуществлении покупок.

Недостаток: коэффициент корреляции Спирмена чувствителен к «выбросам».

ПРИМЕР 1 .

ЗАДАНИЕ. С помощью коэффициента ранговой корреляции установить зависимость между стажем практической работы и временем решения контрольной задачи у 10 программистов на основе следующих данных:

Номера испытуемых	1	2	3	4	5	6	7	8	9	10
Стаж (в мес).	32	15	16	18	20	28	21	29	23	17
Время решения (в мин.)	12	24	23	21	20	9	11	10	15	16

РЕШЕНИЕ.
Присвоим ранги x_i оценкам стажа. Располагаем в возрастающем порядке первые оценки, сохраняя связь между оценками:

x_i	15	16	17	18	20	21	23	28	29	32
y_i	24	23	16	21	20	11	15	9	10	12

Составим таблицу рангов для первого признака X_i - стаж (в месяцах)

X_i	15	16	17	18	20	21	23	28	29	32
Ранг X_i	1	2	3	4	5	6	7	8	9	10

Составим таблицу рангов для второго признака Y_i – время решения (в минутах)

Y_i	9	10	11	12	15	16	20	21	23	24
Ранг Y_i	1	2	3	4	5	6	7	8	9	10

Составляем таблицу соответствующих рангов величин X_i и Y_i

Ранг X_i	1	2	3	4	5	6	7	8	9	10
Ранг Y_i	10	9	6	8	7	3	5	1	2	4

Расчетная таблица

x_i	1	2	3	4	5	6	7	8	9	10	**Сумма**
y_i	10	9	6	8	7	3	5	1	2	4	
$d_i = x_i - y_i$	-9	-7	-3	-4	-2	3	2	7	7	6	
d_i^2	81	49	9	16	4	9	4	49	49	36	**306**

Находим коэффициент ранговой корреляции Спирмена:

$$\rho_B = 1 - \frac{6\sum d_i^2}{n^3 - n} = 1 - \frac{6 \cdot 306}{10^3 - 10} \approx -0,855.$$

Коэффициент достаточно большой по абсолютной величине, связь между стажем работы и временем решения задачи сильная (и обратная по направлению).

Проверим значимость коэффициента. Вычислим значение критерия:

$$t_s = \frac{\rho}{\sqrt{1-\rho^2}}\sqrt{n-2} = \frac{-0,855\sqrt{8}}{\sqrt{1-(-0,855)^2}} \approx -4,663 .$$

Так как $t_{кр} = t(10-2;0,05) = 2,306 < 4,663 = |t_s|$ 975) , гипотезу $H_0 : \rho = 0$ о незначимости коэффициента следует отвергнуть. Связь значима.

Ответ: -0,855, связь значима.

Замечание. Можно доказать, что если провести расчет выборочного коэффициента корреляции Пирсона, считая ранги в качестве выборочных значений случайных величин, то этот рассчитанный коэффициент совпадет с ранговым коэффициентом Спирмена.

2. Коэффициент ранговой корреляции Кендалла

Коэффициент Кендалла онлайн (semestr.ru)

Другой мерой связи между случайными величинами X и Y служит коэффициент ранговой корреляции Кендалла τ. Коэффициент ранговой корреляции Кендалла используется, когда имеются «выбросы», т.к. этот коэффициент к выбросам менее чувствителен, чем ранговый коэффициент Спирмена.

Коэффициент корреляции Кенделла (Kendall tau rank correlation coefficient) — мера линейной связи между случайными величинами. Корреляция Кенделла является ранговой, то есть для оценки силы связи используются не численные значения, а соответствующие им ранги. Коэффициент инвариантен по отношению к любому монотонному преобразованию шкалы измерения.

Для вычисления τ надо упорядочить ряд рангов переменной X , приведя его к ряду натуральных чисел. Затем рассматривают последовательность рангов переменной Y.

Пусть заданы две выборки: будем говорить, что пары наблюдений (x_i, y_i) и (x_j, y_j), $i<j$ согласованы, если порядок сортировки пар (x_i, x_j), (y_i, y_j) согласуется, т.е. выполняется: либо оба $x_i < x_j$ и $y_j < y_j$, либо оба $x_i > x_j$ и $y_j > y_j$.; в противном случае пары считаются несогласованными. Пусть P - число согласованных пар, Q - число несогласованных пар. Тогда, в предположении, что среди x_i и среди y_i нет совпадений, превышение согласованности над несогласованностью есть: P-Q .

Коэффициент Кендалла рассчитывается по формуле:

$$\tau = \frac{2S}{n(n-1)}, \quad \text{где } S = P - Q$$

P - число согласованных пар,

Q - число несогласованных пар.

Коэффициент τ принимает значения из отрезка [-1,1] . Равенство $\tau = 1$ указывает на строгую прямую линейную зависимость, $\tau = -1$ на обратную.

Для вычисления τ надо упорядочить ряд рангов переменной X , приведя его к ряду натуральных чисел. Затем рассматривают последовательность рангов переменной Y.

Пусть заданы две выборки:

Свойства коэффициента ранговой корреляции Кендалла

- $-1 \le \tau \le 1$.
- Если согласие между двумя рангами является идеальным (т.е. Два рейтинга одинаковы), коэффициент равен 1.
- Если расхождение между двумя рангами является идеальным (т.е. один рейтинг является обратным другому), коэффициент имеет значение -1.
- Если случайные величины *X* и *Y* независимы, то коэффициент будет приблизительно равен нулю.

ПРИМЕР 2. Рассчитать коэффициент ранговой корреляции Кендалла τ по таблице рангов депутатов ДО и ПОСЛЕ выборов

Ранг депутатов по экспертной оценке (x)	1	2	3	4	5	6	7	8	9	10
Ранг депутатов по результатам выборов (y)	2	1	7	6	3	4	5	9	10	8

Для нахождения величины S находят P и Q.

При определении P нужно установить, сколько чисел, находящихся справа от каждого из элементов последовательности рангов переменной у, имеют величину ранга, превышающую ранг рассматриваемого элемента. Так, например, первому значению в последовательности рангов переменной у, т.е. числу 2, соответствует 8 чисел (7, 6, 3, 4, 5, 9, 10, 8), которые превышают ранг 2; второму значению 1 соответствует также 8 чисел (7, 6, 3, 4, 5, 9, 10, 8); превышающих 1 и т.д. Суммируя полученные таким образом числа, мы получим число P, которое можно рассматривать как меру соответствия последовательности рангов переменной у последовательности рангов переменных х. P = 35 = 8+8+3+3+5+4+3+1 ; P – число согласованных пар выборки.

Число Q характеризует степень несоответствия последовательности рангов переменной у последовательности рангов переменной х, Q – число несогласованных пар в выборке. Чтобы определить Q подсчитаем, сколько чисел, находящихся справа от каждого из членов последовательности рангов переменной у имеет ранг меньше, чем эта единица. Q = 10= 1+ 0 + 4 +3 +0 +0 +1+ 1. Следовательно, S = P - Q = 35 - 10 = 25.

Коэффициент ранговой корреляции Кендалла τ равен:

$$\tau = \frac{2S}{n(n-1)} = \frac{2\cdot 25}{10\cdot 9} = 0{,}556$$

Статистическая проверка наличия корреляции. Проверим значимость коэффициента τ .
При $n \geq 10$ вычисляют статистику:

$$\tilde{\tau} = \frac{\tau}{\sqrt{D_\tau}}, \quad D_\tau = \frac{2(2n+5)}{9n(n-1)}$$

$$\tau_{экс} = 0{,}556 \,/\, \sqrt{\frac{2\cdot 25}{9\cdot 10\cdot 9}} = 2{,}238$$

Нулевая гипотеза $H_0 : \tau = 0$ - отвергается, т.к. $\tau_{экс}$ =2,238 > $U_{1-\frac{\alpha}{2}}$ = 1, 960 на уровне значимости α =0,05. Гипотезу о незначимости коэффициента Кендалла отвергаем, связь значима.
Ответ: τ =0,556; связь значима.

ПРИМЕР 3 . Требуется оценить корреляционную связь между скоростью чтения первоклассников и их усидчивостью. Скорость чтения первоклассников замерялась секундомером (слов/мин), усидчивость – с помощью экспертного оценивания по специально разработанной пятиточечной шкале: очень высокий (ОВ) – высокий (В) – средний (С) – низкий (Н) – очень низкий (ОН) уровни. Учителю предъявлялись карточки с описанием уровня проявления усидчивости, которые он должен был соотнести с поведением каждого ученика. Результаты измерений приведены в таблице 1. Решение. Скорость чтения первоклассников (переменная X) измерена в реляционной шкале, а их усидчивость (переменная Y) – в порядковой, поэтому определить связь между ними можно с помощью ранговой корреляции, проранжировав обе переменные. Рассчитаем коэффициенты ранговой корреляции Спирмена и Кендалла.

А). Расчет коэффициента корреляции Спирмена. В таблицу 1 с результатами добавим четыре колонки: две для ранговых последовательностей первой и второй переменных, для разности рангов d и для d 2 . Одинаковым значениям по второй переменной присваиваем средние ранговые значения

Табл. 1.

Расчет коэффициента ранговой корреляции Спирмена

Ф.И.	*X*	*Y*	R_x	R_y	$d = R_x - R_y$	d^2
1. Аня К.	12,0	Н	9	8,5	0,5	0,25
2. Боря Л.	18,8	ОН	6	10	-4	16
3. Вася Р.	11,0	В	10	2,5	7,5	56,25
4. Даша В.	29,0	С	2	5,5	-3,5	12,25
5. Зина С.	17,5	Н	7	8,5	-1,5	2,25
6. Игорь М.	23,4	ОВ	4	1	3	9
7. Катя Г.	35,6	С	1	5,5	-4,5	20,25
8. Лёня А.	15,4	С	8	5,5	2,5	6,25
9. Маша Д.	26,1	В	3	2,5	0,5	0,25
10. Яша Б.	20,7	С	5	5,5	-0,5	0,25
						123,00

Коэффициент Спирмена $r_S = 1 - \frac{6*123}{990} = 0,255.$

Б) Расчет коэффициента корреляции Кендалла. Для расчета коэффициента корреляции Кендалла необходимо расчетную таблицу перегруппировать по возрастанию рангов первой переменной. Последний столбец (R) заполняется следующим образом: ниже Кати Г. (Ry = 5,5) имеется 3 ранга Ry, бóльших, чем у Кати (5,5); ниже Даши В. (Ry = 5,5) имеется 3 ранга Ry, бóльших, чем у Даши (5,5); ниже Маши Д. (Ry = 2,5) имеется 5 рангов Ry, бóльших, чем у Маши (2,5). (Т.Е. вычисляем число согласованных пар P).

Таблица2

Расчет коэффициента ранговой корреляции Кендалла

Ф.И.	*X*	*Y*	R_x	R_y	*R*
7. Катя Г.	35,6	С	1	5,5	3
4. Даша В.	29,0	С	2	5,5	3
9. Маша Д.	26,1	В	3	2,5	5
6. Игорь М.	23,4	ОВ	4	1	6
10. Яша Б.	20,7	С	5	5,5	3
2. Боря Л.	18,8	ОН	6	10	0
5. Зина С.	17,5	Н	7	8,5	0
8. Лёня А.	15,4	С	8	5,5	1
1. Аня К.	12,0	Н	9	8,5	0
3. Вася Р.	11,0	В	10	2,5	0
					21

Коэффициент Кендалла: $\tau = \frac{4\sum_{i=1}^{n-1} R_i}{n(n-1)} - 1$ = - 0,067

$$\tau = 1 - \frac{4Q}{N(N-1)}$$

$$\tau = \frac{4P}{N(N-1)} - 1$$