

О коэффициенте корреляции Кендалла*

А. В. Степанов

Балтийский федеральный университет имени И. Канта,
Российская Федерация, 236041, Калининград, ул. Александра Невского, 14

Для цитирования: Степанов А. В. О коэффициенте корреляции Кендалла // Вестник Санкт-Петербургского университета. Математика. Механика. Астрономия. 2025. Т. 12 (70). Вып. 1. С. 91–101. <https://doi.org/10.21638/spbu01.2025.107>

В настоящей статье в непрерывном случае изучается коэффициент корреляции Кендалла. В начале работы рассматриваются коэффициент корреляции Пирсона ρ и его выборочный аналог ρ_n , который при больших n хорошо аппроксимирует ρ , поскольку сходится к нему по вероятности. Далее обсуждаются ранговый коэффициент корреляции Кендалла τ_n и его теоретический аналог τ . В непрерывном случае τ_n определяется в терминах рангов конкомитантов порядковых статистик. В статье показывается, что $E\tau_n = \tau$ и τ_n сходится по вероятности к τ . Таким образом, коэффициент τ_n также хорошо аппроксимирует τ , как это и происходит в случае с коэффициентами ρ_n и ρ . Данное обнаружение объясняет причину, по которой τ тоже можно считать теоретическим коэффициентом корреляции. Во многих работах τ используется в качестве коэффициента корреляции без объяснений, почему его можно считать таковым. Поскольку коэффициент τ мало изучен, далее в работе исследуются его основные свойства, преимущества и недостатки, происходит его сравнение с коэффициентом ρ . Среди преимуществ τ выделим то, что он существует для любых непрерывных распределений. В завершении работы приводятся примеры.

Ключевые слова: двумерные распределения, конкомитанты порядковых статистик, коэффициенты корреляции Пирсона и Кендалла.

1. Введение. Пусть $(X, Y), (X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ — независимые одинаково распределенные непрерывные случайные векторы, имеющие двумерную функцию распределения $F(x, y)$ и маргинальные функции распределения $H(x)$ и $G(y)$. В абсолютно непрерывном случае функции плотностей будем обозначать как $f(x, y), h(x)$ и $g(y)$ соответственно. Обозначение $F_x(x, y)$ будем использовать для записи плотности функции распределения, где плотность описывает поведение величины X , а функция распределения определяет поведение величины Y . Схожим образом будет использоваться запись $F_y(x, y)$. Пусть $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ — порядковые статистики, полученные из выборки $X_1, X_2, \dots, X_n$. Для данных порядковых статистик определим их конкомитанты $Y_{[1]}, Y_{[2]}, \dots, Y_{[n]}$. Пусть $X_i = X_{(j)}$. Тогда $Y_{[j]} = Y_i$ назовем конкомитантом порядковой статистики $X_{(j)}$. Конкомитанты порядковых статистик были предложены в работах [1–2] и обсуждались впоследствии в работах [3–11].

Зависимость случайных величин X и Y характеризуется коэффициентом корреляции Пирсона:

$$\rho = \frac{E(X - EX)(Y - EY)}{\sigma_X \sigma_Y}.$$

*Работа поддержана Министерством науки и высшего образования РФ (номер соглашения 075-02-2025-1789).

© Санкт-Петербургский государственный университет, 2025

$$\rho_n = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

неплохо аппроксимирует ρ при больших n , так как $\rho_n \xrightarrow{P} \rho$ (см., например, [12]). Пусть $Y_{[1]}, \dots, Y_{[n]}$ — конкомитанты порядковых статистик $X_{(1)} \leq \dots \leq X_{(n)}$, а

$$r_j = \sum_{i=1}^{j-1} I_{i,j} = \sum_{i=1}^{j-1} I_{\{Y_{[i]} \leq Y_{[j]}\}} \quad (1 \leq i < j \leq n)$$

(где $I_{i,j}$ — индикаторная функция) — ранг конкомитанта $Y_{[j]}$ среди конкомитантов $Y_{[1]}, \dots, Y_{[j]}$. Следующая величина

$$\tau_n = \frac{4 \sum_{j=2}^n r_j}{n(n-1)} - 1 = \frac{4}{n(n-1)} \sum_{j=2}^n \sum_{i=1}^{j-1} I_{i,j} - 1$$

известна как ранговый коэффициент корреляции Кендалла (см., например, [13]). Ранговый коэффициент корреляции τ_n может использоваться наряду с ρ_n . В отдельных случаях, например тогда, когда выборки X_i и Y_j неизвестны, но ранги r_j известны, только использование τ_n помогает оценить величину зависимости двух выборок. Отметим, что существует еще один ранговый коэффициент корреляции — коэффициент корреляции Спирмана, но мы не будем обсуждать его в настоящей работе.

Основные свойства коэффициента τ_n следующие. Если соответствие (несоответствие) между величинами $X_{(j)}$ и рангами r_j ($1 \leq j \leq n$) сильное, то коэффициент τ_n находится около единицы (минус единицы). Если величины X и Y независимы, то $Y_{[j]}$ ($j = 1, \dots, n$) — независимые одинаково распределенные случайные величины. Очевидно, что тогда $P(Y_{[i]} \leq Y_{[j]}) = 1/2 \forall i, j, i \neq j$. Откуда $\tau_n \rightarrow 0$ п. н. Для коэффициента τ_n были найдены верхние и нижние оценки (см., например, [14–16]).

Дальнейшее содержание работы следующее. Во втором разделе работы находится $E\tau_n$ и показывается, что τ_n сходится по вероятности к $E\tau_n = \tau$. Данный факт указывает на то, что рассмотрение τ в качестве теоретического коэффициента корреляции является оправданным. Во многих научных работах τ уже использовался в качестве теоретического коэффициента корреляции, но не было обоснования того, почему его можно считать таковым. Коэффициент корреляции τ может выступать в качестве альтернативы классическому коэффициенту корреляции ρ . Поскольку τ недостаточно изучен, в разделе 3 настоящей статьи обсуждаются свойства коэффициента τ , его преимущества и недостатки, происходит сравнение с коэффициентом ρ . Среди преимуществ отметим следующее. Для существования τ никакие моментные предположения не требуются, т.е. этот коэффициент корреляции существует для любых невырожденных двумерных непрерывных распределений. В завершающем разделе 4 предлагаются примеры.

2. Математическое ожидание τ_n и сходимость по вероятности.

Лемма 1. Для $n \geq 2$ справедливо

$$E\tau_n = 4EF(X, Y) - 1 = 4 \int_{\mathbb{R}^2} F(x, y)F(dx, dy) - 1 = \quad (1)$$

$$= 1 - 4 \int_{\mathbb{R}^2} F(dx, y)F(x, dy). \quad (2)$$

ДОКАЗАТЕЛЬСТВО. Очевидно, что из (1) вытекает (2). Представим доказательство равенства (1) для случая, когда F — абсолютно непрерывная функция распределения, т. е. покажем справедливость равенства

$$E\tau_n = 4 \int_{x_1 \leq x_2, y_1 \leq y_2} f(x_1, y_1)f(x_2, y_2)dx_1dy_1dx_2dy_2 - 1.$$

Совместная плотность конкомитантов $Y_{[i]}, Y_{[j]}$ ($i < j$) задается формулой (см., например, [5])

$$f_{Y_{[i]}, Y_{[j]}}(y_i, y_j) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} \times \\ \times \int_{x_i \leq x_j} f(x_i, y_i)H^{i-1}(x_i)f(x_j, y_j)(H(x_j) - H(x_i))^{j-i-1}(1 - H(x_j))^{n-j}dx_id x_j.$$

Тогда

$$Er_j = \sum_{i=1}^{j-1} P(Y_{[i]} \leq Y_{[j]}) = \frac{n!}{(j-2)!(n-j)!} \times \\ \times \int_{x_i \leq x_j, y_i \leq y_j} f(x_i, y_i)f(x_j, y_j)H^{j-2}(x_j)(1 - H(x_j))^{n-j}dx_id x_jdy_id y_j.$$

Откуда находим, что

$$E\tau_n = \frac{4 \sum_{j=2}^n Er_j}{n(n-1)} - 1 = 4 \int_{x_i \leq x_j, y_i \leq y_j} f(x_i, y_i)f(x_j, y_j)dx_id x_jdy_id y_j - 1. \quad (3)$$

□

Замечание 1. В случае, когда F — абсолютно непрерывная функция распределения, равенства (1) и (2) можно записать в виде

$$E\tau_n = 4 \int_{\mathbb{R}^2} F(x, y)f(x, y)dx dy - 1 \quad (4)$$

и

$$E\tau_n = 1 - 4 \int_{\mathbb{R}^2} F_x(x, y)F_y(x, y)dx dy. \quad (5)$$

Положим $\tau = E\tau_n$.

Теорема 1. *Справедливо следующее асимптотическое соотношение:*

$$\tau_n \xrightarrow{P} \tau \quad (n \rightarrow \infty).$$

ДОКАЗАТЕЛЬСТВО. Докажем теорему 1 для случая, когда F — абсолютно непрерывное распределение. Из неравенства Чебышёва следует, что для любого $\varepsilon > 0$

$$\begin{aligned} P(|\tau_n - \tau| > \varepsilon) &\leq \frac{D\tau_n}{\varepsilon^2} = \\ &= \frac{16}{\varepsilon^2 n^2 (n-1)^2} \left[\sum_{j=2}^n \sum_{i=1}^{j-1} \sum_{k=2}^n \sum_{l=1}^{k-1} E(I_{i,j} I_{l,k}) - \left(E \sum_{j=2}^n \sum_{i=1}^{j-1} I_{i,j} \right)^2 \right]. \quad (6) \end{aligned}$$

Поскольку выражение (6) неотрицательно, то мы не будем проверять это ниже в доказательстве. Мы только покажем, что выражение, стоящее в квадратных скобках в (6), ведет себя как $O(n^3)$. Из равенства (3) вытекает, что $\left(E \sum_{j=2}^n \sum_{i=1}^{j-1} I_{i,j} \right)^2 = (n(n-1)p)^2$, где $p = \int_{\mathbb{R}^2} F(x, y) f(x, y) dx dy$. Пусть

$$\sum_{j=2}^n \sum_{i=1}^{j-1} \sum_{k=2}^n \sum_{l=1}^{k-1} E(I_{i,j} I_{l,k}) = \sum_{j=2}^n \sum_{i=1}^{j-1} \sum_{k=2}^n \sum_{l=1}^{k-1} P(Y_{[i]} \leq Y_{[j]}, Y_{[l]} \leq Y_{[k]}) = \Sigma_1 + \Sigma_2,$$

где

$$\begin{aligned} \Sigma_1 &= \sum_{1 \leq i < j < l < k \leq n} + \sum_{1 \leq i < l < j < k \leq n} + \sum_{1 \leq l < i < j < k \leq n} + \\ &+ \sum_{1 \leq l < k < i < j \leq n} + \sum_{1 \leq l < i < k < j \leq n} + \sum_{1 \leq i < l < k < j \leq n} \end{aligned}$$

и

$$\begin{aligned} \Sigma_2 &= \sum_{1 \leq i < j = l < k \leq n} + \sum_{1 \leq i = l < j < k \leq n} + \sum_{1 \leq i < l < j = k \leq n} + \\ &+ \sum_{1 \leq i = l < j = k \leq n} + \sum_{1 \leq l < i < j = k \leq n} + \sum_{1 \leq l < k = i < j \leq n} + \sum_{1 \leq l = i < k < j \leq n}. \end{aligned}$$

Здесь в сумме Σ_1 первое слагаемое $\sum_{1 \leq i < j < l < k \leq n}$ обозначает краткую форму записи $\sum_{k=4}^n \sum_{l=3}^{k-1} \sum_{j=2}^{l-1} \sum_{i=1}^{j-1} P(Y_{[i]} \leq Y_{[j]}, Y_{[l]} \leq Y_{[k]})$. Аналогичные обозначения используются для остальных слагаемых суммы Σ_1 . В свою очередь, первое слагаемое $\sum_{1 \leq i < j = l < k \leq n}$ суммы Σ_2 обозначает краткую форму записи $\sum_{k=3}^n \sum_{j=2}^{k-1} \sum_{i=1}^{j-1} P(Y_{[i]} \leq Y_{[j]} \leq Y_{[k]})$ и другие слагаемые суммы Σ_2 обозначены аналогично. Оценим асимптотическое поведение слагаемых сумм Σ_1 и Σ_2 . Рассмотрим первое слагаемое суммы Σ_1 :

$$\sum_{1 \leq i < j < l < k \leq n} = \sum_{k=4}^n \sum_{l=3}^{k-1} \sum_{j=2}^{l-1} \sum_{i=1}^{j-1} P(Y_{[i]} \leq Y_{[j]}, Y_{[l]} \leq Y_{[k]}).$$

Пользуясь совместной плотностью конкомитантов $Y_{[i]}, Y_{[j]}, Y_{[l]}, Y_{[k]}$ (см., например, [5]), имеем

$$\begin{aligned} \sum_{1 \leq i < j < l < k \leq n} &= \sum_{k=4}^n \sum_{l=3}^{k-1} \sum_{j=2}^{l-1} \sum_{i=1}^{j-1} \frac{n!}{(i-1)!(j-i-1)!(l-j-1)!(k-l-1)!(n-k)!} \times \\ &\times \int_{x_i \leq x_j \leq x_l \leq x_k, y_i \leq y_j, y_l \leq y_k} \times \\ &\times f(x_i, y_i) f(x_j, y_j) f(x_l, y_l) f(x_k, y_k) H^{i-1}(x_i) (H(x_j) - H(x_i))^{j-i-1} \times \\ &\times (H(x_l) - H(x_j))^{l-j-1} (H(x_k) - H(x_l))^{k-l-1} \times \\ &\times (1 - H(x_k))^{n-k} dx_i dx_j dx_l dx_k dy_i dy_j dy_l dy_k = \\ &= n(n-1)(n-2)(n-3) \int_{x_i \leq x_j \leq x_l \leq x_k, y_i \leq y_j, y_l \leq y_k} f(x_i, y_i) f(x_j, y_j) f(x_l, y_l) \times \\ &\times f(x_k, y_k) dx_i dx_j dx_l dx_k dy_i dy_j dy_l dy_k = \\ &= n(n-1)(n-2)(n-3) I_{x_i \leq x_j \leq x_l \leq x_k, y_i \leq y_j, y_l \leq y_k}. \end{aligned}$$

Аналогичный результат получаем для оставшихся пяти слагаемых Σ_1 . Так, например,

$$\sum_{1 \leq i < l < j < k \leq n} = n(n-1)(n-2)(n-3) I_{x_i \leq x_l \leq x_j \leq x_k, y_i \leq y_j, y_l \leq y_k}.$$

Откуда

$$\Sigma_1 = n(n-1)(n-2)(n-3) I_{x_i \leq x_j, x_l \leq x_k, y_i \leq y_j, y_l \leq y_k} = n(n-1)(n-2)(n-3) p^2$$

и $\Sigma_1 - \left(E \sum_{j=2}^n \sum_{i=1}^{j-1} I_{i,j} \right)^2 = -2p^2 n(n-1)(2n-3) < 0$. Теперь оценим слагаемые суммы Σ_2 . Пользуясь совместной плотностью конкомитантов $Y_{[i]}, Y_{[j]}$, оценим первое слагаемое. Имеем

$$\begin{aligned} \sum_{1 \leq i=l < j=k \leq n} &= \sum_{j=2}^n \sum_{i=1}^{j-1} P(Y_{[i]} \leq Y_{[j]}) = \\ &= n(n-1) \int_{x_i \leq x_j, y_i \leq y_j} f(x_i, y_i) f(x_j, y_j) dx_i dx_j dy_i dy_j = n(n-1) p = o(n^3). \end{aligned}$$

Несложно показать, что $\sum_{1 \leq i=l < j < k \leq n} = \sum_{1 \leq i=l < k < j \leq n}$. Пользуясь формулой совместной плотности конкомитантов $Y_{[i]}, Y_{[j]}, Y_{[k]}$, находим, что

$$\begin{aligned} \sum_{1 \leq i=l < j < k \leq n} &= \sum_{k=3}^n \sum_{j=2}^{k-1} \sum_{i=1}^{j-1} \frac{n!}{(i-1)!(j-i-1)!(k-j-1)!(n-k)!} \times \\ &\times \int_{x_i \leq x_j \leq x_k, y_i \leq y_j, y_i \leq y_k} f(x_i, y_i) f(x_j, y_j) f(x_k, y_k) H^{i-1}(x_i) (H(x_j) - H(x_i))^{j-i-1} \times \\ &\times (H(x_k) - H(x_j))^{k-j-1} (1 - H(x_k))^{n-k} dx_i dx_j dx_k dy_i dy_j dy_k = \\ &= n(n-1)(n-2) \int_{x_i \leq x_j \leq x_k, y_i \leq y_j, y_i \leq y_k} f(x_i, y_i) f(x_j, y_j) f(x_k, y_k) dx_i dx_j dx_k dy_i dy_j dy_k \leq \\ &\leq n(n-1)(n-2) = O(n^3). \end{aligned}$$

Аналогичным образом можно показать, что

$$\sum_{1 \leq l < i = k < j \leq n} = \sum_{1 \leq i < j = l < k \leq n} = n(n-1)(n-2) \int_{x_i \leq x_j \leq x_k, y_i \leq y_j \leq y_k} \times \\ \times f(x_i, y_i) f(x_j, y_j) f(x_k, y_k) dx_i dx_j dx_k dy_i dy_j dy_k = O(n^3).$$

Схожим образом также находим, что

$$\sum_{1 \leq l < i < j = k \leq n} = \sum_{1 \leq i < l < j = k \leq n} = n(n-1)(n-2) \int_{x_i \leq x_l \leq x_j, y_i \leq y_j, y_l \leq y_j} \times \\ \times f(x_i, y_i) f(x_j, y_j) f(x_l, y_l) dx_i dx_j dx_l dy_i dy_j dy_l = O(n^3).$$

Теперь из (6) вытекает, что $P(|\tau_n - \tau| > \varepsilon) \rightarrow 0$ ($n \rightarrow \infty$). Теорема доказана для абсолютно непрерывного случая. Доказательство в непрерывном случае может быть проведено аналогичными методами. $\square$

3. Коэффициент корреляции Кендалла и его сравнение с коэффициентом Пирсона. В непрерывном случае величина

$$\tau = 4EF(X, Y) - 1 = 4 \int_{\mathbb{R}^2} F(x, y) F(dx, dy) - 1 \quad (1)$$

была предложена (см., например, с. 155 книги [4]) в качестве нового теоретического коэффициента корреляции — коэффициента Кендалла. Данное предложение является обоснованным, поскольку τ , как было показано в разделе 2, хорошо аппроксимируется выборочным коэффициентом корреляции Кендалла τ_n (τ_n сходится к τ по вероятности и $E\tau_n = \tau$). Отметим, что в абсолютно непрерывном случае также справедливы равенства

$$\tau = 4 \int_{\mathbb{R}^2} F(x, y) f(x, y) dx dy - 1 = \quad (2)$$

$$= 1 - 4 \int_{\mathbb{R}^2} F_x(x, y) F_y(x, y) dx dy. \quad (3)$$

Основные свойства τ следующие.

Свойство 1. Очевидно, что если величины X и Y независимы, то $\tau = 0$.

Поскольку τ , так же как и ρ , дает некоторое усредненное значение величины зависимости X и Y , то из того, что $\tau = 0$, не следует, что X и Y независимы. Ниже, в примере 5, приводится функция распределения F , для которой величины X и Y зависимы, но $\tau = 0$.

Свойство 2. Для любой невырожденной двумерной функции распределения F справедливо неравенство $-1 < \tau < 1$.

Доказательство. Докажем свойство для абсолютно непрерывного случая. Поскольку подынтегральные выражения в (2) и (5) неотрицательны, то из (2) и (5)

вытекает справедливость неравенства $-1 \leq \tau \leq 1$. Покажем теперь, что $\tau \neq \pm 1$. Пусть $\tau = 1$. Тогда из (5) следует, что

$$\int_{\mathbb{R}^2} F_x(x, y) F_y(x, y) dx dy = 0.$$

Поскольку $F_x(x, y), F_y(x, y)$ — неотрицательные функции, то либо $F_x(x, y) = 0$, либо $F_y(x, y) = 0$. Откуда $f(x, y) = 0$ для любой точки (x, y) . Но тогда из (2) следует, что $\tau = -1$. Получили противоречие. Аналогичные рассуждения можно провести, предположив вначале, что $\tau = -1$. Таким образом, $\tau \neq \pm 1$. $\square$

Данное свойство также показывает, что коэффициент τ , заданный равенством (1), не определен в случае, когда F — вырожденная функция распределения, т. е. тогда, когда $Y = \varphi(X)$, где φ — некоторая измеримая функция. Доопределим коэффициент τ для этого случая. Отметим, что если $y = \varphi(x)$ — монотонно возрастающая непрерывная функция, то у нее существует обратная монотонно возрастающая непрерывная функция, и тогда события $\{X \leq x\}$ и $\{\varphi(X) \leq \varphi(x)\}$ эквивалентны. Поэтому можно считать, что $f_{X, \varphi(X)}(x, \varphi(x)) = h(x)$ и $F_{X, \varphi(X)}(x, \varphi(x)) = H(x)$. Тогда

$$\tau = 4 \int_{\mathbb{R}^2} F(x, y) f(x, y) dx dy - 1 = 4 \int_{\mathbb{R}} H(x) h(x) dx - 1 = 1.$$

Обратная цепочка рассуждений также справедлива. В результате приходим к следующему свойству.

Свойство 3. Для того чтобы было справедливо равенство $\tau = 1$, необходимо и достаточно, чтобы $Y = \varphi(X)$, где φ — монотонно возрастающая функция. Для того чтобы было справедливо равенство $\tau = -1$, необходимо и достаточно, чтобы $Y = \varphi(X)$, где ψ — монотонно убывающая функция.

В случае, когда F — вырожденная функция распределения и функция φ — функция иного вида, будем считать, что коэффициент τ не определен.

Коэффициенты корреляции τ и ρ имеют определенные преимущества и недостатки.

Преимущества τ :

1. Для существования τ не требуются моментные предположения. Коэффициент τ существует для любого непрерывного F .

2. Для непрерывного случая справедливы соотношения $\tau_n \xrightarrow{p} \tau$ и $\rho_n \xrightarrow{p} \rho$. Таким образом, оба коэффициента τ и ρ хорошо аппроксимируются для больших n выборочными коэффициентами корреляции τ_n и ρ_n . Однако $E\tau_n = \tau$.

Преимущества ρ :

1. Коэффициент корреляции ρ ищется проще. Для нахождения ρ необходимо знать только первые и вторые моменты. Для нахождения τ необходимо знать функцию распределения F . Здесь ρ имеет преимущество, так как при анализе многих статистических экспериментов (например, в теории временных рядов) используются только первые и вторые моменты.

2. Коэффициент корреляции ρ явно присутствует и привычен во многих статистических моделях и формулах. Например, он присутствует в формуле линейной регрессии, в двумерной нормальной функции распределения и т. д.

Дальнейшее обсуждение преимуществ и недостатков ρ и τ происходит на примерах в разделе 4.

4. Примеры. В этом разделе представлены примеры, иллюстрирующие полученные выше теоретические результаты.

Пример 1. Для нормального случая найдем коэффициент τ . Пусть

$$F(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^x \int_{-\infty}^y e^{-\frac{u^2-2\rho uv+v^2}{2(1-\rho^2)}} dudv \quad (x, y \in \mathbb{R})$$

— двумерное нормальное распределение с маргинальными распределениями

$$H(x) = G(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{u^2}{2}} du \quad (x \in \mathbb{R})$$

и маргинальными плотностями $h(x) = g(x) = \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. Отметим, что

$$F_x(x, y) = \phi(x) \Phi\left(\frac{y - \rho x}{\sqrt{1 - \rho^2}}\right) \quad \text{и} \quad F_y(x, y) = \phi(y) \Phi\left(\frac{x - \rho y}{\sqrt{1 - \rho^2}}\right).$$

Из (2) следует, что

$$\tau(\rho) = 1 - 4 \int_{\mathbb{R}^2} \phi(x) \phi(y) \Phi\left(\frac{x - \rho y}{\sqrt{1 - \rho^2}}\right) \Phi\left(\frac{y - \rho x}{\sqrt{1 - \rho^2}}\right) dx dy$$

и

$$\begin{aligned} \tau'(\rho) &= 4 \int_{\mathbb{R}^2} \phi(x) \phi(y) \phi\left(\frac{x - \rho y}{\sqrt{1 - \rho^2}}\right) \left(\frac{y - \rho x}{(1 - \rho^2)^{3/2}}\right) \Phi\left(\frac{y - \rho x}{\sqrt{1 - \rho^2}}\right) dx dy + \\ &+ 4 \int_{\mathbb{R}^2} \phi(x) \phi(y) \phi\left(\frac{y - \rho x}{\sqrt{1 - \rho^2}}\right) \left(\frac{x - \rho y}{(1 - \rho^2)^{3/2}}\right) \Phi\left(\frac{x - \rho y}{\sqrt{1 - \rho^2}}\right) dx dy. \end{aligned}$$

В силу симметрии имеем

$$\tau'(\rho) = 8 \int_{\mathbb{R}^2} \phi(x) \phi(y) \phi\left(\frac{x - \rho y}{\sqrt{1 - \rho^2}}\right) \frac{y - \rho x}{(1 - \rho^2)^{3/2}} \Phi\left(\frac{y - \rho x}{\sqrt{1 - \rho^2}}\right) dx dy.$$

Откуда получаем, что

$$\begin{aligned} \tau'(\rho) &= \frac{8}{\sqrt{2\pi}(1 - \rho^2)} \int_{\mathbb{R}} \frac{e^{-x^2}}{\sqrt{2\pi}} \left(\int_{\mathbb{R}} \phi\left(\frac{y - \rho x}{\sqrt{1 - \rho^2}}\right) \frac{y - \rho x}{1 - \rho^2} \Phi\left(\frac{y - \rho x}{\sqrt{1 - \rho^2}}\right) dy \right) dx = \\ &= \frac{-8}{\sqrt{2\pi}(1 - \rho^2)} \int_{\mathbb{R}} \frac{e^{-x^2}}{\sqrt{2\pi}} \left(\int_{\mathbb{R}} \Phi\left(\frac{y - \rho x}{\sqrt{1 - \rho^2}}\right) d\phi\left(\frac{y - \rho x}{\sqrt{1 - \rho^2}}\right) \right) dx. \end{aligned}$$

Интегрируя по частям и делая замену $z = \frac{y - \rho x}{\sqrt{1 - \rho^2}}$, находим, что

$$\tau'(\rho) = \frac{4}{\pi\sqrt{1 - \rho^2}} \int_{\mathbb{R}} \frac{e^{-x^2}}{\sqrt{2\pi}} dx \int_{\mathbb{R}} \frac{e^{-z^2}}{\sqrt{2\pi}} dz = \frac{2}{\pi\sqrt{1 - \rho^2}}.$$

Откуда $\tau = \frac{2}{\pi} \arcsin \rho$. Следует отметить, что ранее было показано (см., например, [13]), что $E\tau_n = \frac{2}{\pi} \arcsin \rho$. В данной статье мы предлагаем свой более простой вариант нахождения τ .

Пример 2. Пусть

$$F(x, y) = 1 - \frac{1}{(y+1)^t} - \frac{1}{(x+1)^t} + \frac{1}{(x+y+1)^t} \quad (x > 0, y > 0, t > 0).$$

Из (4) следует, что $\tau = \frac{1}{2t+1}$ ($t > 0$). Также можно показать, что $\rho = \frac{1}{t}$ ($t > 2$). Отметим, что в то время как коэффициент ρ при $0 < t \leq 2$ не определен, коэффициент τ добросовестно показывает величину зависимости X и Y .

Пример 3. Пусть

$$F(x, y) = \int_{-\infty}^x \int_{-\infty}^y \frac{1}{2\pi(1+u^2+v^2)^{3/2}} dudv \quad (x, y \in \mathbb{R})$$

— двумерное симметрическое распределение Коши с маргинальными распределениями $H(x) = G(y) = \frac{1}{\pi} \arctg(x)$ ($x \in \mathbb{R}$). Очевидно, что здесь ρ не существует и через него мы не имеем никакой информации о величине зависимости X и Y . Из равенства (2) следует, что

$$\tau = 1 - \frac{1}{\pi^2} \int_{\mathbb{R}^2} \frac{1}{(1+x^2)(1+y^2)} \left(\frac{x}{\sqrt{1+x^2+y^2}} + 1 \right) \left(\frac{y}{\sqrt{1+x^2+y^2}} + 1 \right) dx dy.$$

Пользуясь численными методами, получаем равенство $\tau = 4.154334 \cdot 10^{-12}$. Мы видим, что величина зависимости случайных величин X и Y крайне мала.

Пример 4. Пусть

$$F(x, y) = 1 - e^{-x} - \frac{1}{(1+y)^t} + \frac{e^{-x(1+y)^t}}{(1+y)^t} \quad (x, y > 0).$$

Несложно показать, что $\rho = -\frac{\sqrt{t(t-2)}}{2t-1}$ ($t > 2$) и $\tau = -\frac{1}{2}$ ($t > 0$). С одной стороны, коэффициент ρ не определен при $t \in (0, 2]$, но, с другой стороны, τ не зависит от t , и поэтому при $t > 2$ коэффициент ρ лучше отражает зависимость между X и Y .

Пример 5. Пусть функция распределения F задается равенствами:

$$F(x, y) = \frac{x(y+1)(1+\alpha(1-x)y)}{2} \quad (0 \leq x \leq 1, -1 \leq y \leq 0),$$

$$F(x, y) = \frac{x}{2} + \frac{xy(1+\alpha(1-x)(1-y))}{2} \quad (0 \leq x \leq 1, 0 < y \leq 1),$$

где $\alpha \in [-1, 1]$. Здесь, при $\alpha \neq 0$, величины X и Y зависимы, но $\tau = \rho = 0$.

Наброски настоящей статьи в 2015 г. были выложены автором в архив (см. [17]).

Литература/References

1. David H.A. Concomitants of extreme order statistics. In: *Extreme Value Theory and Applications*. Proceedings of the conference on extreme value theory and applications **1**. J. Galambos, J. Lechner, E. Simiu (eds). Kluwer Academic Publishers, Boston, 211–224 (1994).

2. Bhattacharya B. B. Convergence of sample paths of normalized sums of induced order statistics. *Ann. Statist.* **2**, 1034–1039 (1974).
3. Bairamov I., Stepanov A. Numbers of near-maxima for the bivariate case. *Statistics & Probability Letters* **80**, 196–205 (2010).
4. Balakrishnan N., Lai C. D. *Continuous bivariate distributions*. Second edition, Springer (2009).
5. Balakrishnan N., Stepanov A. Limit results for concomitants of order statistics. *Metrika* **78**, 385–397 (2015).
6. Bhattacharya B. B. Induced order statistics: theory and applications. In: *Handbook of Statistics* **4**, 383–403. Krishnaiah P. R., Sen P. K. (eds), Amsterdam (1984).
7. Chu S. J., Huang W. J., Chen H. A study of asymptotic distributions of concomitants of certain order statistics. *Statist. Sinica* **9**, 811–830 (1999).
8. David H. A., Galambos J. The asymptotic theory of concomitants of order statistics. *J. Appl. Probab.* **11**, 762–770 (1974).
9. Egorov V. A., Nevzorov V. B. Rate of convergence to the normal law of sums of induced order statistics. *Journal of Soviet Mathematics* **25**, 1139–1146 (1984).
10. David H. A., Nagaraja H. N. *Order Statistics*. 3rd ed. New York, John Wiley & Sons (2003).
11. Goel P. K., Hall P. K. On the average difference between concomitants and order statistics. *Ann. Probab.* **22**, 126–144 (1994).
12. Fisher R. A. On the “probable error” of a coefficient of correlation deduced from a small sample. *Metron* **1**, 3–32 (1921).
13. Kendall M. G. *Rank Correlation Methods*. London, Griffin (1970).
14. Daniels H. E. Rank correlation and population models. *Journal of the Royal Statistical Society. Ser. B* **12** (2), 171–191 (1950).
15. Durbin J., Stuart A. Inversions and rank correlation coefficients. *Journal of the Royal Statistical Society. Ser. B* **13** (2), 303–309 (1951).
16. Xu W., Hou Y., Hung Y. S., Zou Y. Comparison of Spearman’s rho and Kendall’s tau in Normal and Contaminated Normal Models. *arXiv:1011.2009v1 [cs.IT]* (2009).
17. Stepanov A. On the Kendall Correlation Coefficient. *arXiv: 1507.01427v1* (2015).

Статья поступила в редакцию 16 марта 2024 г.;
доработана 5 июня 2024 г.;
рекомендована к печати 29 августа 2024 г.

Контактная информация:

Степанов Алексей Васильевич — д-р физ.-мат. наук, проф.;
<https://orcid.org/0000-0001-8979-4748>, alexeistep45@mail.ru

On Kendall’s correlation coefficient*

A. V. Stepanov

Immanuel Kant Baltic Federal University,
14, ul. Aleksandra Nevskogo, Kaliningrad, 236041, Russian Federation

For citation: Stepanov A. V. On Kendall’s correlation coefficient. *Vestnik of Saint Petersburg University. Mathematics. Mechanics. Astronomy*, 2025, vol. 12 (70), issue 1, pp. 91–101.
<https://doi.org/10.21638/spbu01.2025.107> (In Russian)

In the present paper, the Kendall correlation coefficient is studied in the continuous case. In the beginning of the paper, we discuss the Pearson correlation coefficient ρ and its statistical analogue ρ_n , which is a good approximation for ρ for large n since it converges to ρ in probability. We further discuss the Kendall rank correlation coefficient τ_n and its theoretical analogue τ . In the continuous case, τ_n is defined in the terms of ranks of concomitants of order statistics. It is shown that $E\tau_n = \tau$ and that τ_n converges in probability to τ . That

*The work was supported by the Ministry of Science and Higher Education of the Russian Federation (agreement no. 075-02-2025-1789).

way, τ_n is also a good approximation for τ , as well as it is for the coefficients ρ_n and ρ . This finding explains why τ can also be considered a theoretical correlation coefficient. In many works τ was already used as a theoretical correlation coefficient without explanation why it can be considered as such. Since coefficient τ has been little studied, we then discuss the basic properties of τ , advantages and disadvantages, compare it with coefficient ρ . Amongst the advantages of τ we note that τ exists for any continuous distributions. At the end of this work, we present some illustrative examples.

Keywords: bivariate distributions, concomitants of order statistics, Pearson's and Kendall's correlation coefficients.

Received: March 16, 2024

Revised: June 5, 2024

Accepted: August 29, 2024

Author's information:

Alexei V. Stepanov — <https://orcid.org/0000-0001-8979-4748>, alexeistep45@mail.ru