ЛЕКЦИЯ 7

ТЕМА: Корреляционный анализ

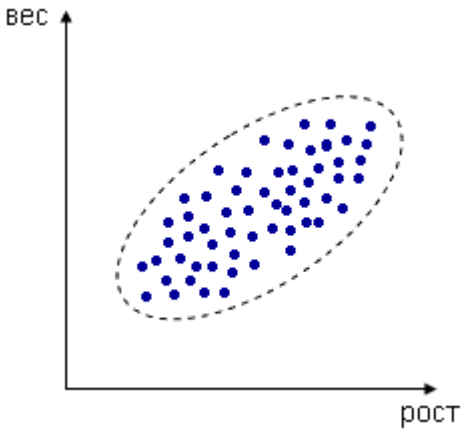
Вопросы темы. Понятие корреляции случайных величин, коэффициент корреляции. Параметрические корреляционые меры. Коэффициент корреляции Пирсона. Непараметрические корреляционые меры. коэффициентов корреляции Спирмена и Кенделла

Корреляционный анализ

Корреляционный анализ- метод, позволяющий обнаружить зависимость между несколькими случайными величинами.

Допустим, проводится независимое измерение различных параметров у одного типа объектов. Из этих данных можно получить качественно новую информацию - о взаимосвязи этих параметров.

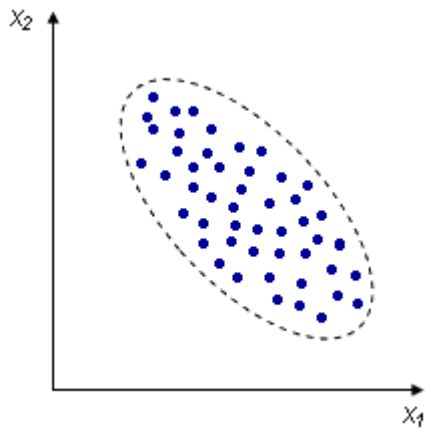
Например, измеряем рост и вес человека, каждое измерение представлено точкой в двумерном пространстве:



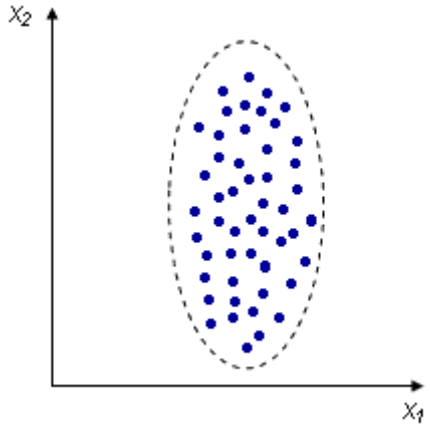
Несмотря на то, что величины носят случайный характер, в общем наблюдается некоторая зависимость - величины коррелируют.

В данном случае это **положительная корреляция** (при увеличении одного параметра второй тоже увеличивается). Возможны также такие случаи:

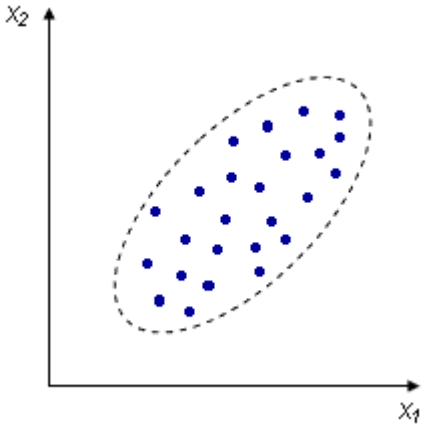
Отрицательная корреляция:

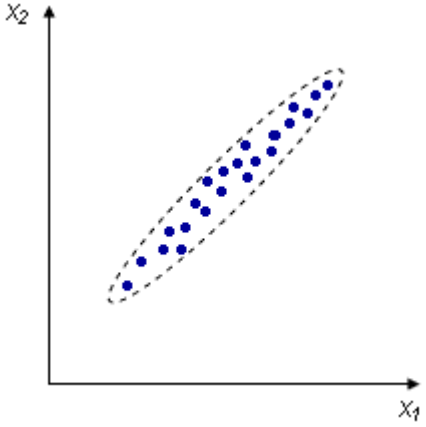


Отсутствие корреляции:



Взаимосвязь между переменными необходимо охарактеризовать численно, чтобы, например, различать такие случаи:





Для этого вводится **коэффициент корреляции**. Он рассчитывается следующим образом:

Коэффициент корреляции Пирсона

\Коэффициент корреляции Пирсона характеризует существование линейной зависимости между двумя величинами.

Пусть даны две выборки $x^m = (x_1, \cdots, x_m),\ y^m = (y_1, \cdots, y_m)$; коэффициент корреляции Пирсона рассчитывается по формуле:

где $\overline{x}, \overline{y}$ – выборочные средние x^m и y^m, s_x^2, s_y^2 – выборочные дисперсии, $r_{xy} \in [-1,1]$.

Коэффициент корреляции Пирсона называют также теснотой линейной связи:

- $|r_{xy}|=1 \Rightarrow x,y$ линейно зависимы,
- $r_{xy}=0 \Rightarrow x,y$ линейно независимы.

Статистическая проверка наличия корреляции

Гипотеза: H_0: отсутствует линейная связь между выборками x и y ($r_{xy}=0$).

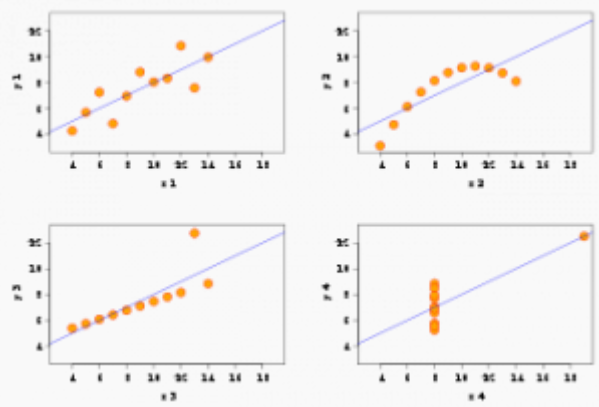
Статистика критерия:

$T=\frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}} \sim t_{n-2}$ – распределение Стьюдента с $n-2$ степенями свободы.

Критерий:

$T \in [t_\alpha, t_{1-\alpha}]$, где t_α есть α-квантиль распределения Стьюдента.

Слабые стороны



Четыре различных набора данных, коэффициент корреляции на которых равен 0.81

- Неустойчивость к выбросам.
- С помощью коэффициента корреляции Пирсона можно определить силу линейной зависимости между величинами, другие виды взаимосвязей выявляются методами регрессионного анализа.
- Необходимо понимать различие понятий "независимость" и "некоррелированность". Из первого следует второе, но не наоборот.

Для того, чтобы выяснить отношение между двумя переменными, часто необходимо избавиться от влияния третьей переменной. Рассмотрим пример 3-х переменных x,y,z. Исключим влияние переменной z:

$r_{xy\backslash z}=\frac{r_{xy}-r_{xz}r_{yz}}{\sqrt{(1-r_{xz})^2(1-r_{yz})^2}}$ – частный коэффициент корреляции.

Для исключения влияния большего числа переменных:

$$r_{ij\backslash vars} = \frac{-R_{ij}}{\sqrt{R_{ii}R_{jj}}},$$
$$R_{ij} = (-1)^{i+j} M_{ij},$$

где M_{ij} – главный минор матрицы коэффициентов корреляции переменных

Ранговая корреляция

Силу связи случайных величин можно оценивать, сравнивая не только численные значения этих случайных величин, но и соответствующие им ранги.

Заданы две выборки $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$, измеренные в ранговых шкалах. Примером выборки, измеренной в ранговых шкалах, могут служить экспертные оценки: эксперт проставляет оценки от 1 до 5 просмотренным n фильмам.

Выборкам x и y соответствуют последовательности рангов:

$R_x = (Rx_1, \ldots, Rx_n)$, где Rx_i — ранг i-го объекта в вариационном ряду выборки x;
$R_y = (Ry_1, \ldots, Ry_n)$, где Ry_i — ранг i-го объекта в вариационном ряду выборки y.

Корреляция последовательностей рангов R_x и R_y называется **ранговой корреляцией**.

Коэффициент корреляции Спирмена

(Spearman rank correlation coefficient) — мера линейной связи между случайными величинами. Корреляция Спирмена 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.

Определение

Заданы две выборки $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$.

Вычисление корреляции Спирмена:

Коэффициент корреляции Спирмена вычисляется по формуле:

$\rho = 1 - \frac{6}{n(n-1)(n+1)} \sum_{i=1}^{n} (R_i - S_i)^2$,[1] где R_i - ранг наблюдения x_i в ряду x, S_i - ранг наблюдения y_i в ряду y.

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

Случай совпадающих наблюдений:

При наличии связок коэффициент корреляции Спирмена следует вычислять следующим образом:

$$\rho = \frac{\sum_{i=1}^{n}(R_i-(n+1)/2)(S_i-(n+1)/2)}{n(n-1)(n+1)-\Delta},$$ [1]

где $$\Delta = \frac{1}{2}\sum_{l=1}^{q} u_i^x((u_i^x)^2-1)+\frac{1}{2}\sum_{i=1}^{f} u_i^y((u_i^y)^2-1)$$.

Здесь q и f — количество связок в выборках x и y, $u_1^x,\ldots,u_q^x$, $u_1^y,\ldots,u_f^y$ — их размеры. Для элементов связок вычисляется средний ранг.

Обоснование критерия Спирмена:

Статистикой критерия Спирмена служит коэффициент корреляции Пирсона ρ ранговых наборов $(R_1\ldots R_n)$ и $(S_1\ldots S_n)$. Он определяется следующей формулой:

$$\rho = \sum_{i=1}^{n}(R_i-\overline{R})(S_i-\overline{S})\Big/\left[\sum_{i=1}^{n}(R_i-\overline{R})^2\sum_{i=1}^{n}(S_i-\overline{S})^2\right]^{1/2}.$$ В этой формуле $$\overline{R}=\overline{S}=\frac{1}{n}\sum_{i=1}^{n} i=\frac{n+1}{2}$$.

Воспользовавшись тем, что $$\sum_{i=1}^{n} i^2=\frac{n(n+1)(2n+1}{6}$$, получим:

$$\sum_{i=1}^{n}(R_i-\overline{R})^2=\sum_{i=1}^{n}(S_i-\overline{S})^2=\sum_{i=1}^{n}\left(i-\frac{n+1}{2}\right)^2=\frac{n(n-1)(n+1)}{12}.$$

Переставив пары $(R_i,\ S_i)$ в порядке возрастания первой компоненты, получим набор $(1,\ T_1)\ldots(n,\ T_n)$. Тогда перепишем коэффициент корреляции Спирмена в виде:

$$\rho = \frac{12}{n(n-1)(n+1)}\sum_{i=1}^{n}\left(i-\frac{n+1}{2}\right)\left(T_i-\frac{n+1}{2}\right).$$

Таким образом, ρ - линейная функция от рангов T_i. Правую часть равенства можно представить в следующем виде:[1]

$$\rho = 1-\frac{6}{n(n-1)(n+1)}\sum_{i=1}^{n}(i-T_i)^2=1-\frac{6}{n(n-1)(n+1)}\sum_{i=1}^{n}(R_i-S_i)^2,$$ который наиболее удобен для вычислений.

Статистическая проверка наличия корреляции

Нулевая гипотеза H_0: Выборки x и y не коррелируют ($\rho = 0$).

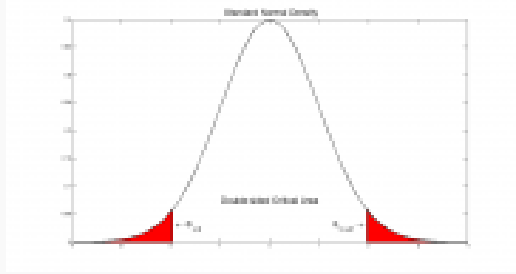
Статистика критерия: ρ.

Критерий (при уровне значимости α):

Против альтернативы H_1: $\rho > 0$:

если ρ больше табличного значения критерия Спирмена p [1] с уровнем значимости $\alpha/2$, то нулевая гипотеза отвергается.

Асимптотический критерий:



Критическая область критерия Спирмена.

Рассмотрим центрированную и нормированную статистику Спирмена:

$\tilde{\rho} = \frac{\rho}{\sqrt{D\rho}}$, где $D\rho = \frac{1}{n-1}$.

Нулевая гипотеза отвергается (против альтернативы H_2 — $|\rho| > 0$), если:

$|\tilde{\rho}| \geq \Phi_{1-\alpha/2}$, [1][1] где $\Phi_{1-\alpha}$ есть $(1-\alpha)$-квантиль стандартного нормального распределения.

Аппроксимация удовлетворительно работает, начиная с $n \geq 50$. [1]

Поправка:

В 1978 году Р. Иман и У. Коновер предложили следующую поправку, значительно повышающую точность аппроксимации. Она использует линейную комбинацию нормальной и стьюдентовской квантилей. Положим:

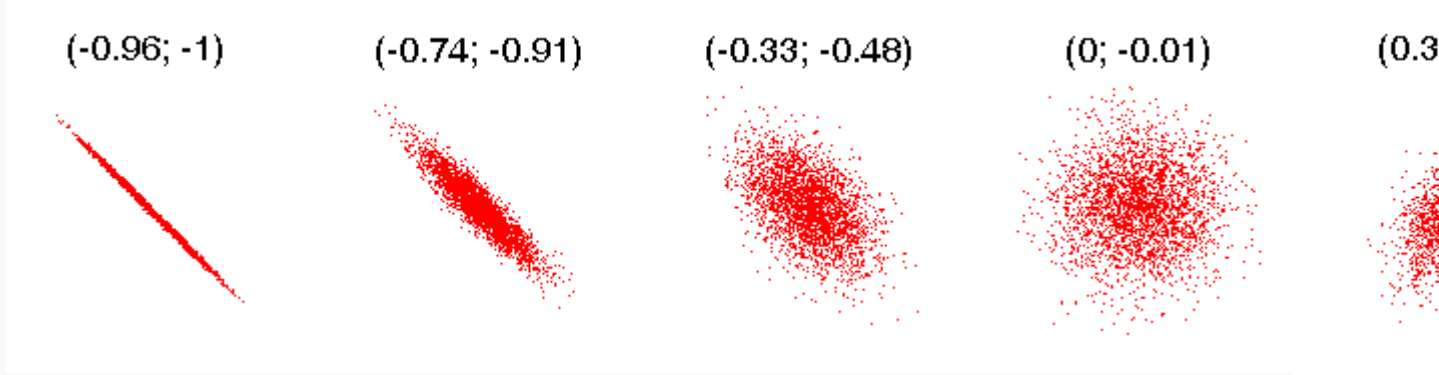
$$\tilde{\rho}^* = \frac{1}{2}\tilde{\rho}\left[\sqrt{n-1} + \sqrt{\frac{n-2}{1-(\tilde{\rho})^2}}\right].$$

Гипотеза H_0 отвергается в пользу альтернативы H_1 $(\rho > 0)$, если $\tilde{\rho}^* \geq (x_{1-\alpha} + y_{1-\alpha})/2$, где $x_{1-\alpha}$, $y_{1-\alpha}$ обозначают соответственно квантили уровня $(1-\alpha)$ стандартного нормального распределения и распределения Стьюдента с $n-2$ степенями свободы.

Примеры

Ниже приведены примеры вычисления корреляций Кенделла и Спирмена. Значения коэффициентов указаны над каждым изображением в виде (τ, ρ), где τ - корреляция Кенделла, ρ - Спирмена. Заметно, что в большинстве случаев $|\rho| > |\tau|$. Объяснение этого эффекта приводится ниже.

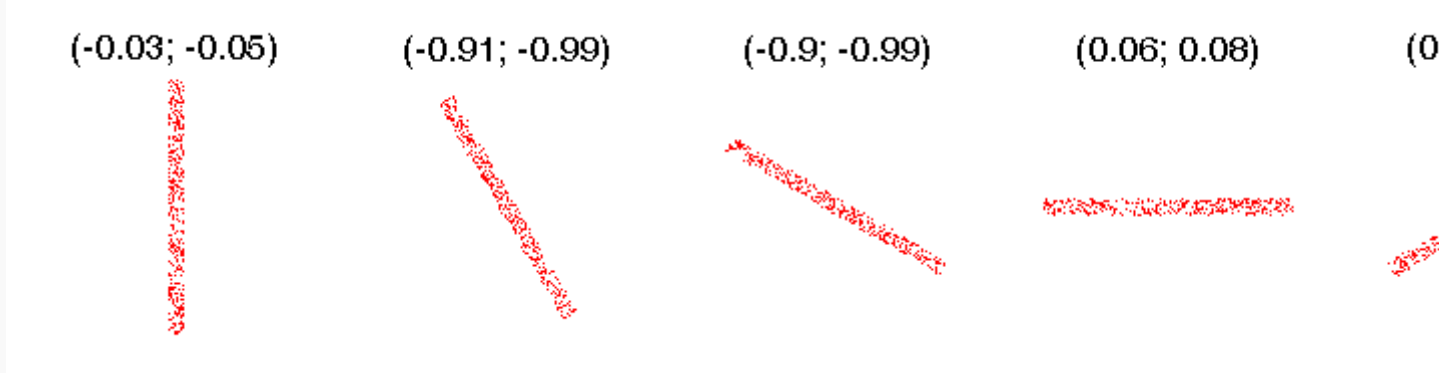
Направление линейной зависимости



Корреляции Кенделла и Спирмена. Нормальные сгущения.

Коэффициенты корреляции реагируют на изменение направления и зашумлённость линейной зависимости между переменными.

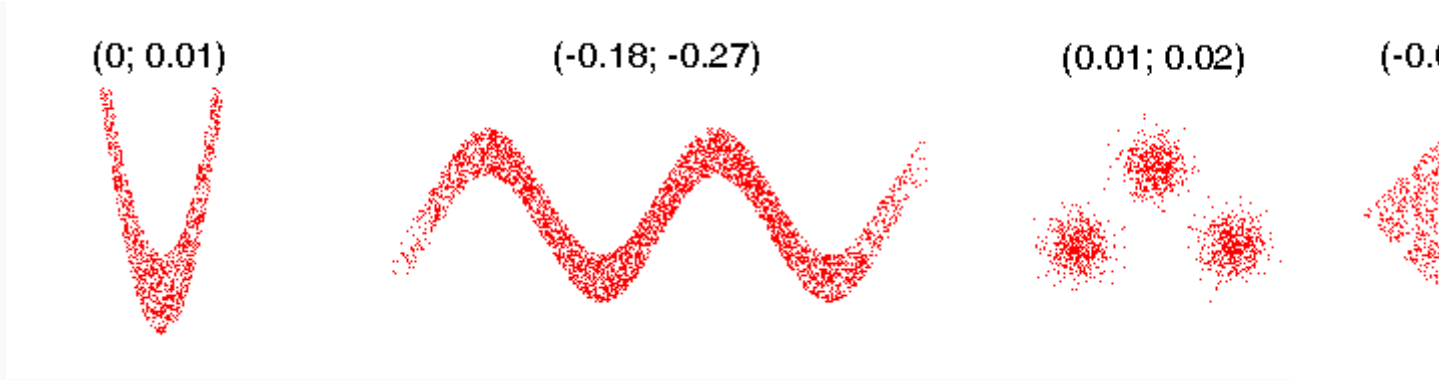
Наклон линейного тренда



Корреляции Кенделла и Спирмена. Вращающаяся полоса.

Коэффициенты корреляции реагируют на изменение направления, но не реагируют на изменение наклона тренда. На первом, четвёртом и седьмом рисунках дисперсия одной из переменных близка к нулю, поэтому не удаётся зафиксировать факт линейной зависимости.

Нелинейная зависимость

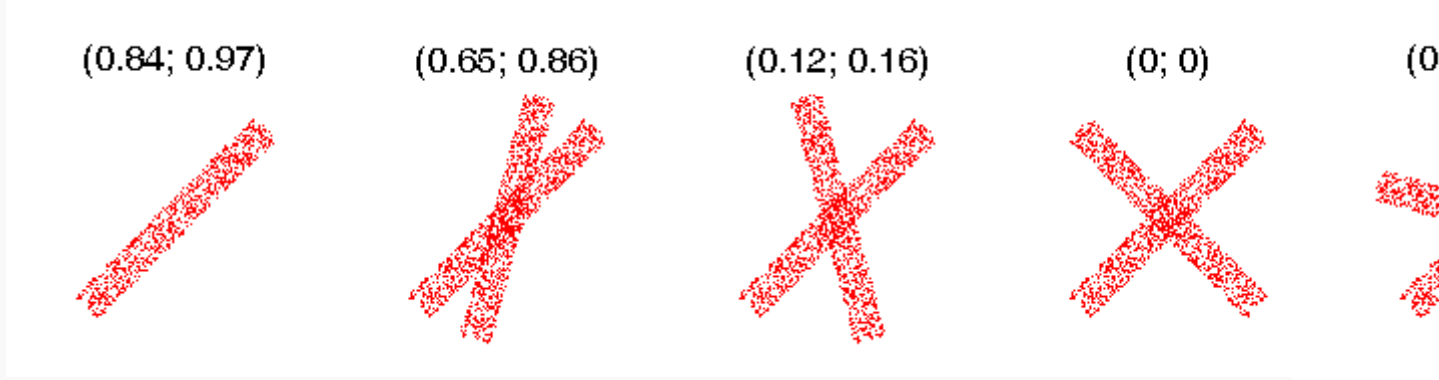


Корреляции Кенделла и Спирмена. Нелинейная зависимость.

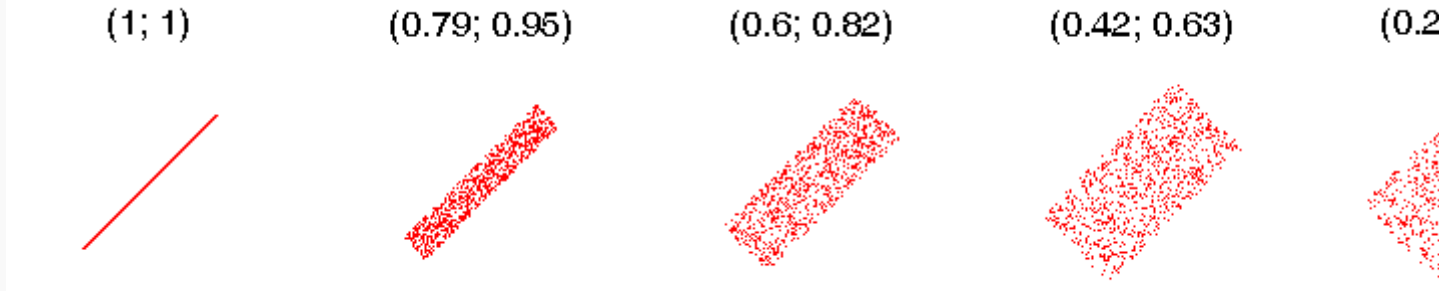
Корреляции Кенделла и Спирмена не отражают меры нелинейной зависимости между переменными.

Линейная и нелинейная зависимости

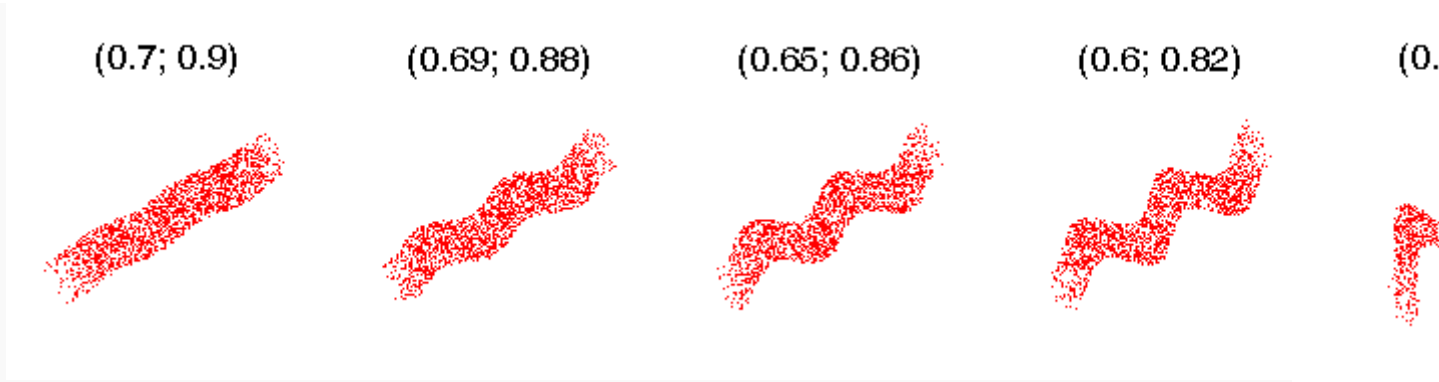
На каждой из приведённых ниже иллюстраций осуществляется переход от линейной зависимости к нелинейной. Коэффициенты корреляции Кенделла и Спирмена реагируют на это одинаковым образом.



Корреляции Кенделла и Спирмена. Перекрещенные полосы.



Корреляции Кенделла и Спирмена. Расширяющаяся полоса.



Корреляции Кенделла и Спирмена. Синусоида с переменной амплитудой.

По мере смены линейной зависимости нелинейной значения коэффициентов корреляции падают.

Связь коэффициентов корреляции Спирмена и Пирсона

В случае выборок из нормального распределения коэффициент корреляции Спирмена ρ может быть использован для оценки коэффициента корреляции Пирсона r по формуле:

$r = 2\sin\frac{\pi}{6}\rho$.[1]

Связь коэффициентов корреляции Спирмена и Кенделла

Выборкам x и y соответствуют последовательности рангов:

$Rx = (Rx_1, \ldots, Rx_n)$, где Rx_i — ранг i-го объекта в вариационном ряду выборки x;

$Ry = (Ry_1, \ldots, Ry_n)$, где Ry_i — ранг i-го объекта в вариационном ряду выборки y.

Проведем операцию упорядочивания рангов.

Расположим ряд значений x_i в порядке возрастания величины: $x_1 \leq x_2 \leq \cdots \leq x_n$. Тогда последовательность рангов упорядоченной выборки x будет представлять собой последовательность натуральных чисел $1, 2, \cdots, n$. Значения y, соответствующие значениям x, образуют в этом случае некоторую последовательность рангов $T = (T_1, \cdots, T_n)$:

$(Rx_i,\ Ry_i) \overset{sort}{\rightarrow} (i,\ T_i),\ i = 1, \cdots, n$.

Коэффициент корреляции Кенделла τ и коэффициент корреляции Спирмена ρ выражаются через ранги [illegible] следующим образом:

$$\rho = 1 - \frac{12}{n^3 - n}\sum_{i<j}(j-i)[T_i > T_j];$$

$$\tau = 1 - \frac{4}{n^2 - 1}\sum_{i<j}[T_i > T_j];$$

Заметно, что в случае ρ инверсиям придаются дополнительные веса $(j-i)$, таким образом ρ сильнее реагирует на несогласие ранжировок, чем τ. Этот эффект проявляется в приведённых выше примерах: в большинстве из них $|\rho| > |\tau|$.

Утверждение.[1] Если выборки x и y не коррелируют (выполняется гипотеза H_0), то величины ρ и τ сильно закоррелированы. Коэффициент корреляции между ними можно вычислить по формуле:

$$corr(\rho, \tau) = \frac{2n+2}{\sqrt{4n^2+10n}}.$$

История