Seminar 12: Коэффициенты корреляции

Carlos Buitrago

Table of contents

1 1. Коэффициент корреляции Пирсона 1

2 2. Коэффициент корреляции Спирмена 3

3 3. Коэффициент корреляции Кендалла 4

4 4. Сравнение трёх коэффициентов 6

5 5. Дополнение: корреляция и двумерное нормальное распределение 6

6 7. Практическая памятка 7

> **Связанные выборки**
>
> Для каждого объекта наблюдается **пара** (X_i, Y_i); пары (X_i, Y_i) независимы и одинаково распределены по i. Проверяем
>
> $$H_0: \quad F_{X,Y}(x,y) = F_X(x)F_Y(y) \quad \forall x, y.$$
>
> Нулевая корреляция сама по себе **не эквивалентна** независимости. Для двумерного нормального распределения нулевая корреляция эквивалентна независимости.

1 1. Коэффициент корреляции Пирсона

> **Определение**
>
> $$\hat{\rho} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}.$$
>
> При существовании конечных вторых моментов и положительных дисперсиях $\hat{\rho} \xrightarrow{P} \rho(X_1, Y_1)$.

Геометрия: $\hat{\rho}$ — косинус угла между центрированными векторами наблюдений. Следовательно, $-1 \leq \hat{\rho} \leq 1$.

> **Точный критерий при нормальности**
>
> Если пары (X_i, Y_i) независимы и одинаково распределены, совместно гауссовские, и при $H_0 : \rho = 0$, то
>
> $$T = \frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}} \sim t_{n-2}, \qquad n > 2.$$

Для двусторонней альтернативы отклоняем H_0, когда $|T| > t_{n-2,1-\alpha/2}$.
Эквивалентно, $|\hat{\rho}| > r_{\text{crit}}$, где

$$r_{\text{crit}} = \frac{t_{n-2,1-\alpha/2}}{\sqrt{t^2_{n-2,1-\alpha/2} + n - 2}}.$$

Важно: этот критерий проверяет нулевую корреляцию; как тест независимости он корректен при совместной нормальности.






```
Эмпирическая доля отклонений при α=0.05: 0.04957142857142857
```

2 2. Коэффициент корреляции Спирмена

Определение

Пусть R_i и S_i — ранги наблюдений X_i и Y_i. Коэффициент Спирмена — корреляция Пирсона **между рангами**:

$$\rho_S = \frac{\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \overline{R})^2 \sum_{i=1}^{n}(S_i - \overline{S})^2}}.$$

При отсутствии совпадающих значений (связок рангов):

$$\boxed{\rho_S = 1 - \frac{6\sum_{i=1}^{n}(R_i - S_i)^2}{n^3 - n}.}$$

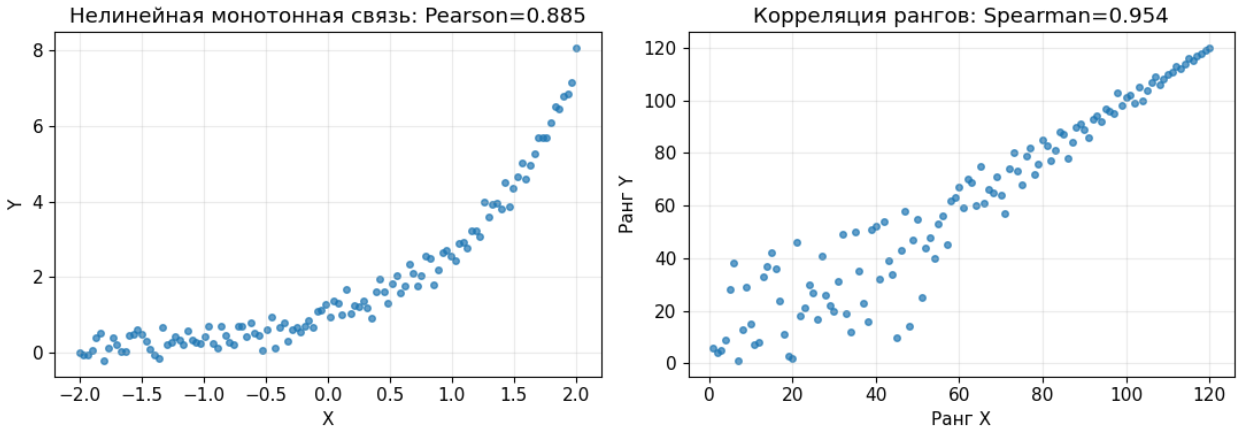
Моменты и асимптотическая нормальность

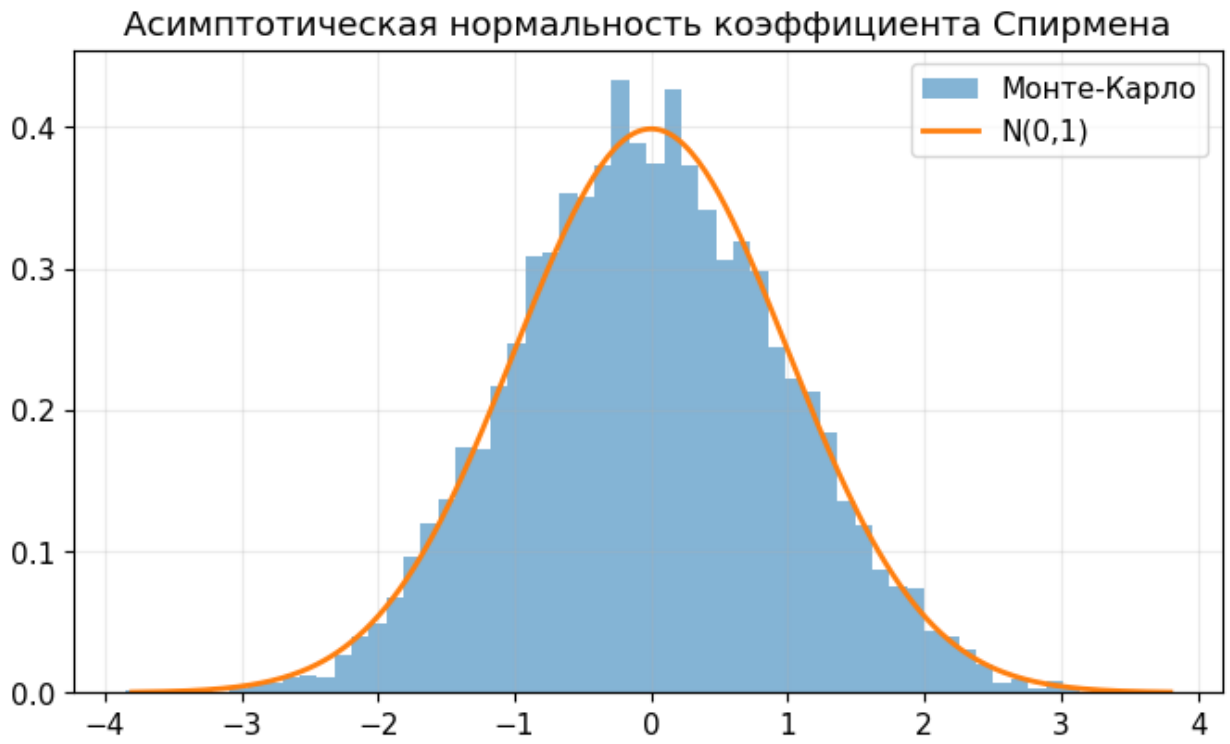
При независимости компонент пары и **непрерывных** распределениях (поэтому связок нет):

$$\mathbb{E}\rho_S = 0, \qquad \mathrm{Var}(\rho_S) = \frac{1}{n-1}, \qquad \sqrt{n-1}\,\rho_S \xrightarrow{d} \mathcal{N}(0,1).$$

Кроме того, $-1 \le \rho_S \le 1$; границы достижимы.

Практика: Спирмен выявляет монотонные, не обязательно линейные связи. При совпадениях рангов используют усреднённые ранги; простая формула через $\sum(R_i - S_i)^2$ и приведённая точная дисперсия требуют отсутствия связок.






```
Среднее: -0.0023; дисперсия: 0.0334; теория: 0.0345
```

3 3. Коэффициент корреляции Кендалла

Согласованные и несогласованные пары

Для $i < j$ пара наблюдений согласованна, если $(X_j - X_i)(Y_j - Y_i) > 0$, и несогласованна, если произведение отрицательно. Пусть S — число согласованных пар, R — число несогласованных. Тогда

$$T = S - R = \sum_{i<j} \text{sign}((X_j - X_i)(Y_j - Y_i)).$$

При отсутствии совпадений $S + R = \binom{n}{2}$.
Коэффициент Кендалла:

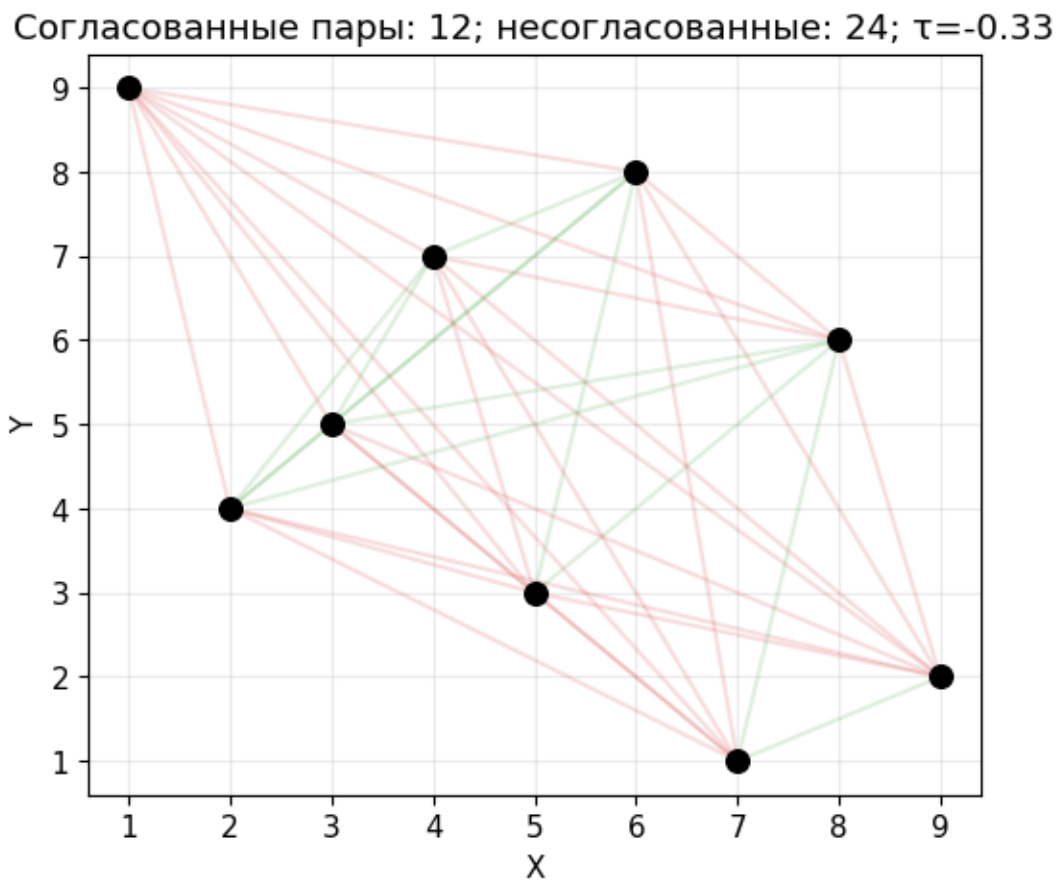
$$\boxed{\tau = \frac{2T}{n(n-1)} = 1 - \frac{4R}{n(n-1)}.}$$

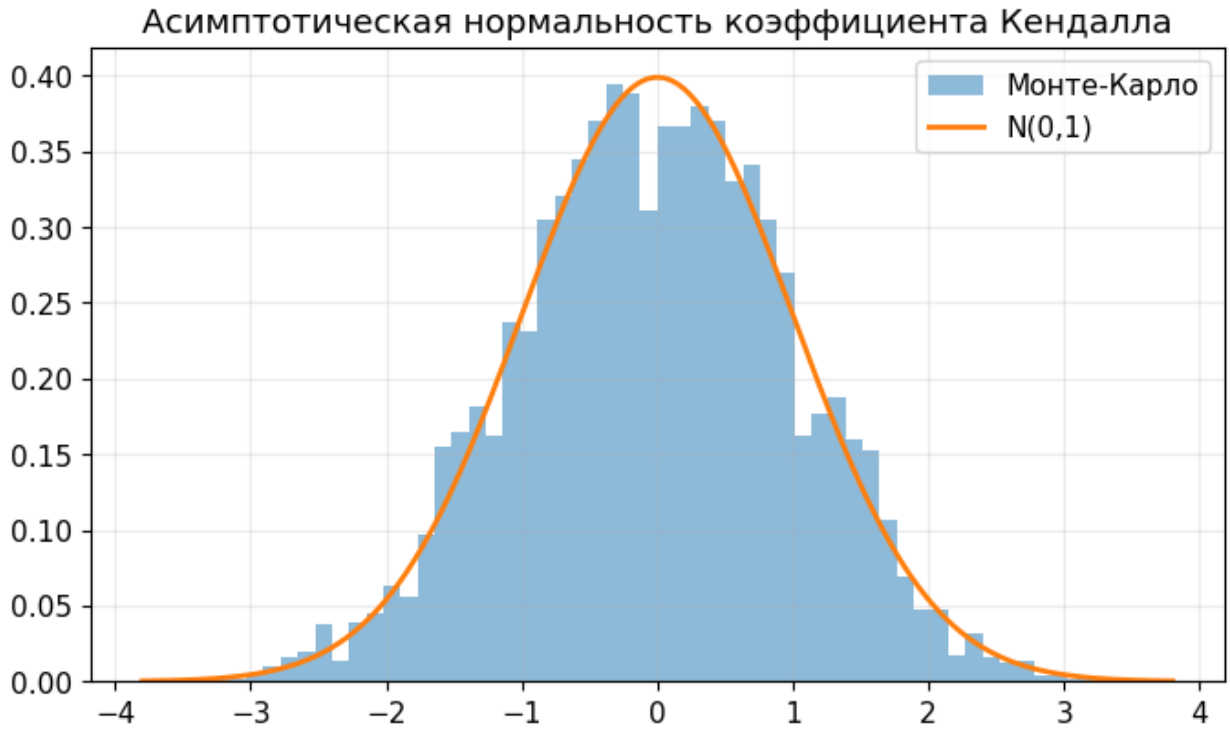
Распределение при независимости

Если X и Y независимы, распределения непрерывны, то

$$\mathbb{E}\tau = 0, \quad \text{Var}(\tau) = \frac{2(2n+5)}{9n(n-1)}, \quad \frac{\tau}{\sqrt{\text{Var}(\tau)}} \xrightarrow{d} \mathcal{N}(0,1).$$

В данных со связками обычно применяют вариант τ_b (`scipy.stats.kendalltau`), у которого нормировка отличается от формулы листка.



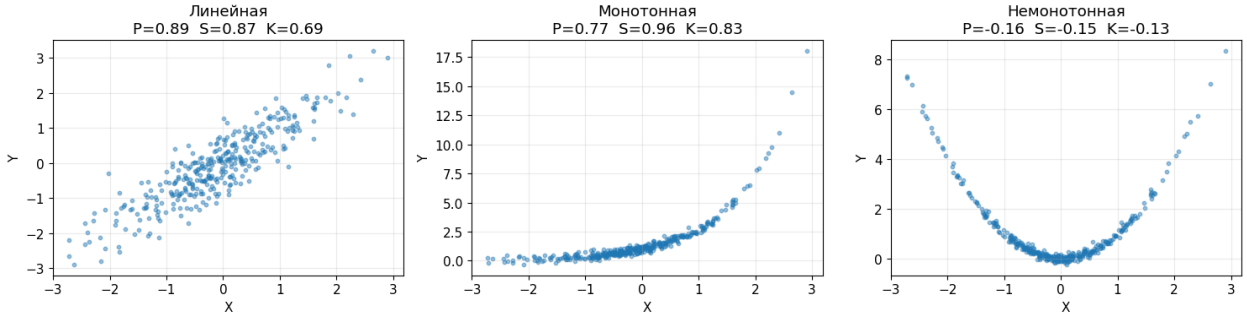


Дисперсия: эмпирическая 0.01584; теоретическая 0.01546

4 4. Сравнение трёх коэффициентов

Коэффициент	Измеряет	Чувствительность	Основное предположение для указанного теста
Пирсона	Линейную связь	К выбросам	Совместная нормальность для точного t-критерия
Спирмена	Монотонную связь рангов	Обычно устойчивее к выбросам	Независимость и непрерывность для указанных моментов
Кендалла	Баланс согласованных и несогласованных пар	Обычно устойчивее к выбросам	Независимость и непрерывность для указанной дисперсии

> ⚠ Независимость и корреляция — разные утверждения
>
> Если $X \sim \mathcal{N}(0,1)$ и $Y = X^2$, то X, Y зависимы, но $\mathrm{Cov}(X,Y) = \mathbb{E}X^3 = 0$. Коэффициенты корреляции могут не обнаружить немонотонную зависимость.



5 5. Дополнение: корреляция и двумерное нормальное распределение

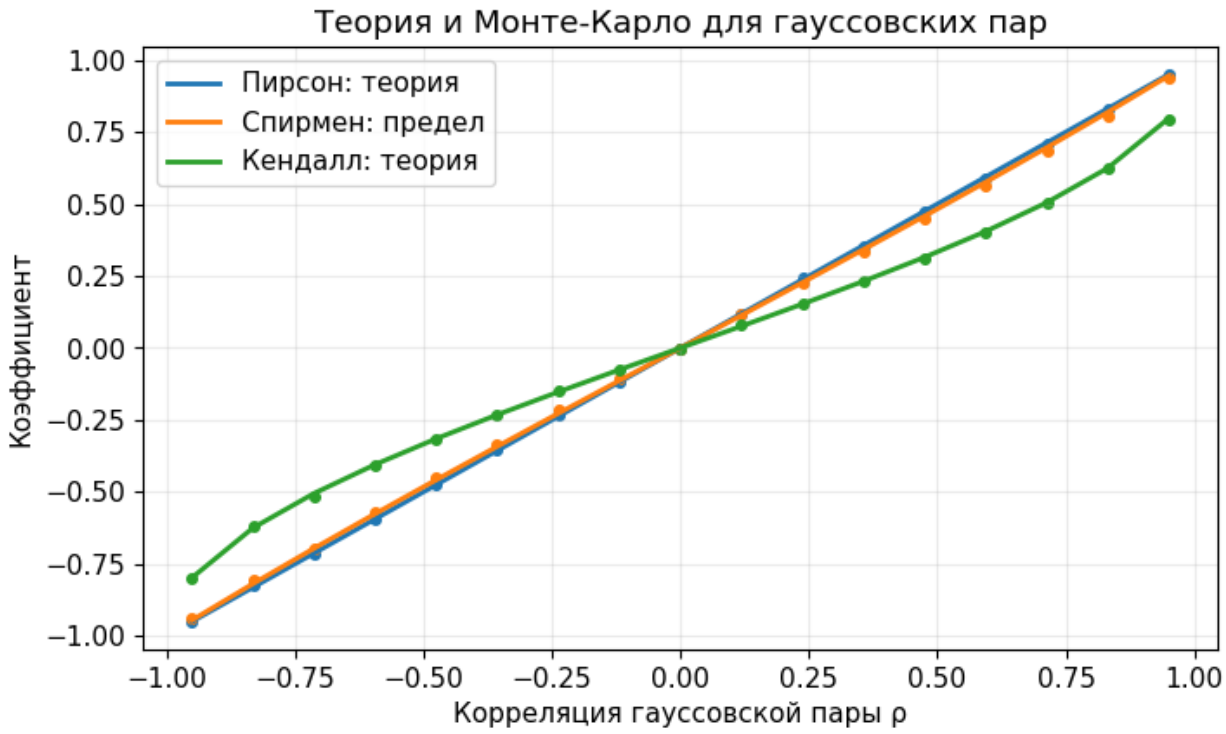
Для двумерного нормального вектора с корреляцией ρ известны популяционные зависимости

$$\tau = \frac{2}{\pi}\arcsin(\rho), \qquad \rho_S = \frac{6}{\pi}\arcsin\left(\frac{\rho}{2}\right).$$

В частности, коэффициент Кендалла является несмещённой оценкой $\frac{2}{\pi}\arcsin\rho$ (без связок). **Для конечного** n коэффициент Спирмена в общем случае не является несмещённой оценкой своего популяционного значения: при гауссовской модели

$$\mathbb{E}\hat{\rho}_S = \frac{6}{\pi(n+1)}\left[(n-2)\arcsin\left(\frac{\rho}{2}\right) + \arcsin(\rho)\right], \quad n \geq 2.$$

Это уточнение полезно для задачи 4(б) из листка.



6 7. Практическая памятка

Выбор критерия

- При совместной нормальности используйте точный t-критерий Пирсона.
- При монотонной связи и отсутствии оснований предполагать нормальность — ранговые методы.
- При связках рангов проверяйте применимость формул; функции `spearmanr` и `kendalltau` учитывают связки, но их p-значения могут использовать приближения.
- Если интересует **любая** зависимость, а не только линейная или монотонная, нужны более общие критерии независимости.