

Введение в статистику



Основные понятия о статистике: медиана, мода, стандартное отклонение, дисперсия. Виды распределений: нормальное, равномерное.

Корреляционный анализ данных. Коэффициенты корреляции Пирсона, Кендалла, Спирмена. Пример матрицы корреляций.

Даниил Корбут

Специалист по Анализу Данных



Даниил Корбут
DL Researcher
Insilico Medicine, Inc

Окончил бакалавриат ФИВТ
МФТИ (Анализ данных) в 2018г
Учусь на 2-м курсе
магистратуры ФИВТ МФТИ
Работал в Statsbot и Яндекс.
Алиса.
Сейчас в Insilico Medicine, Inc,
занимаюсь генерацией
активных молекул и
исследованиями старения с
помощью DL.

Где применяется статистический анализ?

Компьютерное зрение;

Перевод языков;

Генетический анализ данных (молекулярная биология);

Финансовый анализ данных;

Рекомендательные системы;

Моделирование физиологических сигналов;

в любых табличных данных.

Статистика

Рассматривается выборка из случайной величины X :

$$X^n = (X_1, \dots, X_n),$$

где n — объем выборки. Величины X_1, X_2, \dots, X_n — независимые одинаково распределенные случайные величины (*i.i.d.*).

Статистикой $T(X^n)$ называется любая функция от данной выборки.

Основные понятия статистики

Среднее значение;
Медиана;
Мода;
Минимум;
Максимум;
Стандартное отклонение;
Корреляция;
Выбросы.



Среднее

Часто возникает необходимость оценить не всю функцию распределения, а некоторые ее параметры. Самым важным классом параметров распределения являются **средние**. Нестрогое определение можно сформулировать следующим образом: среднее — это значение, вокруг которого группируются все остальные.

Одним из вариантов уточнения данного определения является **матожидание**:

$$EX = \begin{cases} \sum_i a_i p_i, & X — \text{дискретна}, \\ \int_{-\infty}^{+\infty} x f(x) dx, & X — \text{непрерывна}. \end{cases}$$

Квантиль и медиана

Другой характеристикой среднего является медиана. Она определяется с помощью квантиля. **Квантилем** порядка $\alpha \in (0, 1)$ называется величина X_α такая, что:

$$P(X \leq X_\alpha) \geq \alpha, \quad P(X \geq X_\alpha) \geq 1 - \alpha.$$

Медиана — это квантиль порядка 0,5:

$$P(X \leq \text{med } X) \geq 0,5, \quad P(X \geq \text{med } X) \geq 0,5.$$

Медиана

Возьмите ваши наблюдения:

80, 87, 95, 83, 92

Расположите их в
возрастающем порядке:

80, 83, 87, 92, 95

Среднее значение и есть
медиана

↓
80, 83, **87**, 92, 95

Если значений чётное кол-во, то
медианой будет среднее
арифметическое двух средних
значений

89.5
80, 83, **87, 92**, 95, 98

Мода

Еще одной характеристикой среднего является **мода** — самое вероятное значение случайной величины (в нестрогом смысле):

$$\text{mode } X = \begin{cases} a_{\arg\max_i p_i}, & X \text{ — дискретна,} \\ \arg\max_x f(x), & X \text{ — непрерывна.} \end{cases}$$

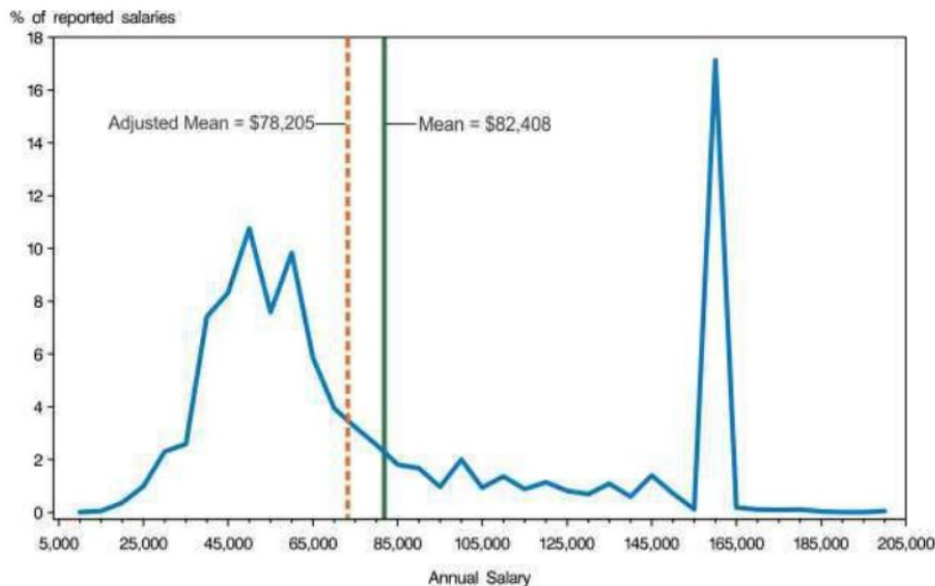
Пример подсчёта

Выборочная мода оценивается по максимуму оценки плотности распределения.

Показателен следующий пример. Рассматривается выборка из 25 человек, для каждого из которых известен годовой доход. В выборке есть десять человек, годовой доход которых равен двум тысячам долларов, один человек с годовым доходом в три тысячи долларов, и так далее. Один человек получает сорока пять тысяч долларов в год. Среднее арифметическое годовых доходов на этой выборке — 5700 долларов. Здесь медиана составляет 3000 долларов, а мода — 2000.

Пример подсчёта

Необходимо заметить, что все рассматриваемые величины называются «средними». Значит, для оптимистичного отчета по данной выборке можно воспользоваться средним арифметическим, а для пессимистичного — модой.



Стандартное отклонение

Мера отклонения значений выборки от среднего

Греческая буква «сигма» используется для обозначения стандартного отклонения

$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$

1. Вычтите каждое наблюдение из среднего значения

2. Возведите каждую разность в квадрат

3. Сложите все разности

4. Разделите сумму на количество наблюдений минус 1

5. Из результата извлеките квадратный корень

The diagram illustrates the five steps to calculate the standard deviation formula. Step 1 points to the subtraction of the mean from each observation in the formula. Step 2 points to the squaring of the differences. Step 3 points to the summation of these squared differences. Step 4 points to the division of the sum by the degrees of freedom (n-1). Step 5 points to the square root operation.

Дисперсия

Квадрат стандартного отклонения. Дисперсия показывает, насколько в среднем значения сосредоточены, сгруппированы около среднего: если дисперсия маленькая - значения сравнительно близки друг к другу, если большая - далеки друг от друга.

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Нахождение математического ожидания и дисперсии

Чему равно математическое ожидание и дисперсия случайной величины?

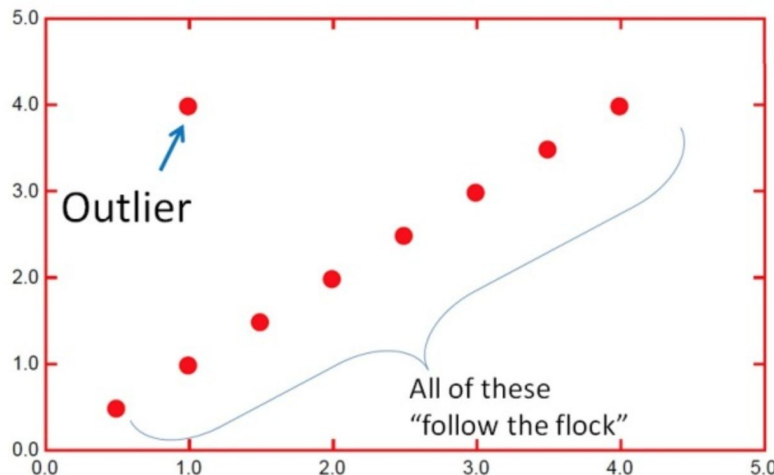
x	2	3	5	6	5	1
---	---	---	---	---	---	---

Математическое ожидание = среднее значение =
 $(2+3+5+6+5+1)/6 = 3.6$

Дисперсия = $1/5 ((2-3.6)^2 + (3-3.6)^2 + (5-3.6)^2 + (6-3.6)^2 + (5-3.6)^2 + (1-3.6)^2) = 4,632$

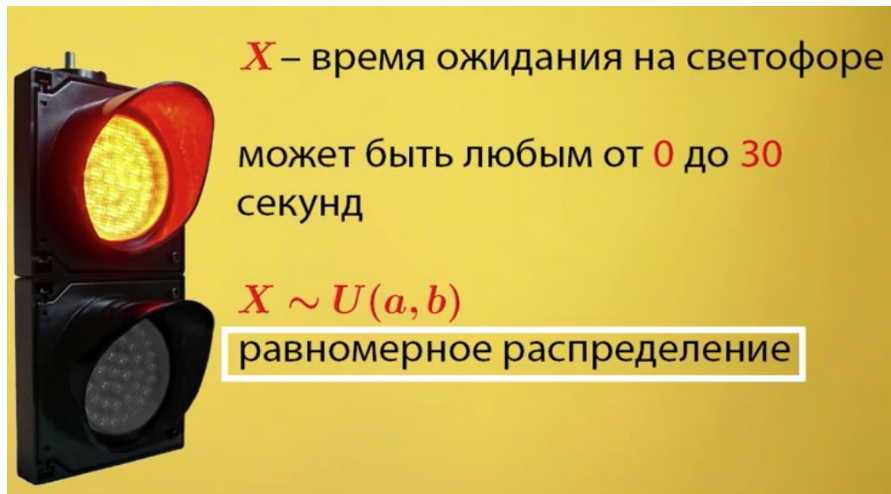
Выбросы

Если в данных есть выбросы — значения, которые имеют слишком большое отклонение от среднего значения, — это может негативно повлиять на анализ.



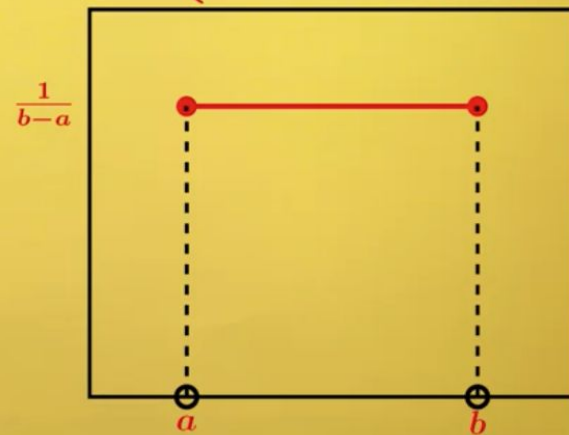
Примеры случайных величин (равномерное распределение)

Ярким примером непрерывной случайной величины, распределённой **равномерно**, является время ожидания перехода дороги со светофором без секунд.



$$X \sim U(a, b)$$

$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a, b] \\ 0, & x \notin [a, b] \end{cases}$$



Примеры случайных величин (нормальное распределение)

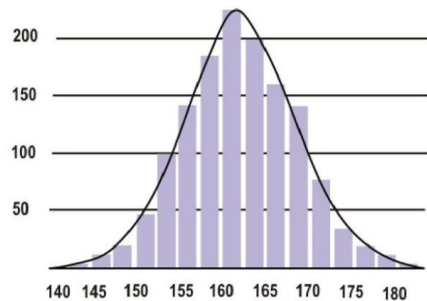
Ярким примером непрерывной случайной величины, распределённой **нормально**, является время прихода на работу, если вы всегда стараетесь приходить в офис, например, около 12:00.

» X – время прихода на работу

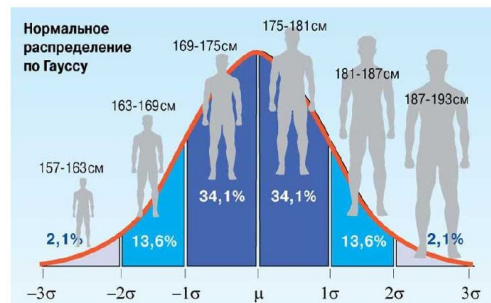
» $X \sim N(\mu, \sigma^2)$

нормальное
(Гауссово)
распределение

Сумма слабо
зависимых
случайных
факторов



Распределение роста

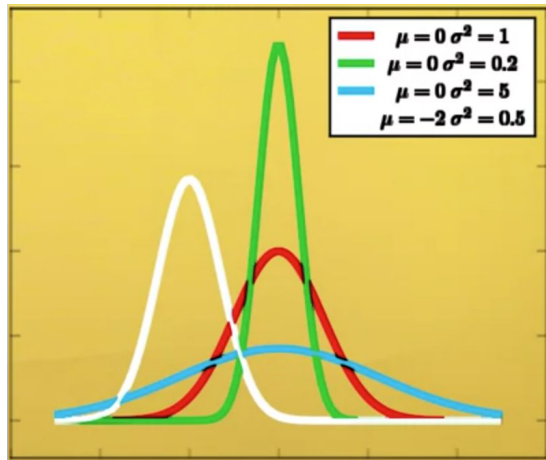


Примеры непрерывных случайных величин (нормальное распределение)

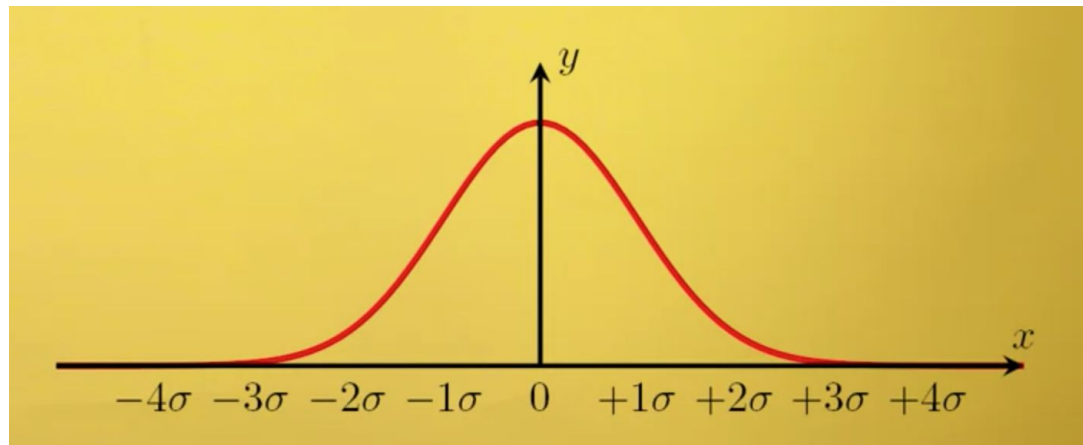
X – время прихода на работу
 $X \sim N(\mu, \sigma^2)$

среднее
время
прихода

разброс
вокруг
среднего

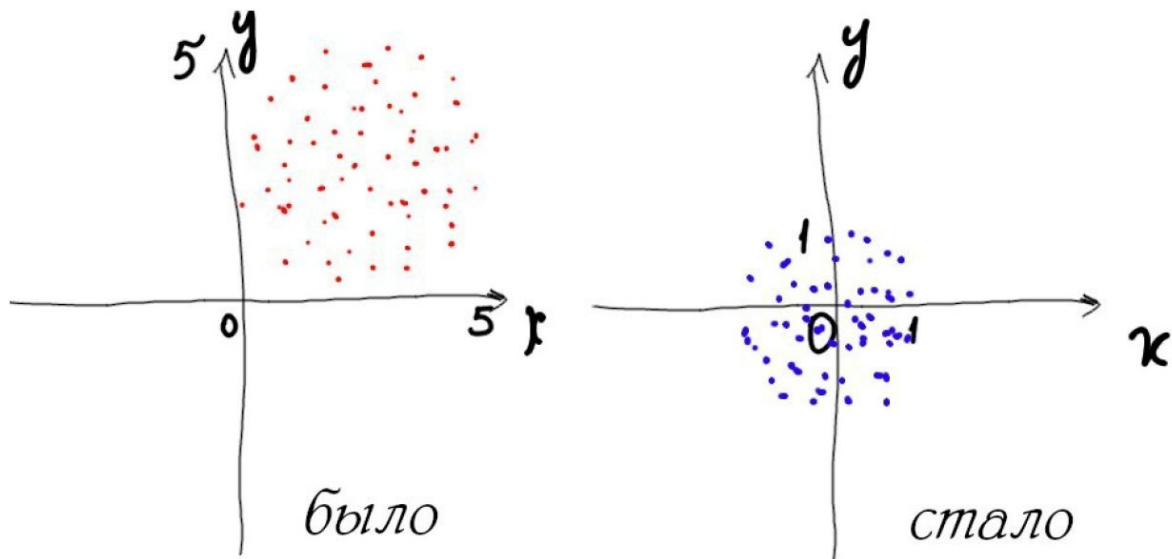


$$X \sim N(\mu, \sigma^2)$$
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Нормализация данных

Часто данные перед анализом необходимо нормализовать.

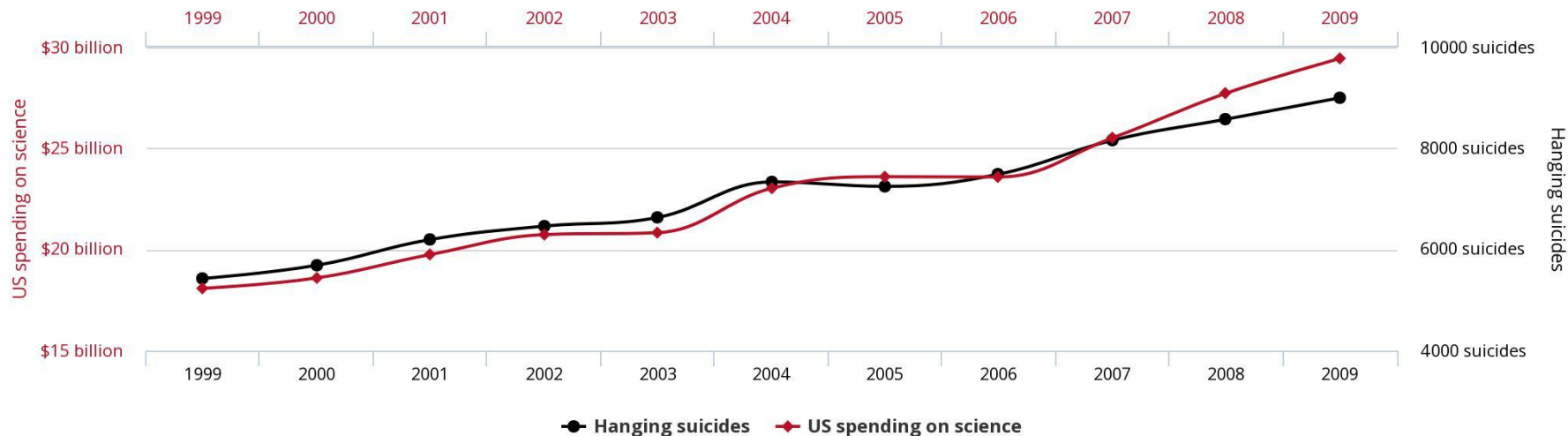


Корреляция

Корреля́ция (от лат. correlatio .соотношение, взаимосвязь.), или корреляционная зависимость, — статистическая взаимосвязь двух или более случайных величин.

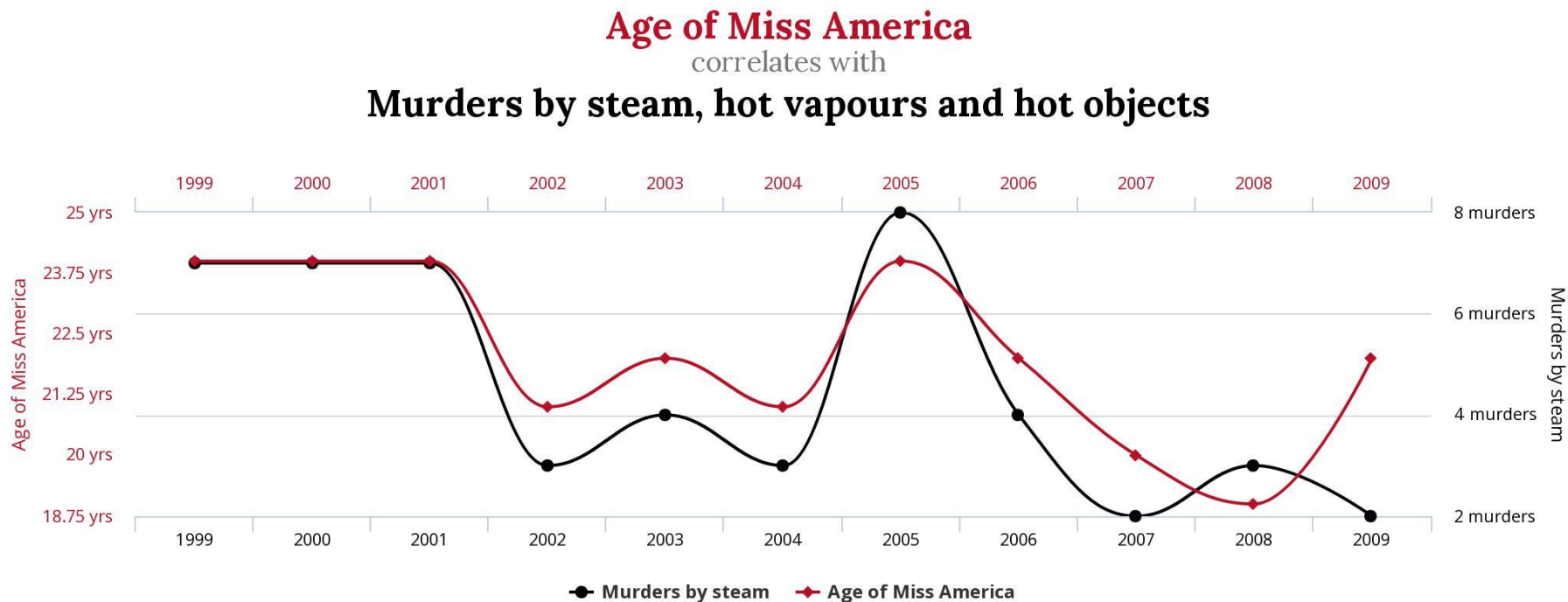
Примеры неожиданной корреляции

US spending on science, space, and technology correlates with Suicides by hanging, strangulation and suffocation



tylervigen.com

Примеры неожиданной корреляции



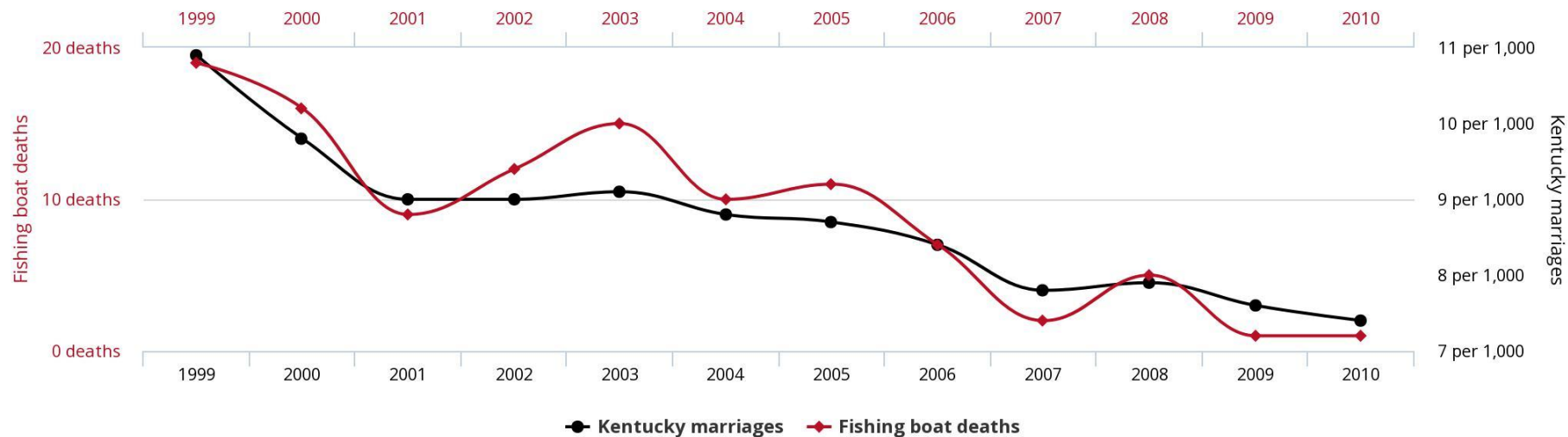
tylervigen.com

Примеры неожиданной корреляции

People who drowned after falling out of a fishing boat

correlates with

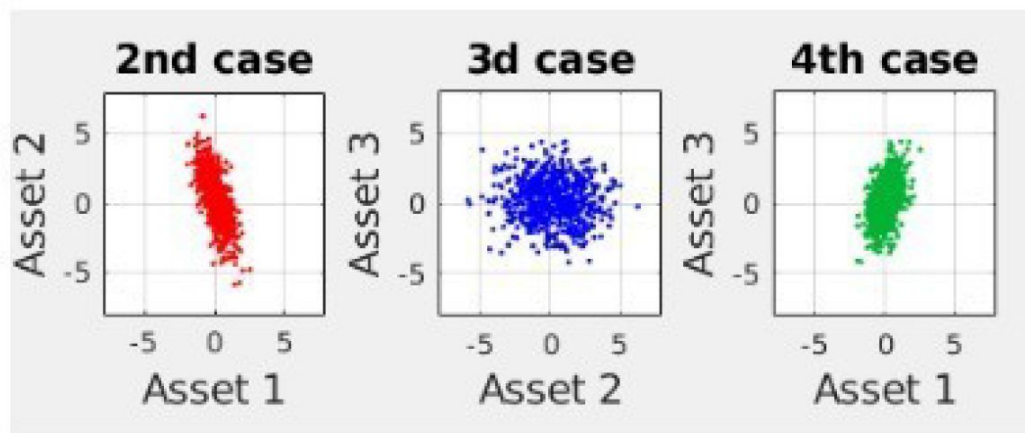
Marriage rate in Kentucky



tylervigen.com

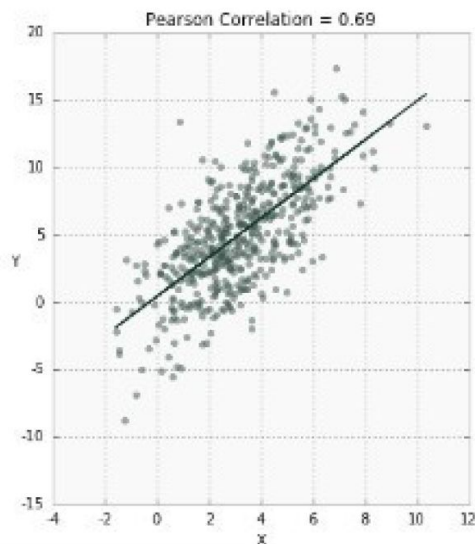
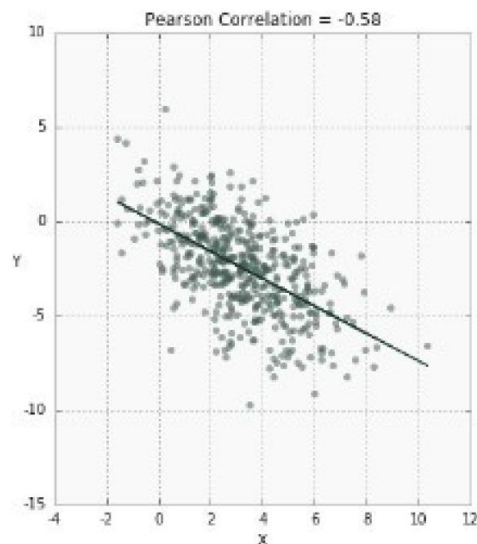
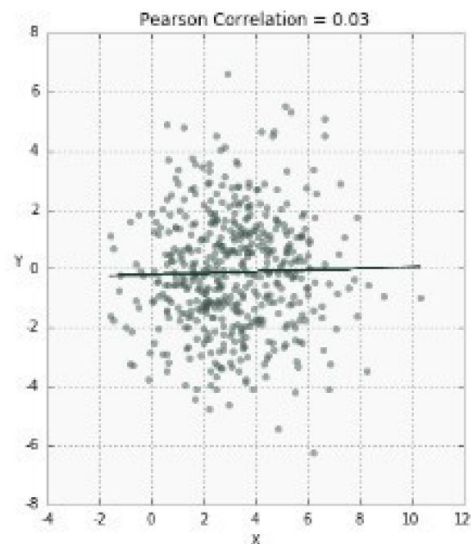
Финансовый анализ данных

Предсказание колебания цены на акции фирмы.
Анализ корреляции необходим для анализа соотношения двух компаний.

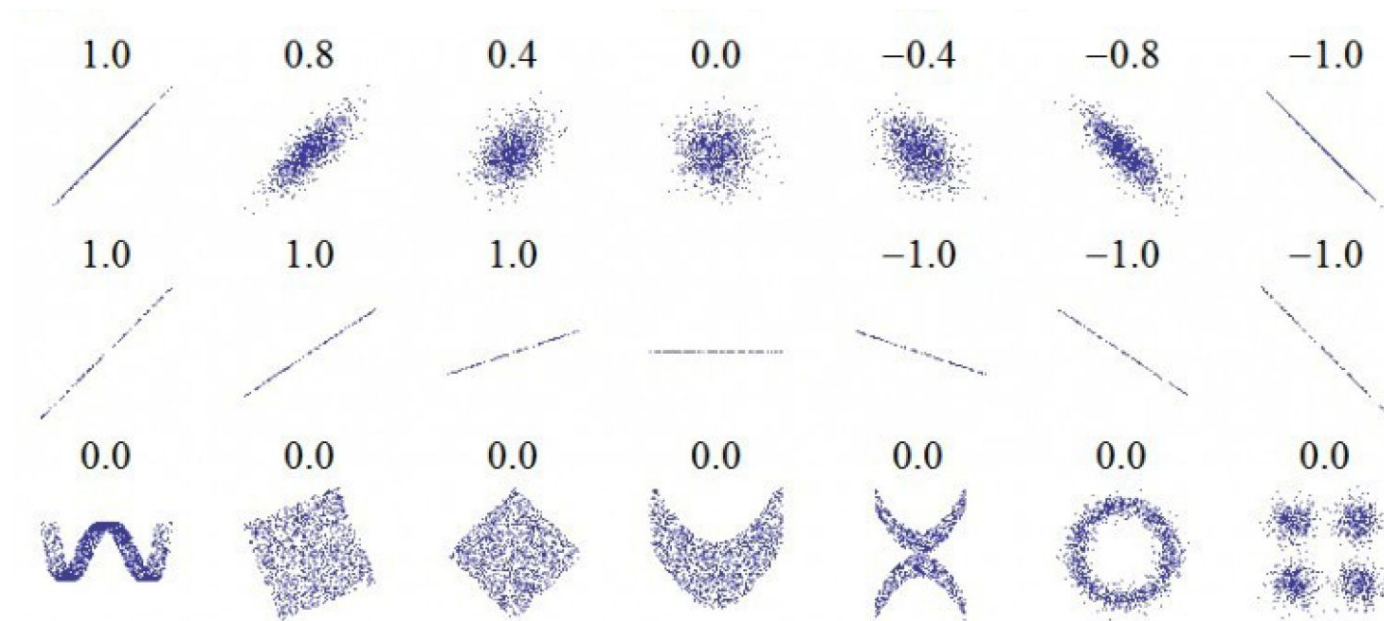


Корреляция Пирсона

$$r_{xy} = \frac{\sum_{i=1}^m (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^m (x_i - \bar{x})^2 \sum_{i=1}^m (y_i - \bar{y})^2}} = \frac{\text{cov}(x, y)}{\sqrt{s_x^2 s_y^2}},$$



Корреляция Пирсона



https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Корреляция Спирмена

Предназначены для определения взаимосвязи между ранговыми Переменными (проверка на нормальность не требуется).

1. Сопоставить каждому из признаков их порядковый номер (ранг) по возрастанию или по убыванию.
2. Определить разности рангов каждой пары сопоставляемых значений (d)
3. Возвести в квадрат каждую разность и суммировать полученные результаты.
4. Вычислить коэффициент корреляции рангов по формуле:

$$\rho = 1 - \frac{6 \cdot \sum d^2}{n(n^2 - 1)}$$

Или использовать библиотеку statistics:
`scipy.stats.spearmanr(x, y)`

<http://medstatistic.ru/theory/spirmen.html>

Корреляция Кендалла

Аналог корреляции Спирмена.

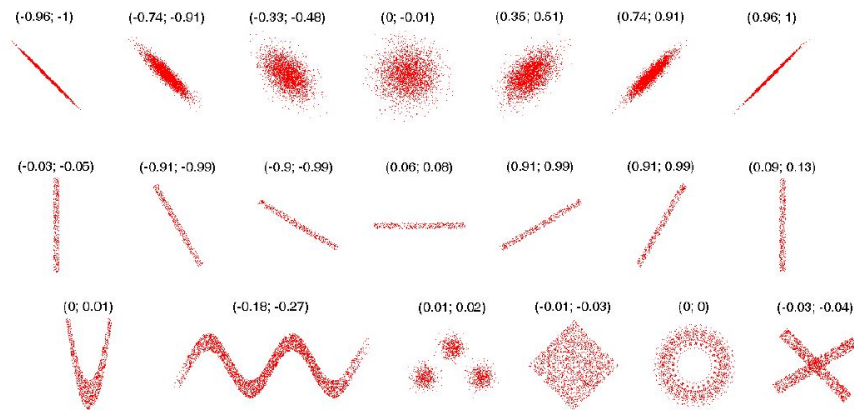
$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

N_c – число совпадений

N_d – число инверсий.

`scipy.stats.kendalltau(x, y)`

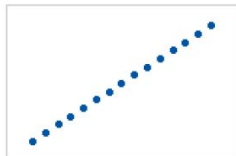
Сравнение коэффициентов Спирмена и Кендалла



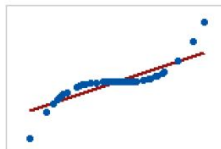
Слева - корреляция Кендалла, справа - корреляция Спирмена

http://www.machinelearning.ru/wiki/index.php?title=Коэффициент_корреляции_Кенделла

Сравнение коэффициентов Пирсона и Спирмена



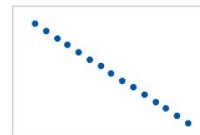
Pearson = +1, Spearman = +1



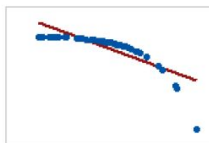
Pearson = +0.851, Spearman = +1



Pearson = -0.093, Spearman = -0.093



Pearson = -1, Spearman = -1



Pearson = -0.799, Spearman = -1

<https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/supporting-topics/basics/a-comparison-of-the-pearson-and-spearman-correlation-methods/>

Матрица корреляций

Для статистического анализа играет наиважнейшую роль.
Строим матрицу корреляций для того, чтобы определить, насколько 2 случайные величины зависят друг от друга.

$$S = \begin{pmatrix} s_x & 0 \\ 0 & s_y \end{pmatrix}$$

s_x — дисперсия переменной x
(среднеквадратичное значение)
 s_y — дисперсия переменной y

Матрица корреляций

Для статистического анализа играет наиважнейшую роль. Строим матрицу корреляций для того, чтобы определить, насколько 2 случайные величины зависят друг от друга.

$$S = \begin{pmatrix} s_x & 0 \\ 0 & s_y \end{pmatrix}$$

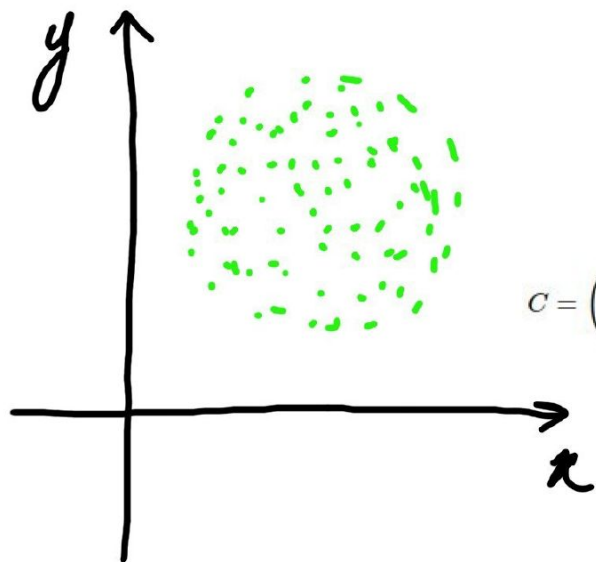
s_x — дисперсия переменной x
(среднеквадратичное значение)
 s_y — дисперсия переменной y

Если 2 случайные величины **зависимы** друг от друга, то матрица корреляций принимает вид:

$$C = \begin{pmatrix} \sigma(x, x) & \sigma(x, y) \\ \sigma(y, x) & \sigma(y, y) \end{pmatrix}$$

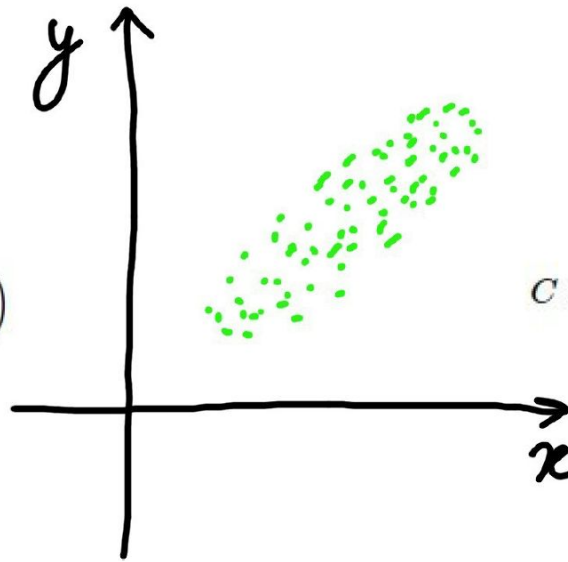
$$\rho_{x, y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Матрица корреляций



$$C = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}$$

Независимые переменные



$$C = \begin{pmatrix} \sigma(x, x) & \sigma(x, y) \\ \sigma(y, x) & \sigma(y, y) \end{pmatrix}$$

Зависимые переменные

Спасибо за внимание!