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.

Юстина Иванова
Специалист по Анализу Данных

Юстина Иванова
студент-аспирант
University of Bolzano

Инженер-программист МГТУ им. Баумана

Master of Science in Artificial Intelligence
University of Southampton

Специалист по компьютерному зрению
в компании Dataplex.

Специалист по анализу данных
в компании ОЦРВ.

НЕТОЛОГИЯ

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:

$$X^n = \big(X_1, \ldots, X_n\big),$$

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

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

Основные понятия статистики

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

Среднее значение случайной величины

Среднее значение случайной величины.
Значение, вокруг которого группируются все остальные.

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

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:

$$P\left(X \leqslant X_\alpha\right) \geqslant \alpha, \quad P\left(X \geqslant X_\alpha\right) \geqslant 1-\alpha.$$

Медиана — это квантиль порядка 0,5:

$$P\left(X \leqslant \operatorname{med} X\right) \geqslant 0{,}5, \quad P\left(X \geqslant \operatorname{med} X\right) \geqslant 0{,}5.$$

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

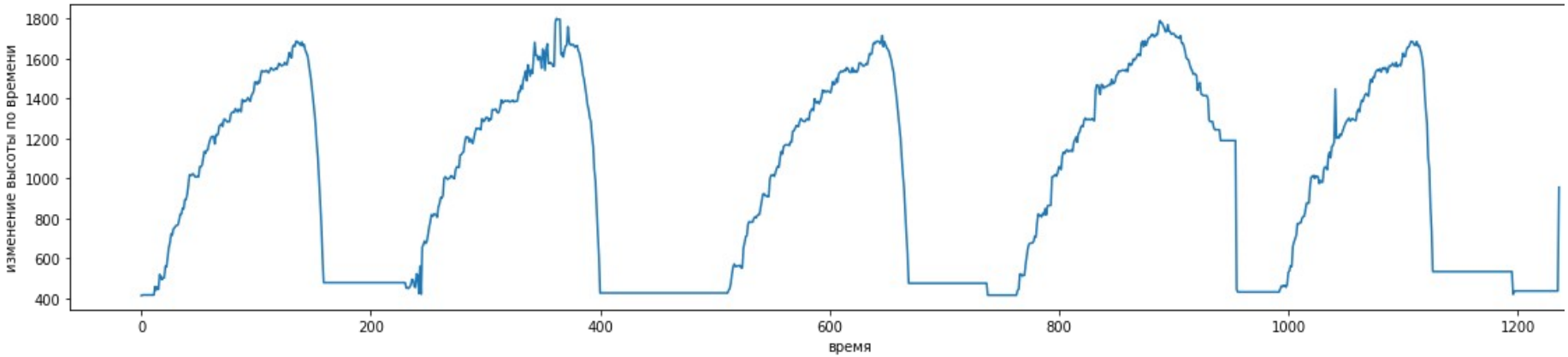
https://tproger.ru/translations/basic-statistics-in-python-descriptive-statistics/

Мода

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

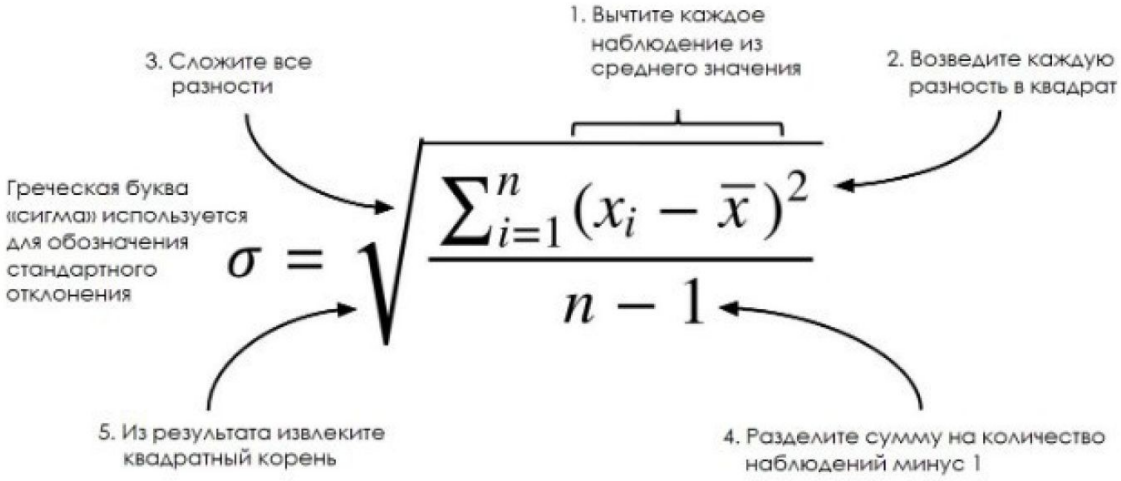
$$\operatorname{mode} X = \begin{cases} a_{\underset{i}{\operatorname{argmax}}\, p_i}, & X \text{ — дискретна,} \\ \underset{x}{\operatorname{argmax}}\, f(x), & X \text{ — непрерывна.} \end{cases}$$

Пример подсчёта



Стандартное отклонение

Мера отклонения значений выборки от среднего



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.

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

Нахождение матожидания и дисперсии

Чему равно математическое ожидание и дисперсия случайной величины?

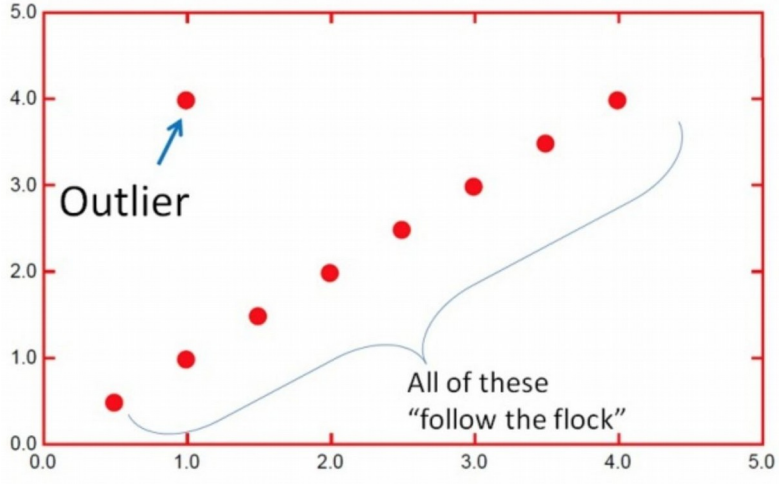
X	2	3	5	6	5	1

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.

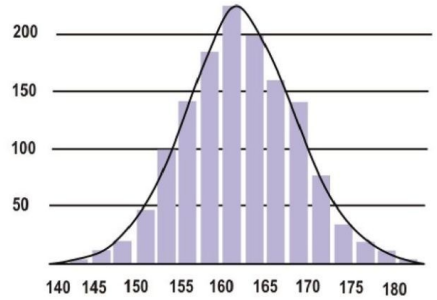


Примеры случайных величин (нормальное распределение)

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

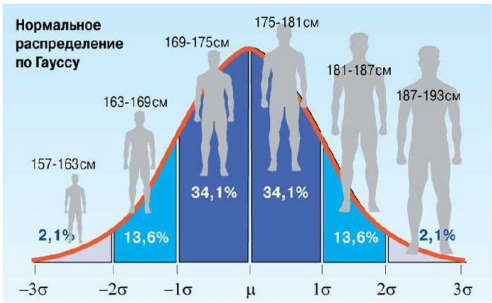
- X – время прихода на работу
- $X \sim N(\mu, \sigma^2)$

нормальное (Гауссово) распределение



Распределение роста

Сумма слабо зависимых случайных факторов



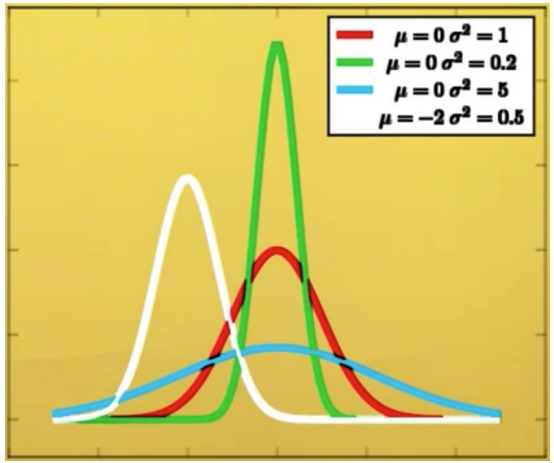
Примеры непрерывных случайных величин (нормальное распределение)

X – время прихода на работу

$X \sim N(\mu, \sigma^2)$

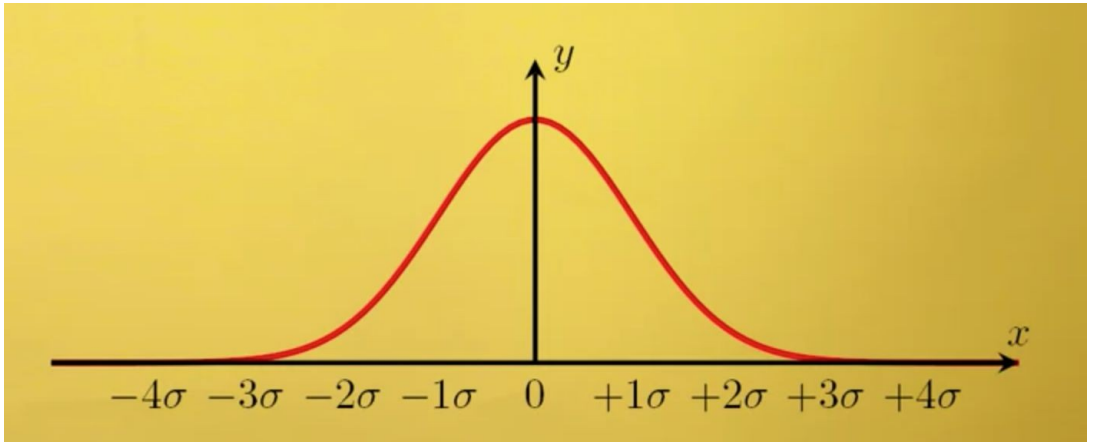
среднее время прихода

разброс вокруг среднего



$$X \sim N(\mu, \sigma^2)$$

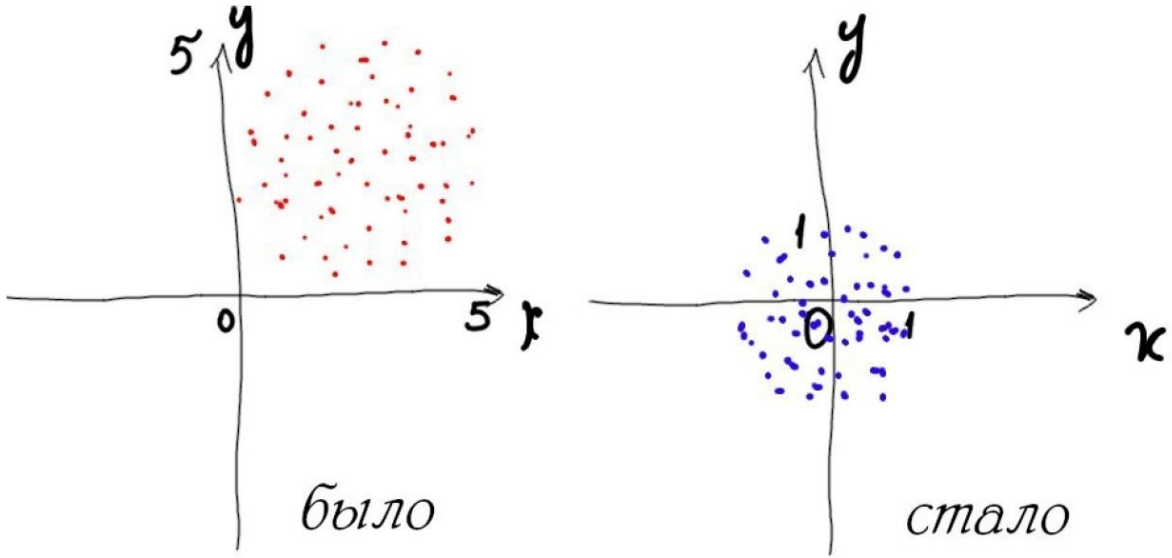
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



НЕТОЛОГИЯ

Нормализация данных

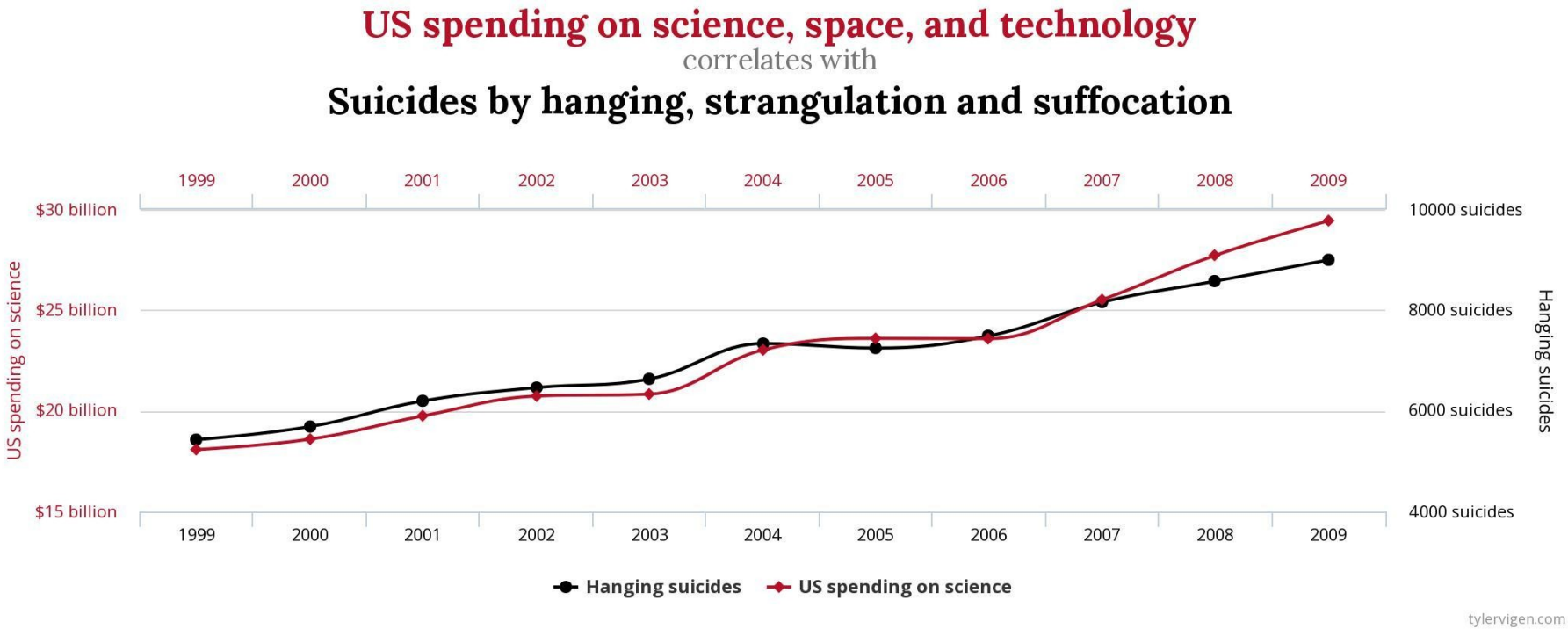
Часто данные перед анализом необходимо нормализовать.



Корреляция

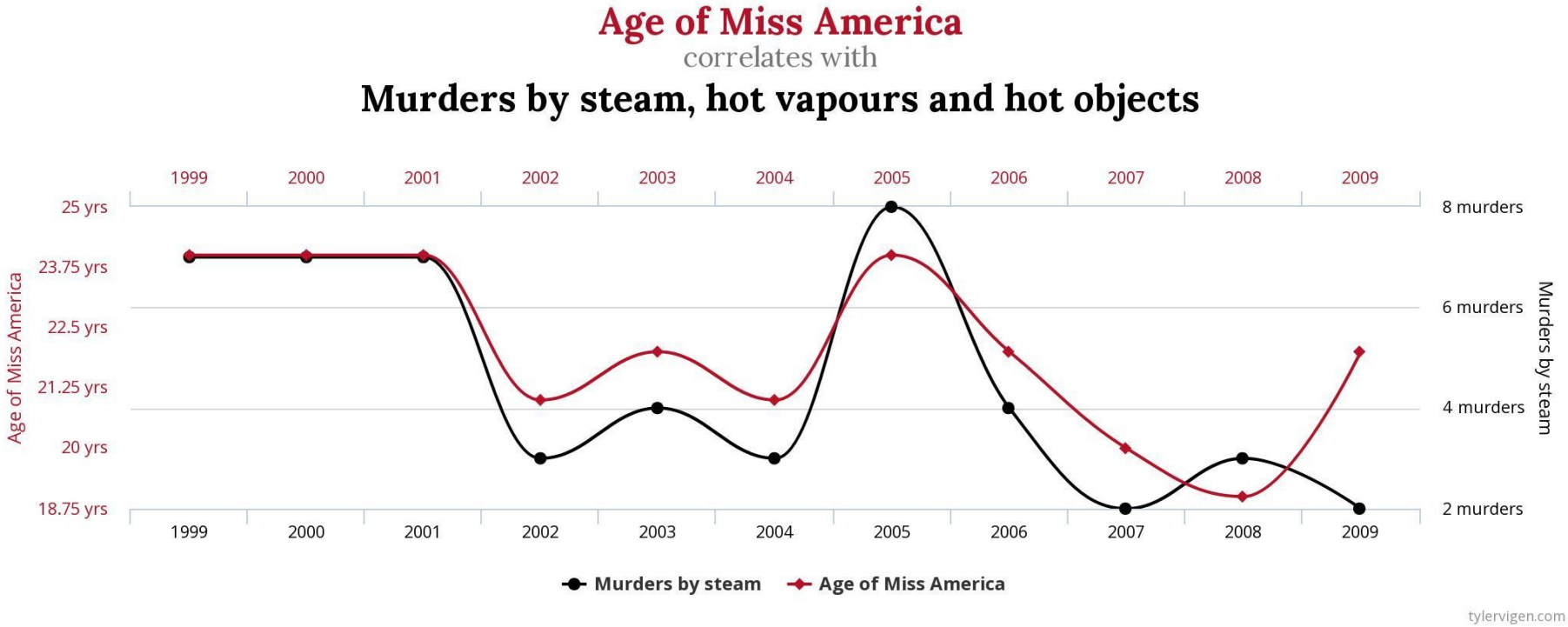
Корреляция (от лат. correlatio «соотношение, взаимосвязь»), корреляционная зависимость, — статистическая взаимосвязь двух или более случайных величин

Примеры неожиданной корреляции



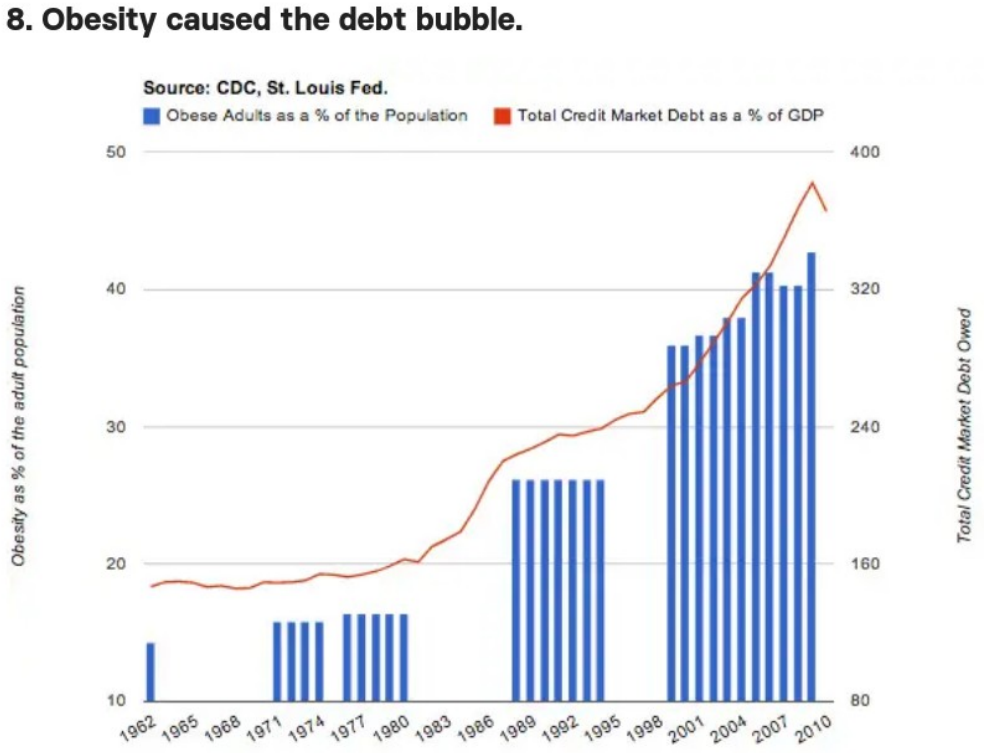
http://www.tylervigen.com/spurious-correlations

Примеры неожиданной корреляции



http://www.tylervigen.com/spurious-correlations

Примеры неожиданной корреляции

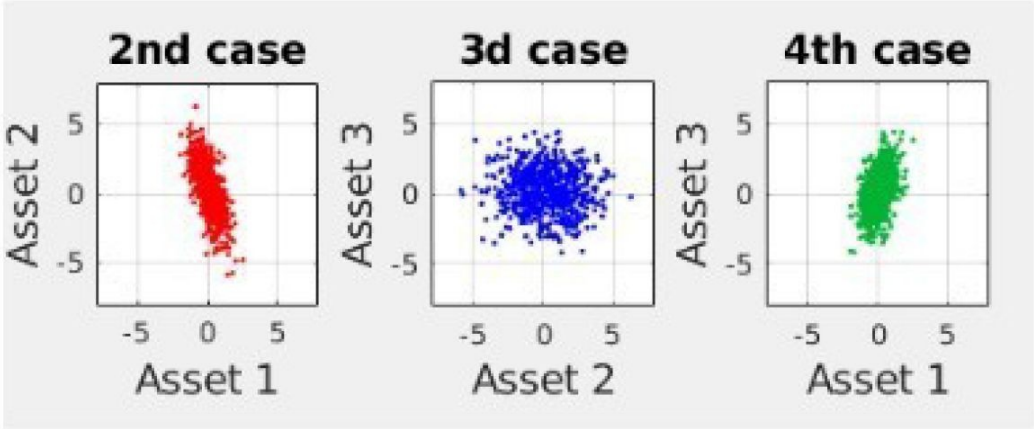


https://www.buzzfeednews.com/article/kjh2110/the-10-most-bizarre-correlations

НЕТОЛОГИЯ

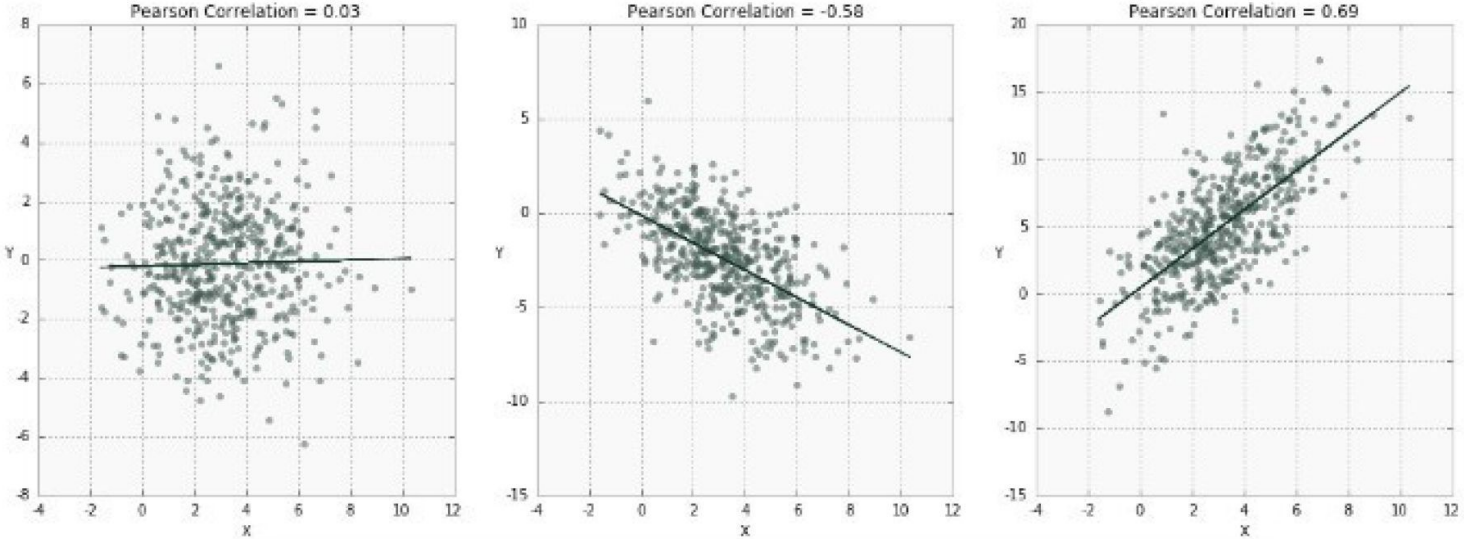
Финансовый анализ данных

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



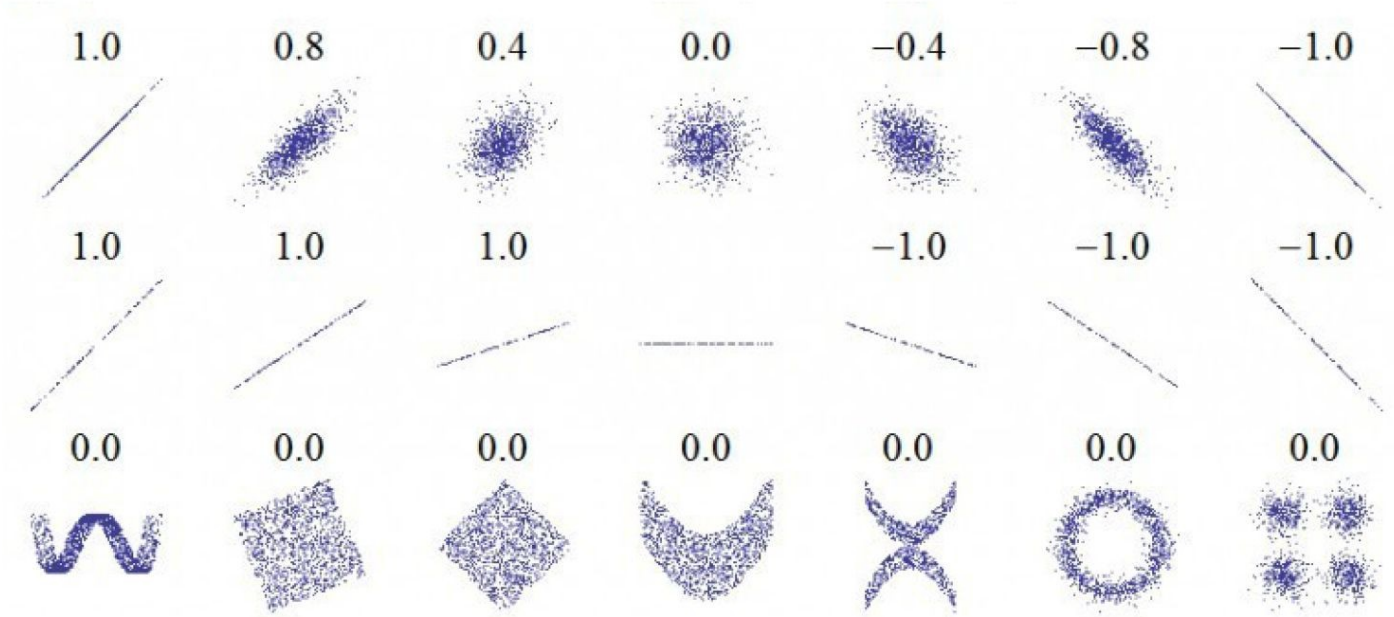
Корреляция Пирсона

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



http://guessthecorrelation.com/

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 значенмй (d)
3. Возвести в квадрат каждую разность и суммировать полученные результаты.
4. Вычислить коэффициент коррялции 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

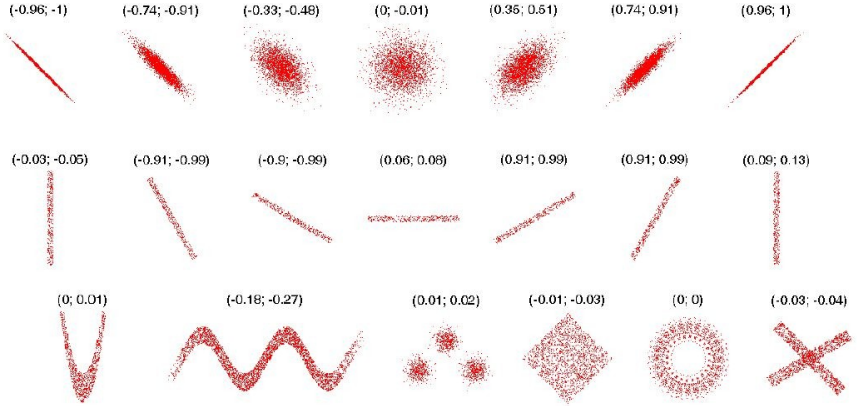
Аналог корреляции Спирмена.

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

Nc – число совпадений
Nd – число инверсий.

scipy.stats.kendalltau(x, y)

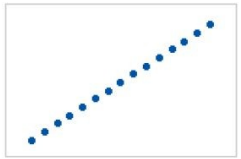
Сравнение коэффициентов Спирмена и Кендалла



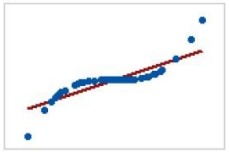
Слева - корреляция Кендалла, справа – корреляция Спирмена

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

Сравнение коэффициентов Пирсона и Спирмена

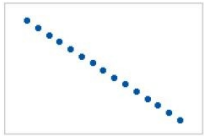


Pearson = +1, Spearman = +1

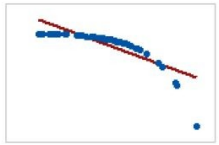


Pearson = +0.851, Spearman = +1

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.

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

Sx — дисперсия переменной x (среднеквадратичное значение)
Sy — 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.

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

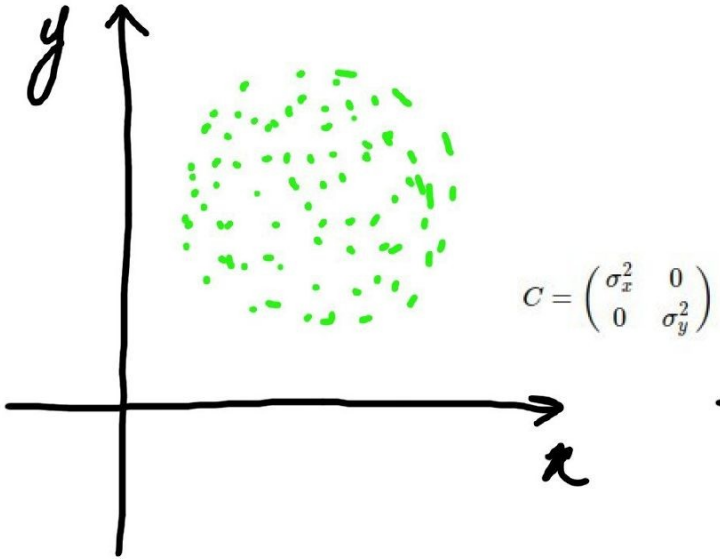
Sx — дисперсия переменной x (среднеквадратичное значение)
Sy — дисперсия переменной y

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

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

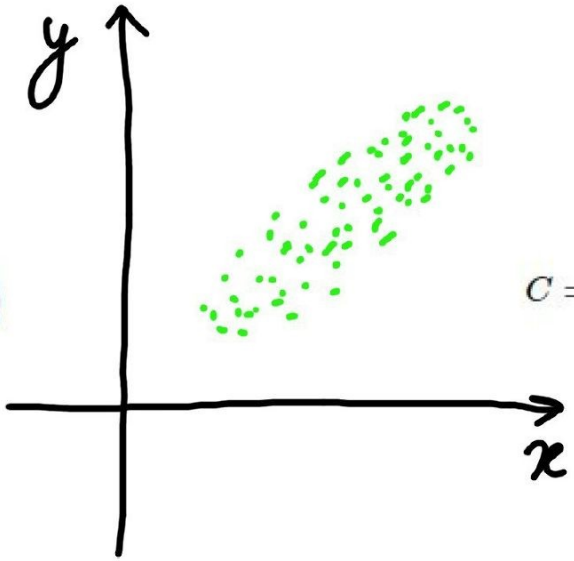
$$\rho_{X,\,Y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

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!

НЕТОЛОГИЯ