

# Юстина Иванова

Программист, data scientist

Основы описательной статистики,  
Основные понятия. Корреляция.  
Коэффициент корреляции Пирсона.  
Виды распределений.  
Нормальное распределение.  
Равномерное распределение.

# Спикер



## Юстина Иванова,

- PhD в университет Больцано (Италия)
- Data scientist по  
компьютерному зрению  
в компании ОЦРВ,
- Выпускница МГТУ им. Баумана
- Магистр по Artificial Intelligence  
В University of Southampton (Англия)

# 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.

# Основные понятия статистики

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

# Математическое ожидание

Среднее значение случайной величины.

Оно же  $\mu$

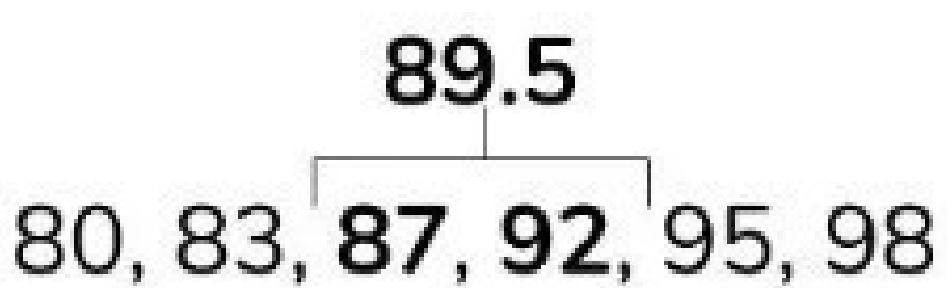
# Медиана

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

Мода определяется как значение, которое наиболее часто встречается в наборе данных.

# Выбросы

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

# Стандартное отклонение

Мера разброса данных (насколько данные варьируются от среднего значения).

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

Греческая буква «сигма» используется для обозначения стандартного отклонения

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

# Дисперсия

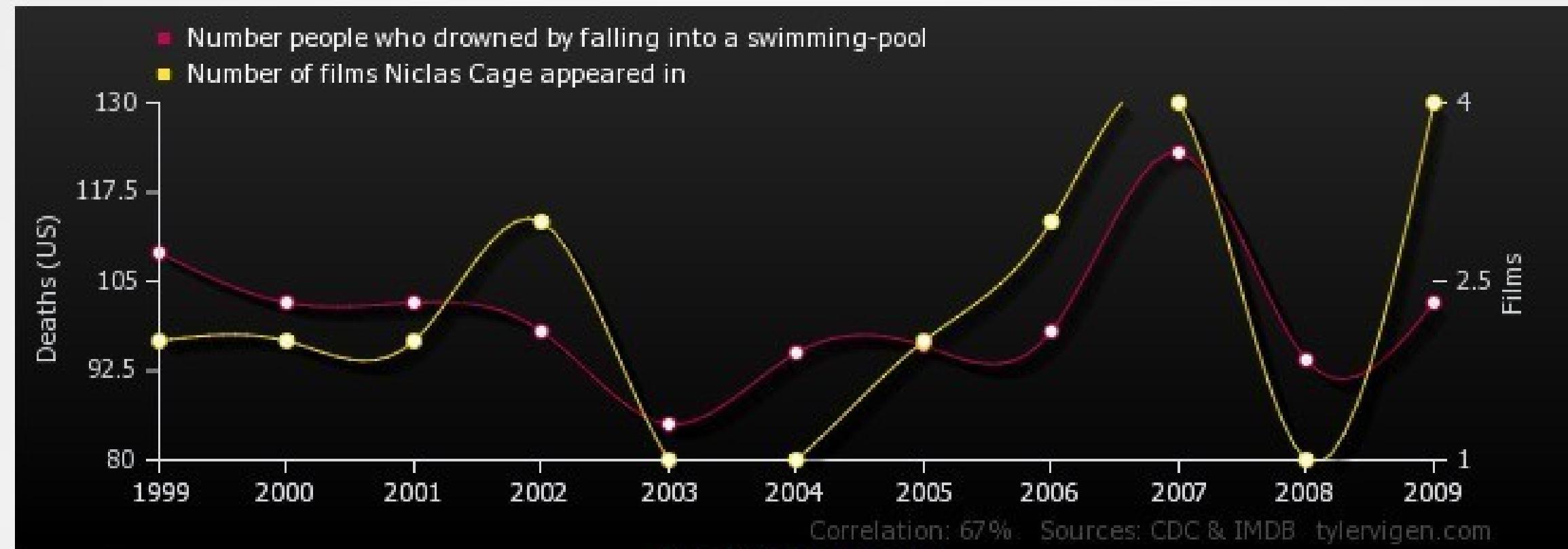
Квадрат стандартного отклонения. Дисперсия показывает, насколько в среднем значения сосредоточены, сгруппированы около среднего: если дисперсия маленькая - значения сравнительно близки друг к другу, если большая - далеки друг от друга (см. примеры нахождения дисперсии ниже).

# Корреляция

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

# Неожиданные случаи корреляции

**Number people who drowned by falling into a swimming-pool**  
 correlates with  
**Number of films Nicolas Cage appeared in**

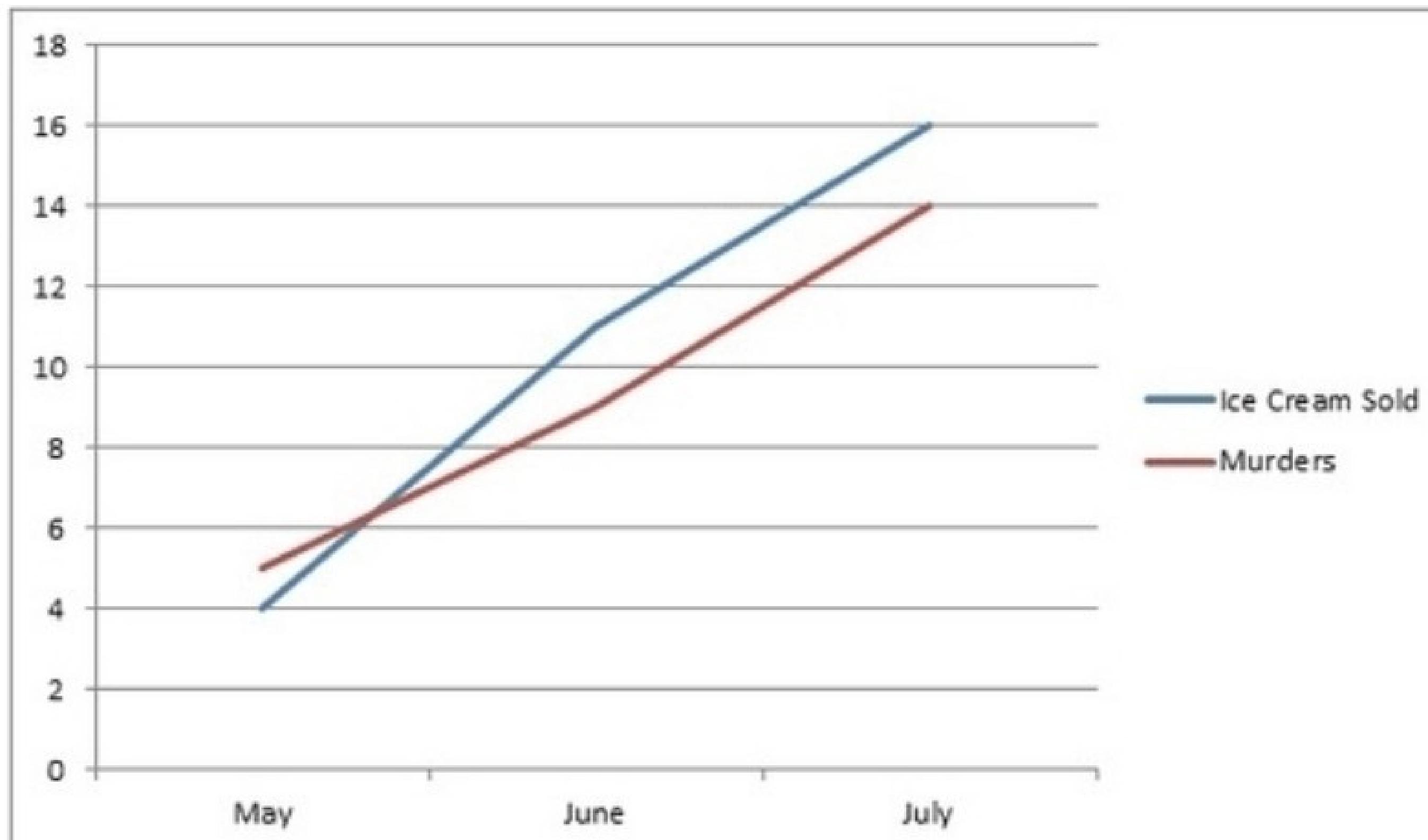


	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Number people who drowned by falling into a swimming-pool Deaths (US) (CDC)	109	102	102	98	85	95	96	98	123	94	102
Number of films Nicolas Cage appeared in Films (IMDB)	2	2	2	3	1	1	2	3	4	1	4

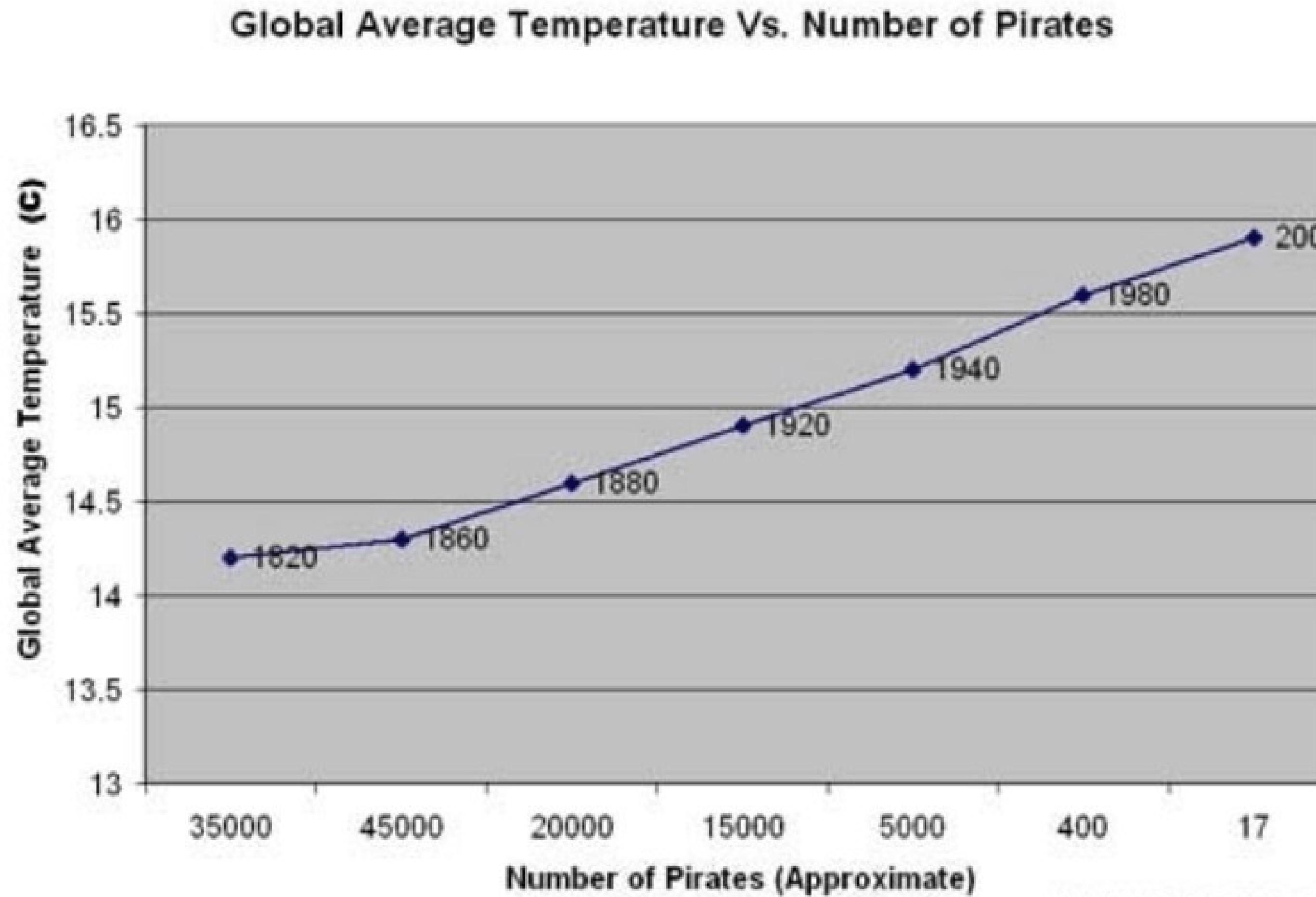
Correlation: 0.666004

# Неожиданные случаи корреляции

1. Ice cream consumption leads to murder.

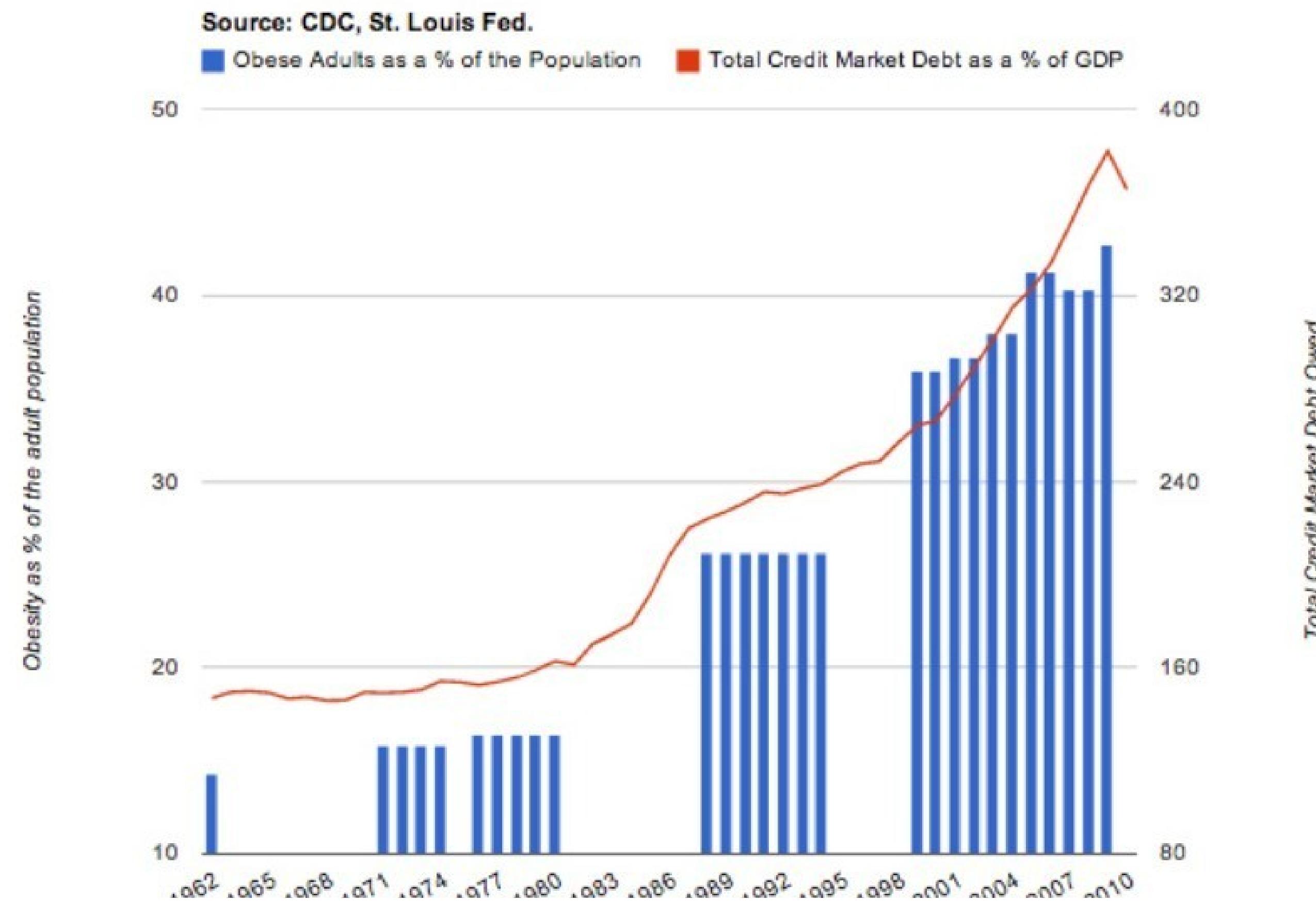


# Неожиданные случаи корреляции



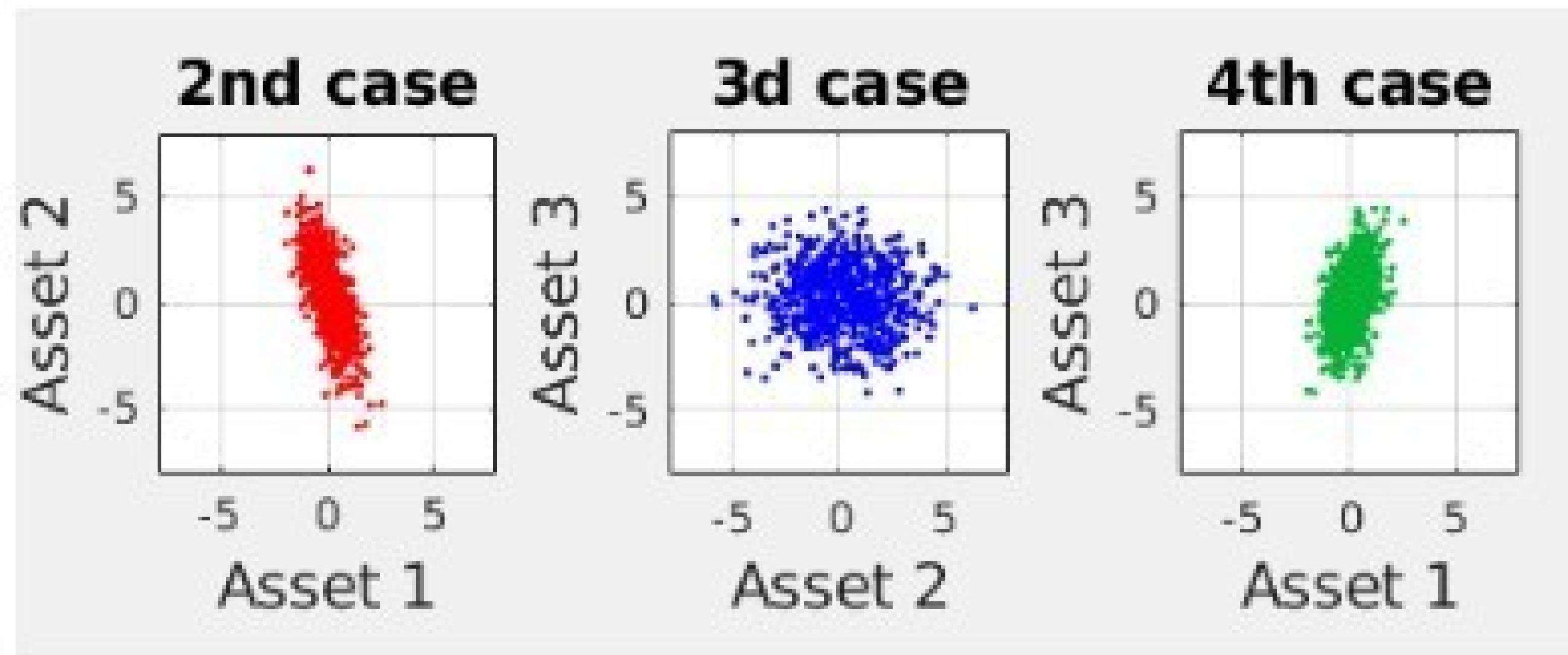
# Неожиданные случаи корреляции

## 8. Obesity caused the debt bubble.



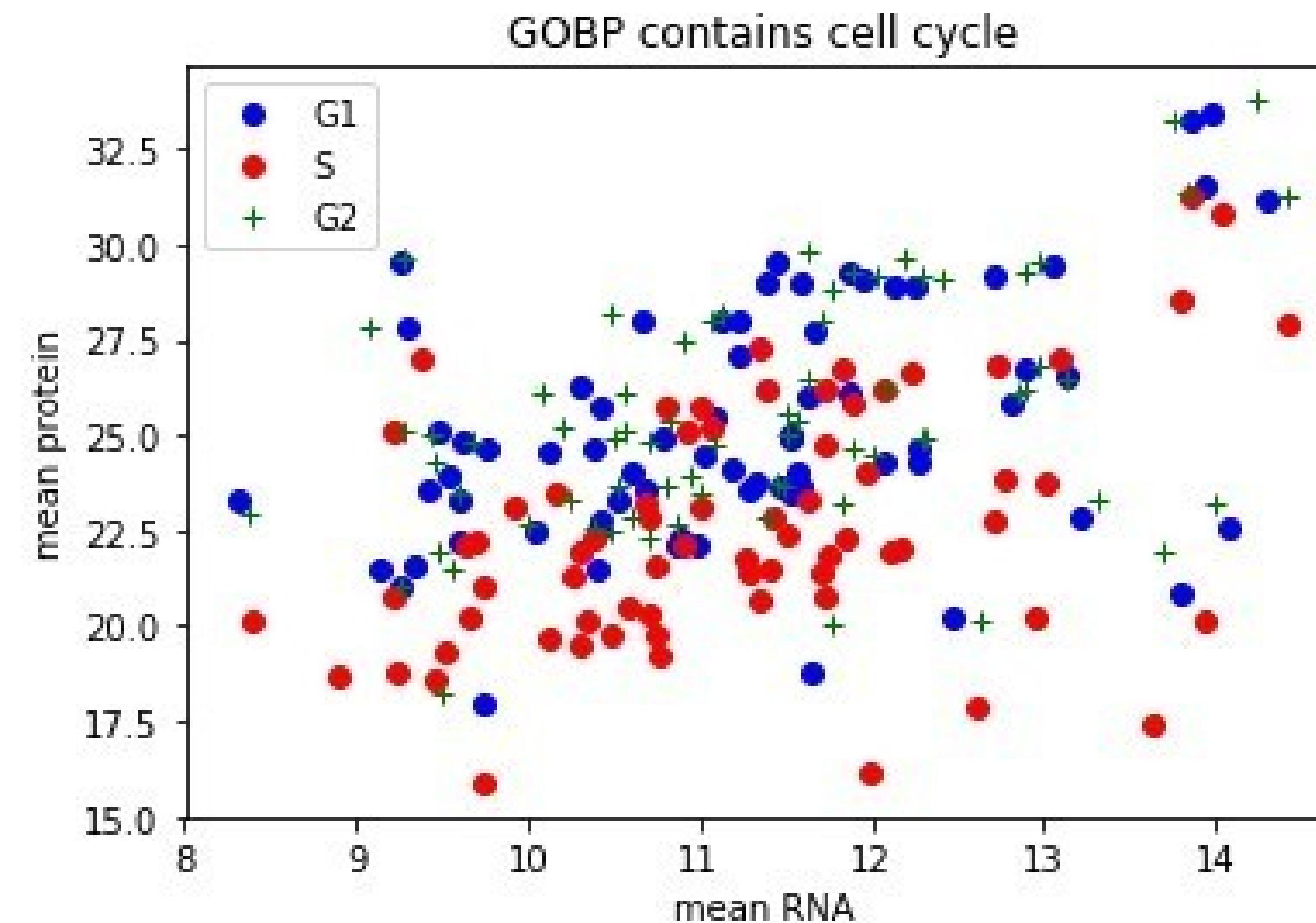
# Финансовый анализ данных

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



# Анализ молекулярной биологии

Насколько соотносится протеины и РНК



# Нормальное распределение Гаусса.

**Функция плотности распределения:**

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

**Математическое ожидание (среднее значение):**

$$\mu, \bar{x}$$

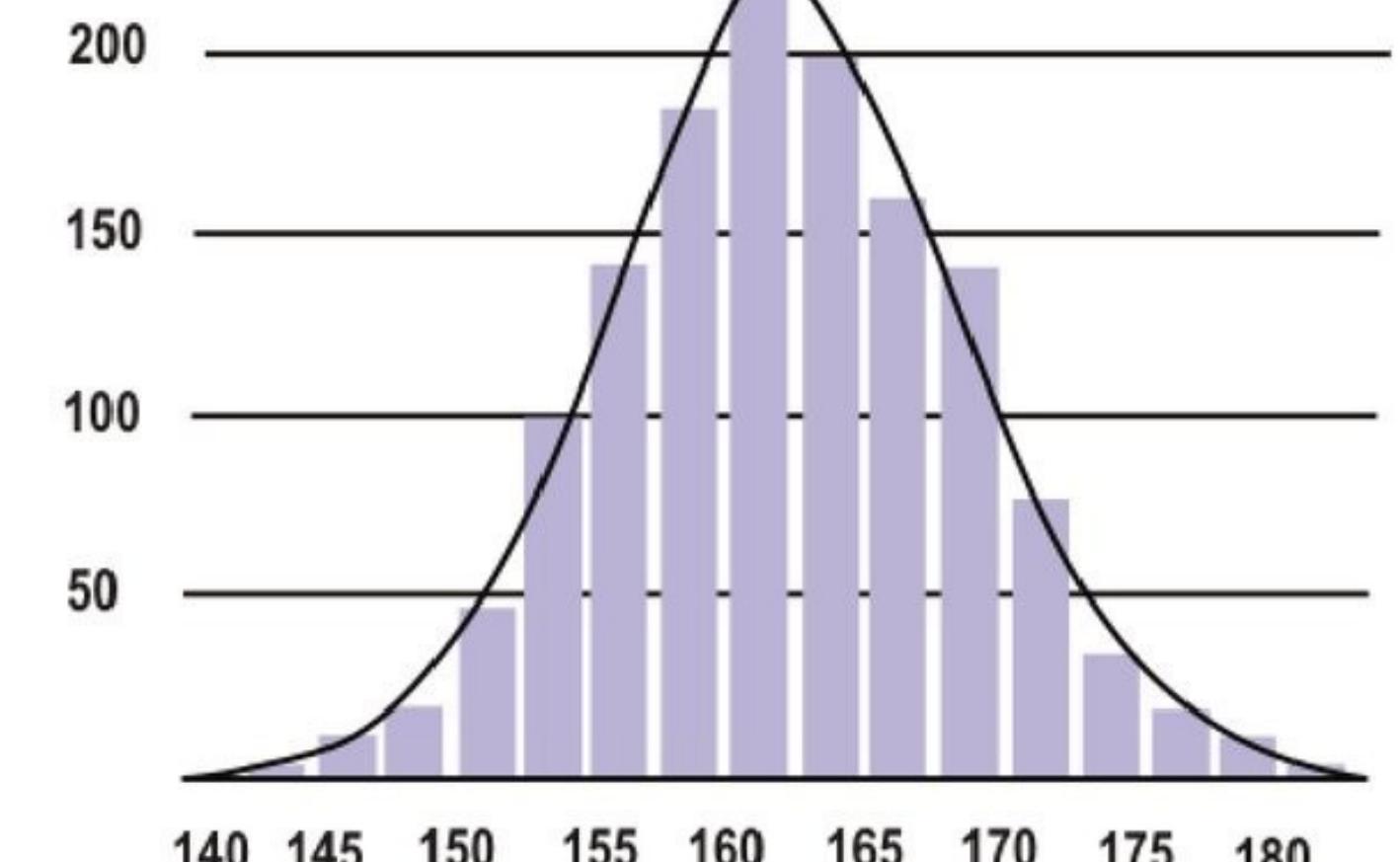
**Дисперсия (среднеквадратичное отклонение):**

$$\sigma^2$$

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

Греческая буква «сигма» используется для обозначения стандартного отклонения

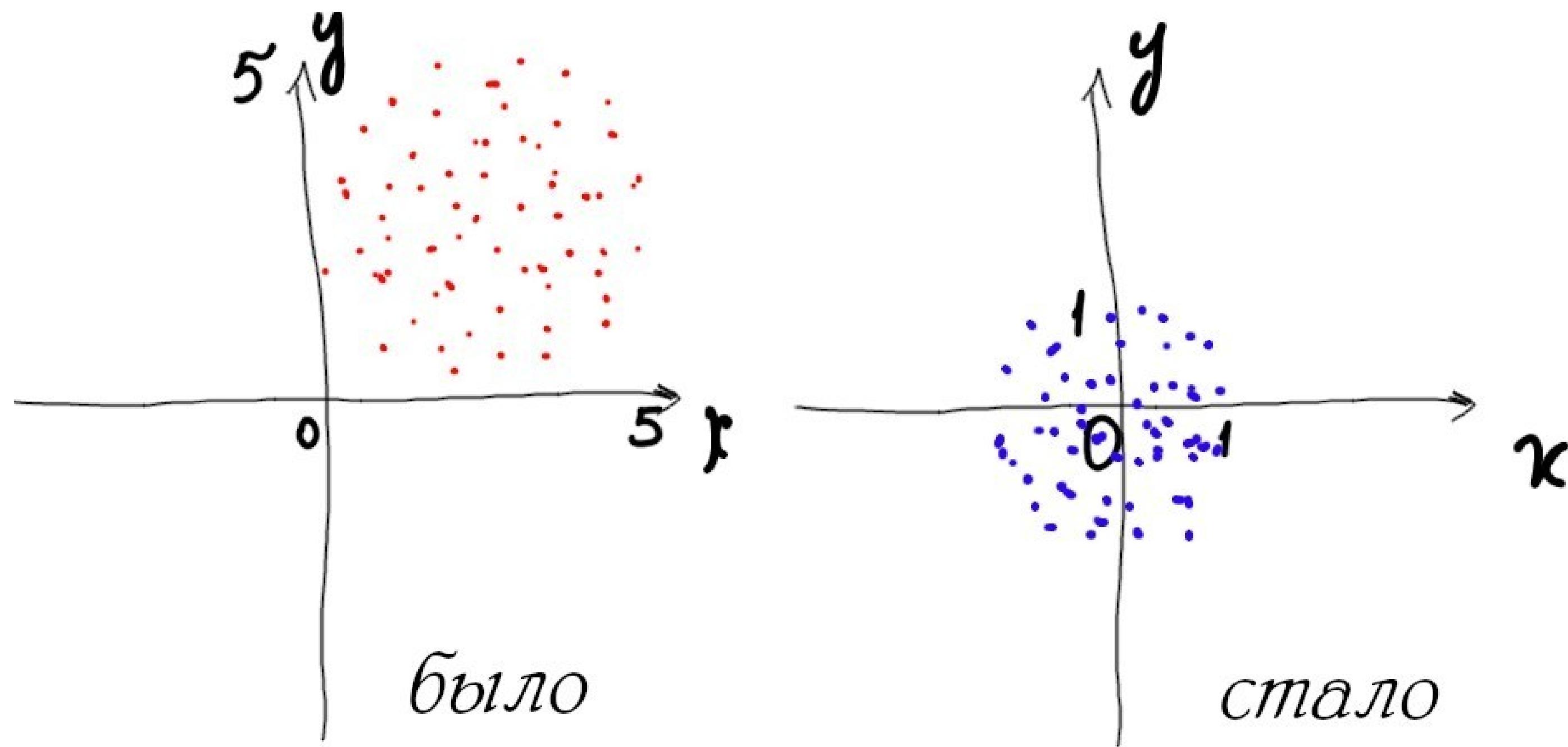
1. Вычтите каждое наблюдение из среднего значения
2. Возведите каждую разность в квадрат
3. Сложите все разности
4. Разделите сумму на количество наблюдений минус 1
5. Из результата извлеките квадратный корень



Распределение роста

# Нормализация данных.

Иногда данные перед анализом необходимо нормализовать.



## Задачи на упражнения.

Чему равно математическое ожидание и дисперсия случайной величины?

X	2	3	5	6	5	1
---	---	---	---	---	---	---

## Задачи на упражнения.

Чему равно математическое ожидание и дисперсия случайной величины?

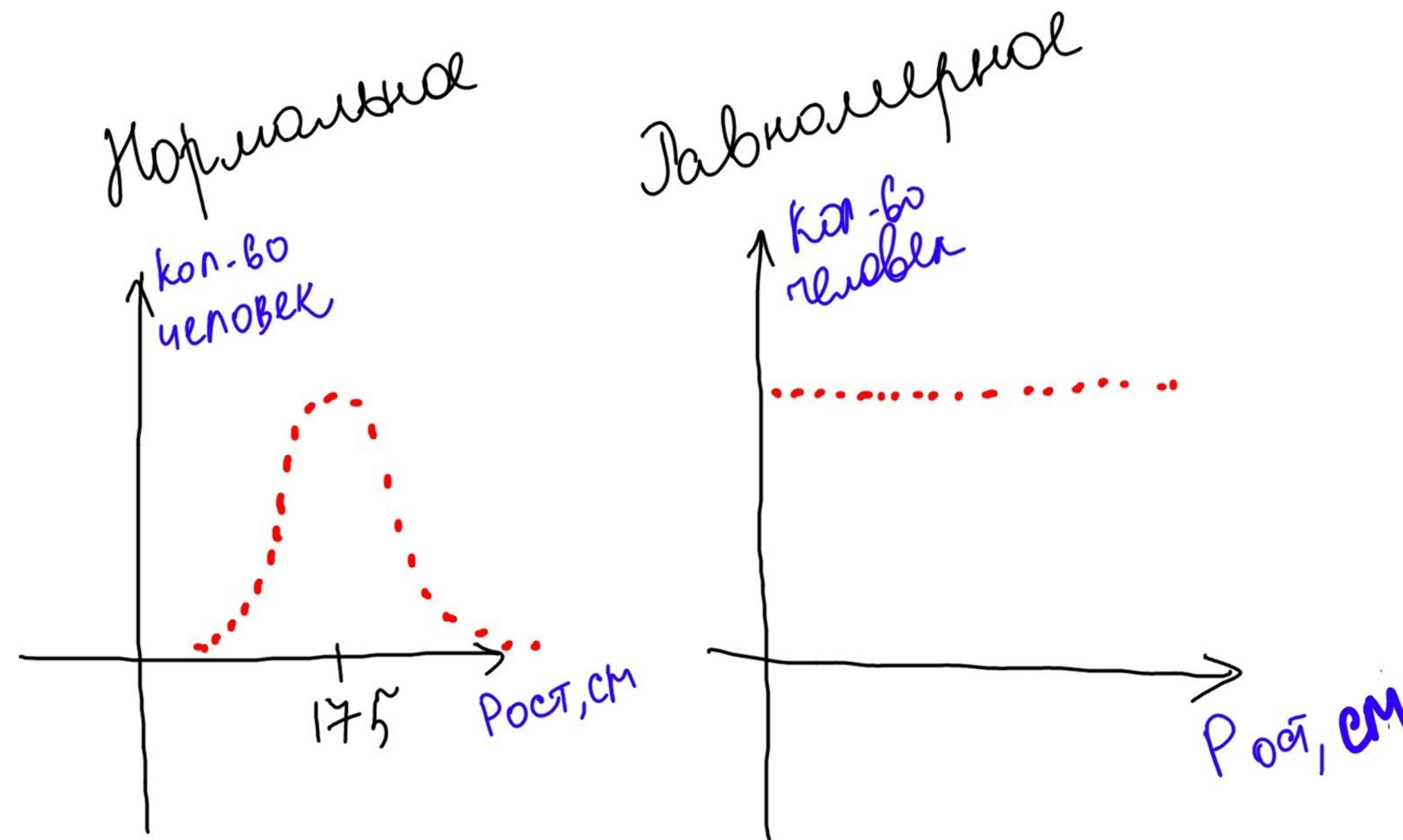
x	2	3	5	6	5	1
---	---	---	---	---	---	---

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

Дисперсия =  $1/5 ((2-3.6)^2 + (3-3.6)^2 + (5-3.6)^2 + (6-3.6)^2 + (5-3.6)^2 + (1-3.6)^2) = 4.632$

# Равномерное распределение.

Отличается тем, что данные распределены равномерно.



# Python: генератор случайных чисел — пример нормального распределения.

Random модуль — пример для генерации случайных чисел.

`random.random` — число от 0 до 1

`random.seed` — настройка генератора

Модуль `numpy` также имеет `random` метод.

`numpy.random.normal` -

<https://docs.scipy.org/doc/numpy/reference/generated/numpy.random.normal.html>

# Модуль (scipy.stats).

<https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.norm.html#scipy.stats.norm>

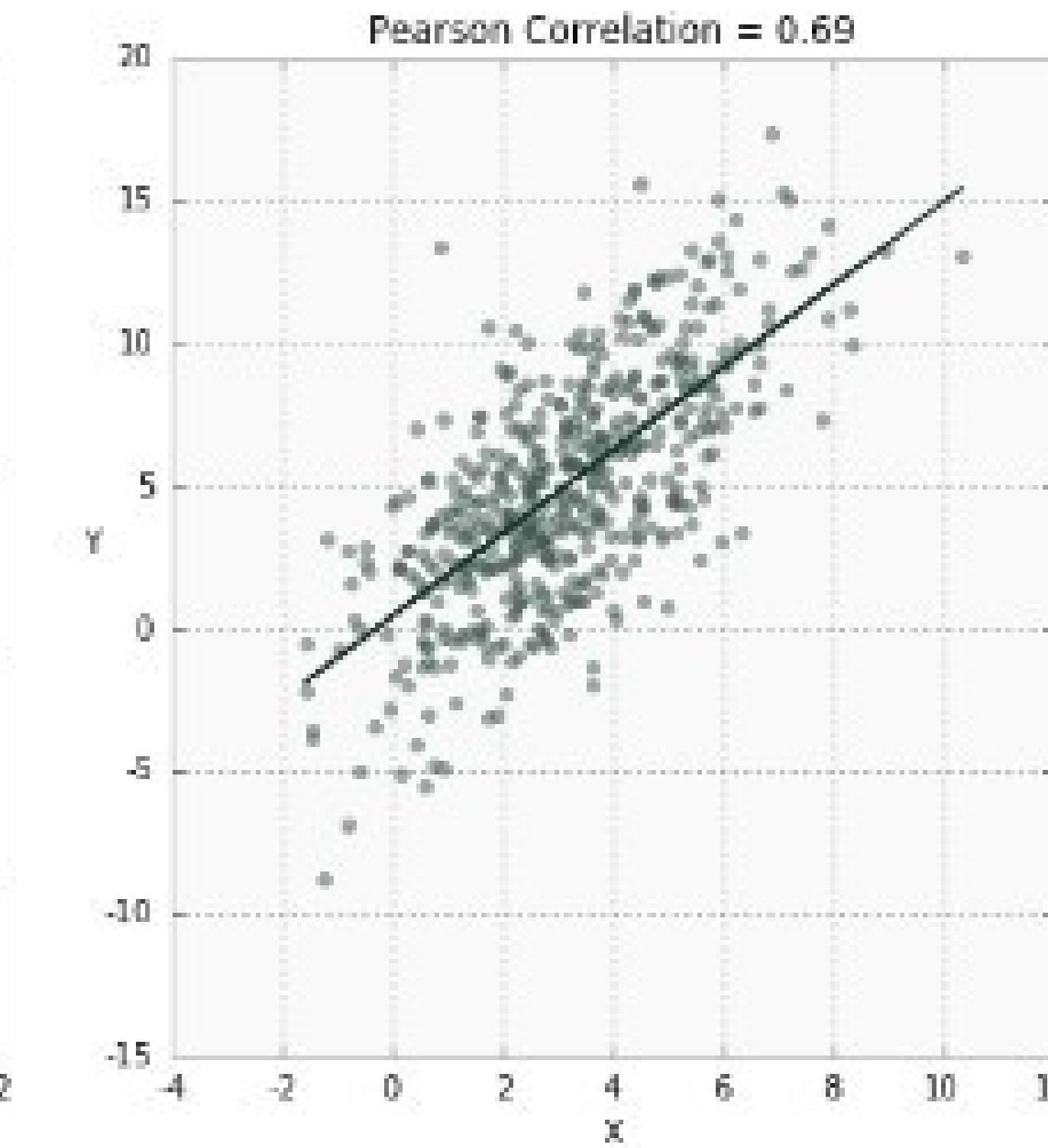
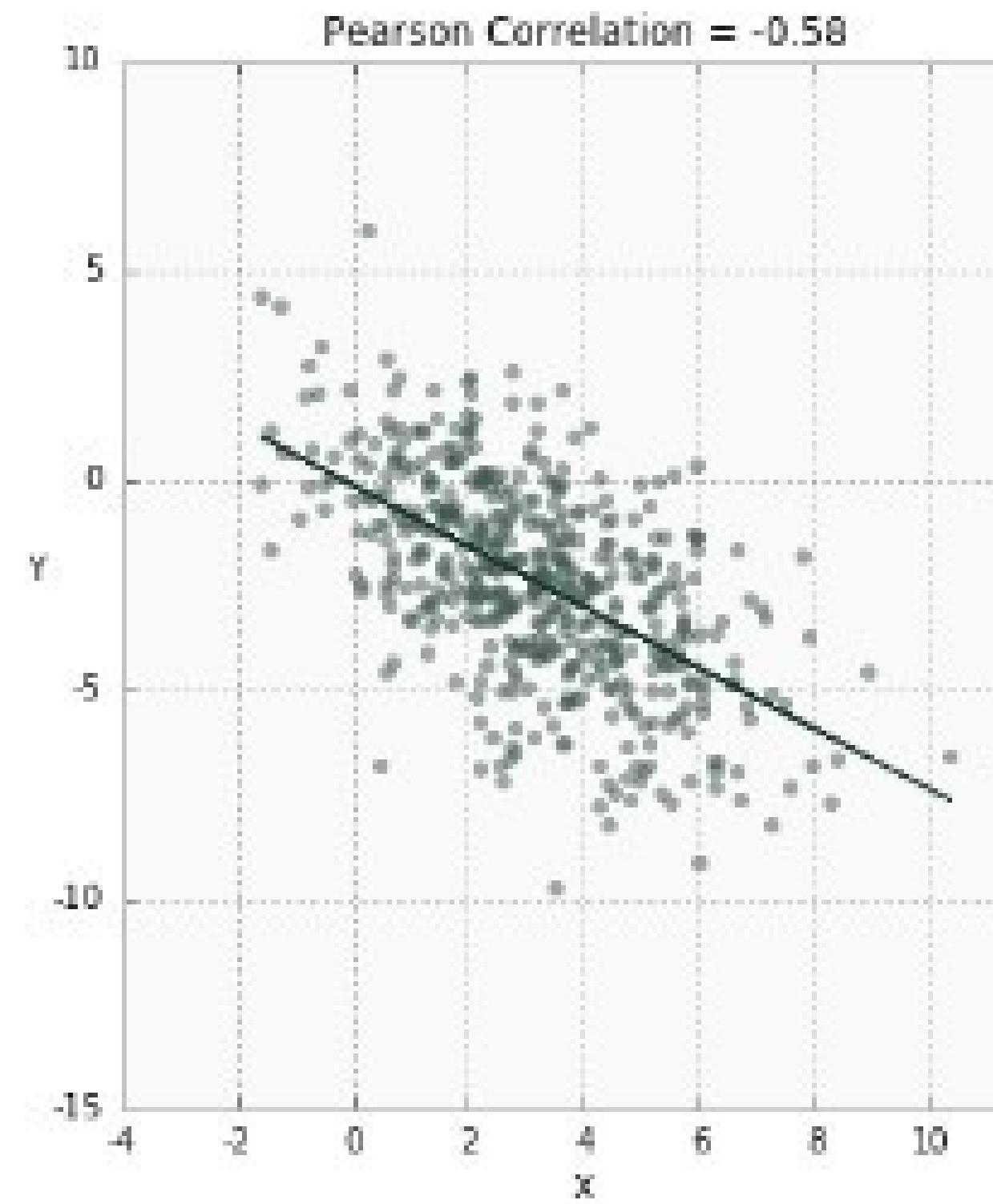
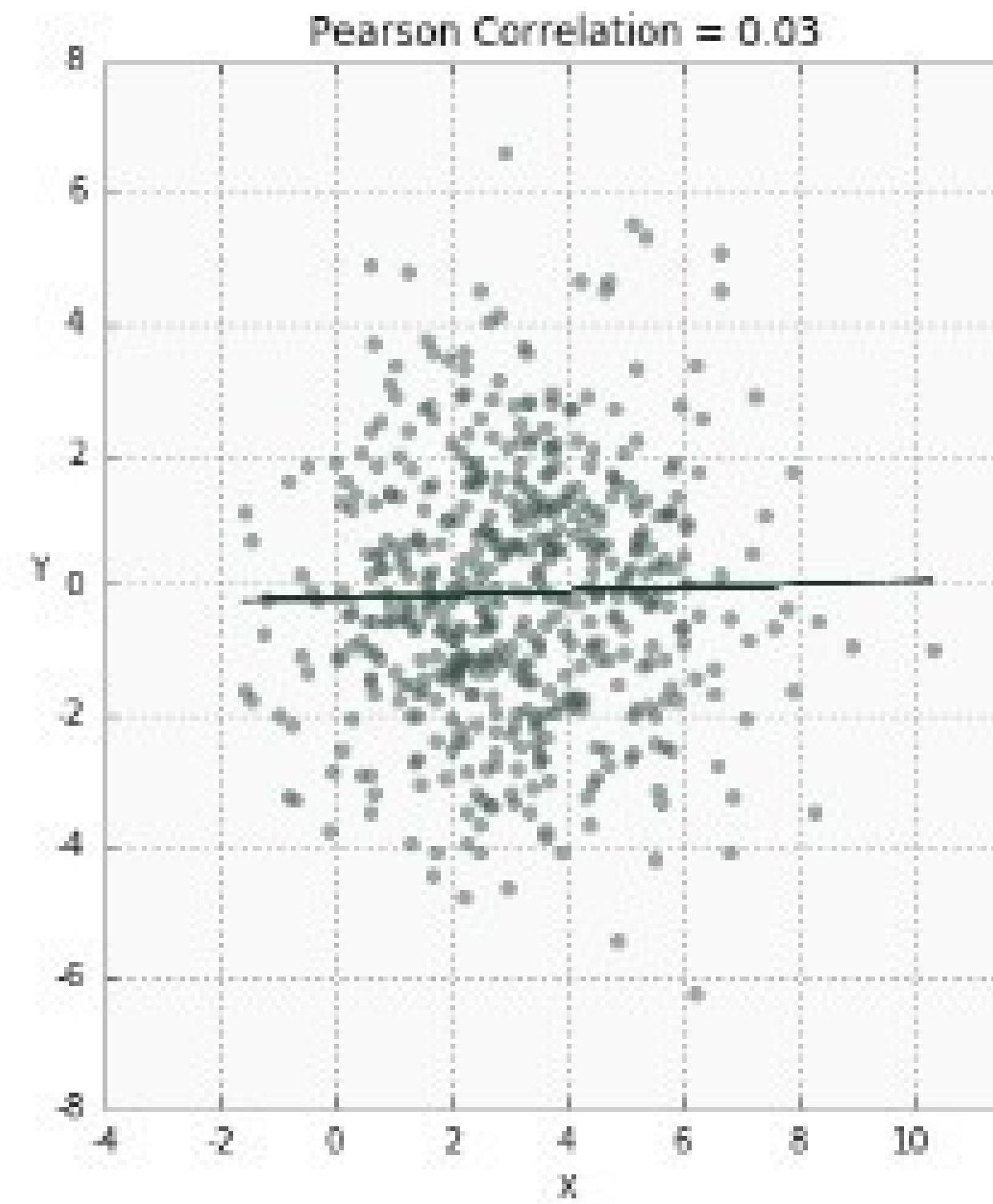
Основные функции для генерации случайных чисел:

stats.norm — создание нормального распределения

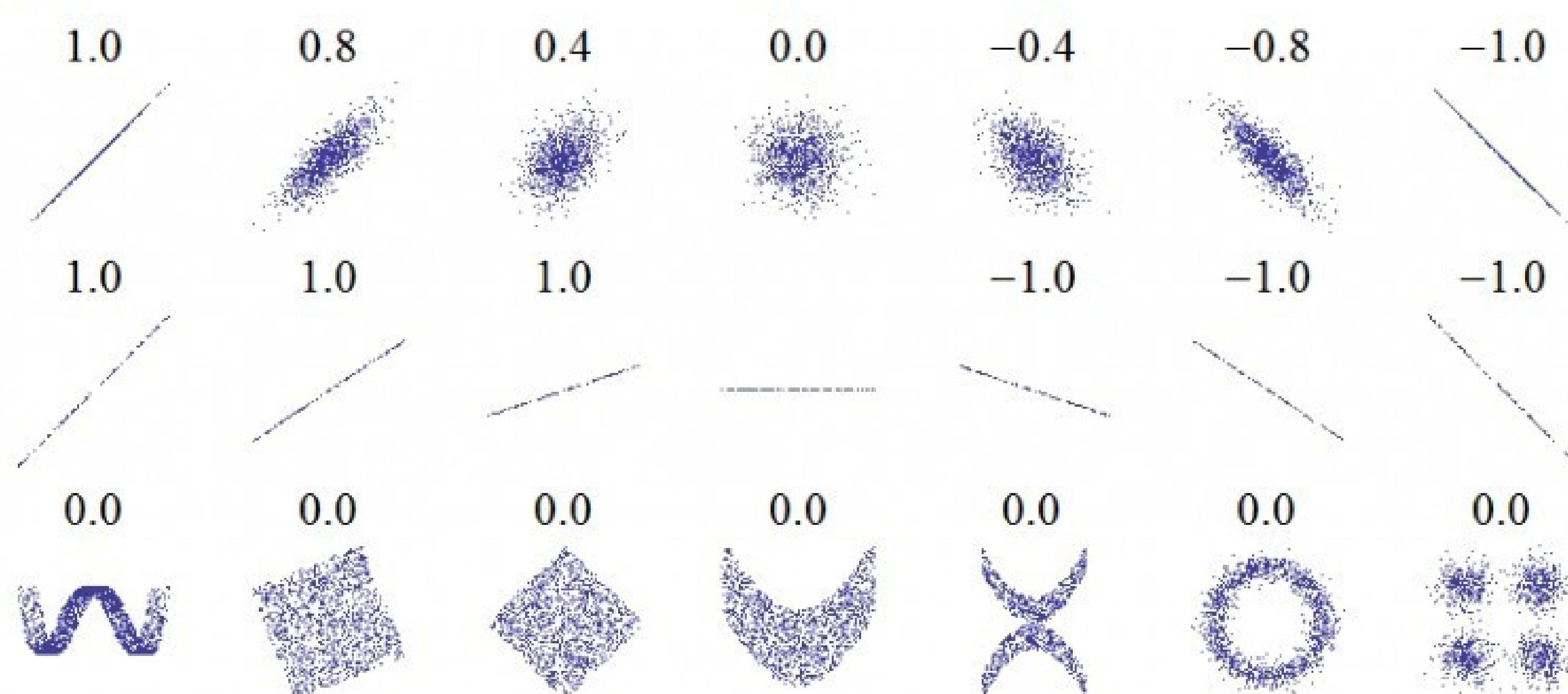
stats.norm.rvs(size=1000) — генерация случайного числа, можно задать дисперсию и математическое ожидание

# Корреляция Пирсона.

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



# Кореляция Пирсона.



[https://en.wikipedia.org/wiki/Pearson\\_correlation\\_coefficient](https://en.wikipedia.org/wiki/Pearson_correlation_coefficient)

# Матрица корреляций.

Для статистического анализа играет наиважнейшую роль.

Строим матрицу ковариаций для того, чтобы определить, насколько 2 случайные величины зависят друг от друга. Эта информация — то, что нужно data scientists.

Если две переменные независимы, матрица ковариаций выглядит так:

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

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

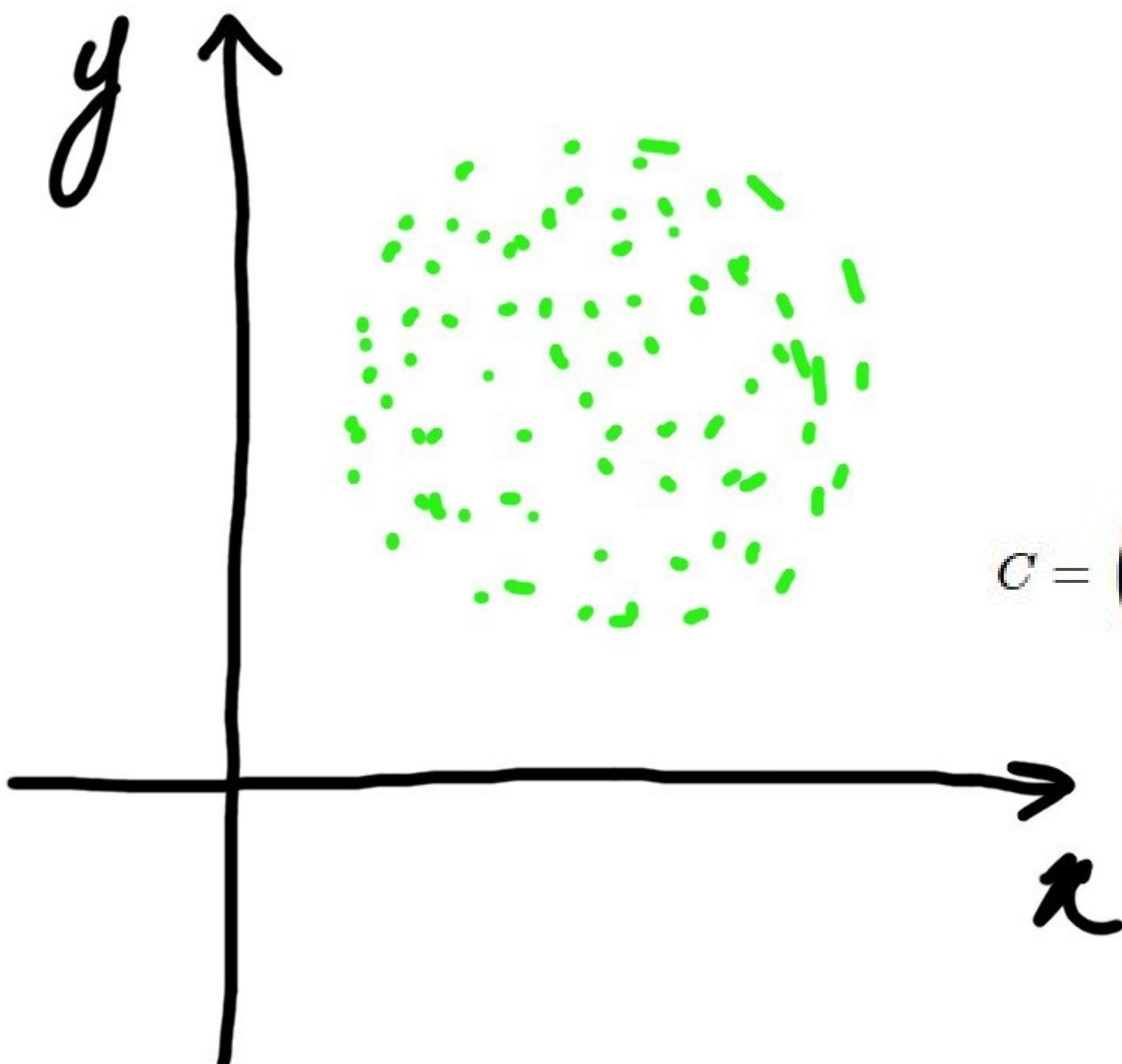
# Матрица корреляций.

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

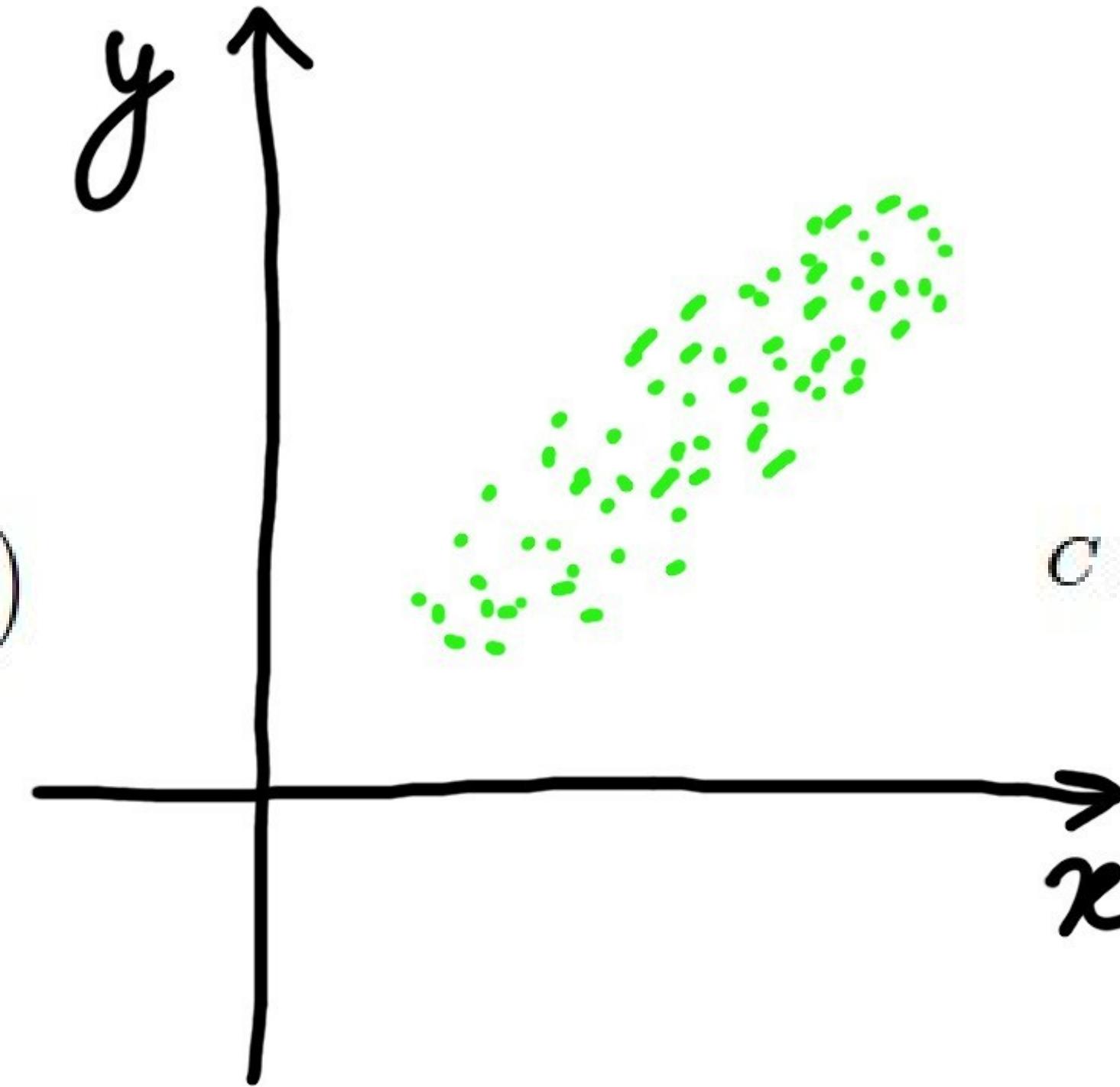
$$C = \begin{pmatrix} \sigma(x, x) & \sigma(x, y) \\ \sigma(y, x) & \sigma(y, y) \end{pmatrix}$$

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

# Матрица корреляций.



Независимые переменные



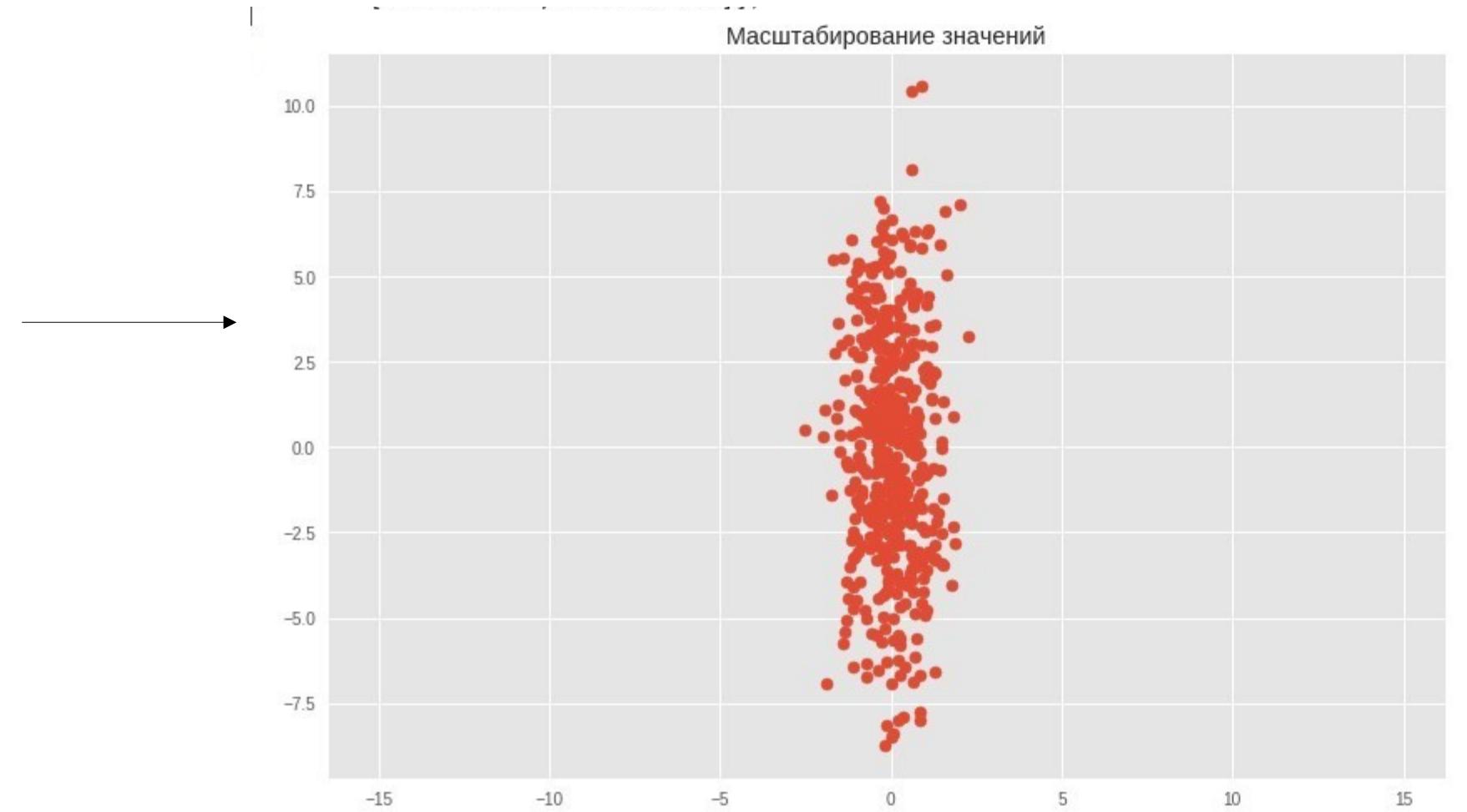
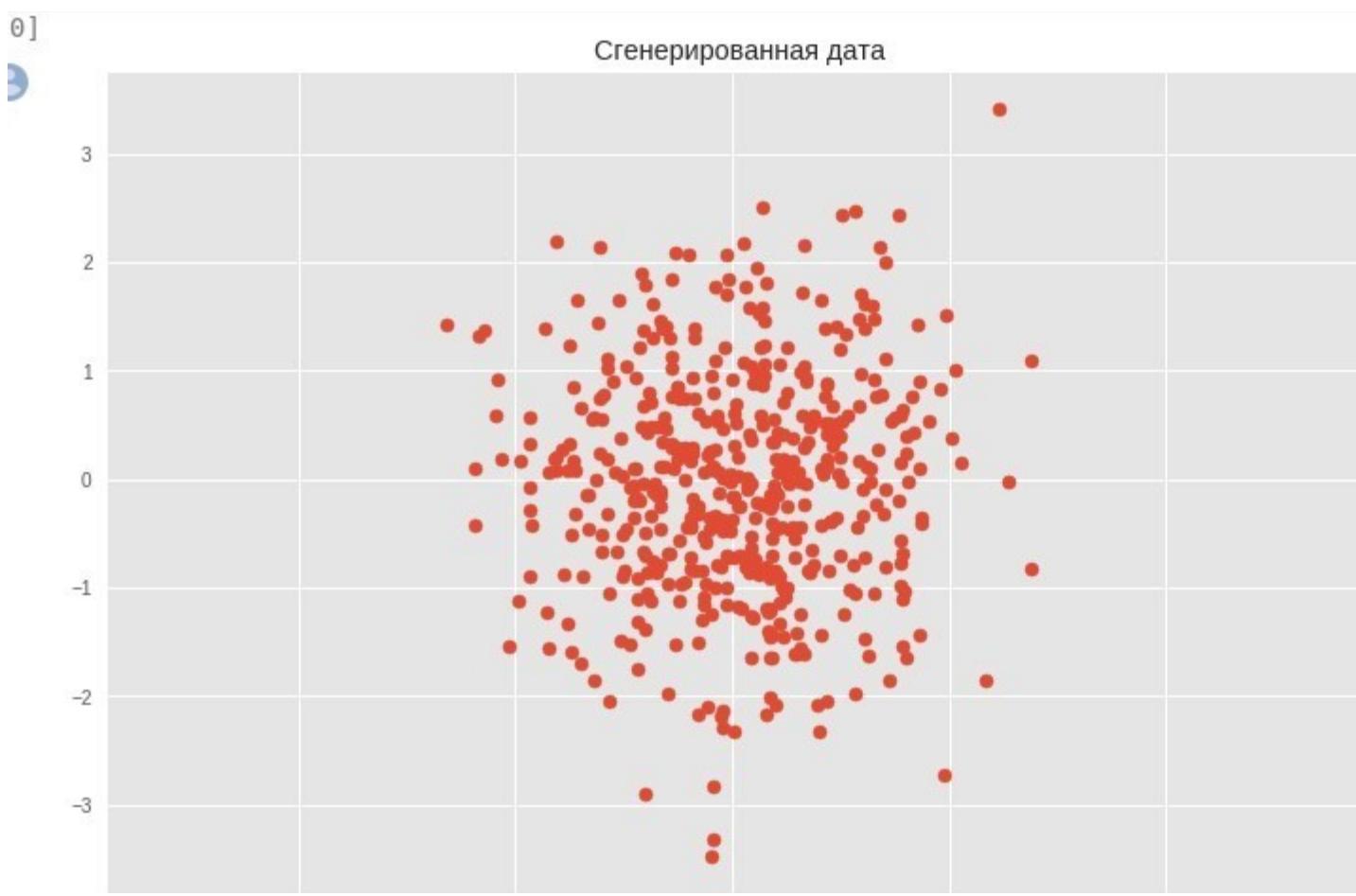
Зависимые переменные

# Матрица корреляций: изменить распределение.

Распределение случайных величин можно оценивать с помощью матрицы ковариаций.

Масштабирование (увеличение или уменьшение разброса)

$$C = \begin{pmatrix} (s_x\sigma_x)^2 & 0 \\ 0 & (s_y\sigma_y)^2 \end{pmatrix}$$



# Перемножение матриц!

Необходимо уметь перемножать матрицы для статистики. Почему? Матрица кореляций нужна для анализа, для определения, насколько переменные зависимы друг от друга. Это можно сделать с помощью линейной алгебры. Сложно, но можно разобраться.

Numpy.sum() - суммирует все элементы

Numpy.ndarray.dot() - умножает одну матрицу на другую

Numpy.ndarray.T — транспонирование матрицы

numpy.vstack((x, y)) — составляем матрицу из двух матриц, вставленных по вертикали

Спасибо за внимание! Вопросы?