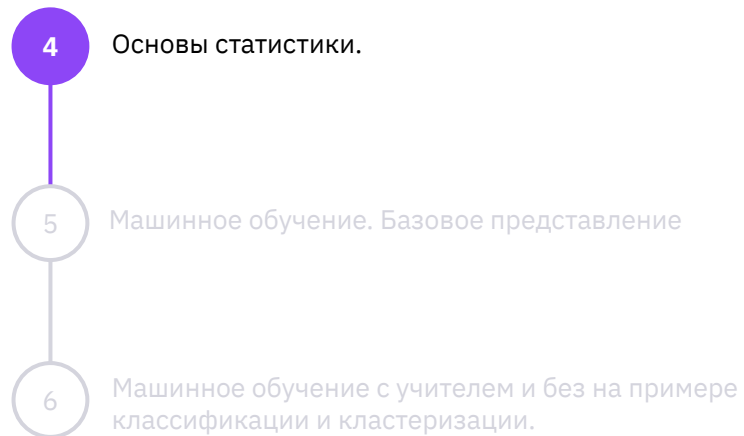


# ОСНОВЫ СТАТИСТИКИ

Урок 4



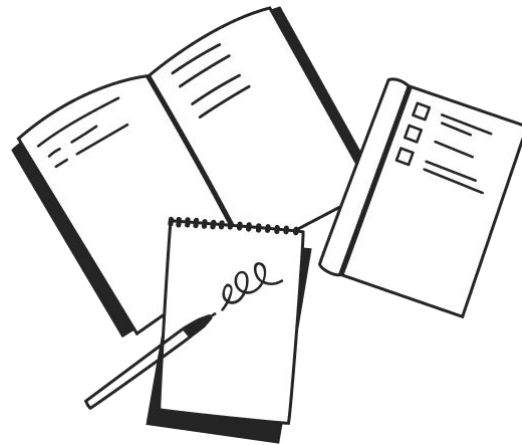
# План курса





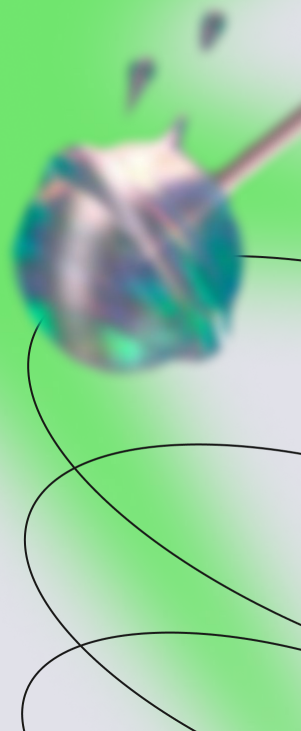
## Что будет на уроке сегодня

- 📌 Почему статистика
- 📌 Наблюдения и переменные
- 📌 Матрица данных и таблица частот
- 📌 Графики и формы распределения
- 📌 Среднее значение, медиана и мода
- 📌 Диапазоны
- 📌 Дисперсия и стандартное отклонение
- 📌 Нормальное распределение, биномиальное распределение и распределение Пуассона





Почему  
статистика?





# Основные шаги в процессе анализа данных

**Шаг 1:** Определитесь с целями или задайте вопрос

**Шаг 2:** Что измерять и как измерять

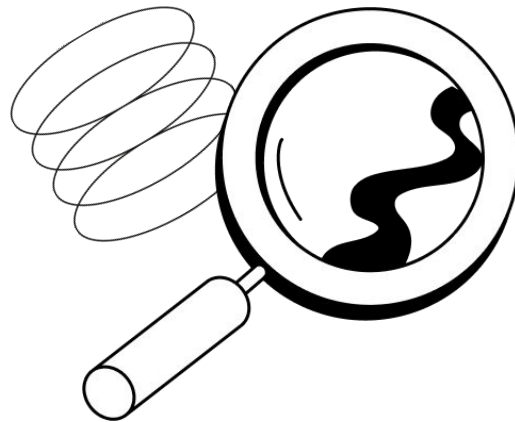
**Шаг 3:** Сбор данных

**Шаг 4:** Очистка данных

**Шаг 5:** Обобщение и визуализация данных

**Шаг 6:** Моделирование данных

**Шаг 7:** Оптимизируйте и повторите





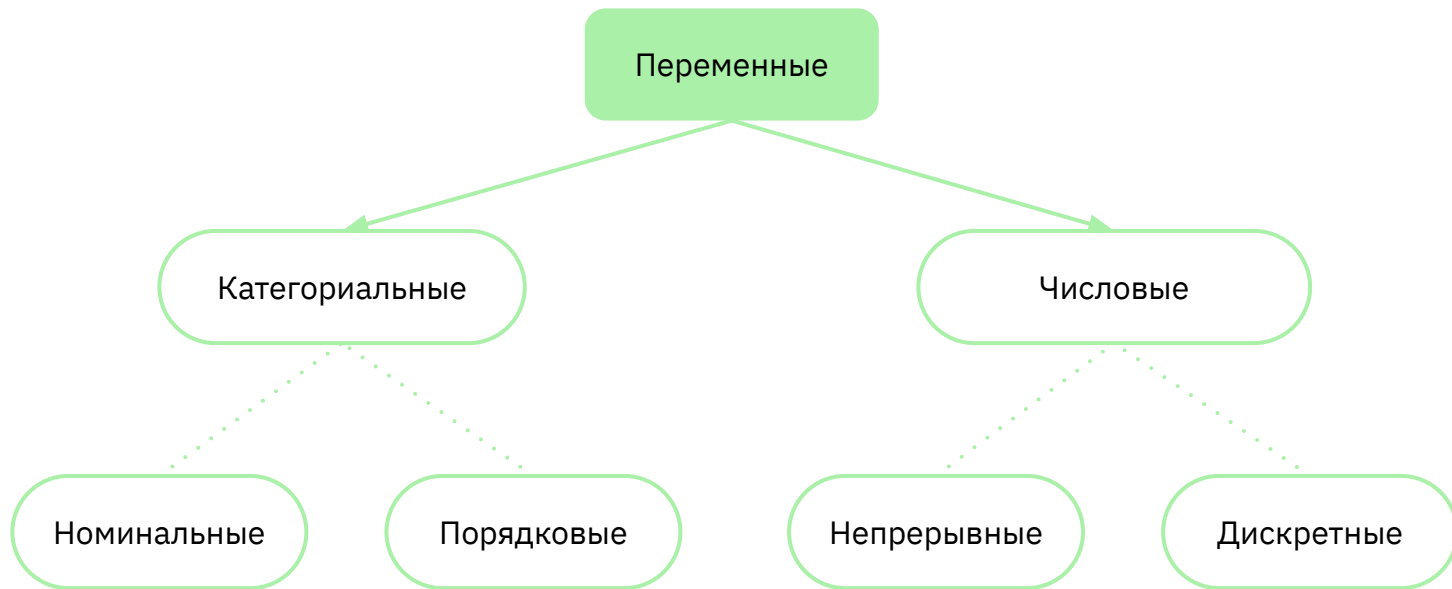
# Наблюдение

**Наблюдение** (кейс) — экспериментальная единица. Это могут быть лица, от которых собираются данные

**Переменная** — это измеряемая характеристика, которая может принимать различные значения.  
Другими словами, что-то, что варьируется между различными наблюдениями.



# Типы переменных





## Примеры переменных

Номинальная переменная

Mode of transportation for travel to work	Number of people
Car, truck, van as driver	9,929,470
Car, truck, van as passenger	923,975
Public transit	1,406,585
Walked	881,085
Bicycle	162,910
Other methods	146,835

Порядковая переменная

Student behaviour ranking	
Behaviour	Number of students
Excellent	5
Very good	12
Good	10
Bad	2
Very bad	1





## Матрица данных и таблица частот

Матрица данных — это прямоугольная таблица или матрица, в которой строки представляют наблюдения или случаи, а столбцы — переменные или атрибуты

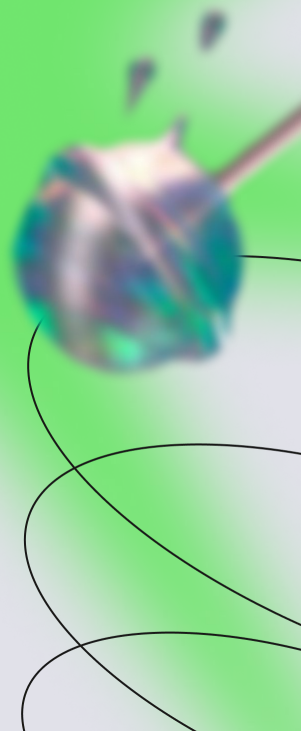
Таблица частот представляет собой табличное представление частотного распределения категориальной переменной. Она показывает количество или частоту наблюдений, которые попадают в каждую категорию переменной

Наблюдения	Переменные				
	sepal length	sepal width	petal length	petal width	class
	5.1	3.5	1.4	0.2	Iris-setosa
	4.9	3	1.4	0.2	Iris-setosa
	6.5	3.2	5.1	2	Iris-virginica
	6.4	2.7	5.3	1.9	Iris-virginica
	6.8	3	5.5	2.1	Iris-virginica
	6.7	3.1	4.4	1.4	Iris-versicolor
	5.6	3	4.5	1.5	Iris-versicolor
	5.8	2.7	4.1	1	Iris-versicolor

Class	Frequency	Percentage	Cumulative Percentage
Iris-setosa	2	25%	25%
Iris-virginica	3	38%	63%
Iris-versicolor	3	38%	100%
Total	8	100%	



# Графики и формы распределения





## Основные понятия

**Центральная тенденция** распределения относится к тому, где собираются данные. Наиболее распространенными мерами центральной тенденции являются среднее значение, медиана и мода

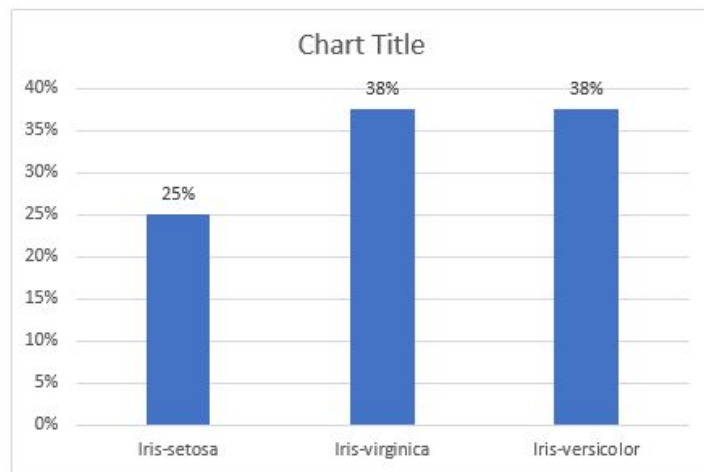
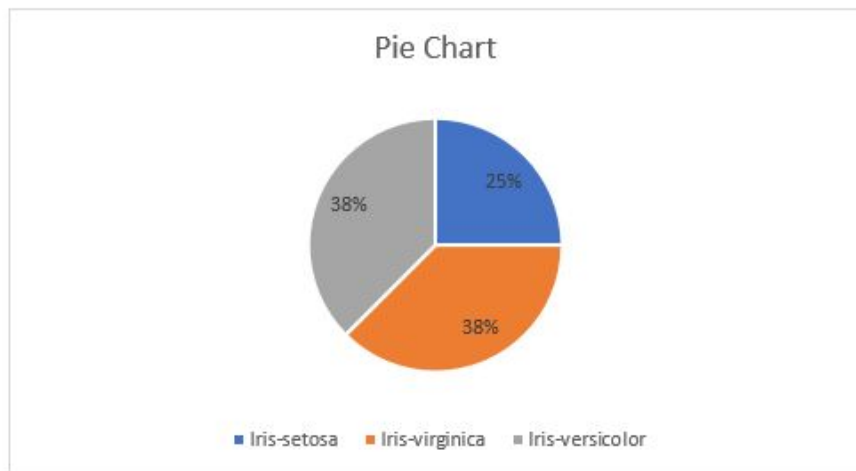
**Изменчивость** распределения относится к тому, насколько разбросаны данные

**Асимметрия** относится к степени асимметрии в распределении



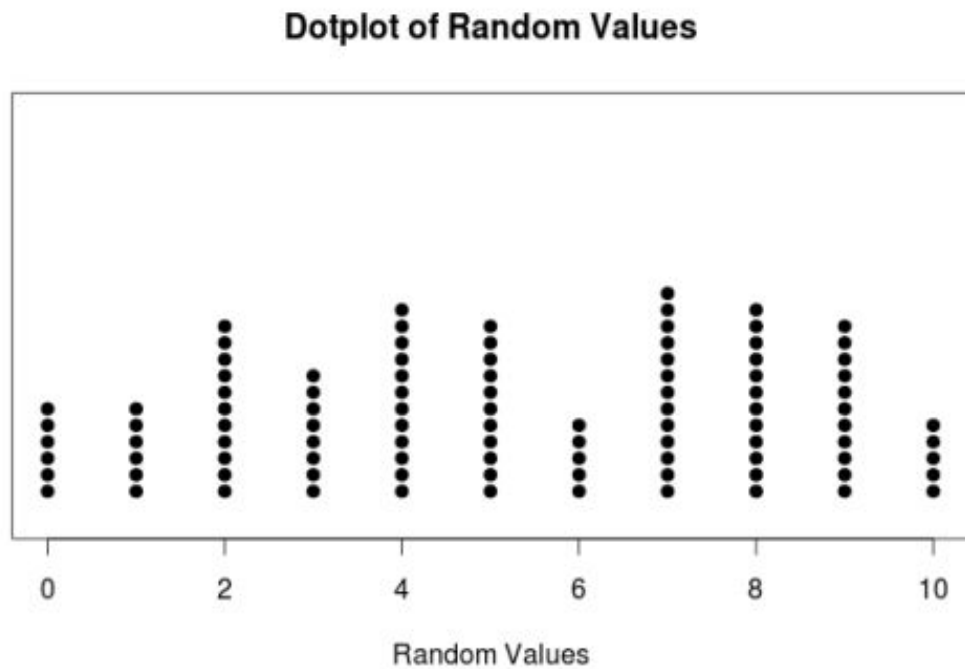
## Для категориальных переменных

Class	Frequency	Percentage	Cumulative Percentage
Iris-setosa	2	25%	25%
Iris-virginica	3	38%	63%
Iris-versicolor	3	38%	100%
Total	8	100%	





## Точечный график

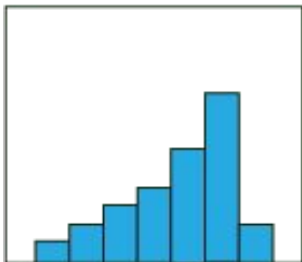




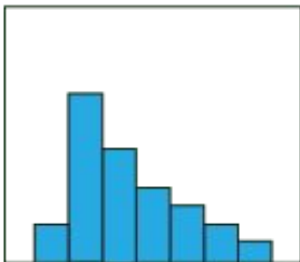
# Гистограмма

Гистограммы могут быть трех разных форм

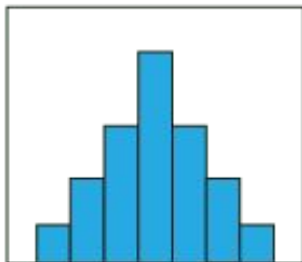
Skewed Left



Skewed Right



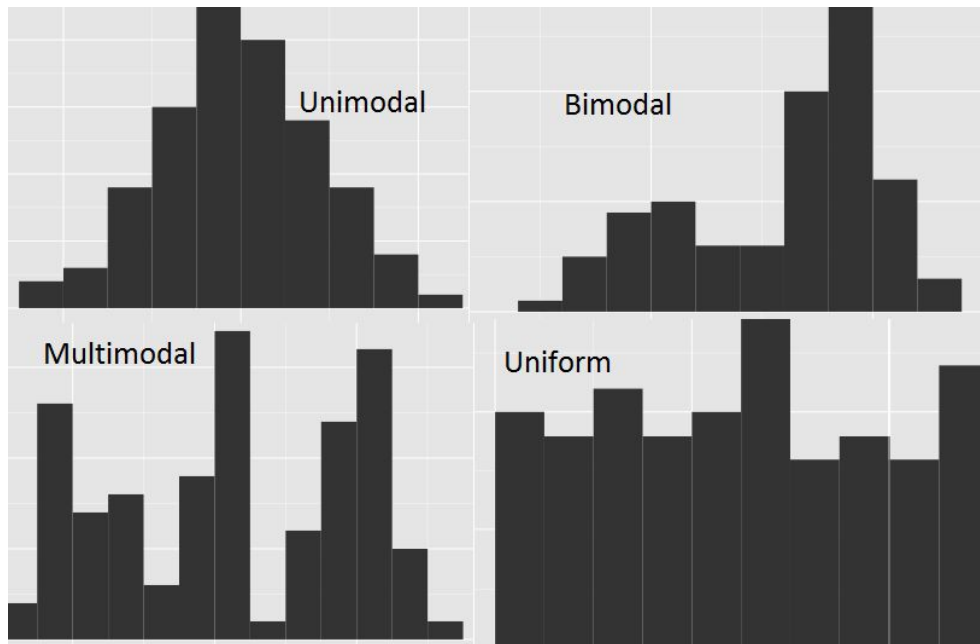
Symmetric





# Гистограмма

Также гистограммы могут представлять один из четырех видов модальности





# Среднее значение, медиана и мода







## Оценка распределения данных





## Мода

5, 6, 5, 7, 5, 8, 9, 5

↓  
Мода

Мода равняется 5,  
потому что это самое  
часто встречающееся  
значение

5, 6, 5, 6, 5, 8, 6, 6, 5

↓  
Мода

Мода равняется 5 и 6  
потому что оба  
значения  
встречаются с  
одинаковой частотой

## Медиана

7,8,7,6,9,8,8

Медиана

~~6 7 7, 8 8 8 9~~

Здесь медиана будет  
равняться 8

7,8,7,6,9,8,8,7

Медиана

~~6 7 7, 7, 8 8 8 9~~

Здесь медиана будет  
равняться  
 $(8+7)/2 = 7.5$



## Среднее значение

7,8,7,6,9,8,8

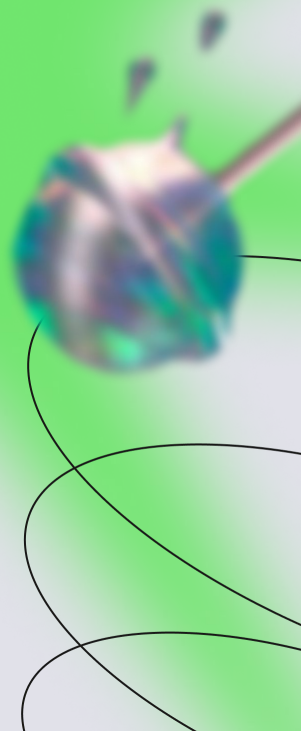
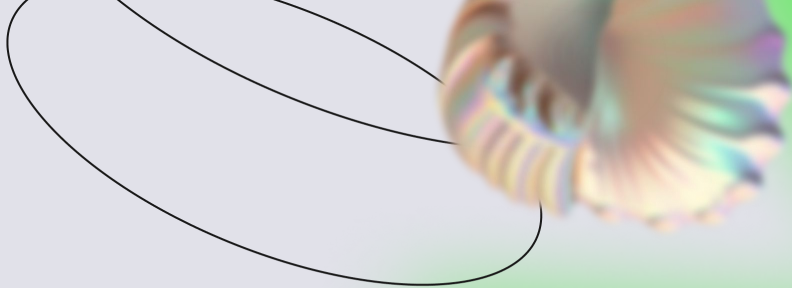


Среднее

$\text{sum}(53)/\text{кол-во}$   
 $\text{измерений}(7) = 7.6$



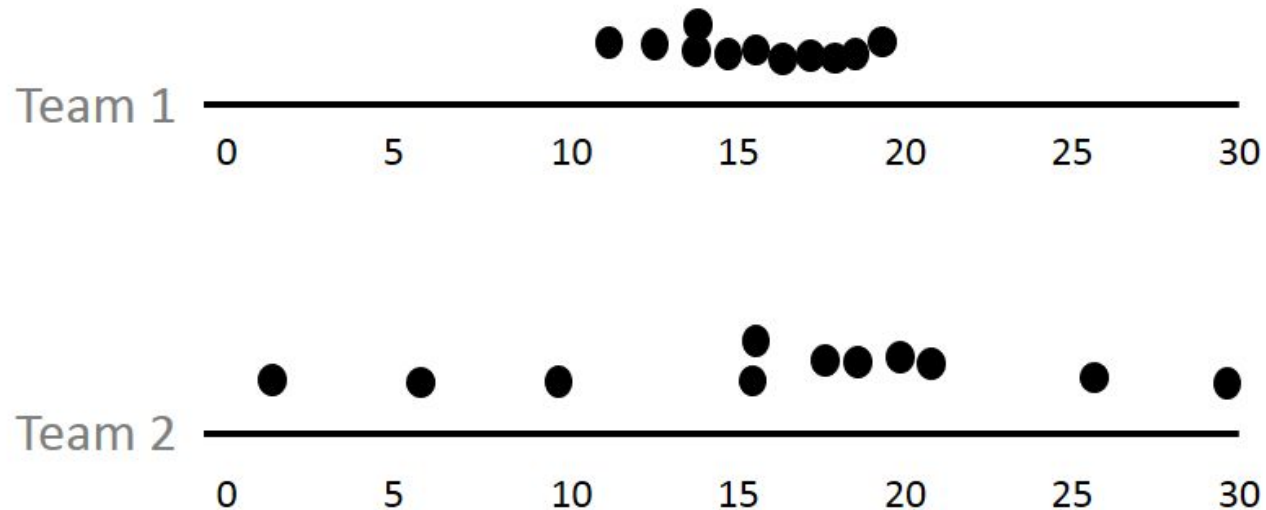
# Меры изменчивости





## Диапазон

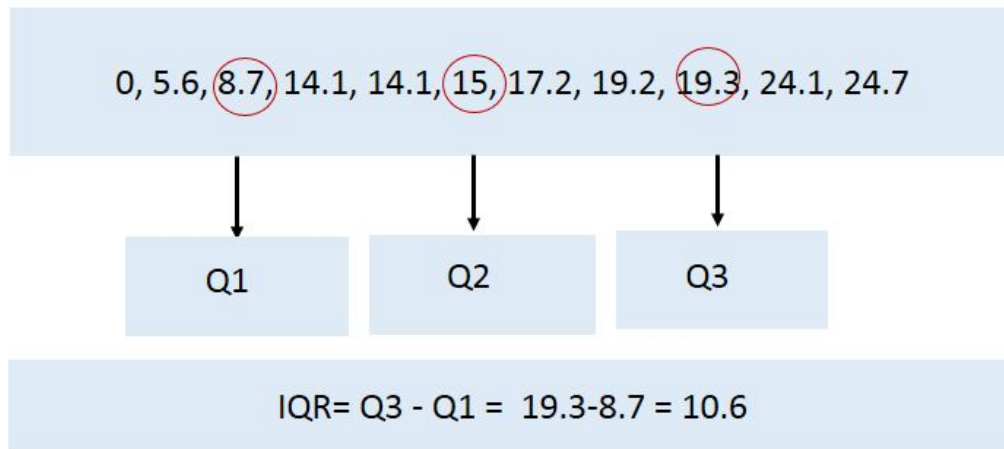
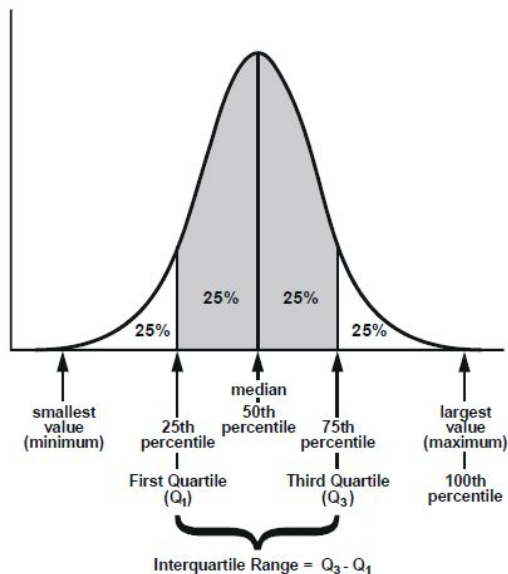
Диапазон — это статистическая мера, которая рассчитывается путем вычитания минимального значения набора данных из максимального значения





## Межквартильный диапазон (размах)

Межквартильный размах (IQR) — это мера статистической дисперсии, основанная на делении набора данных на квартили





## Преимущества межквартильного диапазона

- Основное преимущество IQR заключается в том, что на него не влияют выбросы, поскольку он не принимает во внимание наблюдения ниже  $Q1$  или выше  $Q3$ .
- Тем не менее, может быть полезно искать возможные выбросы в вашем исследовании.
- Как правило, наблюдения можно квалифицировать как выбросы, если они лежат более чем на  $1,5$  IQR ниже первого квартиля или на  $1,5$  IQR выше третьего квартиля.

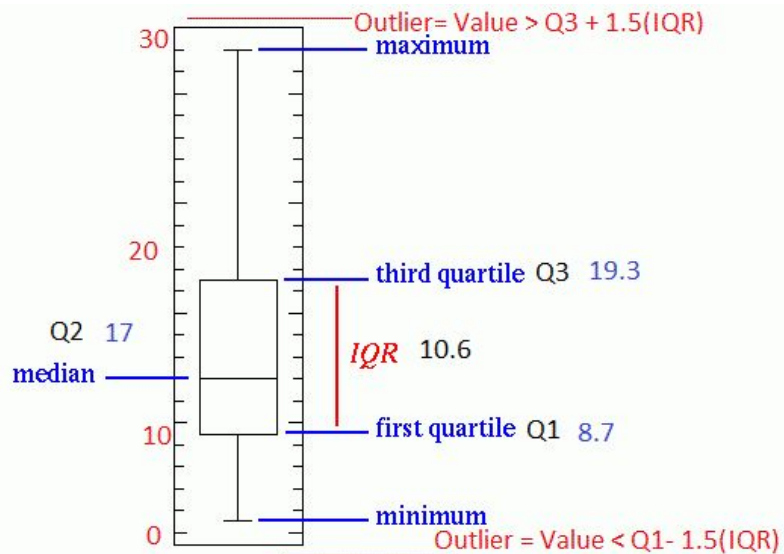
Выбросы =  $Q1 - 1,5 * IQR$  или =  $Q3 + 1,5 * IQR$





## Блочные диаграммы

Блочные диаграммы — это графические представления, которые обычно используются для отображения распределения набора данных и его сводной статистики. Блочные диаграммы отображают медиану, квартили, диапазон и выбросы набора данных.





# Дисперсия

**Дисперсия** измеряет, насколько далеко набор чисел разбросан от их медианного или среднего значения

**Стандартное отклонение** представляет собой квадратный корень из дисперсии и обеспечивает меру отклонения данных от среднего значения

Для выборки

$$variance = s^2 = \sum \frac{(x - \bar{x})^2}{n} - 1$$

$$standard\ deviation = s = \sqrt{s^2}$$

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$$

Для генеральной совокупности

$$variance = \sigma^2 = \sum \frac{(x - \bar{x})^2}{n} - 1$$

$$standard\ deviation = \sigma = \sqrt{\sigma^2}$$

$$\sigma^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n}$$





## Как рассчитать дисперсию

1. Вычислите среднее значение  $\bar{x}$ .
1. Вычтите среднее из каждого наблюдения.  $X - \bar{x}$
1. Возведите в квадрат каждое из полученных наблюдений.  $(X - \bar{x})^2$
1. Сложите эти квадраты результатов вместе.
1. Разделите эту сумму на количество наблюдений  $n$  (в случае совокупности), чтобы получить дисперсию  **$S^2$** . Если вы рассчитываете выборочную дисперсию, разделите на  $n-1$ .
1. Используйте положительный квадратный корень, чтобы получить стандартное отклонение  **$S$** .

$x$	$x - \bar{x}$	$(x - \bar{x})^2$
0	-15	225
24.1	9.1	82.81
5.6	-9.4	88.36
14.1	-0.9	0.81
17.2	2.2	4.84
8.7	-6.3	39.69
19.2	4.2	17.64
14.1	-0.9	0.81
27.7	12.7	161.29
15	0	0
19.3	4.3	18.49
		<b>639.74</b>






## Свойства дисперсии

-  Она всегда неотрицательная, поскольку каждый член суммы дисперсии возводится в квадрат, и поэтому результат либо положительный, либо нулевой
-  Дисперсия всегда имеет квадратные единицы

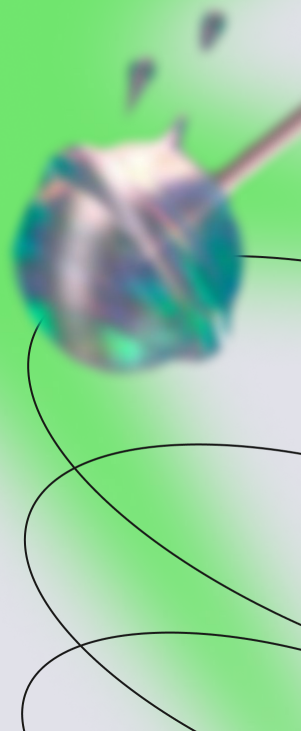


## Свойства стандартного отклонения

-  Оно описывает квадратный корень из среднего значения квадратов всех значений в наборе данных и также называется среднеквадратичным отклонением
-  Наименьшее значение стандартного отклонения равно 0, поскольку оно не может быть отрицательным
-  Когда значения данных группы схожи, стандартное отклонение будет очень низким или близким к нулю. Но когда значения данных меняются друг относительно друга, стандартное отклонение будет высоким или далеко от нуля



Распределения





# Виды распределений



## Нормальное распределение

также известное как распределение Гаусса, представляет собой непрерывное распределение вероятностей, которое часто используется для описания природных явлений, таких как рост и вес



## Биномиальное распределение

это дискретное распределение вероятностей, которое используется для моделирования количества успешных результатов в фиксированном числе независимых испытаний



## Распределение Пуассона

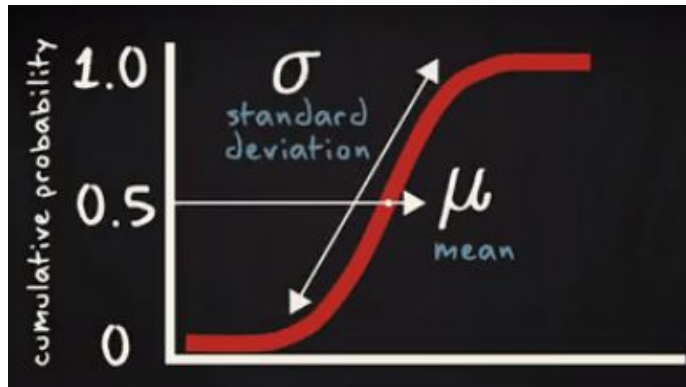
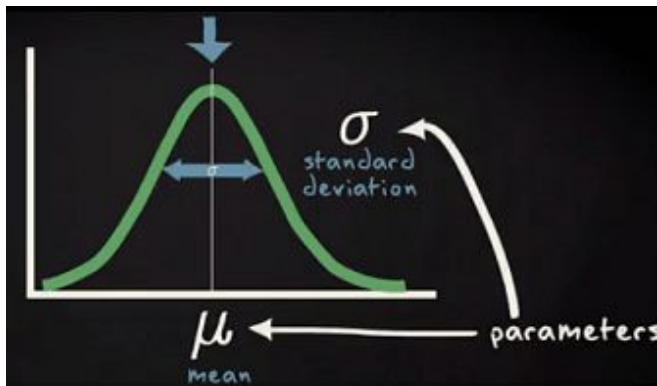
это дискретное распределение вероятностей, которое используется для моделирования количества событий, происходящих за фиксированный интервал времени



## Нормальное распределение

Плотность вероятности нормального распределения:





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







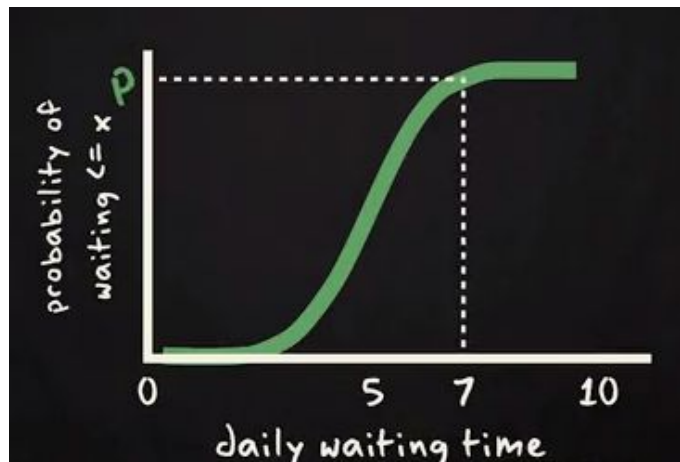
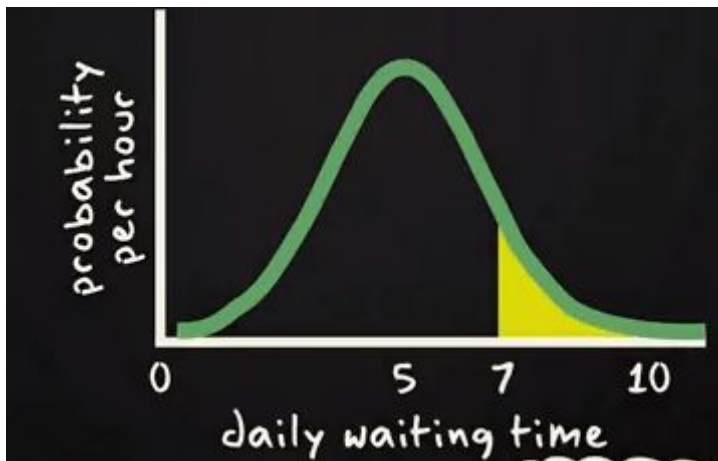
## Свойства нормального распределения

-  Среднее значение, мода и медиана равны.
-  Кривая симметрична в центре (т.е. вокруг среднего,  $\mu$ ).
-  Ровно половина значений находится слева от центра и ровно половина значений — справа
-  Общая площадь под кривой равна 1



## Расчет вероятности нормального распределения

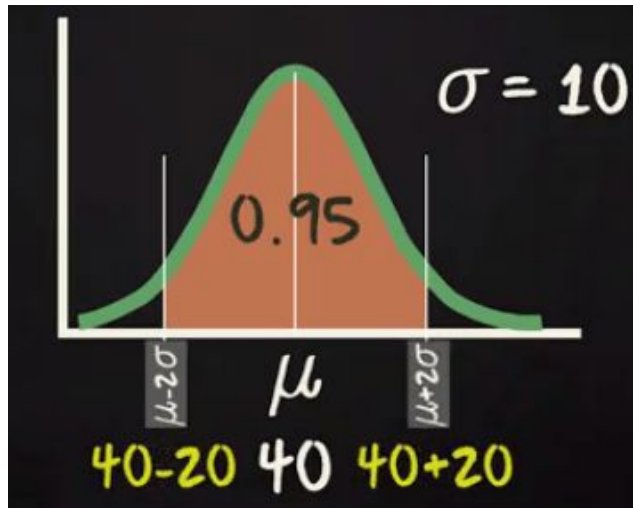
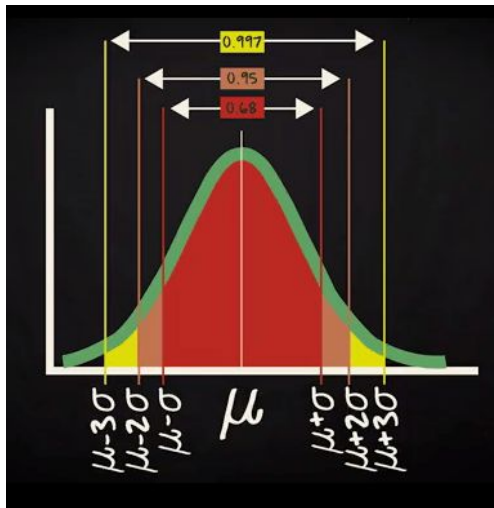
Функция плотности вероятности или p.d.f указывает вероятность на единицу случайной величины





## Распределение в форме колокола и эмпирическое правило

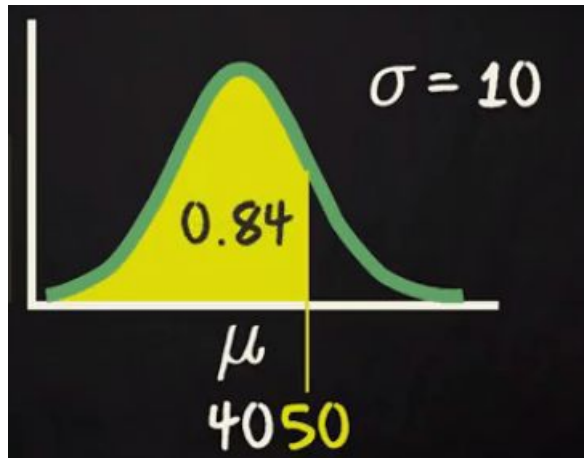
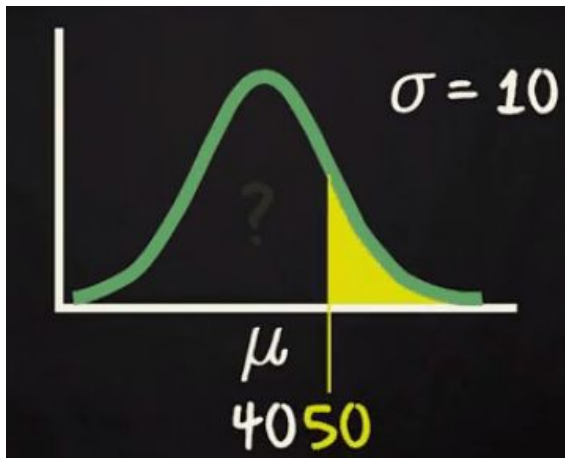
Если распределение имеет форму колокола, то предполагается, что около 68% элементов имеют z-показатель от -1 до 1; около 95% имеют z-показатель от -2 до 2; и около 99% имеют z-показатель от -3 до 3





## Распределение в форме колокола и эмпирическое правило

**Еще один вопрос, на который вы хотите ответить:** какова вероятность того, что вы будете путешествовать более 50 минут?





## Испытание Бернулли и биномиальное распределение

Value of X	$x_1$	$x_2$	$x_3$	...	$x_k$
Probability	$p_1$	$p_2$	$p_3$	...	$p_k$

Испытание Бернулли (или биномиальное испытание) — это случайный эксперимент с ровно двумя возможными исходами, «успехом» и «неудачей», в котором вероятность успеха одинакова при каждом проведении эксперимента .

- Событие (или испытание) приводит только к одному из двух взаимоисключающих исходов – успех/неудача.
- Вероятность успеха известна,  $P(\text{success}) = \pi$



## Биномиальное распределение

Распределение называется биномиальным, если выполняются следующие условия.

1. Каждое испытание имеет бинарный результат (один из двух результатов помечен как «успех»).
2. Вероятность успеха известна и постоянна для всех испытаний.
3. Количество испытаний указано
4. Испытания независимы. То есть результат одного испытания не влияет на результат последующих испытаний.

$$P(X) = \frac{n!}{X!(n-X)!} \pi^X (1-\pi)^{n-X}$$



## Пример биномиального распределения

Какова вероятность того, что при 6 подбрасываниях монеты выпадет 2 орла?

- Успех = «орел»
- $n = 6$  испытаний
- $\pi = 0,5$
- $X$  = количество орлов в 6 бросках, здесь 2.
- $X$  имеет биномиальное распределение с  $n = 6$  и  $\pi = 0,5$ .
- $X \sim B(6, 0,5)$

$$P(X = 2) = \frac{6!}{2!(6-2)!} 0.5^2 (1-0.5)^{6-2} = 15 * 0.5^6 = 0.234$$



## Распределение Пуассона

Распределение Пуассона используется для определения вероятности того, что количество событий произойдет за определенное время или пространство

Подобно биномиальному распределению и нормальному распределению, существует множество распределений Пуассона.

Каждое распределение Пуассона определяется средней скоростью, с которой происходит событие.

Скорость обозначается  $\lambda$

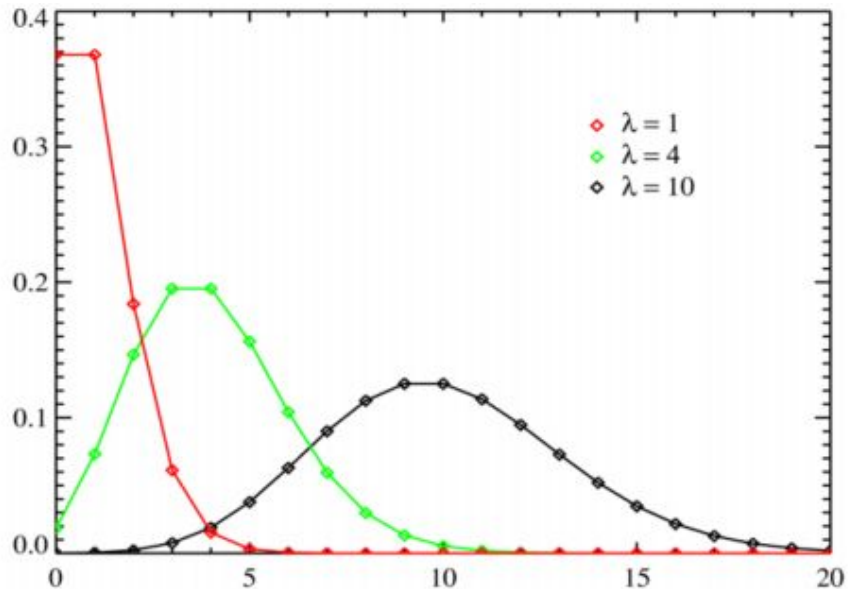
$\lambda$  = «лямбда», греческая буква «L» — для распределения Пуассона есть только один параметр

$$P(X) = \frac{\lambda^x e^{-\lambda}}{X!}$$





## Распределение Пуассона



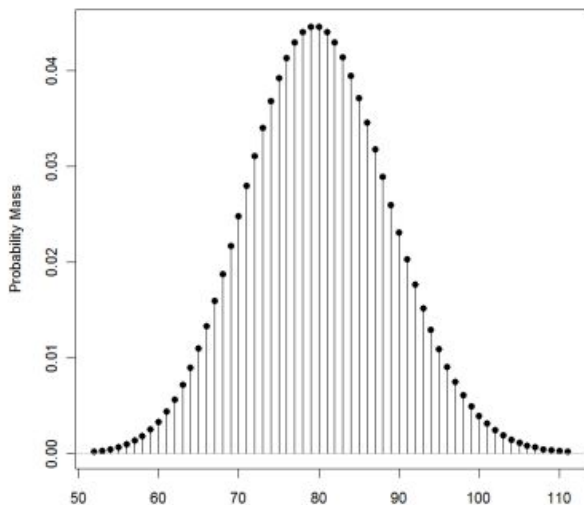
- Если  $\lambda$  равно 10 или больше, нормальное распределение является разумным приближением к распределению Пуассона
- Среднее значение и дисперсия для распределения Пуассона одинаковы и равны  $\lambda$
- Стандартное отклонение распределения Пуассона равно квадратному корню из  $\lambda$



## Пример распределения Пуассона

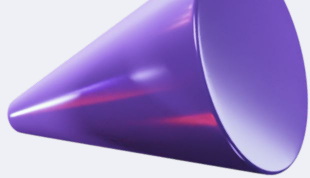
В крупную городскую больницу каждый понедельник поступает в среднем 80 отделений неотложной помощи. Какова вероятность того, что их будет больше 100?

Если мы положим  $\lambda = 80$  и  $x = 100$ , то мы получим значение вероятности как 0,01316885.



- $\lambda$  — скорость допуска / день в понедельник = 80
- мы можем использовать нормальное приближение, так как  $\lambda > 10$

Нормальное приближение имеет среднее значение = 80 и SD = 8,94 (квадратный корень из 80 = 8,94).



**Спасибо за внимание**

