

Элементы статистики для построения гипотез A/B-тестирования

Применение базовых инструментов статистики для оценки гипотезы

Применение базовых инструментов
статистики для оценки гипотезы

Цель урока

Понять, как применять базовые
элементы статистики к данным.

Задачи урока

- ✓ вспомнить, что такое распределения и их виды
- ✓ рассмотреть виды визуализации распределений
- ✓ узнать про базовые статистики и их различия

Распределения

Дискретное

(конечное количество
возможных значений)

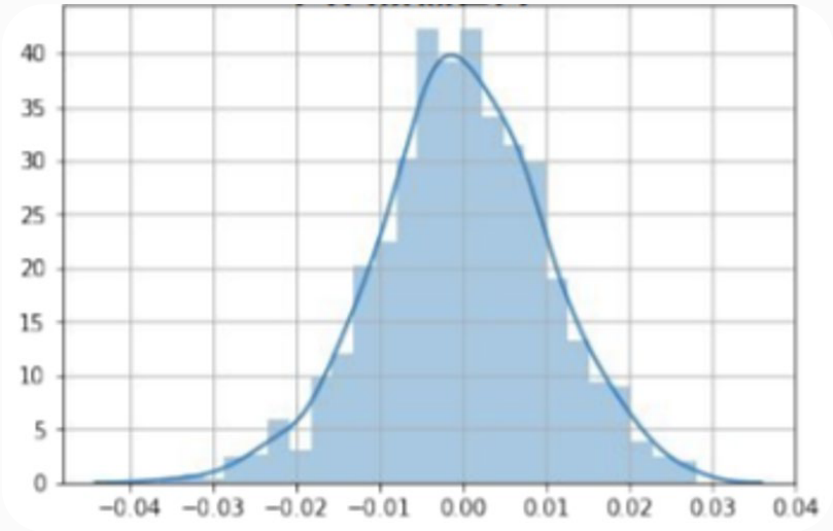
Распределение — закон,
определяющий вероятность
для каждого возможного значения.

X	0	1
$P(X = x)$	0,9	0,1

Непрерывное

(бесконечное количество
возможных значений)

Распределение — функция,
характеризующая вероятность
реализации тех или иных
значений случайной
переменной.



Применение базовых инструментов
статистики для оценки гипотезы

Визуализация распределений

Существует два популярных
и простых способа:

- ✓ Боксплот
- ✓ Гистограмма

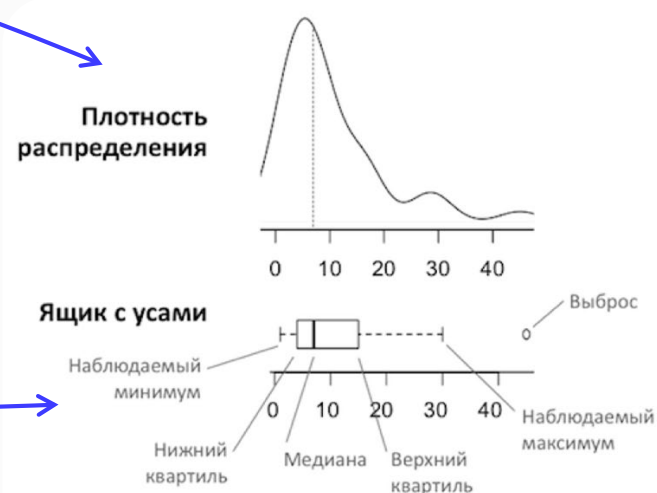
Визуализация распределений

Существует два популярных
и простых способа:

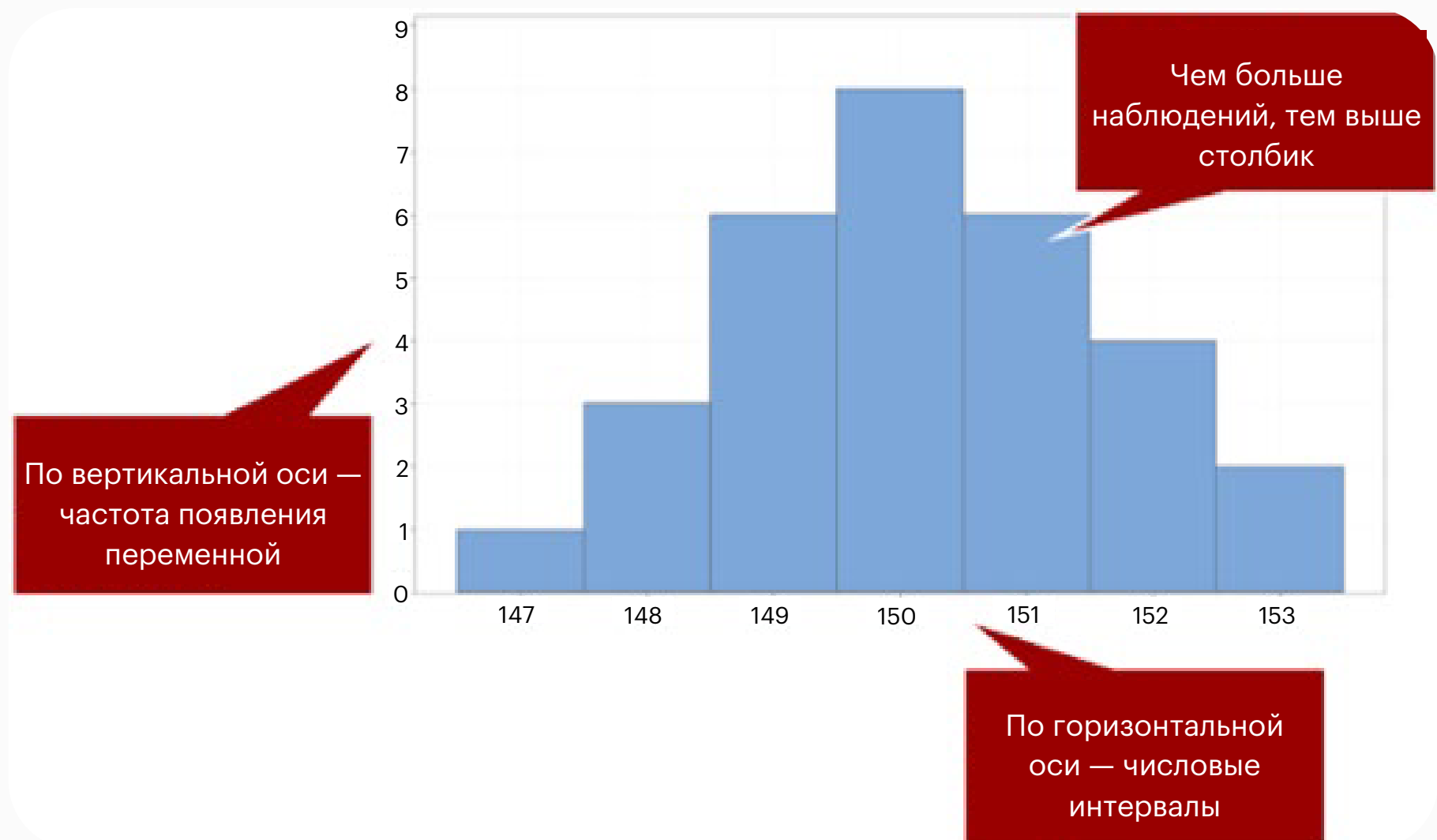
Сопоставление графика
плотности распределения
и ящика с усами

✓ Боксплот

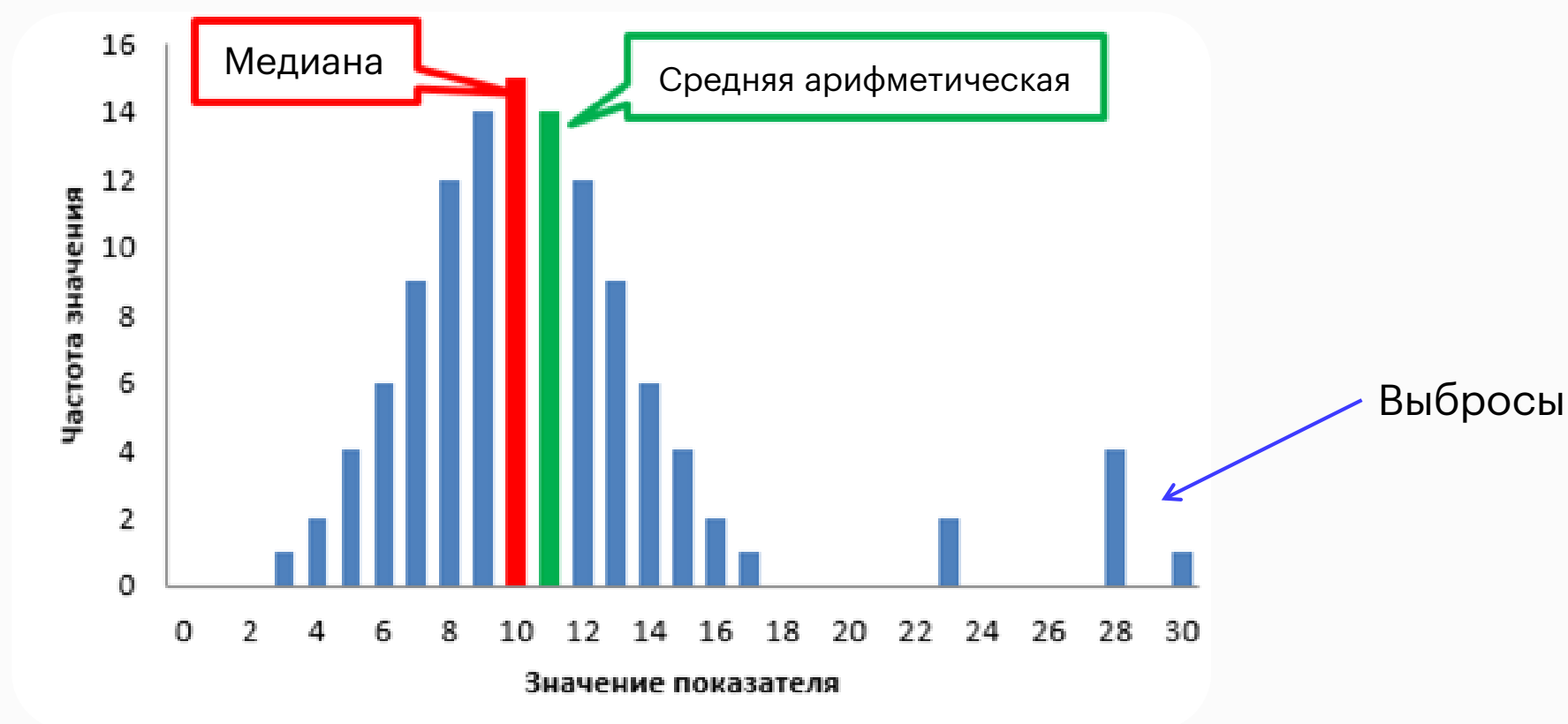
✓ Гистограмма



Гистограмма



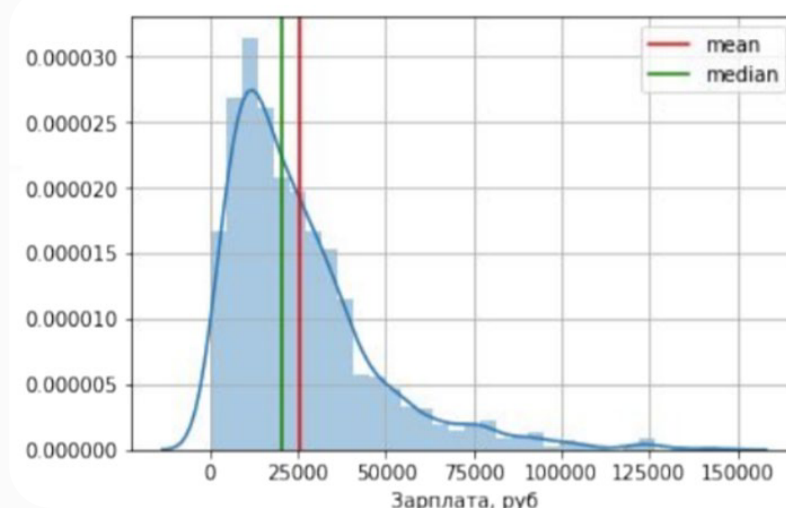
Теперь рассмотрим пару важных статистик



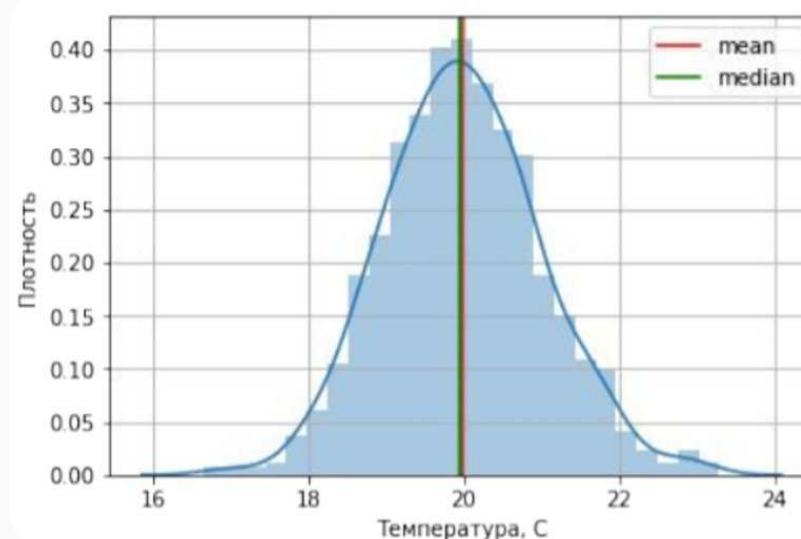
Медиана — это точка, слева и справа от которой лежит равное количество точек. Ключевая её особенность в том, что она слабо смещается из-за выбросов, в отличие от среднего.

Кейс: среднее или медиана?

✓ Общий уровень
зарплата в стране



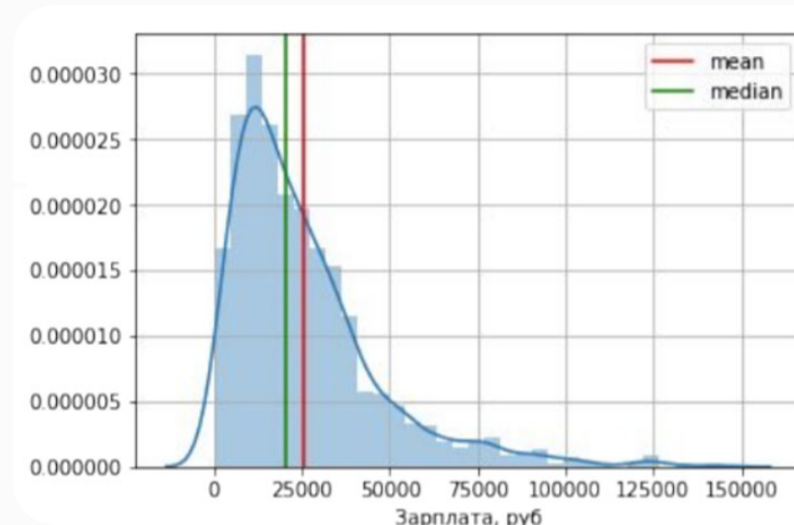
✓ Ожидаемая
температура в июне



Кейс: среднее или медиана?

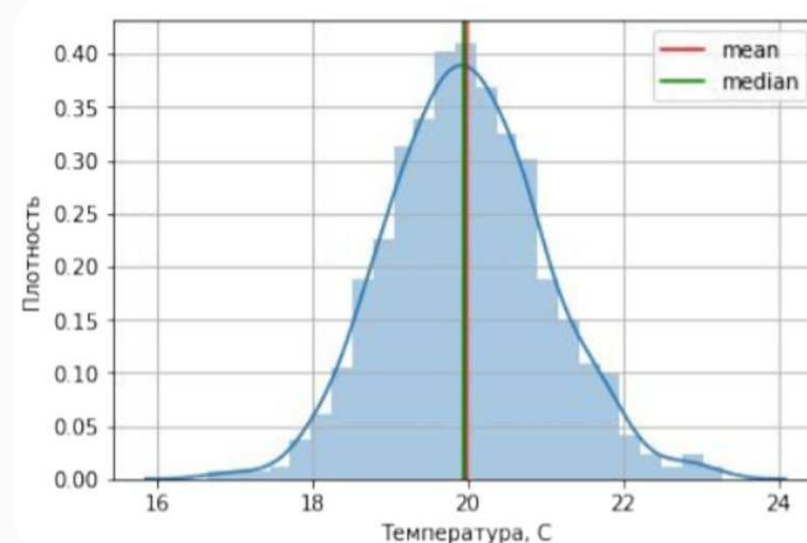
- ✓ Общий уровень
зарплат в стране

Ответ: медиана

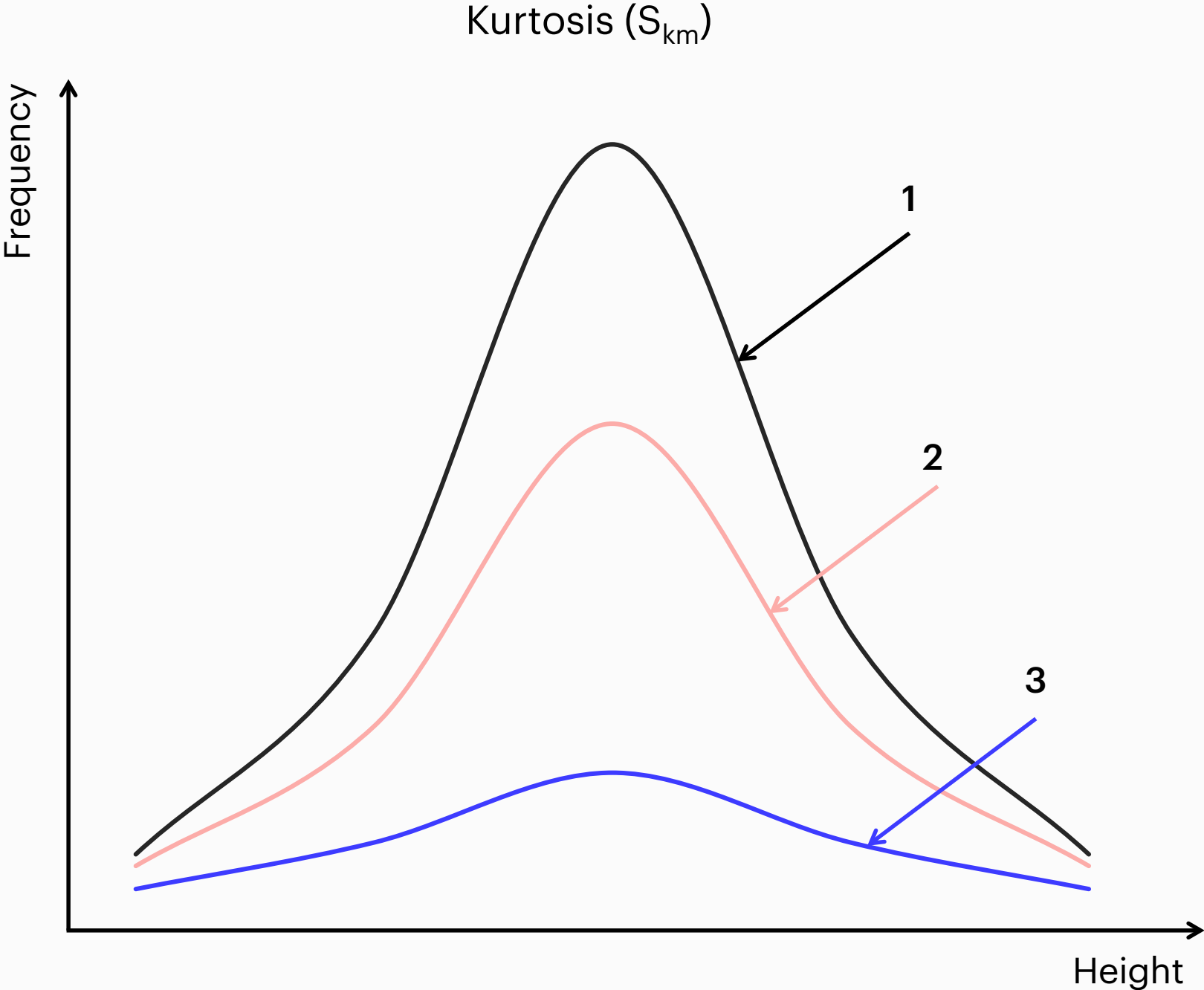


- ✓ Ожидаемая
температура в июне

Ответ: и то, и то



Применение базовых инструментов
статистики для оценки гипотезы



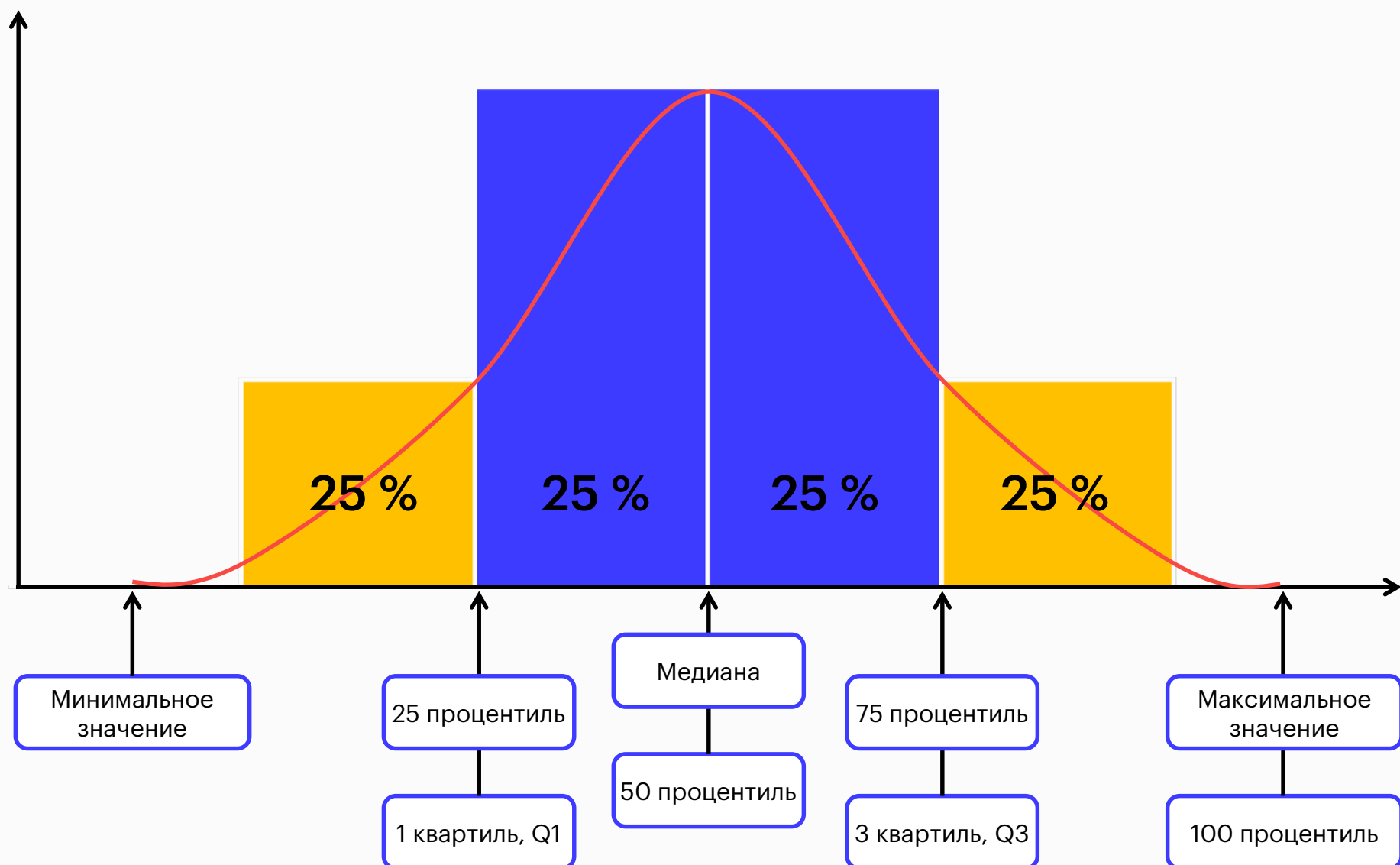
Важное понятие для понимания боксплота — квантиль

Квантиль — значение, которое заданная случайная величина не превышает с фиксированной вероятностью.

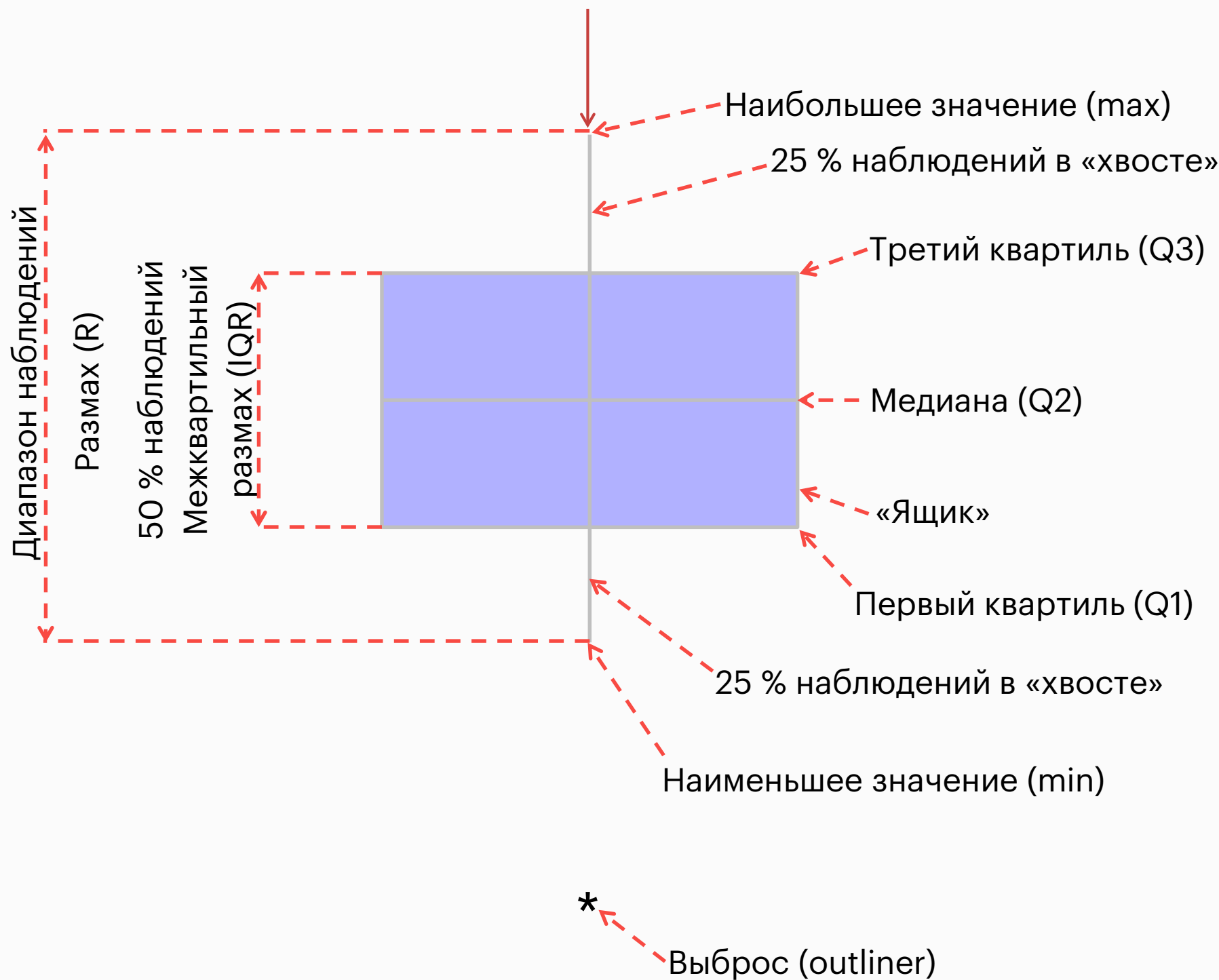
Пример: 80 %-ный квантиль выхода из строя станков на заводе = 3 года.

Это значит, что 80 % станков выходят из строя в течение первых трёх лет, остальные 20 % станков работают больше 3 лет.

Важное понятие для понимания боксплота — квантиль



Ось значений



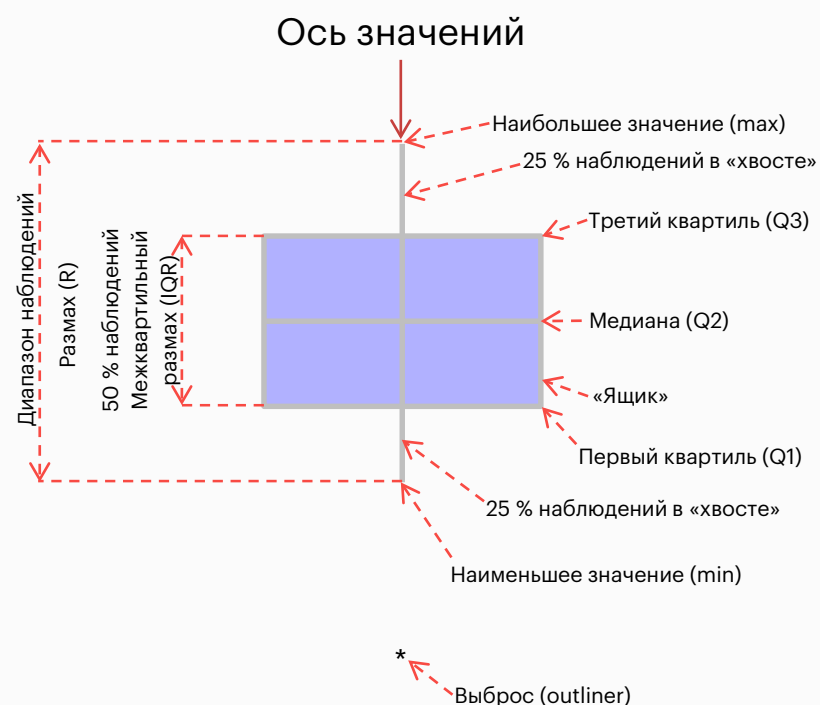
Боксплот («Ящик с усами»)

Зачем?

- Легко читаются основные статистики
- Чётко видно, где выбросы
- Хорошо оценивается степень асимметрии и разброса данных
- Бесценен, когда мало данных

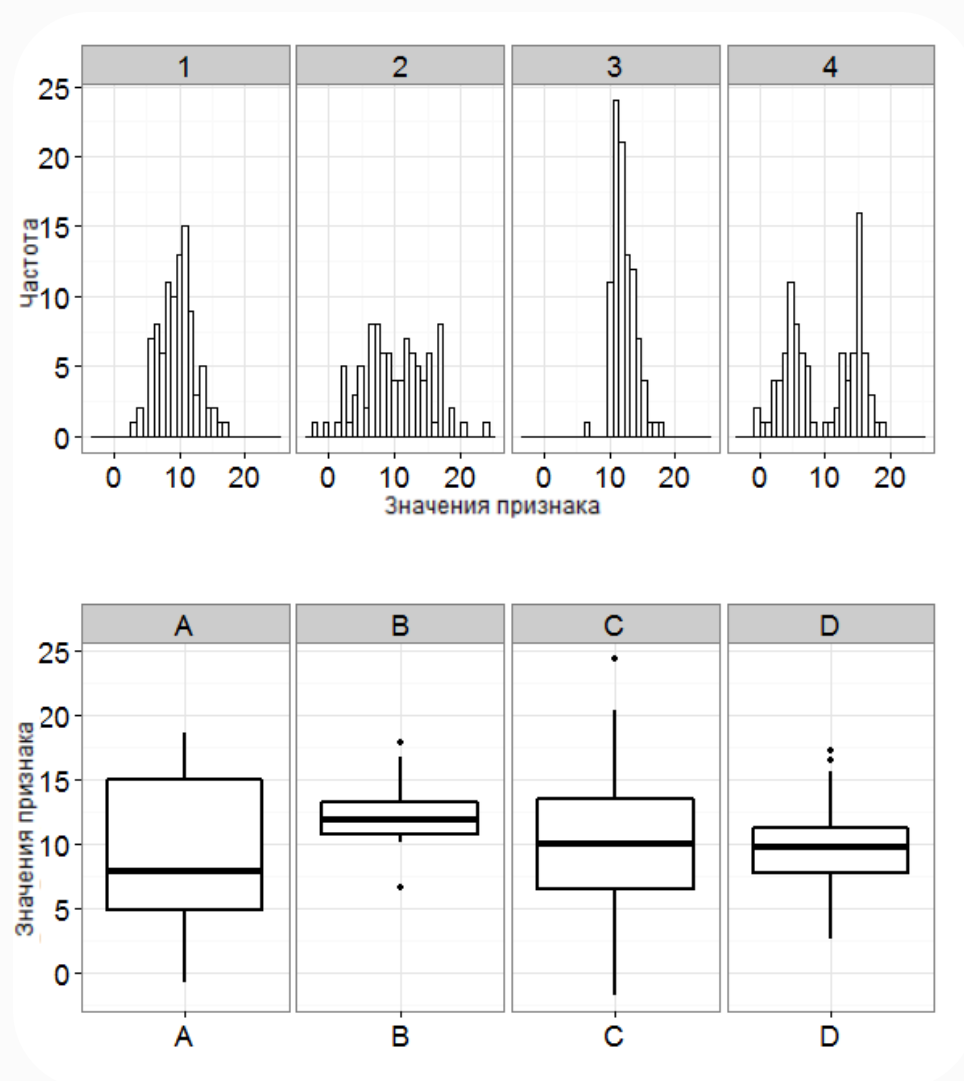
Есть много видов боксплотов, поэтому необходимо уточнять, что чему соответствует.

Один из видов: когда «усы» — это минимум и максимум.



Зачем?

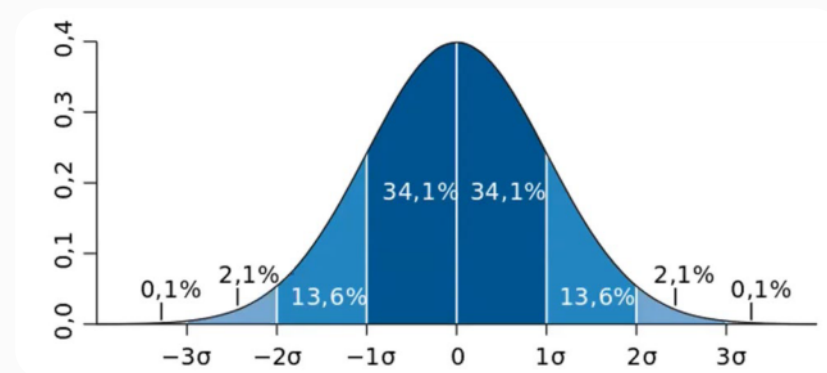
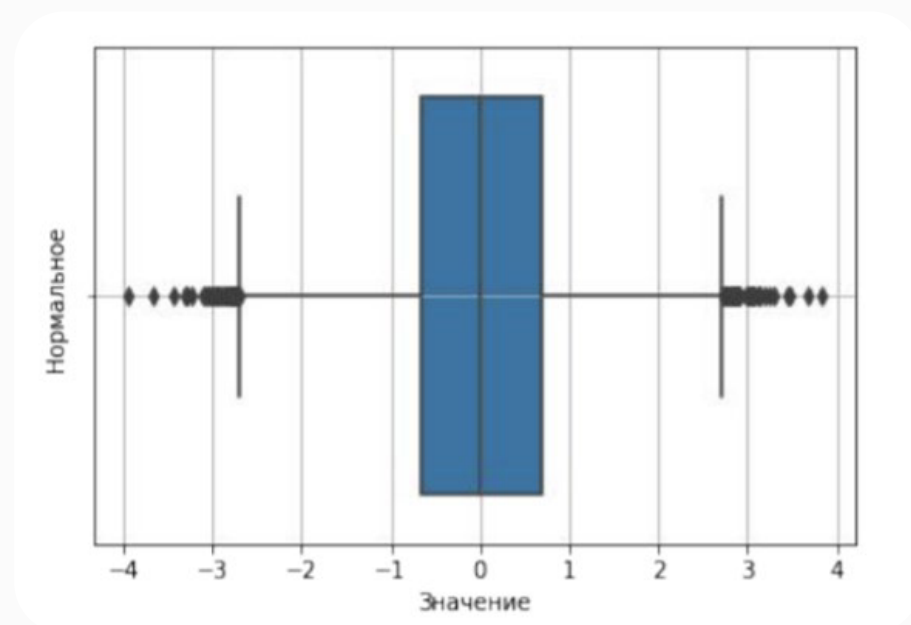
Оба способа визуализации хорошо помогают составить понимание о том, как выглядят ваши данные.



Самое популярное распределение — нормальное распределение

Основные свойства:

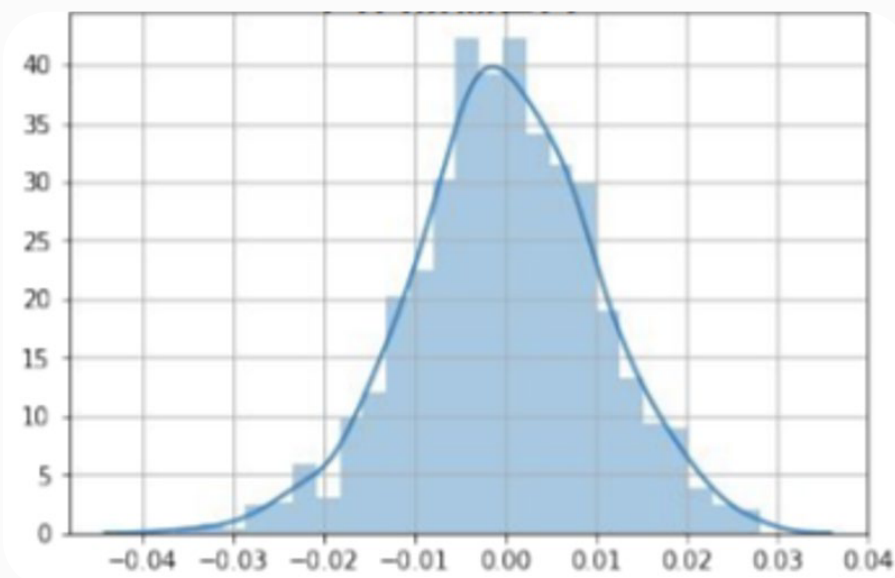
- медиана совпадает со средним
- с помощью окрестностей, кратных среднеквадратичному отклонению, можно выделить большое число наблюдений



Самое популярное распределение — нормальное распределение

Пример: время, проведённое
пользователями в приложении.

Гистограмма этого
распределения выглядит
примерно вот так:

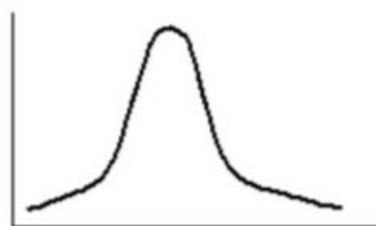


Виды распределений

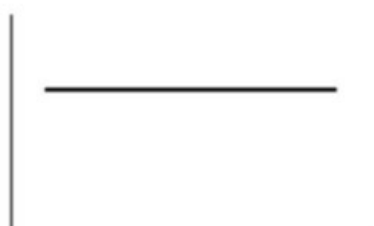
На практике важно одно

У вас распределение нормальное или нет?

Какое именно распределение, уже не так важно.



Нормальное



Равномерное



Логнормальное



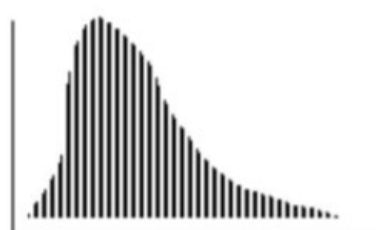
Гамма-распределение



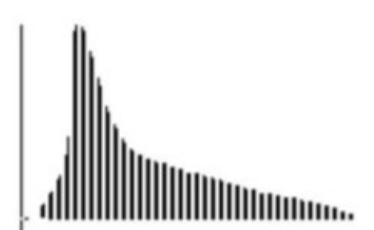
Бета



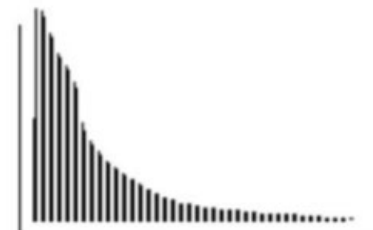
Вейбулла



Биноминальное



Пуассона



Геометрическое

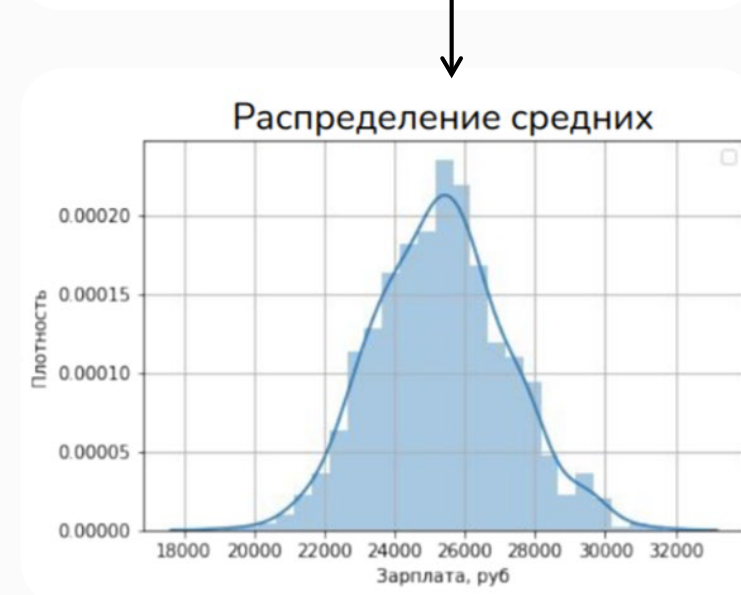
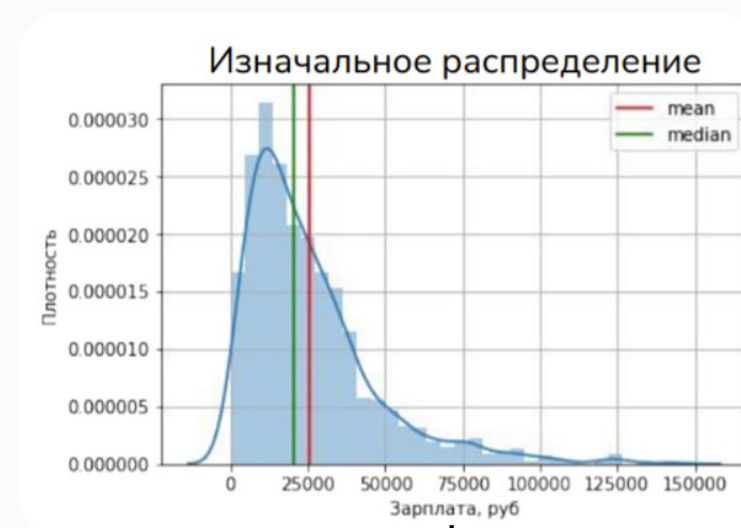
Применение базовых инструментов
статистики для оценки гипотезы

Теперь поговорим про полезные теоремы статистики

Центральная пределльная теорема (ЦПТ)

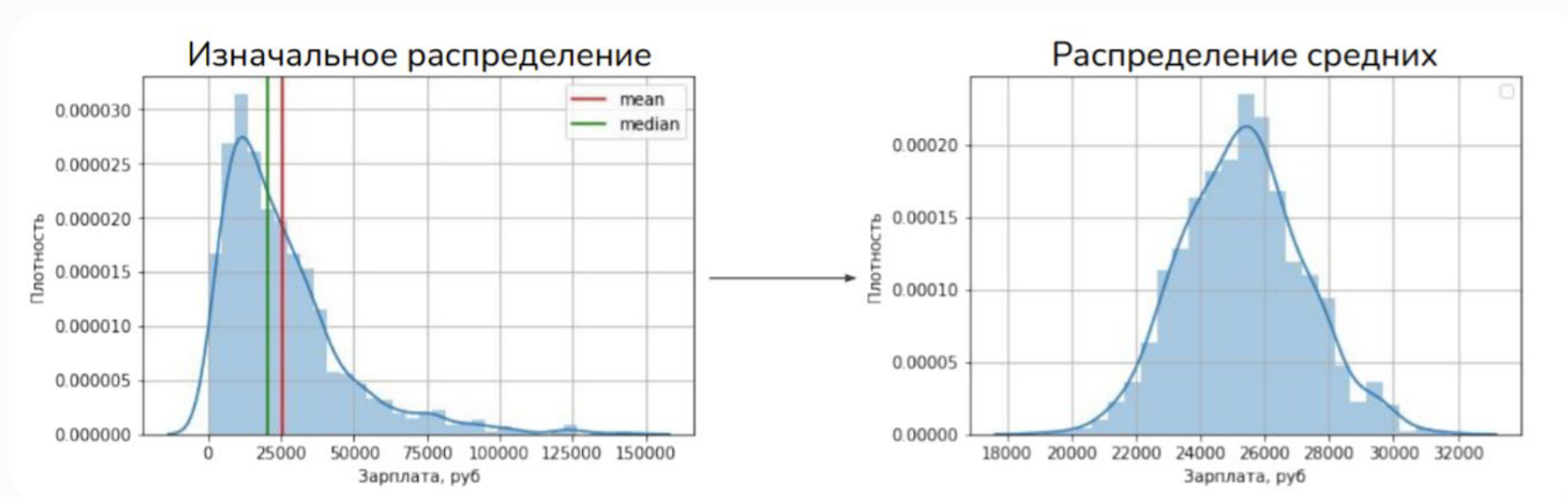
Проиллюстрируем примером:

- 1 Есть распределение, которое распределено ненормально
- 2 Берём выборки и считаем среднее в каждой
- 3 Строим распределение средних
- 4 Оно оказывается нормальным



Центральная пределльная теорема (ЦПТ)

Самое важное свойство — ЦПТ позволяет получить нормальное распределение из ненормального:



Причины невыполнения ЦПТ

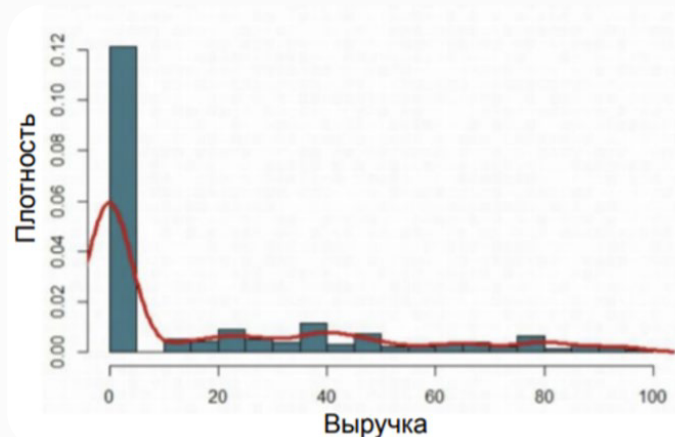
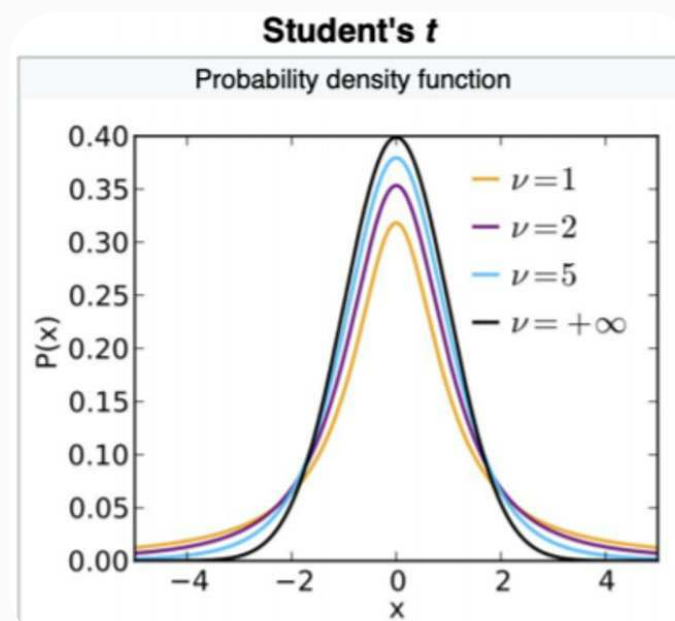
1 Распределение имеет бесконечную дисперсию

Можем наблюдать
её при «тяжёлых хвостах»

Так как часто мы не знаем
формулу распределения,
то любой из «хвостов» может
уйти в бесконечность

2 Имеются выбросы в распределении

Аномальные значения метрики



Итоги урока



Распределение бывает
дискретным и непрерывным



Самое удобное распределение
данных — нормальное



Данные визуализируем
гистограммой и боксплотом



Самые популярные
статистики — медиана,
среднее и дисперсия