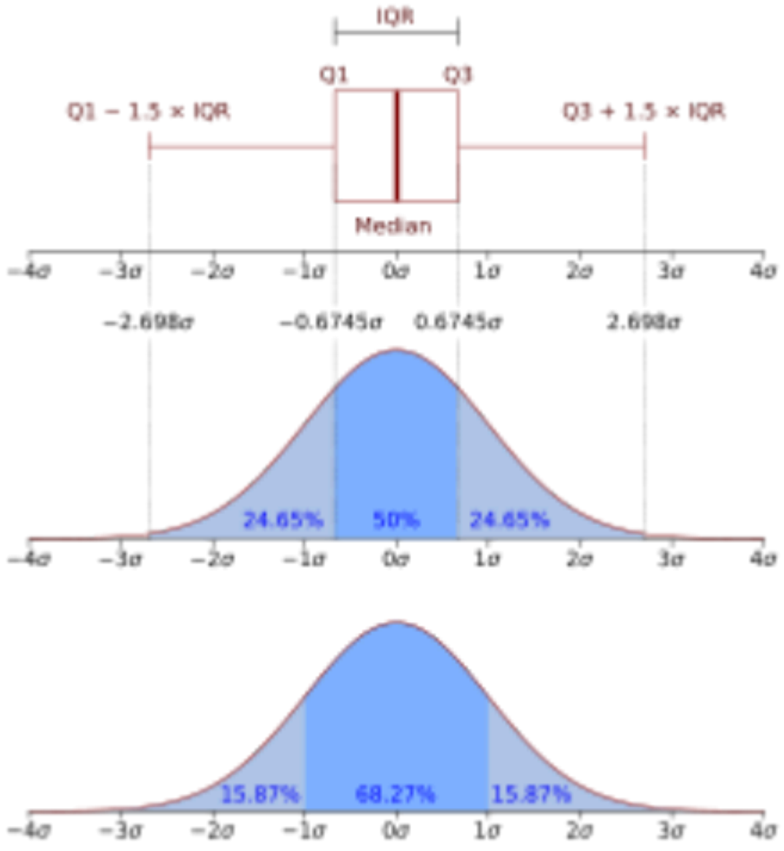
Статистика, меры разброса



Распределение, гистограмма.

Нормальное распределение.

Дисперсия, стандартное отклонение.

Коэффициент вариации.

Диаграмма размаха (ящик с усами).

Размах, межквартильный размах.

Статистический анализ - это нечто большее, чем просто набор вычислений. Не используйте формулы или программы, если не понимаете, почему вы это делаете.

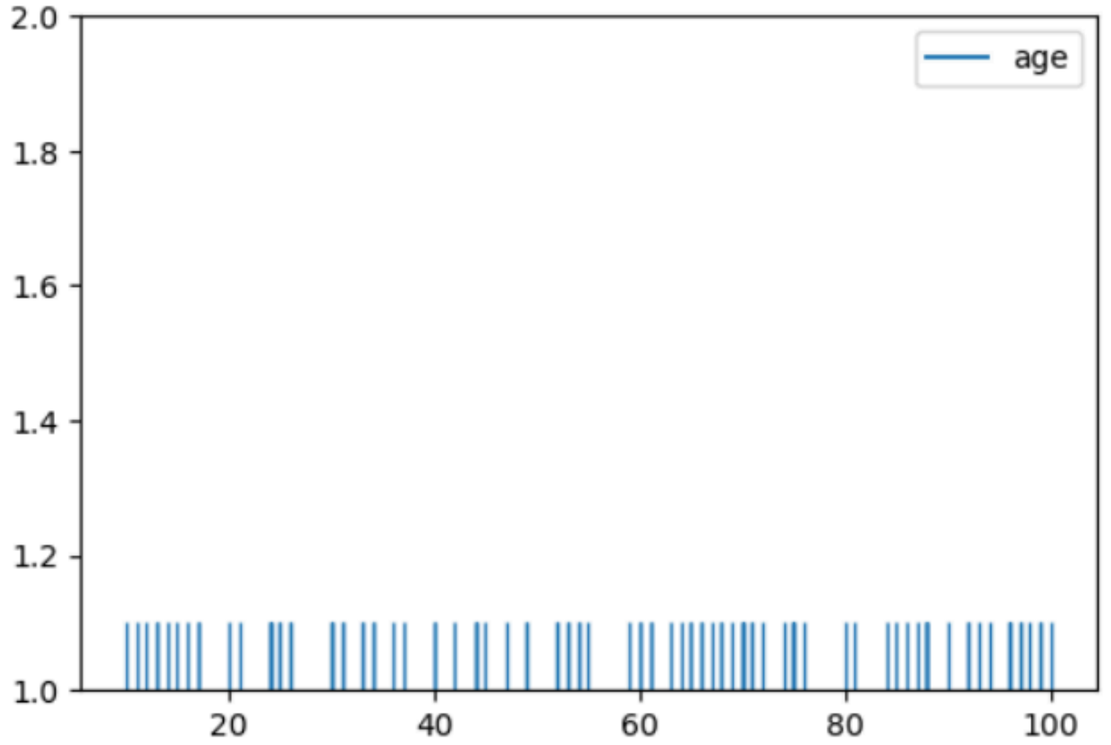
Дэвид Шпигельхалтер,
“Искусство статистики”

Распределение, гистограмма и нормальное распределение – ключевые понятия в статистике, которые помогают понять, как данные "распределяются" в выборке или генеральной совокупности. Разберем их по порядку.

Что такое распределение?

Распределение – это способ описания того, как часто различные значения встречаются в наборе данных. Проще говоря, это отображение того, какие значения встречаются чаще, а какие реже.

index	category	age_count
1	До 20 лет	11
2	От 21 до 40 лет	22
3	От 41 до 60 лет	18
4	От 61 до 80 лет	26
5	Более 80 лет	23

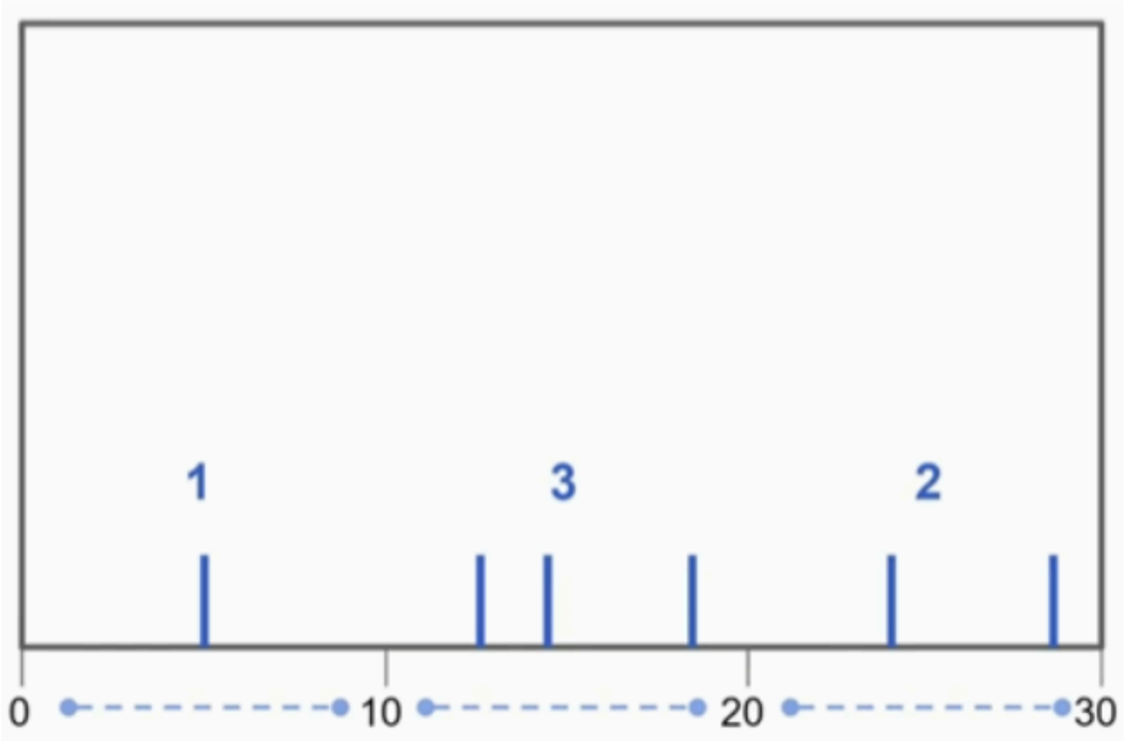


Что такое гистограмма?

Перед тем как работать с гистограммой, давайте сначала разберем, как она создается. Представим некую переменную, для которой мы построили график **rug plot** (ковровая диаграмма).

Здесь всего 6 наблюдений или же шесть точек:

одна точке в интервале от 0 до 10.
три точки в интервале от 11 до 20.
две точки в интервале от 21 до 30.

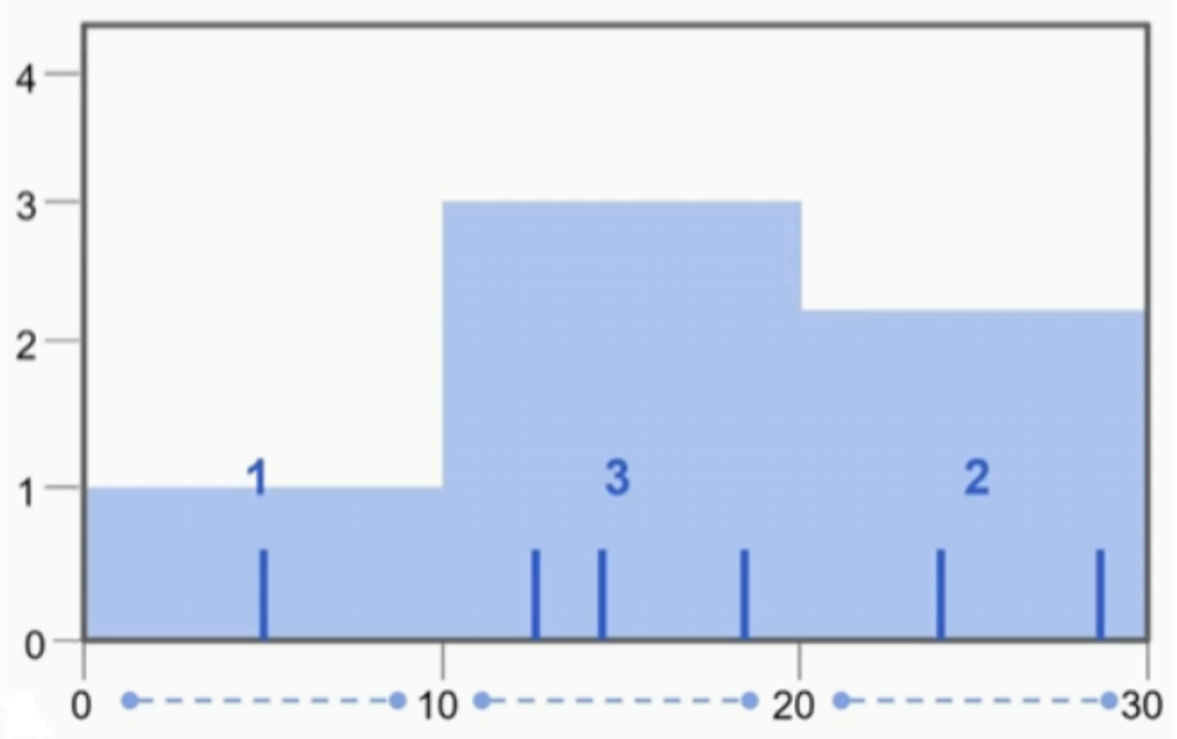


Интервалы, так же называются “бинами” (bins), в нашей диаграмме сейчас три бина.

Гистограмма

Гистограмма – это визуальное представление распределения данных. Она состоит из прямоугольников (столбиков), высота каждого из которых показывает, сколько раз данные встречаются в определённом диапазоне (интервала).

Если для каждого интервала нарисовать прямоугольник, высота которого будет равна количеству наблюдений в этом интервале, мы получим гистограмму.



То есть по **оси X** располагаются значения, а высота каждого столбца равна соответствующему значению на **оси Y**.

Нормальное распределение

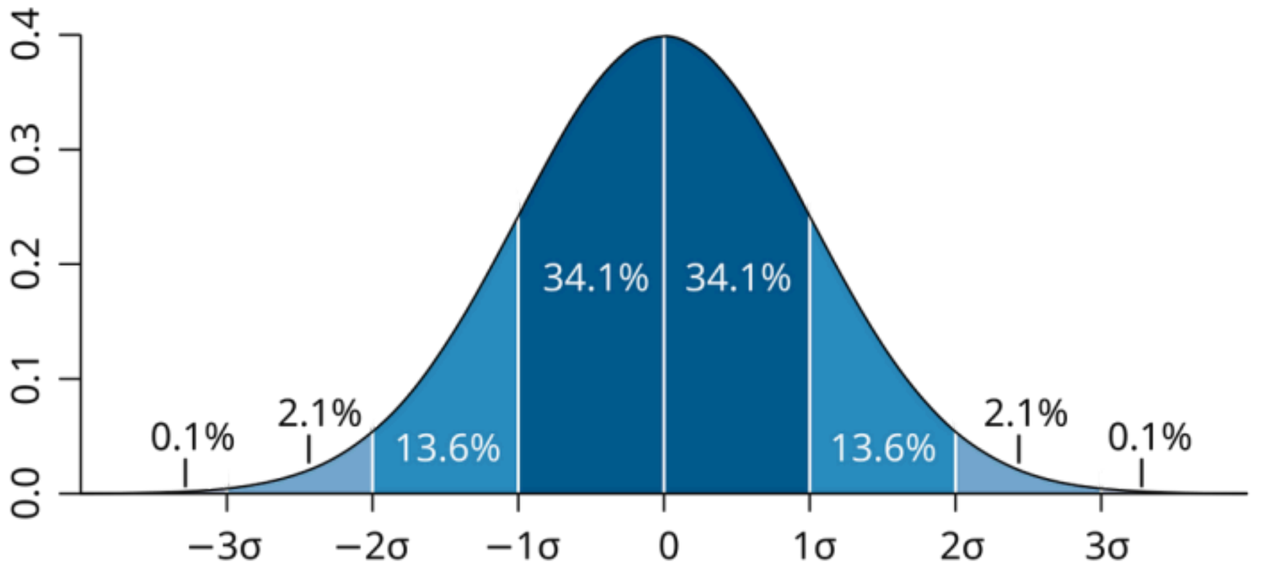
В статистике часто используют распределения, чтобы понять, как часто происходят разные события. По сути, распределение — это связь между значением величины и вероятностью того, что она примет это значение.

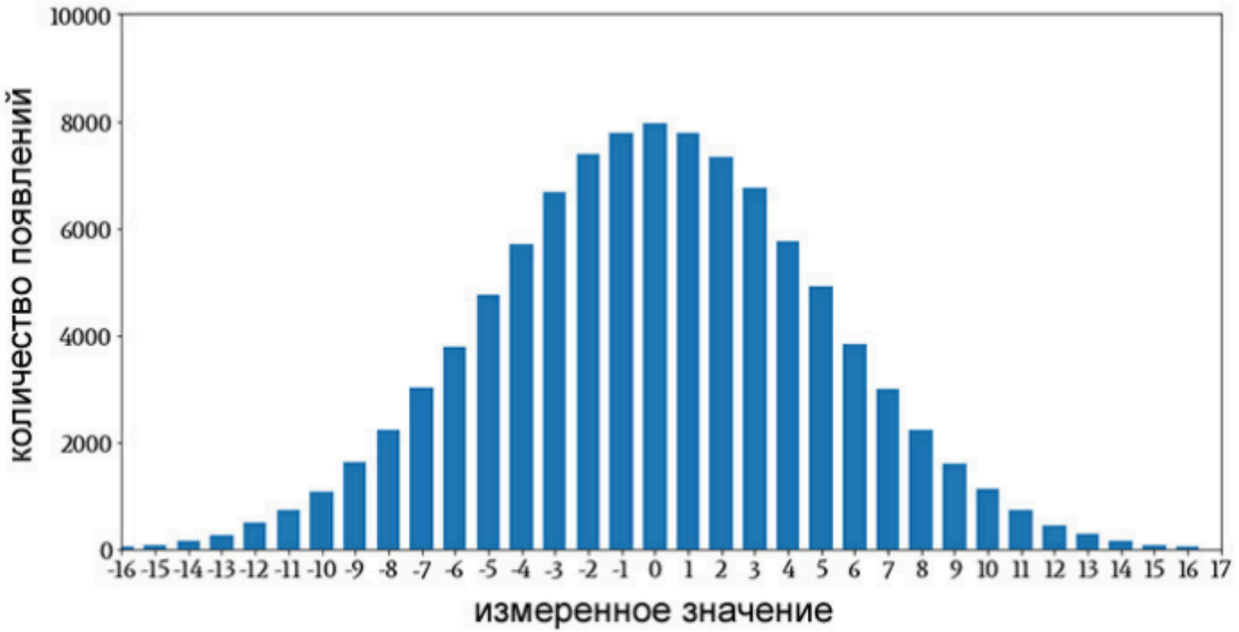
Нормальное распределение — это особый тип распределения, при котором большинство значений сосредоточено около среднего. Его также называют распределением Гаусса или колоколообразным распределением.

Пример нормального распределения – распределение роста людей. Большинство имеют рост, близкий к среднему, а очень высокие или очень низкие люди встречаются реже.

Как понять что распределение “нормальное”?

В большинстве случаев достаточно убедиться, что оно симметричной формы, а среднеарифметическое близко к моде и медиане.


0.0
0.1
0.2
0.3
0.4
0.1%
2.1%
13.6%
34.1%
34.1%
13.6%
2.1%
0.1%
−3σ
−2σ
−1σ
0
1σ
2σ
3σ



Меры изменчивости

Чтобы лучше познакомиться с нормальным распределением, нам нужно изучить две меры изменчивости: Дисперсия и Среднеквадратичное отклонение (или Стандартное отклонение, но терминология может различаться в зависимости от контекста).

Разберем данные термины, на основе следующих данных:

Возраст = [22, 28, 37, 30, 43]

Начнем с Дисперсии. Определение дисперсии звучит так:

Дисперсия — это среднее арифметическое квадратов отклонений значений от среднего.

Чтобы найти дисперсию, последовательно проведите следующие вычисления:

Сначала найдём среднее значение:

$$\text{Средние арифметическое} = \frac{22 + 28 + 37 + 30 + 43}{5} = 32$$

Теперь нужно определить отклонение каждого значения от нашего среднего:

1) 22 - 32 = -10

2) 28 - 32 = -4

3) 37 - 32 = 5

4) 30 - 32 = -2

5) 43 - 32 = 11

Наконец, чтобы вычислить дисперсию, каждую из полученных разностей возводим в квадрат, а затем находим среднее арифметическое этих результатов:

$$\text{Дисперсия} = \frac{(-10)^2 + (-4)^2 + 5^2 + (-2)^2 + 11^2}{5} = 53.44$$

Среднеквадратическое отклонение

Так как же теперь вычислить среднеквадратическое отклонение, зная дисперсию? Нужно извлечь квадратный корень из дисперсии. То есть среднеквадратическое отклонение равно:

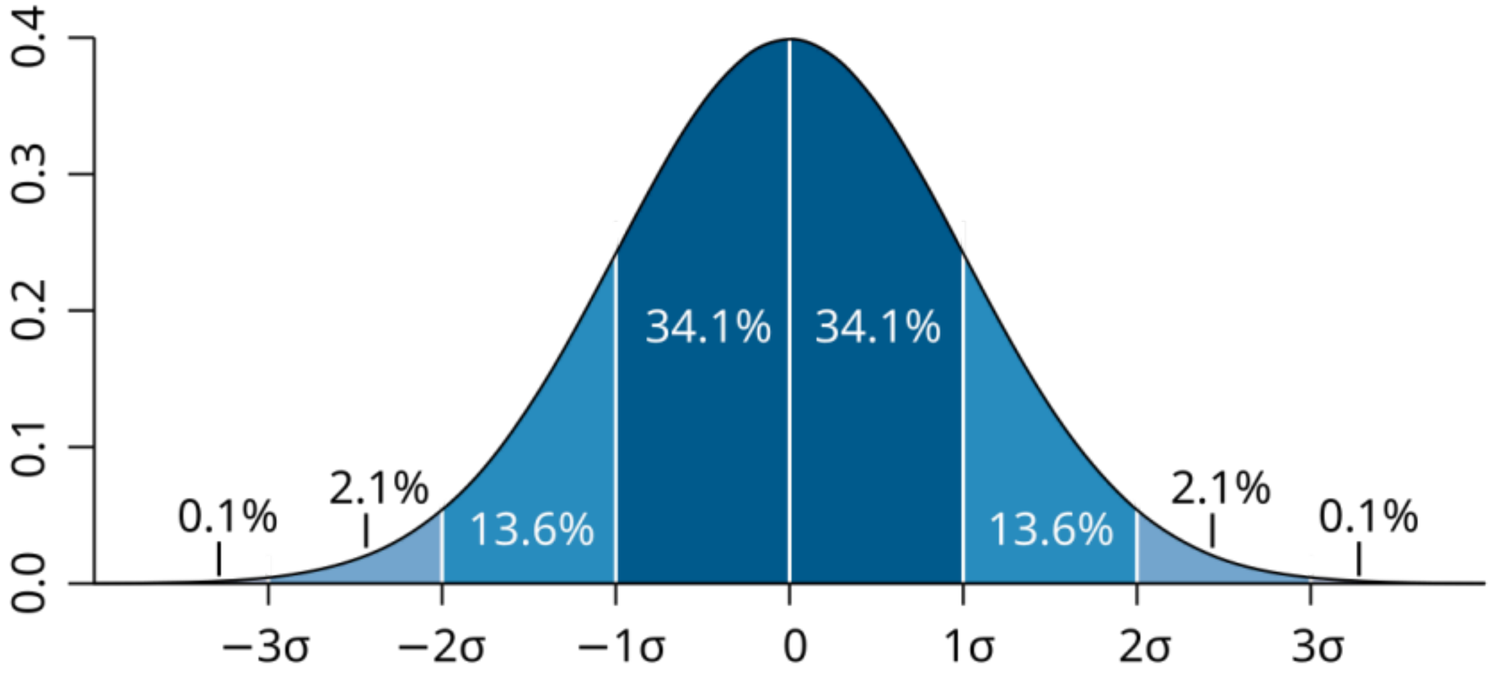
$$\sqrt{53.44} \approx 7.3$$

Среднеквадратичное отклонение несет полезную информацию. Теперь мы можем определить, какие из полученных результатов измерения возраста лежат в пределах интервала, который мы получим, откладывая от среднего значения (в обе стороны от него) среднеквадратическое отклонение.

Среднеквадратическое отклонение (Стандартное отклонение) — это мера изменчивости или разброса данных относительно их среднего значения. Оно показывает, насколько типичное значение данных отклоняется от среднего.

Стандартное отклонение

В нормальном распределении большинство значений находятся в пределах одного стандартного отклонения от среднего (68,2%), двух стандартных отклонений (95,4%), трёх стандартных отклонений (99,7%) и так далее.



Из этого следует, что вероятность встретить в выборке значение, которое отличается от среднего более чем на три стандартных отклонения, составляет менее (0,3%)!

Коэффициент вариации

Коэффициент вариации — это относительный показатель вариации. Он рассчитывается как отношение стандартного отклонения к среднему

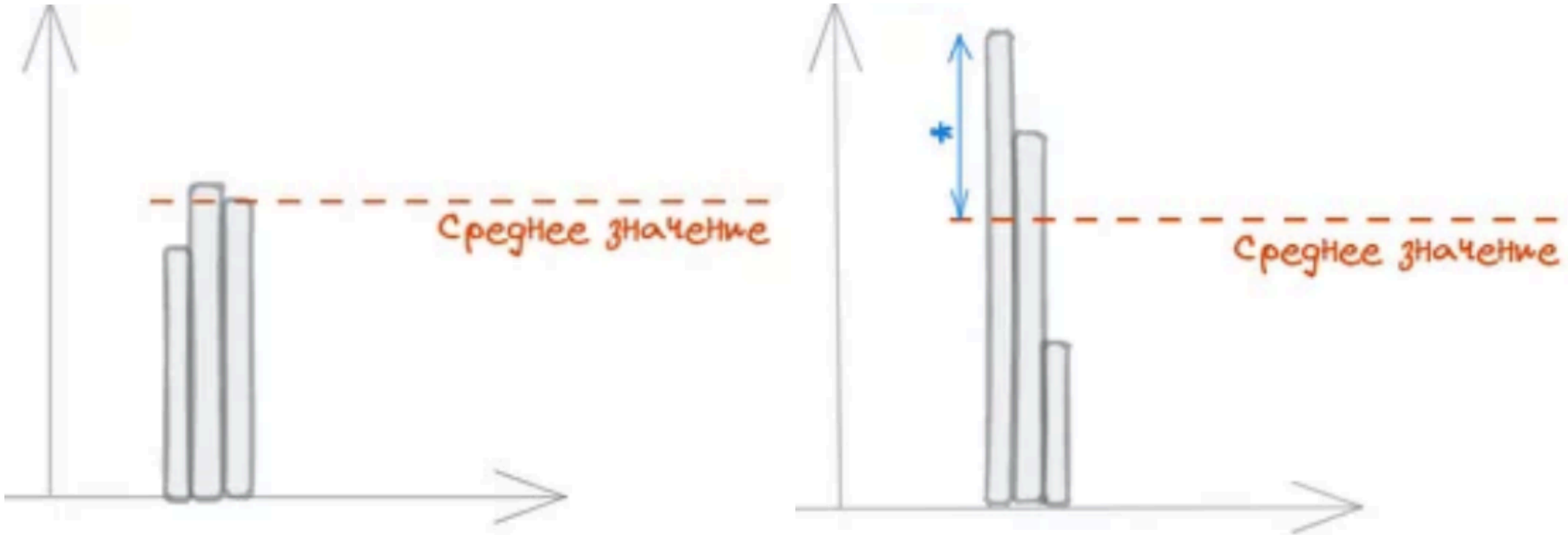
значению. Этот коэффициент подходит для сравнения выборок с разными единицами измерения.

Если средний рост россиян составляет 170 см, а стандартное отклонение — 10 см, то коэффициент вариации будет равен **10/170 = 5,9%**. Это означает, что рост россиян в среднем отклоняется от среднего на **6%**.

Коэффициент вариации

Низкая вариация - рост отдельных элементов похож друг на друга и на средний

Высокая вариация - рост отдельных элементов отличается и не похож на средний



Ассиметричное распределение

Распределение не всегда бывает нормальным.

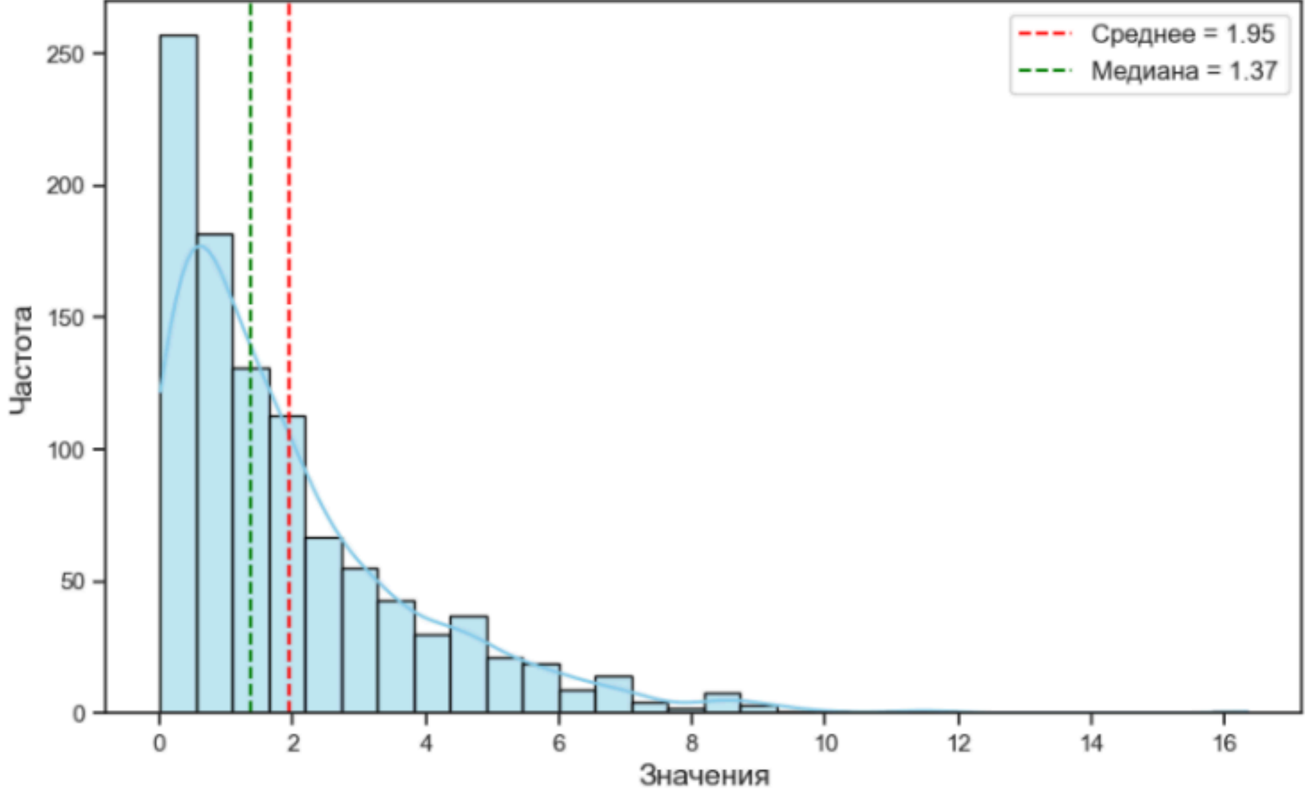
Асимметрия распределения возникает, когда какие-либо факторы действуют сильнее в одном направлении, чем в другом, или когда процесс развития явления подчиняется одной доминирующей причине. Кроме того, природа некоторых явлений такова, что они имеют асимметричное распределение.

Правостороннее (положительный перекос):

В этом случае хвост распределения тянется вправо, а большинство значений сосредоточено на левом конце.

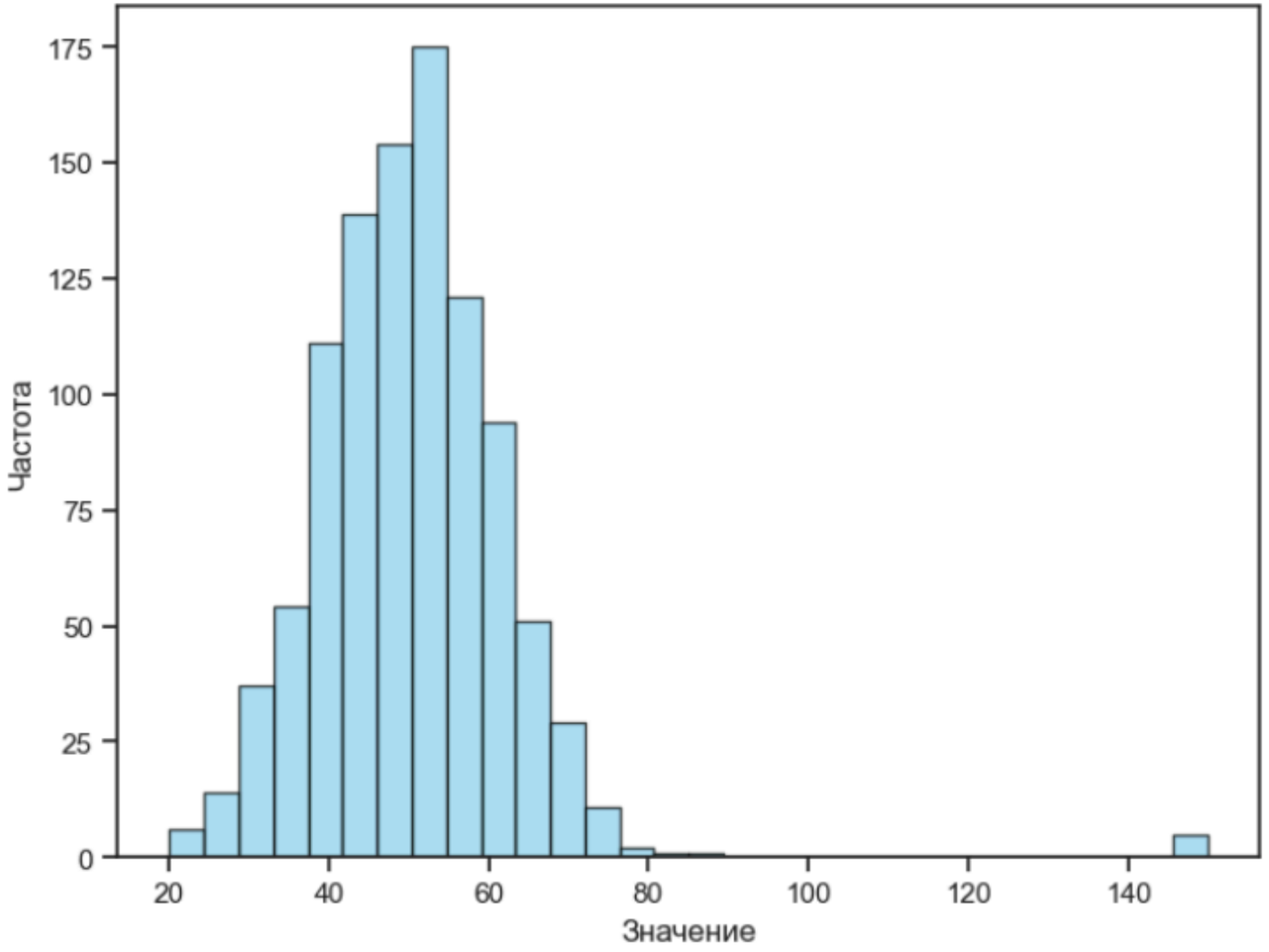
Левостороннее (отрицательный перекос):

В этом случае хвост распределения тянется влево, а большинство значений сосредоточено на правом конце.



Выбросы

Выбросы — это те значения, которые значительно отличаются от остальных. Например, если большинство значений в наборе данных лежат в диапазоне от 10 до 100, а одно значение составляет 180, то это будет выброс.

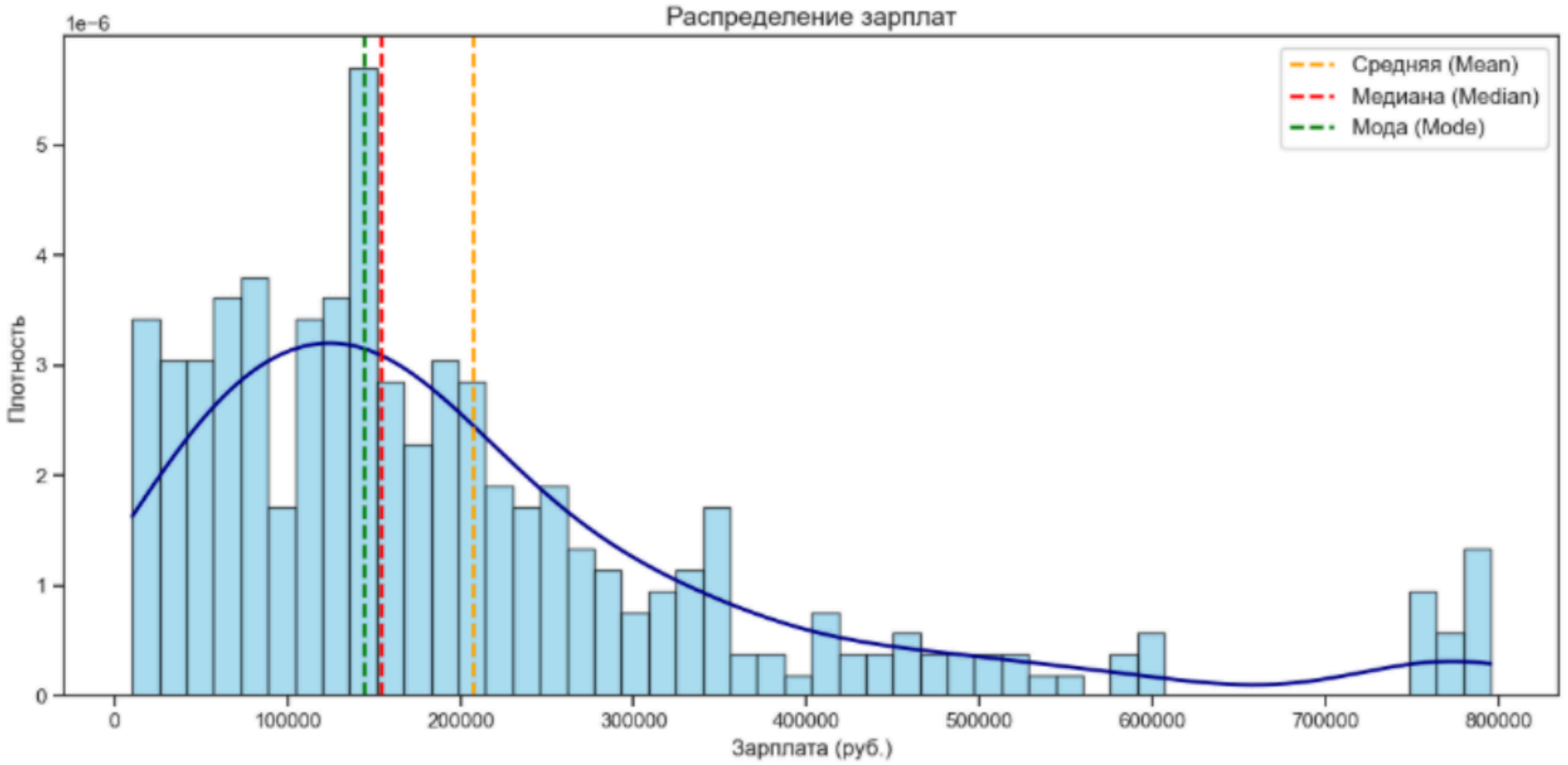


Выброс

Асимметричные распределения и выбросы:

Когда распределение асимметричное (например, правостороннее или левостороннее), выбросы могут оказывать значительное влияние на **среднее арифметическо**е.

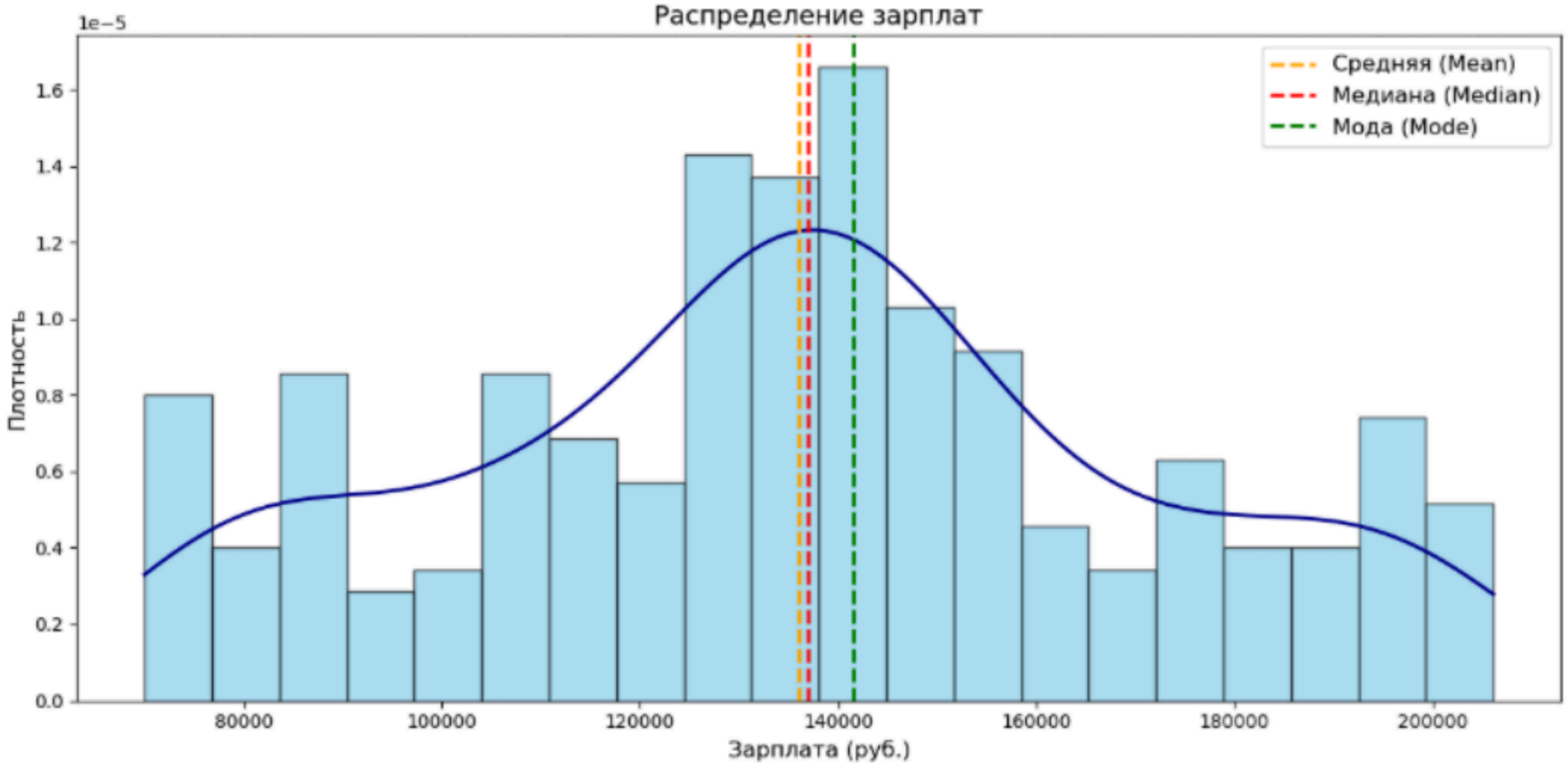
Давайте подробнее рассмотрим это на данном примере:



Почему выбросы "притягивают" среднее арифметическое?

Среднее арифметическое — это сумма всех значений, делённая на их количество. Оно чувствительно к экстремальным значениям (выбросам), поскольку учитывает каждое значение.

В данном примере среднее арифметическое больше медианы, что может означать, что распределение данных асимметрично вправо (или имеет положительный перекос). Если у нас нет выбросов и если распределение напоминает нормальное, то медиана и среднее арифметическое будут примерно одинаковыми.



Существует и формальное правило определения выбросов. Согласно нему выбросами считаются все значения больше, чем:

Выбросы > Третий Квартиль + 1.5 * Межквартильный Размах

Диаграмма размаха

Диаграммы размаха (ящик с усами) – это удобный способ визуального
ых данных через квартили.

Выброс

Верхняя

граница Ус

Верхний квартал

(Q3) Медиана

Нижний квартал (Q1)

Нижняя граница