

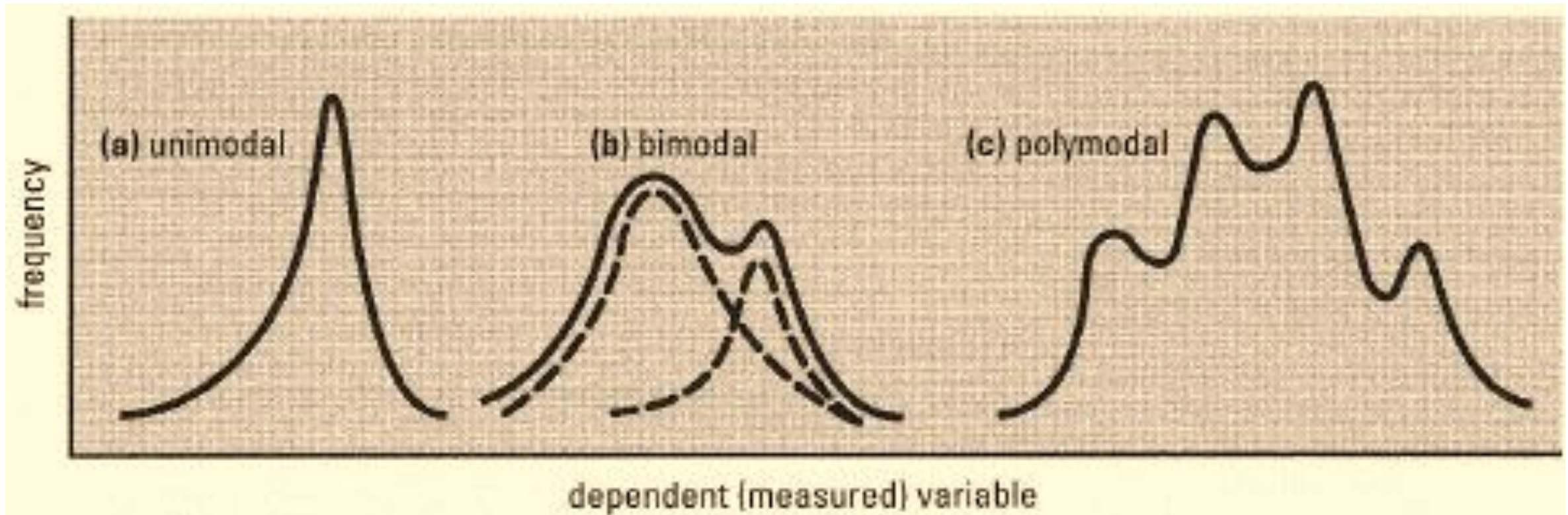
Беседы о прикладной статистике

*Семинар 3. Меры центральной
тенденции. Меры разброса.
Нормальное распределение*

**Фастовец И.
А.**

Меры центральной тенденции. Мода

- Мода – пик, локальный максимум распределения



Среднее

- Сумма всех элементов, разделенная на количество этих элементов
- В случае нормального распределения является *несмещенной оценкой* среднего генеральной совокупности

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

$$\bar{x} = \frac{1}{n} \sum x_i$$

Некоторые свойства среднего

- Если ко всем элементам прибавить одно и то же число, то и к среднему арифметическому будет прибавлено то же число

$$\bar{x}' = \frac{1}{n} \sum_{i=1}^n x'_i = \frac{1}{n} \sum_{i=1}^n (x_i + c) = \frac{1}{n} \sum_{i=1}^n x_i + \frac{1}{n} \sum_{i=1}^n c = \bar{x} + \frac{nc}{n} = \bar{x} + c.$$

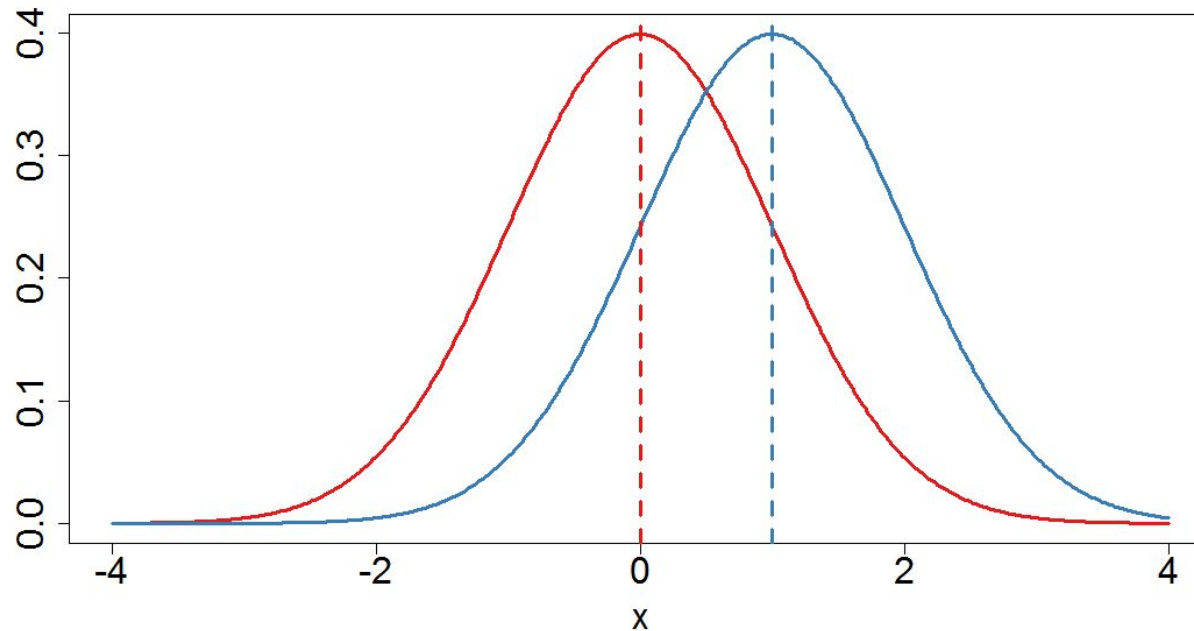
- Если все элементы умножить (разделить) на одно и то же число, то среднее арифметическое умножится (разделится) на то же число

$$\bar{x}' = \frac{1}{n} \sum_{i=1}^n x'_i = \frac{1}{n} \sum_{i=1}^n (x_i \cdot c) = c \cdot \frac{1}{n} \sum_{i=1}^n x_i = c \cdot \bar{x}.$$

Некоторые свойства среднего

- Сумма отклонений элементов от их среднего арифметического равна нулю

$$\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - \sum_{i=1}^n \bar{x} = n\bar{x} - n\bar{x} = 0.$$

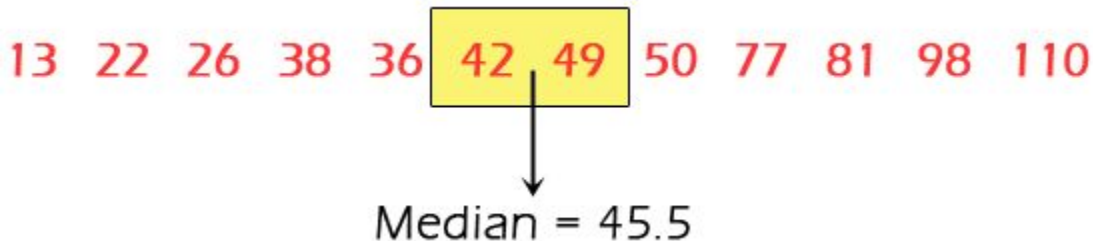
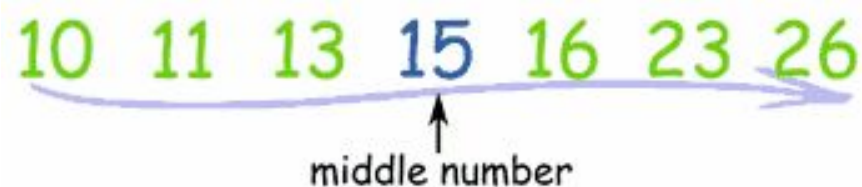


Медиана

Средняя точка распределения. Половина наблюдений больше, а половина меньше медианы

Как вычислить медиану:

- Проранжировать наблюдения от меньшего к большему
- Если n нечетное, то медиана – центральный элемент в ранжированном списке
- Если n четное, то среднее арифметическое двух центральных элементов



Наиболее встречающиеся меры разброса

- Размах – разница между наибольшим и наименьшим значениями. Недостаток – не характеризует распределение целиком, а только крайние значения
- Среднее абсолютное отклонение:

$$\text{average deviation} = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

- Дисперсия и стандартное отклонение

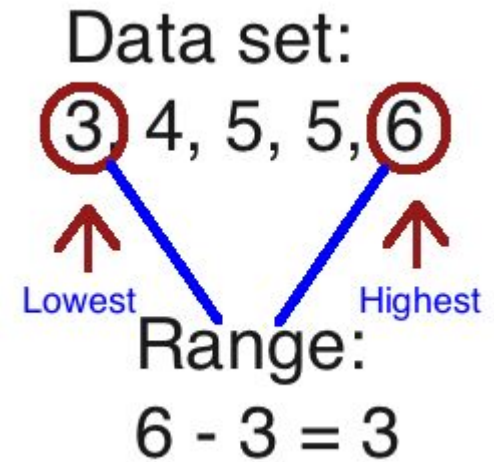
Sample Variance

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

Sample Standard Deviation

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

- Межквартильный интервал (IQR – interquartile range)
- Медианное абсолютное отклонение (MAD)



Дисперсия и стандартное отклонение

- Дисперсия (s^2 , σ^2) – средний квадрат отклонений от среднего арифметического. Стандартное отклонение (СО) – это корень из дисперсии

Sample Variance

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

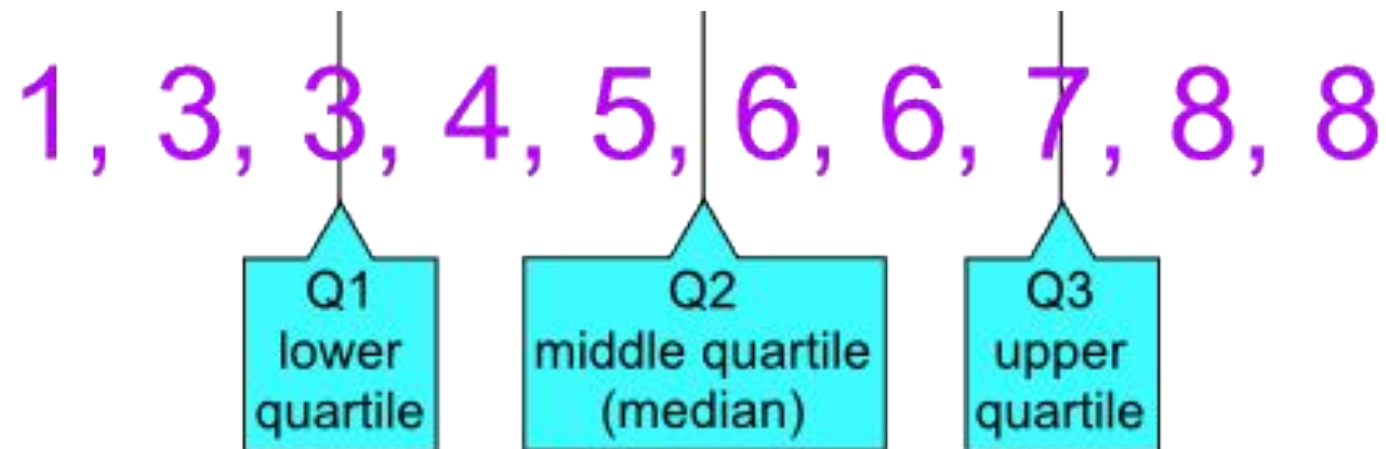
Sample Standard Deviation

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

- Дисперсия и СО по выборке оценивается с учетом степеней свободы ($n-1$). Только тогда они являются *несмещенными* оценками σ^2 и σ генеральной совокупности
- Дисперсия и стандартное отклонение используют только вместе со средним (не с медианой!!!)

Квартили

- Нижний (первый) квартиль Q1 – это медиана левой от медианы группы значений в упорядоченном списке. 25% значений меньше Q1
- Верхний (третий) квартиль Q3 – это медиана правой от медианы группы значений. 25% значений больше Q3
- Второй квартиль Q2 – он же медиана

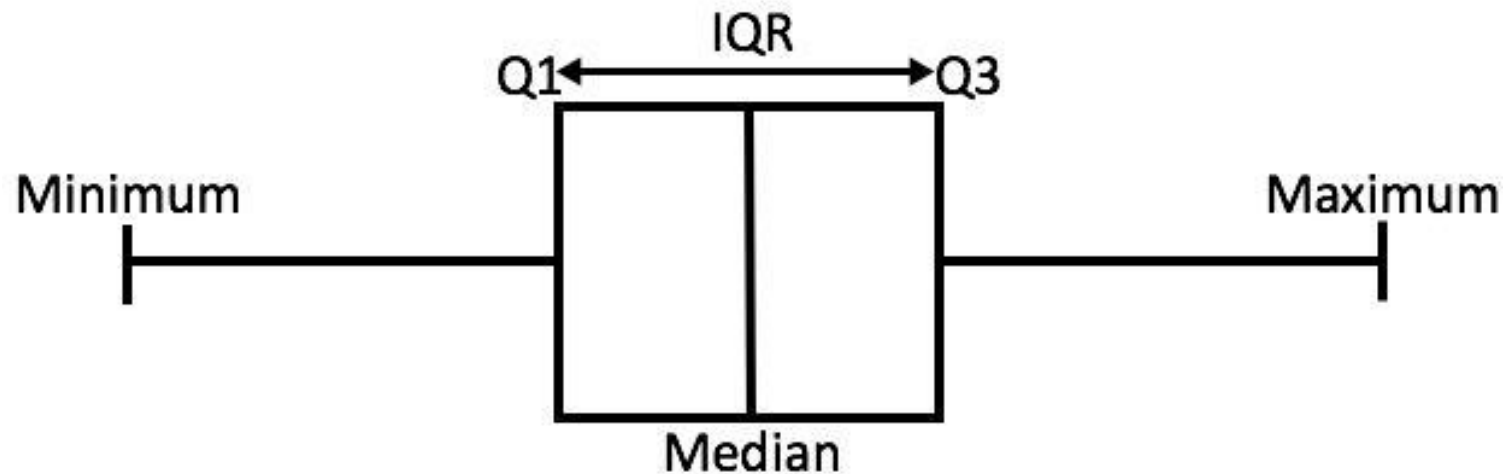


IQR и правило 1.5IQR

- Межквартильный интервал – одна из мер разброса
- Вычисляется как разница третьего и первого квартилей $Q3 - Q1$
- 1.5IQR – правило нахождения выбивающихся значений
- Если значение находится на расстоянии более 1.5IQR над $Q3$ или ниже $Q1$, то это потенциальный выброс
- Five-number summary – непараметрическая форма представления центральной тенденции и разброса распределения:
Минимум – $Q1$ – Медиана – $Q3$ – Максимум

Боксплот

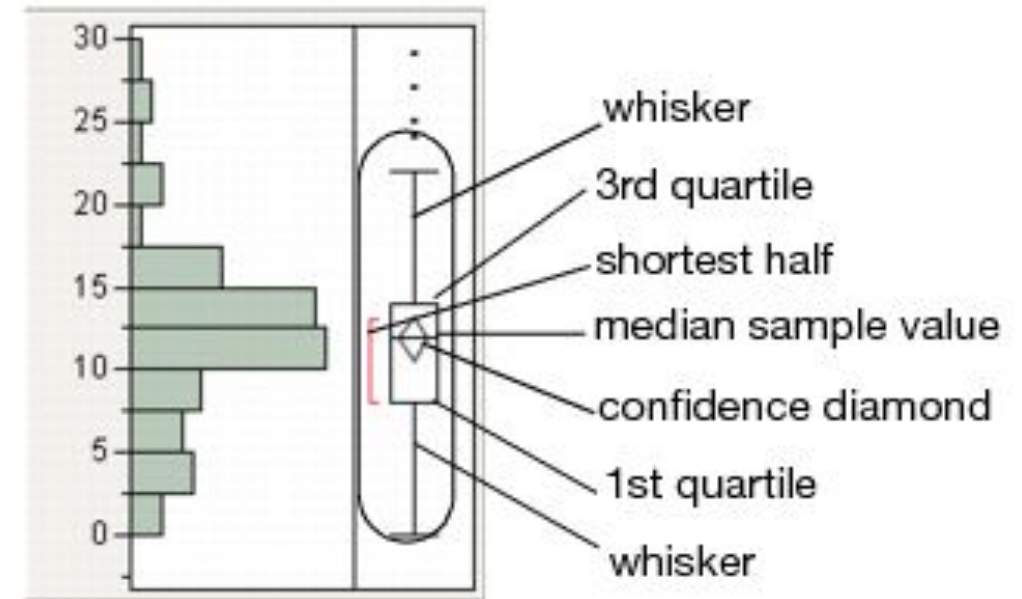
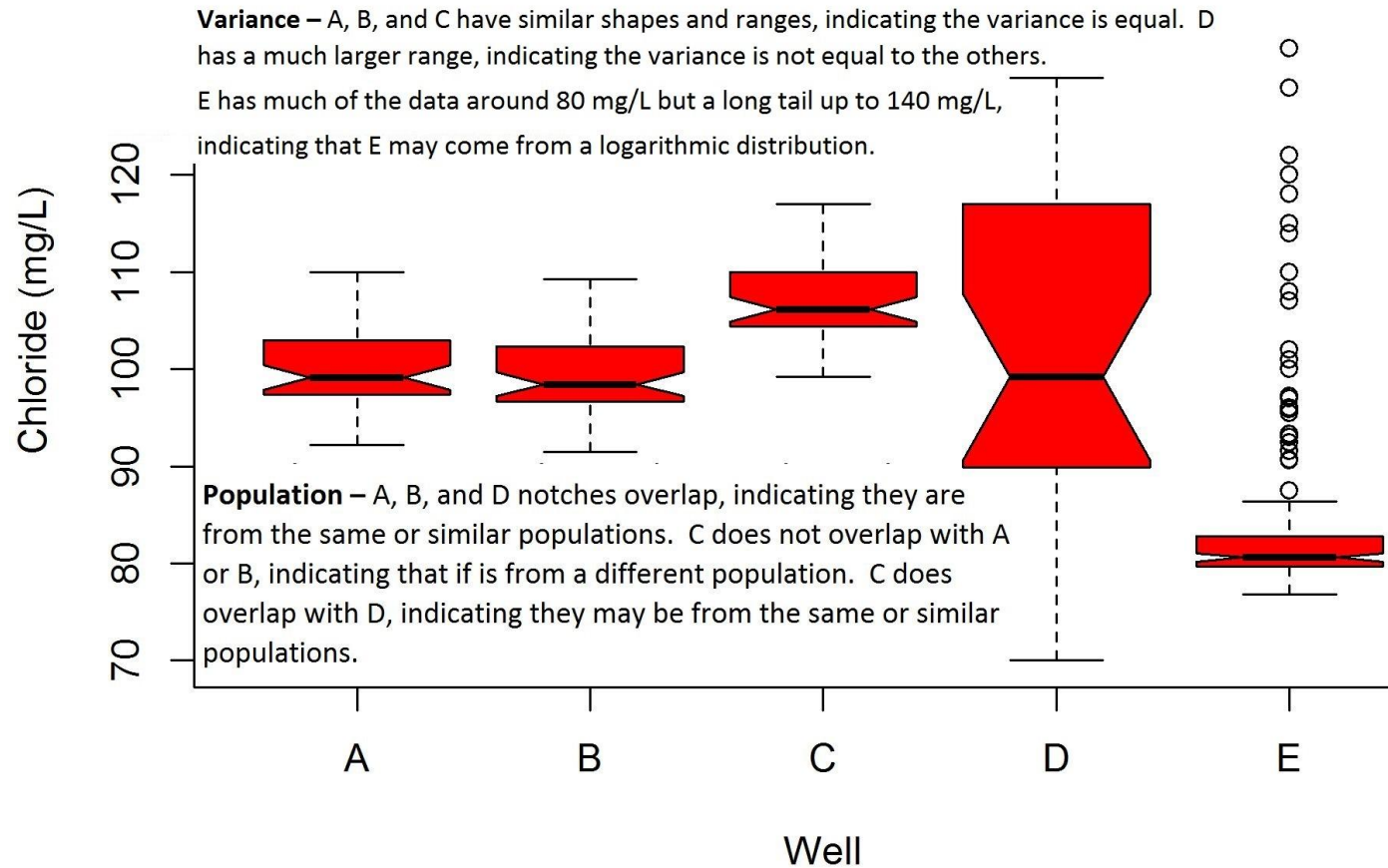
- Диаграмма для представления five-number summary
- В классическом виде коробочка это квартили, а усики – это размах



Модифицированный боксплот

- В модифицированном виде усики – это $1.5IQR$, точки – выбивающиеся значения, а доверительный вырез или алмаз – примерный доверительный интервал для медианы, рассчитываемый как $\pm 1.75 \times \frac{IQR}{\sqrt{n}}$
- Считается, что если вырезы (алмазы) не пересекаются, то имеются значимые различия

Модифицированный боксплот



Медианное абсолютное отклонение (MAD)

- Медиана модулей отклонений от медианы

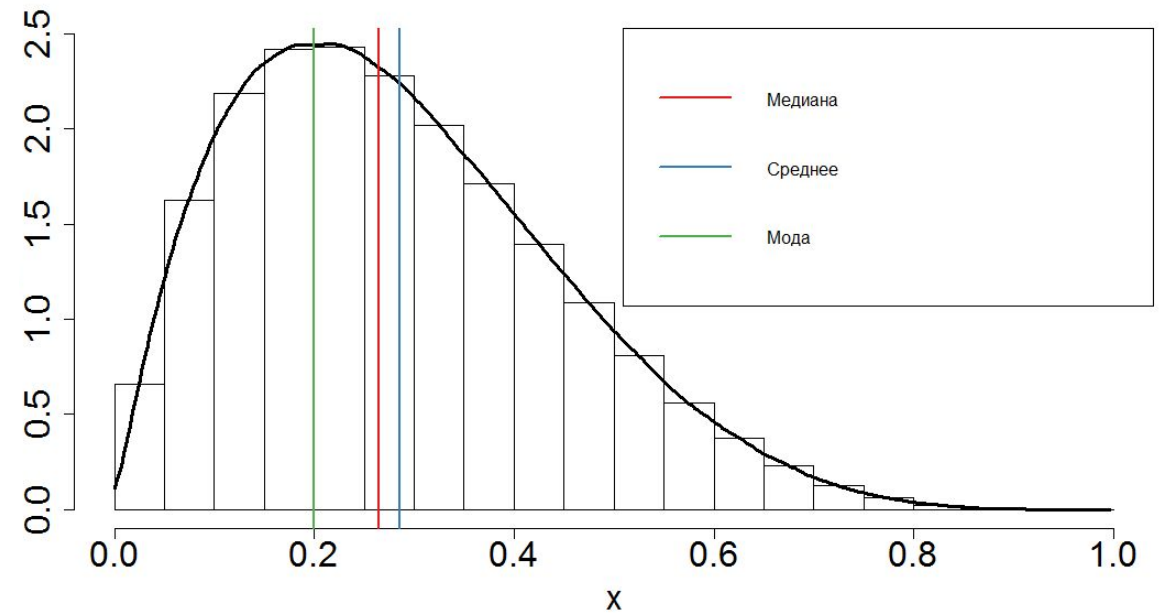
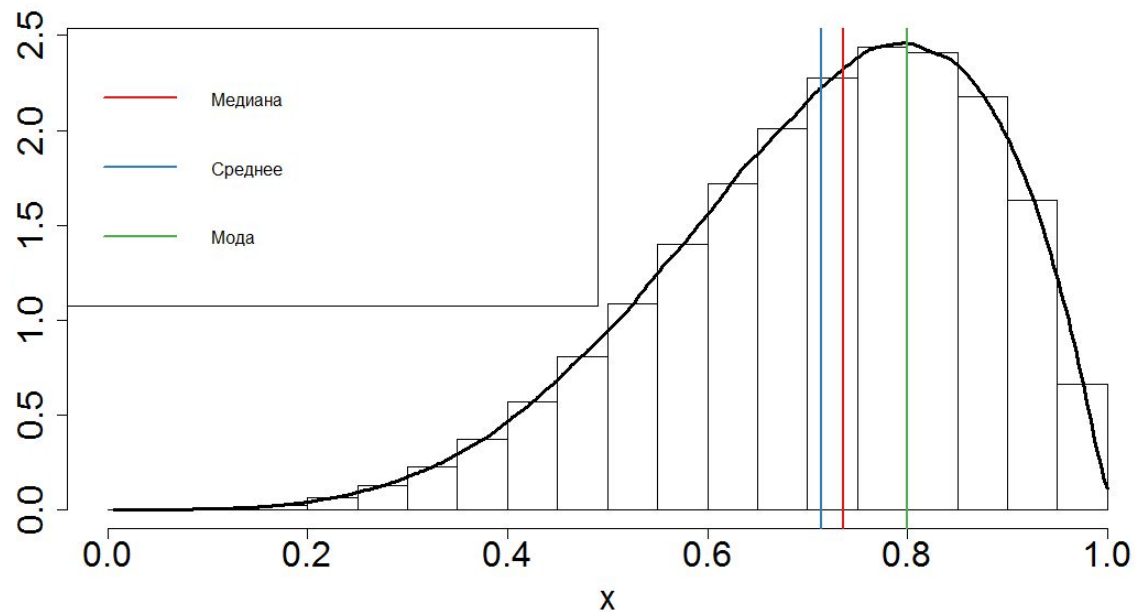
$$\text{MAD} = \text{median}(|X_i - \text{median}(X)|)$$

- Часто умножают на коэффициент 1.4826 . В таком случае представляет собой оценку стандартного отклонения σ , как-будто распределение является нормальным

Чувствительность к выбросам

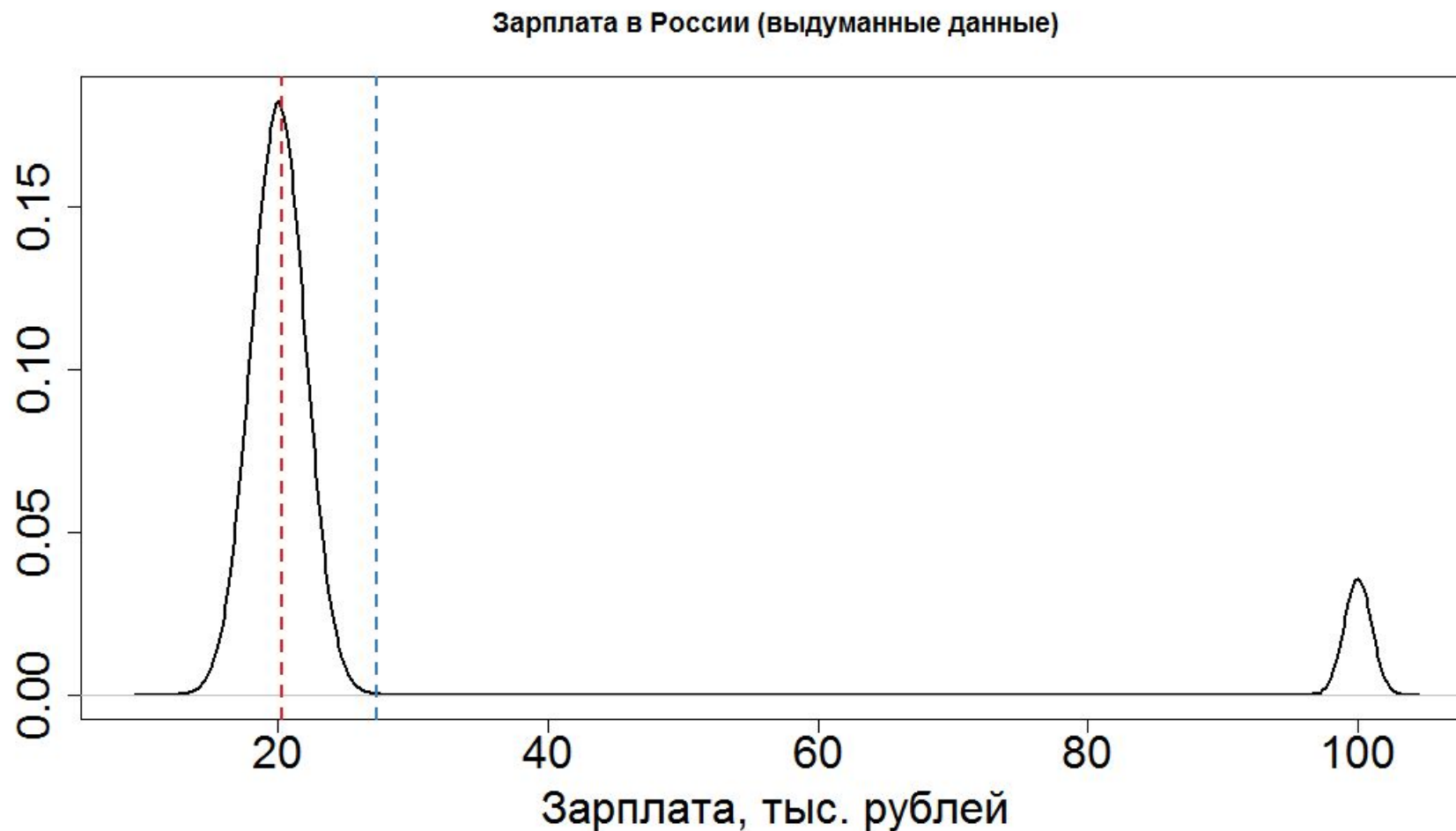
- Различные меры центральной тенденции и разброса характеризуются различной устойчивостью к единичным выбивающимся значениям
- Среднее и особенно дисперсия (стандартное отклонение) являются чувствительными мерами
- Медиана, IQR и MAD характеризуются гораздо меньшей чувствительностью

Среднее, медиана и мода в скошенном унимодальном распределении



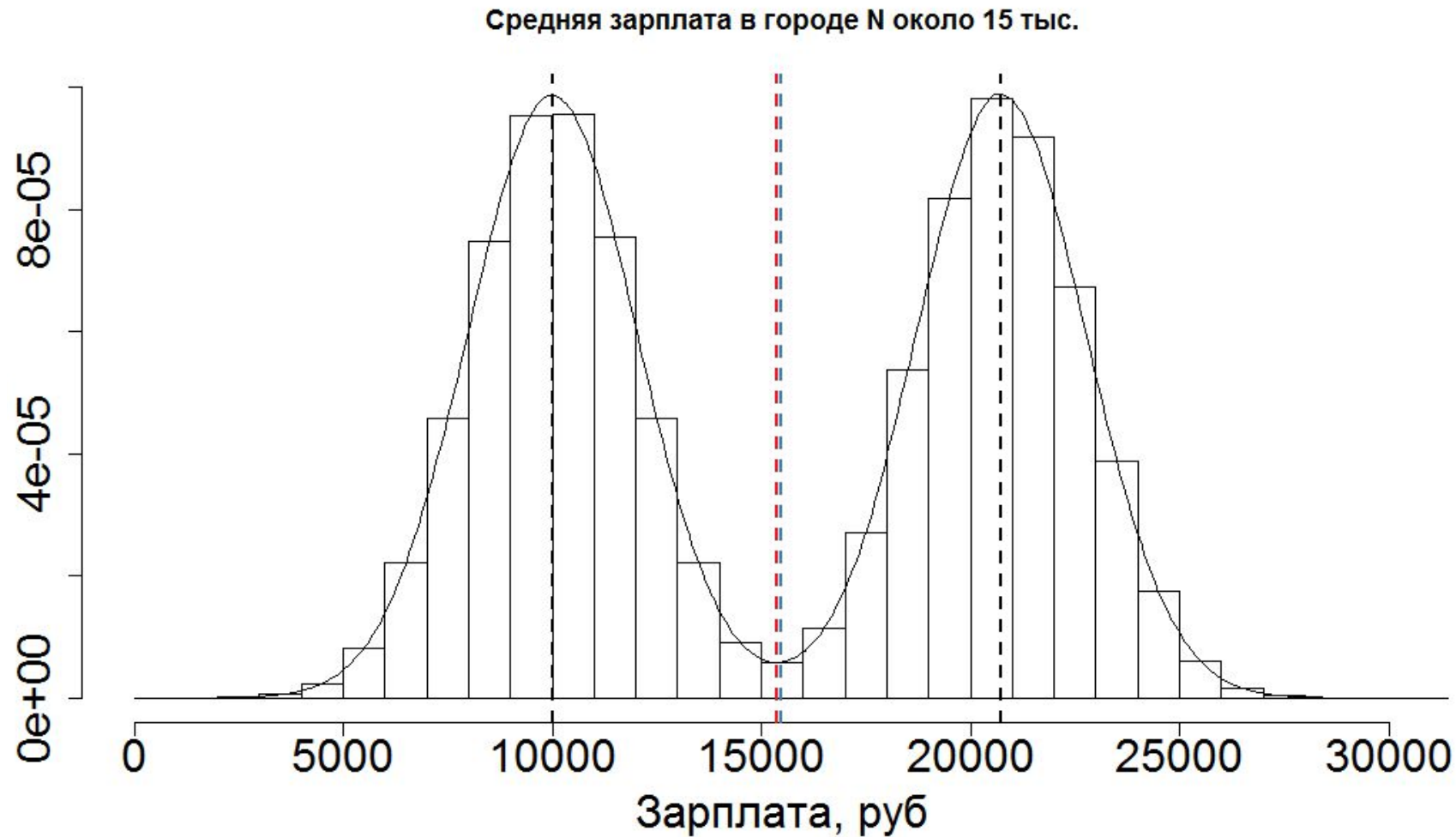
Сильные выбросы

- Средняя зарплата 27.3 тысяч рублей ($s: \pm 23$ тыс.)
- Медианная зарплата 20.2 тысяч рублей (MAD: ± 2.25 тыс.)
- Реальный левый пик: 20 ± 2 тыс.



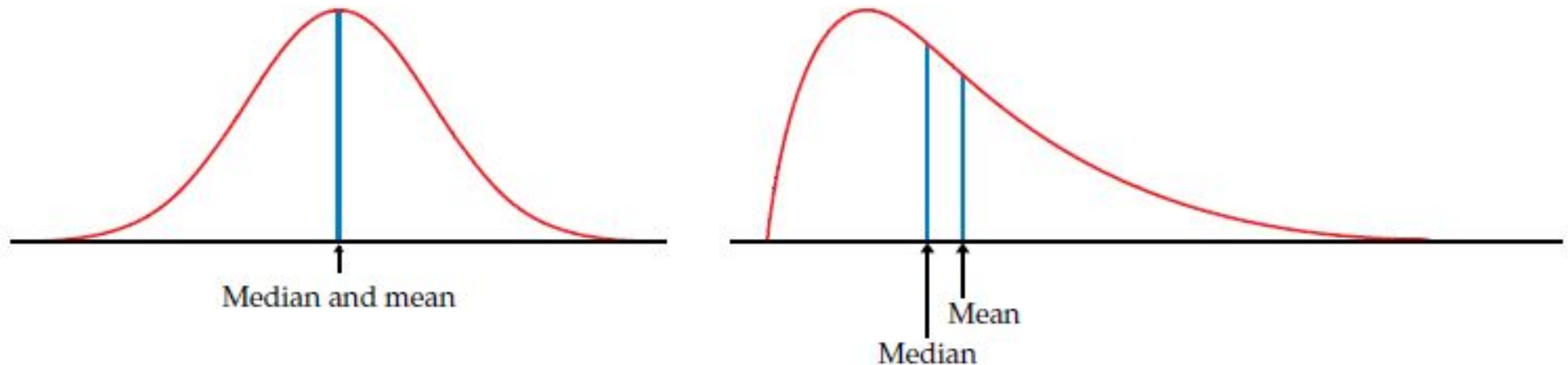
Внимание к модальности!

- Среднее и медиана



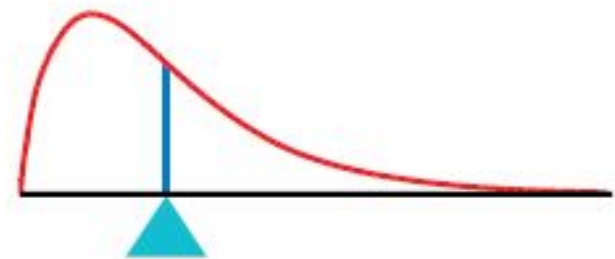
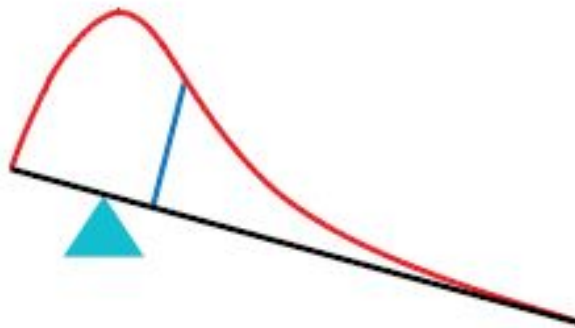
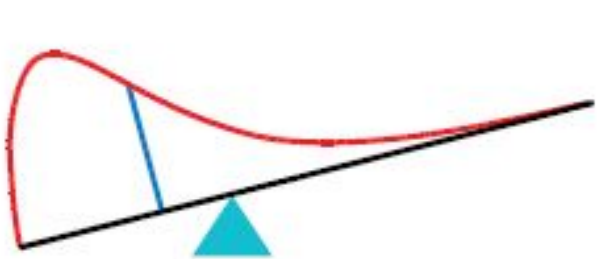
Кривые плотности вероятности

- Описывают общую картину распределения. Площадь под кривой в некотором интервале отражает долю от всех наблюдений, попадающих в этот интервал
- Находится всегда выше горизонтальной оси или на ней
- Имеет площадь под ней, равную 1



Среднее и медиана в контексте кривых плотности вероятности

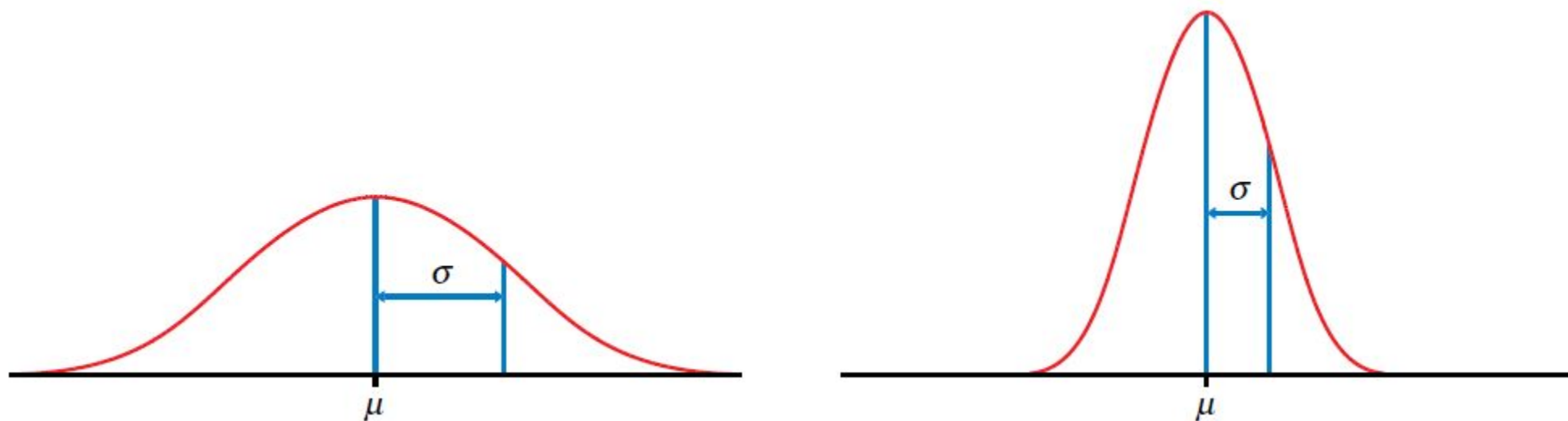
- Медиана делит площадь под кривой плотности вероятности на две равные части по 0.5
- Среднее является «точкой баланса» кривой. Стремится располагаться у более вытянутого хвоста



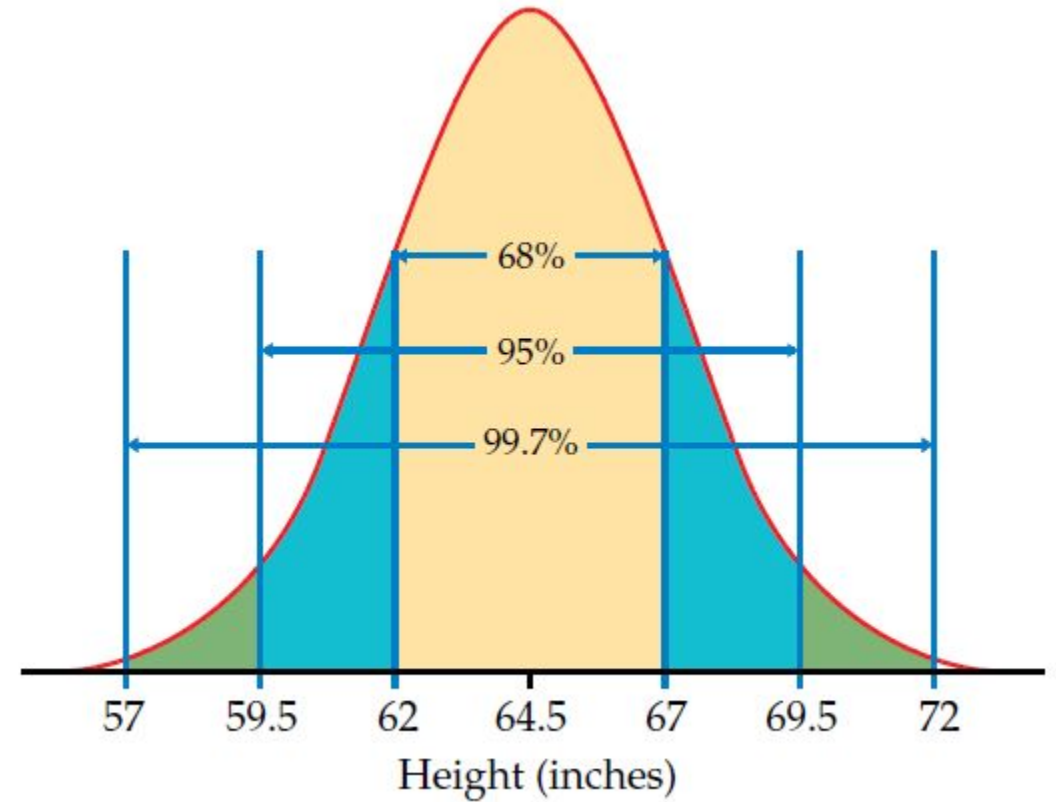
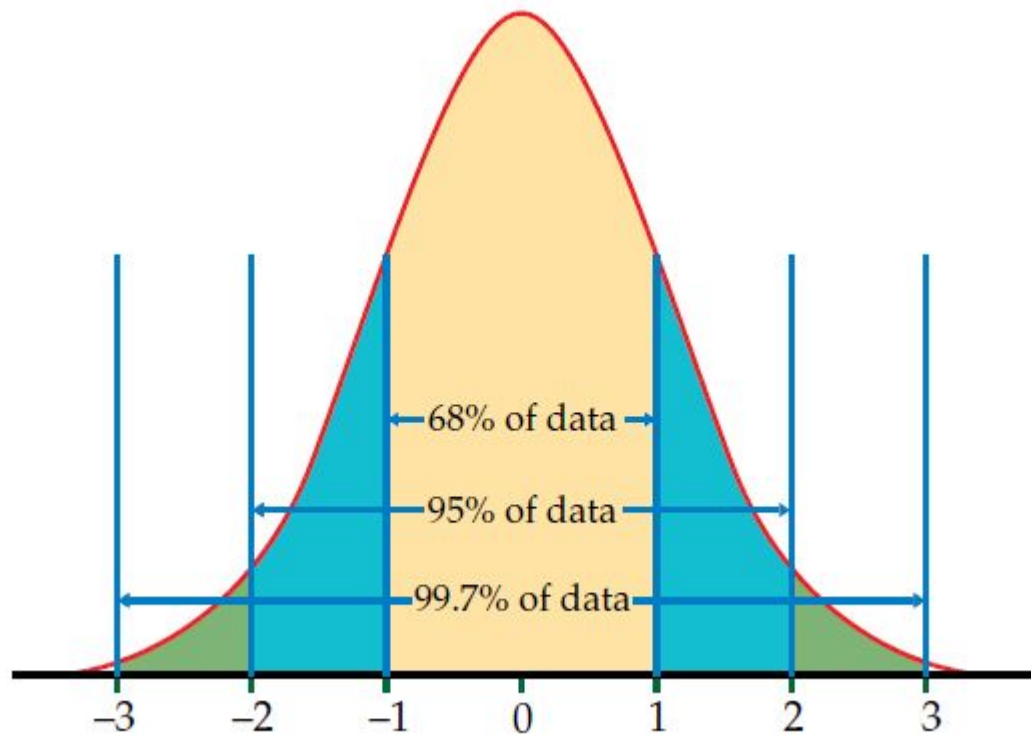
Плотность нормального распределения

- Куполообразное, симметричное распределение
- Задается двумя параметрами: среднее (μ) и стандартное отклонение (σ). Параметры идеального распределения пишутся греческими буквами, как и параметры генеральной совокупности

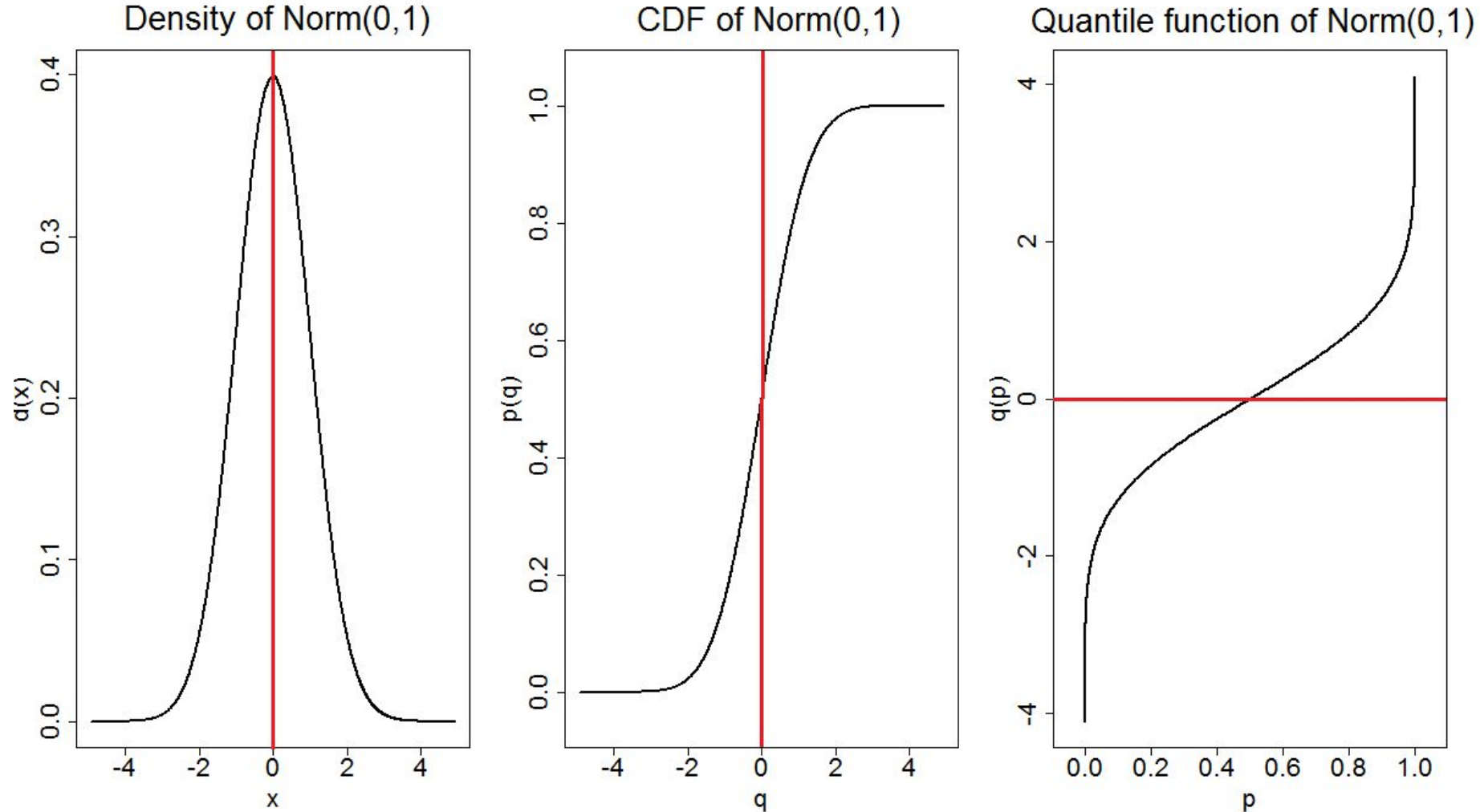
$$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



Правило 68-95-99.7 (трех сигм)



Плотность (PDF) и интегральная функция распределения (CDF)



На следующем семинаре

- Стандартизация и z-шкала
- Параметрические доверительные интервалы
- Проверка гипотез: t-тесты и ранговые тесты Уилкоксона

Student's *t*-distribution

