1
Описательные статистики
Нормальное распределение

План

- Генеральная совокупность и выборка
- Описательные статистики: меры центральной тенденции и разброса
- Описание распределений
- Нормальное распределение

Основные обозначения

X, Y, Z – случайные величины

x, y, z – какие–то конкретные значения

A, B, C – события

$\mathbb{P}$– вероятность

$\mathbb{E}(X)$ – математическое ожидание

$Var(X), \sigma^2$ – дисперсия

$\mathrm{Cov}(X), \rho(X, Y)$ – ковариация и корреляция

Выборка и генеральная совокупность

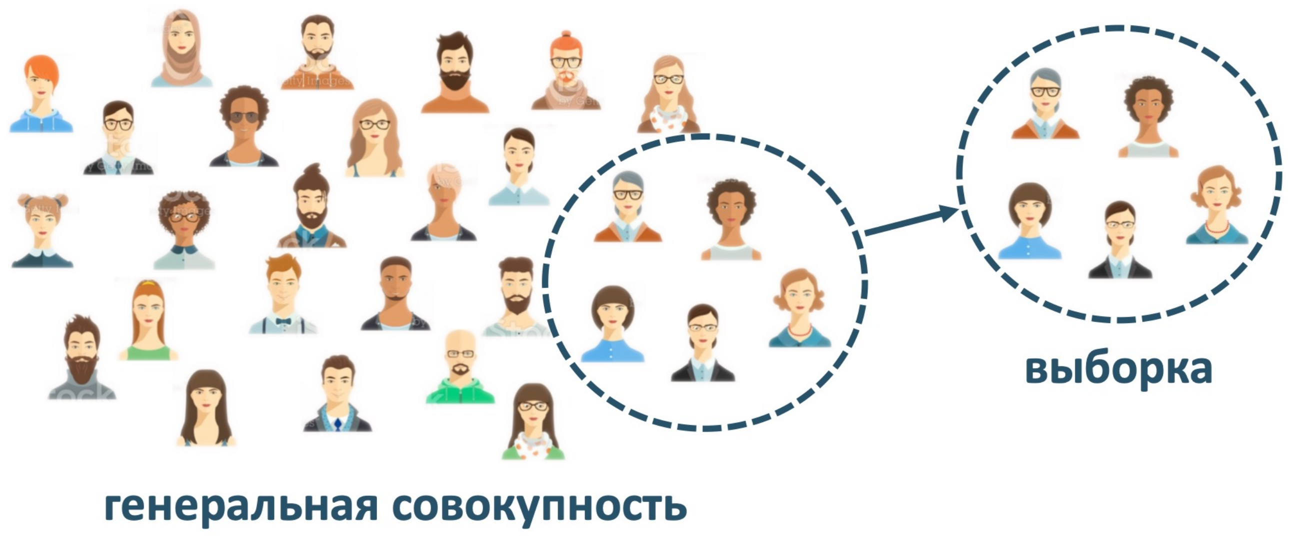
Генеральная совокупность vs Выборка

- **Генеральная совокупность** – это все объекты, которые нас интересуют при исследовании

- **Выборка** – это та часть генеральной совокупности, по которой мы собрали данные для исследования

Генеральная совокупность vs Выборка

- В городе живёт 1 млн. человек
- Провели опрос об уровне дохода (2.5 тыс. человек)
- Опубликовали средний доход по городу
- Опрашивать абсолютно всех людей в городе дорого и долго



Репрезентативность выборки

- Выборки позволяют сделать выводы о всей генеральной совокупности
- Чтобы выводы были корректными, выборка должны быть **репрезентативной**
- **Репрезентативная выборка** – отражает свойства генеральной совокупности
- Способы достижения репрезентативности: случайный отбор, стратифицированный отбор

Пример: Добрыня, Илья и Алёна исследуют рост людей. Чья выборка репрезентативна?

- Добрыня опросил свою баскетбольную команду
- Илья опросил людей на остановке
- Алёна опросила всех своих подруг

Предпосылки

Выборка: $X_1, X_2, \dots X_n$

Одно наблюдение: X_i

• Каждое наблюдение можно рассматривать как случайную величину, которая имеет такое же распределение как и генеральная совокупность

Мы в дальнейшем будем всегда предполагать:

1. Наблюдения $X_1, X_2, \dots X_n$ независимы друг от друга
2. Наблюдения имеют одинаковое распределение (как у генеральной совокупности)

Краткая запись: $X_1, X_2, \dots X_n \sim iid$

➤ iid - *расшифровывается как identically independently distributed (независимы и одинаково распределены)*

Выборка

	Название	Сборы	Год
0	Мстители: Война бесконечности (2018)	2048359754	2018
1	Черная Пантера (2018)	1346913161	2018
2	Мир Юрского периода 2 (2018)	1309484461	2018
3	Суперсемейка 2 (2018)	1242805359	2018

- Строчка таблицы – наблюдение
- Столбец таблицы – переменная

Описательные статистики

Какие бывают переменные

- Категориальные

Принимают значения из какого-то ограниченного множества: пол, цвет машины, страна сборки и т.п.

- Числовые

Могут принимать бесконечное число значений: возраст, вес, цены, кассовые сборы и т.п.

Описательные статистики

- Меры центральной тенденции (МЦТ) - такие значения, которые наилучшим способом описывает *типичное* наблюдение из данных

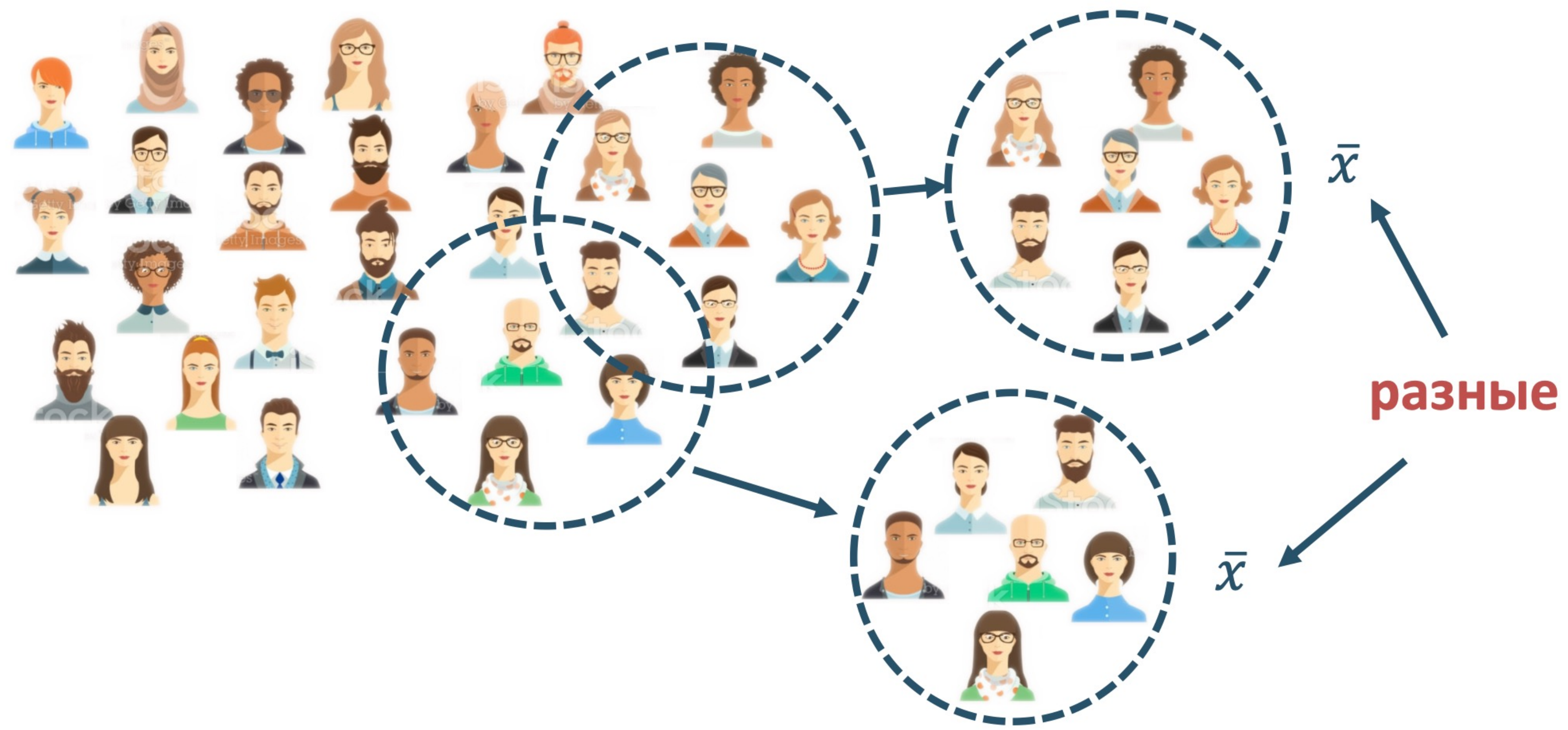
Пример: среднее, медиана, мода

- Меры разброса - это оценка того, насколько данные разбросаны относительно меры центральной тенденции

Пример: дисперсия, стандартное отклонение, IQR

Описательные статистики

Выборка: $X_1, X_2, \dots X_n$
Статистика – любая функция от наблюдений
Примеры: среднее, медиана, максимум и т.п.



МЦТ. Среднее арифметическое

Выборка: $X_1, X_2, \dots X_n$

Среднее арифметическое:

$$\bar{x} = \frac{X_1 + X_2 + \cdots X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Пример:

Выборка: $x_1 = 1\,, x_2 = 5, x_3 = -4, x_4 = 3, x_5 = 0$

Среднее арифметическое:

$$\bar{x} = \frac{1 + 5 + (-4) + 3 + 0}{5} = 1$$

МЦТ. Медиана

Такое значение которое делит выборку пополам. Слева от медианы 50% значений и справа 50%

Пример 1: $x_1 = 1\,, x_2 = 5, x_3 = -4, x_4 = 3, x_5 = 0$

Если в выборке нечетное количество наблюдений

1. Расположим значение по порядку: -4 0 1 3 5
2. Значение по середине медиана: $med = 1$

Пример 2: $x_1 = 1\,, x_2 = 5, x_3 = -4, x_4 = 3$

Если в выборке четное количество наблюдений

1. Расположим значение по порядку: -4 1 3 5
2. Находим среднее двух чисел по середине, это будет медиана: $med = \frac{1+3}{2} = 2$

Среднее vs Медиана

- Среднее более чувствительно к выбросам в данных чем медиана
- Если в выборке нет выбросов они примерно совпадают

Пример 1: $x_1 = 1, x_2 = 5, x_3 = -4, x_4 = 3, x_5 = 100$

Среднее: $\bar{x} = \frac{1 + 5 + (-4) + 3 + 100}{5} = 21$

Медиана:
1. -4 1 3 5 100
2. $med = 3$

Пример 2: $x_1 = 1, x_2 = 5, x_3 = -4, x_4 = 3$

Среднее: $\bar{x} = \frac{1 + 5 + (-4) + 3}{5} = 1$

Медиана:
1. -4 1 3 5
2. $med = \frac{1+3}{2} = 2$

МЦТ. Мода

Значение переменной с самой большой частотой, т.е. самое популярное значение переменной

Пример: $x_1 = 1\,, x_2 = 5, x_3 = -4, x_4 = 3, x_5 = 5$

1. Строим таблицу частотности:

5	2
3	1
1	1
-4	1

2. Чаще всего встречается число 5

3. 5 это мода

Меры разброса. Выборочная дисперсия

Хочется понимать насколько сильно элементы выборки отклоняются от своего типичного значения

$$\hat{\sigma}^2 = \frac{(X_1 - \bar{x})^2 + (X_2 - \bar{x})^2 + \cdots + (X_n - \bar{x})^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})$$

Пример:

	Имя	Возраст
0	Алена	25
1	Настя	23
2	Зина	27

1. Находим среднее: $\bar{x} = \frac{25 + 23 + 27}{3} = 25$

2. Находим выборочную дисперсию:

$$\hat{\sigma}^2 = \frac{(25 - 25)^2 + (23 - 25)^2 + (27 - 25)^2}{3} =$$

$$= \frac{0^2 + (-2)^2 + 2^2}{3} = \frac{0+4+4}{3} = \frac{8}{3} = 2,7$$

Меры разброса. Выборочная дисперсия

Хочется понимать насколько сильно элементы выборки отклоняются от своего типичного значения

$$\hat{\sigma}^2 = \frac{(X_1 - \bar{x})^2 + (X_2 - \bar{x})^2 + \cdots + (X_n - \bar{x})^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})^2$$

Пример:

	Имя	**Возраст**
0	Алена	25
1	Настя	23
2	Зина	27

1. Находим среднее: $\bar{x} = \frac{25 + 23 + 27}{3} = 25$

2. Находим выборочную дисперсию:

$$\hat{\sigma}^2 = \frac{(25 - 25)^2 + (23 - 25)^2 + (27 - 25)^2}{3} =$$

$$= \frac{0^2 + (-2)^2 + 2^2}{3} = \frac{0+4+4}{3} = \frac{8}{3} = 2{,}7$$

Меры разброса. Выборочная дисперсия

Можно использовать еще одну формулу для выборочной дисперсии

$$\hat{\sigma}^2 = \frac{(X_1 - \bar{x})^2 + (X_2 - \bar{x})^2 + \cdots + (X_n - \bar{x})^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - 2X_i^2\bar{x} + \bar{x}^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{2\bar{x}}{n}\sum_{i=1}^{n} X_i + \frac{1}{n}\sum_{i=1}^{n} \bar{x}^2 = \overline{x^2} - 2\bar{x}\frac{\sum X_i}{n} + \bar{x}^2 = \overline{x^2} - 2\bar{x}^2 + \bar{x}^2 = \overline{x^2} - \bar{x}^2$$

Меры разброса. Выборочная дисперсия

$$\hat{\sigma}^2 = \overline{x^2} - \bar{x}^2$$

Пример:

	Имя	**Возраст**
0	Алена	25
1	Настя	23
2	Зина	27

1. $\bar{x}^2 = \left(\frac{25 + 23 + 27}{3}\right)^2 = 625$
2. $\overline{x^2} = \frac{25^2 + 23^2 + 27^2}{3} = 627.7$
3. $\hat{\sigma}^2 = 627.7 - 625 = 2{,}7$

Меры разброса. Стандартное отклонение

Чтобы вернуться от квадратов дисперсии к исходным величинам используют стандартное отклонение. Берут корень из дисперсии:

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2}$$

Пример:

	Имя	Возраст
0	Алена	25
1	Настя	23
2	Зина	27

1. Находим среднее: $\bar{x} = \dfrac{25 + 23 + 27}{3} = 25$

2. Находим выборочную дисперсию:

$$\hat{\sigma}^2 = \frac{(25-25)^2 + (23-25)^2 + (27-25)^2}{3} =$$

$$= \frac{0^2 + (-2)^2 + 2^2}{3} = \frac{0+4+4}{3} = \frac{8}{3} = 2.7$$

3. Находим стандартное отклонение:

$$\hat{\sigma} = \sqrt{2.7} = 1.6$$

Меры разброса. Несмещенная выборочная дисперсия

Обычно на практике используют именно несмещенную выборочную дисперсию:

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

Но мы обсудим это позднее…

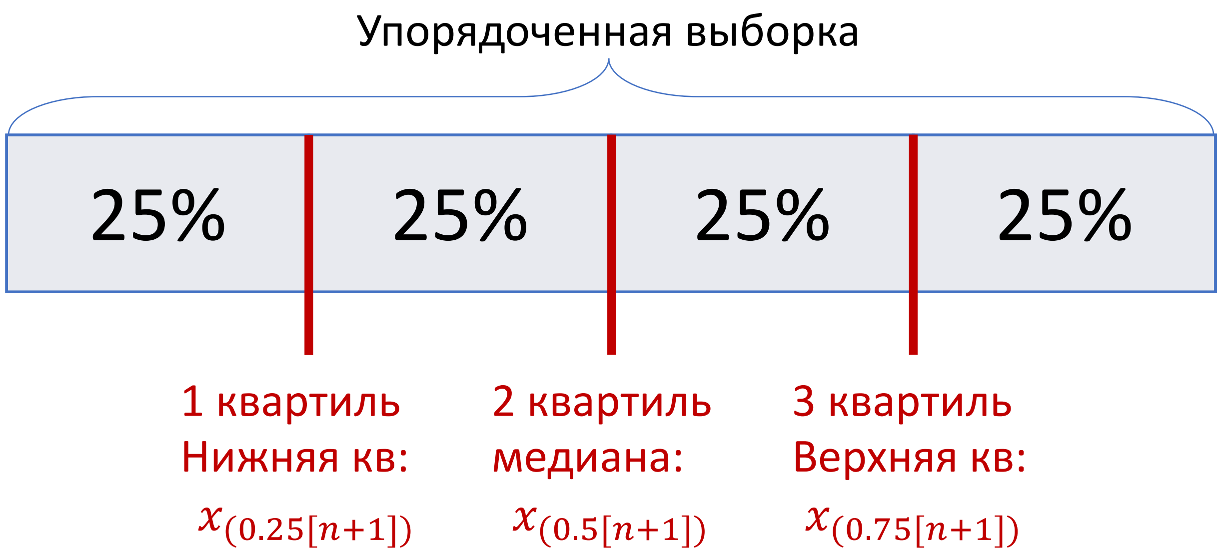
Перцентиль

Перцентиль порядка k – это такое число, что k% выборки меньше этого числа

- Проще всего вычислять его по упорядоченной выборке

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$$

- Квартили – перцентили с шагом в 0.25:



Квартиль

Пример: $x_1 = 1\,, x_2 = 5, x_3 = -4, x_4 = 3, x_5 = 0$

1. Упорядочим выборку

-4 0 1 3 5

2. Найдем 2 квартиль (медиану)

-4 0 1 3 5

3. Найдем 3 квартиль (верхний)

-4 0 1 3 5

3. Найдем 3 квартиль (верхний)

-4 0 1 3 5

Меры разброса. Размах

Разница между максимальным и минимальным значением в наших данных. Амплитуда разброса

Пример 1: $x_1 = 1\,, x_2 = 5, x_3 = -4, x_4 = 3, x_5 = 100$

Минимум: -4

Максимум: 100

Размах: 100-(-4)=104

Пример 2: $x_1 = 1\,, x_2 = 5, x_3 = -4, x_4 = 3, x_5 = 0$

Минимум: -4

Максимум: 5

Размах: 5-(-4)=9

Меры разброса. Интерквартильный размах (IQR)

Разница между верхним и нижнем квартилем

$$IQR = x_{(0.75[n+1])} - x_{(0.25[n+1])}$$

Пример 1: $x_1 = 1, x_2 = 5, x_3 = -4, x_4 = 3,\ x_5 = 100$

$x_{(0.25[n+1])}$: 1

$x_{(0.75[n+1])}$: 5

IQR: 5-1=4

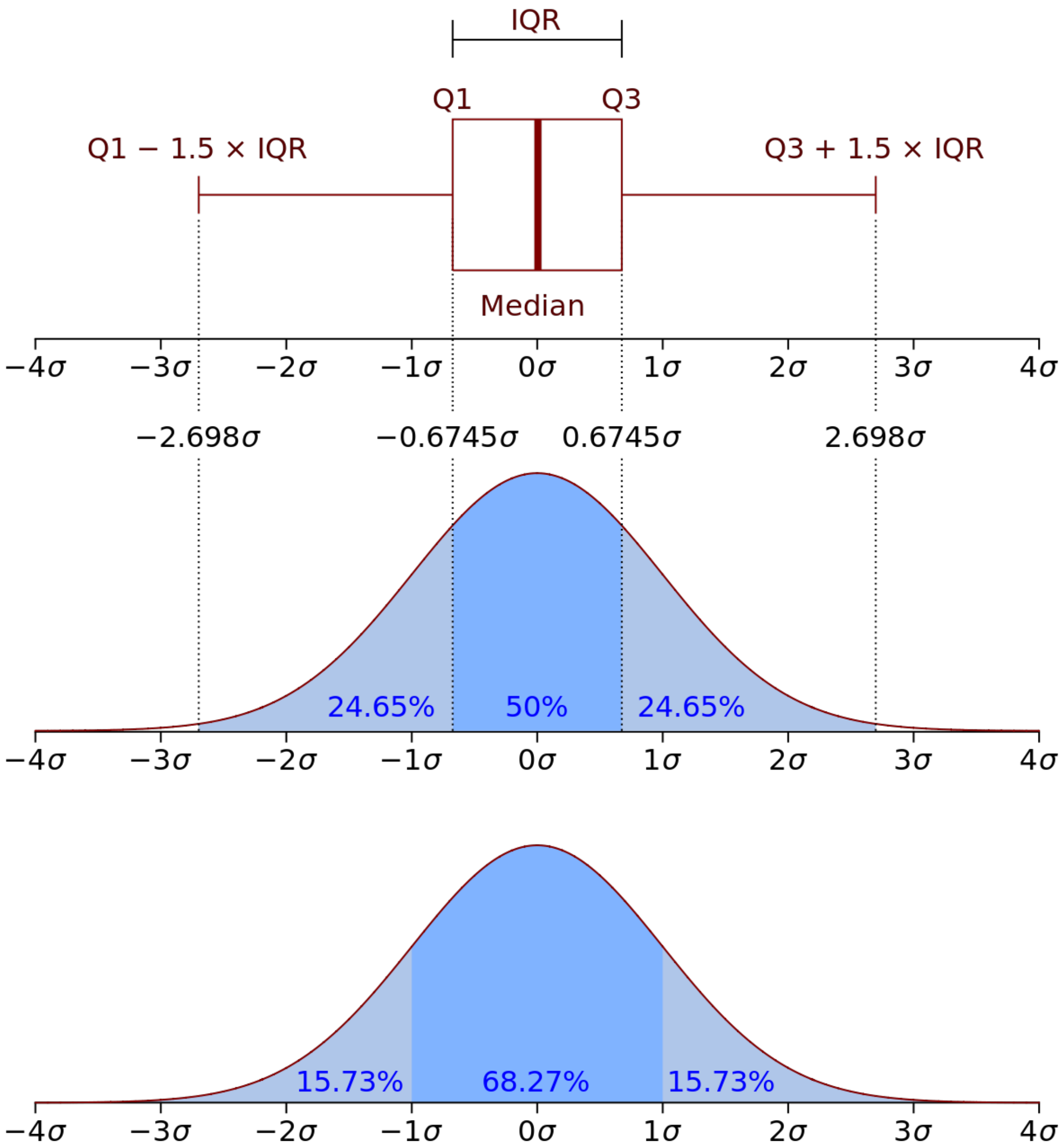
Пример 2: $x_1 = 1, x_2 = 5, x_3 = -4, x_4 = 3,\ x_5 = 0$

$x_{(0.25[n+1])}$: 0

$x_{(0.75[n+1])}$: 3

IQR: 3-0=3

Ящик с усами (Boxplot)



Описание распределения данных

Какие бывают числовые данные

- Дискретные: Множество значений конечно или счётно

число звонков, число очков на игральной кости, число ошибок на страницу текста

- Непрерывные: Принимают бесконечное, континуальное число значений

рост, время ожидания автобуса, вес

Как можно описать непрерывные данные

- Эмпирическая функция распределений
- Функция плотности (гистограмма)
- Ядерная функция плотности

Эмпирическая функция распределения

Эмпирическая функция распределения – функция, которая определяет для каждого x частоту события $X \leq x$, то есть

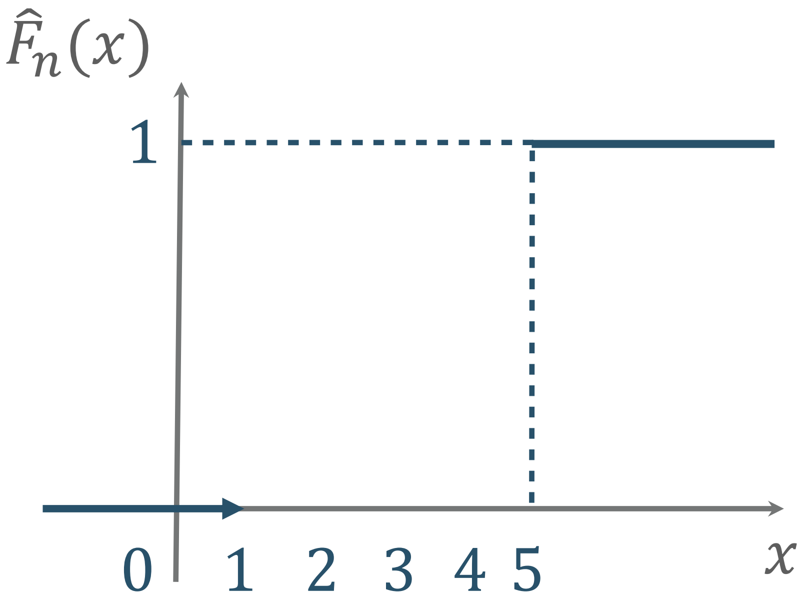
$$\hat{F}_n(x) = \hat{\mathbb{P}}(X \leq x) = \frac{1}{n}\sum_{i=1}^{n}[X_i \leq x],$$

где $[\ ]$ – индикаторная функция, то есть:

$$[X_i \leq x] = \begin{cases} 1, X_i \leq x \\ 0, \text{иначе} \end{cases}$$

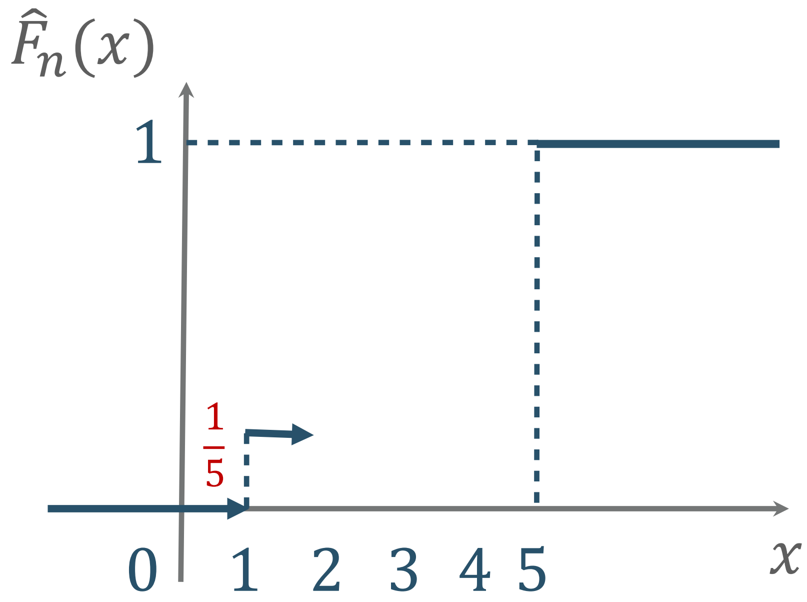
Эмпирическая функция распределения

Пример: $x_1 = 2, x_2 = 5, x_3 = 2, x_4 = 3, x_5 = 1$



Эмпирическая функция распределения

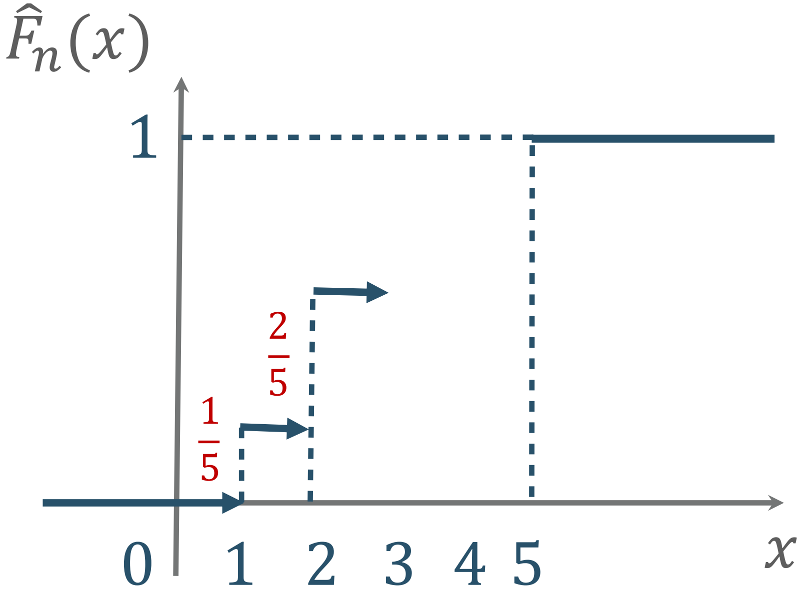
Пример: $x_1 = 2, x_2 = 5, x_3 = 2, x_4 = 3, x_5 = 1$



- Выборочная вероятность того, что некоторая случайная меньше чем 1 равна $\frac{1}{5}$
- Выборочная вероятность того, что некоторая случайная меньше чем 1.5 равна $\frac{1}{5}$

Эмпирическая функция распределения

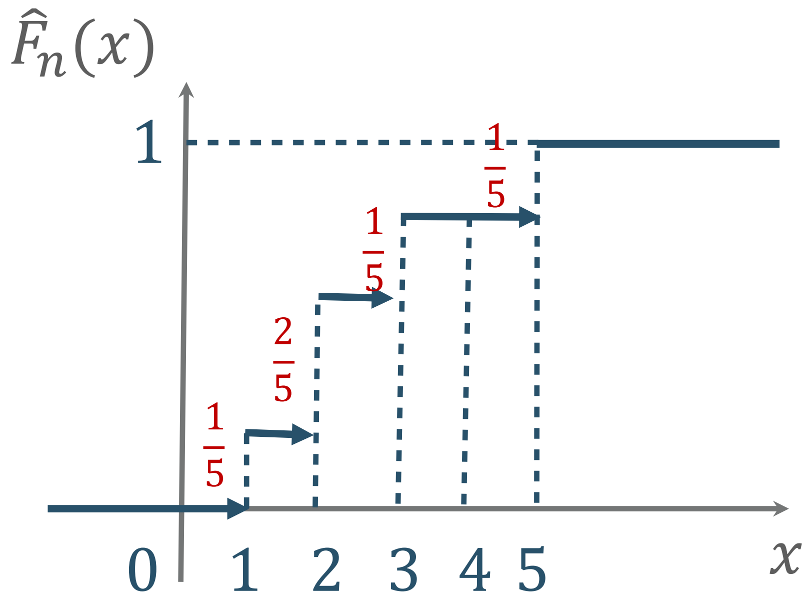
Пример: $x_1 = 2, x_2 = 5, x_3 = 2, x_4 = 3, x_5 = 1$



- Выборочная вероятность того, что некоторая случайная меньше чем 2 равна $\frac{3}{5}$
- Выборочная вероятность того, что некоторая случайная меньше чем 2.6 равна $\frac{3}{5}$

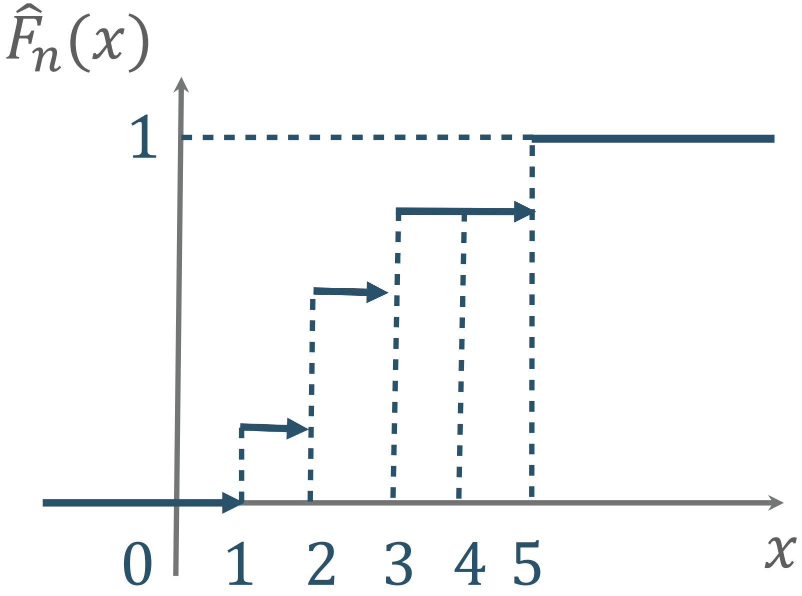
Эмпирическая функция распределения

Пример: $x_1 = 2, x_2 = 5, x_3 = 2, x_4 = 3, x_5 = 1$



Эмпирическая функция распределения

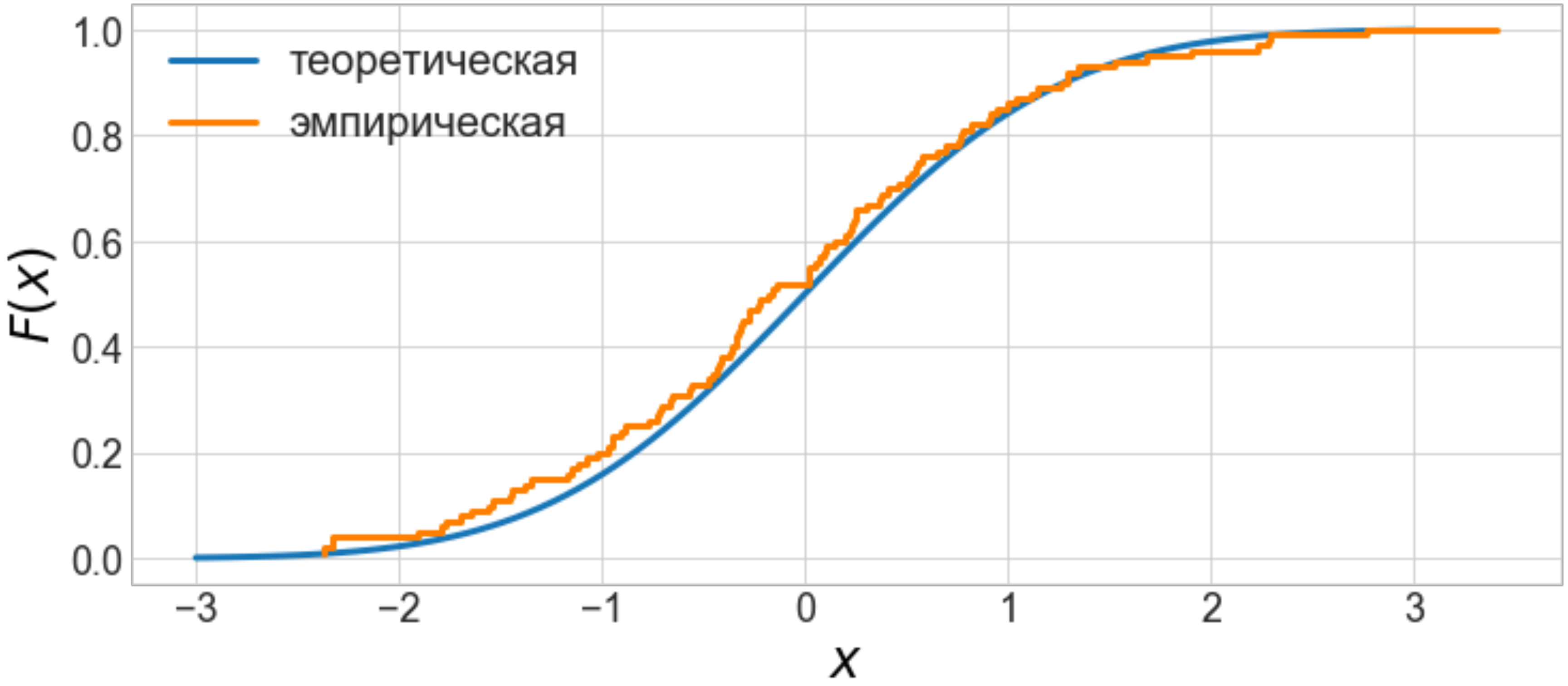
Пример: $x_1 = 2, x_2 = 5, x_3 = 2, x_4 = 3, x_5 = 1$



По аналогии строится теоретическая функция распределения для дискретных случайных величин

Эмпирическая функция распределения

Чем больше выборка, тем чаще ступеньки и тем больше эмпирическая функция распределения похожа на теоретическую

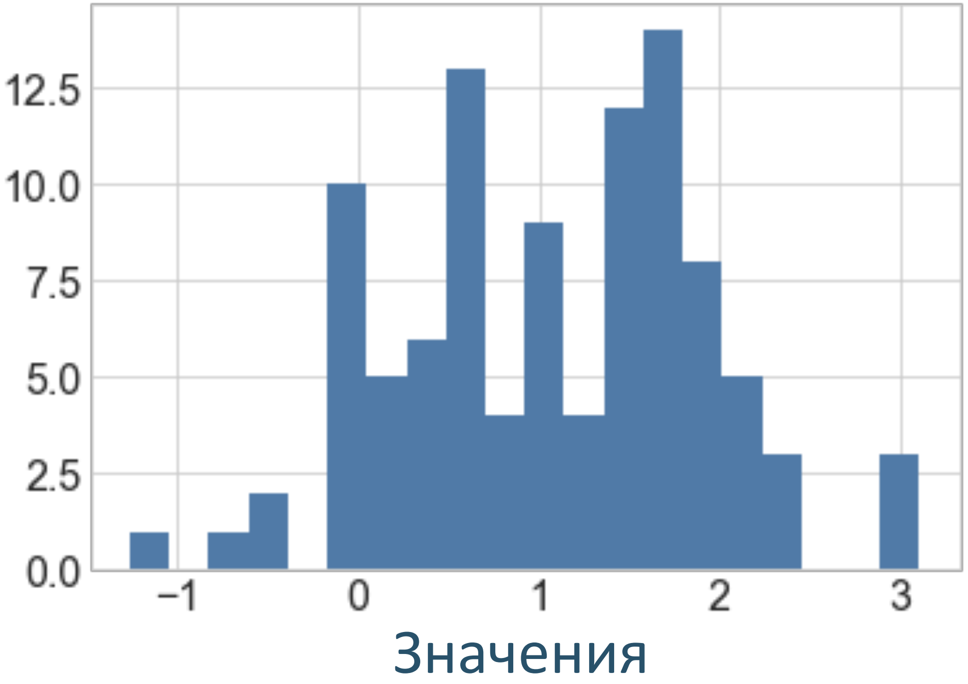


Гистограмма

Гистограмма – эмпирическая оценка плотности распределения. По оси x откладывают значения, по оси y частоты.

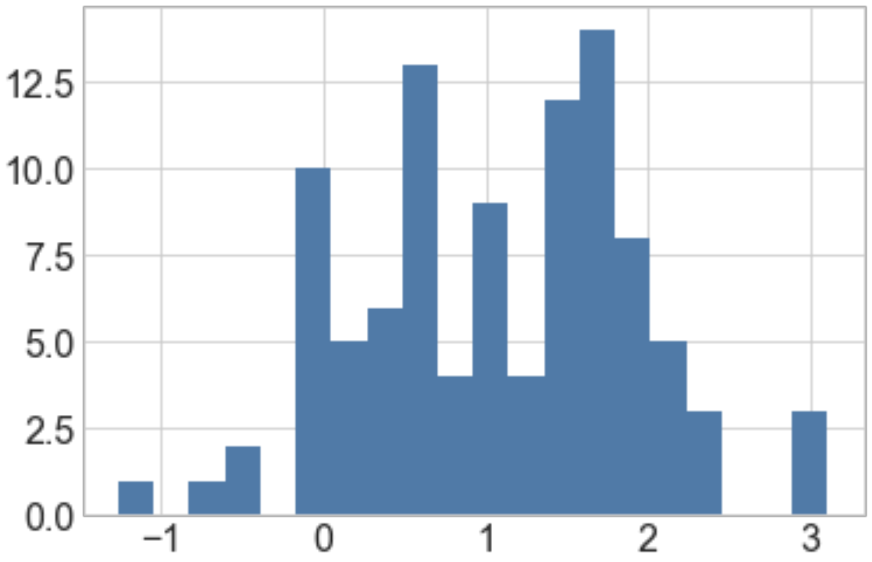
Область возможных значений обычно дробят на отрезки, **бины.** Чем короче бины, тем детальнее рисуется гистограмма.

Сколько значений попали в текущий отрезок (бин)

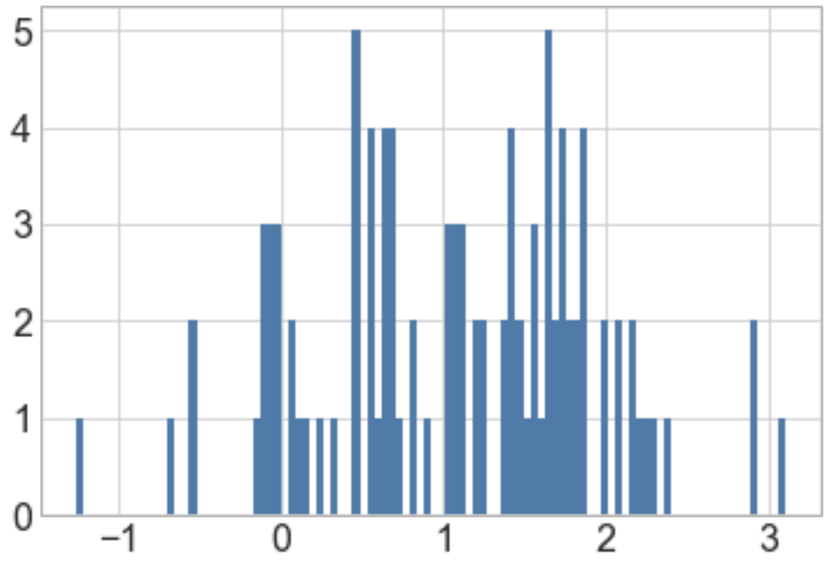


Гистограмма

- Длина интервала h (бина) должна быть достаточно большой, чтобы в него попало существенное число наблюдений
- И при этом достаточно малой, чтобы не потерять важные детали распределения



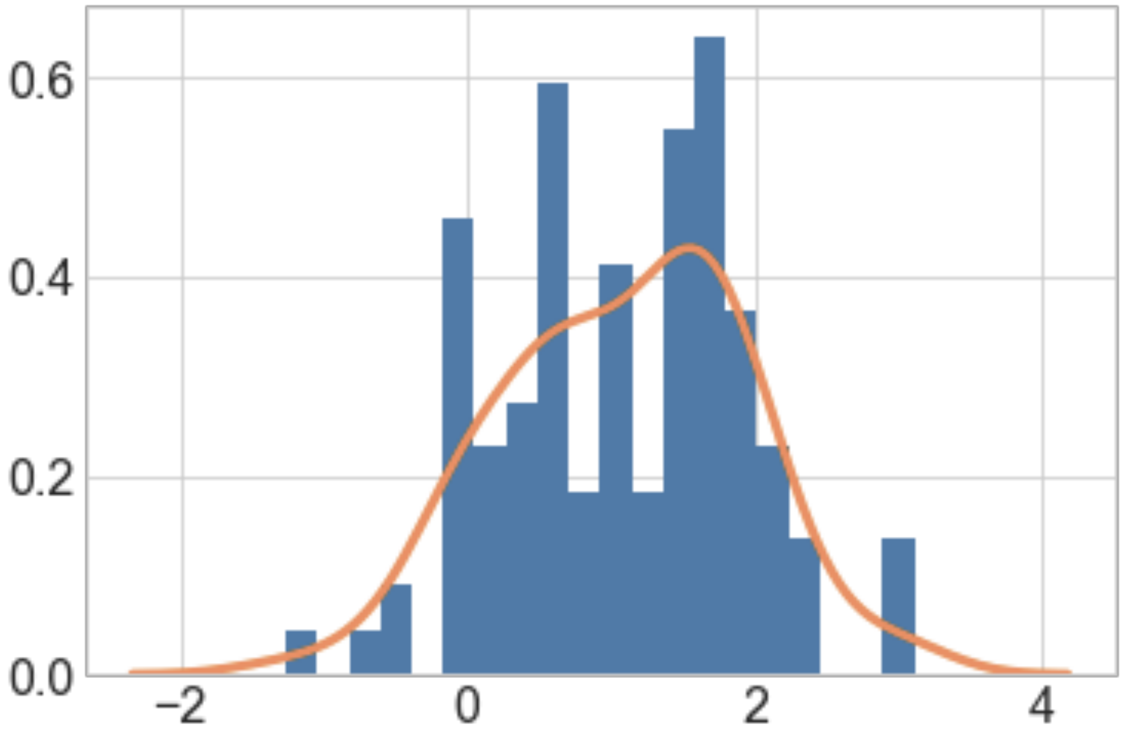
20 бинов



100 бинов

Гистограмма

- Ядерные оценки плотности (kernel density estimation, KDE) позволяют получить график плотности в виде непрерывной кривой



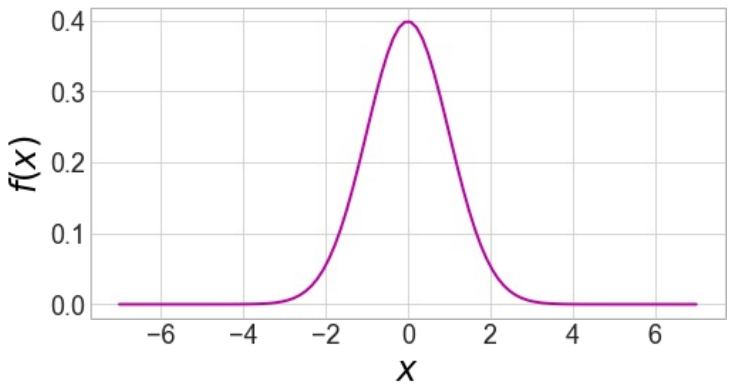
Ядерная оценка плотности

Чтобы взвесить наблюдения, функцию $K(z)$ (ядерную функцию) выбирают так, чтобы:

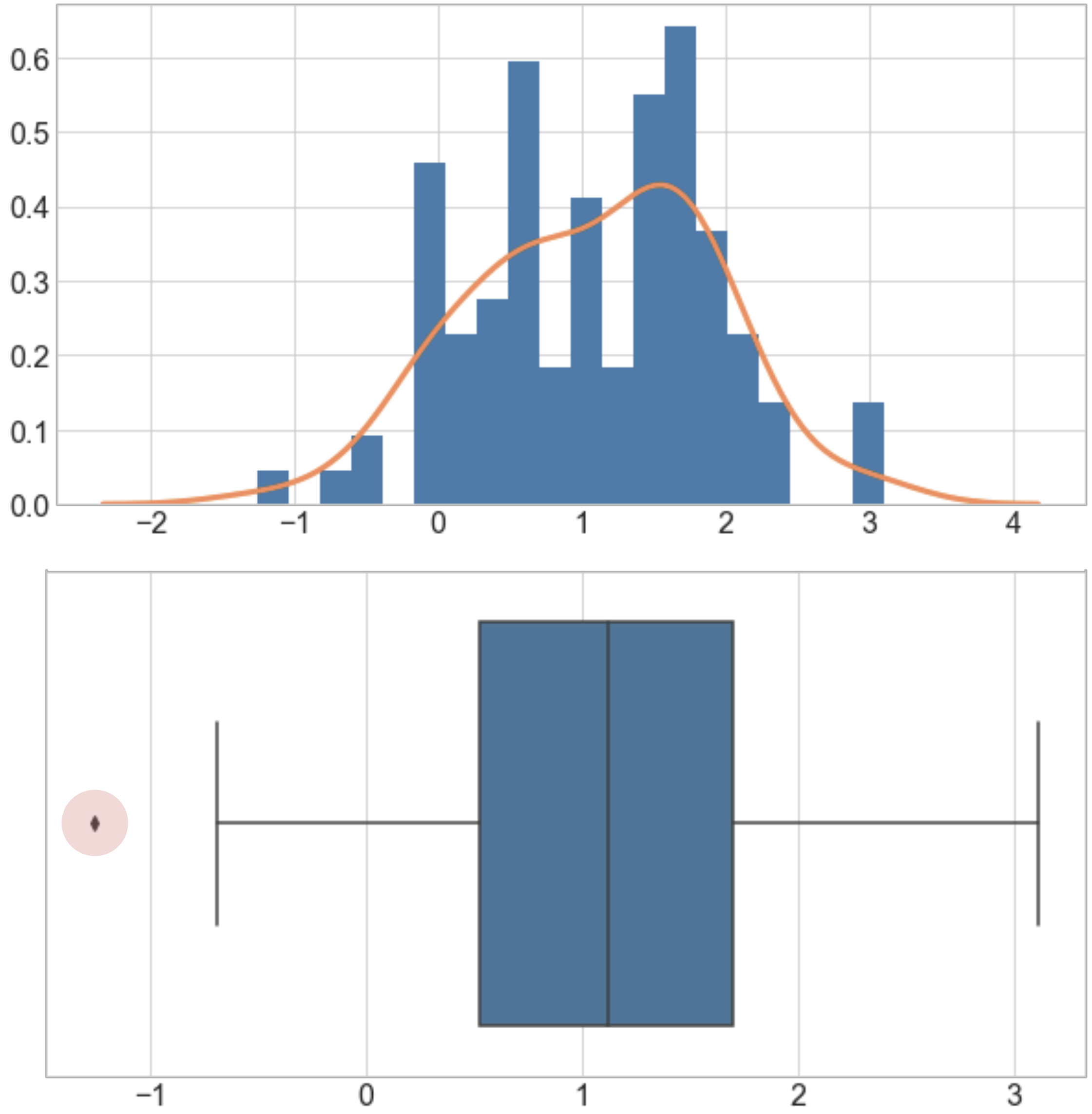
- Она была неотрицательной
- $\int K(z)\,dz = 1$ (сумма всех весов равна 1)

Ядерные функции бывают разными, чаще всего используют Гауссовское ядро:

$$K(z) = \frac{1}{\sqrt{2\,\pi}} \cdot e^{-\frac{x^2}{2}}$$



Ящик с усами



Аномальное
значение

Нормальное распределение

Основные обозначения

X, Y, Z – случайные величины

x, y, z – какие–то конкретные значения

A, B, C – события

$\mathbb{P}$– вероятность

$\mathbb{E}(X)$ – математическое ожидание

$Var(X), \sigma^2$ – дисперсия

$\mathrm{Cov}(X), \rho(X, Y)$ – ковариация и корреляция

Нормальное распределение

- В статистике часто встречается нормальное распределение
- Оно используется для проверки гипотез и для того, чтобы понимать насколько точными у нас получаются прогнозы и оценки
- Его обычно используют, когда у нас есть в распоряжении большая выборка
- Давайте познакомиться с нормальным распределением поближе

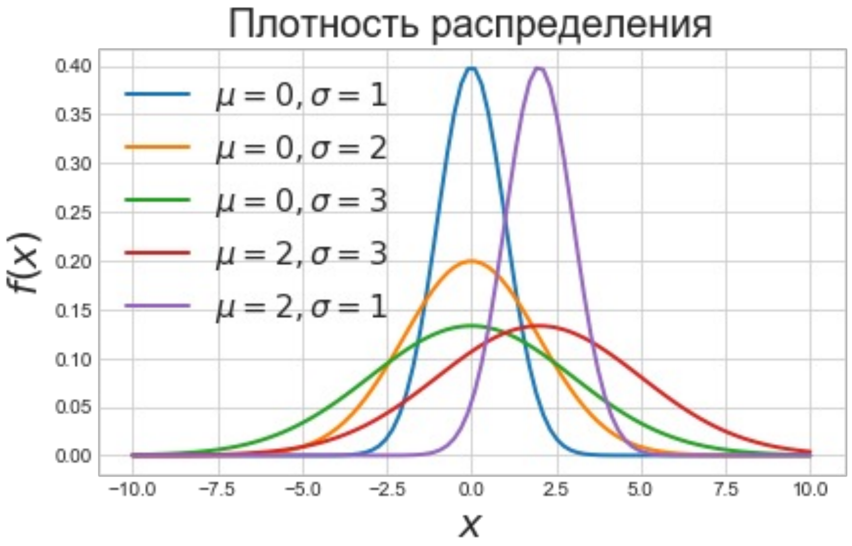
Нормальное распределение

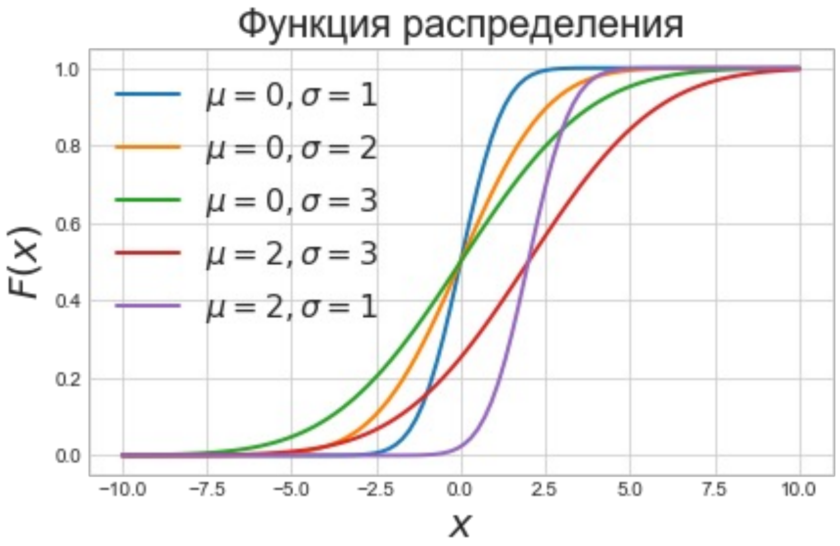
Нормальная случайная величина:

$$X \sim N(\mu, \sigma^2)$$

$$\mathbb{E}(X) = \mu,\ Var(X) = \sigma^2$$

Функцию распределения нельзя найти в аналитическом виде, интеграл не берётся

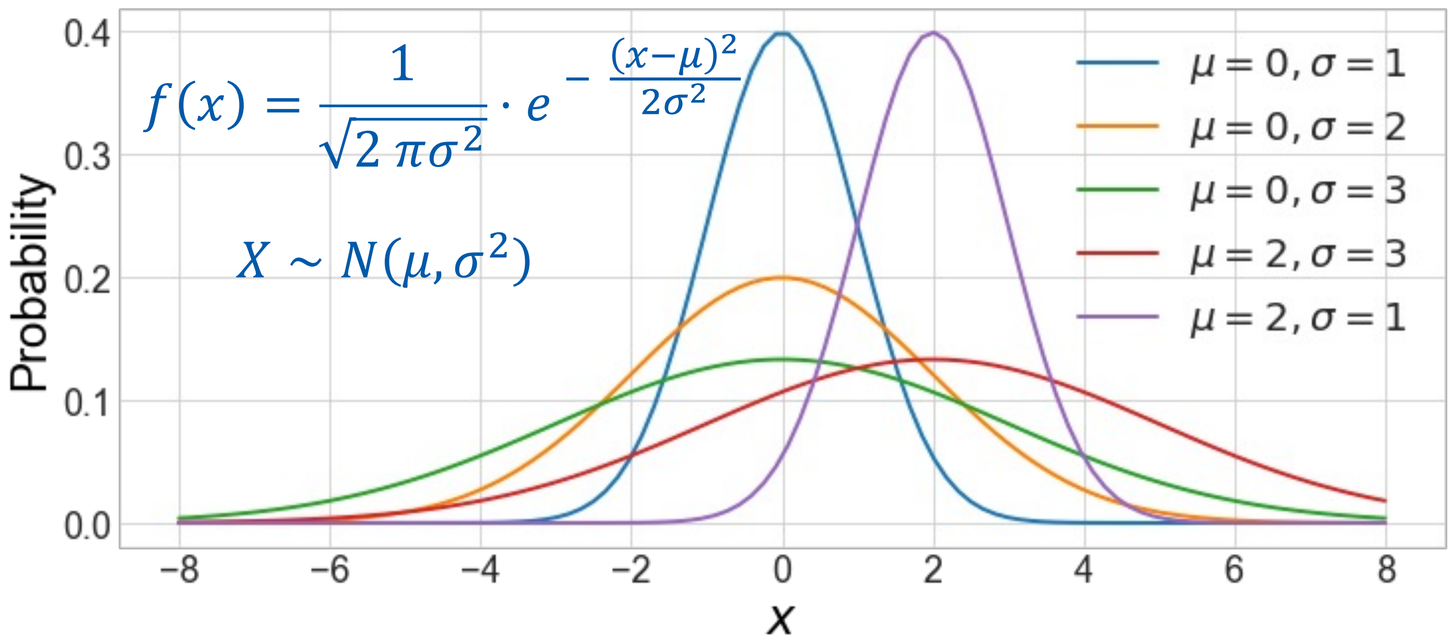




$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$F(x) = \int_{-\infty}^{x} f(x)\ dx$$

Свойства нормального распределения



1. Распределение симметрично относительно точки $\mathbb{E}(X) = \mu$

2. Параметр μ не влияет на форму кривой и отвечает за её сдвиг кривой вдоль оси x, параметр σ определяет степень “размытости” кривой

Свойства нормального распределения

$X \sim N(\mu_x, \sigma_x^2)$ $\quad Y \sim N(\mu_y, \sigma_y^2)$ $\quad a$ — константа

3. $X + Y \sim N(\mu_x + \mu_y,\ \sigma_x^2 + \sigma_y^2)$

4. $X + a \sim N(\mu_x + a,\ \sigma_x^2)$

5. $a \cdot X \sim N(a \cdot \mu_x,\ a^2 \cdot \sigma_x^2)$

Нормальная случайная величина устойчива
к суммированию и линейным преобразованиям

Центрирование и стандартизация

$$X \sim N(\mu, \sigma^2)$$

центрирование

$$X - \mu \sim N(0, \sigma^2)$$

стандартизация

$$\frac{X - \mu}{\sqrt{\sigma^2}} \sim N(0, 1)$$

- Распределение $N(0, 1)$ называется **стандартным нормальным распределением**

Стандартное нормальное распределение

- Функцию распределения для нормального распределения нельзя найти в аналитическом виде
- Для функции распределения случайной величины $N(0, 1)$ составлены таблицы

Как найти вероятность

$X \sim N(7, 16)$

$\mathbb{P}(X \leq 15)$

Искать такую вероятность неудобно, нужны были бы таблицы для всех возможных μ и σ

Как найти вероятность

$$X \sim N(7, 16)$$

$$\mathbb{P}(X \leq 15) = \mathbb{P}\left(\frac{X-7}{4} \leq \frac{15-7}{4}\right)$$

$$= \mathbb{P}(N(0,1) \leq 2) \ = F_{N(0,1)}(2) \ = \Phi(2) \approx 0.98$$

Обозначение
для функции
распределения $N(0,1)$

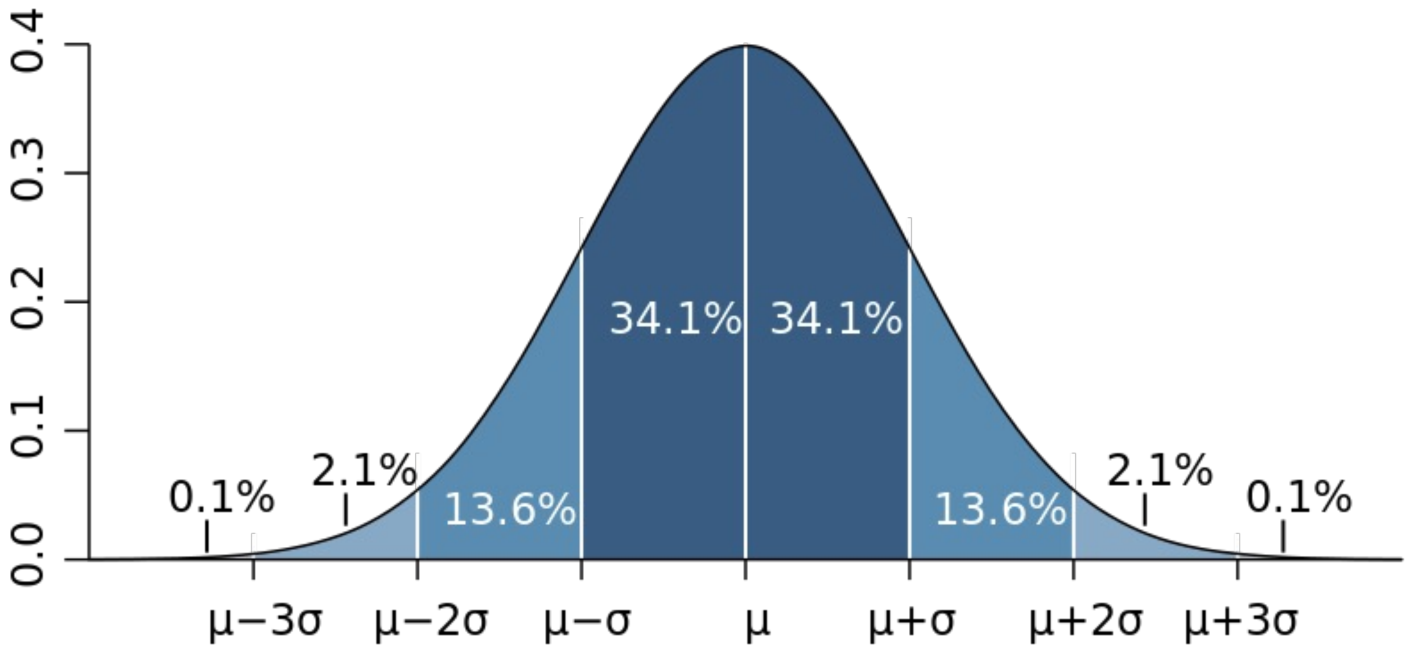
Раньше активно пользовались таблицами для распределения $N(0, 1)$, сегодня для любого распределения расчёты делает компьютер

Правила сигм

$$X \sim N(\mu, \sigma^2)$$

Правило сигмы:

$$\mathbb{P}(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0.68$$

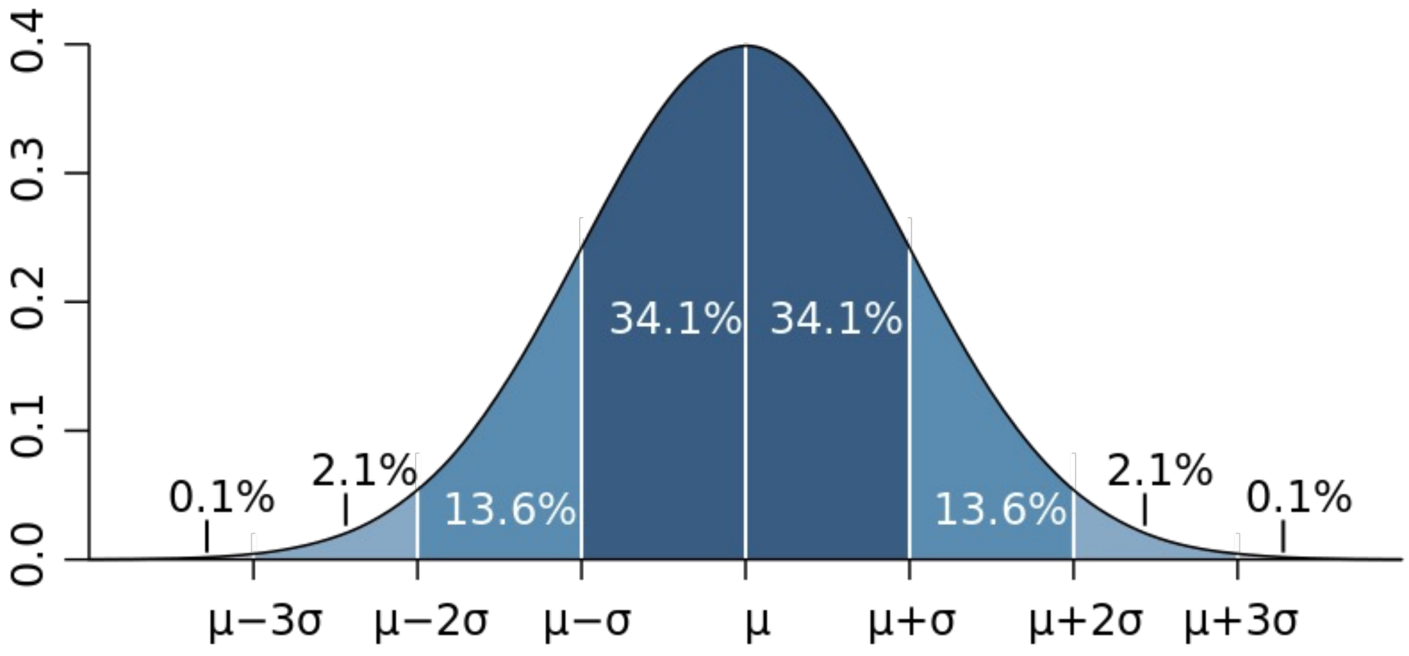


Правила сигм

$$X \sim N(\mu, \sigma^2)$$

Правило двух сигм:

$$\mathbb{P}(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.95$$

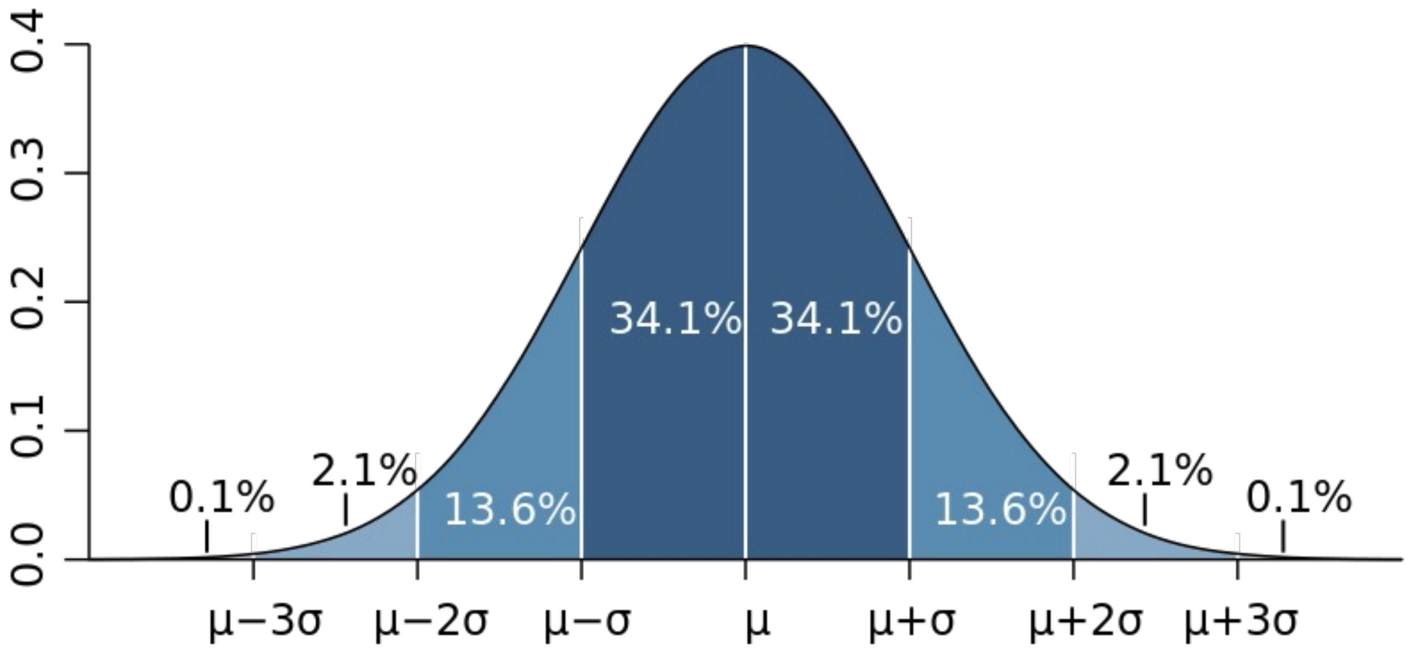


Правила сигм

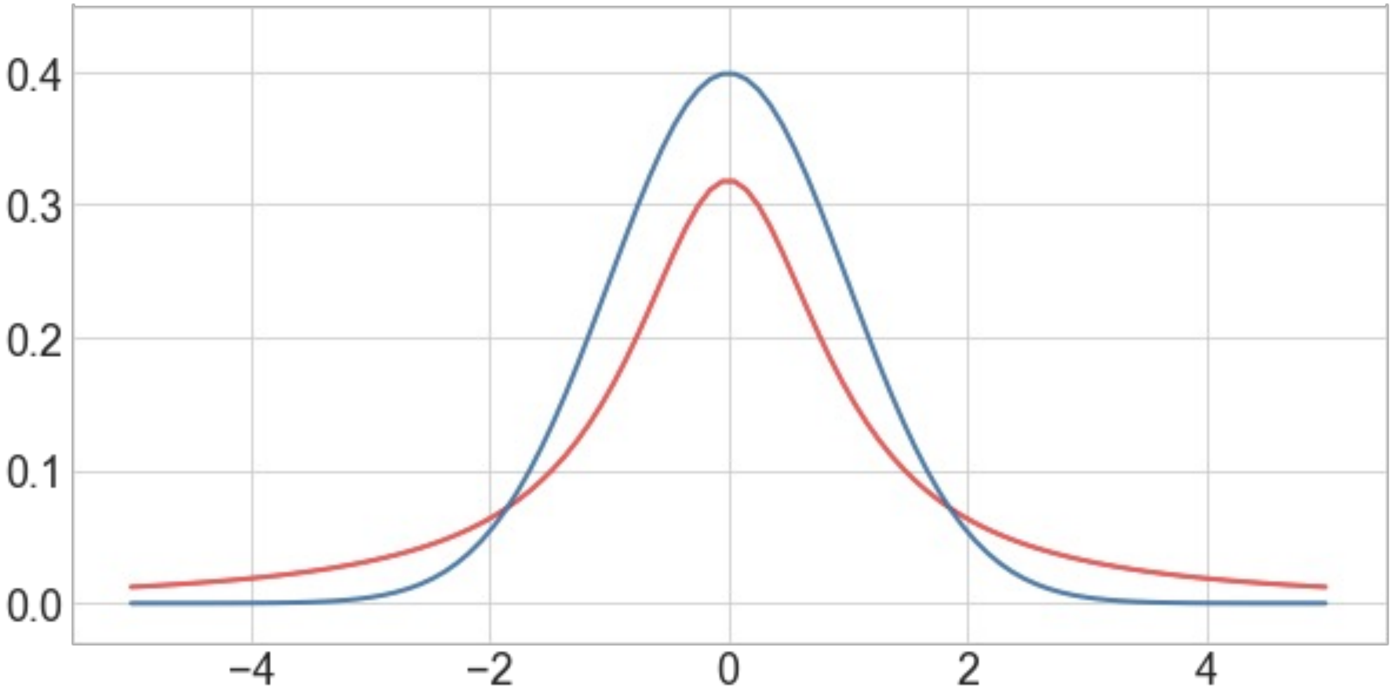
$$X \sim N(\mu, \sigma^2)$$

Правило трех сигм:

$$\mathbb{P}(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.997$$



Тяжёлые хвосты



- Хвосты красного распределения тяжёлые
- Под ними сосредоточена большая вероятностная масса
- События из-под них (выбросы) более вероятны

Эксцесс и куртосис

Эксцессом случайной величины X называют величину

$$\beta_X = \frac{\mathbb{E}[\left(X - \mathbb{E}(X)\right)^4]}{\sigma^4} - 3$$

Куртосис

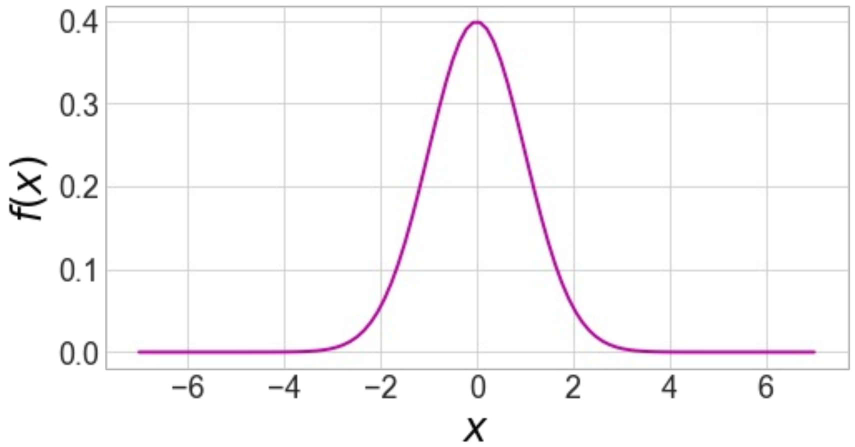
- Число 3 вычитается из куртосиса, чтобы эксцесс нормального распределения был равен нулю
- Если хвосты распределения легче, а пик острее, чем у нормального распределения, тогда $\beta_X > 0$
- Если хвосты распределения тяжелее, а пик более приплюснутый, тогда $\beta_X < 0$

Эксцесс и куртосис

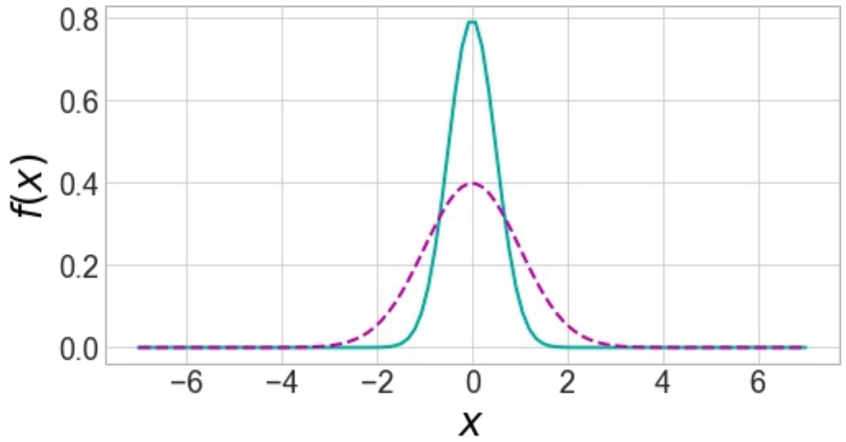
Эксцессом случайной величины X называют величину

$$\beta_X = \frac{\mathbb{E}[(X - \mathbb{E}(X))^4]}{\sigma^4} - 3$$

Куртосис



Нормальное распределение
с нулевым эксцессом



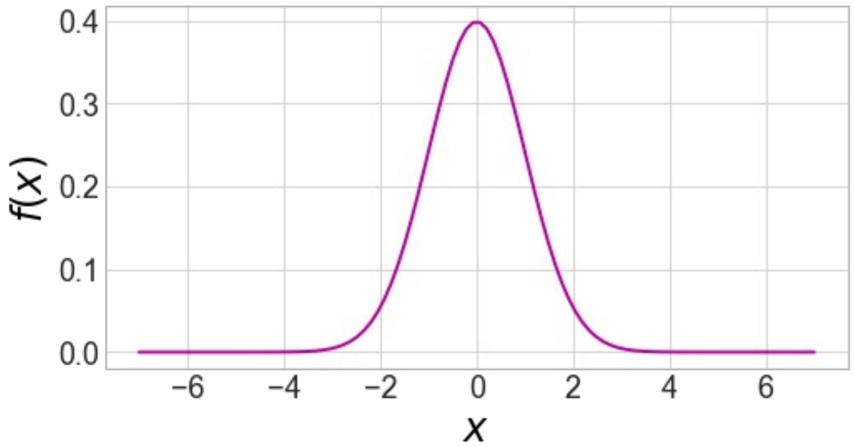
Положительный эксцесс

Эксцесс и куртосис

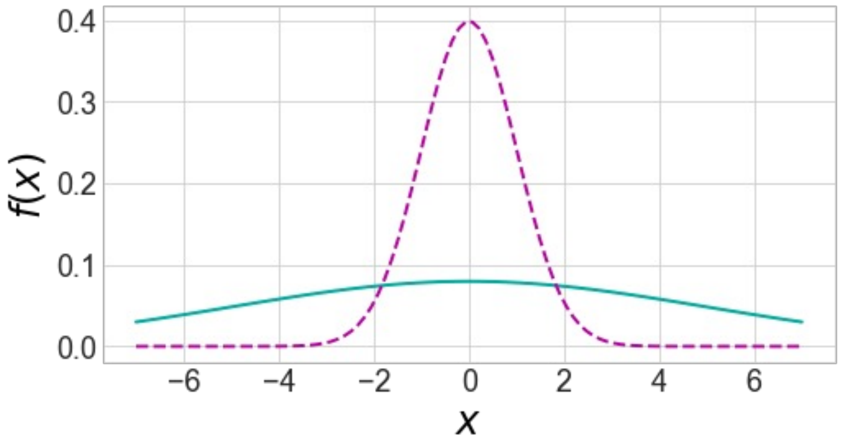
Эксцессом случайной величины X называют величину

$$\beta_X = \frac{\mathbb{E}[(X - \mathbb{E}(X))^4]}{\sigma^4} - 3$$

Куртосис



Нормальное распределение с нулевым эксцессом



Отрицательный эксцесс

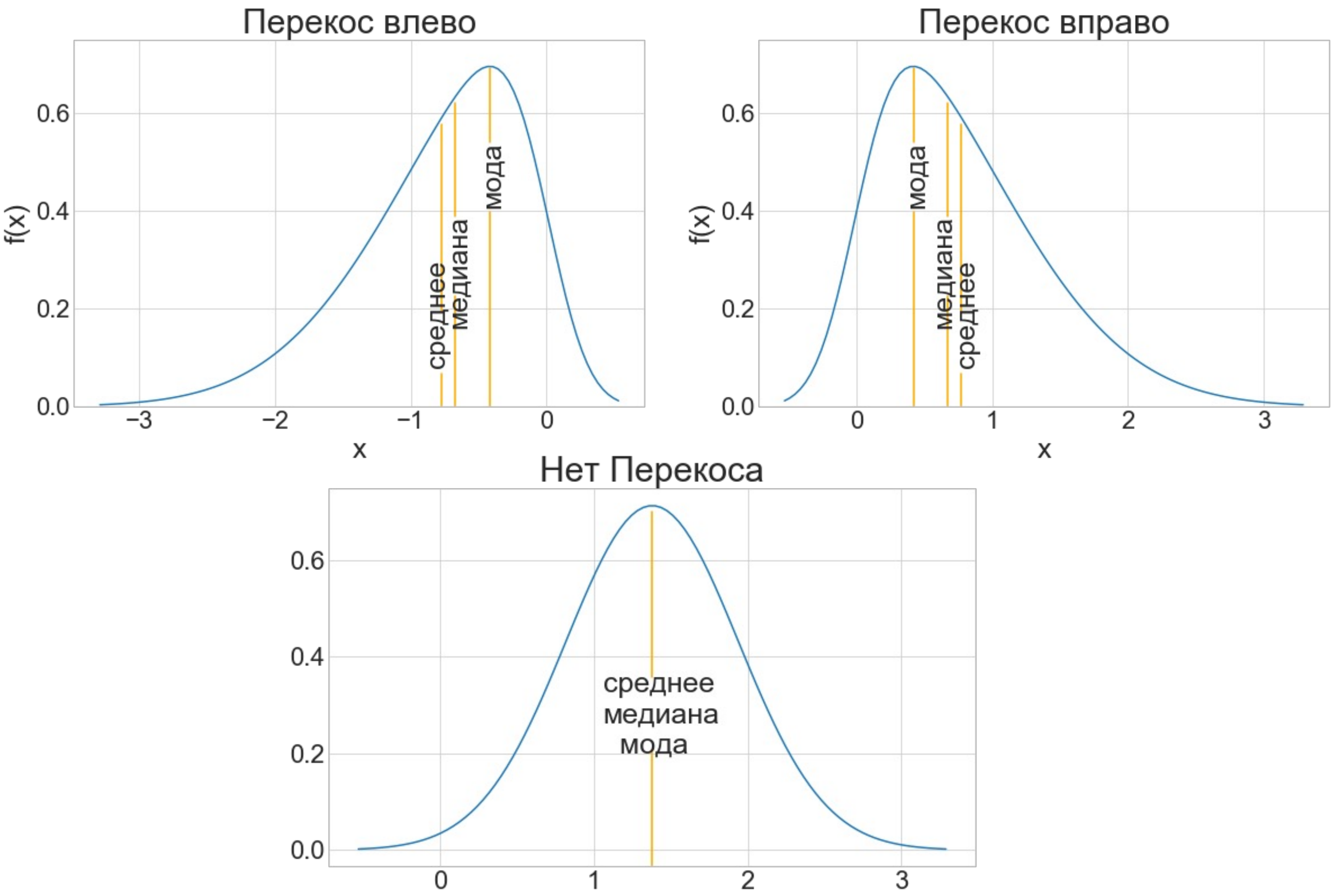
Коэффициент асимметрии (skewness)

Коэффициентом асимметрии случайной величины X называют величину

$$A_X = \frac{\mathbb{E}\left[\left(X - \mathbb{E}(X)\right)^3\right]}{\sigma^3}$$

- Если плотность распределения симметрична, то $\mathrm{A_X} = 0$
- Если левый хвост тяжелее, то $\mathrm{A_X} > 0$
- Если правый хвост тяжелее, то $\mathrm{A_X} < 0$

Коэффициент асимметрии (skewness)



Эксцесс и асимметрия

- Эксцесс оказывается полезным при поиске тяжёлых хвостов
- Большое значение эксцесса сигнализирует о наличии тяжёлых хвостов и выбросов в данных
- Коэффициент асимметрии характеризует перекос в распределении
- Если у распределения сильный перекос, с применением стандартных статистических методов возникают сложности