

Распределения и описательные статистики

План

- Узнаем, есть ли в мире случайности
- Обсудим процесс порождения данных и невежество
- Вспомним основные понятия теории вероятностей
- Поговорим, какими бывают распределения
- Научимся считать описательные статистики
- Разберёмся в выборках и их свойствах

Пакт

Заключим соглашение!

X, Y, Z – случайные величины

x, y, z – какие-то конкретные значения

A, B, C – события

\mathbb{P} – вероятность

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

$\text{Var}(X)$ – дисперсия

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

Бывают ли в мире случайности

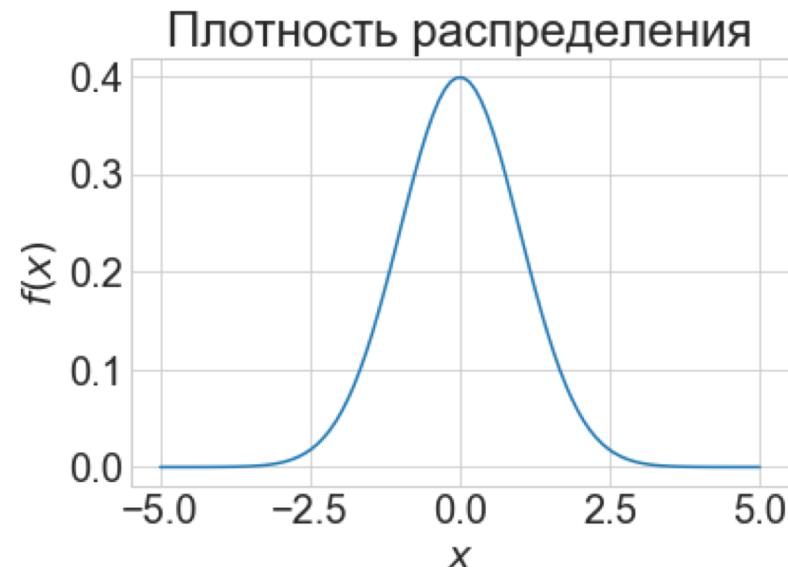
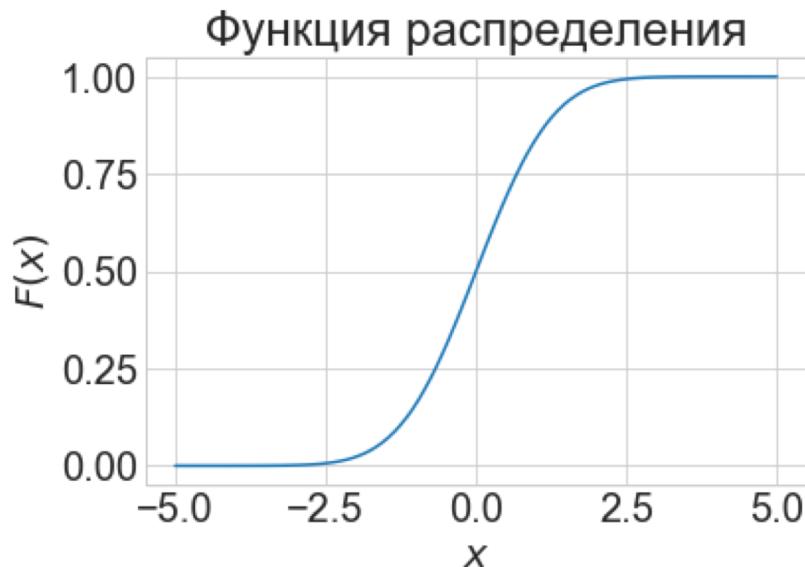
Как устроен мир



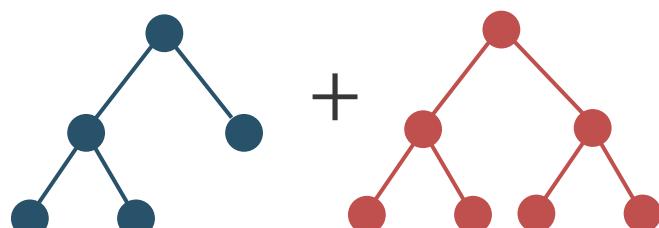
X

- **Сундук** – различные процессы порождения данных. Теория вероятностей изучает этот сундук. В реальности мы не видим его.
- Сундук порождает **выборки**. Математическая статистика изучает их и пытается восстановить внутренности сундука.

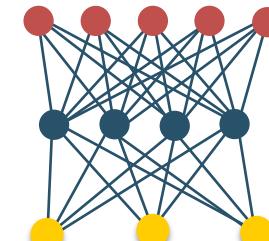
Устройство сундука



Модель – наше предположение о том, как сундук устроен.
За каждой моделью стоят какие-то предпосылки,
описывающие наше незнание.



$$M = U V^T$$



Что извергает сундук

- Сундук порождает выборки
- Мы пытаемся по ним восстановить его структуру
- Делаем это в рамках выбранной модели



Что мы будем делать?

- Изучать выборки из сундука и их свойства
- Предполагать, что внутри сундука, описывать своё незнание с помощью какой-то модели
- Разбираться, насколько наши предположения согласуются с выборками – данными, предоставленными нам сундуком



Как устроены внутренности сундука: распределение

Случайная величина

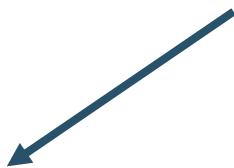


X

- **Сундук** – различные процессы порождения данных. Теория вероятностей изучает этот сундук. В реальности мы не видим его.
- Сундук порождает **выборки**. Математическая статистика изучает их, и пытается восстановить внутренности сундука.

Какими бывают случайные величины

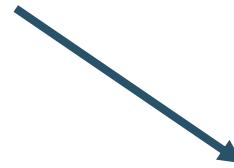
Случайные величины



Дискретные

Множество значений
конечно или счётно

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



Непрерывные

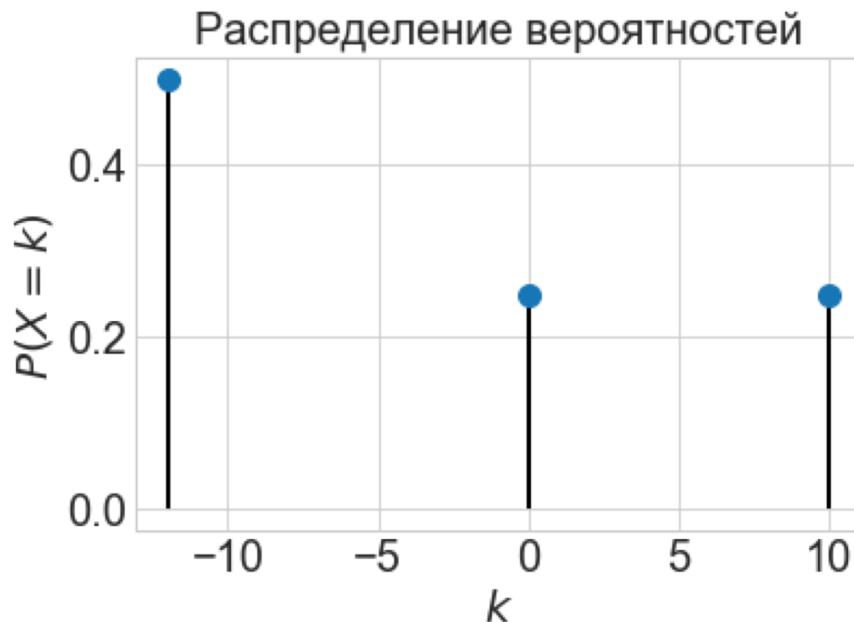
Принимают бесконечное,
континуальное число
значений

(рост, время ожидания
автобуса, вес)

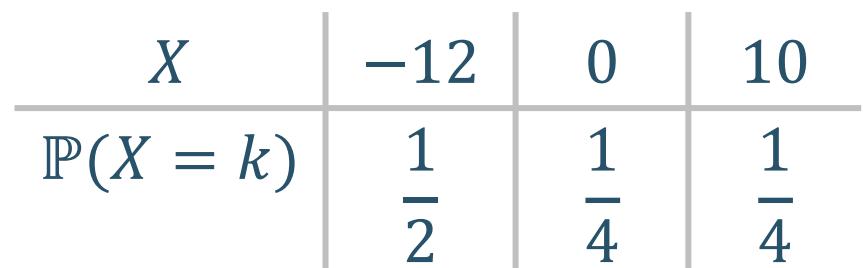
Дискретные случайные величины

Распределение дискретной случайной величины – таблица, которая описывает, какие значения принимает случайная величина с какой вероятностью

Сумма вероятностей должна быть равна 1, каждая вероятность лежит между 0 и 1



Пример: лотерея

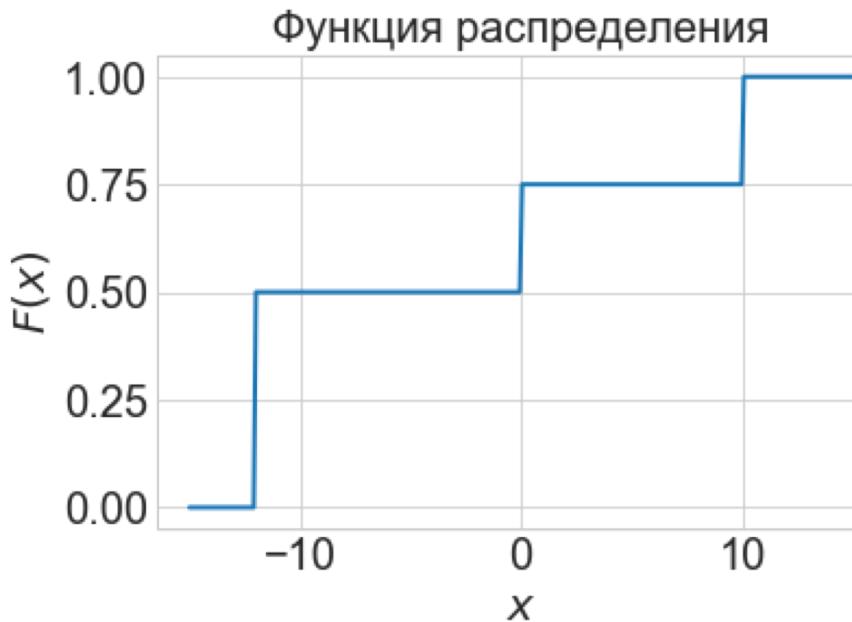


Дискретные случайные величины

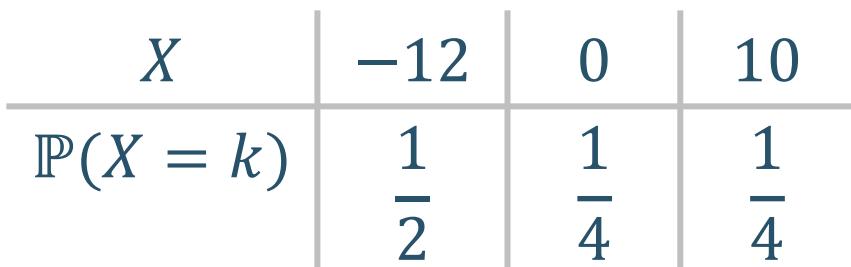
Функция распределения – функция, которая определяет вероятность события $X \leq x$, то есть

$$F(x) = \mathbb{P}(X \leq x) = \sum \mathbb{P}(X = k) \cdot [k \leq x],$$

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

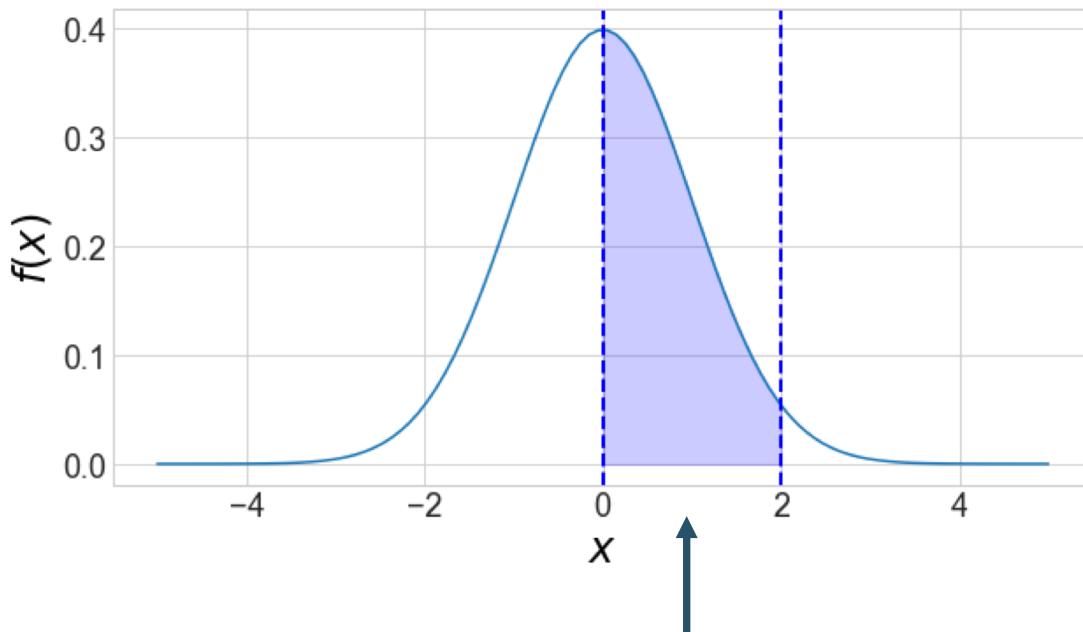


Пример: лотерея



Непрерывные случайные величины

Распределение непрерывной случайной величины описывается **плотностью распределения вероятностей**.



Площадь равна вероятности попасть
на отрезок от нуля до двух

Пример:
нормальное
распределение

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

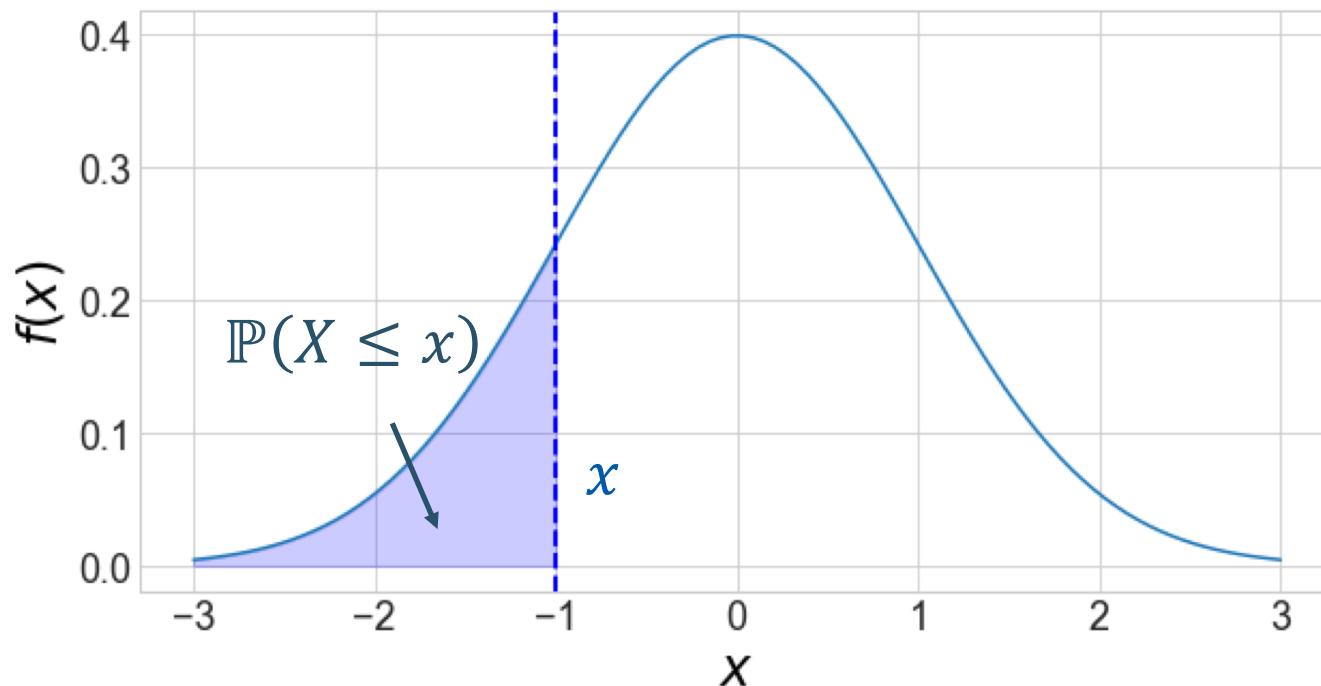
$$= \int_0^2 f(x) \, dx$$

Площадь под всей плотностью должна быть равна 1

Непрерывные случайные величины

Функция распределения – функция, которая определяет вероятность события $X \leq x$, то есть

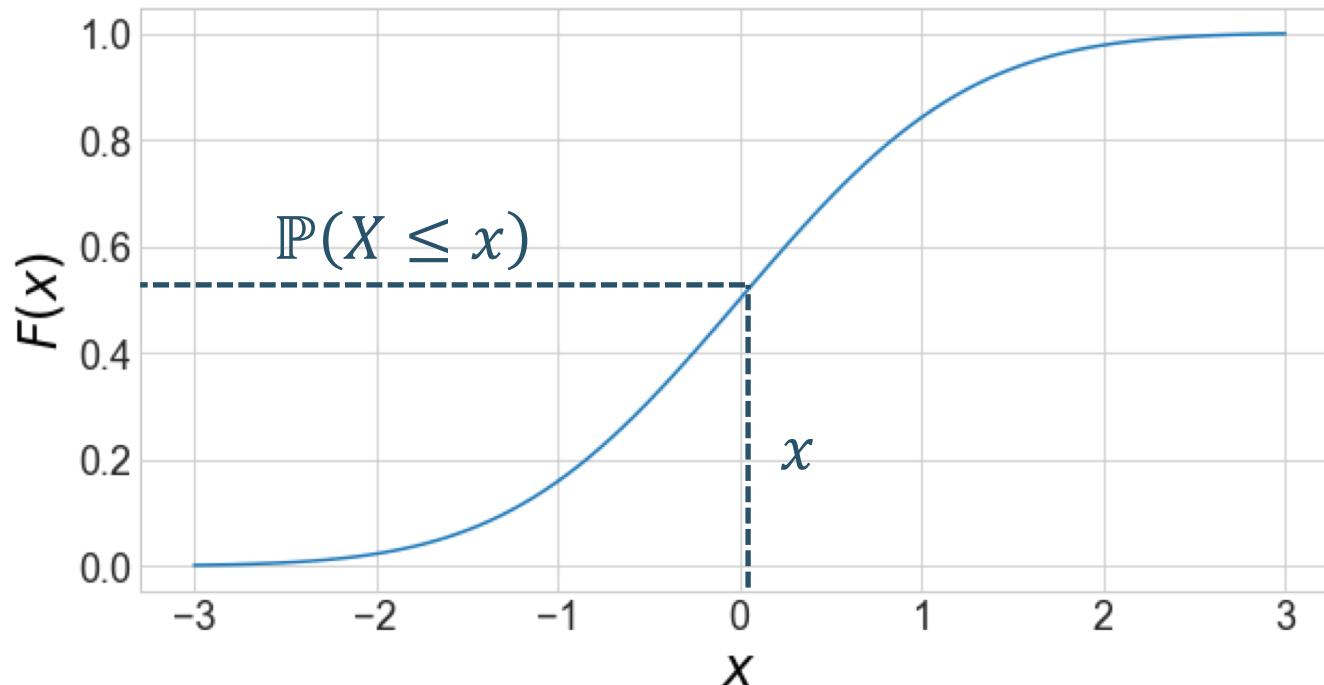
$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x f(t) \, dt, f(t) \text{ – плотность}$$



Непрерывные случайные величины

Функция распределения – функция, которая определяет вероятность события $X \leq x$, то есть

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x f(t) \, dt, f(t) \text{ – плотность}$$



Важные свойства

1. Плотность определена только для непрерывных случайных величин
2. $f(x) = F'(x)$
3. $\int_{-\infty}^{+\infty} f(t) \ dt = 1, \quad f(t) \geq 0 \quad \forall t$
4. $F(x)$ не убывает, лежит между 0 и 1
5. $\mathbb{P}(a \leq X \leq b) = \int_a^b f(t) \ dt = F(b) - F(a)$
6. Вероятность того, что непрерывная случайная величина попадёт в точку, равна нулю

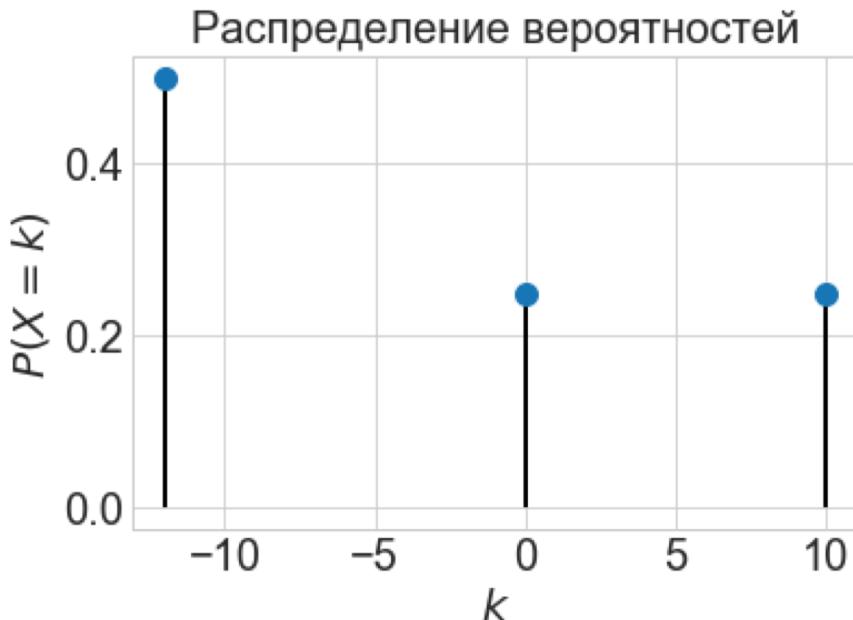
Характеристики случайных величин

Математическое ожидание

Математическое ожидание – среднее значение случайной величины

$$\mathbb{E}(X) = \sum_{k=1}^n k \cdot \mathbb{P}(X = k)$$

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} t \cdot f(t) dt$$



Пример: лотерея

X	-12	0	10
$\mathbb{P}(X = k)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$

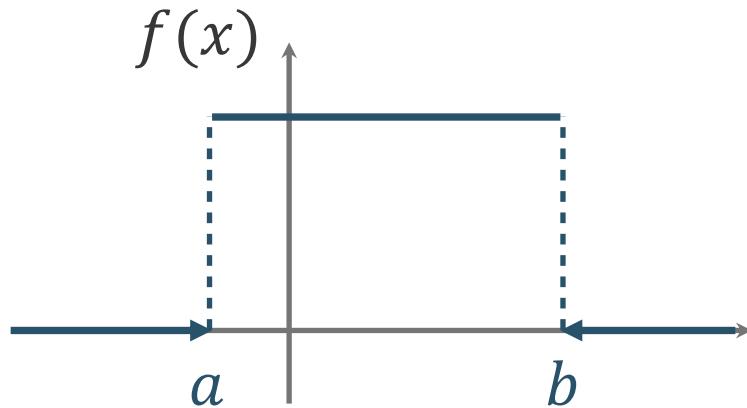
$$\mathbb{E}(X) = -12 \cdot 0.5 + 0 \cdot 0.25 + 10 \cdot 0.25 = -3.5 \text{ рубля}$$

Математическое ожидание

Математическое ожидание – среднее значение случайной величины

$$\mathbb{E}(X) = \sum_{k=1}^n k \cdot \mathbb{P}(X = k)$$

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} t \cdot f(t) dt$$



Пример: равномерное

$$f(x) = \frac{1}{b-a}, x \in [a; b]$$

$$\mathbb{E}(X) = \int_a^b t \cdot \frac{1}{b-a} dt = \frac{1}{b-a} \cdot \frac{t^2}{2} \Big|_a^b = \frac{(b^2 - a^2)}{2(b-a)} = \frac{a+b}{2}$$

Математическим ожиданием оказывается середина отрезка

Свойства математического ожидания

X, Y – случайные величины a – константа

1. $\mathbb{E}(a) = a$
2. $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$
3. $\mathbb{E}(a \cdot X) = a \cdot \mathbb{E}(X)$
4. $\mathbb{E}(X \cdot Y) = \mathbb{E}(X) \cdot \mathbb{E}(Y)$, если независимы
5. Математическое ожидание случайной величины – не случайно
6. $\mathbb{E}(X - \mathbb{E}(X)) = \mathbb{E}(X) - \mathbb{E}(\mathbb{E}(X)) = \mathbb{E}(X) - \mathbb{E}(X) = 0$

Дисперсия

Дисперсия – мера разброса случайной величины вокруг её среднего

$$Var(X) = \mathbb{E}(X - \mathbb{E}(X))^2 = \sum_{k=1}^n (k - \mathbb{E}(X))^2 \cdot \mathbb{P}(X = k)$$

$$Var(X) = \mathbb{E}(X - \mathbb{E}(X))^2 = \int_{-\infty}^{+\infty} (t - \mathbb{E}(X))^2 \cdot f(t) dt$$

Дисперсия

Дисперсия – мера разброса случайной величины вокруг её среднего

Более удобно искать дисперсию по формуле:

$$\begin{aligned}Var(X) &= \mathbb{E}(X - \mathbb{E}(X))^2 \\&= \mathbb{E}(X^2 - 2 \cdot X \cdot \mathbb{E}(X) + \mathbb{E}^2(X)) \\&= \mathbb{E}(X^2) - 2 \cdot \mathbb{E}(X) \cdot \mathbb{E}(\mathbb{E}(X)) + \mathbb{E}^2(X) \\&= \mathbb{E}(X^2) - 2 \cdot \mathbb{E}(X) \cdot \mathbb{E}(X) + \mathbb{E}^2(X) \\&= \mathbb{E}(X^2) - \mathbb{E}^2(X)\end{aligned}$$

Дисперсия

Дисперсия – мера разброса случайной величины вокруг её среднего

$$Var(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X)$$

Пример: лотерея

X	-12	0	10
$\mathbb{P}(X = k)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$

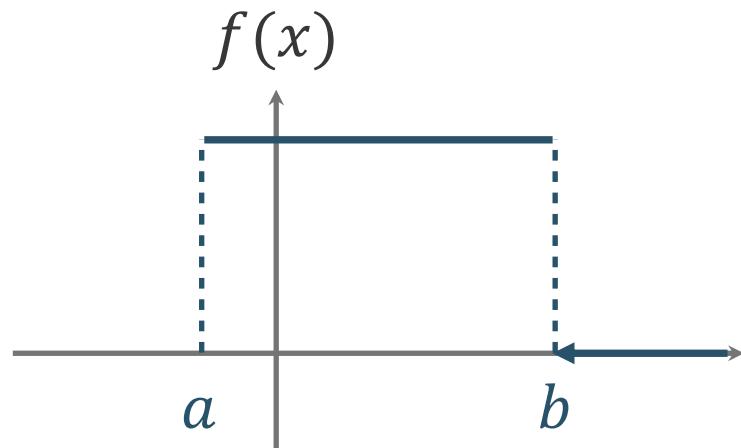
$$\mathbb{E}(X^2) = (-12)^2 \cdot 0.5 + 0^2 \cdot 0.25 + 10^2 \cdot 0.25 = 97 \text{ рублей}^2$$

$$Var(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X) = 97 - 12.25 = 84.75 \text{ рублей}^2$$

Дисперсия

Дисперсия – мера разброса случайной величины вокруг её среднего

$$Var(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X)$$



Пример: равномерное

$$f(x) = \frac{1}{b-a}, x \in [a; b]$$

$$\mathbb{E}(X^2) = \int_{-\infty}^{+\infty} t^2 \cdot \frac{1}{b-a} dt = \frac{1}{b-a} \cdot \frac{t^3}{3} \Big|_a^b = \frac{(b^3 - a^3)}{3(b-a)} = \frac{a^2 + ab + b^2}{2}$$

$$Var(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X) = \frac{a^2 + ab + b^2}{2} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}$$

Среднеквадратическое отклонение

Дисперсия случайной величины имеет размерность, равную квадрату размерности самой величины

Чтобы вернуться к исходной размерности, из дисперсии часто извлекают корень и работают со среднеквадратическим отклонением:

$$\sigma(X) = \sqrt{Var(X)}$$

Свойства дисперсии

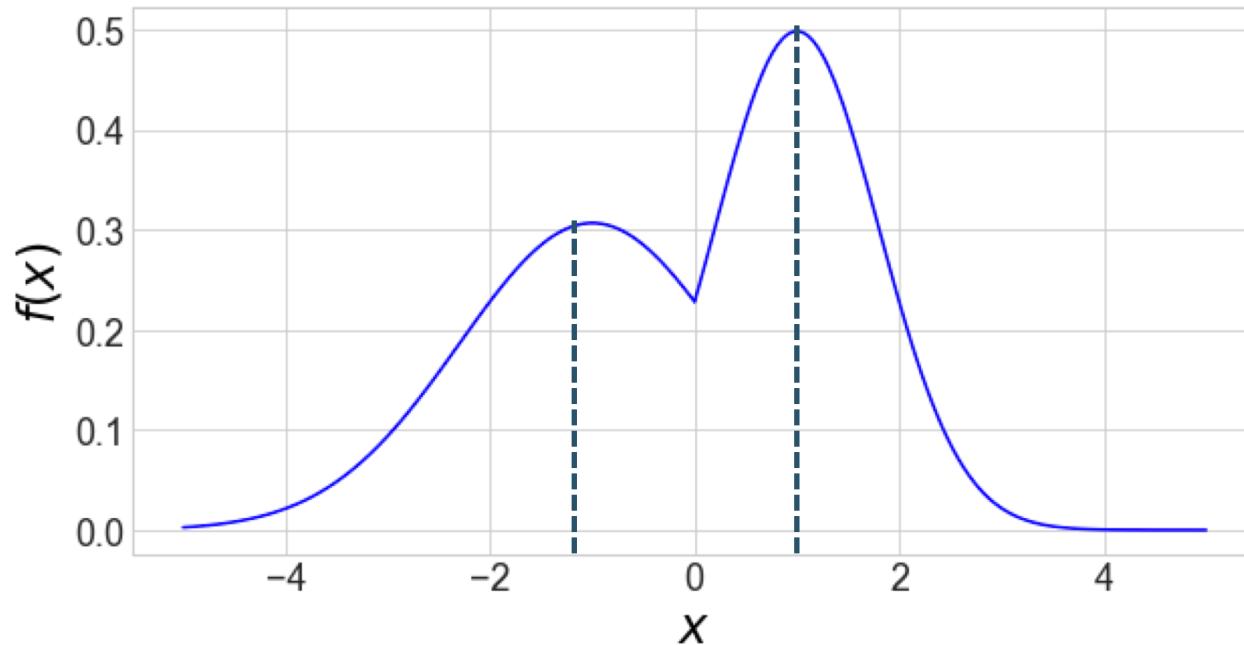
X, Y – случайные величины a – константа

1. $Var(a) = 0$
2. $Var(X + Y) = Var(X) + Var(Y)$, если независимы
3. $Var(a \cdot X) = a^2 \cdot Var(X)$
4. $Var(X - Y) = Var(X) + Var(Y)$, если независимы
5. Дисперсия случайной величины – не случайна

Мода

Мода случайной величины – значение, которому соответствует наибольшая вероятность (для дискретной случайной величины) и локальный максимум плотности распределения (для непрерывной случайной величины)

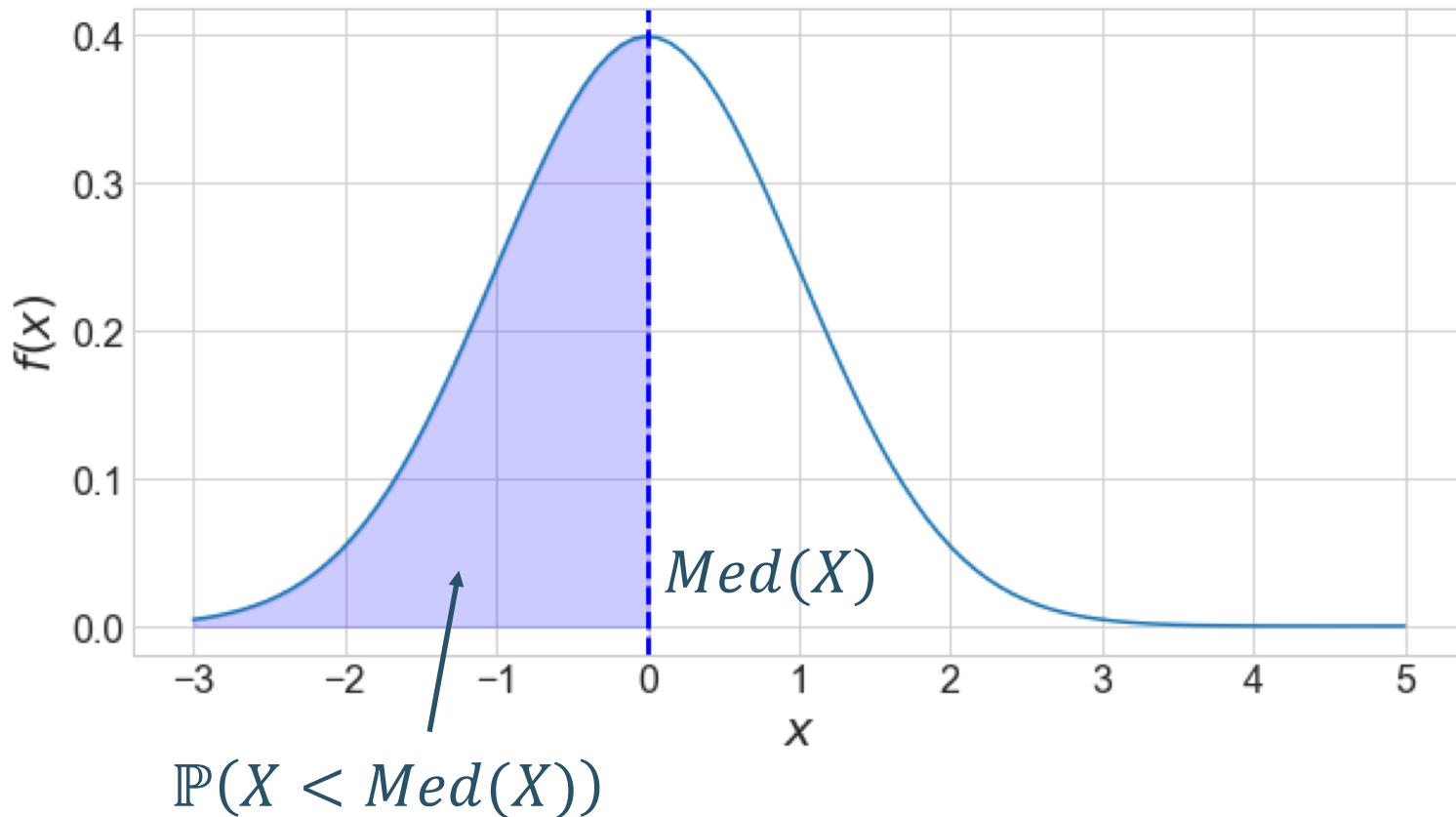
На практике встречаются мультимодальные распределения



Медиана

Медиана случайной величины – такое её значение, что

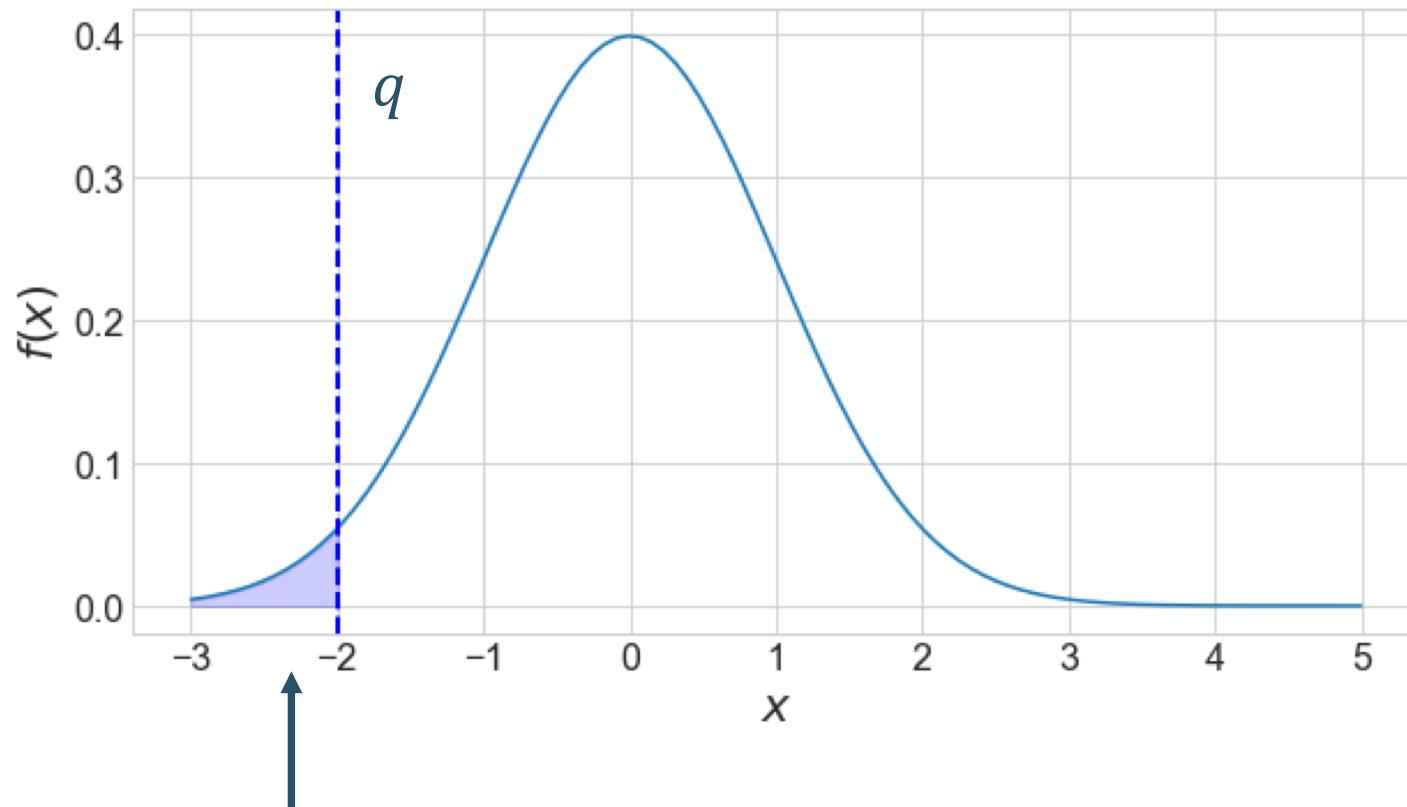
$$\mathbb{P}(X < \text{Med}(X)) = \mathbb{P}(X > \text{Med}(X)) = 0.5$$



Квантиль

Квантиль уровня γ – это такое число q , что

$$\mathbb{P}(X \leq q) = \gamma$$



Вероятность попасть в хвост равна γ

Резюме

- Мы вспомнили основные определения из теории вероятностей
- Мы поговорили про свойства математических ожиданий и дисперсий

Какими бывают случайные величины

Распределение Бернулли

- Пол родившегося ребёнка

		мальчик	девочка
X		0	1
$\mathbb{P}(X = k)$		$1 - p$	p

Распределение Бернулли:

$$X \sim Bern(p)$$

$$\mathbb{E}(X) = 1 \cdot p + 0 \cdot (1 - p) = p$$

$$Var(X) = E(X^2) - E^2(X) = p - p^2 = p \cdot (1 - p)$$

Биномиальное распределение

- Число попаданий в баскетбольную корзину

Биномиальная случайная величина: $X \sim Bin(p, n)$

n – число испытаний

p – вероятность успеха



Futurama s03 e14. Автор Мэтт Грейнинг. FOX Network.

$$\mathbb{P}(X = k) = C_n^k \cdot p^k (1 - p)^{n-k}$$

k принимает значения от 0 до n

Биномиальное распределение

$$Y_i \sim Bern(p)$$

$$\mathbb{E}(X) = n \cdot p$$

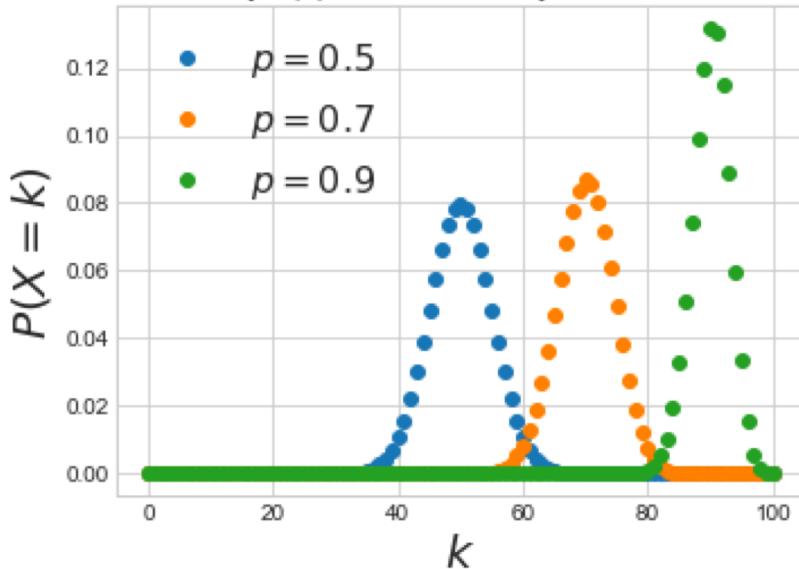
$$X = Y_1 + \dots + Y_n$$

$$Var(X) = n \cdot p \cdot (1 - p)$$

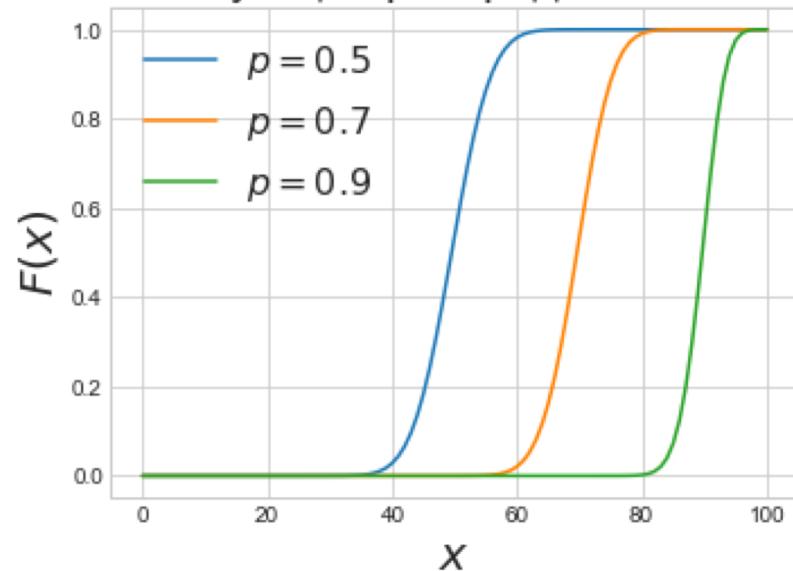
$$X \sim Bin(p, n)$$

$$\mathbb{P}(X = k) = C_n^k \cdot p^k (1 - p)^{n-k}$$

Распределение вероятностей



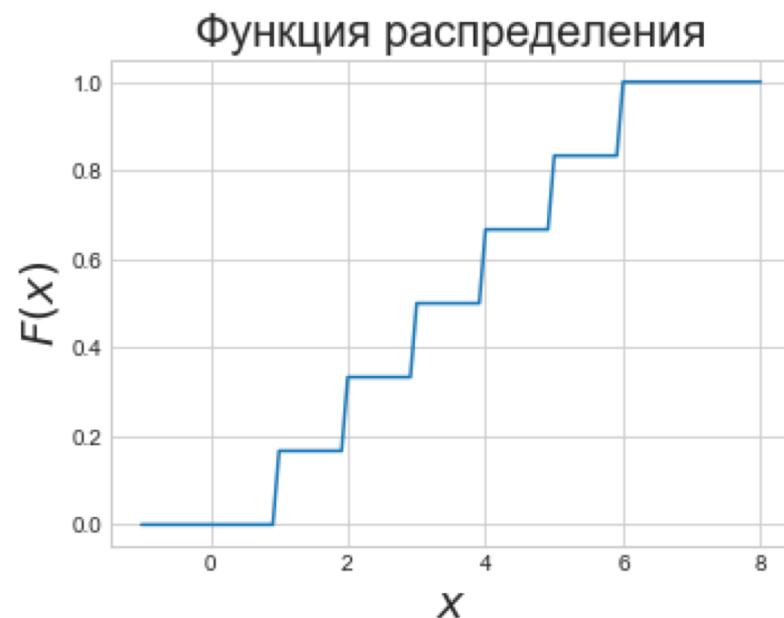
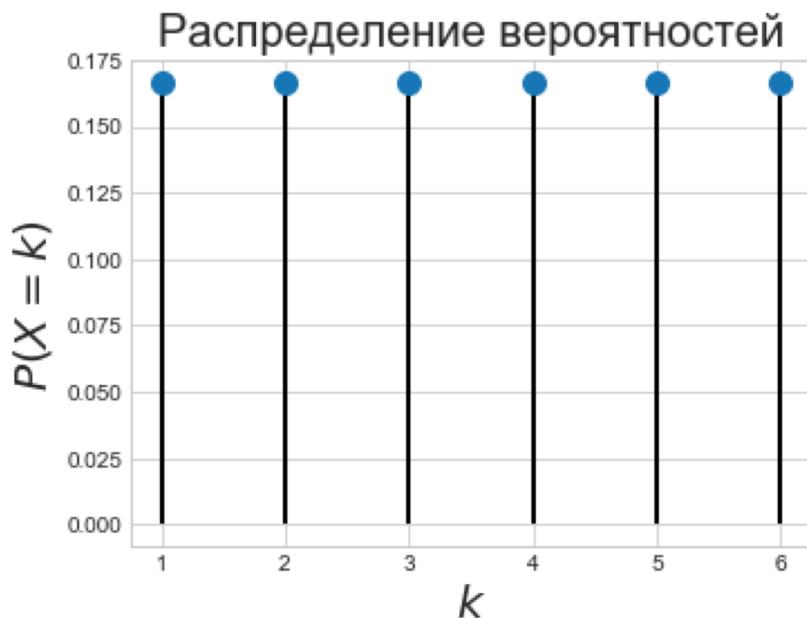
Функция распределения



Произвольное дискретное распределение

- Подбрасывание игральной кости

X	1	2	3	4	5	6
$\mathbb{P}(X = k)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$



Геометрическое распределение

- Номер броска, когда произошло первое попадание в корзину

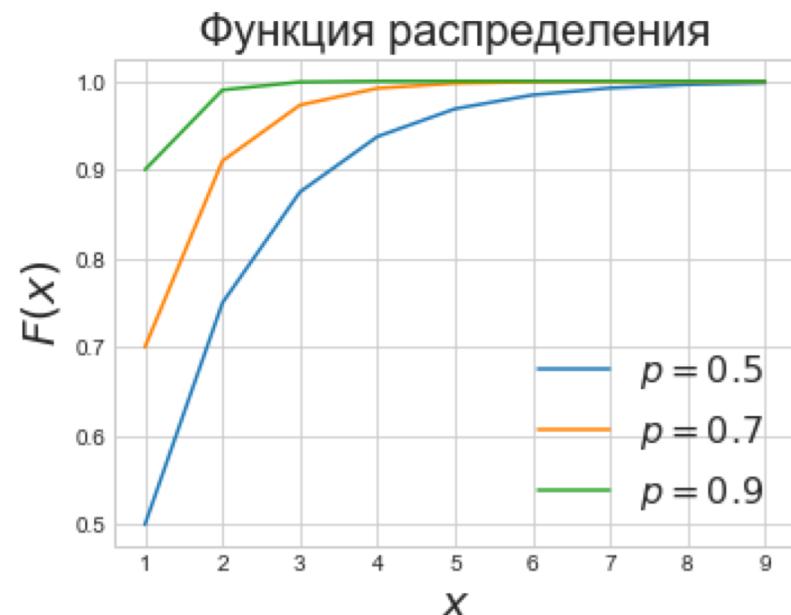
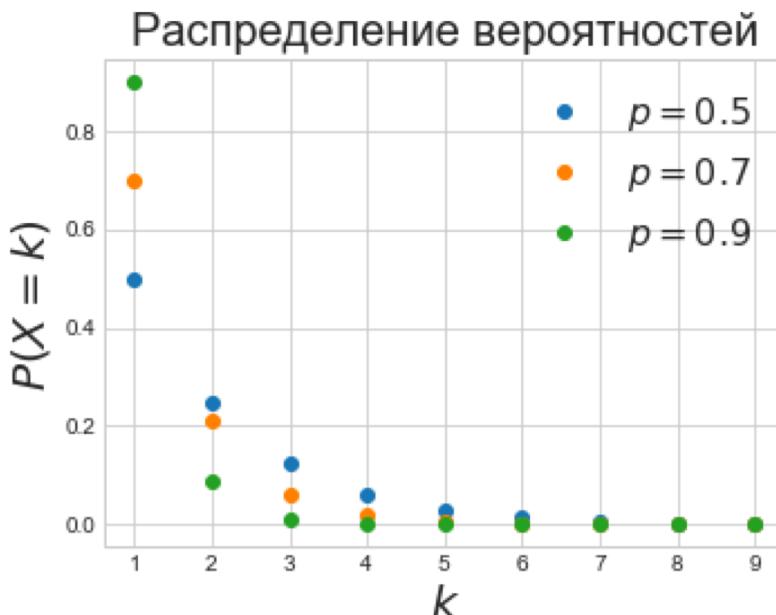
$$\mathbb{E}(X) = \frac{1}{p}$$

Геометрическая случайная величина: $X \sim Geom(p)$

$$Var(X) = \frac{1-p}{p^2}$$

p – вероятность успеха

k принимает значения $1, 2, 3, \dots$ $\mathbb{P}(X = k) = p \cdot (1 - p)^{k-1}$



Счётчики

- Число людей в очереди
- Число лайков под фото
- Число автобусов, проехавших за час мимо остановки

Пуассоновская случайная величина: $X \sim Poiss(\lambda)$

Распределение Пуассона хорошо описывает счётчики



♥ Нравится 15

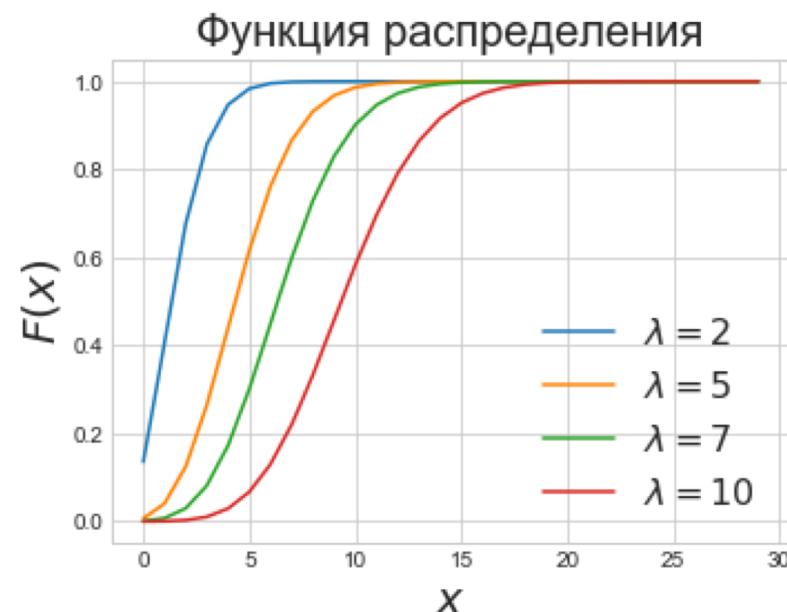
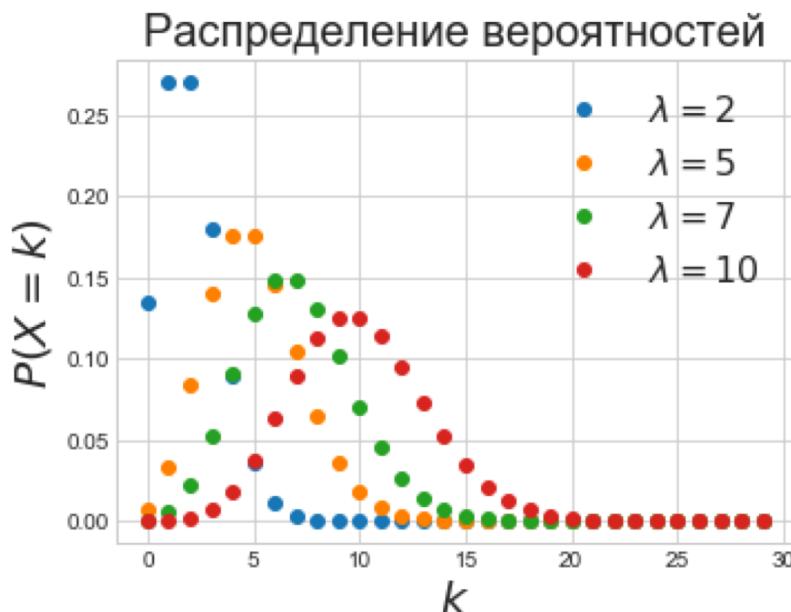
Распределение Пуассона

$$\mathbb{P}(X = k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!} \quad X \sim Poiss(\lambda)$$

- Параметр λ интерпретируется как интенсивность потока событий
- k принимает значения $0, 1, 2, \dots$

$$Var(X) = \lambda$$

$$\mathbb{E}(X) = \lambda$$



Время до ...

- Время ожидания трамвая
- Время до прихода нового человека в очередь
- Время до поломки механизма

Экспоненциальная
случайная величина:
 $X \sim Exp(\lambda)$

- Интервалы времени между событиями
- Модели времени жизни



Равномерное распределение

- Время рождения ребёнка

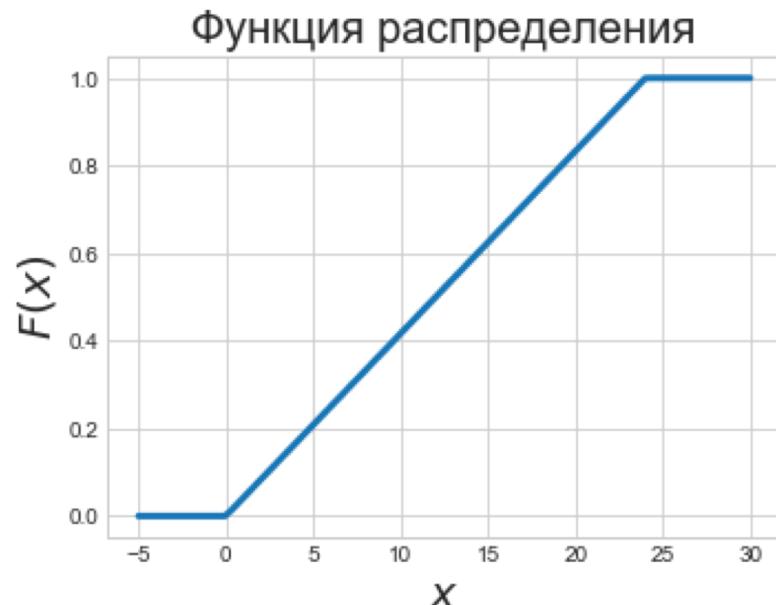
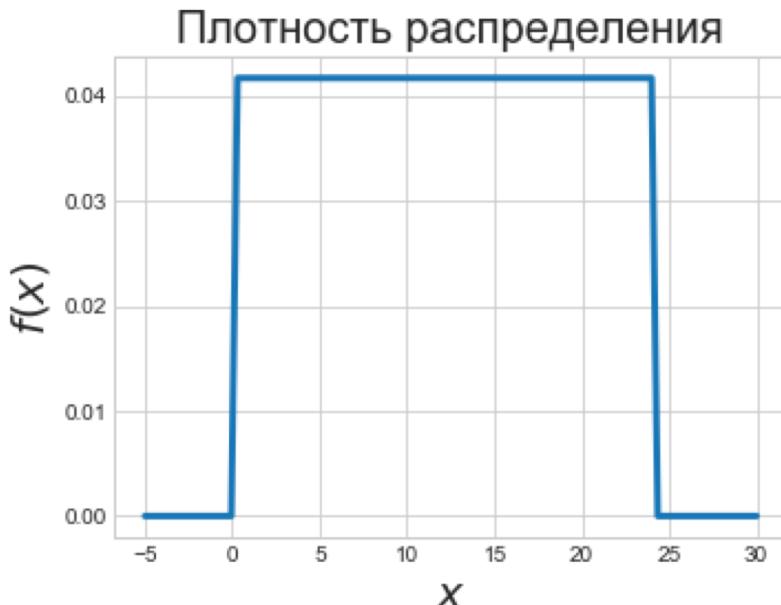
Равномерная случайная величина: $X \sim U[a; b]$

$$f_X(x) = \frac{1}{b-a}, x \in [a; b]$$

$$\mathbb{E}(X) = \frac{a+b}{2}$$

$$\text{Var}(X) = \frac{(b-a)^2}{12}$$

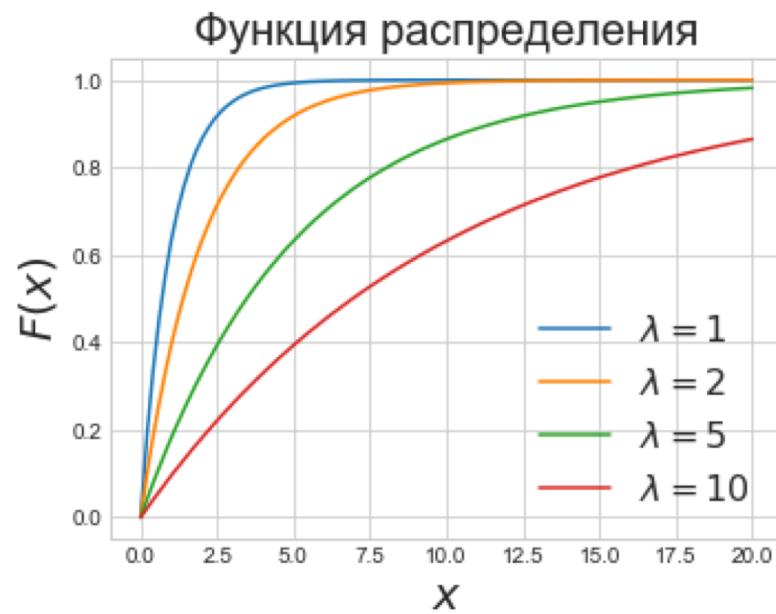
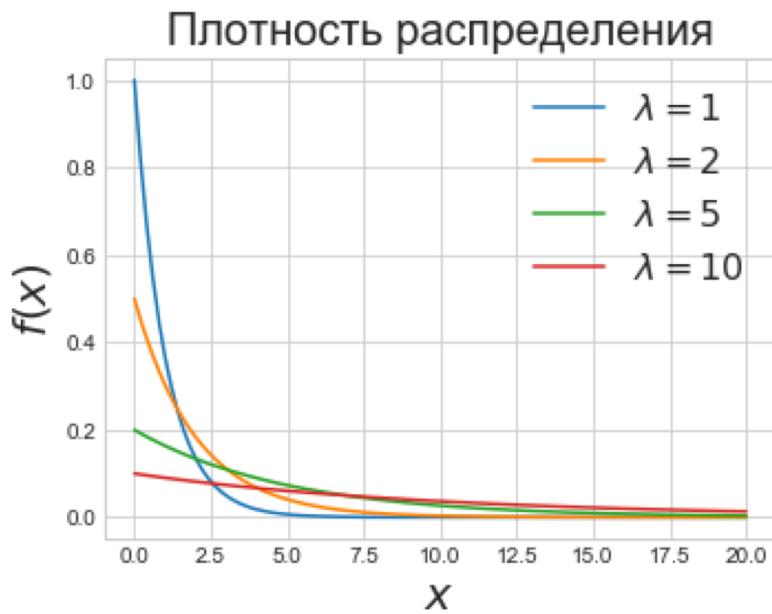
$$F_X(x) = \frac{x-a}{b-a}, x \in [a; b]$$



Экспоненциальное распределение

$$f_X(x) = \lambda \cdot e^{-\lambda \cdot x}, x \geq 0$$

$$F_X(x) = 1 - e^{-\lambda \cdot x}, x \geq 0$$



У экспоненциального распределения нет памяти. Автобусы приходят на остановку случайно. Время, которое осталось ждать не зависит от того, сколько уже прошло времени.

$$\mathbb{E}(X) = \frac{1}{\lambda}$$
$$\text{Var}(X) = \frac{1}{\lambda^2}$$

Нормальное распределение

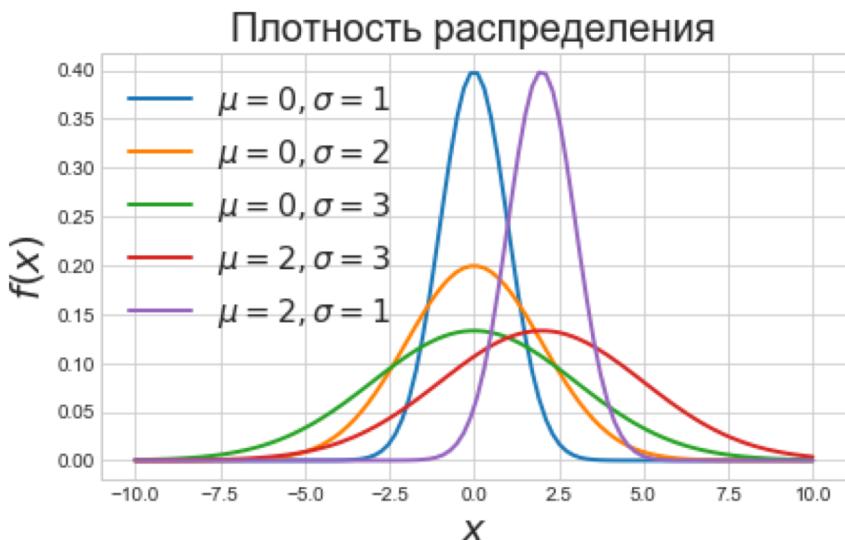
- Погрешность весов

Нормальная случайная величина:

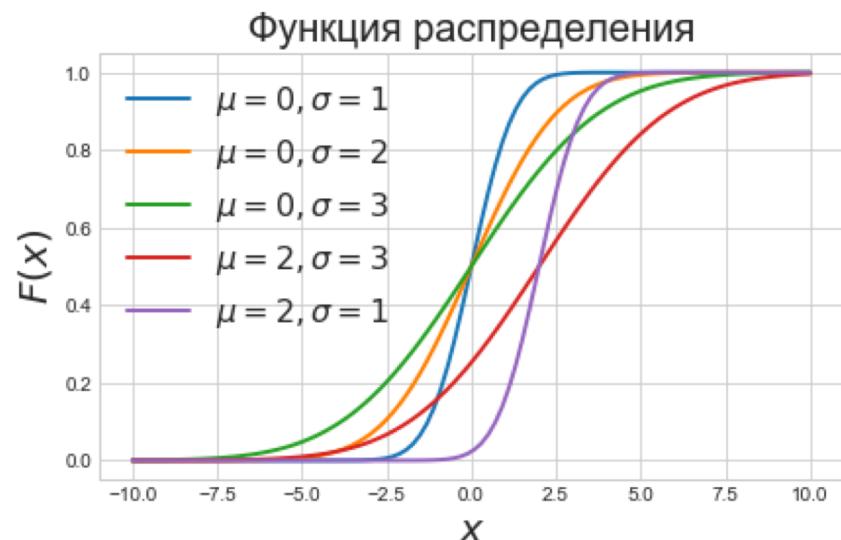
$$X \sim N(\mu, \sigma^2)$$

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

Функцию распределения нельзя найти в аналитическом виде, интеграл не берётся



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



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

Распределения бывают разными

Случайная величина	Распределение
Пол ребенка	$Bern(p)$
Попадания в корзину	$Binom(n, p)$
Число бросков до первого попадания	$Geom(p)$
Число людей в очереди	$Poiss(\lambda)$
Подбрасывание кости	Дискретное
Время между событиями	$Exp(\lambda)$
Время до поломки часов	$Exp(\lambda)$
Время рождения ребенка	$U[0; 24]$
Погрешность весов	$N(0, \sigma^2)$

Резюме

- Моделировать внутренности сундука можно с помощью различных законов распределения
 - Наиболее подходящий закон выбирается с помощью здравого смысла
- !** Мы перечислили лишь один из вариантов моделировать незнание с помощью случайной величины. Эти распределения не истина в последней инстанции
- Все предпосылки, связанные с выбранным законом, должны проверяться по данным, в будущем мы научимся это делать

Описательные статистики

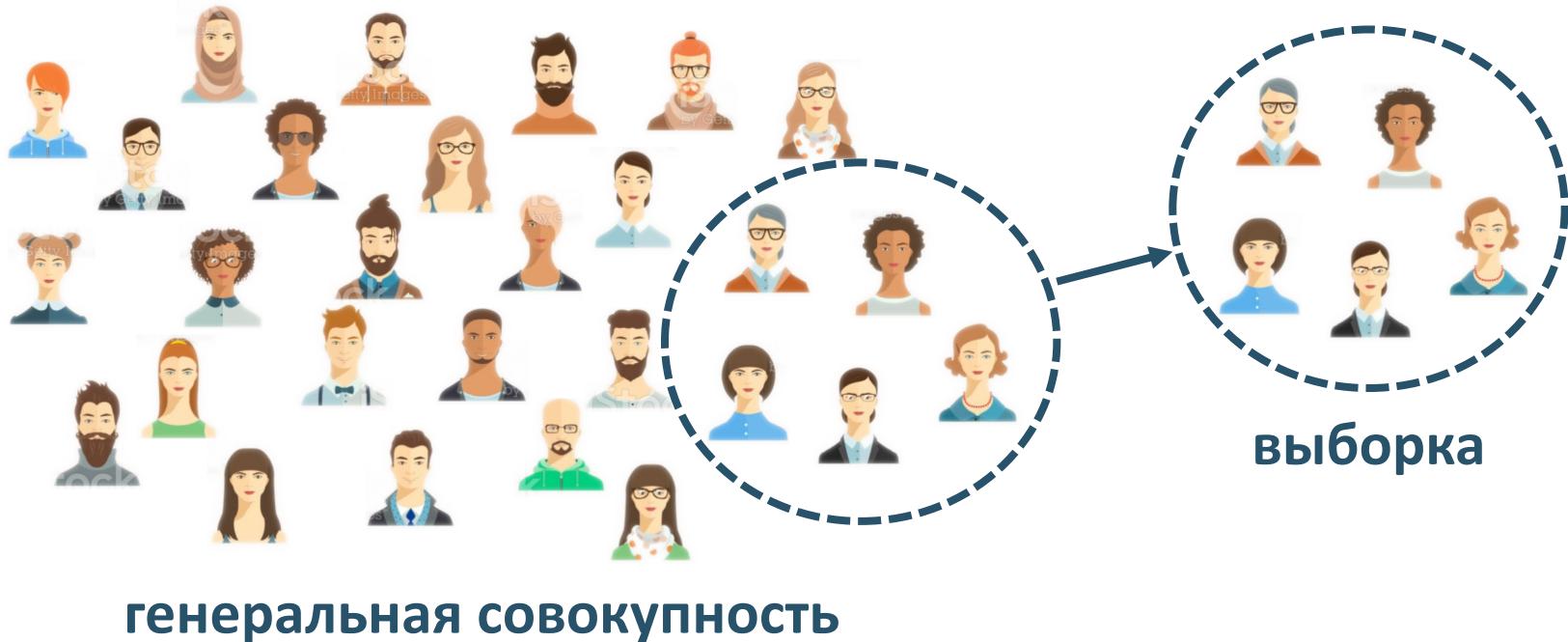
Генеральная совокупность и выборка

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 и 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 долго



Репрезентативность

- Выборки позволяют сделать выводы о всей генеральной совокупности
- Чтобы выводы были корректными, выборка должны быть **репрезентативной**
- **Репрезентативная выборка** – отражает свойства генеральной совокупности

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

Репрезентативность

- Выборки позволяют сделать выводы о всей генеральной совокупности
- Чтобы выводы были корректными, выборка должны быть **репрезентативной**
- **Репрезентативная выборка** – отражает свойства генеральной совокупности

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

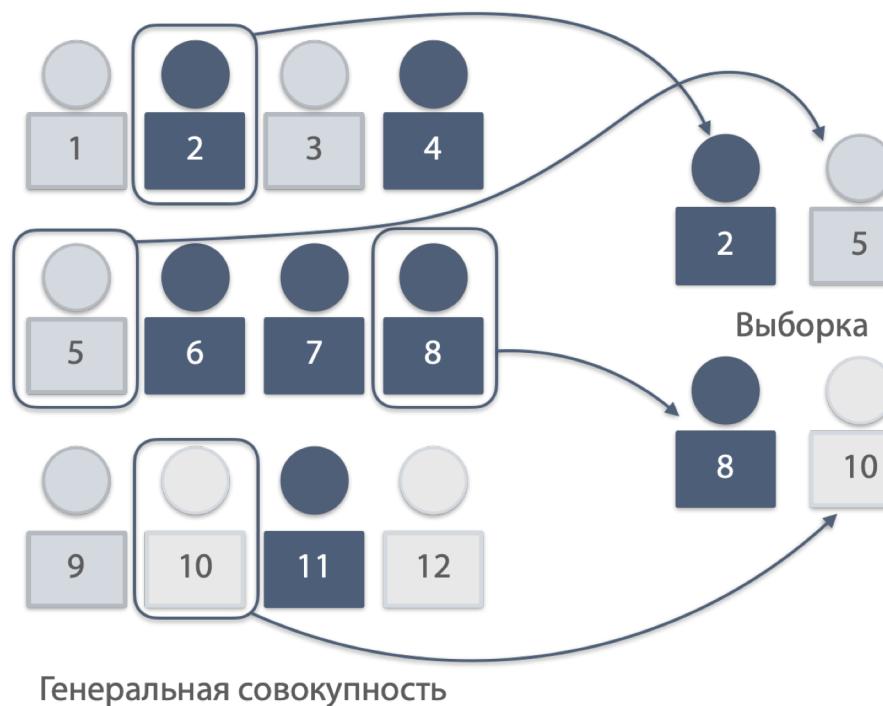
нет • Добрыня опросил свою баскетбольную команду

да • Илья опросил людей на остановке

нет • Алёна опросила всех своих подруг

Выборку бы, да репрезентативную бы

- Репрезентативность выборки определяет, насколько корректно делать выводы о всей генеральной совокупности, опираясь только на неё
- Один из способов достижения репрезентативности: случайный отбор наблюдений



Предпосылки

Выборка: X_1, X_2, \dots, X_n

Размер выборки

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)

Статистика

Выборка: $X_1, X_2, \dots, X_n \sim iid$

Статистика – любая функция от наблюдений

Примеры: среднее, медиана, максимум и т.п.



Статистика

Каждая статистика – случайная величина, так как она вычисляется на основе случайной выборки, т.е. на основе других случайных величин



Статистика

Каждая статистика – случайная величина, так как она вычисляется на основе случайной выборки, т.е. на основе других случайных величин

- ! Мы будем рассматривать любые статистики, посчитанные на основе выборки как случайные величины и изучать их свойства

Выборка из сундука

	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

Какими бывают переменные

Переменные



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

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

Какими бывают описательные статистики

Описательные статистики



Меры центральной тенденции

Отвечают на вопрос
“а на что похожи типичные
наблюдения из выборки”

Примеры: среднее, мода,
медиана



Меры разброса

Отвечают на вопрос
“а как сильно значения
в выборке могут отличаться от
типовых значений”

Примеры: дисперсия,
стандартное отклонение,
интерквантильный размах

Среднее

Выборочный аналог математического ожидания, рассчитывается по формуле:

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

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

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

Медиана

Чтобы найти медиану, данные нужно расположить в порядке возрастания. Медианой будет значение, которое оказалось в середине.

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

$$-4, 0, \textcolor{pink}{1}, 3, 5 \Rightarrow med = 1$$

Если число значений чётное, берётся среднее двух значений, которые «окружают» середину

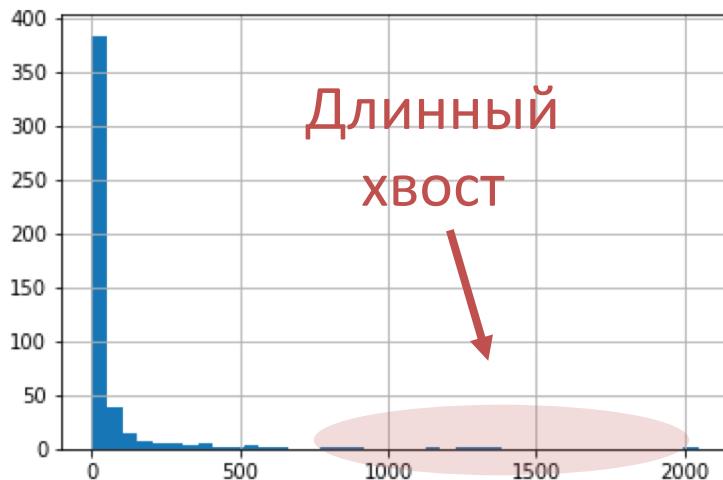
Пример 2: $x_1 = 1, x_2 = 5, x_3 = -4, x_4 = 3$

$$-4, \textcolor{pink}{1}, 3, 5 \Rightarrow med = \frac{1+3}{2} = 2$$

Среднее и медиана

Среднее чувствительно к выбросам в данных, медиана не чувствительна

- Среднее и медиана отражают типичное значение
- Если в выборке нет выбросов, они примерно совпадают



Высокая зарплата сильно исказит среднее, но не медиану

Выборочная дисперсия

- Хочется понимать, насколько сильно элементы выборки отклоняются от своего типичного значения

Алёна: 18 лет,

Карина: 22 года

Отклонения от среднего:

$$\bar{x} = \frac{18 + 22}{2} = 20$$

$$x_1 - \bar{x} = 18 - 20 = -2$$

$$x_2 - \bar{x} = 22 - 20 = 2$$

Среднее отклонение:

$$\frac{-2 + 2}{2} = 0$$

~~⇒ в выборке нет
неопределённости~~



Как быть?

Выборочная дисперсия

- Хочется понимать, насколько сильно элементы выборки отклоняются от своего типичного значения

Алёна: 18 лет,

Карина: 22 года

$$\bar{x} = \frac{18 + 22}{2} = 20$$

Отклонения от среднего:

$$x_1 - \bar{x} = 18 - 20 = -2$$
$$x_2 - \bar{x} = 22 - 20 = 2$$

Среднее отклонение:

$$\frac{-2 + 2}{2} = 0$$

Выход №1:

$$\frac{| -2 | + | 2 |}{2} = \frac{4}{2} = 2$$

Выборочная дисперсия

- ! Проблема меры, основанной на модуле в том, что она недифференцируема

Модуль неудобно использовать
при теоретических выкладках

Выборочная дисперсия

- Хочется понимать, насколько сильно элементы выборки отклоняются от своего типичного значения

Алёна: 18 лет,

Карина: 22 года

$$\bar{x} = \frac{18 + 22}{2} = 20$$

Отклонения от среднего:

$$x_1 - \bar{x} = 18 - 20 = -2$$
$$x_2 - \bar{x} = 22 - 20 = 2$$

Среднее отклонение:

$$\frac{-2 + 2}{2} = 0$$

Выход №2:

$$\frac{(-2)^2 + 2^2}{2} = \frac{4 + 4}{2} = 4$$

Выборочная дисперсия

- ! Для квадратичной функции всегда есть производная
 - Она обладает хорошими статистическими свойствами
 - Её удобно использовать для теоретических выкладок

Выборочная дисперсия

- Это мера разброса. Показывает, насколько сильно элементы выборки отклоняются от своего типичного значения

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

Пример: $x_1 = 1, x_2 = 5, x_3 = -4, x_4 = 6 \quad \bar{x} = 2$

$$\hat{\sigma}^2 = \frac{(1 - 2)^2 + (5 - 2)^2 + (-4 - 2)^2 + (6 - 2)^2}{4}$$

$$\hat{\sigma}^2 = \frac{1 + 9 + 36 + 16}{4} = 15.5$$

Выборочная дисперсия

Удобнее искать дисперсию по более простой формуле:

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (X_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n [X_i^2 - 2 \cdot X_i \cdot \bar{x} + \bar{x}^2] = \\ &= \frac{1}{n} \sum_{i=1}^n X_i^2 - \frac{2 \bar{x}}{n} \sum_{i=1}^n X_i + \frac{1}{n} \sum_{i=1}^n \bar{x}^2 \\ &= \overline{x^2} - 2 \bar{x}^2 + \bar{x}^2 = \overline{x^2} - \bar{x}^2\end{aligned}$$

Выборочная дисперсия

Удобнее искать дисперсию по более простой формуле:

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

Пример: $x_1 = 1, x_2 = 5, x_3 = -4, x_4 = 6 \quad \bar{x} = 2$

$$\bar{x^2} = \frac{1^2 + 5^2 + (-4)^2 + 6^2}{4} = 19.5$$

$$\hat{\sigma}^2 = 19.5 - 4 = 15.5$$

Стандартное отклонение

- Дисперсия измеряется в квадратных величинах
- Чтобы вернуться назад к исходным величинам, можно взять из неё квадратных корень

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

лет = $\sqrt{\text{лет в квадрате}}$

Несмешённая выборочная дисперсия

- Обычно на практике используют другую формулу:

$$s^2 = \frac{(X_1 - \bar{x})^2 + \dots + (X_n - \bar{x})^2}{n - 1} = \frac{1}{n - 1} \sum_{i=1}^n (X_i - \bar{x})^2$$

- Это несмешённая дисперсия
- Что это значит, мы подробно обсудим в будущих неделях нашей специализации

Перцентиль

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

- Перцентиль это выборочный аналог квантиля
- Проще всего вычислять его по упорядоченной выборке

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

- Квартили – перцентили с шагом в 0.25:

$$x_{(0.25 \cdot [n+1])} \quad x_{(0.5 \cdot [n+1])} \quad x_{(0.75 \cdot [n+1])}$$

медиана

Интерквантильный размах:

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

Перцентиль

Пример:

1, 5, 3, -1, -4, 3, 3, -10, 2, -1

-10, -4, -1, -1, 1, 2, 3, 3, 3, 5

упорядочили
выборку

$$\Rightarrow med = \frac{1+2}{2} = 1.5$$

позиция медианы:

$$0.5 \cdot (10 + 1) = 5.5$$

Перцентиль

Пример:

1, 5, 3, -1, -4, 3, 3, -10, 2, -1

-10, -4, -1, -1, 1, 2, 3, 3, 3, 5

упорядочили
выборку

$$\Rightarrow \frac{3 + 3}{2} = 3$$

позиция верхней квартили: $0.75 \cdot (10 + 1) = 8.25$

- Перцентили в спорных случаях можно считать по-разному, каждая библиотека предоставляет разные варианты

Резюме

- Репрезентативная выборка – отражает свойства генеральной совокупность
- Мы будем в дальнейшем предполагать, что все наблюдения, которые мы делаем, не зависят друг от друга

Теоретическая величина	Выборочный аналог
Математическое ожидание	Выборочное среднее
Дисперсия	Выборочная дисперсия
Квантиль	Перцентиль
Медиана	Выборочная медиана
Мода	Выборочная мода

Гистограмма и эмпирическая функция распределения

Эмпирическая функция распределения

Функция распределения – функция, которая определяет вероятность события $X \leq x$, то есть

$$F(x) = \mathbb{P}(X \leq x)$$

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

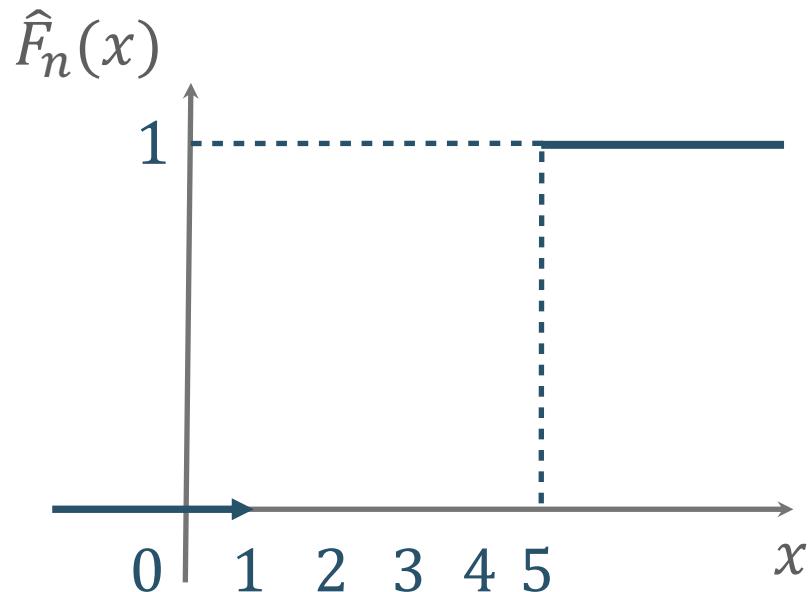
$$\hat{F}_n(x) = \widehat{\mathbb{P}}(X \leq x) = \frac{1}{n} \sum_{i=1}^n [X_i \leq x],$$

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

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

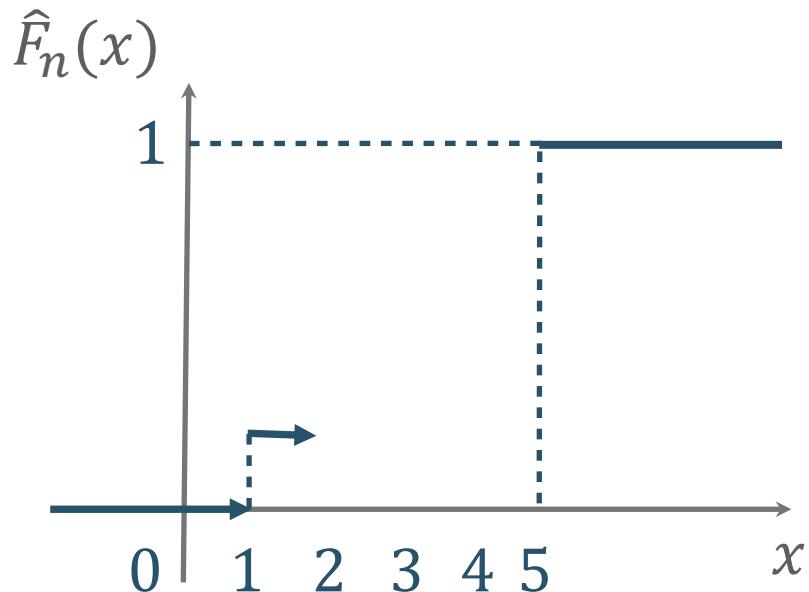
Эмпирическая функция распределения

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



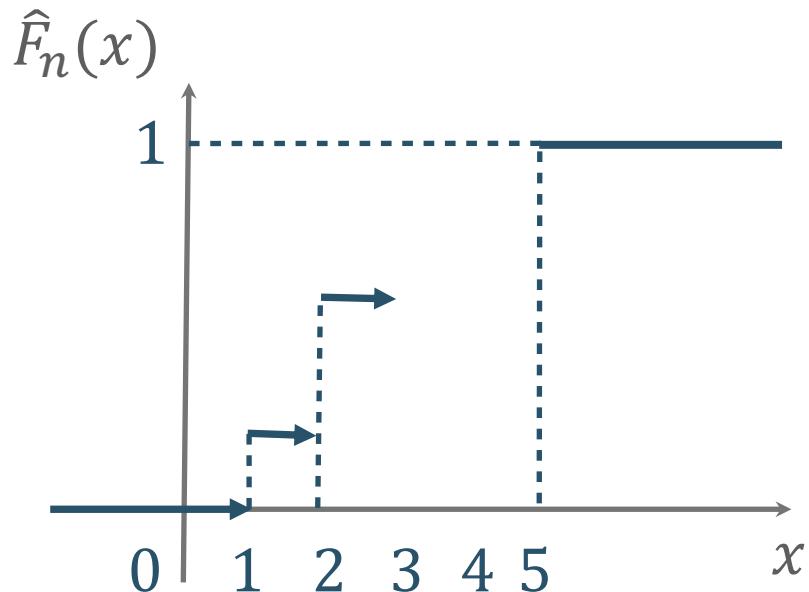
Эмпирическая функция распределения

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



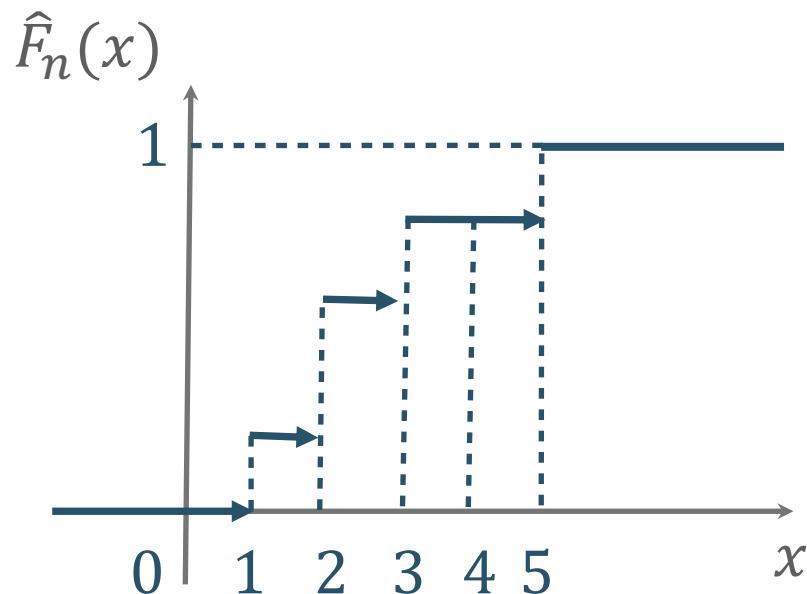
Эмпирическая функция распределения

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



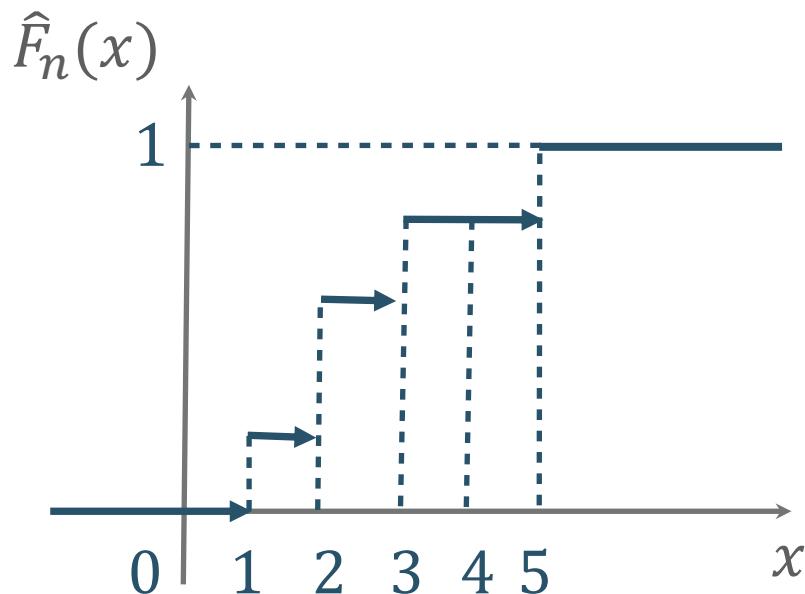
Эмпирическая функция распределения

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



Эмпирическая функция распределения

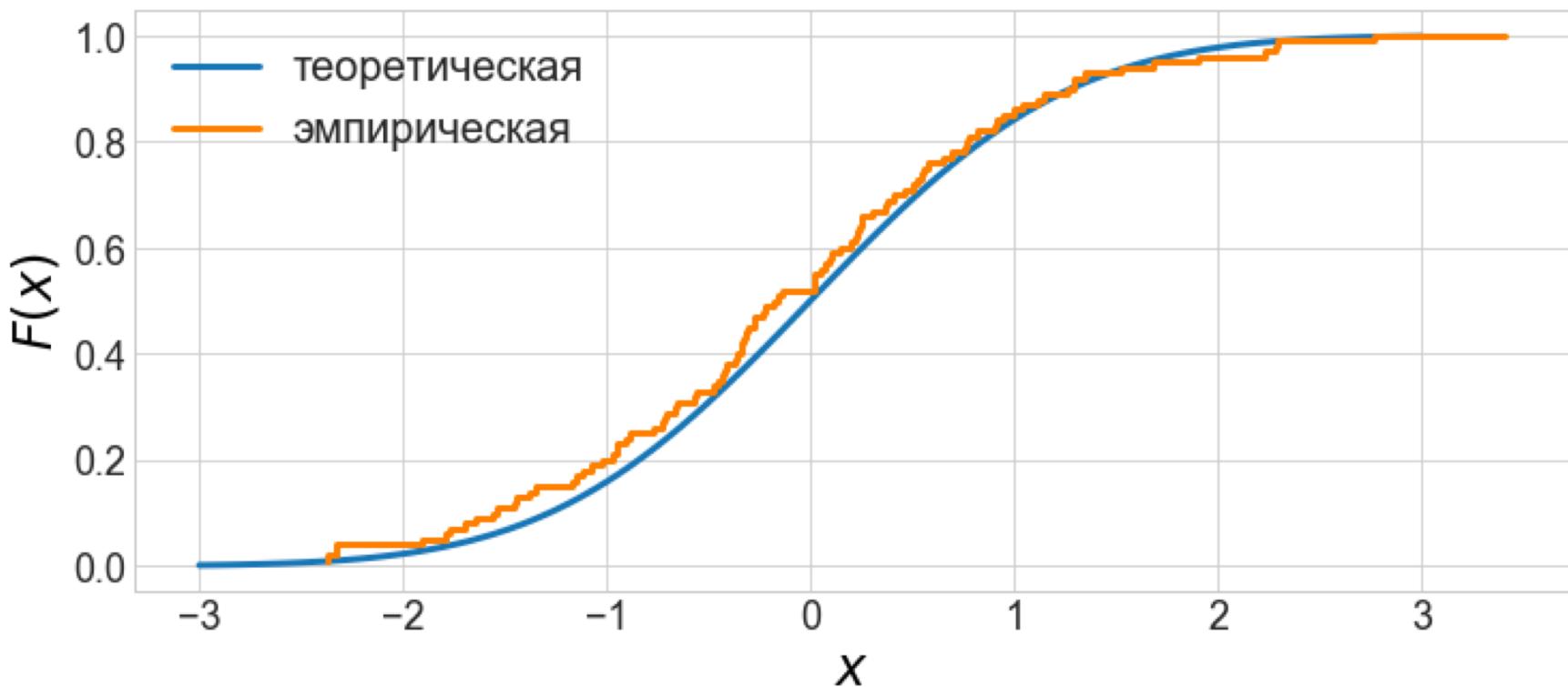
Пример: $x_1 = 2, x_2 = 5, x_3 = 2, x_4 = 3, x_5 = 1$



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.

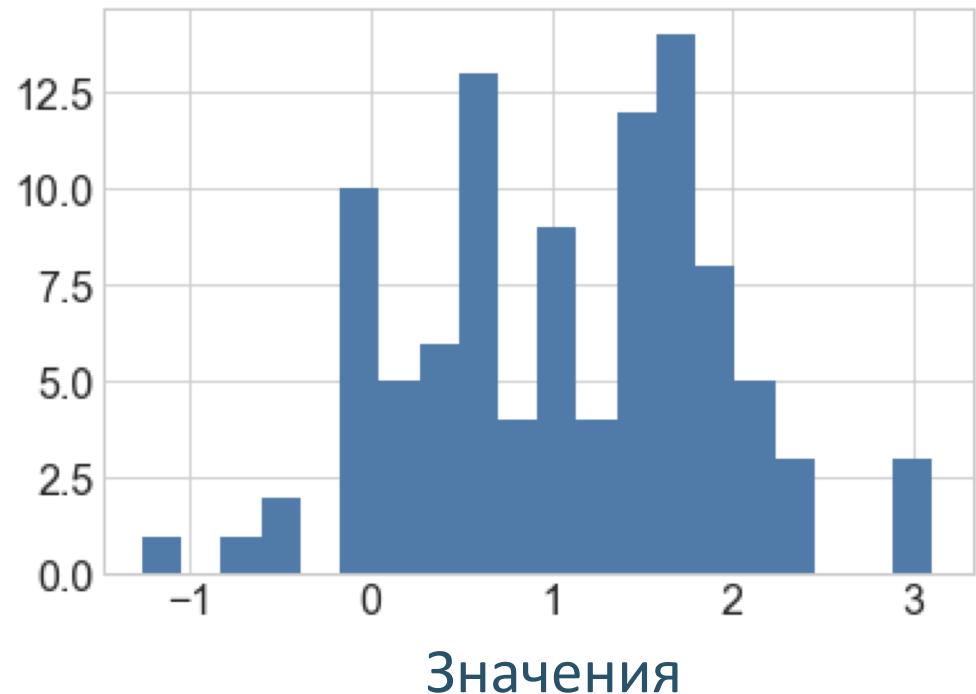


Гистограмма

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

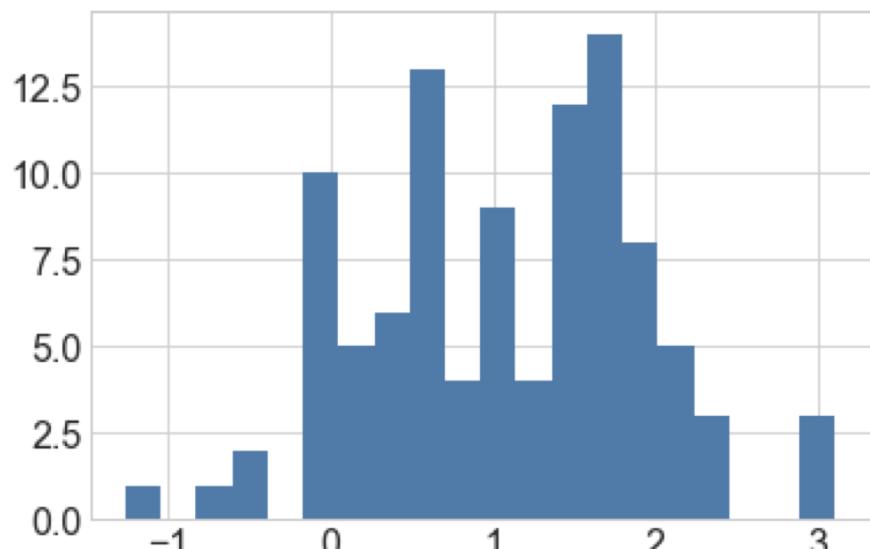
Область возможных значений обычно дробят на отрезки, **бины**.
Чем короче бины, тем детальнее рисуется гистограмма.

Сколько значений
попали в текущий
отрезок (бин)

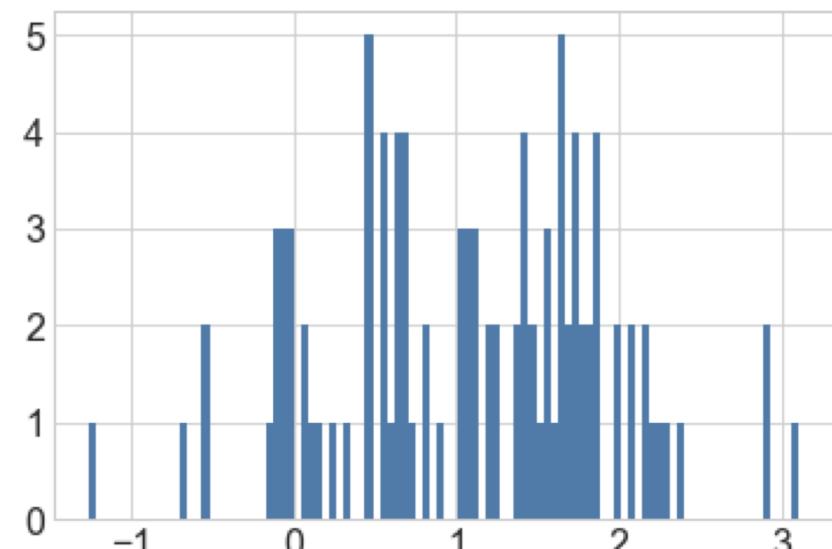


Гистограмма

- Чем короче бины, тем чувствительнее гистограмма к шуму
- Выборка объёма 100 из нормального распределения



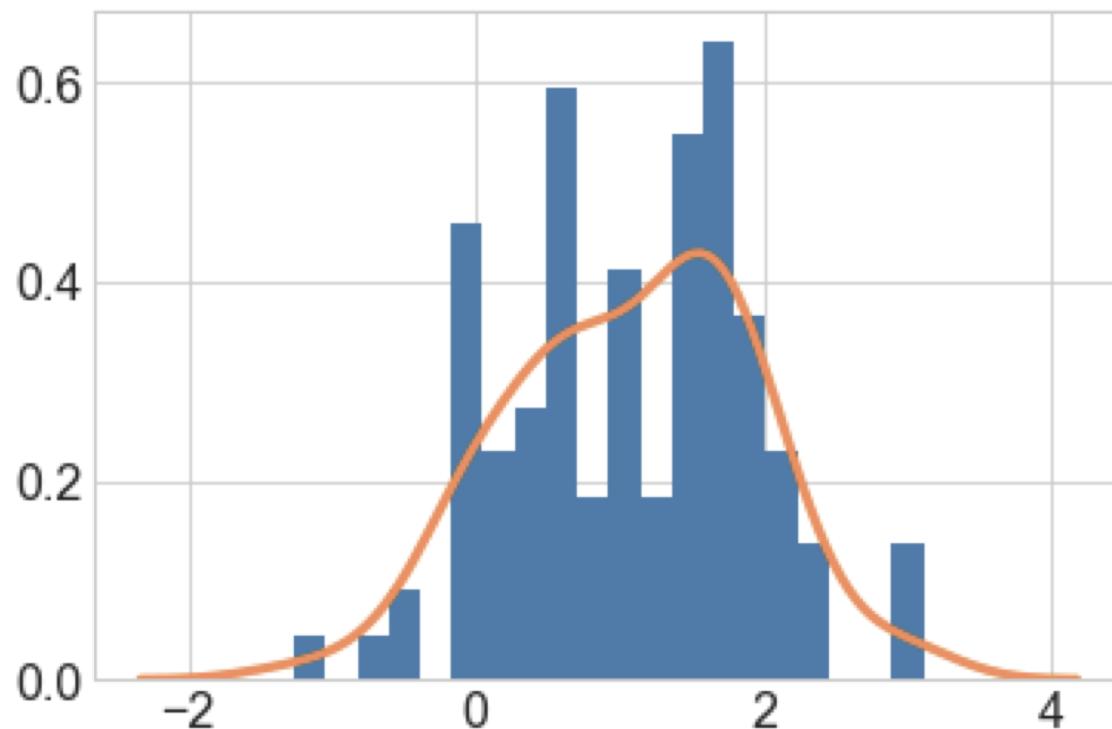
20 бинов



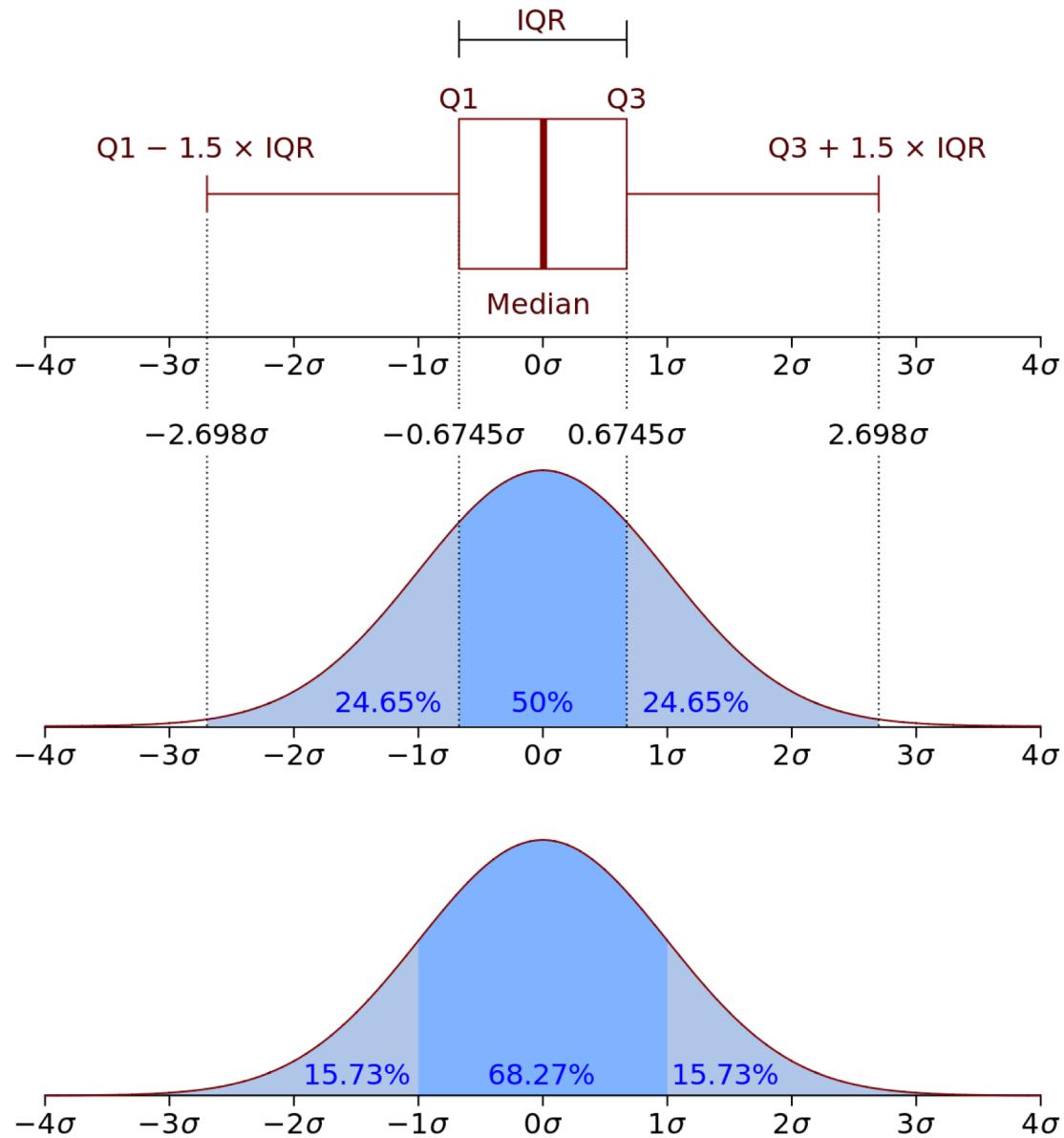
100 бинов

Гистограмма

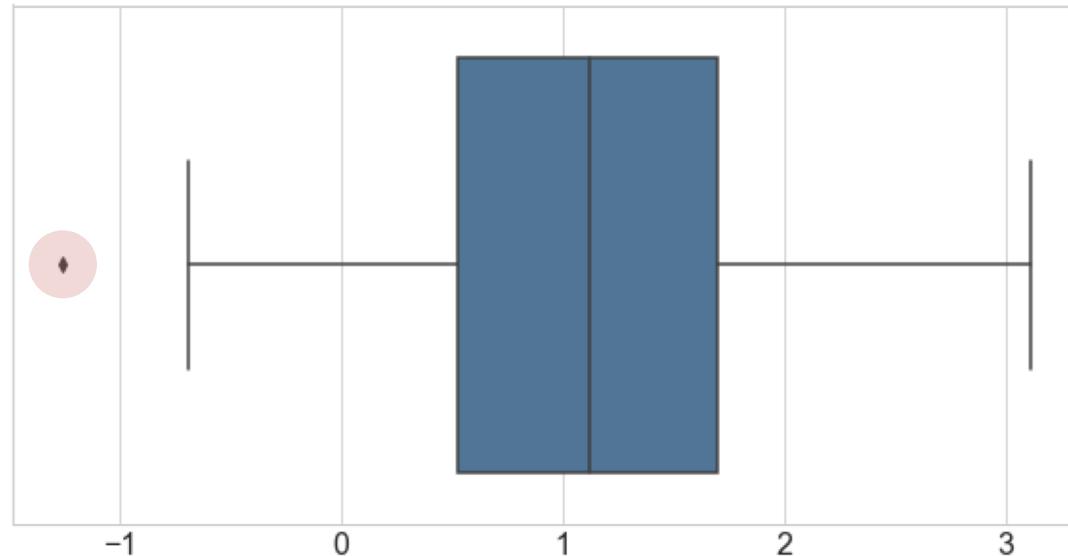
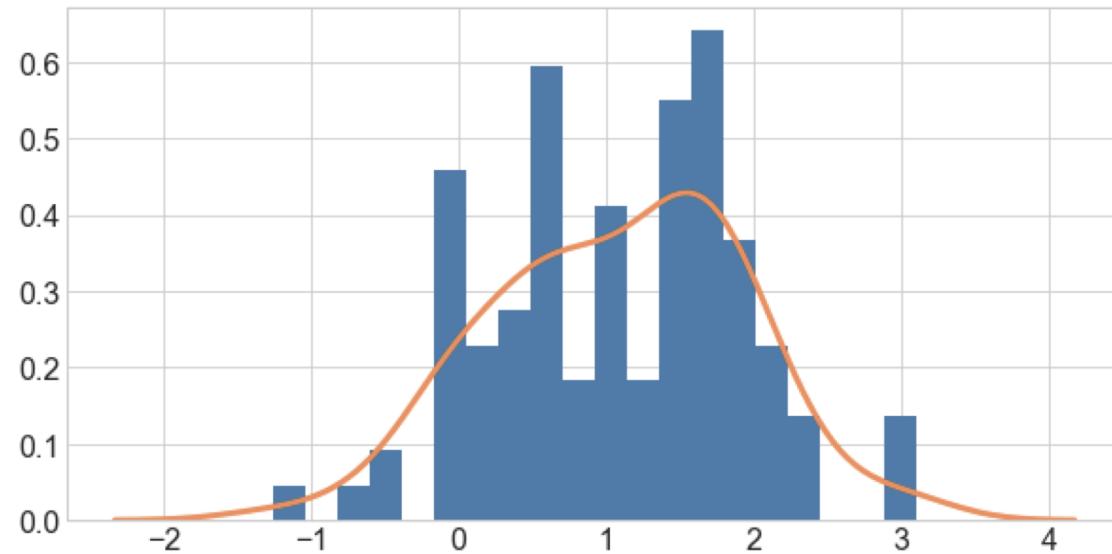
- По гистограмме можно попытаться оценить плотность распределения случайной величины
- Позже мы подробнее поговорим про методы, которые позволяют это сделать, ядерное сглаживание (kernel density estimation)



Ящик с усами (Boxplot)



Ящик с усами



Аномальное
значение

Ящик с усами (Boxplot)

- Позволяет обобщить данные
- Показывает наличие выбросов
- Даёт некоторое представление о симметрии данных
- Позволяет сравнить несколько переменных между собой

Резюме

- Плотность распределения вероятностей и функцию распределения также можно оценить по выборке
- Ящик с усами позволяет визуализировать основные описательные статистики

Теоретическая величина	Выборочный аналог
Функция распределения	Эмпирическая функция распределения
Плотность распределения	Гистограмма