

---

# Случайные величины





# Олег Булыгин

IT-аудитор, ментор и наставник

---

## Аккаунты в соц.сетях

 [fb.com/obulygin91](https://fb.com/obulygin91)

 [vk.com/obulygin91](https://vk.com/obulygin91)

 [obulygin91@ya.ru](mailto:obulygin91@ya.ru)

 [linkedin.com/in/obulygin](https://linkedin.com/in/obulygin)

 [@obulygin91](https://t.me/obulygin91)



---

# Содержание

1

Мотивация

2

Случайные события и случайные величины

3

Вероятность и частота

4

Среднее значение и риск

5

Распределения

6

Условная вероятность

7

Полная вероятность и теорема Байеса



# Мотивация

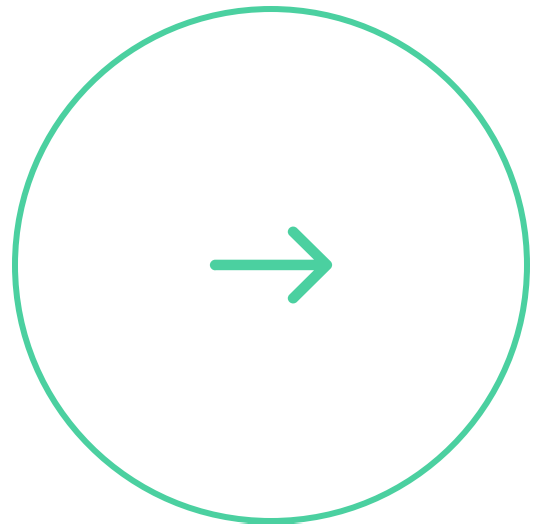
## Случайные величины полезно понимать:

- Инженерам – чтобы обрабатывать неточные показания приборов
- Врачам – чтобы верно интерпретировать результаты анализов
- Маркетологам – чтобы строить надежные прогнозы спроса
- Чиновникам – чтобы оценивать требуемые резервы бюджета

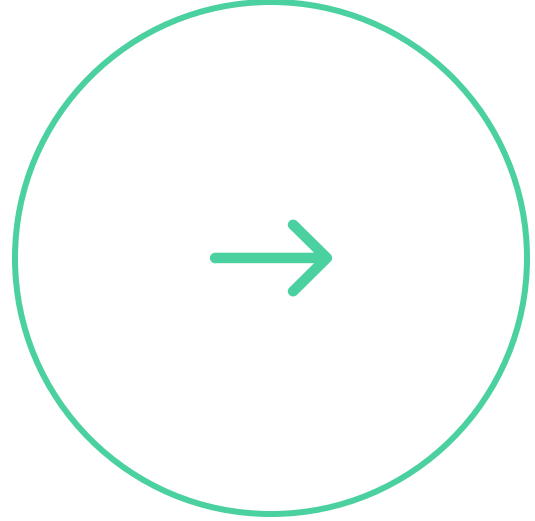


# Мотивация

## Примеры задач, которые мы научимся решать:



**Задача «для врача»:** тестирование на COVID-19. Ни один тест не является абсолютно достоверным. Поэтому можно поставить вопрос: какова вероятность того, что человек заражен, если тест дал положительный результат?



**Задача «для маркетолога»:** прогнозирование эффективности рекламы в соцсетях. Есть несколько возрастных категорий покупателей, причем вероятность интереса к продукту различна у разных категорий. Есть сообщество в соцсети, где предполагается разместить рекламу. Известен возраст участников сообщества. Какова вероятность того, что случайный участник заинтересуется продуктом?



# Случайные события

**Случайное событие** – событие, о наступлении которого мы не можем определенно сказать заранее

**Н.В.** Почти все случайные события в мире, кроме событий внутри атомов, выглядят случайными лишь из-за ограниченности нашего знания

## Примеры

- Выпадение трёх очков на игральной кости
- Заражение коронавирусом
- Падение метеорита
- Сдача экзамена на «отлично»



# Случайные величины

**Случайная величина** – величина, о значении которой мы не можем определённо сказать заранее

То, что случайная величина приняла какое-то значение, является случайным событием

## Примеры дискретных величин

- Количество очков, выпавших на игральной кости
- Оценка на экзамене

## Примеры непрерывных величин

- Курс доллара
- Время в дороге до работы
- Продолжительность жизни





# Случайные величины

Генерация случайных величин возможна с помощью функций пакета `numpy.random`:

- **`numpy.random.randint`** (low, high, size) – дискретная целая величина из диапазона [low, high). Если задан size, возвращает массив величин размера size
- **`numpy.random.random`** (size) – непрерывная величина из диапазона [0, 1) или массив таких величин
- **`numpy.random.uniform`** (low, high, size) – непрерывная величина из диапазона [low, high) или массив таких величин

**N.B.** Все функции на самом деле возвращают псевдослучайные величины. Они рассчитываются по алгоритму, но для тех, кто не видит алгоритм, его результат неотличим от случайного





# Вероятность

Если в опыте одинаково возможны  $N$  исходов, из которых  $M$  исходов приводят к событию  $A$ , то вероятность события  $A$  – это

$$P(A) = \frac{M}{N}$$

**Пример.** Какова вероятность выпадения нечетного числа очков на игральной кости? Возможны  $N = 6$  исходов, из них  $M = 3$  исхода (выпадение 1, 3, 5 очков) приводят к требуемому событию  $A$ . Поэтому  $P(A) = 3/6 = 0,5$



# Частота

Не всегда можно перечислить все исходы и воспользоваться определением вероятности напрямую. Нам нужен универсальный практический способ вычисления вероятности.

Можно многократно повторить (или смоделировать на компьютере) опыт и вычислить **частоту** события  $A$  – долю «успешных» опытов, в которых происходит событие. Это **метод Монте-Карло**

При большом количестве повторений опыта **частота близка к вероятности:**

$$P(A) \approx \frac{\text{Количество успехов}}{\text{Количество опытов}}$$

**Jupyter Notebook:** задача 1



# Среднее значение

Если дискретная случайная величина  $X$  принимает значения  $X_1, X_2 \dots$  с вероятностями  $P_1, P_2 \dots$ , то ее **среднее значение** равно

$$\bar{X} = P_1 X_1 + P_2 X_2 + \dots$$

**Пример.** Эта формула объясняет, например, необходимость содержания противопожарных служб, несмотря на крайне низкую вероятность пожаров в современных городах. Хотя вероятность очень мала, но очень велик потенциальный ущерб, и их произведение не будет малым. Пусть вероятность пожара  $P_1 = 0,001$ , а ущерб при пожаре  $X_1 = 1$  млрд руб. Тогда средний ущерб  
Именно произведение вероятности на ущерб считают определением **риска**

$$\bar{X} = P_1 X_1 + (1 - P_1) \cdot 0 = 1 \text{ млн руб.}$$

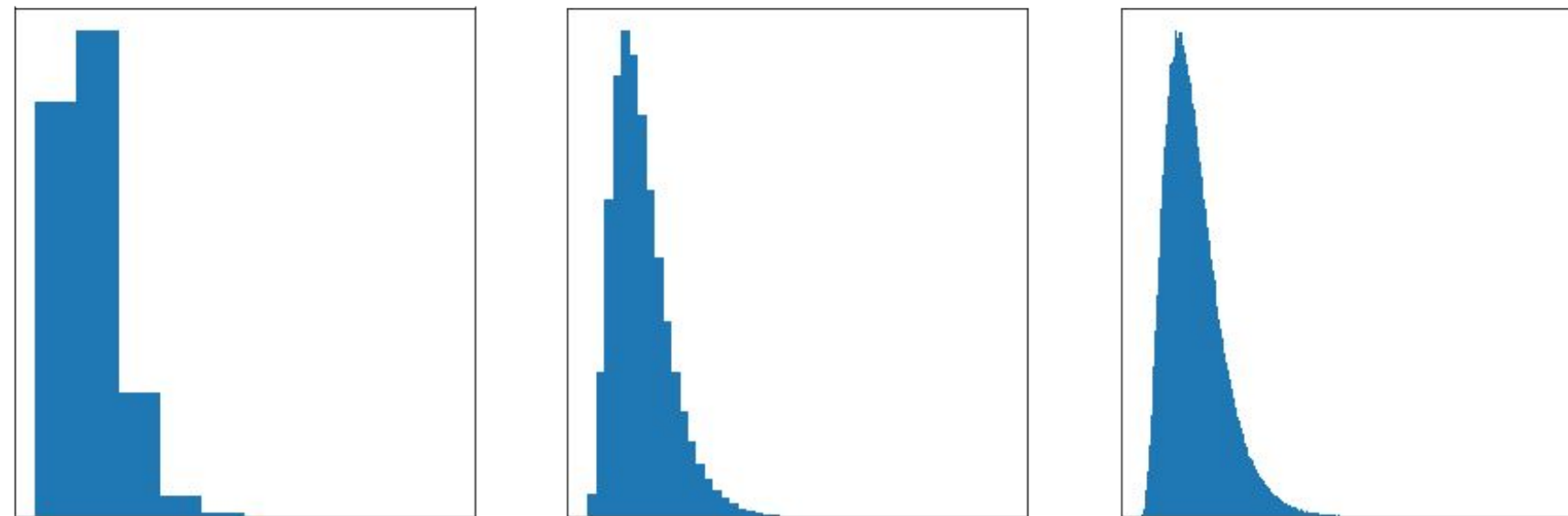
Именно произведение вероятности на ущерб считают определением **риска**



# Распределения

Если спрашивать у случайных прохожих их рост или возраст, изучать суммы ресторанных чеков, то окажется, что очень большие и очень малые значения встречаются редко, а близкие к средним – часто

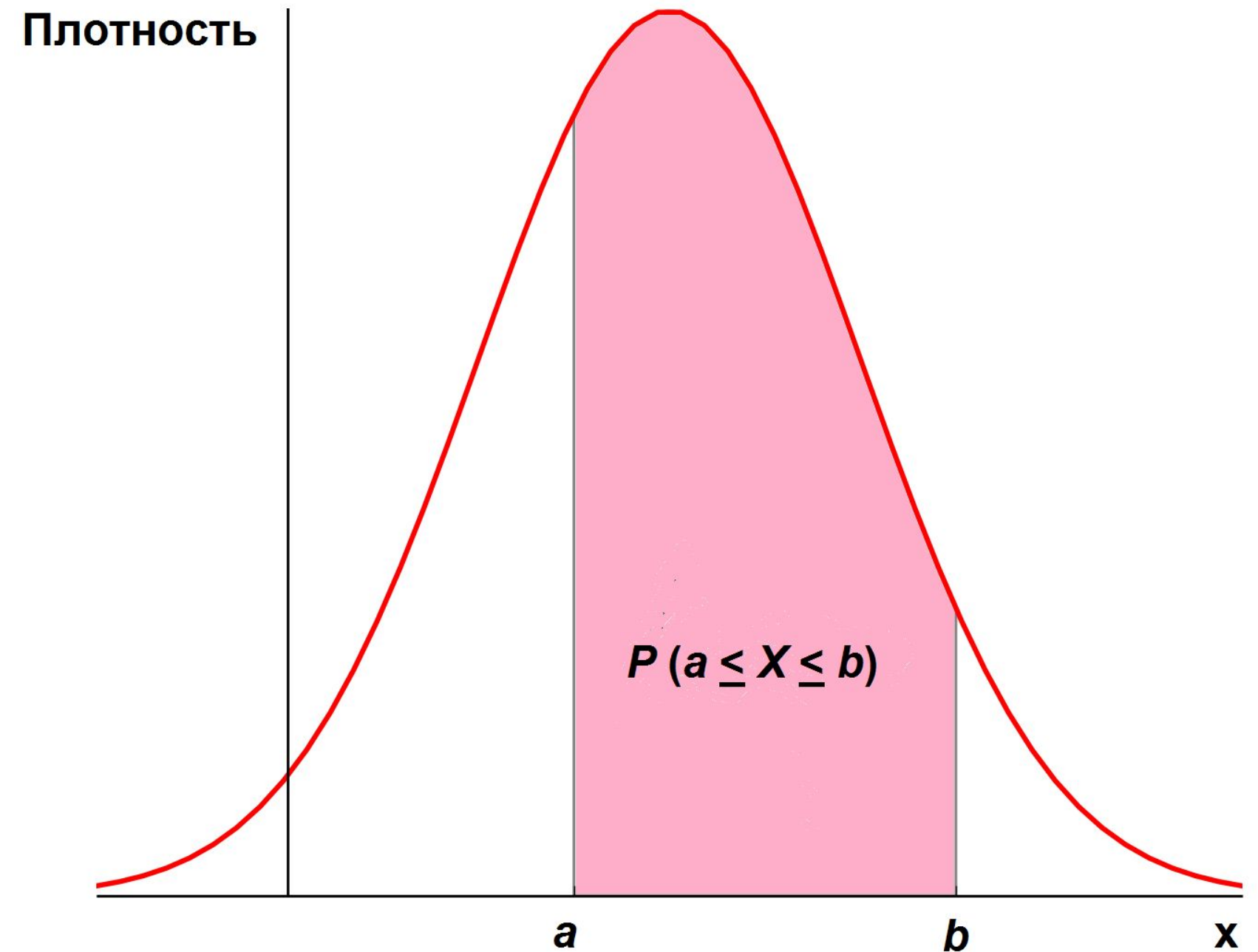
Возьмем какую-нибудь непрерывную случайную величину и посмотрим, как часто её значения попадают в тот или иной интервал. Для этого построим **гистограмму**. С ростом числа значений случайной величины и количества интервалов гистограмма будет превращаться в гладкую кривую – **плотность распределения**



# Распределения

Плотность распределения связана с вероятностью: **площадь фигуры под графиком плотности равна вероятности** того, что случайная величина  $X$  примет значение из интервала  $(a, b)$ , которым ограничена фигура

**N.B.** Вероятность того, что непрерывная случайная величина примет точно заданное значение, равна нулю. Фигура под графиком плотности схлопнется в вертикальный отрезок, имеющий нулевую площадь



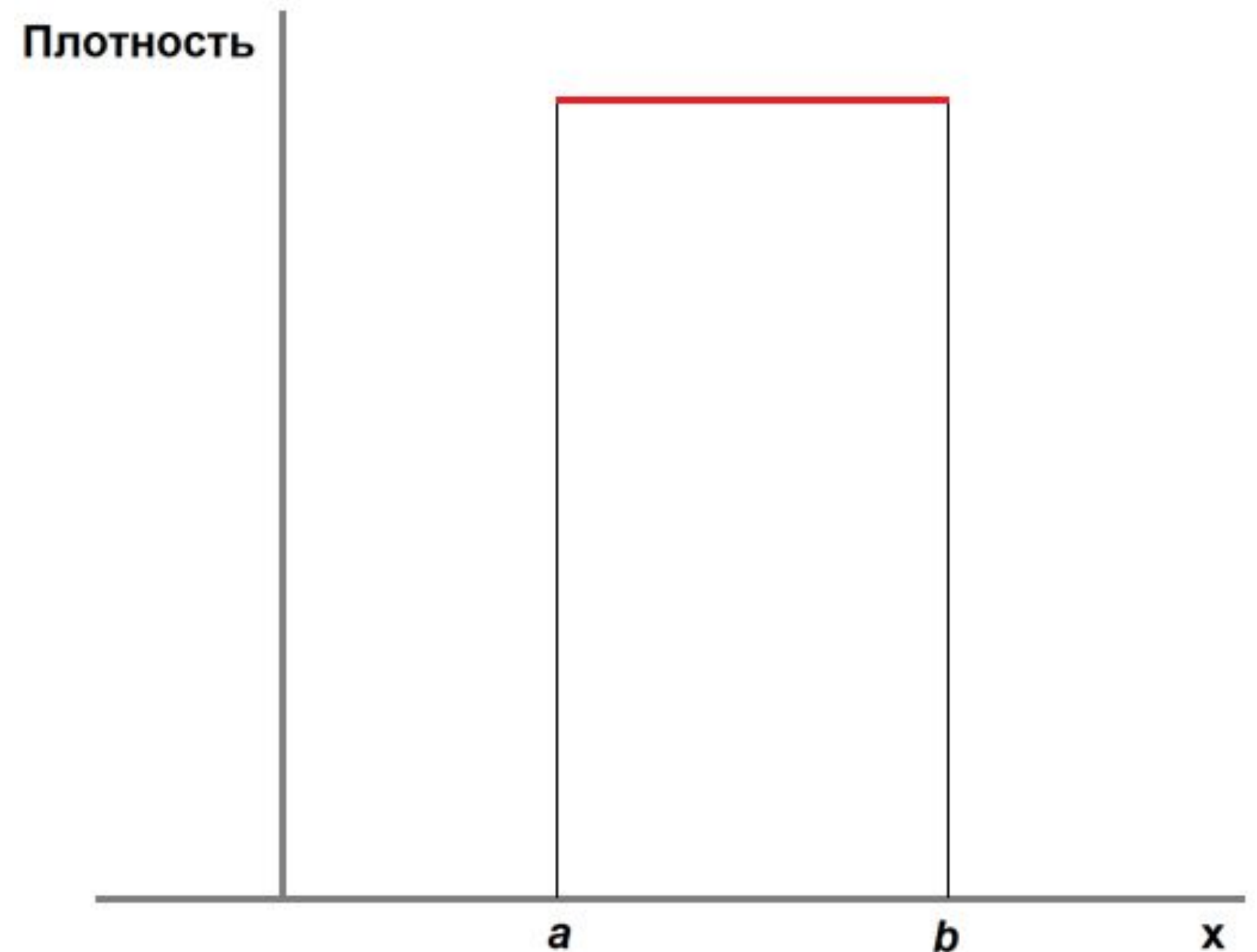
# Равномерное распределение

Случайная величина  $X$  равновероятно принимает любые значения в интервале  $(a, b)$

**Пример.** Если поезда метро следуют с одинаковыми промежутками 2 мин, то время ожидания поезда равномерно распределено между 0 и 2 мин

**Генерация**

`numpy.random.uniform(low, high, size)`



# Нормальное распределение

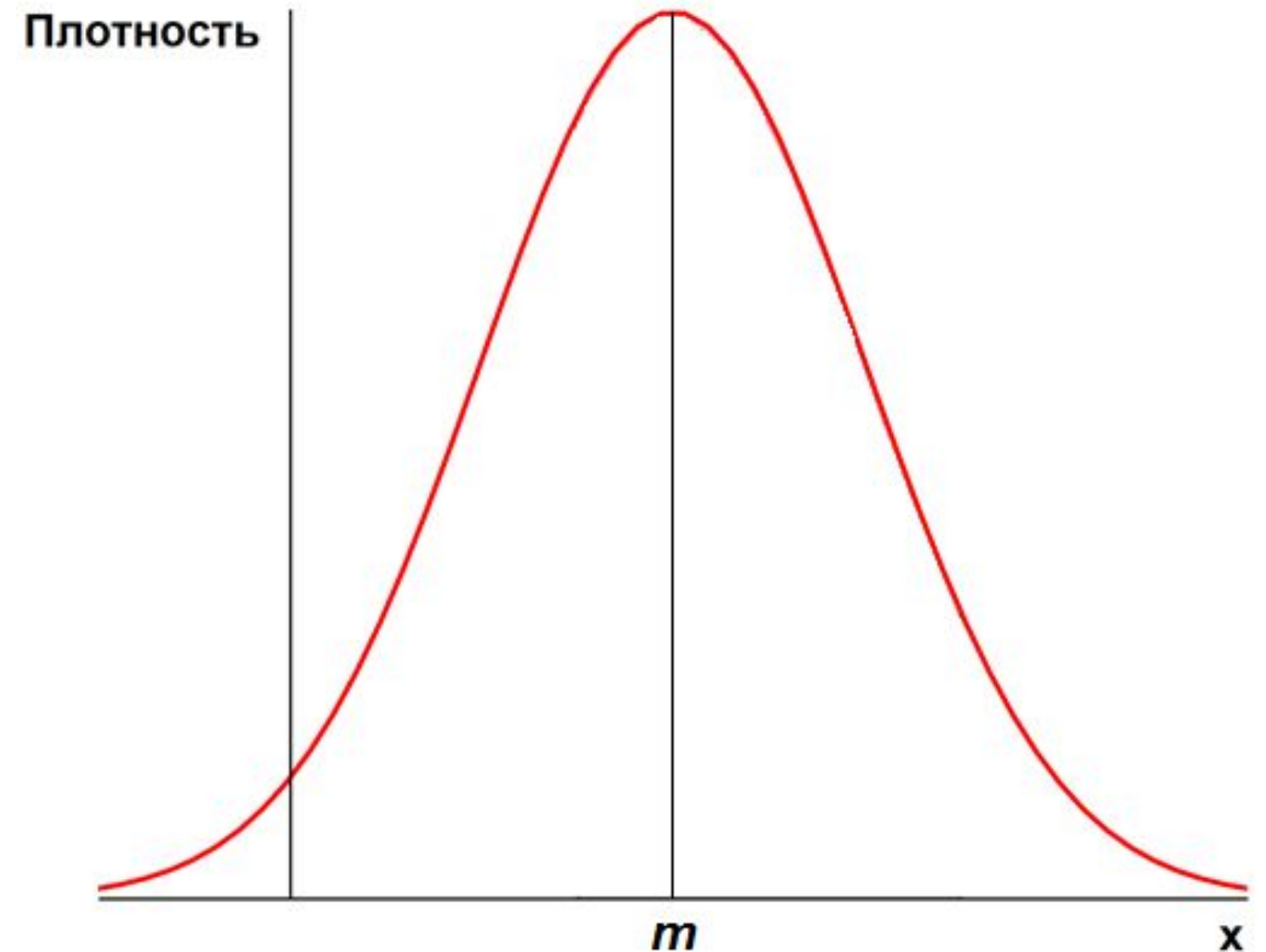
Случайная величина  $X$  принимает любые значения, однако вероятность тем меньше, чем дальше значение от среднего значения  $m$ . Чем меньше среднеквадратичное отклонение  $\sigma$ , тем сильнее значения «концентрируются» вокруг среднего, тем выше «пик» и тоньше «хвосты»

**N.B.** Вероятности попадания в интервалы:

$$P(m - 1\sigma \leq X \leq m + 1\sigma) = 0,68$$

$$P(m - 2\sigma \leq X \leq m + 2\sigma) = 0,95$$

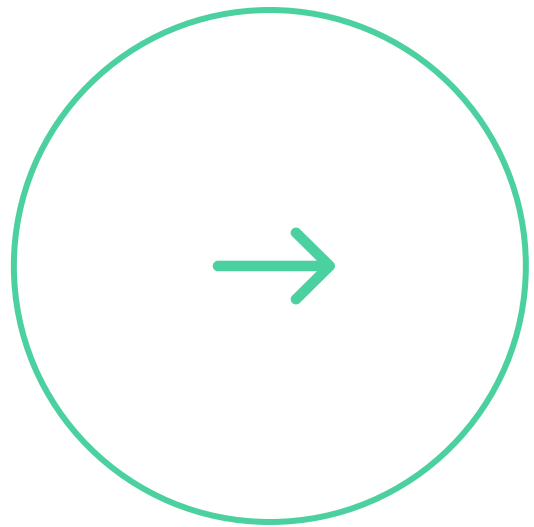
$$P(m - 3\sigma \leq X \leq m + 3\sigma) = 0,997$$





# Нормальное распределение

Исключительная роль нормального распределения связана с **законом больших чисел: сумма большого числа случайных величин с произвольным распределением имеет нормальное распределение**



**Пример.** При стрельбе по мишени на величину промаха влияет множество случайных факторов: дрожь руки, дыхание, ветер, сбитый прицел. Мы не знаем, как распределен каждый из этих факторов, однако в сумме они приводят к тому, что отверстия от пуль распределены по нормальному закону

Генерация: `numpy.random.normal(m, sigma, size)`

Jupyter Notebook: задачи 2, 3



# Условная вероятность

Бывают события, наступление которых изменяет вероятность других событий.  
Вероятность события  $A$  при условии наступления события  $B$  обозначается

$$P(A|B)$$

**Пример.** Изменится ли вероятность извлечения прибыли в бизнесе, если вдруг станет известно, что конкуренты резко снизили цены?

$A$  = "Бизнес принесет прибыль"

$B$  = "Конкуренты снизили цены"

Здравый смысл подсказывает, что в данном случае  $P(A|B) < P(A)$



# Полная вероятность

Пусть известно, что некоторые случайные события («гипотезы»)  $H_1, H_2 \dots$  приводят к одному и тому же событию  $A$  с разными вероятностями  $P(A|H_1), P(A|H_2) \dots$

Известны вероятности осуществления самих гипотез  $P(H_1), P(H_2) \dots$ , причём всегда осуществляется ровно одна гипотеза. **Какова вероятность события  $A$ ?**

Ответ даёт формула полной вероятности:

$$P(A) = P(A|H_1)P(H_1) + P(A|H_2)P(H_2) + \dots$$

**Н. В.** Поскольку всегда какая-то гипотеза осуществляется, то

$$P(H_1) + P(H_2) + \dots = 1$$



# Полная вероятность

**Пример:** вернемся к задаче «для маркетолога». Пусть вероятность интереса к продукту равна 20 % для людей младше 30 лет и 10 % для людей 30 лет и старше. Пусть 60 % участников сообщества младше 30 лет. Какова вероятность того, что случайный участник сообщества заинтересуется продуктом?

**События:**

$H_1$  = "Человеку до 30 лет"

$H_2$  = "Человеку 30 лет или больше"

$A$  = "Человек заинтересовался продуктом"

**Вероятности:**

$$P(H_1) = 0,6; P(H_2) = 0,4; P(A|H_1) = 0,2; P(A|H_2) = 0,1$$

**Решение:**

$$P(A) = 0,2 \cdot 0,6 + 0,1 \cdot 0,4 = 0,16 = 16\%$$



# Теорема Байеса

Вернёмся к гипотезам  $H_1, H_2 \dots$ . Предположим, что событие  $A$  случилось. Как это повлияет на наше доверие к гипотезам?

Формализуем вопрос: насколько отличается апостериорная вероятность гипотезы  $P(H_i|A)$  от априорной вероятности  $P(H_i)$ ?

Ответ даёт теорема Байеса:

$$P(H_i|A) = \frac{P(A|H_i)P(H_i)}{\underbrace{P(A)}_{\text{Вычислить по формуле полной вероятности}}}$$



# Теорема Байеса

**Пример:** вернемся к задаче «для врача» – анализу теста на COVID-19. Какова вероятность того, что человек заражен, если тест дал положительный результат?

Исходные данные

- **Распространенность инфекции (Prevalence)** – доля зараженных в популяции
- **Чувствительность теста (Sensitivity)** – способность теста реагировать на вирус (вероятность положительного результата при наличии вируса)
- **Специфичность теста (Specificity)** – способность теста не реагировать ни на что, кроме вируса (вероятность отрицательного результата при отсутствии вируса)



# Теорема Байеса

## События:

$H_1$  = "Человек заражен"

$H_2$  = "Человек не заражен"

$A$  = "Тест дал положительный результат"

## Вероятности:

$P(H_1)$  = Prevalence

$P(H_2)$  = 1 – Prevalence

$P(A|H_1)$  = Sensitivity

$P(A|H_2)$  = 1 – Specificity

## Решение:

$$P(H_1|A) = ?$$





# Теорема Байеса

Решение: применим теорему Байеса:

$$P(H_i|A) = \frac{\overbrace{P(A|H_i)}^{\text{Sensitivity}} \overbrace{P(H_i)}^{\text{Prevalence}}}{\underbrace{P(A|H_1)}_{\text{Sensitivity}} \underbrace{P(H_1)}_{\text{Prevalence}} + \underbrace{P(A|H_2)}_{1-\text{Specificity}} \underbrace{P(H_2)}_{1-\text{Prevalence}}}$$

$$P(H_i|A) = \frac{\text{Sensitivity} \times \text{Prevalence}}{\text{Sensitivity} \times \text{Prevalence} + (1 - \text{Specificity}) \times (1 - \text{Prevalence})}$$

Jupyter Notebook: задача 4

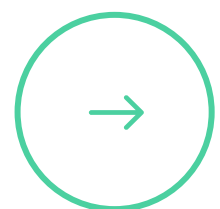
**Вывод:** вероятность наличия вируса повышается, если становится известно о положительном результате теста, но не достигает 100 % из-за несовершенства самого теста



# Домашнее задание

Смоделировать игру против лотерейного автомата типа "**777**". Игрок платит 1 руб., после чего выпадает случайное целое число, равномерно распределенное от 0 до 999. При некоторых значениях числа игрок получает выигрыш (см. справа)

- Выгодна ли игра игроку?
- Сколько в среднем приобретает или теряет игрок за одну игру?



**Дополнительное задание повышенной сложности.** Теоретически рассчитать средний выигрыш (проигрыш) и сравнить с результатами моделирования

**777:** 200 руб.

**999:** 100 руб.

**555:** 50 руб.

**333:** 15 руб.

**111:** 10 руб.

**\*77:** 5 руб.

**\*\*7:** 3 руб.

**\*00:** 2 руб.

**\*\*0:** 1 руб.

\* – любая цифра



---

# Случайные величины

Вопросы?