Конспект по теме «Теория вероятностей»

Что такое вероятность: эксперименты, элементарные исходы и события

Эксперимент — это повторяемый опыт, который может окончиться разными исходами, или, как принято говорить, **элементарными исходами**: исход либо случился, либо не случился.

В простейшем случае эти исходы не отличаются: мы не можем предпочесть один из них другому, — и значит вероятность каждого из них одинакова. Такие исходы называются **равновероятными**. В честном эксперименте (эксперимент с равновероятными исходами) с **n** элементарными исходами вероятность каждого исхода одинакова и равна 1/**n**.

Множество всех элементарных исходов эксперимента принято называть **вероятностным пространством**. Из него можно выделить подмножества, содержащие в себе некоторое количество элементарных исходов - **события**.

Невозможное событие - событие, которое не произойдёт никогда, вероятность его появления равна 0. **Достоверное событие** - событие, которое точно произойдёт, вероятность его которого равна 1. Вероятность появления других событий находится в промежутке от 0 до 1.

При сохранении условия равновероятности всех элементарных исходов **вероятность события** — количество исходов, входящих в это событие, делённое на общее количество исходов, то есть на размер вероятностного пространства.

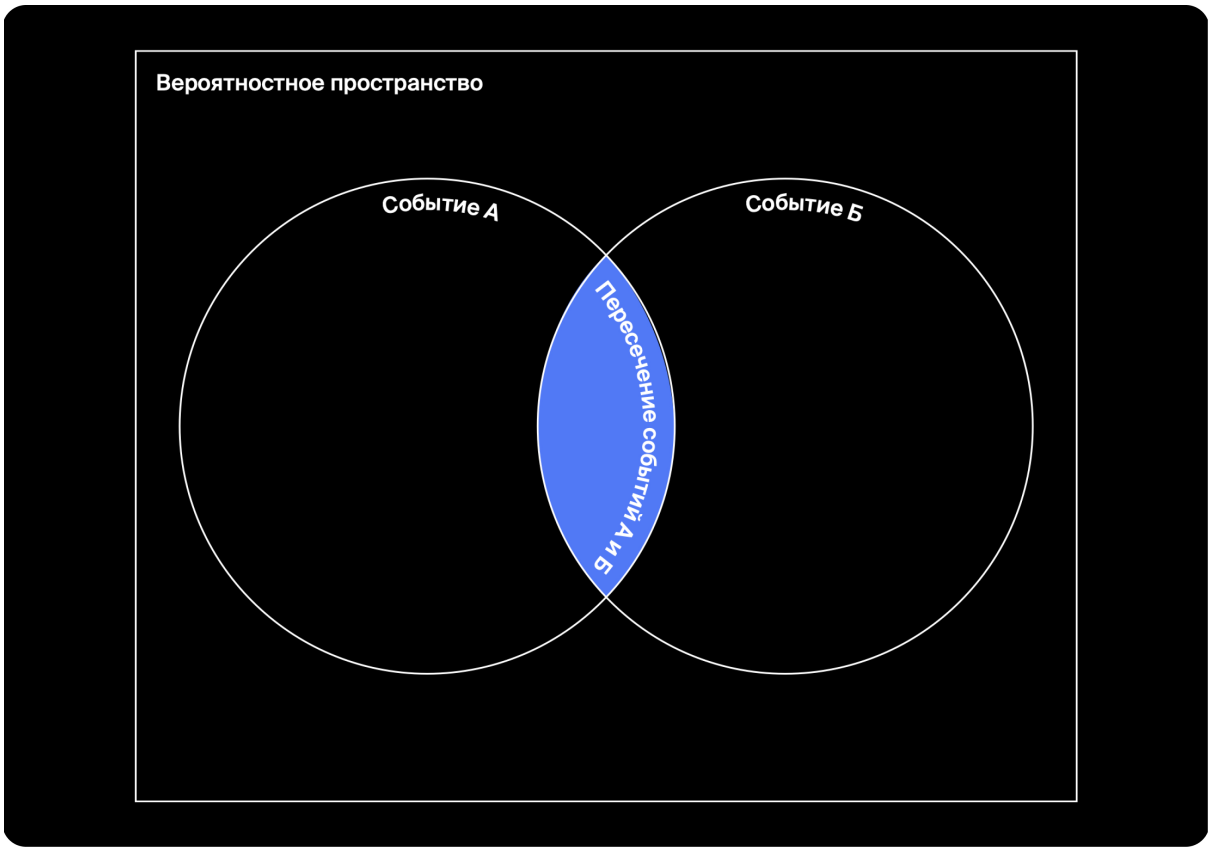
Закон больших чисел

Закон больших чисел: чем больше раз повторяется эксперимент, тем ближе частота заданного на этом эксперименте события будет к его вероятности.

Это правило можно использовать и в обратную сторону: если мы не знаем вероятность какого-то события, но можем много раз повторить эксперимент, по частоте выпадания исходов, входящих в это событие, можно судить о его вероятности.

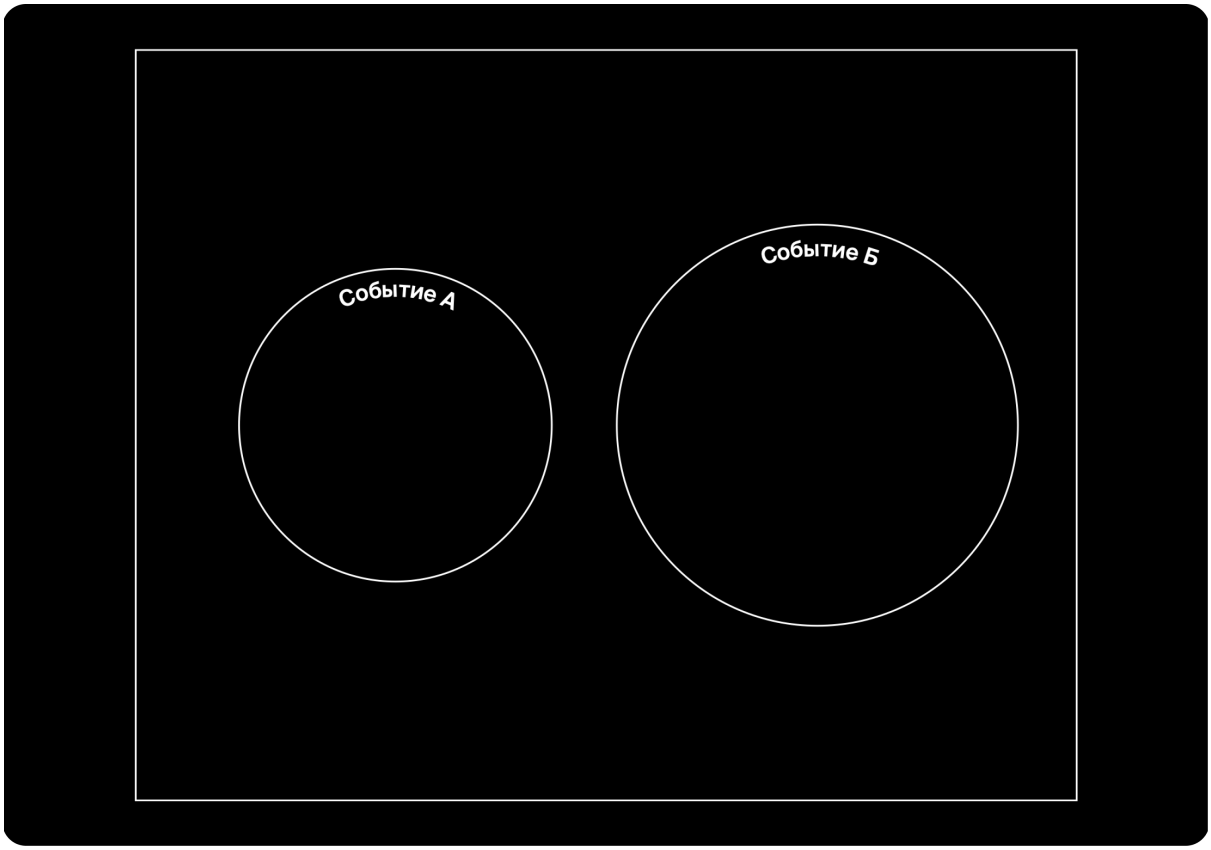
Взаимоисключающие и независимые события, умножение вероятностей

Для отображения пересечения между событиями, используется **диаграмма Эйлера-Венна**:



События А и Б пересекаются, значит существуют элементарные исходы, входящие и в А, и в Б.

Взаимоисключающими называют события, которые не могут произойти одновременно при проведении эксперимента — на диаграмме Эйлера-Венна они не пересекаются:



Вероятность взаимоисключающих событий равна нулю.

События называются **независимыми**, если наступление одного из них не влияет на вероятность другого. Если события независимы, то вероятность их пересечения равна произведению их вероятностей. Это правило работает и в обратную сторону.

Если взаимоисключающие события охватывают всё вероятностное пространство, сумма их вероятностей равна единице.

Взаимоисключаемость событий видна на диаграмме Эйлера-Венна. А вот независимость так просто не обнаружишь, нужно проверять условие равенства произведения вероятностей событий вероятности их пересечения.

Случайные величины, распределение вероятностей и интервалы значений

Случайная величина — это переменная, которая принимает **случайные значения** - те значения, которые нельзя предсказать до проведения эксперимента. У эксперимента есть исходы, которые могут описываться как количественно, так и качественно. Случайная же величина определяется на этих исходах *численно*. Это способ спроецировать исходы эксперимента, как бы они ни определялись, на числовую ось.

Как и все количественные переменные, случайная величина может быть **дискретной** или **непрерывной**.

Распределением вероятности случайной величины называется таблица, содержащая всевозможные значения случайной величины и вероятности их появления.

Для хранения числовых таблиц, используется тип данных **numpy array** из библиотеки Numpy:

```
table = np.array([[2,3,4,5,6,7],
[3,4,5,6,7,8],
...
[7,8,9,10,11,12]])
```

Если при работе со словарём вам необходимо получить список всех ключей словаря, то это можно сделать с помощью метода `keys()`. А список всех значений словаря — с помощью метода `values()`:

```
dictionary = {...}
print(dictionary.keys())
print(dictionary.values())
```

Математическое ожидание и дисперсия

Для эксперимента можно задать случайную величину и найти численное значение, к которому она будет в среднем стремиться при многократном повторе эксперимента. Это значение называется **математическим ожиданием** случайной величины.

Если эксперимент состоит из *равновероятных элементарных исходов*, заданных численно, математическое ожидание будет равно *среднему* возможных значений.

Математическое ожидание случайной величины — сумма всех значений случайной величины, помноженных на их вероятности:

$$E(X) = \sum p(x_i)x_i$$

Математическое ожидание — аналог метрики локации, только не для датасета, а для случайной величины. Оно показывает, вокруг какого значения распределена случайная величина, и — по закону больших чисел — к какому значению она будет в среднем стремиться при повторе эксперимента.

Поскольку случайная величина распределена вокруг этой «метрики локации», можно найти и меру её разброса. Для этого нужно найти математическое

ожидание квадрата случайной величины — это несложно если учесть, что значения меняются, а их вероятности — нет.

Если мы знаем математическое ожидание самой случайной величины и её квадрата, **дисперсию** находят по формуле:

$$Var(X) = E(X^2) - (E(X))^2$$

Вероятность успеха в биномиальном эксперименте

Эксперименты с двумя возможными исходами называются **биномиальными экспериментами**. Обычно один из результатов называют успехом, а второй, соответственно, неудачей. Если вероятность успеха равна `p`, то вероятность неудачи `(1 - p)`.

Биномиальное распределение

Количество способов выбрать **k** успехов из **n** повторений эксперимента находят по формуле:

$$C_n^k = \frac{n!}{k!(n-k)!}$$

где <число>! (читается как <число> факториал) равно произведению всех натуральных чисел от 1 до этого числа: n! = 1 · 2 · 3 · 4 · ... · (n-1) · n.

Для вычисления факториала используется библиотека *math* и её метод *factorial*:

```
from math import factorial
x = factorial(5)
```

Рассмотрим задачу о биномиальном эксперименте (повторении эксперимента с двумя исходами *n* раз) в общем виде. Если вероятность успеха `p` и неуспеха `1 - p`, а эксперимент был повторен *n* раз, то вероятность любого количества успехов *k* из этих *n* экспериментов:

$$P(k \text{ успехов из } n \text{ попыток}) = C_n^k p^k (1-p)^{n-k}$$

Условия, при которых можно утверждать что случайная величина распределена биномиально:

- проводится конечное фиксированное число попыток *n*;
- каждая попытка — простой биномиальный эксперимент ровно с двумя исходами;
- попытки независимы между собой;

- вероятность успеха *p* одинаковая для всех *n* попыток.

Нормальное распределение

Ключевая теорема в статистике — **центральная предельная теорема**. Если немного упростить, она гласит: «Много независимых случайных величин, сложенных вместе, дают нормальное распределение».

Нормальное распределение описывает множество реальных непрерывных величин. Нормальное распределение определяют два параметра — среднее и дисперсия:

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

Эта запись читается так: переменная *X* распределена нормально со средним мю (μ) и дисперсией сигма в квадрате (σ^2), то есть стандартным отклонением сигма.

Для того, чтобы по известным параметрам распределения найти вероятность попадания в те или иные интервалы, вызовем два метода из пакета *scipy.stats:* **norm.ppf** и **norm.cdf**:

- ppf — функция процентных значений;
- cdf — кумулятивная функция распределения.

Обе работают с нормальным распределением, заданным своими средним и стандартным отклонением.

- Функция *norm.ppf* выдаёт *значение переменной* для известной *вероятности* интервала слева от этого значения.
- Функция *norm.cdf,* наоборот, выдаёт для известного значения *вероятность* интервала слева от этого значения.

Чтобы задать нормальное распределение, используется метод *norm()* из пакета *scipy.stats* с двумя аргументами: математическим ожиданием и стандартным отклонением. Найдём вероятность получить некоторое значение **x**:

```
from scipy import stats as st

# задаем нормальное распределение
distr = st.norm(1000, 100)

x = 1000

result = distr.cdf(x) # считаем вероятность получить значение x
```

С помощью функции **norm.cdf** можно посчитать вероятность получить значение в промежутке от **x1** до **x2:**

```
from scipy import stats as st

# задаем нормальное распределение
distr = st.norm(1000, 100)

x1 = 900
x2 = 1100

result = distr.cdf(x2) - distr.cdf(x1) # считаем вероятность получить значение между x1 и x2
```

Для того, чтобы по вероятность получить значение, воспользуемся методом **norm.ppf**:

```
from scipy import stats as st

# задаем нормальное распределение
distr = st.norm(1000, 100)

p1 = 0.841344746

result = distr.ppf(p1)
```

Нормальная аппроксимация биномиального распределения

При большом количестве повторений биномиального эксперимента биномиальное распределение приближается к нормальному.

Для дискретного биномиального распределения, заданного числом повторов эксперимента `n` и вероятностью успеха `p`, математическое ожидание равно `n·p`, а дисперсия: `n·p·(1-p)`.

Если *n* больше 50, эти параметры биномиального распределения можно взять как среднее и дисперсию для нормального распределения, которое будет достаточно близко описывать биномиальное. Максимально близкое к биномиальному нормальное распределение задаётся его математическим ожиданием `n·p` в качестве среднего значения и дисперсией `n·p·(1-p)`.