KARPOV.COURSES >>>
КОНСПЕКТ

> Конспект > 1 урок > Зачем нужна статистика и A/B тесты

> Оглавление

> **Оглавление**

> Введение в A/B тестирование

> Теория вероятностей. Вероятность

- Что такое вероятность?
- Частотный(статистический подход) к вычислению вероятности

> Свойства вероятности

- Вероятность несуществующего события:
- Вероятность противоположного события:
- Вероятности вложенных событий:

> Условная вероятность

- Независимость событий:
- Формула Байеса:
- Формула полной вероятности:
- Разберем пример:

> Случайная величина

- Типы случайных величин:
- Характеристики случайных величин:

> Что такое статистика

- Понятия из статистики:

> Описательные статистики

> Дискретные распределения

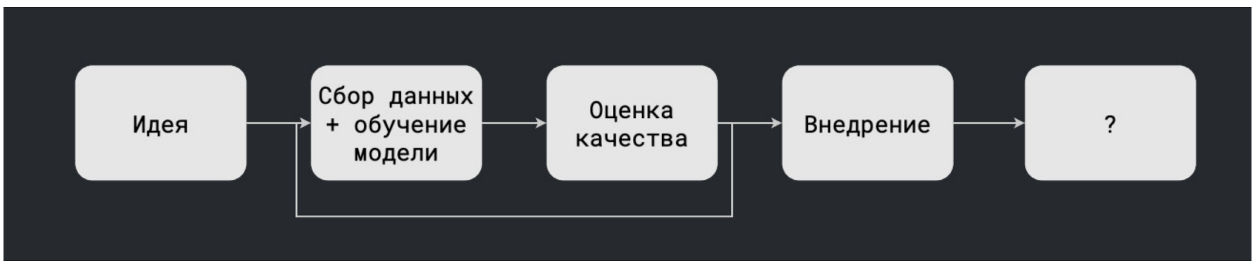
- Распределение Бернулли
- Биномиальное распределение

Равномерное распределение
Распределение Пуассона

> Введение в A/B тестирование

У вас уже есть знания, как подготавливать данные для обучения модели, какие модели можно обучать и как это можно делать, как в последующем внедрять полученные модели.

Чему осталось научиться: как *оценивать* внедряемые ML-решения с точки зрения реальных пользователей.



Зачем это нужно? Разберем на примере:

Предположим, что существует приложение, в котором мы показываем видео пользователям. Пользователь смотрит видео, если оно ему не понравилось или он его досмотрел — переходит к следующему видео. Мы, разработчики, со своей стороны хотим показывать пользователю релевантные видео, чтобы он проводил как можно больше времени в приложении.

Допустим, у нас уже есть готовая модель, которая выбирает следующие видео для показа пользователям. Но со временем мы придумали новую модель, которая учитывает еще какой-то дополнительный параметр.

Возникает вопрос: насколько она лучше того, что у нас было раньше?

Может показаться, что мы уже это проходили: учились оценивать качество модели. Например, можно посмотреть на метрики ранжирования (*precision@k*, *mAP@k* и т.д).

На самом деле компании не так важно, какое значение любой из вышеперечисленных метрик мы получили. Как правило, это важно нам или нашим коллегам. Компании важно, как пользователь взаимодействует с приложением.

Например, компании больше интересны такие метрики, как:

- Сколько видео посмотрел пользователь за сессию?
- Сколько времени провел пользователь в приложении?
- Сколько лайков оставил пользователь?
- Вернулся ли пользователь завтра?

Возможное решение: смотреть на наши новые метрики.

Проблема: Мы не можем показать одному и тому же пользователю два разных алгоритма.

Так мы подошли к сути A/B экспериментов: одному пользователю показать результат одного алгоритма, а другому пользователю — результат другого, и уже их метрики сравнить между собой.

Основная проблема, которую решают A/B эксперименты:

Как оценить изменения, если наши метрики считаются только онлайн.

Этому и будем посвящен текущий блок.

> Теория вероятностей. Вероятность

К этому моменту у вас появилось представление о том, какую проблему решают A/B эксперименты, и в каких случаях они могут быть нам полезны.

Изучение этих вопросов мы начнем с основ Теории Вероятностей и Математической Статистики.

Рассмотрим пример.

- Автобусы приезжают последовательно на остановку
- Есть расписание, но невозможно ему идеально следовать

- Приезд автобуса в конкретный момент — случайность (может приехать сейчас, а может позднее).

Случайность — результат некоторого процесса, который нельзя точно предсказать.

Какие ещё могут быть случайности:

- Сколько времени займёт дорога на такси в аэропорт?
- В какую сторону изменится курс валюты завтра?
- Понравится ли предложенная песня пользователю?
- Выплатит ли компания кредит?

Теория Вероятностей — дисциплина, которая изучает случайность с математической точки зрения.

В общем виде в теории вероятностей рассматриваются следующие вопросы:

- Какие модели могут описывать случайности?
- Какие свойства есть у этих моделей?
- Как можно описать явление?
- Какую новую информацию о явлениях можно получить?

Что такое вероятность?

Вероятность — количественная оценка возможности наступления некоторого события.

Событие — нечто, что может произойти или не произойти (состоит из элементарных исходов). Например, «сегодня пойдет снег».

Частотный(статистический подход) к вычислению вероятности

Вероятность выпадения орла у монетки с двумя гранями - 50%. Но откуда берутся эти 50%?

- Кидаем монетку 10 раз, встретили орёл 4 раза - 4/10 = 40 %
- Кидаем монетку 100 раз, встретили орёл 48 раз - 48/100 = 48 %

- Кидаем монетку 1000 раз, встретили орёл 495 раз - 495/1000 = 49.5 %

…Как можно заметить, чем больше число экспериментов проводим, тем ближе соотношение выпадения орла приближается к 0.5

- Определим вероятность события А:

$$P(A) = lim_{N \to \infty} \frac{N(A)}{N}$$

N — общее количество наблюдений

$N(A)$ — количество наблюдений события А

Больше информации

> Свойства вероятности

Вероятность несуществующего события:

$$P(\oslash) = 0$$

«Завтра будет идти снег при ясной погоде в +30»

Вероятность противоположного события:

$$P(\overline{A}) = 1 - P(A)$$

«Если вероятность дождя завтра 60%, то шансы на отсутствие дождя — 40%»

Вероятности вложенных событий:

$$A \subseteq B \Rightarrow P(A) \leq P(B)$$

«Сильный дождь с ветром имеет меньшую вероятность, чем любой дождь»

Возможные значения вероятностей:

$$0 \leq P(A) \leq 1$$

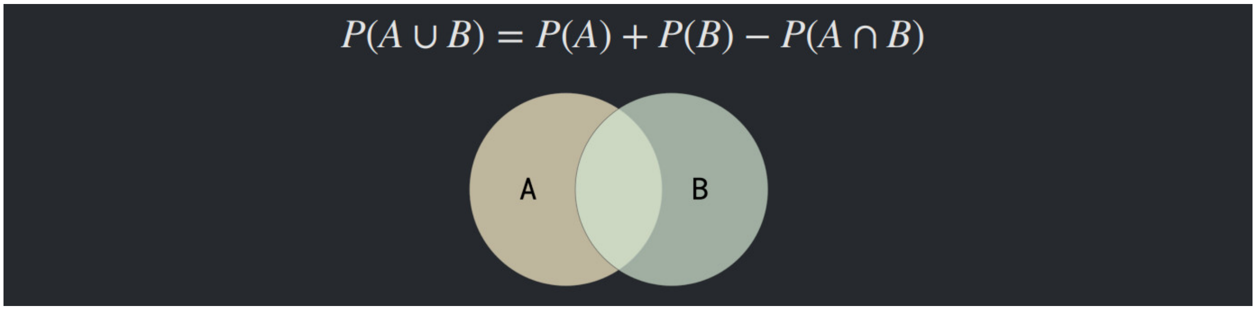
«Шансы на дождь не могут быть меньше 0% и больше 100%»

Вероятность разности событий:

$$A \subseteq B \Rightarrow P(B \backslash A) = P(B) - P(A)$$

«Если вероятность дождя 60%, а вероятность сильного дождя — 20%, то вероятность слабого дождя — 40%»

Вероятность объединения событий:



$$P(A \cup B) = A + B$$

$$P(A \cap B) = A * B$$

Больше информации

> Условная вероятность

Условная вероятность события A при условии события B — вероятность наступления события A при условии наступления события B. Например, какая вероятность, что автобус приедет в течение 5 минут(A), если идёт дождь(B)? Или какая вероятность, что акции компании вырастут, если будет опубликована хорошая финансовая отчётность?

Обозначается как $P(A|B)$ и вычисляется по формуле:

$$P(A \mid B) * P(B) = P(A \cap B)$$

Если умножить условную вероятность события A при условии события B на вероятность события B, то получится вероятность одновременного наступления события A и B.

Условная вероятность — уточнение априорной (без условий) вероятности.

Например, вероятность того, что автобус приедет в течение 5 минут, можно уточнить, если узнать:

- Какая сейчас погода?
- Где находится автобус?
- Насколько большой трафик?

Независимость событий:

События называются независимыми, если условная вероятность события A при условии события B совпадает с вероятностью события A:

$$P(A \mid B) = P(A)$$

То есть вероятность наступления события B никак не влияет на наступление события A

«Выпадение снега завтра не зависит от того, надел ли я сегодня шапку»

Как найти условную вероятность?

Формула Байеса:

Если поменять местами события в формуле условной вероятности, то получится соотношение. Из него можно выразить одну условную вероятность через другую:

$$P(A \mid B) {*} P(B) = P(A \cap B) = P(B \mid A) {*} P(A)$$

Данная формула называется формулой Байеса, и может быть записана еще по-другому:

$$P(A \mid B) = \frac{P(B)P(B|A)}{P(A)}$$

$P(A)$ и $P(B)$ — это вероятности событий A и B, соответственно.

$P(A \mid B)$ — вероятность события A, **если наступило событие B.**

$P(B \mid A)$ — вероятность события B, **если наступило событие A.**

Очень важно не путать $P(A \mid B)$ и $P(B \mid A)$! Они могут казаться похожими, но очень редко совпадают в реальной жизни. Скажем, если с вами рядом летает комар, то довольно высока вероятность получить укус — **P(получить укус|рядом комар)**. Но если вы получили укус, то это необязательно сделал именно комар — вас могли укусить любая другая пакость, собака или змея — **P(рядом комар|получить укус)**.

Формула полной вероятности:

Предположим, что есть несколько событий B_i, и они не пересекаются. И хотя бы одно из этих событий всегда происходит. Тогда можно выразить вероятность события A, используя формулу:

$P(A) = \sum_i P(A \mid B_i) * P(B_i)$

Допустим, мы замерили, с какой вероятностью автобус приезжает в течение 5 минут во время дождя (0.5) и без осадков (0.8)

Вероятность дождя — 0.2, какая вероятность приехать в течение 5 минут для любой погоды?

Решение: P(в течение 5 минут) = P(в течение минут / дождь) P(дождь) + P(в течение минут / нет осадков) P(нет осадков) =

= 0.5 * 0.2 + 0.8 * 0.8 = 0.74

Разберем пример:

Есть 2 партии изделий:

- размер первой — 2000, вероятность брака 0.1
- размер второй — 5000, вероятность брака 0.2

Если мы взяли случайное изделие и оно оказалось бракованным, то с какой вероятностью мы взяли изделие из первой партии?

$P(batch_1) = \frac{2000}{2000+5000} = \frac{2}{7}$

$P(defect|batch1) = 0.1$

$P(batch_2) = 1 - P(batch_1) = \frac{5}{7}$

$P(defect|batch_2) = 0.2$

$P(batch_1|defect) = \frac{P(defect|batch1) \cdot P(batch1)}{P(defect)} = \frac{P(defect|batch1)\ cdot P(batch1)}{P(defect|batch1)P(batch1)+P(defect|batch2)P(batch2)}$

$= \frac{0.1 \cdot \frac{2}{7}}{0.1 \cdot \frac{2}{7} + 0.2 \cdot \frac{5}{7}} = \frac{1}{6}$

Для дополнительного погружения в идею формулы Байеса можно почитать эту заметку.

> Случайная величина

Случайная величина — величина, принимающая некоторое числовое значение в зависимости от исхода с определенной вероятностью.

До этого мы ввели понятие случайности, как же случайность соотносится со случайной величиной? Если случайность — это приедет ли автобус в некоторый момент времени, то **количество** минут до ближайшего автобуса — случайная величина.

Распределение случайной величины — какие значения случайная величина принимает и с какими вероятностями.

«Количество очков, выпавшее на игральном кубике от 1 до 6 с вероятностью 1/6»

Типы случайных величин:

Количественные – измеренные значения некоторого признака:

- непрерывные – могут принимать любое значение на определенном промежутке. *Пример: рост человека.*
- дискретные – могут принимать определенные значения. *Пример: число детей в семье (целые неотрицательные числа, то есть 3.5 ребенка быть не может).*

Качественные (номинативные / категориальные) – делят наши объекты на группы.

Пример: кодировка пола человека (0 — мужчина, 1 женщина)

Также переменные могут быть ранговыми, например, когда мы смотрим на результаты марафона: 1 – прибежал первым, 2 – вторым и так далее. Мы не знаем, насколько различаются результаты участников между собой, но знаем их порядок.

Важно отметить, что некоторые переменные, в зависимости от того, в какой шкале они представлены, могут относиться к разным категориям. Одним из таких примеров является переменная возраста.

Количественная непрерывная — возраст, измеренный в днях/месяцах/годах.

Ранговая переменная — возраст разбит на группы (очень часто встречается в анкетировании) — от 14-17 лет; 18-29 лет и так далее.

Характеристики случайных величин:

Математическое ожидание — среднее случайной величины, которое считается как среднее арифметическое с вероятностями в качестве весов.

$$E(x) = \sum_{i=1}^{N} x_i P(x_i)$$

Дисперсия — показатель разброса случайной величины.

$$D(x) = \frac{1}{N} \sum_{i=1}^{N} P(x_i)(x_i - E(x))^2$$

> Что такое статистика

Математическая статистика — дисциплина, которая изучает, как собирать данные, как систематизировать, описывать и делать из них выводы.

Цель статистики — описывать процессы со случайностью (определить, какому распределению принадлежат данные).

Понятия из статистики:

Генеральная совокупность — совокупность всех объектов, относительно которых предполагается делать выводы в конкретной задаче:

- **Все** автобусы, приезжающие на остановку
- **Все** пользователи нашего сервиса
- **Все** произведённые детали на заводе

Выборка — часть генеральной совокупности, с которой будем работать. (Часть деталей, которые мы будем проверять).

- Наш замер приезда автобусов в конкретный день
- Выборка пользователей, которая пользовалась приложением на этой неделе
- Выборка деталей из партии, которую мы можем проверить

Репрезентативность — свойство соответствия характеристик выборки характеристикам генеральной совокупности (описывает ли выбранная нами часть устройства все остальные).

Например, описывает ли детали для проверки все производимые детали. Таким образом можем судить о генеральной совокупности только если выборка репрезентативна.

Больше информации

> Описательные статистики

Статистика — любая измеримая функция от выборки.

Мода — наиболее часто встречающееся значение выборки.

`df.column_1.mode()`

Медиана — элемент выборки, который больше половины выборки и меньше другой половины выборки (среднее арифметическое двух соседних значений для четного размера выборки).

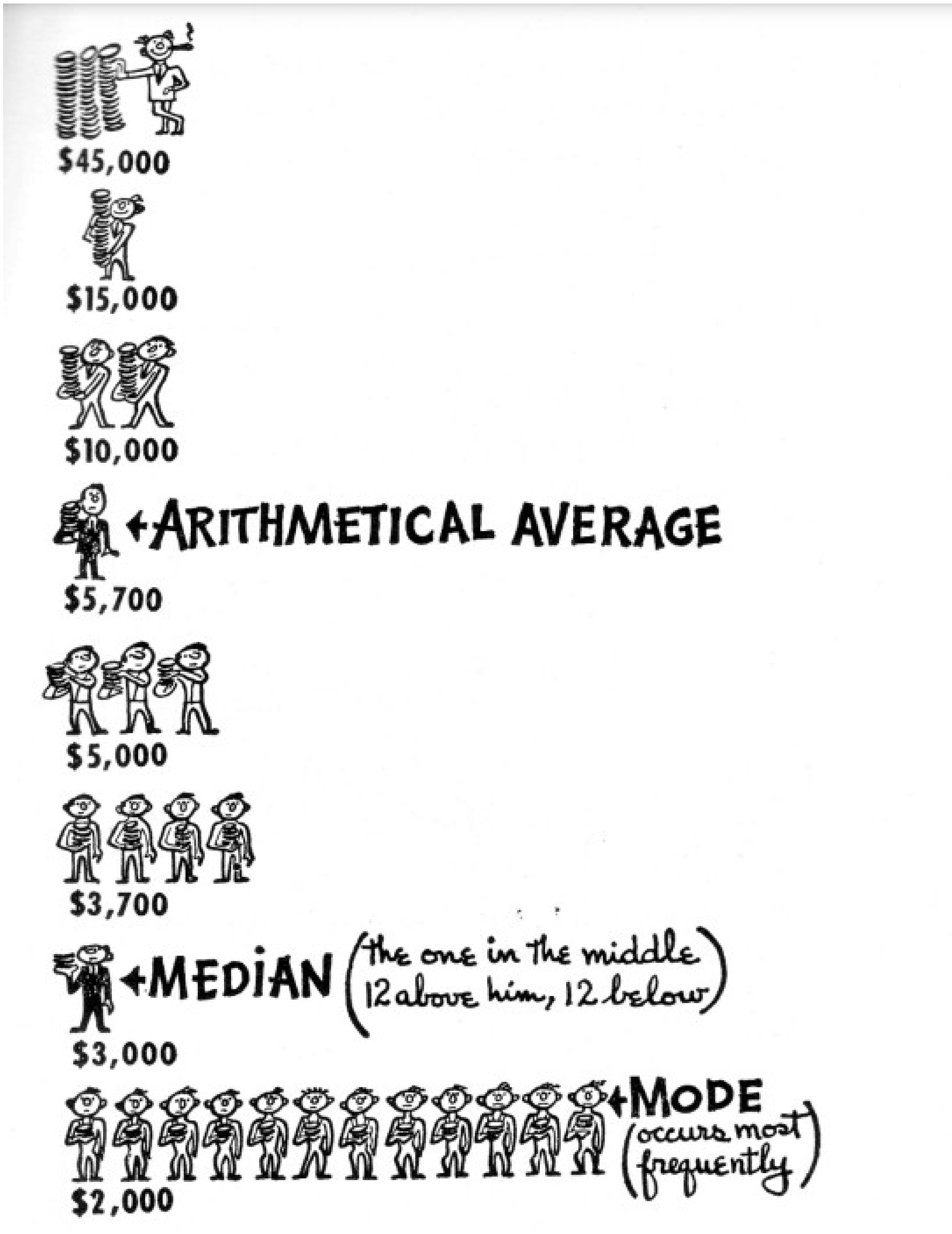
`df.column_1.median()`

Среднее — среднее арифметическое значений (оценка на математическое ожидание случайной величины).

$\overline{X} = \frac{x_1 + x_2 + \cdots + x_n}{n}$

$\overline{X}$ - среднее выборки

`df.column_1.mean()`



"Как лгать с помощью статистики"(Иллюстрация из книги)

Когда не стоит использовать среднее значение, а лучше брать моду или медиану:

- явная асимметрия

- заметные выбросы
- несколько мод

Со «средним» понятно, а как оценить разброс значений в выборке?

Дисперсия — средний квадрат отклонений индивидуальных значений признака от их средней величины.

$D(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2$

```
df.A.var()
np.var(df.A)
```

Среднеквадратичное отклонение - корень из дисперсии:

```
# среднеквадратичное отклонение
std = np.mean((data - avg) ** 2) **0.5
print(std)
```

- А если посчитать «ширину распределения» и «как широко сконцентрирована основная часть значений»?
- Понадобится описывать расположение «частей выборки» —квантили (до какого значения находится некоторый процент выборки)

α-квантиль — значение выборки, которое больше α части выборки (0.3 квантиль больше 30% выборки).

Например, в распределении [3, 4, 4, 5, 6, 7, 10] — «3.1» больше 1/7 выборки, а «4.1» больше 3/7 (то есть больше 30% выборки)

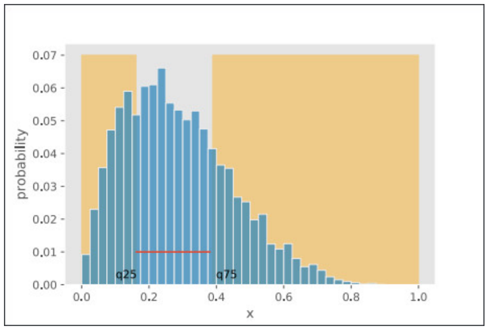
```
df.quantile(q=0.75)
```

Как измерить ширину распределения?

Интерквартильный размах — разность между 0.75 и 0.25 квантилями.

- Посчитаем расстояние между 2 квантилями
- «Отрежем левый и правый хвост»

- Оценим ширину «середины распределения»




```
#квантили и межквартильный размах
q25, q75 = np.quantile(data, [0.25,0.75])

# медиана ==0.5-квантиль
q50 = np.quantile(data, q=0.5)
median = np.median(data)
```

> Дискретные распределения

Распределение Бернулли

- Всего два возможно значения («успех» и «неуспех»)
- Вероятность «успеха» — p, тогда у «неуспеха» — (1 - p)
- $E(x) = p$
- $D(x) = p(1 - p)$

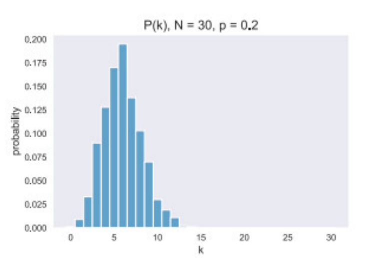
«Какой стороной упадет монетка?»

«Будет ли завтра дождь?»

Биномиальное распределение

- N независимых одинаково распределенных по Бернулли величин
- Подсчитываем число «успехов» k из N испытаний
- Биномиальное распределение моделирует число «успехов» k из N испытаний

$$P(k) = C_N^k p^k (1-p)^{N-k}$$
$$E(x) = p^N$$
$$D(x) = p(1-p)^N$$



Пример:

- В коробке N деталей
- Каждая деталь оказывается бракованной с вероятностью p
- Сколько бракованных деталей есть в коробке?
- Какая вероятность, что в коробке нет бракованных деталей?

Равномерное распределение

- N возможных значений случайной величины (от a до b)
- Каждое значение встречается с одинаковой вероятностью (одинаково часто)

$$\mathbb{P}(k) = \frac{1}{N}$$
$$\mathbb{E}(x) = \frac{a+b}{2}$$
$$\mathbb{D}(x) = \frac{N^2-1}{12}$$

Пример:

- Игральный кубик с 6 гранями
- Каждая грань выпадает с одинаковой вероятностью
- Какая вероятность встретить 2?
- Какая вероятность встретить чётное число?

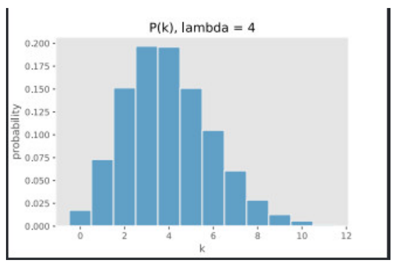
Распределение Пуассона

- Некоторые события происходят с фиксированной интенсивностью и независимо друг от друга
- Распределение Пуассона моделирует число событий k за фиксированный промежуток времени
- Пусть λ — интенсивность (число событий в промежуток времени)

$P(k) = \frac{\lambda^k}{exp^{-\lambda}}$

$E(x) = \lambda$

$D(x) = \lambda$



Пример:

- Автобусы приезжают на остановку с некоторой интенсивностью (в среднем 3 автобуса за 10 минут)
- Приезд каждого автобуса не зависит от других автобусов
- Какая вероятность, что за 10 минут не приедет ни одного автобуса?

Больше информации про распределения здесь