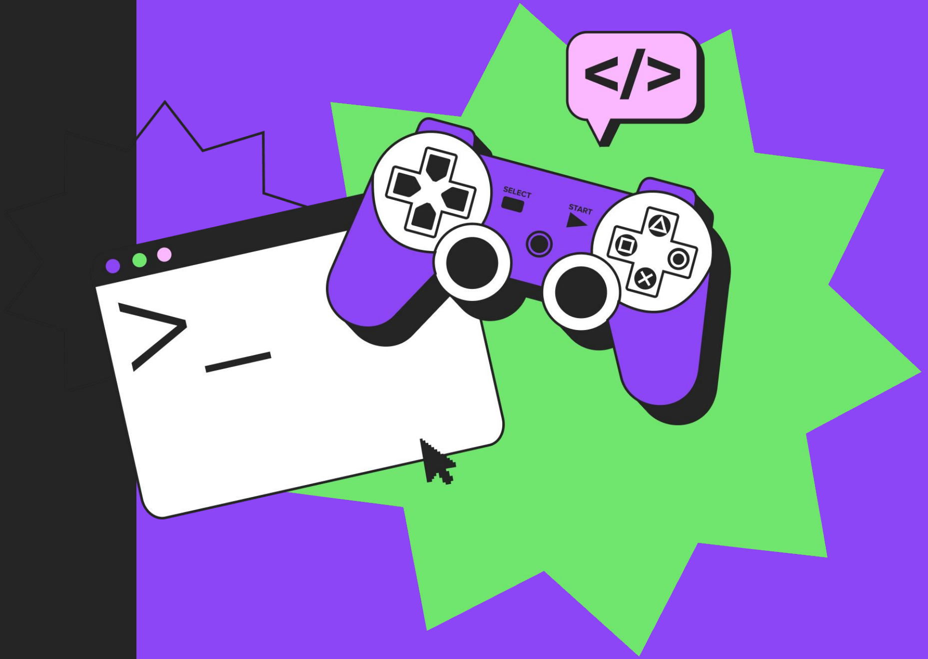
GeekBrains

Теория вероятностей и статистика

Познакомимся с понятиями теории вероятностей и математической статистики. Изучим закон распределения случайных величин и технику проверки статистических гипотез.



Максим Кулаев

Team Lead Data Scientist, VK

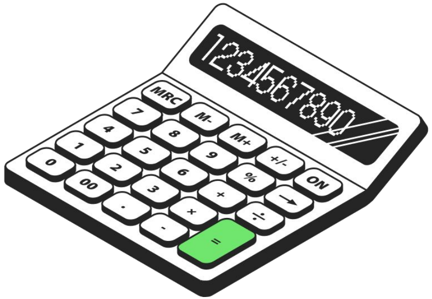
Окончил с отличием НИЯУ МИФИ по направлению «Информационно-аналитические системы безопасности». Аспирант в НИУ ВШЭ

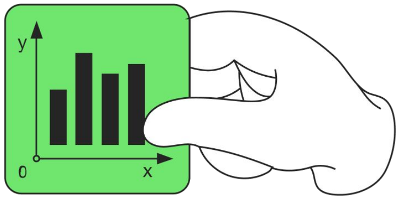
💥В компании VK прошёл путь от младшего аналитика до руководителя команды.

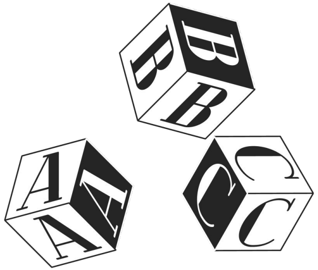
💥Работал в Ренессанс Страховании, НИУ ВШЭ;

💥Сфера научных интересов: обработка текстов (NLP) на низкоресурсных языках.

Ответьте на несколько вопросов сообщением в чат







Какой у вас опыт изучения математических дисциплин?

Какой раздел математики кажется вам самым сложным?

Какие темы по теории вероятностей и статистике вы когда-либо изучали?

Что будет на уроке сегодня

- 📌 Основные понятия теории вероятностей
- 📌 Распределения случайных величин
- 📌 Основные понятия математической статистики
- 📌 Проверка статистических гипотез

Классическое распределение вероятности

Пусть мы хотим найти вероятность некоторого события А. Например, вероятность того, что на игральной кости при броске выпадет значение «6».

Нам необходимо:

- Определить множество возможных исходов (Ω). В нашем примере их шесть – по числу граней кубика.
- Определить множество благоприятных (нужных нам) исходов. У нас он всего один — выпадание шестёрки.
- Разделить число благоприятных исходов на число возможных исходов.

В результате получаем 1/6 в ответе.

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

Задача на закрепление

Предположим, что мы хотим найти вероятность того, что на игральной кости при броске выпадет значение больше четырёх.

Необходимо:

- Определить множество возможных исходов (Ω). В нашем примере их будет **?**
- Определить множество благоприятных (нужных нам) исходов. В нашем примере оно будет **?**
- Поделить число благоприятных исходов на число возможных исходов. В результате получаем **?**

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

Операции с вероятностями

1

Отрицание

Обозначим вероятность того, что на кубике выпадет 6 за P(A).

Тогда противоположным событием будет такое событие B, при котором на кубике выпадет любое число, кроме шести.

Вероятность этого события получаем по формуле:

$$P(B) = 1 - P(A) = 1 - \frac{1}{6} = \frac{5}{6}$$

2

Сложение

Если два события не могут появиться одновременно, их вероятности можно складывать.

Например, событие A (выпало «5») и событие B (выпало «6») не могут произойти одновременно при одном броске. Значит, их вероятности можно складывать.

3

Умножение

Если два события не связаны друг с другом, их вероятности можно перемножать.

Например, событие A (при первом броске выпало «5») и событие B (при втором броске выпало «5») независимы, Значит, их вероятности можно перемножать.

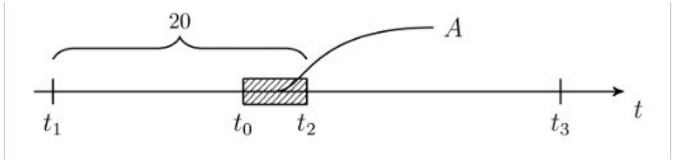
Классическое определение вероятности в ином ракурсе

Предположим, что мы пришли на автобусную остановку. Мы знаем, что наш автобус ходит раз в 20 минут.

Задача: найти вероятность того, что, придя на остановку, мы прождём не более пяти минут.

Что надо сделать:

- Определить множество возможных исходов (Ω). Максимально можем ждать автобус 20 минут, если пришли сразу после того, как уехал предыдущий.
- Определить множество благоприятных (нужных нам) исходов. Мы хотим ждать не более 5 минут.
- Поделить число благоприятных исходов на число возможных исходов.
  В ответе получим: $\frac{5}{20} = \frac{1}{4}$



$$P(A) = \frac{t(A)}{t(\Omega)}$$

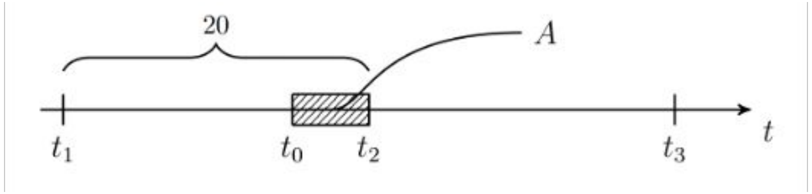
Классическое определение вероятности в ином ракурсе

Предположим, что мы пришли на автобусную остановку. Мы знаем, что наш автобус ходит раз в 20 минут.

Задача: найти вероятность того, что, придя на остановку, мы прождём не более пяти минут.

Что надо сделать:

- Определить множество возможных исходов (Ω). Максимально можем ждать автобус 20 минут, если пришли сразу после того, как уехал предыдущий.
- Определить множество благоприятных (нужных нам) исходов. Мы хотим ждать не более 5 минут.
- Поделить число благоприятных исходов на число возможных исходов.
  В ответе получим: $\frac{5}{20} = \frac{1}{4}$



$$P(A) = \frac{t(A)}{t(\Omega)}$$

Классическое определение вероятности в ином ракурсе

1

Размещение

Число упорядоченных комбинаций подмножества k объектов из множества n объектов.

Без повторений: $A_n^k = \frac{n!}{(n-k)!}$

С повторениями: $rep.\ A_n^k = n^k$

2

Перестановка

Сколькими комбинациями мы можем отсортировать наши n объектов.

Формула: $P_n = A_n^n = n!$

Факториал: $n! = 1 * 2 * \cdots * n$
$0! = 1$

3

Сочетание

Размещение при игнорировании порядка. Для этой величины наборы (A, B) и (B, A) идентичны.

Без повторений: $C_n^k = \frac{n!}{k!\,(n-k)!}$

С повторениями:

$$rep.\,C_n^k = C_{n+k-1}^k = \frac{(n+k-1)!}{k!\,(n-1)!}$$

Примеры на комбинаторику

Предположим, что мы хотим купить фрукты на рынке. Всего существует 3 типа фруктов: яблоко (#1), персик (#2) и груша (#3). Но мы планируем купить только 2 фрукта (НЕ типа, а две фруктовые единицы).

- Если мы хотим разнообразить наш рацион, то фрукты не должны повторяться (первая строка). В противном случае у нас в корзине могут оказаться 2 фрукта одного типа (вторая строка).
- Дальше надо решать, важен ли нам порядок съедания фруктов. Если да, нас интересует первый столбец, иначе — второй столбец.
- Если у нас уже есть 2 разных фрукта и мы выбираем порядок их съедания, число вариантов будет равно 2!=2.

$$A_n^k = \frac{n!}{(n-k)!} = \frac{3!}{(3-2)!} = 6$$

$$C_n^k = \frac{n!}{k!\,(n-k)!} = \frac{3!}{2!} = 3$$

	Упорядоченное	Неупорядоченное
Без повторений	(1 2) (1 3) (2 1) (2 3) (3 1) (3 2)	(1 2) (1 3) (2 3)
С повторениями	(1 1) (1 2) (1 3) (2 1) (2 2) (2 3) (3 1) (3 2) (3 3)	(1 1) (1 2) (1 3) (2 2) (2 3) (3 3)

$$rep.\ A_n^k = n^k = 3^2 = 9$$

$$rep.\ C_n^k = C_{n+k-1}^k = \frac{(n+k-1)!}{k!\,(n-1)!} = \frac{4!}{2! * 2!} = 6$$

Условная вероятность и формула Байеса

Смысл условной вероятности – оценивать вероятность некоторого события при условии, что осуществилось иное событие, предшествующее финальному.

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

Важные аспекты:

- A&B обозначает составное событие, при котором выполнились и A, и B. Причём, поскольку B наступает раньше A, вероятность составного события определяется как произведение условной вероятности «A при B» на вероятность B.

- Формула Байеса позволяет оценивать условную вероятность через срабатывание «B при A».

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

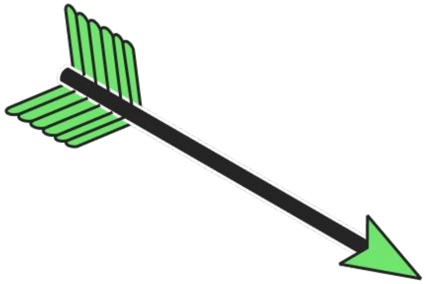
Задача на формула Байеса

Из тридцати стрелков попадают в цель:

- двенадцать — с вероятностью 0,6
- восемь — с вероятностью 0,5
- десять — с вероятностью 0,7

Случайно выбранный стрелок произвёл выстрел и поразил цель. К какой группе вероятнее всего принадлежал этот стрелок?

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

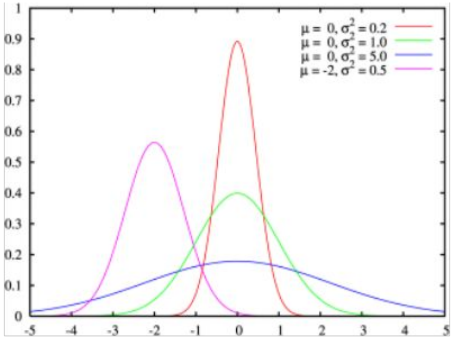


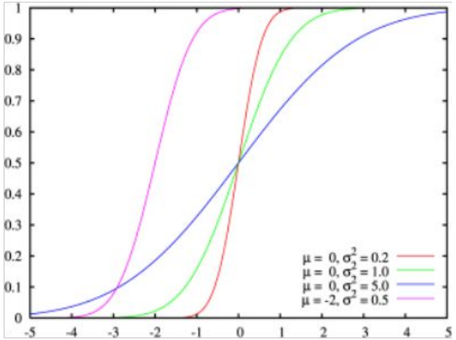
Случайная величина и её распределение

Случайная величина – это функция, значения которой представляют собой исходы случайного эксперимента.

- Распределение — это закон, описывающий область значения случайной величины. Определяется функцией плотности вероятности.
- Функция распределения — это вероятность того, что случайная величина примет значение, меньшее некоторого x.

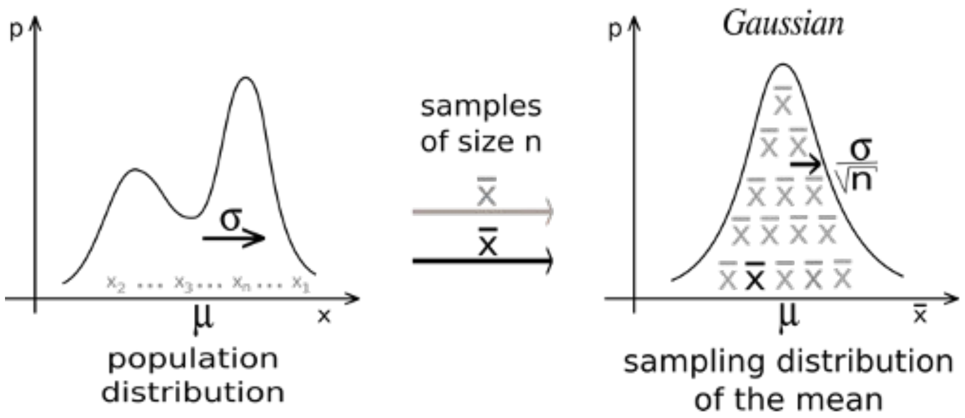
На рисунках представлены функции для нормального распределения.





Центральная предельная теорема

Сумма независимых и одинаково распределённых случайных величин с конечным математическим ожиданием имеет распределение, близкое к нормальному.



Теория вероятностей vs Математическая статистика

Генеральная совокупность — это совокупность всех объектов, имеющих общие характеристики.

- Математическое ожидание;
- Дисперсия.

$$D[X] = \mathrm{M}[\mathrm{X}^2] - (\mathrm{M}[\mathrm{X}])^2$$

Выборка — наблюдаемое подмножество объектов генеральной совокупности.

- Среднее арифметическое значений выборки;
- Среднее арифметическое квадратов отклонений от среднего.

$$S = \frac{1}{n}\sum\left(X_i - \overline{X}\right)^2$$

Статистические гипотезы

Гипотеза о свойствах случайной величины и о виде распределения, проверяемая путём применения статистических методов к данным выборки.

- Нулевая гипотеза — основное утверждение, которое необходимо проверить
- Альтернативная гипотеза— иное утверждение о параметрах

Понятия

1

Уровень значимости

Значение, задаваемое перед проверкой гипотезы и определяющее максимально допустимую ошибку.

2

Ошибка первого рода

Отказ от нулевой гипотезы при условии, что она верна.

3

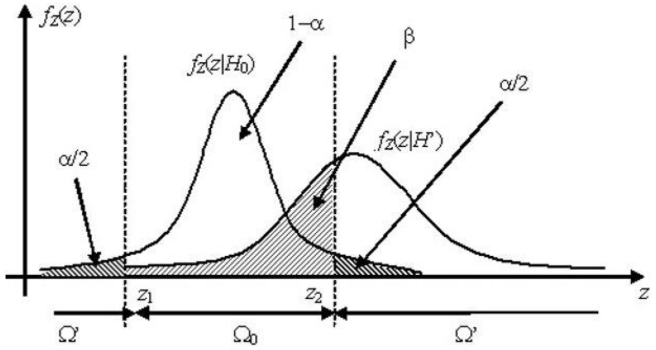
Ошибка второго рода

Принятие нулевой гипотезы несмотря на то, что она неверна.

Гипотезы о виде распределения

Позволяют проверить, принадлежит ли рассматриваемая выборка к некоторому закону распределения.

- Фактически происходит сравнение эмпирической и теоретической функций распределения.
- Критерий Колмогорова-Смирнова универсально работает для любого распределения.
- Критерии Стьюдента, Манна-Уитни и Пирсона направлены на оценку похожести (с точки зрения распределений и их параметров) двух выборок независимо от их распределения.



Корреляция

Показатель взаимосвязи между случайными величинами.

- Принимает значения от -1 до 1;
- 1 — полная взаимосвязь,
  0 — отсутствие взаимосвязи,
  -1 — обратная зависимость;
- Позволяет определить, является ли направление изменений и их степень однонаправленной и насколько сильно.

$$cor_{xy} = \frac{cov(x)}{\sigma_x \sigma_y} = \frac{M(xy) - M(x)M(y)}{\sigma_x \sigma_y} =$$

$$= \frac{\sum (x - \overline{x})(y - \overline{y})}{\sqrt{\sum (x - \overline{x})^2 (y - \overline{y})^2}}$$

Вопросы?

Вопросы?

Вопросы?