Лекция 1

**Расчет вероятности случайных событий**

Здравствуйте. Меня зовут Елена Неизвестных. Вместе мы будем изучать курс «Теория вероятностей и математическая статистика».

Чтобы понять, насколько важен данный курс, достаточно лишь посмотреть, как часто мы сталкиваемся с вероятностями в жизни.

Например, проснулись мы утром, глянули за окно, погода изменилась. Тогда, предположим, мы заходим в поисковик, и начинаем вводить слово «погода», чтобы узнать прогноз на день.

Но только мы ввели первую букву «П», как система автозаполнения уже предлагает нам в выплывающем списке несколько вариантов ввода. И на самом первом месте, вверху списка, видим слово «погода». Т.е. можем предположить, что есть какой-то алгоритм, который учитывает все слова на букву «П», введенные за какой-то период. Т.е. слова «погода» в этой массе различных слов на букву «П» встречается чаще всего. Или иными словами, это слово наиболее вероятное. Т.е. имеет самую большую долю из всех слов на букву «П». Т.о. вероятность- это не что иное, как доля. И мы с каждой лекцией только будем в этом убеждаться снова и снова.

Или другой жизненный пример, пациент пришел в лабораторию сдать анализ. Например, на госпитализацию всегда требуют анализ на ВИЧ. Это серьезное заболевание, следовательно, точность данного анализа должна быть очень высокой. И здесь мы сталкиваемся с такими понятиями, как чувствительность и специфичность. Это тоже вероятности – условные вероятности. Например, вероятность, что тест покажет «отрицательно» при том, что болезни и, правда, нет. Эта вероятность называется специфичностью. Или вероятность, что тест покажет положительный результат, при том, что болезнь есть. Это чувствительность.

Чувствительность > 99% P(+| болен)

Специфичность > 99% P(- |здоров)

Когда мы строим, какие-то математические модели, которые прогнозируют результат, в основу ложится понятие вероятности. Мы постоянно взаимодействуем с вероятностью. Например, логистическая регрессия, которую мы будем проходить в дальнейшем, в тестировании гипотезы мы будем сталкиваться с термином p-value, alpha , мощность теста. Все это тоже вероятности определенных событий.

Для понимания законов распределений мы тоже обращаемся к вероятности, мы строи 95% интервалы. И этот процент тоже показатель определенной вероятности. Об этом все будем говорить позднее.

Помимо того, что нам нужно построить модель, нам нужно ее оценить. И здесь опять мы обращаемся к условной вероятности.

Потому что даже самый эффективный алгоритм может допустить ошибку, поскольку в реальной жизни очень часто бывает, что разные результирующие (например, разные виды лягушек), будут иметь схожие признаки (вес, длина тела), по которым мы будем предсказывать вид лягушки.

Теперь, замотивировав Вас изучать этот курс, можем переходить к основному понятию.

**Виды случайных событий**

Начнем с термина «случайное событие». **Это событие, которое при определенных условиях может произойти либо не произойти.**

Примером случайного события может быть бросок двух игральных костей, где на одной выпала тройка, а на другой единица.

Или клиент банка не вернул кредит. С точки зрения банка это тоже случайное событие.

Мы будем рассматривать случайные события, скажем так, по разным шкалам. Возможное и невозможное, зависимое и независимое, совместное и несовместное.

Итак, **достоверным** событием является такое событие, когда в результате испытания оно обязательно произойдет.

Например, при подбрасывании монетки, выпадет орел или решка. Это достоверное событие. Или при стократном подбрасывании монеты решка выпала не более 100 раз.

А **невозможным** мы называем событие, которое никогда не произойдет. Например, при броске двух игральных костей сумма выпавших чисел составила 1. Такого быть не может. Потому что минимальная сумма будет 2. Или другой пример, при стократном подбрасывании орел выпал 56 раз, а решка 44 раза. Это также невозможное событие, т.к. 56 и 44 в сумме дают 110 раз.

**События бывают зависимые и независимые.**

**Независимые события -** появление одного не влияет на появление другого. Примером могут служить две игральные кости. Если на первой выпала 1, то это событие никак не исключает, что и на другой выпадет 1 или какое-то другое число.

Вероятность **одновременного** появления двух независимых событий:

**Зависимые события** - это когда появление одного влияет на появление другого.

Вероятность наступления двух зависимых событий:

Выражение P(B|A) означает вероятность наступления события при том, что событие уже наступило.

Например, у нас есть корзина с белыми и черными мячами. Вытащить сначала белый мяч, а потом черный - это два зависимых события. Т.е. второе наступает при условии наступления первого события.

Вернемся к этим формулировкам чуть позже, когда изучим понятие вероятности.

События также разделяются на совместные и несовместные.

**Совместные события** – это события, когда одно не исключает появление другого. Например, кидаем две 2 игральные кости. На одной выпадает 1, а на другой выпадает 2-ка. Это независимые **совместные** события.

А вот **несовместные события** – это события, где появление одного из них исключает появления другого. Пример: бросаем одну игральную кость. И выпадение, например, единицы будет исключать выпадение какой – либо другой грани. Выпадение 1 и выпадение, например, 2-ки – это два несовместных события, при условии , что бросаем одну кость.

**Вероятность**

Для случайного события часто приходится рассчитывать вероятность его наступления. Сейчас мы познакомимся с 3-мя понятиями. Относительная частота, статистическая вероятность и в заключении дадим классическое определение вероятности.

Относительная частота. Рассчитывается как отношение числа наступления событий к общему числу испытаний m/n. Это экспериментальная характеристика. Т.е. то, что мы рассчитали в ходе эксперимента. Для нее свойственно небольшое число испытаний.

Пример: Подбрасываем игральную кость 60 раз. Нас интересует, с какой вероятностью мы увидим 3. Для этого нам нужно посчитать, сколько раз из 60 выпала 3, а потом это количество разделить на 60



И относительная частота получится 0.1 в данном примере.

Статистическая вероятность. Это также экспериментальная характеристика. Но для нее свойственно большое число испытаний. Если мы многократно увеличим число испытаний, то результирующее значение практически не будет изменяться. Оно будет стремиться к какой-то постоянной величине.

А теперь дадим классическое определение вероятности.

Представим, что у нас есть игральная кость. Нас интересует вероятность выпадение 3-ки. Кость имеет форму кубика, 6 граней. Каждая грань – это равновозможный элементарный исход. Тройка нарисована только на одной грани. Соответственно, эта единственная грань будет благоприятствовать нашему событию (выпадение тройки). И эту грань будем называть благоприятствующим исходом. Если бы кость была нечестной, а предположим, на 2 гранях имелась бы тройка, тогда благоприятствовали бы нашему событию 2 грани.

Итак, классическое определение вероятности гласит: вероятность события — это отношение числа элементарных исходов, благоприятствующих данному событию, к числу всех равновозможных исходов опыта, в котором оно может появиться. Т.е. для нашего примера с честной костью – это 1/6.

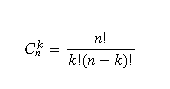
Классическим определением вероятности мы можем воспользоваться, если нам заранее известны все равновозможные исходы. Но в жизни так бывает не всегда. Например, вероятность попадания спортсменом в мишень. Для ее расчета необходимо разделить число удачных выстрелов на общее их количество. В данном случае мы воспользуемся статистической вероятностью, когда будет большое число выстрелов за длительный период.

**Комбинаторика.**

На собеседованиях задачи на комбинаторику встречаются достаточно часто, т.к. умение посчитать число различных комбинаций распространённая задача в сфере IT. Также комбинаторику будем использовать для расчета вероятностей. Рассмотрим сочетания, размещения, перестановки.

**Сочетания** – это набор, состоящий из k элементов, выбранных из множества, содержащего n различных элементов. В сочетаниях неважен порядок. Например, когда вы играете в карты, у вашего соперника 6 карт. Вас не интересует, в какой последовательности он их держит, а вас интересует только их содержание: какие масти, есть ли тузы и т.д.

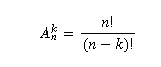
Формула для расчета числа сочетаний



**Размещения** – это набор k элементов, выбранных из множества n, которые отличаются либо самими элементами, либо порядком их расположения. Здесь участвуют не все элементы и важен порядок. Пример: у нас есть 4 спортсмена и нам их нужно расставить на пьедестал, т.е. есть 1-е, 2-е и 3-е место. Для спортсмена важно, какое место он займет. Мы можем из 4х взять первых 3х спортсменов. И расставлять их различными способами, меняя местами.

123, 132, 213, 231, 312, 321

А можем взять другую тройку спортсменов 124 и начать их также по-всякому расставлять. Вот все возможные комбинации мы можем посчитать через формулу для расчета размещений.

,

где n =4, а k=3

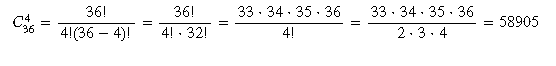
**Перестановки.** Перестановки - комбинации из n элементов, отличающиеся их порядком. Т.о. здесь важен порядок и участвуют все элементы. Например, сколькими способами мы можем расставить трех спортсменов на пьедестал. Воспользуемся формулой для расчета числа перестановок n! Т.о. 3! = 3\*2\*1 = 6

Разберем 3 задачи

1. Сколькими способами можно выбрать из колоды, состоящей из 36 карт, 4 карты?

Здесь имеется в виду, что важно только содержимое этих 4 карт. Порядок неважен нам был в сочетаниях, следовательно, это сочетания. Сочетания из 36 по 4 дадут нам 58905 сочетаний.





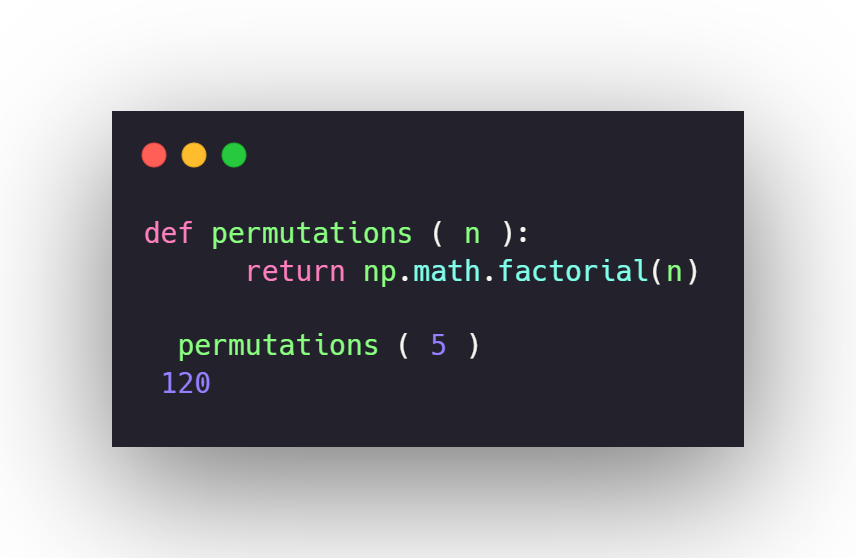
1. В магазине 20 покупателей. Сколькими способами они могут образовать очередь из 5 человек?

В этой задаче сначала определим, важен ли порядок. Очередь занимаю неспроста, именно потому, что порядок в ней важен. Но здесь участвуют не все элементы. Следовательно, это размещения.

Посчитаем в Python

1. Сколькими способами 5 покупателей могут образовать очередь?

Здесь также порядок важен, но участвуют все элементы.



Или воспользуемся формулой n!, получаем 5! = 5\*\*4\*3\*2\*1= 120

Теперь рассмотрим более сложную задачу:

Из колоды, состоящей из 36 карт, случайным образом выбраны 5. Сколько можем получить комбинаций, содержащих 2 туза без учета порядка.

Т.е. мы должны посчитать все комбинации, где будет 2 туза и 3 нетуза.

Алгоритм решения следующий. Поскольку в вопросе задачи есть разбиение на тузы и нетузы, то мы разобьем все 36 карт на 2 множества. Множество тузов (состоит из 4 карт) и нетузов (оставшиеся 32 карты).

В конечной комбинации должно быть 2 туза среди 5 карт.

Найдем все возможные сочетание из 4х по 2. Получим 6 возможных сочетаний из 2 тузов.

Теперь возьмем множество нетузов. В конечной комбинации 3 нетуза из 5. Находим все возможные сочетания из 3 нетузов. Для этого берем подмножество нетузов и находим число сочетаний из 32 по 3, получаем 4960.

А теперь представим, что возьмем 1 из 4960 сочетаний из 3 нетузов. Пусть это будет определенные три семерки. Чтобы получить нам комбинацию из 5 карт, где будет 3 нетуза и 2 туза, мы доложим к этим трем семеркам сначала 1-е возможное сочетание из 2 тузов, потом 2-е сочетание из 2 тузов. Помним, что всего сочетаний из 2х тузов 6.

Т.о. вместо трех семерок у нас появляется 6 сочетаний из трех семерок и 2 тузов. Их стало теперь в 6 раз больше. И так сделаем с каждой тройкой нетузов из 4960 комбинаций. И их количество увеличится в 6 раз. 4960\*6= 29760 возможных сочетаний из 2 тузов и 3 нетузов. Это и будет ответ на поставленную задачу.

*А теперь зададим вопрос, а какова вероятность, что из 36 карт мы возьмем 5 карт и среди них окажется 2 туза.*

Когда мы сталкиваемся с вероятностью, помним, что вероятность – это всегда доля! Доля благоприятствующих исходов из всех возможных исходов. Данному событию благоприятствуют все сочетания из 5 карт, где есть ровно 2 туза. Мы рассчитали их число в предыдущей задаче, 29760. А общее число исходов из 5 карт посчитаем, как число сочетаний из 36 по 5. Получим 376992. Теперь находим долю благоприятствующих исходов, это и будет искомая вероятность.

Р= 29760 / 376992 = 0,0789

**Произведение и сложение событий.**

Есть такое понятие, как произведение событий, т.е. их одновременное наступление. Когда нужно найти вероятность, что два события наступят одновременно, необходимо перемножить эти вероятности.

Например, для двух совместных событий (одновременное выпадение 1 и 2 на двух разных игральных костях при броске) – эта вероятность равна 1/6 \* 1/6

А для 2х несовместных событий (выпадение 1 и 2 на одной и той же кости) будет равна 1/6 \* 0 = 0. Вероятность наступление второго события равна 0, т.к. первое исключает наступление второго события.

**Сложение событий**. Вероятности складываем, когда строго или события A наступает, или событие В, но никак не одновременно.

Возьмем два **несовместных** события. Нас интересует событие выпадение 1 или 2 на одной игральной кости. Одновременно эти события произойти не могут. Но нас устроит вариант 1 и вариант 2. Т.е. нам благоприятствуют 2 грани (два элементарных равновозможных исхода), а вероятность, как мы помним – это отношение числа благоприятствующих исходов к общему числу исходов. Т.е. 2/6 = 1/3.

Или мы могли 1/6+1/6, потому что ожидается или событие А, или событие В.

А для двух **совместных** событий возможно одновременное их наступление. Но нас интересует или только событие А, или только событие В, поэтому, чтобы посчитать вероятность наступление одного из двух совместных событий, мы должны сложить их вероятности и вычесть произведение вероятностей наступления этих двух событий.

Сделаем таблицу для наших двух вышеприведенных примеров

|  |  |
| --- | --- |
| Несовместные события | Совместные события |
| Произведение событий | |
|  |  |
| Сложение событий | |
|  |  |

**Для простоты запоминания «или» – это «+», «и» - это «\*»**

**Условная вероятность**

В теории вероятностей есть также понятие, как **условная вероятность**. Т.е. рассчитанная для события, которое произойдет при условии, что другое событие уже наступило. Мы такие упоминали это событие, когда разделяли события по шкале зависимые и независимые.

Давайте рассмотрим поиск условной вероятности на примере следующей задачи:

Есть три корзины. В 1й 3 красных и 5 зеленых мячей, во второй все красные, в 3-й корзине все зеленые мячи. Случайным образом выбирается корзина и случайным образом из нее берут мяч

Какова вероятность, что мяч окажется зеленым?



Событием А будем обозначать – вытащить зеленый мяч, а событием будем обозначать событие - вытащить ту или иную корзину.

Что вообще может произойти такого, чтобы в итоге мы вытащили зеленый мяч?

Мы можем взять 1-ю корзину **И** (знак \*) вытащить из нее зеленый мяч , **ИЛИ** (знак +) взять вторую корзину и вытащить из нее зеленый мяч, или взять третью корзину и вытащить из нее зеленый мяч.

Или иными словами вероятность взять первую корзин \* вероятность взять зеленый мяч, при условии, что взяли 1 корзину + (или) вероятность взять вторую корзину \* вероятность взять зеленый мяч, при условии, что вытащили эту вторую корзину + вероятность вытащить 3ю корзину \* вероятность вытащить зеленый мяч, при условии, что взяли 3ю корзину.

Запишем это на языке математики, помня, что \* - это И, «+» - это ИЛИ, а обозначение в условной вероятности читается, как вероятность наступления события А, при условии, что событие В уже наступило. Или для нашей задачи - это вероятность вытащить зеленый мяч, при условии, что взяли определенную корзину.

P(A) = P(

Вероятность P( = P( = P( = 1/3

Вероятность = 5/8 (т.к. для 1 корзины благоприятствует 5 зеленых мячей из 8),

– во второй корзине нет зеленых мячей,

, т.к. в 3-ей корзине все зеленые мячи.

Подытожим: если событие А может наступить только при наступлении событий , , …, , образующих полную группу событий \*,

то вероятность события А вычисляется по формуле:

P(A) = P(

\* полная группа событий означает, что при отдельном испытании обязательно произойдет одно из них.

**Формула Байеса**

**Чтобы определить вероятность события B при условии, что событие A уже произошло, используют формулу Байеса.**

Здесь P(B) – априорная вероятность (определяется до испытания)

P(B|A) - апостериорная вероятность

Для осознания формулы, давайте немного изменим предыдущую задачу:

У нас есть все те же три корзины. Случайным образом выбирается корзина и случайным образом из нее берут зеленый мяч. Какова вероятность, что зеленый мяч окажется из 3-ей корзины?



Здесь нас интересует найти вероятность наступления некого события (в данном случае вероятность вытащить 3ю корзину), при условии, что мяч мы уже вытащили зеленый.

Когда нам нужно найти вероятность события, при том, что какое-то событие уже наступило, начинаем решение задачи с записи формулы Байеса. Запишем конкретно для третьей корзины, добавив индекс тройку к событию В.

Помним, что А – вытащить зеленый мяч, В - вытащить определенную корзину.

Вероятность вытащить зеленый мяч, при условии, что взяли третью корзину = 1 (все мячи зеленые в ней), вероятность вытащить 3ю корзину = 1/3 (с одинаковой вероятностью мы можем взять 1 из трех корзин), ну и полная вероятность события «вытащить зеленый мяч» была рассчитана в предыдущей задаче 13/24.

ДОПОЛНИТЕЛЬНАЯ ЗАДАЧА НА ФОРМУЛУ БАЙЕСА

Пациент приходит к врачу, чтобы сделать тест на конкретное заболевание.

Вероятность положительного теста при том, что болезнь есть 90%.

Вероятность отрицательного теста, при том, что человек здоров 95 %.

Вероятность заболевания 2%.

*Используя теорему Байеса, найти вероятность, что человек болен, при условии, что тест положительный?*

Почему надо использовать формулу Байеса? Напоминаю, когда нам нужно найти вероятность события, при том, что какое- то событие уже произошло, мы используем формулу Байеса.

И в этой задаче нам нужно найти вероятность, что человек болен , при условии, что событие «тест положительный» уже наступило.

Выпишем формулу, и разберемся в обозначениях

– это и есть наш вопрос задачи. Вероятность, что человек болен (получается, что это событие В), при том, что тест уже положительный (это событие А). Разобрались, что есть событие А, а что есть событие В.

Пойдем дальше:

– это вероятность положительного теста (событие А), при том, что человек болен (событие В). Теперь надо посмотреть, а какая вероятность (помним, вероятность - это доля) положительного теста среди больных пациентов. Вернемся к условию задачи:

«Вероятность положительного теста при том, что болезнь есть 90%»

вероятность события В, т.е вероятность вообще, что человек является носителем этой болезни, иными словами, доля больных в популяции. Из условия задачи мы берем эту цифру 2% или 0,02

Осталось определиться со знаменателем.

– это вероятность положительного теста. А тест может быть положительным как среди больных пациентов, так и среди здоровых. Т.о. сделав тест и получив положительный результат, мы учтем следующее: вероятность, что тест положительный \* вероятность ,что брали больного пациента или тест положительный \* вероятность , что взяли для теста здорового пациента.

Из условия задачи: «Вероятность положительного теста при том, что болезнь есть 90%» или 0,9 в долях. Вероятность болезни из условия задачи 0.02 (эта вероятность есть доля больных в популяции)

Первая часть выглядит так :

Чтобы рассчитать вторую часть , нам нужно посчитать вероятность положительного теста, при условии, что он проводился на здоровых людях. Из условия задачи знаем «Вероятность отрицательного теста, при том, что человек здоров 95 %» Изобразим это:

Значит, на долю положительных тестов среди здоровых приходится 0,05 (посчитали как 1-0.95).

А вероятность, что человек здоров – это доля здоровых во все популяции 1- 0.02= 0.98.

Теперь мы знаем все, чтобы ответь на вопрос задачи, какова вероятность, что человек болен, при условии, что тест положительный.

В итоге ответ: вероятность, что пациент болен, при условии, что тест положительный равна 26,9%

**Совет от преподавателя.**

1. Начинайте решение задач с того, что выписываете формулу, по которой будете решать задачу. Это будет как план решения задачи. Затем смотрите, как посчитать те или другие неизвестные.
2. Помним, что вероятность это доля
3. Помним, что для расчета вероятности необходимо разделить число благоприятствующих исходов на общее число исходов.
4. Если мы находим вероятность через комбинаторику, то будьте внимательны, с тем, какие комбинации вы рассчитываете. Если у вас в числителе не учитывается порядок, т.е. вы решаете через сочетания, то и в знаменателе тоже порядок не учитывается. Т.е. находится общее число сочетаний. Нельзя, например, находить долю сочетаний из всех возможных размещений.