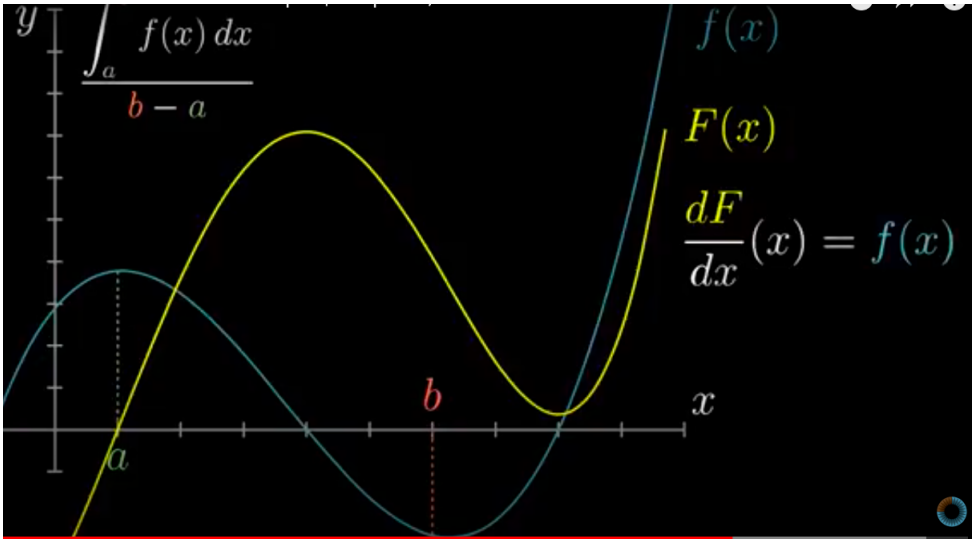
1. Первообразная, интеграл.

Рассмотрим обратную задачу: у нас есть функция $f(x)$ (график или формула), которая является производной какой-то функции $F(x)$. Вопрос: как найти эту функцию $F(x)$?

Из определения производной можем найти только разность двух состояний:

$$\lim_{x \to x_0} \frac{F(x) - F(x_0)}{x - x_0} = f$$

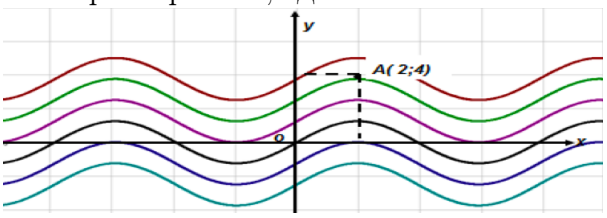


Попробуем интерпретировать задачу в терминах физики: получается, у нас есть скорость $V(t)$, которая является изменением координаты точки на заданном участке пути. Но как всегда есть нюанс: мы знаем только скорость, но эта скорость может быть на любом участке пути, с одной и той же скоростью мы можем ехать по трассе Москва-Санкт-Петербург или Воронеж-Сочи. То есть сама координата нам не даст никакой информации, так как мы не знаем начало системы отсчёта (напоминает принцип Неопределённости Гейзенберга). А вот разность двух состояний/координат покажет, путь какой длины мы проехали.

Ещё одна аналогия из физики: потенциал сам по себе не несёт существенной информации - это работа по перемещению из бесконечно удалённой точки в точку A, а работу по перемещению заряда из точки A в точку B как раз представляет разность потенциалов $\phi(B) - \phi(A)$.

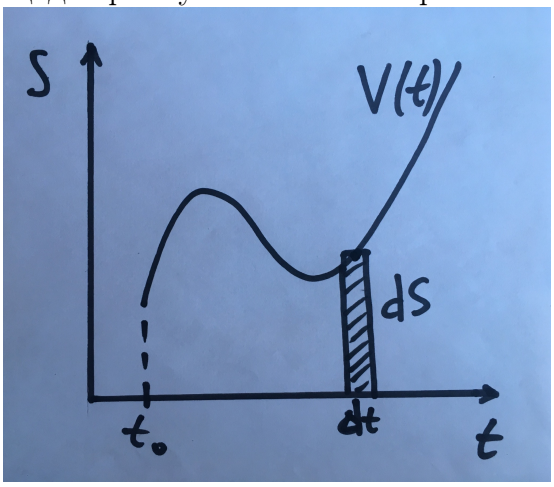
Итого, первообразная $F(x)$ - просто математически введённая функция, которая «подгоняется» как обратное действие к операции дифференцирования при помощи

таблицы производных. Общей формулы для нахождения первообразных, аналогичной формуле для вычисления производной частного и произведения, не существует. Отметим, что такая функция определена с точностью до константы или представляет собой семейство параллельных функций: если $F(x)$ - первообразная, то и $F(x)+c$ - тоже первообразная, где $c = const$.



Строгое определение: первообразной для функции $f(x)$ называется такая функция $F(x)$, определённая на $(a;b)$, что $F'(x) = f(x)$ для любого $x \in (a;b)$.

Вернёмся к задаче о нахождении пройденного пути через известную скорость. В физических терминах $\frac{dS}{dt} = V$, $Vdt = dS$, а вот тут мы как раз используем дифференциал - разбиваем весь путь на очень маленькие промежутки по времени, и на каждом из таких промежутков рассматриваем движение как равномерное - с одной и той же скоростью (приближение линейной функцией), на графике это как раз будет площадь прямоугольника со сторонами V и dt.



А весь пройденный путь как раз равен сумме таких кусочков: $S = \int_{t_1}^{t_2} V(t)\,dt$. А процесс суммирования в непрерывном случае назвали интегрированием. Отсюда получаем известный всем факт, что определённый интеграл равен площади под графиком функции. Воспользуемся площадью как функцией: берём одну из первообразных $F(x) = \int_a^x f(t)\,dt$ - зафиксировали начальную точку и движемся от неё вправо, получим функцию от «конца пути» или верхнего предела интегрирования,

которая как раз покажет путь от a до x.

Отметим, что мы описали путь поиска первообразной и такое представление $F(x)$ через интеграл - это пока наше предположение, надо математически доказать, что $F'(x) = f(x)$ для любого x.

Немного отвлечёмся на факт, который называется «первая теорема о среднем», строгое доказательство можно найти в любом учебнике, нас интересует суть и физическая интерпретация:

Первая теорема о среднем: пусть $f(x)$ - непрерывная и ограниченная на $[a; b]$ функция, тогда на этом отрезке найдётся такая точка c, что

$$\int_a^b f(x)\,dx = f(c)\int_a^b dx = f(c)(b-a)$$

.

Пройденный путь можно найти как *среднюю скорость* $f(c)$, умноженную на затраченное время $(b-a)$ - а это просто следует из физического определения средней скорости. Причём существует такой момент времени c, в который величина скорости как раз равнялась средней - весьма понятный интуитивно факт.

Итак, строго математически докажем, что функция $F(x) = \int_a^x f(t)\,dt$ будет первообразной для функции $f(x)$:

$$F'(x_0) = \lim_{x\to x_0}\frac{F(x)-F(x_0)}{x-x_0} = \lim_{x\to x_0}\frac{\int_a^x f(t)\,dt - \int_a^{x_0} f(t)\,dt}{x-x_0} =$$
$$= \lim_{x\to x_0}\frac{\int_{x_0}^x f(t)\,dt}{x-x_0} = \lim_{x\to x_0}\frac{f(c)(x-x_0)}{x-x_0} = \lim_{x\to x_0} f(c) = f(x_0)$$

Последний переход обосновывается тем, что точка c лежит между x_0 и x, и при стремлении $x \to x_0$ все три точки «сжимаются» в одну.

2. Дискретные и непрерывные случайные величины. Функции плотности и распределения.

Начнём с того, что теория вероятности занимается массовыми явлениями. Когда проводится большое количество одинаковых экспериментов, и на основе полученных данных можно делать какие-либо выводы.

Случайная величина ξ - это функция, которая ставит в соответствие какому-либо происходящему событию число. То есть у нас произошло какое-то событие, и мы можем его численно интерпретировать.

Дискретной случайной величиной называется случайная величина, которая в результате испытания принимает отдельные значения с определёнными вероятностями. Проще говоря, дискретные случайные величины — это величины, множество значений которых не более, чем счётно. Число возможных значений дискретной случайной величины может быть конечным и бесконечным. Примеры дискретной случайной величины: выпадение орла или решки при подбрасывании монеты или выпадение определённого числа на игральном кубике.

Непрерывной случайной величиной называют случайную величину, которая в результате испытания принимает все значения из некоторого числового промежутка. Число возможных значений непрерывной случайной величины бесконечно. Пример непрерывной случайной величины: измерение скорости перемещения любого вида транспорта или температуры в течение конкретного интервала времени.

Вероятность можно рассматривать как обобщение логики на рассуждения в условиях неопределенности. Логика дает нам набор формальных правил, позволяющих определить, истинно некоторое высказывание или ложно, в зависимости от предположения об истинности или ложности других высказываний. Теория вероятностей предлагает набор формальных правил для определения правдоподобия высказывания при условии правдоподобия других высказываний.

Чтобы оценить правдоподобие высказывания, делают серию наблюдений, результаты записывают, и дальше встаёт вопрос: как обработать все данные и делать какие-либо выводы? Разумно записать в таблицу все значения случайной величины и сколько раз встречается каждое значение - это частота событий. Но, как мы понимаем,

просто частота не даёт информации: например, «в рулетке зеро выпал 20 раз» - на основании такой информации сложно сделать какой-либо вывод, а вот «в рулетке зеро выпал 20 раз из 21» и «в рулетке зеро выпал 20 раз из 200» - важное дополнение, при помощи которого можно сделать ставку. Поэтому более информативно узнать долю, которую занимает интересующее нас значение из общего количества всех значений случайной величины - это относительная частота (из определения понятно, что она заключена в отрезке [0;1]). Отметим, что относительная частота рассчитывается исключительно ПОСЛЕ опытов на основе фактически полученных данных и совпадает с вероятностью. Из определения следует, что если все события независимы, то сумма относительных частот (вероятностей) равна 1.

Если мы имеем дело с непрерывными случайными величинами, то нам нужно как-то оценить количественно исходы событий, а раньше мы обсуждали, что за количество элементов в множестве отвечает функция меры, поэтому под вероятностью в общем случае логично понимать отношение

$$\text{функция вероятности} = \frac{\text{мера количества удачных для нас исходов события}}{\text{мера общего количества исходов события}}$$

Мерой чаще всего выступает длина отрезка, площадь или объём фигуры.

Теперь разберёмся, как распределена случайная величина: в жизненных задачах редко нужно знать, в какое единственное значение попадает случайная величина, чаще нас интересует интервал или отрезок. Математики придумали следующую конструкцию: зафиксируем левый конец интервала $-\infty$, правый будет переменной x, будем двигать правый конец небольшими шагами вправо и «смотреть», сколько новых значений случайной величины попадает в каждый такой интервал. Очевидно, что их количество будет не уменьшаться, и общее количество точек можно представить как «вес» или «массу» множества значений случайной величины.

Более строго математически: вводим новую случайную величину $(\xi < x)$, её вероятность $P(\xi < x)$ будет новой функцией, которая как раз показывает вероятность попадания значений случайной величины в нужный нам интервал $(\infty; x)$, или показывает, как распределены на числовой оси значения случайной величины ξ (отметим, что сама случайная величина и её значения фиксированы). Итого получаем функцию распределения случайной величины $F(x) = P(\xi < x)$, которая обладает удобными свойствами: неубывающая, $F(-\infty) = 0$; $F(+\infty) = 1$.

Из такой конструкции легко понять, как находить вероятность попадания точки в полуинтервал $[a;b)$, у которого оба конца зафиксированы. Рассмотрим три случайных события $a \leq \xi < b$, $\xi < a$, $\xi < b$, последнее есть сумма первых двух, тогда по теореме о сложении вероятности $P(\xi < a) + P(a \leq \xi < b) = P(\xi < b)$, в терминах функции распределения $F(a) + P(a \leq \xi < b) = F(b)$ или $P(a \leq \xi < b) = F(b) - F(a)$, что весьма напоминает формулу Ньютона-Лейбница и наталкивает на мысль о представлении вероятности через интеграл.

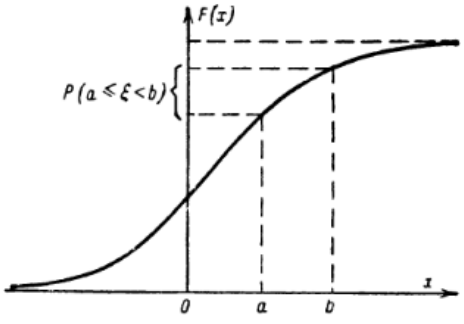
Ещё раз обратим внимание на названия «плотность» и «масса». В физике мы интегрируем плотность для получения массы. Если думать о функции распределения как о массе, то для её получения как раз и нужно проинтегрировать плотность. Осталось дать понятие термину «плотности» в терминах вероятности.

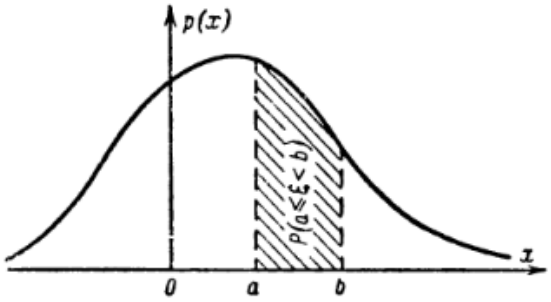
Плотность из физики показывает насколько близко друг к другу расположены точки, свяжем это с изменением функции распределения $F(x)$: если функция распределения растёт быстро, то есть точки «прибывают», они расположены близко - плотность высокая, если растёт медленно, то новых точек мало, а если остаётся неизменной - новых точек нет совсем

Также если функция распределеня дифференцируема, то плотность как раз будет производной функции распределения:

$$\frac{dF}{dx} = p(x), \quad F(x) = \int_{-\infty}^{\infty} xp(z)dz$$

Смысл функции $p(x)$: если событие = «случайная величина ξ поппадет в малый интервал Δx», то вероятность этого события $P(x < \xi < x + \Delta x) \approx F'(x)dx = p(x)\Delta x$. При этом $\int_{-\infty}^{\infty} p(x) = 1$, так как попадание ξ в неограниченный интервал гарантировано.





Итого получили формулы взаимосвязи:

$$F(x) = \int_{\infty}^{x} p(z)dz$$

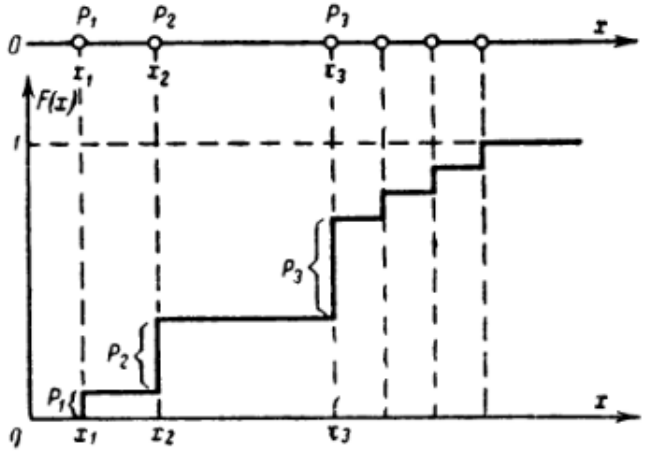
$$p(x) = \frac{dF(x)}{dx}$$

$$P(a \le \xi < b) = \int_a^b p(x)\,dx = F(b) - F(a)$$

$$\int_{-\infty}^{+\infty} p(x)\,dx = \int_{-\infty}^{+\infty} \frac{dF(x)}{dx}\,dx = F(+\infty) - F(-\infty) = 1 - 0 = 1$$

Последнее условие есть условие нормировки. Состояния физической системы всегда однозначны, то есть образуют полную совокупность событий. Условие нормировки для вероятности состояния физической системы отражает факт: если физическая система существует, то она находится в одном из доступных ей состояний.

Заметим, что для дискретной случайной величины понятие плотности вероятности отсутствует, а функция распределения будет ступенчатой. Поэтому с дискретными случайными величинами удобнее работать с набором вероятностей или таблицей, в то время как для непрерывных случайных величин проще анализировать график или уравненение функции плотности и функции распределения.



3. Байесовский подход к вероятности.

Для лучшего понимания рекомендую прочитать статью «Скажи Байесу «да!». Забудь про интуицию — просто думай, как Байес завещал»

В предыдущей главе мы обсуждали случаи с частотной вероятностью - когда у нас есть эксперимент, который мы можем повторить, например, подбрасывание монеты или кубика. А когда, например, говорят про диагноз у конкретного человека, мы не можем «размножить» человека, заразить его разными болезнями и сравнить результаты эксперимента. То есть в данном случае под вероятностью следует понимать степень доверия и восприятия информации: 0 - абсолютно невозможно, 1 - точно истино. Подход, основанный на качественном уровнем уверенности, предложил Байес. Байес по сути говорит, что вновь полученная информация влияет на наше восприятие какого-то события, и вероятность по сути - численное представление личного уровня доверия, который может кардинально измениться вследствие количества наступивших событий. Старое знание + новый опыт = новое, более полное знание.

Рассмотрим принцип Байеса на конкретном примере обработки спама, получаемого по электронной почте. Мы получаем какое-то письмо, в котором содержатся какие-то слова в каком-то количестве. Сначала мы просто подсчитываем разные слова, входящие в это письмо, а потом определяем, является письмо спамом или нет. Проделав это некоторое количество раз, мы соберем базу слов вместе с частотой их появления в спаме и в обычных письмах. В итоге получаем табличку, где записаны слово, количество его упоминаний в спаме и общее количество упоминаний. Теперь введем понятие «веса» слова — вероятность того, что сообщение с таким словом является спамом. Например, такой оценкой может быть частота появлений этого слова в спаме, поделенная на частоту появлений этого слова в любом произвольном письме. Теперь скажем, что «вес» всего письма — это усредненный вес всех слов, которые в нем содержатся. Дальше мы просто говорим, что, например, если этот вес больше 80%, то будем считать это сообщение спамом. Мы получили новое письмо, определили спам это или не спам, и к известным нам данным добавилось новое знание про слова, встретившиеся нам в этом письме, поэтому мы запишем в нашу базу новые показатели и пересчитаем «веса».

Ещё раз подчеркнём, что ключевое отличие состоит в том, что считать случай-

ной величиной. В частотном или фриквентистском подходе мы под такой величиной подразумеваем значение, которое мы не можем спрогнозировать, не проведя какого-то количества экспериментов. В байесовском же подходе случайная величина — это строго определенный процесс, который можно сначала спрогнозировать целиком, просто мы знаем не все начальные факторы, которые могут влиять на исход. Но после «запуска» этого процесса, мы получаем новые знания, которые позволяют «подкрутить настройки» и сделать процесс более эффективным, тем самым повысив наш уровень уверенности в получаемых результатах.

Когда все события независимые, всё просто - вероятности складываются, а если по предположению Байесу одно событие влияет на другое, возникает вопрос, что делать в таком случае?

Колмогоров вводит условную вероятность по определению как $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$.

Далее по Байесу вводятся две вероятности:

Априорная вероятность - предполагаемая вероятность до проведения эксперимента $P(A)$.

Апостериорная вероятность - вероятность, полученная после проведения экспериментов и получения новой информации $P(A \mid B)$.

Теорема Байеса предполагает, что событие B известно ($P(B) \neq 0$), и нужно понять, как знание о событии B влияет на уверенность в том, что произойдёт событие A:

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

. Формула Байеса позволяет «переставить причину и следствие»: по известному факту события вычислить вероятность того, что оно было вызвано данной причиной. Доказательство следует напрямую из определения Колмогорова:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, P(B \mid A) = \frac{P(B \cap A)}{P(A)}, \quad \text{поделим:} \quad \frac{P(A \mid B)}{P(B \mid A)} = \frac{P(A)}{P(B)}$$

Встаёт вопрос: условная вероятность была определена Колмогоровым в XX веке, а Байес вывел свою теорему в XVIII веке? Единственное логичное объяснение, которое удалось найти: Байес больше рассуждал в терминах философии, и пришёл к Теореме о взаимосвязи явлений при помощи словесных логических рассуждений, а Колмогоров вводил строгую аксиоматику и определения таким образом, что Теорема Байеса логично вписалась во введённую им терминологию.

Из определения условной вероятности следует цепное правило для подсчёта совместного распределения вероятности нескольких случайных величин:

$P(x_1, \ldots, x_n) = P(x_1) \prod_{i=2} P(x_i \mid x_1, \ldots, x_{i-1})$, подробнее:

$$P(a, b, c) = P(a \mid b, c)P(b, c) \qquad P(b, c) = P(b \mid c)P(c)$$

$$P(a, b, c) = P(a \mid b, c)P(b \mid c)P(c)$$

Полной противоположностью условному распределению является маргинальное распределение подмножества набора случайных величин — это распределение вероятностей переменных, содержащихся в этом подмножестве. Это даёт возможность представить вероятности различных значений переменных в подмножестве без указания на другие значения переменных. То есть снять зависимость одной величины от всех остальных:

$$P(X = x) = \sum_y P(X = x, Y = y) = \sum_y P(X = x \mid Y = y)P(Y = y)$$