Плотность вероятности

В этом уроке мы определим понятие *плотности* и разберёмся, в чём его смысл. Нужно нам это будет, чтобы уметь находить *матожидание* или *дисперсию* непрерывной случайной величины. Как мы узнаем в следующих уроках, для их нахождения нам нужно будет

- знать функцию плотности случайной величины
- уметь вычислять интегралы

Находить интегралы мы уже умеем, так что теперь будем разбираться с плотностью вероятности :)

Гистограмма

Гистограмма — это инструмент описательной статистики. Сначала мы изучим гистограммы, а через них поймём *плотность вероятности* непрерывной случайной величины.

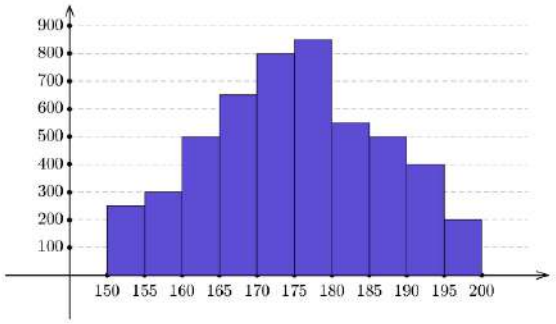
Пример

Вы проводите медицинское исследование. В ходе него у 5000 участников исследования были измерены различные физические показатели, в том числе рост. Важная часть работы с данными — это визуализация. Глядя на таблицу или список из 5000 полученных значений сложно сделать какие-то выводы. Гистограмма — один из простых и наглядных способов представления данных. Вот как она строится:

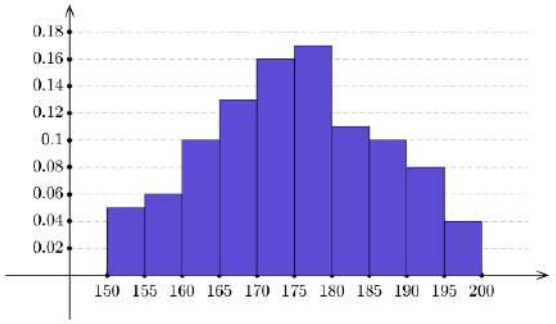
1. Промежуток значений, которое может принимать измеряемая величина, разбивается на несколько интервалов — по-английски их называют *bins*, по-русски — карманы / корзины. Чаще всего эти интервалы берут одинаковыми.

   Допустим, в нашем исследовании рост участников варьируется от 151 до 197 сантиметров. Разделим промежуток от 150 до 200 на десять равных интервалов длиной 5 — это будут наши карманы.
2. Отложим полученные интервалы на горизонтальной оси. Над каждым карманом изобразим прямоугольник с высотой равной количеству участников, чей рост попал в данный карман.

Результат будет выглядеть примерно так:

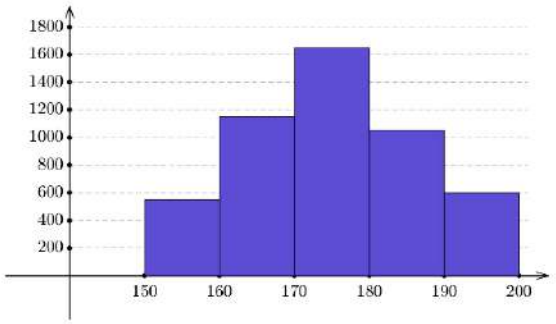


Часто для большей наглядности высоту столбиков берут равной доле, а не количеству значений, попавших в данный карман. Гистограмма будет выглядеть так:



Изменилась только шкала по вертикальной оси: чтобы перейти от количества к долям нужно поделить высоту столбиков на количество элементов выборки — в нашем случае на количество участников исследования, то есть на 5000.

Можно было взять карманы длины не 5, а 10. То есть разбить промежуток от 150 до 200 на равные интервалы длины 10. Тогда гистограмма бы выглядела так:



Выводы

Гистограмма показывает, какие диапазоны значений более частые в выборке, а какие менее. Особенно актуально это, если диапазон значений величины очень большой.

Например, представим себе, что мы хотим построить гистограмму зарплат в Иркутске. Зарплата может быть очень разной и вряд ли нам интересно знать, сколько людей получают ровно 34.350 рублей в месяц. Если нарисовать гистограмму с шагом (шириной карманов) в $2 - 3$ тысячи, то общее представление получить можно.

С ростом числа карманов растет детальность гистограммы, но может падать информативность. Например, если при построении гистограммы зарплат жителей Иркутска в качестве карманов выбрать отрезки вида $[n, n+1]$, где n — натуральное число, информативность такой гистограммы немногим лучше, чем просто таблица значений.

Мы построили гистограмму возрастов 5000 участников нашего исследования. Для этого мы взяли карманы по 3 года. Над каждым карманом мы нарисовали столбец с высотой, равной доле участников, чей возраст попал в этот карман.

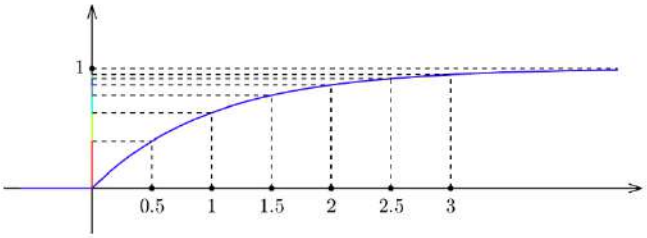
Чему равна сумма высот всех столбцов?

Введите численный ответ

Введите число

Функция распределения и гистограммы

Посмотрим ещё раз на экспоненциальное распределение с показателем $\lambda = 1$. Функция распределения равняется $F_\xi(x) = 1 - e^{-x}$. Разобъём числовую ось на отрезки длиной 0.5 и для каждого такого отрезка посмотрим на вероятность того, что значение ξ лежит на нем.



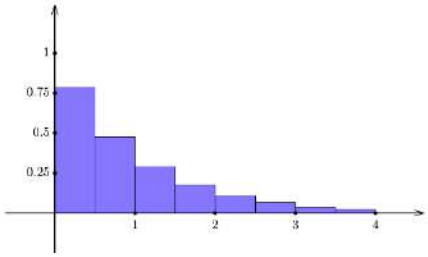
Эта вероятность будет равняться разности значений функции распределения F_ξ на концах отрезков. Например, вероятность того, что ξ попадёт в отрезок от 0.5 до 1 равна $F_\xi(1) - F_\xi(0.5)$. Это число в точности равно длине зелёного отрезка, отложенного на вертикальной оси.

Для всех $x < 0 : P(\xi < x) = 0$, поэтому нам их рассматривать неинтересно. А вот для отрезков на положительной полуоси вероятности будут разные. Вероятность каждого из них отмечена цветом на вертикальной оси. Из картинки, например, видно, что:

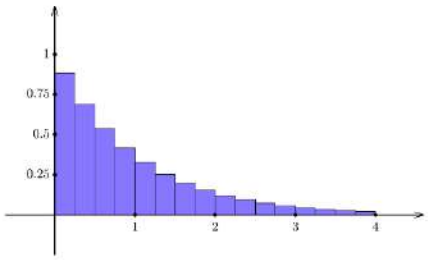
$$P(0 \leqslant \xi \leqslant 0.5) > P(0.5 \leqslant \xi \leqslant 1) > P(1 \leqslant \xi \leqslant 1.5) > P(1.5 \leqslant \xi \leqslant 2) > P(2 \leqslant \xi \leqslant 2.5) > \dots$$

Псевдо-гистограмма

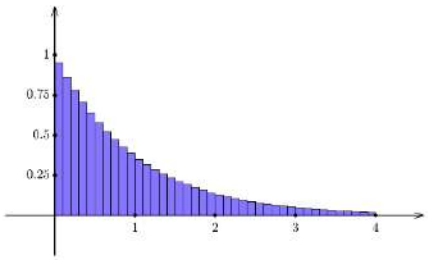
На основе такого разбиения мы можем нарисовать что-то вроде гистограммы — назовём её псевдо-гистограммой. На ней высота столбика — это вероятность попадания в данный интервал, поделённая на длину интервала. То есть высота столбика над интервалом (a, b) равняется $\frac{1}{b-a} \cdot P(a < \xi < b)$. Таким образом, **площадь столбца равняется вероятности того, что случайная величина попадёт на этот интервал.** Действительно, длина горизонтального интервала равна $b - a$, а высота столбца над ним равна $\frac{1}{b-a} \cdot P(a < \xi < b)$. Значит, площадь столбца равна $(b-a) \cdot \frac{1}{b-a} \cdot P(a < \xi < b) = P(a < \xi < b)$.



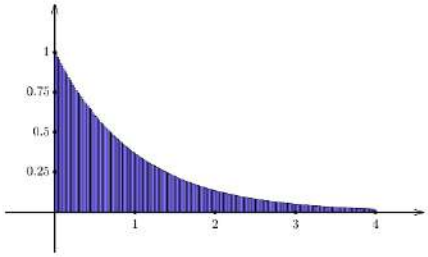
В отличие от гистограммы выборки или гистограммы для дискретных случайных величин, псевдо-гистограмма непрерывной случайной величины при выборе более мелких интервалов становится не только детальнее, но и информативнее.



Ещё детальнее



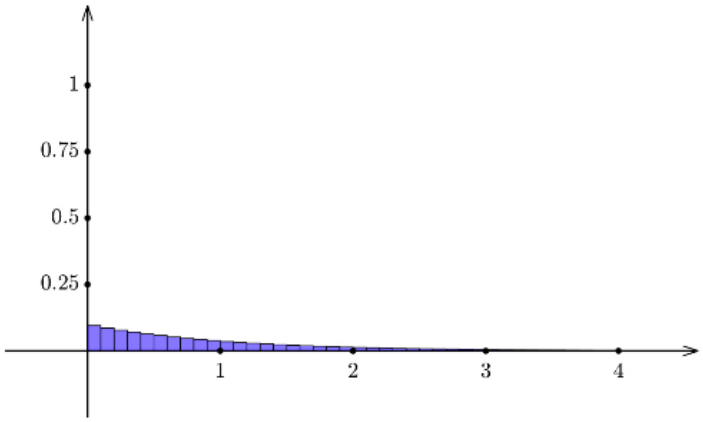
И ещё детальнее



Почему мы выбираем такую высоту столбиков в псевдо-гистограмме

Мы хотим, чтобы высота столбика над интервалом отражала вероятность попадания случайной величины на этот интервал. Если один столбик выше другого, то вероятность попадания в интервал, соответствующий первому, больше, чем вероятность попадания в интервал, соответствующий второму. То есть, как и в случае с гистограммой, наша псевдо-гистограмма отражает частоту попадания в выбранный интервал. Поэтому высота столбиков пропорциональна вероятности попадания в него $P(a < \xi < b)$.

Если бы в качестве высоты столбика над интервалом (a, b) мы выбрали просто $P(a < \xi < b)$, то при стремлении длины интервала к 0 вероятность тоже стремилась бы к 0 просто из свойств непрерывности функции распределения. Например, для интервалов длиной 0.1 выглядела бы так:

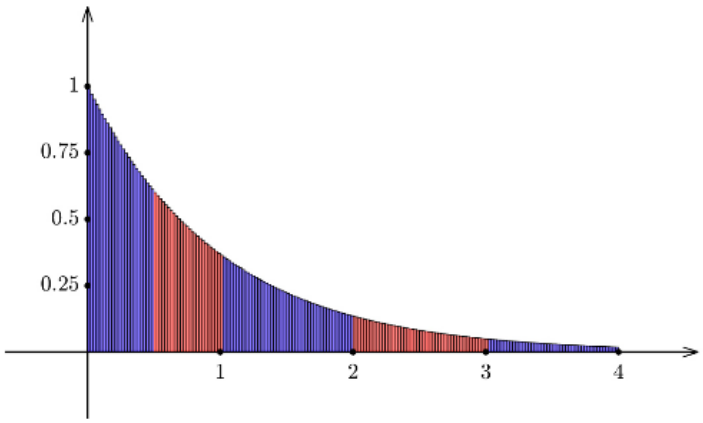


Картинка схлопывается в линию, нам такое не нравится.

Поэтому мы делим вероятность на длину интервала. Так картинка становится детальнее, но не схлопывается в линию. И ещё одно важное свойство:

Псевдо-гистограмма позволяет считать вероятность (приближенно)

Вероятность попадания на некоторый промежуток может быть приближена суммой площадей столбиков над интервалами, которые попадают в данный промежуток. Об этом подробнее поговорим на следующих шагах с теорией, а пока вот картинка:



Задача с проверкой. Плотность распределения 1

Задача. Чему будет равняться сумма площадей всех столбиков псевдо-гистограммы для экспоненциального распределения с параметром $\lambda = 1$?

Проверка. Введите ответ.

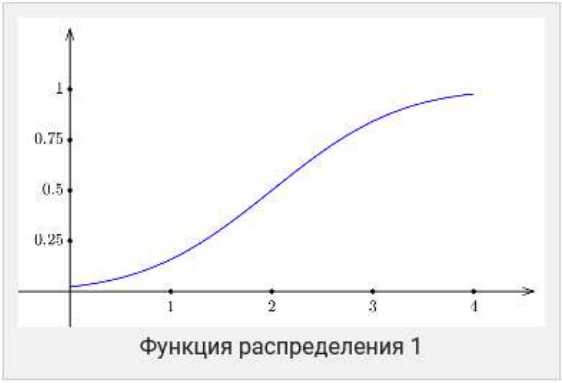
Если ваш ответ не является числом, введите в поле ответа

- введите `depends`, если ответ зависит от длины интервалов в разбиении
- введите `infinity`, если сумма бесконечно большая

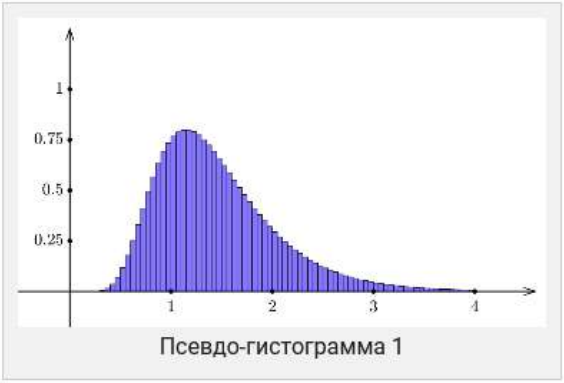
Напишите текст

Напишите ваш ответ здесь...

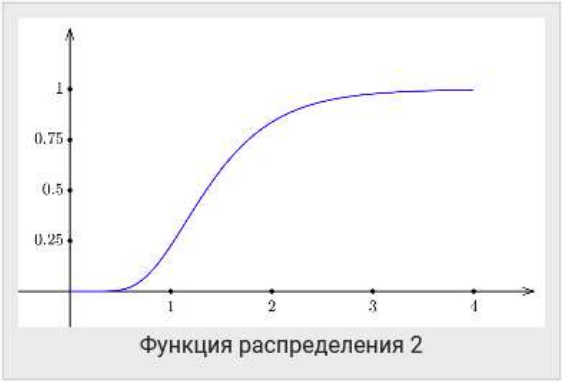
Сопоставьте графики функций распределения с псевдо-гистограммами.



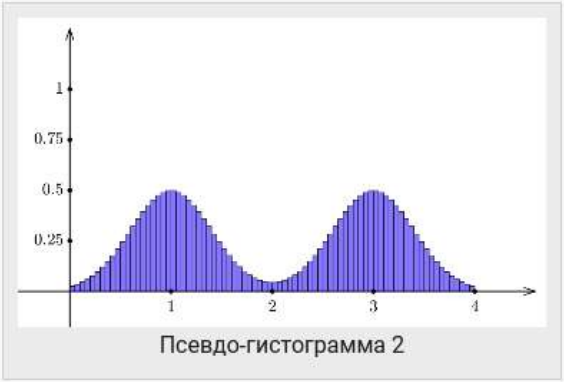
Функция распределения 1



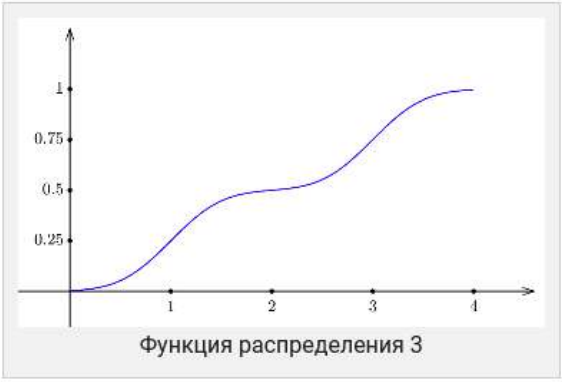
Псевдо-гистограмма 1



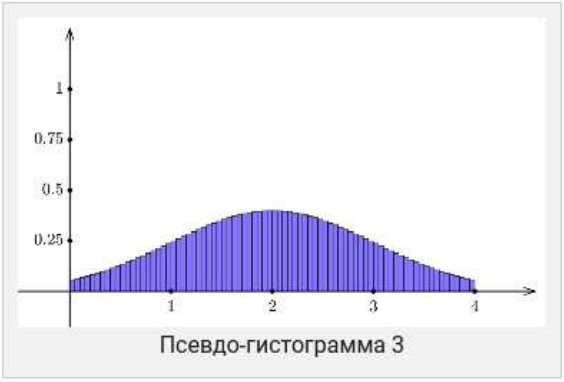
Функция распределения 2



Псевдо-гистограмма 2



Функция распределения 3



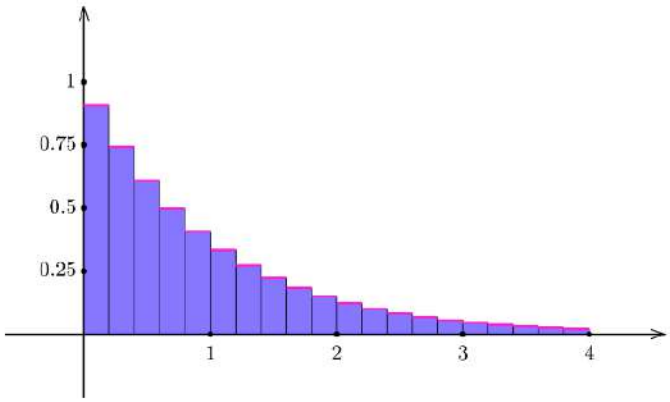
Псевдо-гистограмма 3

Сопоставьте значения из двух списков

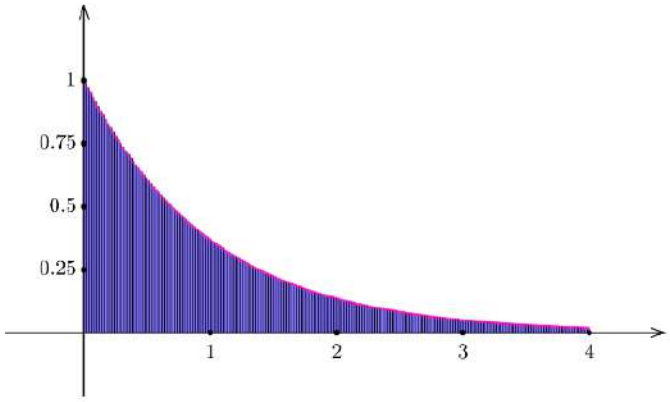
Функция распределения 1	Псевдо-гистограмма 2
Функция распределения 2	Псевдо-гистограмма 1
Функция распределения 3	Псевдо-гистограмма 3

Плотность — неформально

Если посмотреть на верхушки столбцов гистограммы, то получится ступенчатая разрывная функция:



Однако, чем меньше длина интервала при построении нашей псевдо-гистограммы, тем больше эти ступеньки становятся похожи на связную линию:



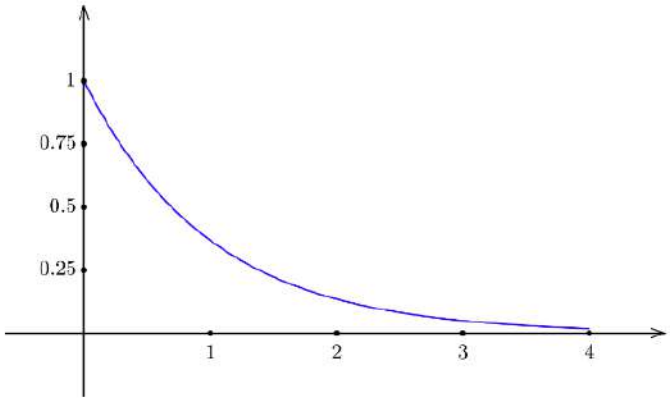
Производная функции распределения — плотность

Если устремить к 0 ширину интервалов в разбиении горизонтальной оси, то мы получим не что иное, как производную функции распределения F_ξ.

Действительно, рассмотрим столбик над интервалом $(a, a+\varepsilon)$. Высота столбика — это $\frac{1}{\varepsilon}(F_\xi(a+\varepsilon) - F_\xi(a))$. Если устремить длину интервала к 0, то высота столбика в точке a будет равняться

$$\lim_{\varepsilon \to 0} \frac{F_\xi(a+\varepsilon) - F_\xi(a)}{\varepsilon}$$

То есть в точности производная F_ξ в точке a. Таким образом, при стремлении ширины интервалов к 0 в пределе псевдо-гистограмма превратится в производную F_ξ, вот как она будет выглядеть:

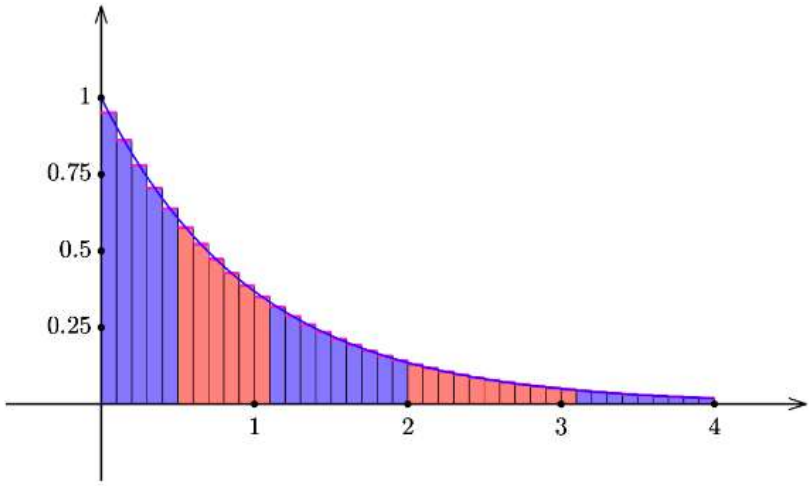


Для случайной величины ξ такую функцию называют *плотность вероятности* — формально мы про неё поговорим через шаг. А сейчас попробуем неформально понять её смысл.

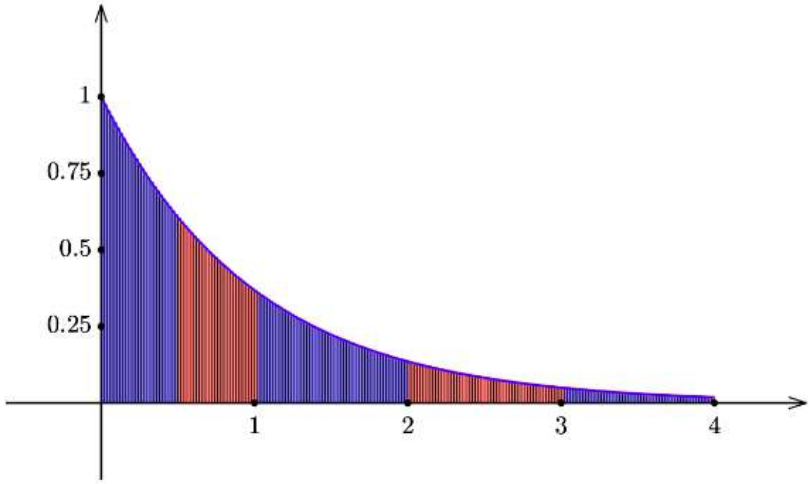
Интеграл плотности по отрезку — это вероятность попадания на этот отрезок

Как мы уже говорили, чтобы приближенно вычислить вероятность того, что случайная величина попала на некоторый промежуток, нужно просуммировать площадь столбиков над интервалами, которые входят в отрезок.

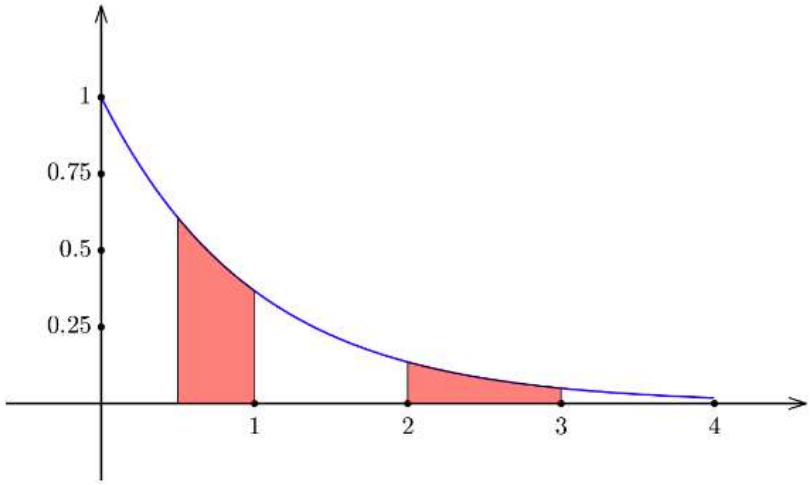
Например, на рисунке ниже суммарная площадь красных столбцов примерно равна вероятности того, что ξ приняла значение из множества $[0.5, 1] \cup [2, 3]$.



При стремлении длины интервалов к 0 приближенная вероятность, вычисленная таким образом, будет стремиться к точной вероятности. Согласитесь, что сумма площадей столбиков очень похожа на интегральную сумму — мы её определили на этом шаге.



Таким образом, вероятность попадания на промежуток — это площадь под графиком *плотности вероятности* на данном промежутке. То есть интеграл *плотности вероятности* по промежутку.



Плотность — формально

Определение. Случайная величина ξ с функцией распределения F_ξ называется *абсолютно непрерывной*, если существует функция $p_\xi : \mathbb{R} \to \mathbb{R}_{\geqslant 0}$ такая, что для всех $a \in \mathbb{R}$ выполнено

$$F_\xi(a) = \int\limits_{-\infty}^{a} p_\xi(x)dx.$$

В этом случае функция p_ξ называется *плотностью вероятности*.

(За $\mathbb{R}_{\geqslant 0}$ обозначено множество неотрицательных действительных чисел.)

Заметим, что по свойствам интеграла выполнено

$$F'_\xi(a) = p_\xi(a).$$

То есть p_ξ из определения выше — это как раз то, что мы получили на предыдущем шаге как предел псевдо-гистограмм.

Может показаться, что существование плотности — это какое-то редкое явление. На самом деле среди непрерывных случайных величин "экзотика" — это скорее те, для которых плотность не существует. Эквивалентно можно переформулировать требование существования плотности так: функция распределения F_ξ непрерывной случайной величины должна быть дифференцируема во всех точках за исключением, быть может, конечного или счётного числа точек.

Непрерывная и в то же время не дифференцируемая в более чем счётном числе точек функция — это, как правило, искусственно построенные функции, а не те, которые естественным образом возникают в практических задачах.

Вероятность события $\xi \in [a, b]$

Как мы помним, для любой непрерывной случайной величины ξ выполнено $P(\xi \in [a, b]) = F(b) - F(a)$. Если у ξ существует функция плотности, то выполнено:

$$P(\xi \in [a, b]) = F(b) - F(a) = \int\limits_{-\infty}^{b} p_\xi(x)dx - \int\limits_{-\infty}^{a} p_\xi(x)dx = \int\limits_{a}^{b} p_\xi(x)dx.$$

Тем самым, вероятность того, что $\xi \in [a, b]$ равна интегралу функции плотности по отрезку $[a, b]$.

Задача с проверкой. Плотность распределения 2

Задача. Нарисуйте график плотности вероятности для равномерного распределения

1. на отрезке $[0, 1]$
2. на отрезке $[3, 5]$
3. на отрезке $[1, x]$, где $x > 1$

Проверка. Ответы округлите до 3 знаков после запятой.

Заполните пропуски

1. Пусть ξ равномерно распределена на отрезке $[0, 1]$. Тогда $p_\xi(-2) =$, $p_\xi(0.25) =$, $p_\xi(1.5) =$

2. Пусть ξ равномерно распределена на отрезке $[4, 9]$. Тогда $p_\xi(3) =$, $p_\xi(5) =$

3. Пусть ξ равномерно распределена на отрезке $[0.1,\ 0.2]$. Тогда $p_\xi(0.13) =$ (обратите внимание на этот ответ, он может показаться странным)

Пусть ξ имеет экспоненциальное распределение с коэффициентом $\lambda = 1$, то есть имеет функцию распределения $F_\xi(x) = 1 - e^{-x}$.

Найдите формулу для функции плотности распределения p_ξ.

$p_\xi(x) =$

Введите математическую формулу

Напишите ваш ответ здесь...

Пример. Пусть на отрезке $[5, 8]$ плотность случайной величины ξ задана формулой $p_\xi(x) = \frac{1}{x^3}$ (при этом мы не знаем, какова плотность ξ вне этого отрезка). Найдите $P(\xi \in [5, 8])$.

Решение.

$$P(\xi \in [5, 8]) = \int\limits_5^8 p_\xi(x)dx = \int\limits_5^8 \frac{1}{x^3}dx$$

Первообразная функции $\frac{1}{x^3}$ это $\frac{-1}{2x^2}$. Поэтому

$$\int\limits_5^8 \frac{1}{x^3}dx = \frac{-1}{2 \cdot 8^2} - \frac{-1}{2 \cdot 5^2} = -\frac{1}{128} + \frac{1}{50} \approx \ 0.012.$$

То есть $P(\xi \in [5, 8]) \approx 0.012$.

Задача. Пусть на отрезке $[6, 7]$ плотность случайной величины ξ задана формулой $p_\xi(x) = \cos(x)$ (при этом мы не знаем, какова плотность ξ вне этого отрезка). Найдите $P(\xi \in [6, 7])$.

Ответ округлите до 3 знаков после запятой.

Введите численный ответ

Введите число

Пусть ξ это абсолютно непрерывная случайная величина. Её функция распределения всюду дифференцируема. Тогда для плотности вероятности ξ обязательно выполнены следующие утверждения:

Выберите все подходящие ответы из списка

$\int\limits_{-\infty}^{+\infty} p_\xi(x)dx = 1$

$\forall x : F_\xi(x)' = p_\xi$

$p_\xi \in [0, 1]$

$\forall x : p_\xi(x)' = F_\xi(x)$

Дополнительный материал

Сингулярные распределения: лестница Кантора

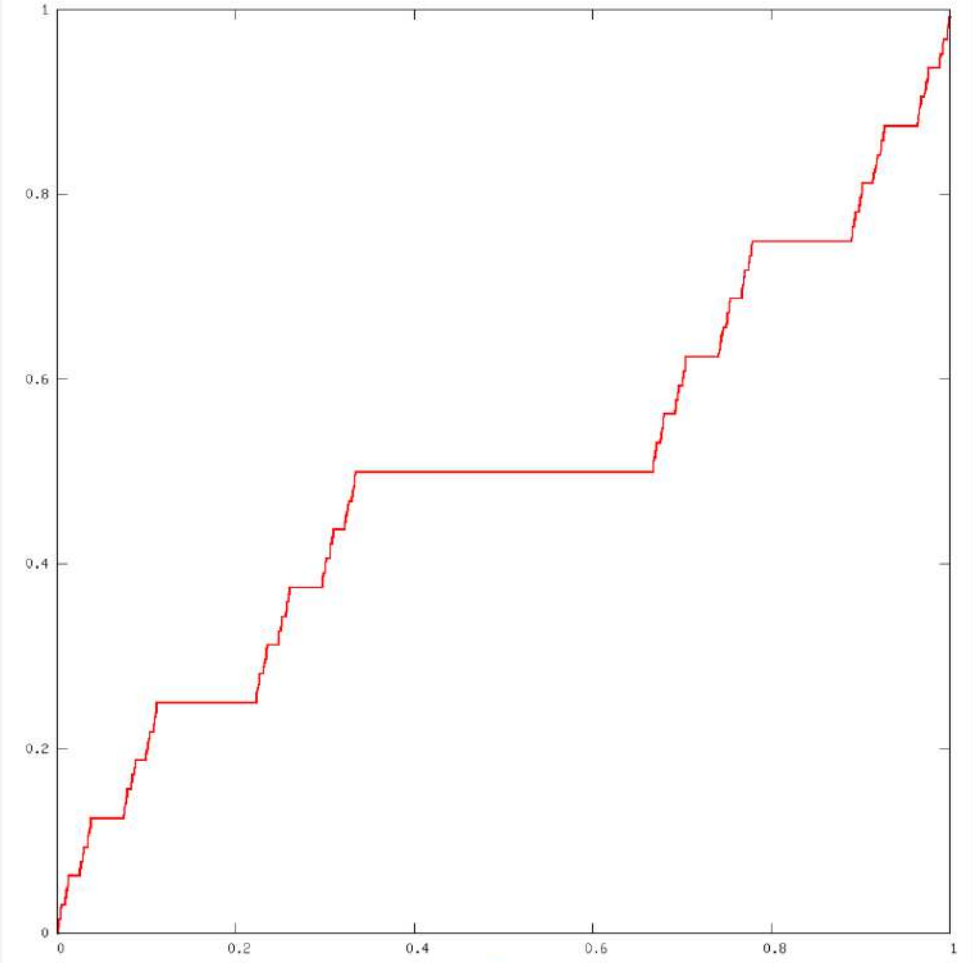
Есть отдельный подкласс непрерывных распределений — *сингулярные*. Уже само название подсказывает, что хороших свойств от этих распределений ждать не стоит :) Зато интересно.

Классический пример сингулярной функции распределения — это Канторова лестница. Строится она следующим образом. Полагается $F_\xi(0) = 0$ и $F_\xi(1)$, после чего рекурсивно повторяется процедура:

- Интервал $(0, 1)$ разбивается на 3 равных сегмента
- Для всех точек x на среднем сегменте $[\frac{1}{3}, \frac{2}{3}]$ полагаем $F_\xi(x) = \frac{1}{2}(F_\xi(0) + F_\xi(1)) = \frac{1}{2}$
- Для левого и правого сегмента повторяем процедуру рекурсивно:
    - Разбиваем на три равных сегмента
    - На среднем сегменте полагаем значение F_ξ равным среднему арифметическому значений на концах сегмента-родителя: то есть для левого $\frac{1}{4}$, а для правого $\frac{3}{4}$
    - Для оставшихся четырех сегментов, где функция ещё не определена, повторяем процедуру рекурсивно:
        - Разбиваем на три равных сегмента

        и т.д.

Получается вот такой график. Ниже мы сформулируем некоторые её свойства.



Источник: Википедия

Вот несколько фактов про эту функцию F_ξ:

- Она непрерывна.
- Она дифференцируема почти всюду. То есть дифференцируема во всех точках за исключением множества точек, имеющих суммарную длину 0. (А точнее меру 0 по Лебегу. Неформально говоря, для $\mathbb{R}$ мера по Лебегу — это обобщение понятия длины).
- Производная равна 0 во всех точках, где производная определена.
- Постоянна (равна некоторой константе) почти всюду, то есть за исключением множества точек, имеющих суммарную длину 0. Аналогично, тут подразумевается под этим мера 0 по Лебегу.
- Тем не менее множество точек, в которых она не дифференцируема, более чем счётно (континуально, то есть равномощно $\mathbb{R}$).
- Для функции распределения F_ξ нельзя построить функцию плотности.

Что мы прошли на этом уроке

- Обсудили, что такое гистограмма и псевдогистограмма
- Ввели три формулы, связывающие функцию распределения F_ξ и плотность p_ξ:

1. $F'_\xi(a) = p_\xi(a)$
2. $F_\xi(a) = \int\limits_{-\infty}^{a} p_\xi(x)dx$
3. $P(\xi \in [a, b]) = \int\limits_{a}^{b} p_\xi(x)dx.$

- В дополнительном материале немного поговорили про случайные величины, которые являются непрерывными, но не абсолютно непрерывными. Далее все непрерывные случайные величины, которые будут нам встречаться, будут абсолютно непрерывными

Что нас ждёт на следующем уроке

На следующем уроке мы

- научимся считать математическое ожидание для абсолютно непрерывных случайных величин
- разберём несколько примеров часто встречающихся распределений