Содержание

Описательная статистика

Статистический анализ = описательная статистика + индуктивная статистика

Описательная статистика — это сжатая и концентрированная характеристика изучаемого явления, представленная в виде графиков, таблиц, схем и числовых выражений.

Описательная статистика противопоставляется индуктивной в том смысле, что не делает выводов о генеральной совокупности на основании результатов исследования частных случаев.

Описательная статистика	Индуктивная статистика
методы описания статистических данных, представления их в форме таблиц, распределений и пр.	обработка данных, полученных в ходе эксперимента, и формулировка выводов, имеющих прикладное значение для конкретной области человеческой деятельности тесно связана с теорией вероятностей и базируется на её математическом аппарате

➔ Как сделать описательную статистику?

При решении статистических задач придерживаются следующего порядка:

1. Собирают исходные данные. При этом учитывают размер выборки. Чтобы получить достоверные данные, минимальное число не может быть меньше 1000. Чем оно будет больше, тем точнее получится итоговый результат
2. Строят вариационный ряд. Все полученные данные упорядочивают по возрастанию. Чтобы это было удобнее выполнить, находят минимальный и максимальный элементы, а затем относительно них переписывают его в нужной последовательности. В некоторых случаях для упрощения процедуры обработки допускается вычитание из каждого элемента ряда минимального значения. Таким образом, работа дальше ведётся не с конкретными размерами, а только с их отклонениями
3. Проводят группировку данных. Для этого их разбивают на R интервалов, число которых соотносят с количеством наблюдений
4. Определяют частости и эмпирические плотности вероятностей. Частость используется для того, чтобы заменить частоты при составлении вариационных рядов

5. Строят полигон. Но для этого первоначально определяют масштаб по осям
6. Строят гистограмму и эмпирическую функцию распределения
7. Используя данные из гистограммы, рассчитывают параметры распределения
8. Оформляют результат, который сводят в таблицу, схему, гистограмму, график или прочее

Далее рассмотрим основные показатели, которые используют для осуществления методов описательной статистики.

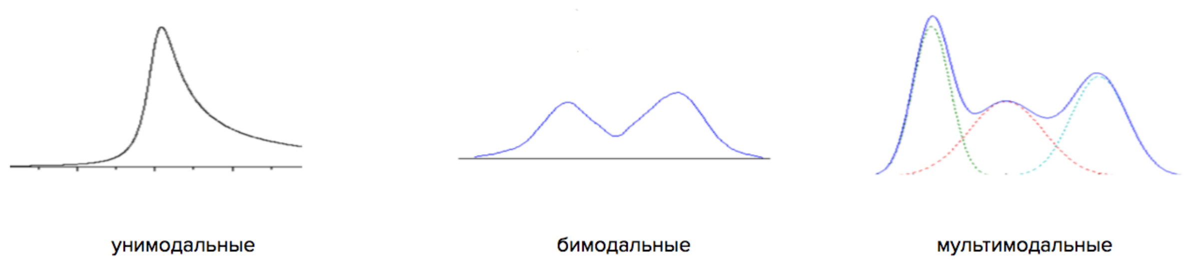
Меры центральной тенденции

В описательной статистике мы рассматриваем только выборку объектов, а затем расширяем выводы о ней на множество всех объектов. Описать группу объектов одним числом позволяют меры центральной тенденции — различные средние.

Среднее арифметическое нельзя посчитать для нечисловых данных. Кроме того, она искажается асимметрией данных и выбросами. Например, если у нас 10 человек получают зарплату 10 тысяч рублей, а одиннадцатый получает миллион, то среднее арифметическое получается 100 тысяч. Как в анекдоте: «Чиновники едят мясо, я — капусту. В среднем мы едим голубцы».

Медиана не зависит от крайних величин — не искажается асимметрией данных и такими экстремальными значениями, как в предыдущем примере.

Мода показывает горб распределения — нашу максимальную точку распределения.



Источник иллюстрации

Унимодальное распределение – это распределение, имеющее только одну моду. Бимодальное распределение – это распределение, имеющее две моды (т.е. два "пика"). Мультимодальное распределение – это распределение, имеющее несколько мод (т.е. два или более "пика").

Мультимодальность распределения выборки часто является показателем того, что распределение не является нормальным, а также даёт важную информацию о природе исследуемой переменной. Например, если переменная представляет собой предпочтение или отношение к чему-то, то мультимодальность может означать, что существует несколько определённых высказываемых мнений или несколько определённо различных мнений.

Мультимодальность часто может показывать, что выборка не является однородной и наблюдения порождены двумя или более «наложенными» распределениями. Иногда мультимодальность распределения означает, что выбранные инструменты не подходят для измерения. Например «проблемы разметки» в естественных науках, «смещённые ответы» в социальных.

При этом медиана и мода при использовании теряют большую часть информации о выборке. Однако эти величины можно применять и для нечисловых данных: цвет, бренд и т.д.

У каждой средней величины есть свои плюсы и минусы.

Тип среднего	Плюсы	Минусы
Среднее арифметическое	известно просто вычисляется	определено только для числовых данных искажается асимметрией данных и выбросами
Медиана	не искажается асимметрией данных и выбросами	при использовании теряется большая часть информации о выборке определено не только для числовых данных
Мода	показывает горб распределения	при использовании теряется большая часть информации о выборке определено не только для числовых данных

Меры разброса

Для более полного описания результатов эмпирического исследования используются меры разброса данных, характеризующие степень индивидуальных отклонений от центральной тенденции или от среднего.

Обычно рассчитывают следующие показатели:

- простой размах
- межквартильный размах
- дисперсия
- стандартное отклонение (корень из дисперсии)
- коэффициент вариации

Как и у мер центральной тенденции, у каждой меры разброса есть свои плюсы и минусы.

Мера разброса	Плюсы	Минусы
Размах	легко вычислить	очень неустойчив к выбросам зависит от размера выборки (с ростом увеличивается) не использует значения внутри выборки
Межквартильный размах	устойчив к выбросам не зависит от размера выборки хорошо работает с асимметрично распределёнными данными	использует больше значений внутри выборки, но не все
Дисперсия	использует каждое наблюдение выборки	восприимчива к выбросам плохо работает с асимметрично распределёнными данными

Стандартное отклонение	использует каждое наблюдение выборки легко объяснимо (имеет те же единицы измерения, что и у исходных данных)	восприимчиво к выбросам плохо работает с асимметрично распределёнными данными
Коэффициент вариации	подходит для сравнения наборов данных с различными единицами измерения или сильно отличающимися средними величинами	чувствителен к небольшим изменениям среднего (когда среднее значение близко к нулю, коэффициент вариации будет приближаться к бесконечности) нельзя использовать непосредственно для построения доверительных интервалов для среднего значения (в отличие от стандартного отклонения)

Квантили, квартили и выбросы

Существенную роль для корректного отображения получаемых данных играют также квантили, квартили и выбросы.

Квантиль в математической статистике — значение, которое заданная случайная величина не превышает с фиксированной вероятностью.

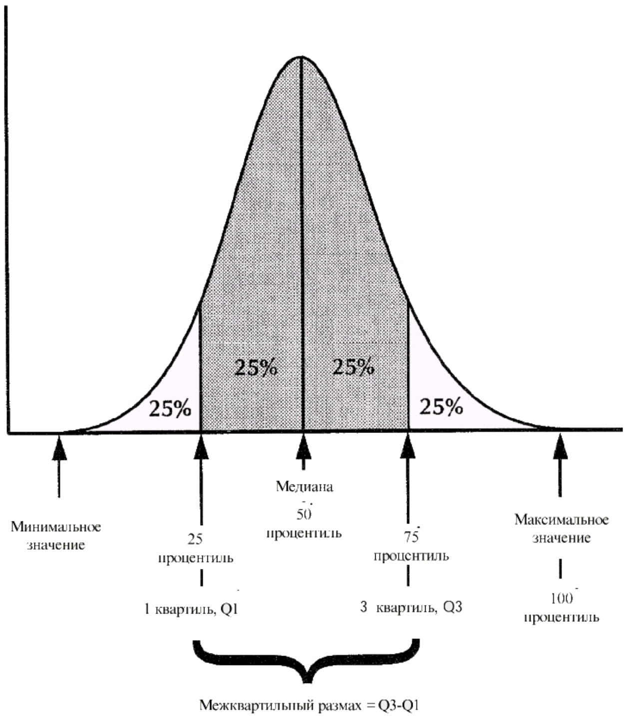
Если вероятность задана в процентах, то квантиль называется **процентилем** или перцентилем. Квантиль нужен для оценки «масштабов» выбросов.

Также важную информацию информацию о структуре вариационного ряда представляют **квартили**. Это значения, которые делят таблицу данных (или её часть) на четыре группы, содержащие приблизительно равное количество наблюдений.

Значение	Единицы измерения
Квантили	0...100
Процентили / перцентили	0...100%

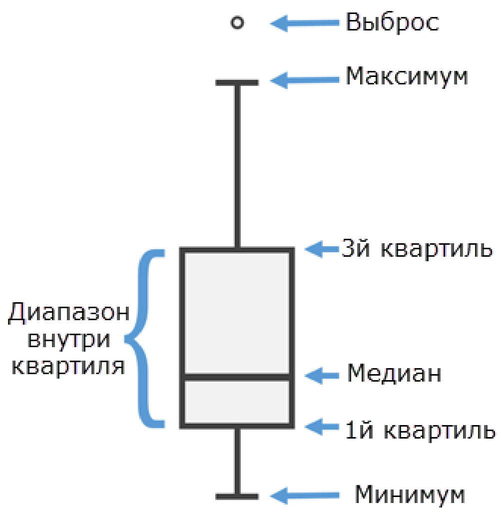
<table>
<tr>
<td>Квартили</td>
<td>
<table>
<tr><td>0Q</td><td>1Q</td><td>2Q</td><td>3Q</td><td>4Q</td></tr>
<tr><td>0%</td><td>25%</td><td>50%</td><td>75%</td><td>100%</td></tr>
</table>
</td>
</tr>
</table>

На графике ниже показаны 25, 50, 75 и 100 процентили. Случаи 25 и 75-ого процентиля, включающие четверть и три четверти выборки соответственно, называются квартилями.



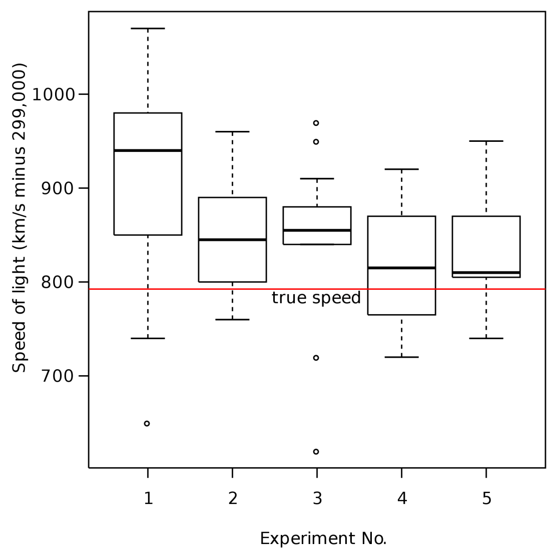
Источник иллюстрации

Межквартильным размахом называется разность между третьим и первым квартилями. Размах, а также нижний и верхний квартили, минимальное и максимальное значение выборки и выбросы хорошо показывает диаграмма размаха. Также такие диаграммы называют «ящики с усами» (boxplot).



Источник иллюстрации

Несколько таких ящиков можно нарисовать бок о бок, чтобы визуально сравнивать одно распределение с другим — их можно располагать как горизонтально, так и вертикально. Расстояния между различными частями ящика позволяют определить степень разброса (дисперсии) и асимметрии данных, выявить выбросы.



Источник иллюстрации

Данные, выходящие за границы усов (выбросы), отображаются на графике в виде точек, маленьких кружков или звёздочек. Иногда на графике отмечают среднее арифметическое и его доверительный интервал — «зарубка» на ящике. Иногда зарубками обозначают доверительный интервал для медианы.

В связи с тем, что не существует общего согласия относительно того, как конкретно строить «ящик с усами», при виде такого графика необходимо искать информацию в сопроводительном тексте относительно того, по каким параметрам ящик с усами строился.

Несмотря на свою простоту и удобство, первоначальная форма ящика с усами обладает и некоторыми недостатками. Один из таких существенных недостатков — отсутствие на графике информации о количестве наблюдений по выборке.

Действительно, ящик с усами позволяет сравнить медианы, квартили, минимумы и максимумы по различным выборкам, но если мы захотим сделать вывод об общей медиане по всей совокупности выборок, то мы не сможем этого сделать, не прибегая к расчётам на исходных данных.

➔ Как искать выбросы?

Выбросы — наблюдения, которые выделяются из общей выборки. Например, все предметы на кухне имеют температуру около 22–25 грудусов по Цельсию, а — духовка 220.

Поскольку многие статистические методы неустойчивы к выбросам, их необходимо находить и исключать из выборки. Простейшим способом является **метод «заборы Тьюки»**, который основан на межквартильном размахе. Похожий способ — **правило трёх сигм**, основанное на стандартном отклонении.

Однако существуют и более тонкие критерии:

- Критерий Шовене
- Критерий Граббса
- Q-тест Гибсона
- Модифицированный тау-тест Томпсона
- Критерий Пирса (итеративный сходящийся алгоритм)

Автоматизировать поиск выбросов позволяют обучаемые алгоритмы:

- Алгоритмы кластеризации типа k-means
- DBScan (тоже кластеризация)
- Isolation Forest
- Random Cut Forest
- Карты Шухарта

Популярные распределения

Распределение вероятностей — это закон, описывающий область значений случайной величины и соответствующие вероятности появления этих значений.

Дискретная случайная величина — это случайная величина, множество значений которой не более чем счётно. Очевидно, значения дискретной случайной величины не содержат какой-либо непрерывный интервал на числовой прямой. Примеры: любая случайная величина, принимающая целочисленные значения.

Непрерывной случайной величиной — это случайная величин, которая в результате испытания принимает все значения из некоторого числового промежутка. Число возможных значений непрерывной случайной величины бесконечно.

Дискретные	Непрерывные
Равномерное Бернулли Биноминальное Пуассоновское Геометрическое	Равномерное Нормальное (Гауссовское) Логнормальное Гамма-распределение Экспоненциальное Лапласса Коши Бета-распределение Хи-квадрат Стюдента Фишера Рэлея Вейбулла Логистическое Вигнера Парето

➔ Дискретное равномерное распределение

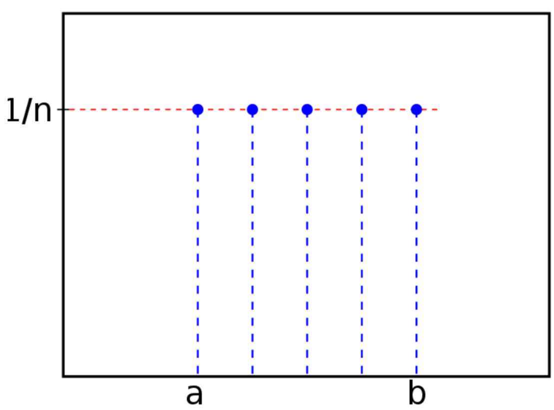
В теории вероятностей случайная величина имеет дискретное **равномерное распределение**, если она принимает конечное число n значений с равными вероятностями, соответственно, вероятность каждого значения равна 1/n.

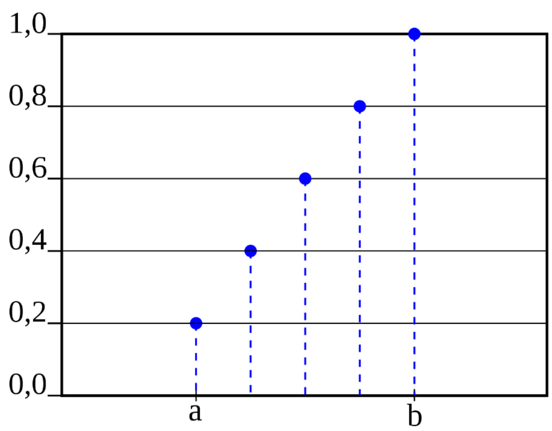
Пример: При бросании игральной кости случайная величина — число точек на грани — принимает одно из 6-и возможных значений: 1, 2, 3, 4, 5, 6.

Вероятность выпадения одной точки из шести равна 1/6, одинакова для каждой точки, поэтому случайная величина имеет дискретное равномерное распределение.

Характеристики дискретного равномерного распределения: функция вероятности, функция распределения, мат ожидание, медиана и дисперсия. Моды у данного типа распределения нет (потому что все значения могут выпасть равновероятно).

Функция	Формула		
Функция вероятности	$\frac{1}{n}, a \le k \le b$		$0, else$
Функция распределения	$0, k < a$	$1, k > b$	$\frac{k-a+1}{n}, a \le k \le b$
Математическое ожидание	$\frac{a+b}{2}$		
Медиана	$\frac{a+b}{2}$		
Мода	нет		
Дисперсия	$\frac{n^2-1}{12}$		





n=5, где n=b-a+1
Функция вероятности

n=5, где n=b-a+1
Функция распределения

Источник иллюстрации

→ Распределение Бернули

Распределение Бернулли в теории вероятностей и математической статистике — дискретное распределение вероятностей, моделирующее случайный эксперимент произвольной природы при заранее известной вероятности успеха или неудачи.

Случайная величина X имеет распределение Бернулли, если она принимает всего два значения: 1 и 0 с вероятностями p и q = 1-p соответственно.

Принято говорить, что событие X=1 соответствует «успеху», а событие X=0 — «неудаче». Эти названия условные, и в зависимости от конкретной задачи могут быть заменены на противоположные.

Пример: В урне 20 белых и 10 черных шаров. Вынули 4 шара, причём каждый вынутый шар возвращают в урну перед извлечением следующего, и шары в урне перемешивают. Найти вероятность того, что из четырех вынутых шаров окажется 2 белых.

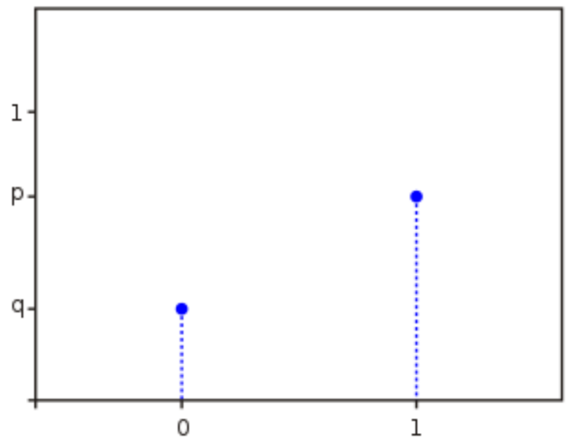
Решение:
Событие A – достали белый шар. Тогда вероятности:

$$P(A) = \frac{2}{3}, P(\bar{A}) = \frac{1}{3}$$

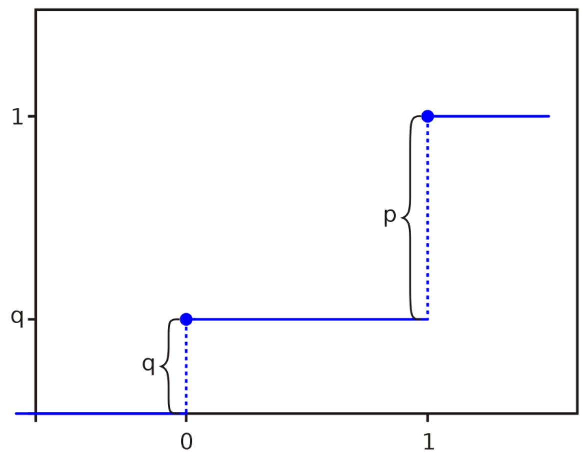
По формуле Бернулли требуемая вероятность равна:

$$P_4(2) = C_4^2 \left(\frac{2}{3}\right)^2 \left(\frac{1}{3}\right)^2 = \frac{8}{27} \approx 0,3$$

Функция	Формула	
Функция вероятности	q	$k = 0$
	p	$k = 1$
Функция распределения	0	$k < 0$
	q	$0 \leq k < 1$
Математическое ожидание	1	$k \geq 1$
Медиана	$\begin{cases} 0, & q > p \\ 0, 1, & q = p \\ 1, & q < p \end{cases}$	
Дисперсия	pq	



Функция вероятности



Функция распределения

Источник иллюстрации

➔ Биноминальное распределение

Биноминальное распределение в теории вероятностей — распределение количества «успехов» в последовательности из n независимых случайных экспериментов, таких, что вероятность «успеха» в каждом из них постоянна и равна p.

Пример: При автоматической наводке орудия вероятность попадания по быстро движущейся цели равна 0,9. Задача: найти наивероятнейшее число попаданий при 50 выстрелах.

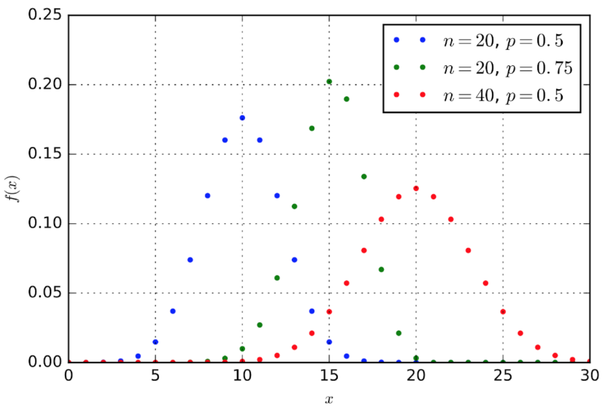
Решение:
Здесь n = 50, p = 0,9, q = 1 - 0,9 = 0,1. Поэтому имеем неравенства:

$50 \times 0,9 - 0,1 \leq k \leq 50 \times 0,9 + 0,9$

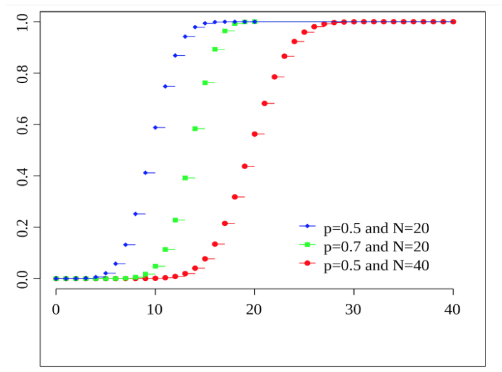
$44,9 \leq k \leq 45,9$

Следовательно, k = 45

Функция	Формула
Функция вероятности	$\binom{n}{k} p^k q^{n-k}$
Функция распределения	$I_{1-p}(n - \lfloor k \rfloor, 1 + \lfloor k \rfloor)$
Математическое ожидание	np
Медиана	одно из $\{\lfloor np \rfloor - 1, \lfloor np \rfloor, \lfloor np \rfloor + 1\}$
Мода	$\lfloor (n+1)p \rfloor$
Дисперсия	npq



Функция вероятности



Функция распределения

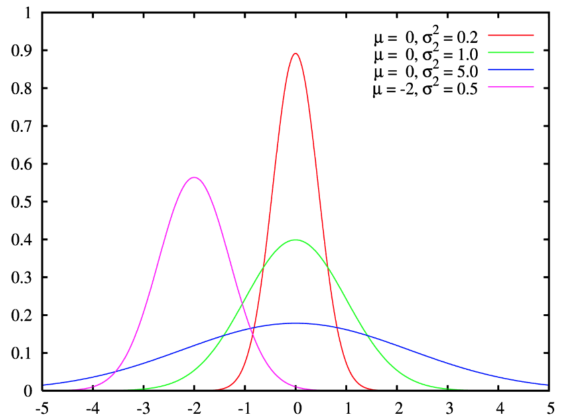
Источник иллюстрации

→ Нормальное распределение

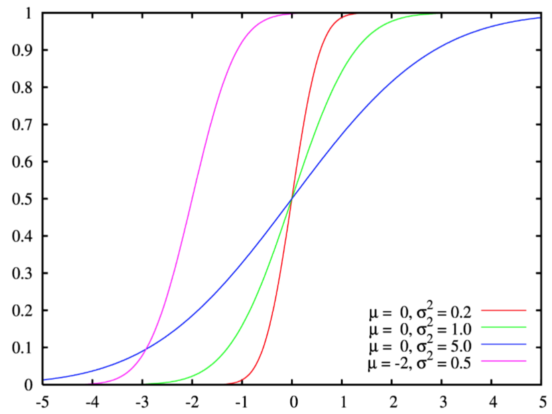
Нормальное распределение, также называемое распределением Гаусса — распределение вероятностей, которое в одномерном случае задаётся функцией плотности вероятности, совпадающей с функцией Гаусса.

Нормальное распределение характеризуется тем, что крайние значения признака в нём встречаются редко, а значения, близкие к средней величине, — часто.

Функция	Формула
Плотность вероятности	$\frac{1}{\sigma\sqrt{2\pi}}\,exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
Функция распределения	$\frac{1}{2}\left[1+erf\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right)\right]$
Математическое ожидание	μ
Медиана	μ
Мода	μ
Дисперсия	σ^2



Плотность вероятности



Функция распределения

Источник иллюстрации

➔ Экспоненциальное распределение

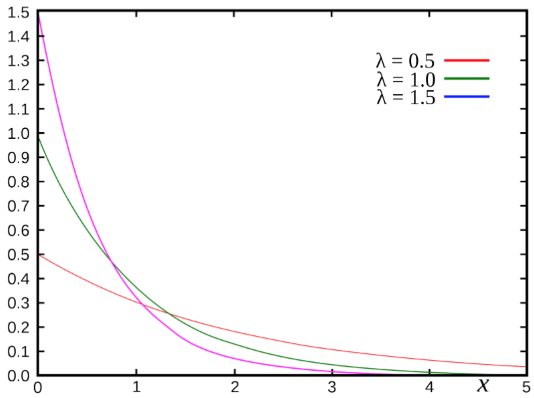
Экспоненциальное распределение — абсолютно непрерывное распределение, моделирующее время между двумя последовательными свершениями одного и того же события.

Типичные примеры, где реализуется экспоненциальное распределение:

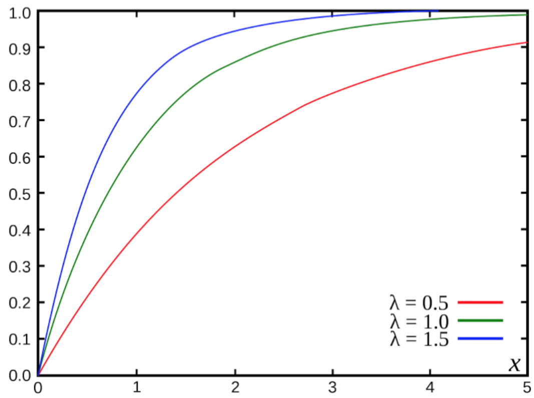
- теория обслуживания, при этом X, например, — время ожидания при техническом обслуживании

- теория надежности, здесь X, например, — срок службы оборудования до отказа, промежуток времени между поломками

Функция	Формула
Плотность вероятности	$\lambda e^{-\lambda x}$
Функция распределения	$1 - e^{-\lambda x}$
Математическое ожидание	λ^{-1}
Медиана	$ln(2)/\lambda$
Мода	0
Дисперсия	λ^{-2}



Плотность вероятности



Функция распределения

Источник иллюстрации

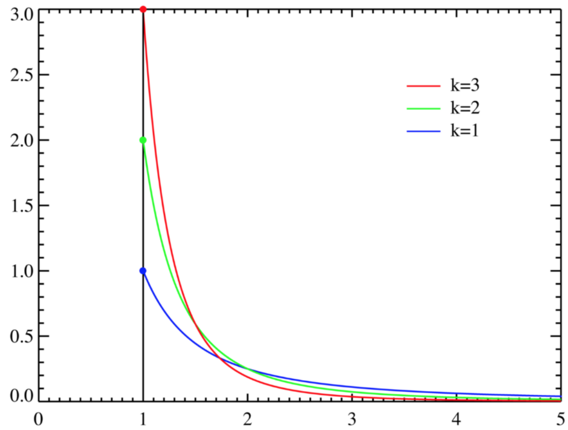
➔ Распределение Парето

Распределение Парето в теории вероятностей – двухпараметрическое семейство абсолютно непрерывных распределений, являющихся степенными.

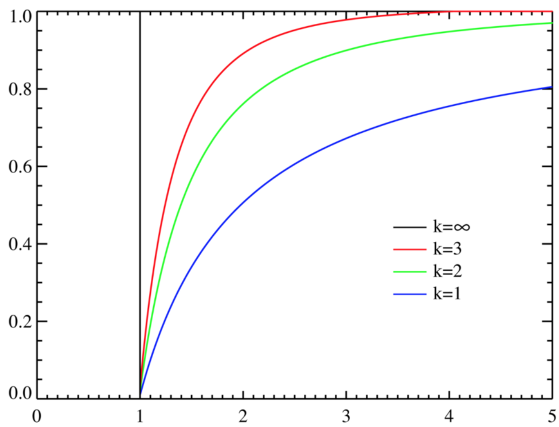
Встречается при исследовании различных явлений, в частности, социальных, экономических, физических и других.

Одно из самых известных распределений. Сейчас встречается в формулировке «20% прилагаемых нами усилий обеспечивают 80% результата и, следовательно, всего лишь 20% результата дают 80% усилий»

Функция	Формула
Кумулятивное распределение Парето вероятности	$\frac{kx_m^k}{x^{k+1}}$
Функция распределения	$1-\left(\frac{x_m}{x}\right)^k$
Математическое ожидание	$\frac{kx_m}{k-1}$ если $k>1$
Медиана	$x_m\sqrt[k]{2}$
Мода	x_m
Дисперсия	$\left(\frac{x_m}{k-1}\right)^2\frac{k}{k-2}$ при $k>2$



Плотность вероятности



Функция распределения

Источник иллюстрации