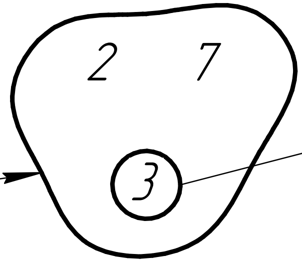
База теории вероятностей

Дискретное вероятностное пространство

<u>Пространство элементарных исходов</u> (множество Ω) – множество, содержащее все возможные взаимоисключающие результаты эксперимента

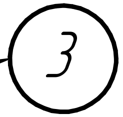


Событие A

<u>Событие A</u> (обычно события обозначают буквой лат. алфавита) – подмножество множества Ω, содержащее часть результатов эксперимента

F – <u>сигма-алгебра событий</u>, множество **всех** событий

(мощность равна 2^{Ω}, т.е. $|F| = 2^{\Omega}$)



Элементарный исход w_i

Элементарный исход w_i – единичный мельчайший неделимый результат эксперимента

С чем мы работаем? → С непрерывным Ω → $P(\{\omega_0\}) = 0$

Вероятность для любой конкретной точки равна 0

С чем мы работаем? → С дискретным Ω →

Вероятность события – числовая мера возможности наступления этого события:

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

Примечание: $|A|$ – мощность подмножества A

В теории вероятностей работают с вероятностью **событий**, т.к. вероятность множества Ω равна единице, а вероятность элементарного исхода равна вероятности события, состоящего из одного элемента.

Вероятностное пространство включает все, с чем мы ознакомились (F, Ω, P)

Из всего сказанного выходим на аксиомы теории вероятностей:

1. P: F -> [0,1] – функция, которая каждому событию ставит в соответствие его вероятность

2. $P(\Omega) = 1$ (вероятность того, что эксперимент произойдет)
3. Для любого конечного или счётного набора попарно непересекающихся событий $A_1, A_2, \ldots$ выполняется:

$$P\left(\bigcup_i A_i\right) = \sum_i P(A_i)$$

Случайная величина

Случайная величина – это функция $X\colon \Omega \to R$, измеримая относительно сигма-алгебры F и борелевской сигма-алгебры на R. Звучит страшно, однако по факту это число, которое получается в результате случайного эксперимента. Выпадение орла при броске монеты, например (обозначим выпадение орла за 1, а решки за 0).

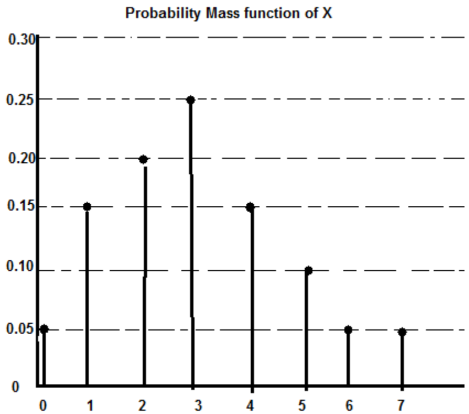
Закодировали орел как 1, решку как 0, но как еще можно представить случайную величину?

дискретный случай

Табличный способ (PMF)

X	x_1	x_2	$\cdots$	x_n
$P(X=x_i)$	p_1	p_2	$\cdots$	p_n

P.S. Конечно же, $\sum p_i = 1$



Функция распределения (CDF)

$F_X(x) = P(X \le x)$

Свойства:

1. Неубывающая
2. $\lim\limits_{x\to-\infty} F_X(x) = 0, \quad \lim\limits_{x\to+\infty} F_X(x) = 1$
3. Справа непрерывна

Значение функции распределения F(x) показывает долю всех возможных исходов эксперимента, при которых случайная величина X принимает значения, меньшие или равные x. Например, если мы говорим о температуре воздуха, то F(20°C) покажет вероятность того, что температура окажется ниже или равной 20°C

Плотность распределения (PDF)

(для непрерывного случая)

$$f_X(x) = \frac{dF_X(x)}{dx}$$

– производная ф-ии распределения

Тогда вероятность случайной величины попасть в промежуток от a до b:

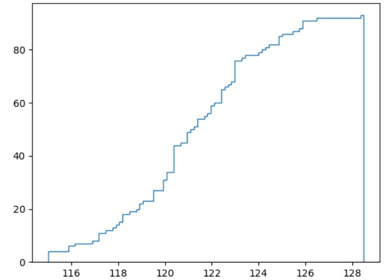
$$P(a < X \le b) = \int_a^b f_X(x)dx$$

дискретный случай

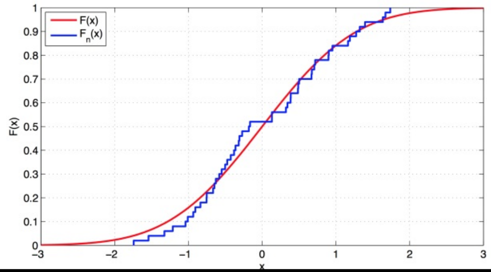
непрерывный случай

Эмпирическая функция распределения

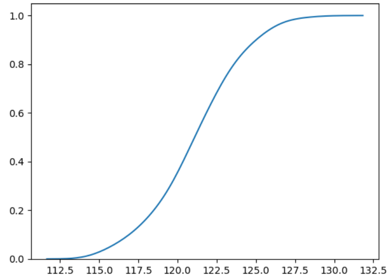
много-много раз проделали эксперимент – получили теор. ф-ию распределения:



По теореме Гливенко – Кантелли эмпирическая функция распределения сходится к истинной равномерно почти наверное



Теоритическая функция распределения



Числовые характеристики случайных величин

Что еще за моменты?

Моменты нужны для определения поведения случайной величины. 1-ый момент показывает среднее по вероятности значение, 2-ой – разброс данных, 3-ий – насколько кривая распределения налоняется вправо или влево (ассиметрия), 4-ый – насколько острый пик у кривой (эксцесс).

Начальные моменты
k-ым начальным моментом называется мат. ожидание от случайной величины в степени k: $\nu_k = \mathbb{E}\left[X^k\right]$
Например, первый начальный момент – то самое мат. ожидание

Центральные моменты
k-ым центральным моментом назыв-ся следующая величина:
$\mu_k = \mathbb{E}\left[(X - \mathbb{E}[X])^k\right]$
Например,
2-ой центральный момент – var(X)
3-ий – ассиметрия, 4-ый – эксцесс
т.е. центральный момент описывает поведение с.в. относительно среднего значения

Первый момент

Математическое ожидание (дискретное)
Смысл: средневзвешенный по вероятности арифметический результат повторяющегося эксперимента. (складываем произведения возможных результатов с соответствующей вероятностью)

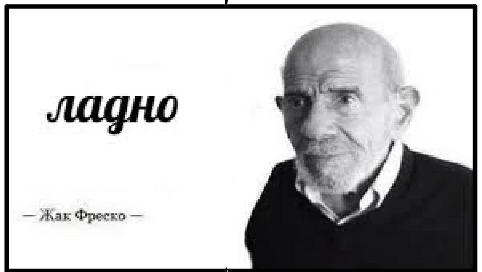
$$\mathbb{E}[X] = \sum_{k=1}^{n} x_k P(X = x_k)$$

Математическое ожидание (непрерывное)
Смысл: "центр тяжести" под графиком плотности.
По сути (из определения интеграла) мы точно также складываем произведения случайной величины и ее вероятности.

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x f_X(x) dx = \int x \, dF_X(x)$$

Свойства мат. ожидания:

- $Mc = c$
- $M(cX) = cMX$
- $M(X + Y) = MX + MY$
- $M(X \cdot Y) = MX \cdot MY$



Второй момент

Дисперсия
– мера разброса данных

$$\operatorname{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

+ среднеквадратичное отклонение – корень из дисперсии, данная статистика интерпритируема, так как имеет ту же размерность, что и случайная величина.

$$\sigma_X = \sqrt{\operatorname{Var}(X)}$$

Свойства дисперсии:

- $\operatorname{Var}(c) = 0$
- $\operatorname{Var}(aX) = a^2 \operatorname{Var}(X)$
- $\operatorname{Var}(X + c) = \operatorname{Var}(X)$
- $X = \mathbb{E}[X]$ п.н. $\iff \operatorname{Var}(X) = 0$
- $\operatorname{Var}(X + Y) = \operatorname{Var}(X) + \operatorname{Var}(Y)$

моментов бесконечно много, но самые частоиспользуемые – первые два

Наиболее полезные распределения случайных величин

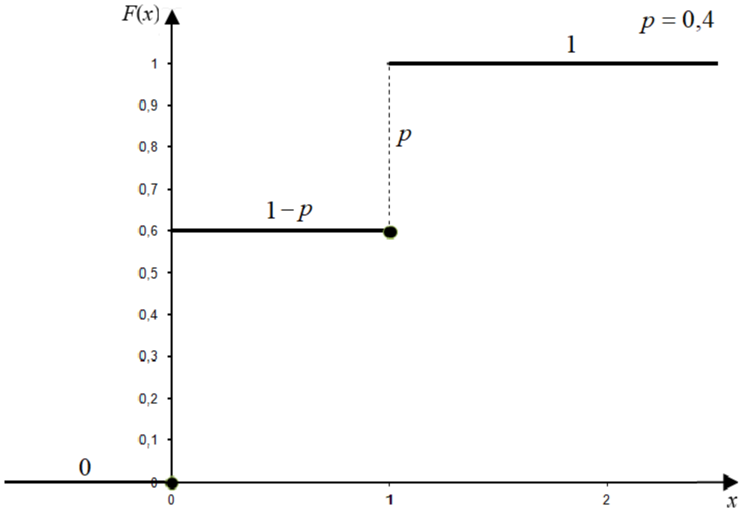
Распределение Бернулли

Случайная величина X ~ Bernoulli(p), если принимает только два значения: 1 ("успех") с вероятностью p, 0 ("неудача") с вероятностью q = 1 – p

Функция вероятности (PMF):

$$P(X=k)=\begin{cases} p & \text{при } k=1 \\ 1-p & \text{при } k=0 \\ 0 & \text{иначе} \end{cases} \quad \text{или} \quad P(X=k)=p^k(1-p)^{1-k} \quad \text{для} \quad k\in\{0,1\}$$

Функция распределения вероятности (CDF):



Числовые характеристики:

E[X] = p

Var(X) = p(1 – p)

Применение:

- Моделирование <u>одного</u> испытания с бинарным исходом
- Анализ успеха/неудачи <u>единичного</u> события

Относится к классу дискретных случайных величин

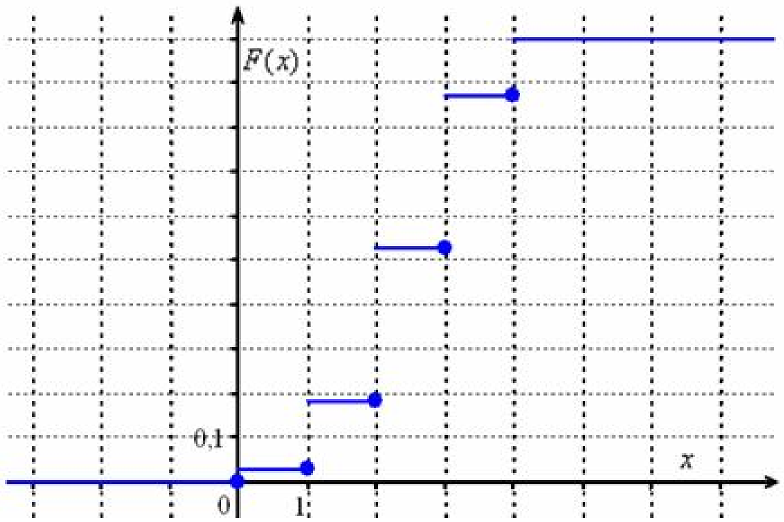
Биномиальное распределение

Случайная величина X ~ Bin(n, p), если описывает число успехов в n независимых испытаниях. Бернулли с одинаковой вероятностью успеха p в каждом.

Функция вероятности (PMF):

$$P(Y=k)=\binom{n}{k}p^k(1-p)^{n-k}, \quad k=0,1,2,\ldots,n$$

Функция распределения вероятности (CDF):



Числовые характеристики:

E[X] = np

Var(X) = np(1 – p)

Применение:

- Моделирование <u>нескольких</u> (n) испытаний с бинарным исходом
- Число успешных операций из n проведенных

Относится к классу дискретных случайных величин

частный случай

n = 1

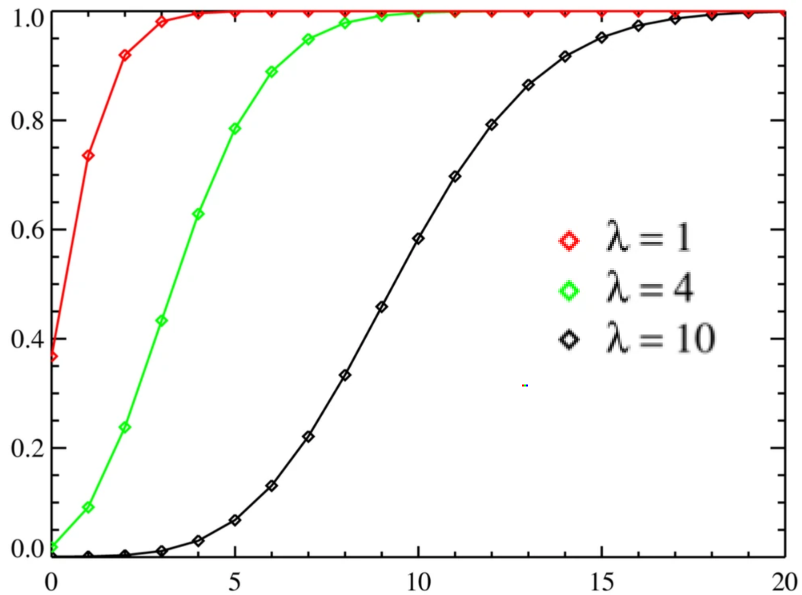
Здесь важно понимать, что речь идет о **независимых** экспериментах. Две случайные величины X и Y называются **независимыми**, когда результат одной не влияет на результат второй. Более формально совместная вероятность двух событий равна произведению вероятностей: P(X,Y)=P(X) · P(Y)

Пуассоновское распределение

Что будет, если взять биномиальное распределение при $n \to \infty$, $p \to 0$, $\lambda = np = const$ (мат. ожидание/среднее количество успехов из бин. распр.) и подставить все это добро в формулу биномиального распределения? Все верно, мы получим распределение Пуассона, PMF для которого выглядит следующим образом:

$$P(X = m) = \frac{\lambda^m e^{-\lambda}}{m!}, \quad m = 0, 1, 2, \dots$$

Функция распределения (CDF)



Числовые характеристики:

$E[X] = \lambda$

$Var(X) = \lambda$

Применение:

- Системы массового обслуживания: вероятность m звонков за время t:

$$P(N_t = m) = \frac{(\lambda t)^m e^{-\lambda t}}{m!}$$

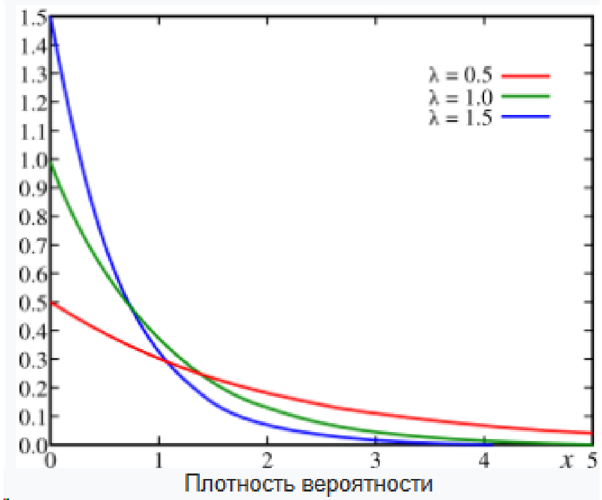
Относится к классу дискретных случайных величин

Экспоненциальное распределение

На этот раз уже используем пуассоновский поток с интенсивностью λ интервалы между событиями T имеют экспоненциальное распределение $T \sim Exp(\lambda)$

Функция распределения CDF:

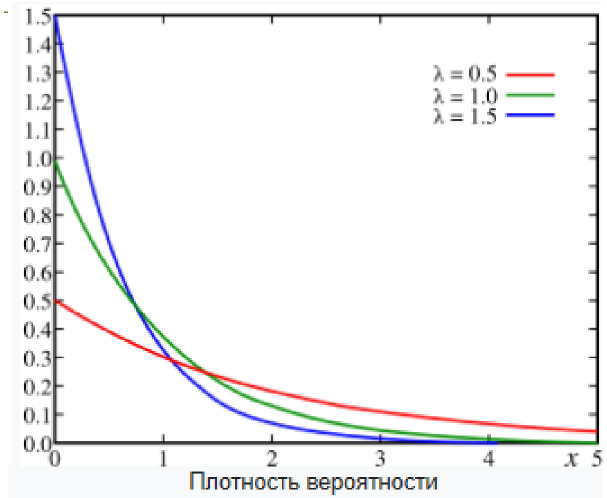
$$F_T(t) = 1 - e^{-\lambda t}$$



Важной особенностью этого распределения является **отсутствие памяти.** Например, вероятность, что звонок придет через 2 минуты, не зависит от того, ждали ли вы уже 5 минут.

Плотность вероятности PDF:

$$f_T(t) = \frac{d}{dt} F_T(t) = \lambda e^{-\lambda t}, \quad t \geq 0$$



Числовые характеристики:

$E[X] = \dfrac{1}{\lambda}$

$Var(X) = \dfrac{1}{\lambda^2}$

Применение:

- Системы массового обслуживания: каждый оператор обрабатывает звонок за экспоненциальное время

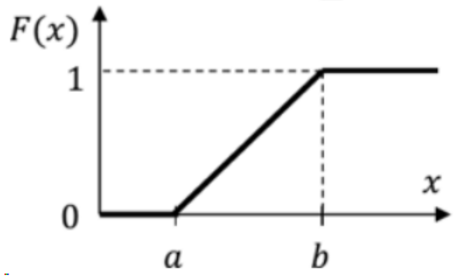
Относится к классу дискретных случайных величин

Равномерное распределение

Равномерное распределение описывает ситуацию, где все возможные исходы равновероятны.

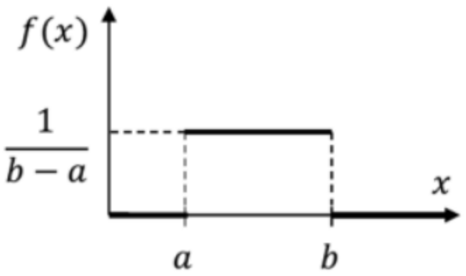
CDF:

$$F(x) = \begin{cases} 0 & x \le a \\ \dfrac{x-a}{b-a} & x \in [a,b]. \\ 1 & x \ge b \end{cases}$$



PDF:

$$f(x) = \begin{cases} \dfrac{1}{b-a} & x \in [a,b] \\ 0 & x \notin [a,b] \end{cases}.$$



Числовые характеристики:

$E[X] = \dfrac{a+b}{2}$

$Var(X) = \dfrac{(b-a)^2}{12}$

Применение:

- Генерация случайных чисел
- Равномерное заполение области точками для создания текстур

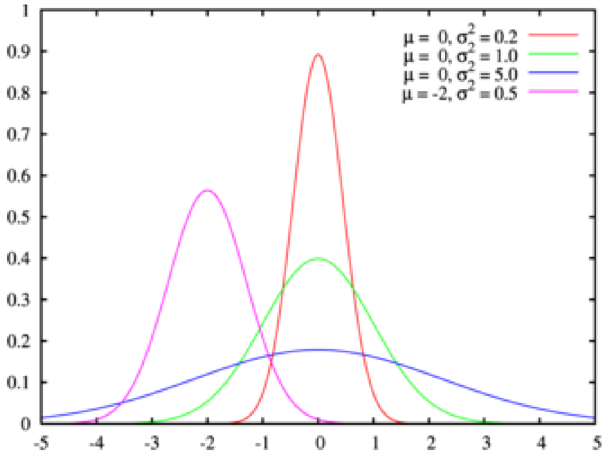
Относится к классу дискретных случайных величин

Нормальное распределение (Гаусс)

Нормальное распределение – очень важное распределений в теории вероятностей и мат. статистике.

Случайная величина $X \sim N(\mu, \sigma^2)$, если её **функция плотности PDF** имеет вид:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}$$



Нормальное распределение характеризуется μ и σ^2

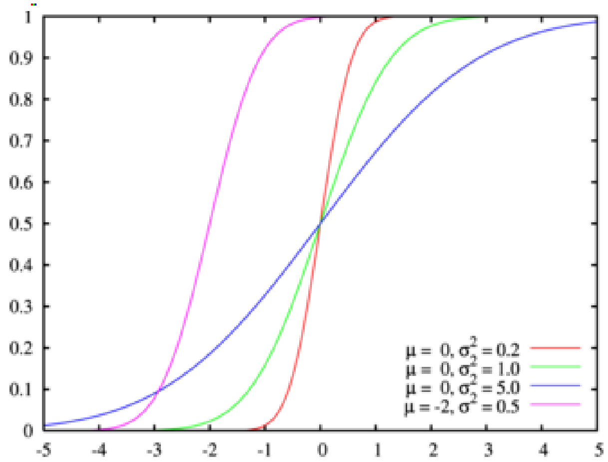
Если $\mu = 0, \sigma^2 = 1$

Его плотность:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Функция распределения вероятностей CDF:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt$$



Числовые характеристики:

$E[X] = \mu$

$Var(X) = \sigma^2$

Применение:

- В измерениях (ошибки часто распределены нормально)
- В физике и инженерии (шум)
- В финансах (доходности акций часто моделируются нормальными распределениями)

Относится к классу непрерывных случайных величин

Распределение Парето

Распределение Коши

Распределение Коши

Распределение Коши возникает, например, при рассмотрении отношения двух независимых нормальных случайных величин.
Пусть X N(0, 1), Y N(0, 1), X, Y независимы, тогда рассмотрим:

$$Z = \frac{X}{Y}$$

PDF для Z: $f_{X,Y}(x,y) = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}$

many, many mathematical transformations later...

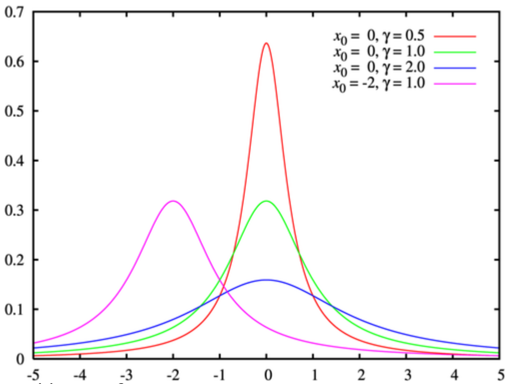
Таким образом, **плотность распределения Коши**:

$$f(x) = \frac{1}{\pi(1+x^2)},$$

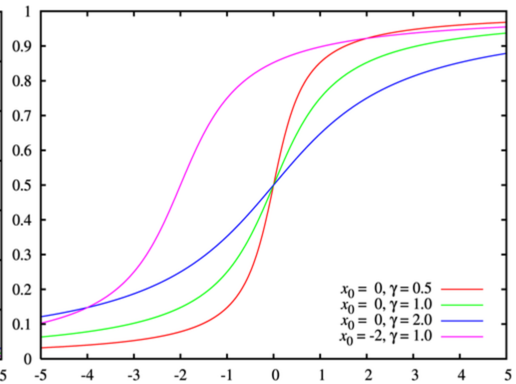
$x \in \mathbb{R}$.

график плотности распределения похож на нормальное р-ние, только с тяжелыми хвостами

PDF



CDF



Числовые характеристики:
E[X] – отсутствует
Var(X) – отсутствует

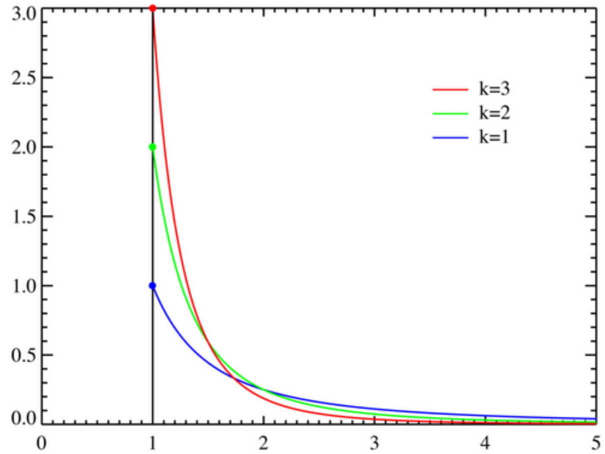
Применение:
- Статистика: устойчивые распределения с тяжёлыми хвостами, моделирование данных с выбросами.
- Математическое моделирование: отношение нормальных величин.

Распределение Парето

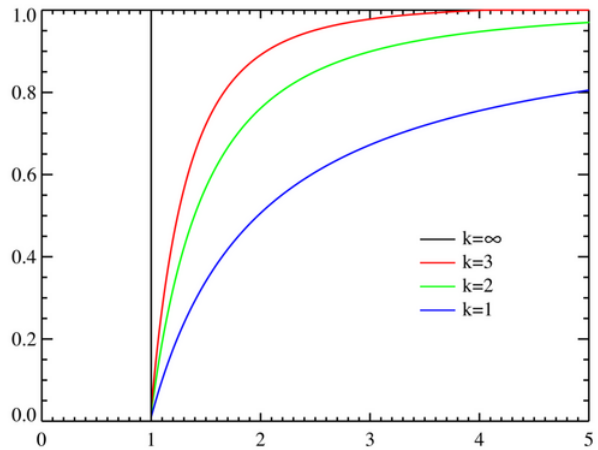
Распределение Парето используется для описания величин, имеющих закон 80/20. Его плотность PDF:

$$f_X(x) = \begin{cases} \dfrac{k x_m^k}{x^{k+1}}, & x \geqslant x_m \\ 0, & x < x_m \end{cases}$$

PDF:



CDF:



Числовые характеристики:
E[X] – существует только при k > 1

$$\mathbb{E}[X] = \frac{k x_m}{k-1}$$

Var(X) – существует только при k > 2

$$\mathrm{D}[X] = \left(\frac{x_m}{k-1}\right)^2 \frac{k}{k-2}$$

Применение:
- Экономика: распределение богатства (которое гласит: 20% популяции владеет 80% богатства)
- Интернет-трафик: распределение размеров файлов.
- Природные явления: распределение размеров городов, землетрясений.

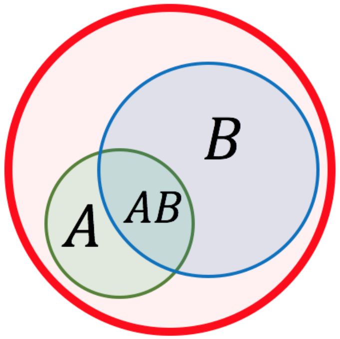
Теорема Байеса

<u>Условная вероятность</u>

Какова вероятность правильного ответа на вопрос "какое животное я загадал"? Наверное, достаточно маленькая. Однако как изменится эта вероятность, если я добавлю условие "у этого животного есть крылья и клюв". Кажется, вероятность успеха выросла. Это и есть условная вероятность, иначе, вероятность события А при условия выполнения условия В:

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

Как можно интерпритировать формулу выше? Например, на кругах Эйлера:



mA – количество исходов множества A

mB – количество исходов множества B

mAB – количество исходов пересечения A и B

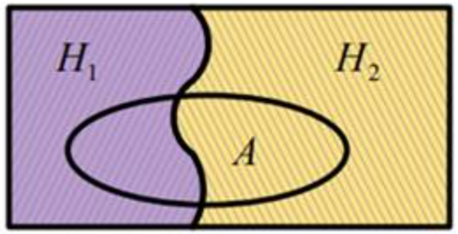
n – общее количество исходов

$P(A) = \frac{mA}{n}$, $P(B) = \frac{mB}{n}$, $P(AB) = \frac{mAB}{n}$

Но по логике $P(A|B) = \frac{mAB}{mB}$ (ведь мы знаем, что событие B точно произошло, тогда n -> mB)

Заметим, что $\frac{P(AB)}{P(B)} = \frac{mAB}{mB}$, вот и формула!

<u>Формула полной вероятности</u>



Как можно представить вероятность события А? Через условную вероятность!

$$P(A) = \sum_{i=1}^{n} P(H_i) \cdot P(A \mid H_i)$$

<u>Формула умножения вероятностей</u>

Справедливо можно заметить, что числитель P(AB) встречается и в P(A|B), и в P(B|A), тогда:

$$P(A \cap B) = P(A \mid B) \cdot P(B) =$$
$$= P(B \mid A) \cdot P(A)$$

<u>Теорема Байеса</u>

Объединим предыдущие шаги и получим ее величество теорему Байеса позволяет пересчитывать вероятность гипотезы Hi при условии наступления события А:

$$P(H_i \mid A) = \frac{P(H_i) \cdot P(A \mid H_i)}{\sum_{j=1}^{n} P(H_j) \cdot P(A \mid H_j)}$$

<u>В машинном обучении</u> Теорема Байеса имеет следующий вид:

ϑ – параметры модели (например, веса нейросети)

D – данные

P(ϑ / D) – апостериорное распределение (что мы знаем о параметрах после наблюдения данных)

P(D / ϑ) – правдоподобие (насколько хорошо модель объясняет данные)

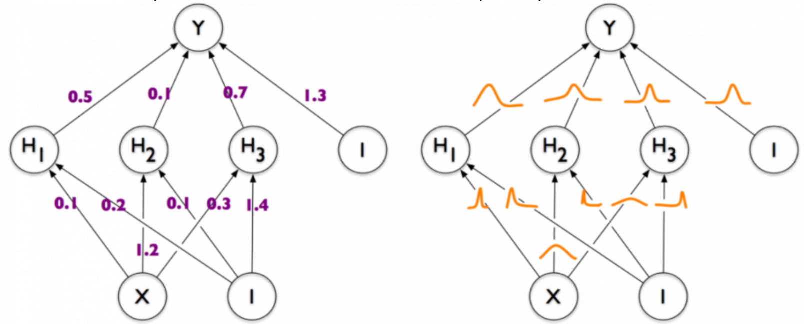
P(ϑ) – априорное распределение (наши знания до наблюдения данных)

P(ϑ / D) – то, что мы стараемся максимизировать

Формула, которую можно найти в байесовском выводе: $p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)}$

<u>В машинном обучении</u> Байес используется в наивкой Байесовской модели (буквально фильтруем данные с помощью теоремы байеса с допущением о независимости признаков)

<u>В глубоком обучении</u> существуют байесовские нейросети, где в каждом нейроне зашиты не веса, а распределения с.в.:



Марков, Чебышев, ЗБЧ, ЦПТ

Неравество Маркова

Пусть X – неотрицательная случайная случайная величина ($X \geq 0$) с конечным математическим ожиданием $\mathbb{E}X$. Тогда для любого $\varepsilon > 0$:

$$P(X \geqslant \varepsilon) \leqslant \frac{\mathbb{E}X}{\varepsilon}$$

Суть можно понять через пример: пусть X – время ожидания в очереди с EX = 10 мин. Вероятность того, что время ожидания превысит 30 мин:

$$P(X \geqslant 30) \leqslant \frac{10}{30} = \frac{1}{3}$$

Важные заметки по поводу Маркова:

1. Неравенство Маркова не требует знания распределения X, только его среднее.
2. Оно не всегда точное, но полезно как грубая верхняя оценка.
3. Работает только для неотрицательных случайных величин.

Неравентво Чебышева

Пусть X – случайная величина с конечным математическим ожиданием $\mu = EX$ и конечной дисперсией $\sigma^2 = VarX$. Тогда для любого $\varepsilon > 0$:

$$P(|X - \mu| \geqslant \varepsilon) \leqslant \frac{\sigma^2}{\varepsilon^2}$$

Оба неравенства являются грубыми, но быстрыми оценками (Чебышев более точен), оба оценивают вероятность выбросов с.в.:

Диапазон	Вероятность
$\mu \pm \sigma$	$\geqslant 0$
$\mu \pm 2\sigma$	$\geqslant 1 - \frac{1}{4} = 0.75$
$\mu \pm 3\sigma$	$\geqslant 1 - \frac{1}{9} \approx 0.888$
$\mu \pm k\sigma$	$\geqslant 1 - \frac{1}{k^2}$

Марков и Чебышёв – это «аварийные» границы, когда у вас нет данных о распределении, но нужно гарантировать, что модель не выйдет за рамки допустимого риска.

Закон больших чисел (ЗБЧ)

С ростом размера выборки <u>среднее арифмитическое</u> сходится по вероятности к <u>мат ожиданию</u> ген. совокупности

Более формализованно:

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n} \xrightarrow{p} \mathbb{E}(X)$$

Примечание: сходимость по вероятности означает

$$\forall \varepsilon > 0 \quad \mathbb{P}(|X_n - X| < \varepsilon) \to 1 \text{ при } n \to \infty$$

то есть

$$\lim_{n \to \infty} \mathbb{P}(|X_n - X| < \varepsilon) = 1$$

Визуализация ЗБЧ: Среднее значение по кубику в зависимости от размера выборки

Центральная предельная теорема (ЦПТ)

Распределение суммы случайных величин при увеличении их количества имеет распределение, близкое к нормальному

Более формализованно:

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow{d} N\left(\mathbb{E}(X), \frac{Var(X)}{n}\right)$$

Примечание: сходимость по распределению означает

$$\lim_{n \to \infty} F_{X_n}(x) = F_X(x)$$

то есть при увеличении n эмпирическая случайная величина стремится стать теоретической

Даже если отдельные наблюдения Xi имеют несимметричное, дискретное или "экзотическое"распределение, сумма многих таких наблюдений часто ведёт себя как нормально распределённая величина.

Для разных распределений подходящее значение n будет разным!

ЦПТ не работает, если:

1. Выборка слишком мала
2. Сильная ассиметрия данных
3. Бесконечная дисперсия
4. Зависимые случайные величины
5. Слишком тяжелые хвосты