

Элементы теории вероятностей. Обработка  
результатов экспериментов.

26 ноября 2019 г.

## 0.1 Вводные понятия

Случайной величиной (с.в.) называется измеримая функция  $X : \Omega \mapsto E$  из пространства возможных событий  $\Omega$  в измеримое пространство  $E$ . Как правило, случайные величины вещественно-значны, т.е.  $E = \mathbb{R}$ .

Вероятность того, что  $X$  примет значение из измеримого подмножества  $S \subseteq E$

$$P(X \in S) \equiv P(\{\omega \in \Omega | X(\omega) \in S\}).$$

Если множество возможных событий счётно (то есть, каждому элементу  $\omega \in \Omega$  можно сопоставить натуральное число  $i \in \mathbb{N}$ ), то множество значений  $\text{range}(X) = \{x_1, x_2, \dots, x_n\}$  тоже счётно. Такая с.в. называется **дискретной**. Если же  $\text{range}(X)$  не счётно, то  $X$  – непрерывная с.в.

### 0.1.1 Операции с вероятностью

- логическое или ( $\vee$ ):  $P(A \vee B) = P(A) + P(B)$  (Если  $A$  и  $B$  – взаимно-исключающие события);
- логическое ( $\wedge$ ):  $P(A \wedge B) = P(A) \cdot P(B)$ .

## 0.2 Функция и плотность распределения случайной величины

### 0.2.1 Дискретная с.в.

Пусть  $X$  – дискретная с.в. Тогда можно говорить о принятии величиной  $X$  некоторого конкретного значения  $x_i$ , а следовательно

$$P(X \in \{x_i\}) \equiv P(X = x_i) = p_i.$$

Функция  $p : \mathbb{N} \mapsto \mathbb{R}$  называется **распределением** вероятности дискретной величины.

Отсортируем значения  $E = \{x_1, x_2, \dots, x_n\}$  и положим их на численную ось. Вероятность того, что  $X$  примет любое из значений  $x_i \leq x_j$

$$P(X \leq x_j) \equiv F_X(x_j) = \sum_{i=1}^j p_i,$$

и называется **функцией распределения** (cumulative distribution function) с.в.  $X$ .

### 0.2.2 Непрерывная с.в.

В случае непрерывной с.в., в качестве распределения вероятности используют **плотность распределения** вероятности  $f(x)$ . Тогда, вероятность

наблюдать величину  $X$  в бесконечно-малом диапазоне значений  $dx$ :

$$P(x \in (x_0, x_0 + dx)) = f(x)dx,$$

а в конечном диапазоне  $[a, b]$ , соответственно,

$$P(X \in [a, b]) = \int_a^b f(x)dx.$$

Очевидно, что определение функции распределения непрерывной с.в.

$$F_X(x) = \int_{\min E}^x f(\xi)d\xi,$$

где  $\min E$  – наименьший элемент множества  $E$ .

**Замечание 1.** Из определения кумулятивного распределения непрерывной функции следует, что

$$P(x \in [a, b]) = F_X(b) - F_X(a).$$

### 0.3 Числовые характеристики случайной величины

**Математическое ожидание**

- Дискретной с.в.:  $M[X] = \sum_i x_i p_i$ .
- Непрерывной с.в.:  $M[X] = \int_{-\infty}^{+\infty} x f(x) dx$ .

**Моментами** распределения называют математические ожидания следующих функций с.в.  $X$ :

$$\begin{aligned} \mu_1 &= M[X] \equiv m, & (\text{матожидание}) \\ \mu_2 &= M[(X - m)^2] \equiv \sigma^2, & (\text{дисперсия}) \\ \mu_3 &= M[(X - m)^3], & (\text{асимметрия}) \\ \mu_4 &= M[(X - m)^4], & (\text{kurtosis}) \end{aligned}$$

и так далее.

**Замечание 2** (Терминология). В русскоязычной литературе, то, что я обозвал kurtosis называют *эксцесс*. Но, это не правильно, потому что есть kurtosis, и есть *excess* kurtosis. Они отличаются следующим образом:

$$\mu_4^e = \mu_4 - 3.$$

Почему 3? 3 – это куртосис нормального распределения. А вот *эксцесс* куртосиса – это на сколько куртосис распределения превосходит куртосис нормального распределения. Чем куртосис больше, тем острее распределение, и наоборот: чем он меньше, тем оно более пологое.

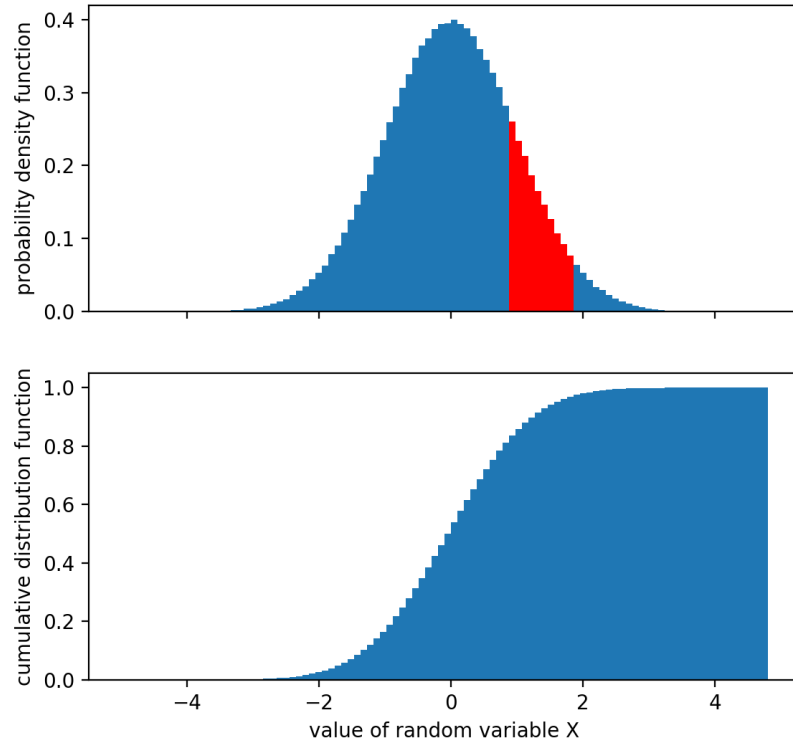


Рис. 1: Плотность (верхняя панель) и функция (нижняя панель) распределения случайной величины  $X \sim N(0, 1)$ .

**Замечание 3.** (Что характеризует кurtosis?) По сути, кurtosis говорит о вероятности наблюдения экстремального значения. *Чем кurtosis больше, тем более вероятно наблюдение экстремального значения* – т.е. тем "толще" хвосты распределения. А поскольку интеграл распределения вероятности равен 1, чем толще хвост, тем уже середина, тем острее распределение. И наоборот. См. Рис. 2

**Замечание 4.** Строго говоря, во всех формулах выше, вместо  $m$  может стоять любое число  $x_0$ , включая ноль. Если  $x_0 = 0$ , то такие моменты называются *начальными*; если  $x_0 = M[X]$ , такие моменты называются *центрными*. Последнее потому, что, для *симметричного* распределения (например, нормального распределения  $N(\mu, \sigma)$ ), матожидание является центральной точкой. При этом, для асимметричных распределений (например,  $\chi^2$ -распределения) это не так.

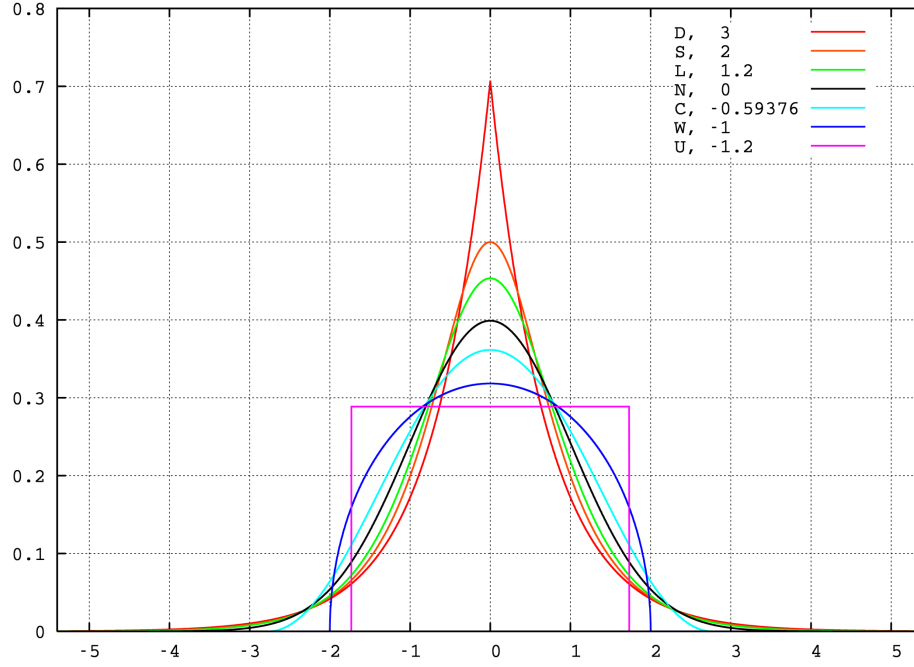


Рис. 2: Плотности распределения вероятности с различными значениями эксцесса кurtosis (на легенде).

Таблица 1: Параметры распределений из Рис. 3

Распределение	mean	median	skewness	kurtosis
$t$	$9.54 \cdot 10^{-3}$	$1.65 \cdot 10^{-2}$	$8.31 \cdot 10^{-2}$	8.07
$\chi^2_4$	4.03	3.34	1.38	5.71

## 0.4 Доверительный интервал. Доверительная вероятность

Предположим, вы измерили некоторую величину  $X$   $n$  раз, и получили какую-то выборку  $D = \{x_1, x_2, \dots, x_n\}$ . Построив гистограмму, вы решили, что  $X \sim N(\mu, \sigma)$ . Теперь начальник требует от вас некоторую оценку величины  $X$ . Вы можете, конечно, дать ему значение  $\langle X \rangle = M[X]$ . Такая оценка будет называться *точечной*. Более информативно будет дать что-то вроде  $\langle X \rangle \pm \sigma_X$ .

Так вот, интервал  $I_{CI} = [\langle X \rangle - \sigma_X, \langle X \rangle + \sigma_X]$  называется *доверительным интервалом* (confidence interval), а вероятность  $p$  обнаружить очередное измерение  $x_{n+1}$  внутри  $I_{CI}$  называется *доверительной вероятностью*.

**Замечание 5.** В рассматриваемом примере, кстати, доверительная веро-

ятность будет 67%.

## 0.5 Двумерная случайная величина

Если на пространстве событий  $\Omega$  заданы *две* случайные функции  $X$  и  $Y$ , говорят, что задана *двумерная* с.в.

Соответственно, кумулятивное распределение  $F_{X,Y}(x, y) = P(X < x \wedge Y < y)$ . Функция плотности распределения (joint probability density function)  $f(x, y)$  определяется аналогично одномерному случаю.

### 0.5.1 Ковариация и коэффициент корреляции

В одномерном случае у нас был параметр дисперсии  $\sigma_X$ , такой, что вариация

$$\begin{aligned}\text{var}[X] &= \sigma_X^2 = \langle (X - \langle X \rangle)^2 \rangle \\ &= \langle (X - \langle X \rangle)(X - \langle X \rangle) \rangle.\end{aligned}$$

Аналогично, для двух с.в. можно определить ко-вариацию

$$\text{cov}[X, Y] = \sigma_{X,Y} = \langle (X - \langle X \rangle)(Y - \langle Y \rangle) \rangle.$$

Иными словами

$$\text{var}[X] = \text{cov}[X, X].$$

**Коэффициент корреляции** – это нормированная ковариация:

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}.$$

## 0.6 Метод наименьших квадратов

### 0.6.1 Метод максимального правдоподобия

Предположим, мы измерили некоторую с.в.  $X \sim f(\theta)$  несколько раз, получили набор значений  $\{x_1, x_2, \dots, x_n\}$ , и теперь хотим определить значения параметров  $\theta = (\theta_1, \theta_2, \dots)$ .<sup>1</sup>

Это можно сделать методом максимального правдоподобия, который заключается в следующем:

1. Составляем функцию правдоподобия  $L(\theta|x) = f(x|\theta)$ .<sup>2</sup>
2. Пытаемся подобрать вектор параметров  $\theta_0$ , который бы максимизировал  $L(\theta|x)$ .

---

<sup>1</sup>Именно эта задача возникает, когда мы пытаемся профитировать данные какой-то функцией.

<sup>2</sup>По сути, функция правдоподобия – это вероятность наблюдения данного набора данных, при данном значении параметра; но интерпретируется как функция  $\theta$ , а данные считаются параметром.

**Замечание 6** (independent identically distributed). Поскольку в нашей задаче рассматриваются отдельные измерения одной и той же величины, логично предположить, что  $x_{i+1}$  независимо от  $x_i$ , и что все измерения происходят из одного и того же распределения. Тогда

$$L(\boldsymbol{\theta}|\mathbf{x}) = f(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^n f(x_i|\boldsymbol{\theta}).$$

### 0.6.2 Статистика $\chi^2$

Сделаем *ещё одно* предположение: пускай

$$\begin{aligned} f(x_i|\boldsymbol{\theta}) &= f(x_i|(\mu, \sigma)) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left[ -\frac{1}{2} \left( \frac{x_i - \mu}{\sigma} \right)^2 \right]. \end{aligned}$$

В таком случае, удобнее не **максимизировать**  $L(\boldsymbol{\theta}|\mathbf{x})$ , а **минимизировать** функцию  $-\ln L(\boldsymbol{\theta}|\mathbf{x})$ :

$$\begin{aligned} -\ln L(\boldsymbol{\theta}|\mathbf{x}) &= -\ln \left( \prod_i f(x_i|\boldsymbol{\theta}) \right) \\ &= -\ln \left[ K_1 \cdot \exp \left( -\frac{1}{2} \sum_i \left( \frac{x_i - \mu}{\sigma} \right)^2 \right) \right] \\ &= \ln K + \frac{1}{2} \sum_i \left( \frac{x_i - \mu}{\sigma} \right)^2. \end{aligned}$$

Иными словами, нужно **минимизировать сумму квадратов** отклонений измерений от ожидания

$$\sum_i \left( \frac{x_i - \mu}{\sigma} \right)^2.$$

**Замечание 7** (Наконец-то  $\chi^2$ ). Пусть у нас есть набор с.в.  $X_i \sim N(\mu, \sigma)$ ; тогда с.в.

$$Y_i = \frac{X_i - \mu}{\sigma} \sim N(0, 1), \quad (1)$$

а сумма

$$Z = \sum_{i=1}^n Y_i^2 \sim \chi_n^2 \quad (2)$$

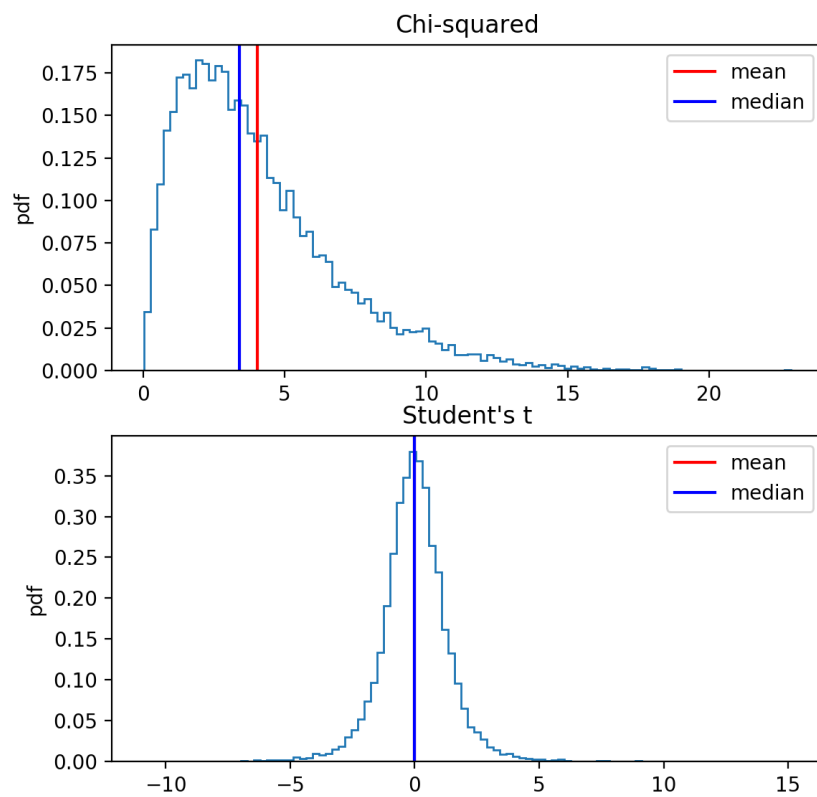


Рис. 3: Примеры асимметричного и heavy-tailed распределений (см. Таблицу 1). Как можно наблюдать, для асимметричного распределения медиана – более хороший эстиматор центральной тенденции, чем среднее; для симметричного распределения они совпадают в пределах погрешности.