Здравствуйте!

Функция распределения и плотность вероятностей

Оценка функции распределения

Рассмотрим теперь вопросы, связанные с оценкой функции распределения случайной величины и проверки гипотез относительно её.

Пусть имеется случайная величина ξ с функцией распределения $F(x)$, которую мы не знаем. Нам необходимо оценить её по опытным данным.

Пусть мы провели n опытов и получили выборку $x_1, x_2, x_3, \ldots, x_n$ объёма n. Тогда в качестве оценки неизвестной функции распределения $F(x)$ берётся так называемая **эмпирическая функция распределения** $F_n(x)$, которая определяется следующим образом:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n}\theta(x - x_i),$$

где

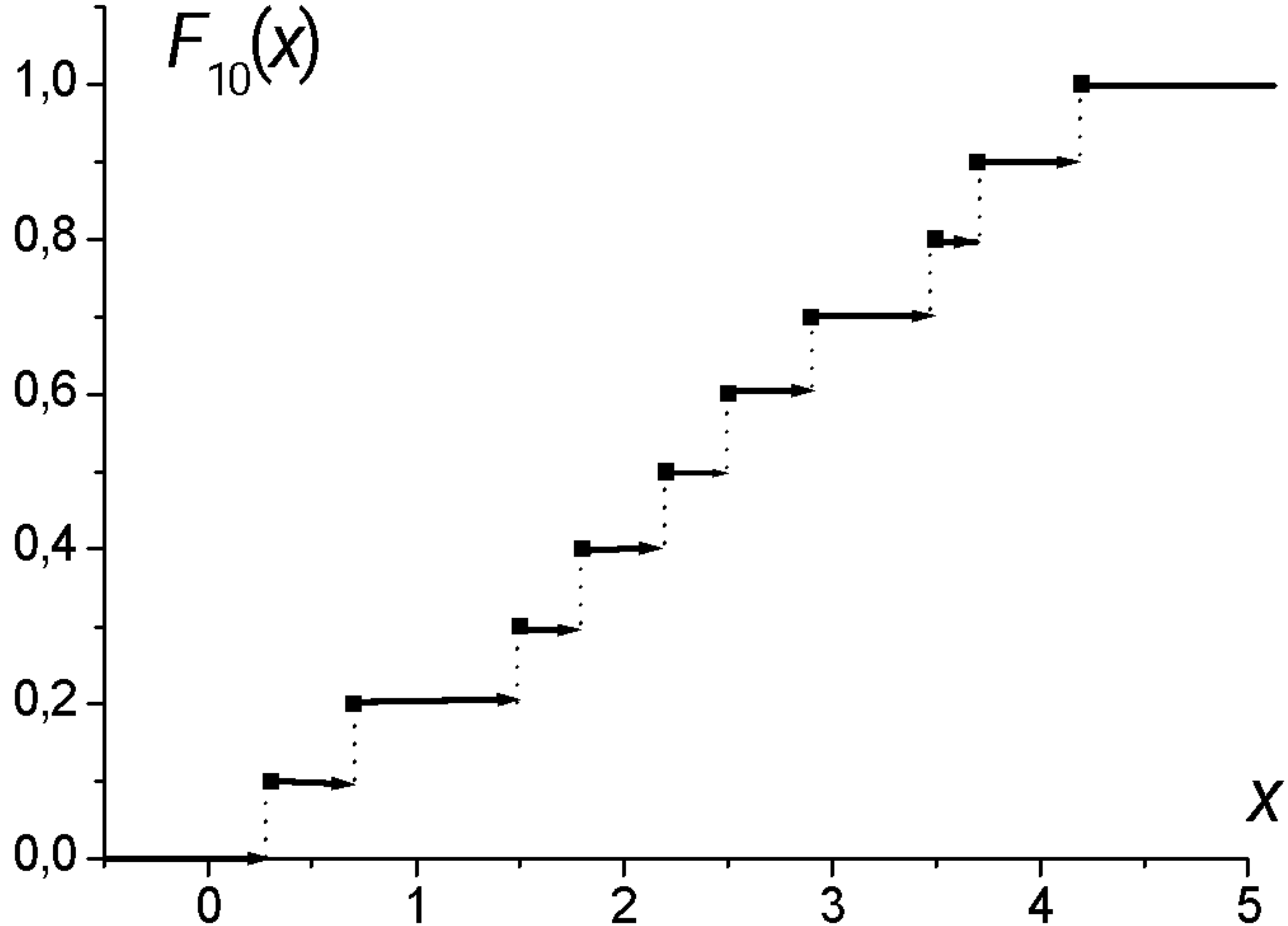
$$\theta(x) = \begin{cases} 1, & \text{если } x \geq 0, \\ 0 & \text{если } x < 0 \end{cases}$$

есть так называемая функция Хевисайда или функция единичного скачка.

Её определяют еще и следующим образом. Представим себе, что мы нашу выборку x_1, x_2, x_3, …, x_n упорядочили **в порядке возрастания** и получили упорядоченную выборку $x_1^{(n)}, x_2^{(n)}, x_3^{(n)}, \ldots, x_n^{(n)}$, так что $\forall i = \overline{1, n-1}$ $x_i^{(n)} \le x_{i+1}^{(n)}$. Тогда $F_n(x)$ есть ступенчатая функция со скачками, равными $1/n$ в точках $x_i^{(n)}$.

Её можно записать и так:

$$F_n(x) = \begin{cases} 0, & \text{если } x < x_1^{(n)}, \\ m/n, & \text{если } x_i^{(n)} \le x < x_{i+1}^{(n)}, 1 \le i \le n-1, \\ 1, & \text{если } x \ge x_n^{(n)}. \end{cases}$$



F10(x)
1,0
0,8
0,6
0,4
0,2
0,0
0
1
2
3
4
5
x

Эта оценка обладает следующими свойствами

$$\mathrm{M}\{F_n(x)\} = F(x),$$

то есть она является несмещенной оценкой, и

$$\mathrm{D}\{F_n(x)\} = \frac{1}{n}F(x)(1-F(x)) \xrightarrow[n\to\infty]{} 0,$$

то есть она сходится, по меньшей мере, в средне квадратичном смысле. Можно показать, что она сходится также и почти наверное.

Критерии Колмогорова и Смирнова

Следующие две популярные задачи связаны с проверкой статистических гипотез.

Первая из них выглядит следующим образом. На основании каких-то соображений мы предполагаем, что функция распределения исследуемой случайной величины есть $F(x)$. Мы провели n опытов и по полученной выборке построили эмпирическую функцию распределения $F_n(x)$. Можно ли сказать, что наши экспериментальные данные не противоречат нашей гипотезе?

Более точно наша гипотеза имеет вид

$$H_0: \quad \forall x \quad \mathrm{M}\{F_n(x)\} = F(x)$$

при альтернативе

$$H_1: \quad \exists x \quad \mathrm{M}\{F_n(x)\} \neq F(x).$$

Для проверки этой гипотезы используют так называемый **критерий Колмогорова**. Согласно ему, для проверки этой гипотезы необходимо вычислить величину

$$D_n = \sup_{|x|<\infty} |F_n(x) - F(x)|.$$

Реально эта величина считается по следующим формулам

$$D_n^+ = \max_{1 \le m \le n} \left(\frac{m}{n} - F(x_m^{(n)}) \right),$$

$$D_n^- = \max_{1 \le m \le n} \left(F(x_m^{(n)}) - \frac{m-1}{n} \right),$$

$$D_n = \max(D_n^+, D_n^-).$$

Решающее правило выглядит следующим образом: если окажется, что

$$\sqrt{n}D_n \le \lambda_\alpha,$$

то следует **принять** гипотезу H_0, то есть считать, что наши опытные данные не противоречат гипотезе о том, что функция распределения исследуемой случайной величины равна $F(x)$; если же выполнено противоположное неравенство, то гипотеза должна быть **отвергнута**, то есть считается, что опытные данные ей противоречат.

Сама величина λ_α даётся следующей таблицей:

α_0	0,05 (5%)	0,01 (1%)	0,001 (0,1%)
λ_α	1,36	1,63	1,95

Другой популярной проблемой является задача о проверке однородности двух выборок. Она ставится следующим образом. Пусть имеется две случайные величины ξ и η с функциями $F(x)$ и $G(x)$ соответственно. Необходимо проверить гипотезу

$$H_0:\ \forall x\ \ F(x) = G(x)$$

при альтернативе

$$H_1:\ \exists x\ \ F(x) \neq G(x)$$

Пусть мы провели две серии опытов и получили две выборки случайных величин ξ и η объёмами m и n соответственно. Представим себе, что по ним мы построили две эмпирические функции распределения $F_m(x)$ и $G_n(x)$. Тогда для проверки сформулированной гипотезы надо найти величину

$$D_{mn} = \sup_{|x|<\infty} |F_m(x) - G_n(x)|,$$

и вычислить

$$\sqrt{\frac{mn}{m+n}} D_{mn}.$$

Если окажется, что эта величина не превосходит λ_α, то следует принять гипотезу H_0, то есть считать, что опытные данные не противоречат тому, что функции распределения в обеих выборках одинаковы; при выполнении противоположного неравенства гипотеза должна быть отвергнута.

Этот критерий носит название критерия Колмогорова–Смирнова.

Оценка плотности вероятностей. Гистограмма

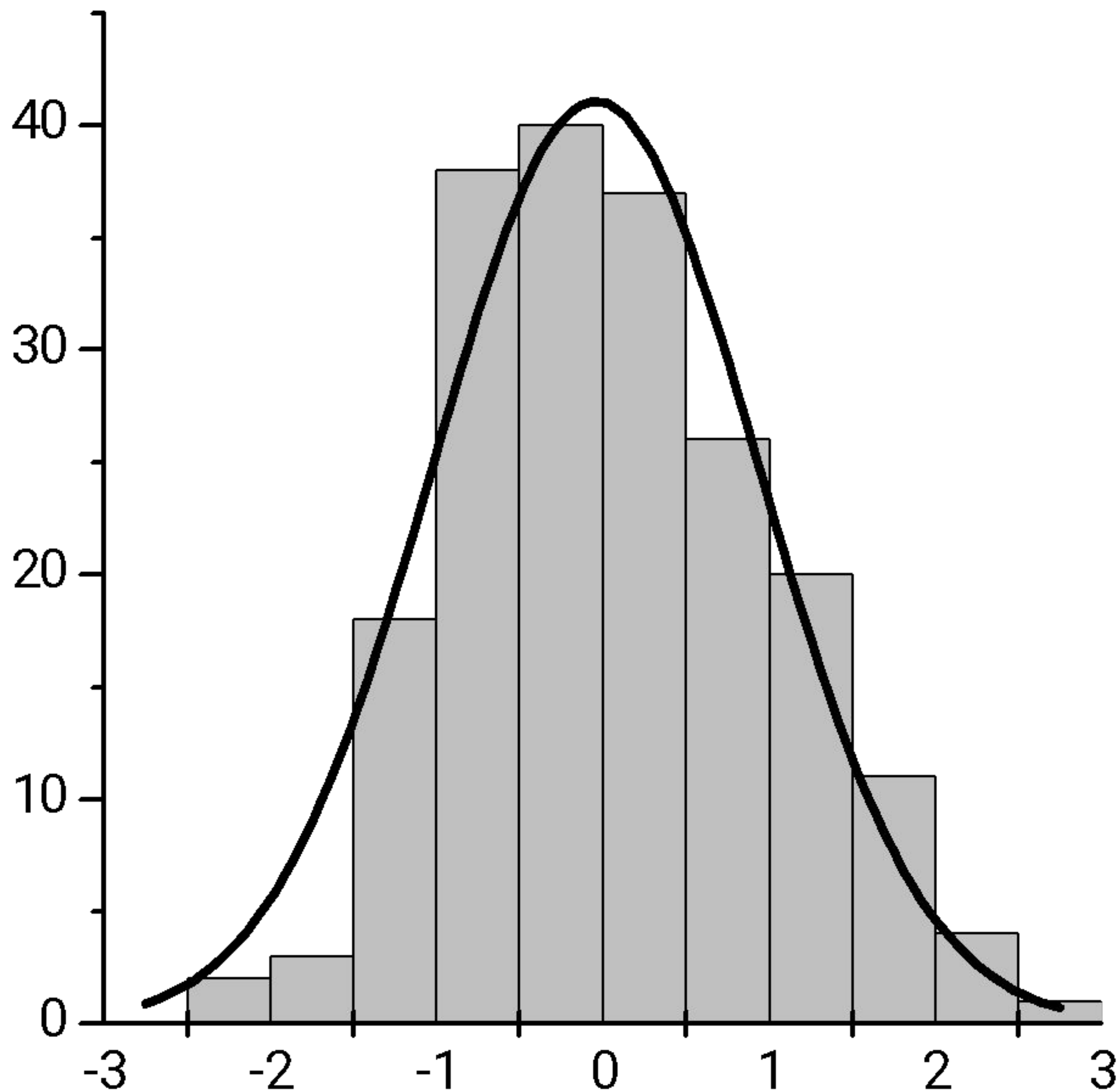
Рассмотрим, наконец, оценку плотности вероятностей случайной величины. Она выполняется с помощью так называемых гистограмм.

Пусть мы имеем выборку $x_1, x_2, x_3, \ldots, x_n$ объёма n. Разобьём всю ось OX на отрезки длиной h. Эту величину рекомендуется находить по формуле

$$h = \frac{x_{\max} - x_{\min}}{1 + 3{,}3\log n},$$

округляя её до ближайшего удобного числа. Здесь $x_{\max}$ ($x_{\min}$) есть максимальное (минимальное) значение из выборочных данных.

Пусть эти отрезки (их обычно называют классами) будут $[a, a+h]$, $[a+h, a+2h]$, … , $[a+(k-1)h, a+kh]$, так что $x_{\min}$ попадает в первый отрезок $[a, a+h]$, а $x_{\max}$ – в последний, k-й отрезок $[a+(k-1)h, a+kh]$. После этого производится так называемая разноска по классам. Она состоит в том, что считается, сколько выборочных значений попало на тот или иной отрезок. Пусть эти числа будут n_1, n_2, …, n_k. Тогда на плоскости XOY строится система прямоугольников с основаниями на выбранных отрезках и площадями, пропорциональными величинам n_1, n_2, …, n_k. Этот рисунок и называется гистограммой.



0
10
20
30
40
-3
-2
-1
0
1
2
3

Эта гистограмма и является некоторым образом плотности вероятностей изучаемой случайной величины. Дальнейшая работа с ней сводится к следующему.

Обычно существуют некоторые соображения о виде этой плотности вероятностей с точностью до некоторого количества s неизвестных параметров. Тогда компьютеру поручается сделать оценку этих неизвестных параметров и построить выравнивающую кривую. В современных пакетах программ для статистической обработки данных обычно имеется большой набор подобных плотностей вероятностей.

Затем проверяется гипотеза о том, что выбранная плотность вероятностей согласуется с выборочными значениями. Для этого считаются вероятности p_i попадания в каждый класс и по ним вычисляется величина

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i},$$

и полученное значение сравнивается с пороговым значением χ^2_α, найденным по таблицам критерия хи-квадрат по заданному уровню значимости α_0 и числу степеней свободы $k - s - 1$. Если будет выполнено неравенство $\chi^2 \le \chi^2_\alpha$, то считается, что выборочные значения не противоречат выбранной плотности вероятностей и работу можно считать законченной. При выполнении противоположного неравенства приходится выравнивать к другой кривой.