# Математическая Статистика

17 марта 2014 г.

функция
распределения!неизвестная
функция
распределения!эмпирическая
функция
распределения!выборочная

# Глава 1

# Основы

## 1.1 Методы оценок характеристик распределения наблюдаемых случайных величин

$x_1, \ldots, x_n$ — независимые одинаково распределённые случайные величины с неизвестной функцией распределения $F$. Логично, что вероятность выпадения каждого $x_k$ (вероятность того, что наугад взятый из выборки $x$ будет равен $x_k$) одинакова

$$P(x = x_k) = \frac{1}{n}$$

Цель — найти $F$ или сказать что-то о её свойствах.

### 1.1.1 Эмпирическая функция распределения

**Определение 1.1.1** (Эмпирическая функция распределения)**.** *Эмпирической (выборочной) функцией распределения, построенной по выборке $x_1, \ldots, x_n$ называется функция*

$$F_n\left(x\right) = \frac{1}{n} \cdot \sum_{k=1}^{n} \mathbb{1}\left(x_k \leq x\right)$$

**Теорема 1.1.1.** *Неизвестная функция распределения $F(x)$ может быть сколь угодно точно восстановлена по выборке достаточно большого объёма [1, стр. 25].*

$$\mathbb{P}\left(F_n\left(x\right) \underset{n\to\infty}{\Longrightarrow} F\left(x\right)\right) = 1$$

*Идея доказательства.* Вспомним, чему равна эмпирическая функция распределения

$$F_n\left(x\right) = \frac{1}{n} \cdot \sum_{k=1}^{n} \mathbb{1}\left(x_k \leq x\right)$$

Заметим, что индикаторы $\mathbb{1}\left(x_k \leq x\right)$ являются независимыми одинаково распределёнными случайными величинами, а функцию распределения $F(x)$ можно записать следующим образом

$$F\left(x\right) = \mathbb{P}\left\{x_1 \leq x\right\} = M\mathbb{1}\left(x_1 \leq x\right)$$

Так как эмпирическая функция распределения является средним арифметическим индикаторов, то по усиленному закону больших чисел она сходится к неизвестной функции распределения почти наверное при устремлении длины выборки к бесконечности

$$F_n(x) = \frac{1}{n} \cdot \sum_{k=1}^{n} \mathbb{1}(x_k \le x) \xrightarrow[n\to\infty]{a.s.} M\mathbb{1}(x_1) = F(x)$$

Теорема доказана

$$F_n(x) \overset{a.s.}{\underset{n\to\infty}{\rightrightarrows}} F(x)$$

□

### 1.1.2 Гистограмма

Как можно попытаться отследить плотность распределения? Постараемся найти функцию распределения, а потом и плотность.

Допустим, $F$ имеет хорошую (непрерывную) плотность. Как тогда из $F$ получить $p$?

Мы знаем, что $F' = p$, но это никому не нужно, так как $F_n'$ — производная ступенчатой функции, которая почти везде будет равна нулю.

Но также мы помним, что

$$F(b) - F(a) = \int\limits_a^b p(x)\,dx$$

Положим $a = x$ и введём $\Delta_x = b - x$

$$F(x+\Delta_x) - F(x) = \int\limits_x^{x+\Delta_x} p(y)\,dy$$

Делим обе части на $\Delta_x$.

$$\frac{1}{\Delta_x} \cdot \int\limits_x^{x+\Delta_x} p(y)\,dy = \frac{F(x+\Delta_x) - F(x)}{\Delta_x}$$

Несложно заметить, что при достаточно малых значениях $\Delta_x$ получаем плотность распределения $p(x)$

$$\frac{\Delta F(x)}{\Delta_x} \xrightarrow{\Delta_x\to 0} \frac{dF(x)}{dx} = p(x)$$

Значит, можем заменить $p(x)$ не производной, а такой разностью.

$$p(x) \approx \frac{F(x+\Delta) - F(x)}{\Delta}$$

Возьмём $m$ полуинтервалов на числовой прямой $I_j = (a_{j-1}, a_j]\,, i = \overline{1,m}$ таких, что все значения выборки попадают в один из них. Для этого определим пару свойств точек, ограничивающих эти интервалы:

1. Каждая следующая точка строго правее (больше) предыдущей. (так как зачем нам одинаковые точки?) гистограмма

$$a_0 < a_1 < \dots < a_m$$

2. Каждое значение выборки должно попадать ровно в один полуинтервал. Очевидно, что данные полуинтервалы $I_j$ не пересекаются между собой. Значит, осталось потребовать, чтобы крайнее левое значение было меньше минимального значения из выборки, а крайнее правое — не больше максимального

$$a_0 < min\left(X\right) \leq max\left(X\right) \leq a_m$$

Введём функцию $q\left(y\right)$

$$q\left(y\right) = \sum_{j=1}^{m} \frac{F\left(a_j\right) - F\left(a_{j-1}\right)}{a_j - a_{j-1}} \cdot \mathbb{1}\left(y \in I_j\right)$$

Определим последовательность функций $q_n\left(y\right)$, заменив $F(x)$ на $F_n\left(x\right)$ в предыдущем определении

$$q_n\left(y\right) = \sum_{j=1}^{m} \frac{F_n\left(a_j\right) - F_n\left(a_{j-1}\right)}{a_j - a_{j-1}} \cdot \mathbb{1}\left(y \in I_j\right) \tag{1.1}$$

Отметим, что $q_n$ сходится к $q$ почти наверное (согласно закону больших чисел), а $q$ в свою очередь сходится к $p$ (согласно центральной предельной теореме)

$$q_n\left(y\right) \xrightarrow[n\to\infty]{a.s.} q\left(y\right) \xrightarrow[m\to\infty]{} p\left(y\right)$$

Функция $q_n$ называется **гистограммой**.

Избавимся от $a_j$ в формуле, а для этого вспомним, чему равно $F_n\left(x\right)$

$$F_n\left(x\right) = \frac{1}{n} \cdot \sum_{k=1}^{n} \mathbb{1}\left(x_k \leq x\right)$$

Теперь посмотрим, чему равна разность $F_n\left(a_j\right) - F_n\left(a_{j-1}\right)$, которая, как мы видим, является вероятностью того, что $x$ попало в отрезок $I_j$

$$F_n\left(a_j\right) - F_n\left(a_{j-1}\right) =$$
$$= \frac{1}{n} \cdot \sum_{k=1}^{n} \mathbb{1}\left(x_k \leq a_j\right) - \frac{1}{n} \cdot \sum_{k=1}^{n} \mathbb{1}\left(x_k \leq a_{j-1}\right)$$

Сгруппируем слагаемые и получим чуть более компактную запись разности

$$F_n\left(a_j\right) - F_n\left(a_{j-1}\right) =$$
$$= \frac{1}{n} \cdot \sum_{k=1}^{n} \left[\mathbb{1}\left(x_k \leq a_j\right) - \mathbb{1}\left(x_k \leq a_{j-1}\right)\right] \tag{1.2}$$

Рассмотрим возможные значения индикаторов

Если оба индикатора равны единице, это значит, что $x_k$ не больше $a_j$ и не больше $a_{j-1}$. Поскольку $a_{j-1} \le a_j$, то можно обойтись тем, что $x \le a_{j-1}$

$$\begin{cases} \mathbb{1}\left(x_k \le a_j\right) = 1 \\ \mathbb{1}\left(x_k \le a_{j-1}\right) = 1 \\ a_{j-1} \le a_j \end{cases} \Rightarrow \begin{cases} x_k \le a_j \\ x_k \le a_{j-1} \\ a_{j-1} \le a_j \end{cases}$$
$$\Rightarrow x_k \le a_{j-1} \le a_j \Rightarrow x_k \le a_{j-1}$$

Такая ситуация, что $x$ больше, чем $a_j$, но не больше, чем $a_{j-1}$, невозможна, так как $a_{j-1}$ не больше, чем $a_j$, а признать возможной такое положение дел $(a_j < x_k \le a_{j-1})$ означало бы то, что $a_j < a_{j-1}$

$$\begin{cases} \mathbb{1}\left(x_k \le a_j\right) = 0 \\ \mathbb{1}\left(x_k \le a_{j-1}\right) = 1 \\ a_{j-1} \le a_j \end{cases} \Rightarrow \begin{cases} x_k > a_j \\ x_k \le a_{j-1} \\ a_{j-1} \le a_j \end{cases}$$
$$\Rightarrow \begin{cases} a_j < x_k \le a_{j-1} \\ a_{j-1} \le a_j \end{cases}$$

Если оба индикатора равны нулю, то это значит, что $x$ строго больше как $a_j$, так и $a_{j-1}$. Опять же, поскольку $a_{j-1} \le a_j$, то достаточно сказать, что $x > a_j$.

$$\begin{cases} \mathbb{1}\left(x_k \le a_j\right) = 0 \\ \mathbb{1}\left(x_k \le a_{j-1}\right) = 0 \\ a_{j-1} \le a_j \end{cases} \Rightarrow \begin{cases} x_k > a_j \\ x_k > a_{j-1} \\ a_j \ge a_{j-1} \end{cases}$$
$$\Rightarrow x_k > a_j \ge a_{j-1} \Rightarrow x_k > a_j$$

Если же $x$ больше, чем $a_{j-1}$, но не больше, чем $a_j$, то $x$ попадает в полуинтервал $(a_{j-1}, a_j]$

$$\begin{cases} \mathbb{1}\left(x_k \le a_j\right) = 1 \\ \mathbb{1}\left(x_k \le a_{j-1}\right) = 0 \\ a_{j-1} \le a_j \end{cases} \Rightarrow \begin{cases} x_k \le a_j \\ x_k > a_{j-1} \\ a_j \ge a_{j-1} \end{cases}$$
$$\Rightarrow a_{j-1} < x_k \le a_j$$

Вспомним формулу (1.2)

$$F_n\left(a_j\right) - F_n\left(a_{j-1}\right) = \\ = \frac{1}{n} \cdot \sum_{k=1}^{n} \left[\mathbb{1}\left(x_k \le a_j\right) - \mathbb{1}\left(x_k \le a_{j-1}\right)\right]$$

Очевидно, что нас интересуют те пары, разность которых не равна нулю. Это значит, что те случаи, когда $x > a_j$ или $x \le a_{j-1}$, нас не интересуют. Поскольку такой случай, что $a_j < x \le a_{j-1}$ невозможен, то его тоже отбросим. Значит, остался только тот вариант, когда $x$ попадает в полуинтервал $(a_{j-1}, a_j]$

$$\frac{1}{n} \cdot \sum_{k=1}^{n} \left[\mathbb{1}\left(x_k \le a_j\right) - \mathbb{1}\left(x_k \le a_{j-1}\right)\right] = \frac{1}{n} \cdot \sum_{k=1}^{n} \mathbb{1}\left(x_k \in (a_{j-1}, a_j]\right)$$

Видим знакомые полуинтервалы $(a_{j-1}, a_j] = I_j$. Воспользуемся этим

$$\frac{1}{n} \cdot \sum_{k=1}^{n} \mathbb{1}\left(x_k \in (a_{j-1}, a_j]\right) = \frac{1}{n} \cdot \sum_{k=1}^{n} \mathbb{1}\left(x_k \in I_j\right)$$

Получаем компактную запись для разности функций распределения

$$F_n\left(a_j\right) - F_n\left(a_{j-1}\right) = \frac{1}{n} \cdot \sum_{k=1}^{n} \mathbb{1}\left(x_k \in I_j\right) \tag{1.3}$$

Вернёмся к уравнению (1.1)

$$q_n\left(y\right) = \sum_{j=1}^{m} \frac{F_n\left(a_j\right) - F_n\left(a_{j-1}\right)}{a_j - a_{j-1}} \cdot \mathbb{1}\left(y \in I_j\right)$$

Воспользовавшись тем, что $(a_j - a_{j-1})$ — длина полуинтервала $I_j$, а разность $F_n\left(a_j\right) - F_n\left(a_{j-1}\right)$ была только что переписана через индикаторы, получаем такую формулу

$$q_n\left(y\right) = \sum_{j=1}^{m} \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}\left(x_k \in I_j\right) \cdot \frac{1}{|I_j|} \cdot \mathbb{1}\left(y \in I_j\right)$$

Упростим, введя функцию $\nu_j\left(X\right)$ [1, стр. 68], которая считает количество элементов выборки $X = x_1, \ldots, x_n$, попавших в интервал $I_j$. Это будет сумма индикаторов того, что элемент $x_k$ попал в $I_j$

$$\nu_j\left(X\right) = \sum_{x \in X} \mathbb{1}\left(x \in I_j\right) = \sum_{k=1}^{n} \mathbb{1}\left(x_k \in I_j\right)$$

Поскольку $\mathbb{1}\left(y \in I_j\right)$ зависит от $j$ и не зависит от $k$, то его можно перенести во внешнюю сумму. Получаем следующую формулу

$$q_n\left(y\right) = \sum_{j=1}^{m} \frac{\mathbb{1}\left(y \in I_j\right)}{n \cdot |I_j|} \cdot \nu_j\left(X\right)$$

У этой суммы только один ненулевой элемент, так как $y$ может попасть только в один полуинтервал. Тогда обозначим номер отрезка, в который попал $y$, через $k$ $(y \in I_k)$, а функцию $q_n\left(y\right)$ запишем как $q_n^k$

$$q_n^k = \frac{\nu_k\left(X\right)}{n \cdot |I_k|} \tag{1.4}$$

Что мы тут видим? Теперь $k$ — номер “столбика” гистограммы (номер интересующего нас полуинтервала — того, в который попал $y$).

“Высота” столбика (значение функции на определённом полуинтервале) пропорциональна количеству элементов, попавших в этот отрезок (что логично). Кроме того, происходит деление на общее количество элементов. Деление нужно, чтобы $q\left(y\right)$ сходилось к $p\left(y\right)$.

Делителю же $|I_k|$ отведена особая роль — он предотвращает искажение гистограммы при различных длинах отрезков. То есть, чем длиннее отрезок, тем ниже столбик, так как элементы более “размазаны” по отрезку, что тоже логично.

неизвестный параметр
статистика
оценка

Представим, что значение функции — это высоту прямоугольника, а длина отрезка — его ширина (графически это изображается именно так). Тогда отношение количества элементов, попавших в полуинтервал, к количеству всех элементов выборки (вероятность того, что случайно взятый элемент из выборки попадёт в $k$-ый отрезок [1, стр. 24]), является площадью прямоугольника

$$S_k = \frac{\nu_k(X)}{n} = \mathbb{P}_n(x \in I_k)$$

Введём замену в формуле (1.4) и умножим обе части на длину отрезка

$$\mathbb{P}_n(x \in I_k) = q_n^k \cdot |I_k|$$

Если устремить количество полуинтервалов к бесконечности ($m \to \infty$), то каждый полуинтервал будет сжиматься в точку. При этом вероятность попадения $x$ в отрезок будет стремиться к вероятности попадения $x$ в точку $y$. Введём обозначения $|I_j| = \delta$, $I_j = \Delta_y$

$$\mathbb{P}_n(x = y) \approx \mathbb{P}_n(x \in \Delta_y) = q_n(y) \cdot \delta, \qquad m \to \infty$$

Очень напоминает ситуацию с плотностью распределения непрерывной случайной величины $\xi$

$$\mathbb{P}(\xi = x) \approx p(x) \cdot \delta, \qquad \delta \to 0$$

Нужно отметить, что количество элементов выборки должно стремиться к бесконечности ($n \to \infty$), так как плотность может быть лишь у непрерывных случайных величин. Чем больше будет элементов, тем плотнее они будут стоять на числовой прямой.

### 1.1.3 Оценка неизвестных параметров

Снова у нас есть $x_1, \ldots, x_n$ — выборка из распределения $F_\theta$, где $\theta$ — неизвестный параметр из множества $\Theta$

**Пример 1.1.1.** *Имеем нормальное распределение с известным СКО $\sigma = 1$ и неизвестным математическим ожиданием $a$ — $N(a, 1)$. Тогда $\theta$ — математическое ожидание $a$*

**Пример 1.1.2.** *Есть нормальное распределение, в котором неизвестны оба параметра. Тогда $\theta$ будет парой $(a, \sigma)$*

Главный вопрос — определение основных параметров распределения выборки.

**Определение 1.1.2** (Статистика)**.** *Статистикой называют функцию $S$ от выборки $X = (x_1, x_2, \ldots, x_n)$*

$$S(X) = S(x_1, x_2, \ldots, x_n)$$

**Определение 1.1.3** (Оценка)**.** *Статистику, значение которой заменяет неизвестный параметр, называют оценкой*

**Пример 1.1.3.** *Предположим, что выборка сделана из распределения Бернулли, то есть $\{x_i\}$ — набор одинаково распределённых случайных величин, причём*

$$x_i = \begin{cases} 1, & p \\ 0, & 1-p \end{cases}$$

*Тогда неизвестный параметр — величина $p$ (вероятность удачного эксперимента)*

$$\theta = p \in [0;1] = \Theta$$

*Введём разные оценки $\hat{p}$*

$$\hat{p}_1 = \frac{1}{n}\sum_{k=1}^{n} x_k$$
$$\hat{p}_2 = x_1$$
$$\hat{p}_3 = \frac{2}{n}\sum_{k=1}^{\left\lfloor \frac{n}{2} \right\rfloor} x_k$$

Замечание: Поскольку $\hat{p}$ — случайная величина, то может оказаться, что она не равна настоящему параметру $p$

$$\mathbb{P}\left\{\hat{p} = p\right\} = 0$$

1. Возникает мысль о том, что разность $\hat{p} - p$ должна быть "маленькой". Например, чтобы $M\left(\hat{p} - p\right)^2$ было самое маленькое из возможных.

2. Также логично желать того, чтобы оценка $\hat{p}$ сходилась к истинному значению параметра $p$ по вероятности ($\hat{p} \xrightarrow[n\to\infty]{\mathbb{P}} p$) или почти всюду ($\hat{p} \xrightarrow[n\to\infty]{a.s.} p$)

3. При многократном повторении эксперимента даже самая (на первый взгляд) плохая оценка может оказаться полезной

$$M\hat{p}_1 = p$$
$$M\hat{p}_2 = p$$
$$M\hat{p}_3 = p$$

   Например, если целый год каждый день дают набор чисел, а статистик считает значение параметра $p$ с помощью оценки $\hat{p}$, то в среднем за год у него получится величина, близкая к истинному $p$.

**Определение 1.1.4** (Состоятельная оценка)**.** *Оценка $\hat{\theta}$ называется состоятельной, если стремится к истинному значению $\theta$ по вероятности*

$$\hat{\theta} \xrightarrow[n\to\infty]{\mathbb{P}} \theta$$

оценка!несмещённая

**Определение 1.1.5** (Сильно состоятельная оценка)**.** *Оценка $\hat{\theta}$ называется сильно состоятельной, если стремится к истинному значению $\theta$ почти наверное*

$$\hat{\theta} \xrightarrow[n\to\infty]{a.s.} \theta$$

**Пример 1.1.4.** *Оценка $\hat{p_1}$ из прошлого примера является сильно состоятельной.*

**Определение 1.1.6** (Несмещённая оценка)**.** *Оценка $\hat{\theta}$ несмещённая, если*

$$\forall \theta \in \Theta : M_\theta \hat{\theta} = \theta$$

**Замечание 1.** *Несмещённая оценка существует не всегда*

**Определение 1.1.7.** *Несмещённая оценка $\hat{\theta} \in K$ называется оптимальной в классе квадратично интегрируемых оценок $K$, если для всякой другой несмещённой оценки $\tilde{\theta} \in K$*

$$D_\theta \hat{\theta} \le D_\theta \tilde{\theta}, \qquad \forall \theta \in \Theta$$

*или же*

$$M_\theta \left(\hat{\theta} - \theta\right)^2 \le M_\theta \left(\tilde{\theta} - \theta\right)^2, \qquad \forall \theta \in \Theta$$

**Замечание 2.** *В учебнике Боровкова А. А. “Математическая статистика” оценка, удовлетворяющая этим условиям, носит название* ***эффективная оценка*** *[1, стр. 130], но у нас этот термин будет использоваться далее в другом смысле*

**Пример 1.1.5.** *Сравним $\hat{p_1}$ и $\hat{p_3}$*

$$D_p \hat{p_1} = \frac{1}{n^2} \cdot n \cdot p \cdot (1-p) = \frac{p \cdot (1-p)}{n}$$
$$D_p \hat{p_3} = \frac{2 \cdot p \cdot (1-p)}{n}$$

### 1.1.4 Выборочные оценки. Метод моментов

Как восстановить неизвестный параметр $\theta \in \Theta$, имея и так далее функцию распределения $F_\theta(x)$?

Вспомним распределения и их параметры

1. Нормальное распределение $N\left(a, \sigma^2\right)$. В нём параметр $a$ является средним, а параметр $\sigma^2$ — дисперсией

2. Пуассоновское распределение $Poi(\lambda)$. Тут параметр $\lambda$ является и средним, и дисперсией

3. Экспоненциальное распределение $Exp(\lambda)$. $\dfrac{1}{\lambda}$ — среднее, $\dfrac{1}{\lambda^2}$ — дисперсия

И так далее... выборочное среднее

Как правило, неизвестный параметр $\theta$ можно искать следующим образом:

$$\exists\varphi \in C\left(\mathbb{R}\right) : \int\limits_{\mathbb{R}} \varphi\left(x\right) dF_\theta\left(x\right) = g\left(\theta\right)$$

Значит, у нас есть уравнение для поиска оценки $\hat{\theta}$

$$g\left(\hat{\theta}\right) = \int\limits_{\mathbb{R}} \varphi\left(x\right) dF_n\left(x\right) \tag{1.5}$$

**Пример 1.1.6.** *Если $\theta$ — среднее, то $\varphi\left(x\right) = x$*

$$\int\limits_{-\infty}^{+\infty} x dF_\theta\left(x\right) = \theta = g\left(\theta\right)$$

**Теорема 1.1.2.** *Пусть функция $\varphi\left(x\right)$ в (1.5) ограничена и строго монотонная. Тогда оценка $\hat{\theta}$ существует и является сильно состоятельной.*

*Доказательство.* Имеем формулу (1.5)

$$g\left(\hat{\theta}\right) = \int\limits_{\mathbb{R}} \varphi\left(x\right) dF_n\left(x\right)$$

Поскольку функция $g\left(\hat{\theta}\right)$ непрерывна и монотонна, то она имеет обратную функцию $g^{-1} : g^{-1}\left(g\left(\hat{\theta}\right)\right) = \hat{\theta}$.

Применим обратную функцию к обеим частям уравнения

$$\hat{\theta} = g^{-1}\left(\int\limits_{\mathbb{R}} \varphi\left(x\right) dF_n\left(x\right)\right)$$

Поскольку выборочная функция распределения почти всюду равна неизвестной функции распределения при достаточно большом объёме выборки, то

$$\int\limits_{\mathbb{R}} \varphi\left(x\right) dF_n\left(x\right) \xrightarrow[n\to\infty]{a.s.} \int\limits_{\mathbb{R}} \varphi\left(x\right) dF_n\left(x\right)$$

Функция $g^{-1}\left(x\right)$ непрерывна

$$\hat{\theta} = g^{-1}\left(\int\limits_{\mathbb{R}} \varphi\left(x\right) dF_n\left(x\right)\right) \xrightarrow[n\to\infty]{a.s.} g^{-1}\left(\int\limits_{\mathbb{R}} \varphi\left(x\right) dF_n\left(x\right)\right) = \theta$$

Теорема доказана

$$\hat{\theta} \xrightarrow[n\to\infty]{a.s.} \theta$$

$\square$

выборочная дисперсия

**Определение 1.1.8** (Выборочное среднее)**.** *Выборочное средние обозначается через $\overline{x}$ и считается по следующей формуле*

$$\overline{x} = \int\limits_{\mathbb{R}} x dF_n(x)$$

*Поскольку все элементы выборки равновероятны, получаем математическое ожидание дискретной равномерно распределённой случайной величины, принимающей $n$ значений*

$$\overline{x} = \int\limits_{\mathbb{R}} x dF_n(x) = \frac{1}{n} \cdot \sum_{k=1}^{n} x_k$$

**Определение 1.1.9** (Выборочная дисперсия)**.** *Выборочная дисперсия $\overline{\sigma^2}$ считается формуле*

$$\overline{\sigma^2} = \int\limits_{\mathbb{R}} (x - \overline{x})^2 dF_n(x) = \frac{1}{n} \cdot \sum_{k=1}^{n} (x_k - \overline{x})^2$$

## 1.2 Свойства оценок

# Литература

[1] Боровков А. А. Математическая статистика. Санкт-Петербург: Лань, 2010. 705 с.

# Предметный указатель

# Оглавление