Математическая Статистика

8 марта 2014 г.

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 наблюдения

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

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

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

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

Доказательство. Вспомним, чему равна эмперическая функция распределения

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

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

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

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

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

□

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

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

Тогда, положив $a = x$, взяв некую Δ, и постановив $b = x + \Delta$, получаем следующее

$$F(x+\Delta) - F(x) = \int_x^{x+\Delta} p(y)\,dy$$

Делим обе части на Δ и при достаточно малых его значениях получаем

$$\frac{1}{\Delta} \cdot \int_x^{x+\Delta} p(y)\,dy = \frac{F(x+\Delta) - F(x)}{\Delta} \approx \frac{dF(x)}{dx} = p(x)$$

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

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

Возьмём выборку из m случайных величин в порядке возрастания $a_1, \ldots, a_m$, обозначим отрезки $I_j = [a_{j-1}, a_j]$ и введём функцию $q(y)$

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

Теперь введём последовательность функций $q_n(y)$ и видим, что она сходится к $q_n(y)$ почти наверное согласно закону больших чисел, а та в свою очередь имеет сходимость порядка $\frac{1}{n}$ к плотности распределения $p(y)$

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

Отметим, что q_n сходится к q почти наверное, а q в свою очередь сходится к p

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

Функция q_n называется **гистограммой**.

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

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

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

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

Сгруппируем слагаемые и получим чуть более компактную запись разности

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

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 $x \le a_{j-1}$

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

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 $a_j < a_{j-1}$

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

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 $x > a_j$.

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

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 $(a_{j-1}, a_j]$

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

Пренебрегаем тем, что у нас полуинтервал, и будем считать, что вероятность попадения x чётко на границу интервала пренебрежимо мала и заменим индикатор на более удобный.

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

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 отрезка I_j, а разность $F_n\left(a_j\right) - F_n\left(a_{j-1}\right)$ была только что компактизирована, 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 формулу. Введём функцию $\nu_j\left(x\right)$ [1, стр. 68], которая считает количество элементов выборки $x_1, \ldots, x_n$, попавших в интервал I_j. Это будет сумма индикаторов того, что элемент x_k попал в интервал 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 отрезок (пренебрегаем возможностью его попадения на границу между двумя отрезками). Тогда обозначим номер отрезка, в который попал y, как k, а функцию $q_n\left(y\right)$ 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 отрезка — номер отрезка, в который попал y).

"Высота" столбика (значение функции на определённом отрезке) пропорциональна количеству элементов, попавших в этот отрезок (что логично). Кроме того, происходит деление на общее количество элементов, которое возникло, 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. Получается, что, чем длиннее отрезок, тем ниже столбик, так как элементы более "размазаны" по отрезку — тоже логично.

Если рассматривать значение функции как высоту прямоугольника, а длину отрезка как его ширину (графически это так и изображается), то оказывается, что отношение количества элементов, попавших в отрезок, к количеству всех элементов выборки (вероятность того, что случайно взятый элемент из выборки попадёт в k-ый отрезок [1, стр. 24]) яется площадью прямоугольника

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

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

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

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

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

Очень напоминает ситуацию с плотностью распределения непрерывной случайной величины ξ

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

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 $x_1, \dots, x_n$ — выборка из распределения F_θ, где θ — неизвестный параметр из множества Θ

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

Пример 1.1.2. *Есть нормальное распределение, в котором неизвестны оба параметра. Тогда θ будет парой (a, σ)*

Главный вопрос — определение основных параметров распределения выборки.

Определение 1.1.2. *Функцию от выборки, значение которой заменяет неизвестный параметр, назвают 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 $\hat{p}$

$$\begin{aligned} \hat{p}_1 &= \frac{1}{n} \sum_{k=1}^{n} x_k \\ \hat{p}_2 &= x_1 \\ \hat{p}_3 &= \frac{2}{n} \sum_{k=1}^{\lfloor \frac{n}{2} \rfloor} x_k \end{aligned}$$

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

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

1. Возникает мысль о том, что разность $\hat{p} - p$ должна быть “маленькой”. Например, чтобы $M(\hat{p} - p)^2$ 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

$$M\hat{p_1} \approx p$$
$$M\hat{p_2} \approx p$$
$$M\hat{p_3} \approx p$$

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

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

Определение 1.1.4 (Сильно состоятельная оценка). *Оценка $\hat{\theta}$ называется сильно состоятельной, если стремится к истинному значению θ почти наверное*

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

Литература

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

Оглавление