Лекция 2. Выборка. Эмпирическая функция. Выборочные моменты

Кохович Дарья Игоревна

12 февраля 2024 г.

Статистическая модель

- $X_1, X_2, \ldots, X_n$ – выборка (n – объем выборки);
- $\mathcal{P} = \{P\}$ – семейство возможных распределений.

Def.

Статистика – любая функция от выборки.

Характеристики выборки

- Эмпирическое среднее:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

- Эмпирическая дисперсия:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X})^2.$$

- Выборочный момент порядка k:

$$X^k = \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

Вариационный ряд

Def.

Пусть $X_1, X_2, \ldots, X_n$ – выборка. Тогда $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ – вариационный ряд из выборки, если

$$X^{(1)} \leq X^{(2)} \leq \cdots \leq X^{(n)}$$

Вариационный ряд

Def.
Пусть $X_1, X_2, \ldots, X_n$ – выборка. Тогда $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ – вариационный ряд из выборки, если

$$X^{(1)} \leq X^{(2)} \leq \cdots \leq X^{(n)}$$

$X^{(1)} = \min(X_i)$, $X^{(n)} = \max(X_i)$.

Def.
$X^{(i)}$ называется порядковой статистикой.

Def.
Статистики, основанные на рангах, называются ранговыми статистиками.

Характеристики выборки

Выборочная медиана:

- если $n = 2k + 1$, то $m_n = X^{(k+1)}$;
- если $n = 2k$, то $m_n = X^{(k+1)}$.

Характеристики выборки

Выборочный квантиль порядка p:

- если np – целое, то $q_{n,p} = X^{(np)}$;
- если np – не целое, то $q_{n,p} = X^{([np]+1)}$.

Эмпирическая функция распределения

- Эмпирическая мера:

$$\nu_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}.$$

- Эмпирическая функция распределения:

$$F_n(r) = \frac{\#\{X_i : X_i \le r\}}{n}, \qquad -\infty < r < \infty.$$

Эмпирическая функция распределения

Если Y имеет распределение равной эмпирической мере (то есть $\mathbb{P}(Y \leq r) = F_n(r)$), то

$$\mathbb{E}Y = \bar{X}, \qquad \mathbb{D}Y = S^2.$$

Гистограмма

Эмпирическим аналогом плотности распределения является так называемая *гистограмма*. Гистограмма строится по группированным данным. Область на прямой, занимаемую элементами выборки, делят на k интервалов. Пусть $A_1, \ldots, A_k$ — интервалы на прямой, называемые интервалами группировки . Обозначим для $j = 1, \ldots, k$ через v_j число элементов выборки, попавших в интервал A_j.

Гистограмма

На каждом из интервалов A_j строят прямоугольник, площадь которого пропорциональна v_j. Общая площадь всех прямоугольников должна равняться единице. Поэтому высота f_j прямоугольника над интервалом A_j равна

$$f_j = \frac{v_j}{nl_j},$$

где через l_j обозначена длина интервала A_j. Полученная фигура, состоящая из объединения прямоугольников, называется **гистограммой**.

Гистограмма

Пример. Имеется вариационный ряд:

$$(0; 1; 1; 2; 2,6; 2,6; 2,6; 3,1; 4,6; 4,6; 6; 6; 7; 9; 9).$$

Разобьём отрезок $[0, 10]$ на четыре равных отрезка. Отрезку $[0, 2, 5)$ принадлежат четыре элемента выборки, отрезку $[2, 5, 5)$ — шесть, отрезку $[5, 7, 5)$ — три, и отрезку $[7, 5, 10]$ — два элемента выборки.

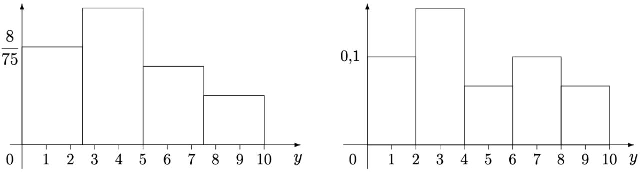

Рис. 1: Гистограммы для $k = 4$ и $k = 5$.

Классы задач мат. статистики

1. Оценки параметров распределения.
   Есть некоторый параметр $\theta : \mathcal{P} \to \mathbb{R}$, необходимо найти оценку $T_n(X_1, X_2, \ldots, X_n) \approx \theta$ при всех $P \in \mathcal{P}$.
2. Проверка гипотез.
   Например: Верно ли, что $\mathcal{P} = \{N(a, \sigma^2)\}$? Дана двумерная выборка, верно ли, что ее координаты независимы?

Несмещенные оценки

Def.

Оценка T_n называется несмещенной для θ, если

$$\mathbb{E}_P T_n(X_1, X_2, \ldots, X_n) = \theta(P) \qquad \forall P \forall n.$$

Проверим несмещенность $\theta(P)$, равной дисперсии.

Несмещенные оценки

Проверим несмещенность $\theta(P)$, равной дисперсии.

$$\mathbb{E}_P S^2 = \mathbb{E}_P\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right) = \frac{1}{n}\cdot n\cdot\left(\mathbb{E}_P(X_i^2)\right) - \mathbb{E}_P\frac{(\sum_{i=1}^{n} X_i)^2}{n} =$$

$$= \mathbb{E}_P(X_i^2) - \frac{1}{n^2}\left[n(n-1)\left(\mathbb{E}_P X_i\right)^2 + n\mathbb{E}_P(X_i)^2\right] =$$

$$= \left(1-\frac{1}{n}\right)\left(\mathbb{E}_P X_i^2 - \left(\mathbb{E}_P X_i\right)^2\right) = \left(1-\frac{1}{n}\right)\mathbb{D}_p X_i.$$

Несмещенные оценки

Таким образом, S^2 – смещенная оценка для дисперсии.

$$S_0^2 = \frac{n}{n-1}S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)$$

несмещенная оценка для дисперсии.

Def.

Оценка T_n называется асимптотически несмещенной, если

$$\lim_{n\to\infty} \mathbb{E}_P T_n(X_1, \ldots, X_n) = \theta(P) \qquad \forall P \in \mathcal{P}.$$

Несмещенные оценки

Утверждение

Эмпирическая функция распределения является несмещенной оценкой для истинной функции распределения F.

Доказательство.

$$\mathbb{E}F_n(r) = \mathbb{E}\frac{\#\{j : X_j \le r\}}{n} = \frac{1}{n}\mathbb{E}\sum_{j=1}^{n} \mathbb{I}_{(X_j \le r)} =$$

$$\frac{1}{n}\sum_{j=1}^{n} \mathbb{P}(X_j \le r) = \frac{1}{n}\sum_{j=1}^{n} F(r) = F(r) \qquad \forall r \forall P \in \mathcal{P}.$$

$\square$

Состоятельность

Def.
Оценка T_n называется состоятельной для θ, если

$$\forall P \forall \epsilon > 0 \quad \mathbb{P}(|T_n(X_1, \ldots, X_n) - \theta(P)| > \epsilon) \to 0 \text{ при } n \to \infty$$

или

$$T_n(X_1, \ldots, X_n) \xrightarrow{\mathbb{P}} \theta(P).$$

Состоятельность

Утверждение

1. $\bar{X}$ – *состоятельная для* $\mathbb{E}X$.
2. S^2, S_0^2 – *состоятельные для* $\mathbb{D}X$.
3. $F_n(r)$ – *состоятельная для* $F(r)$.

Состоятельность

Доказательство.

1. Следует напрямую из ЗБЧ.
2. Из ЗБЧ:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i)^2 \xrightarrow{\mathbb{P}} \mathbb{E}X_i^2; \qquad \bar{X}^2 \xrightarrow{\mathbb{P}} (\mathbb{E}X_i)^2.$$

   Следовательно:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i)^2 - \bar{X}^2 \xrightarrow{\mathbb{P}} \mathbb{E}X_i^2 - (\mathbb{E}X_i)^2 = \mathbb{D}X.$$

3. Также следует из ЗБЧ:

$$F_n = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{(X_i \le r)} \xrightarrow{\mathbb{P}} F(r)\forall r.$$

Теорема Гливенко-Кантелли

Theorem (Гливенко-Кантелли)

$$\mathbb{P}\left(\sup_r |F_n(r) - F(r)| \xrightarrow{n\to\infty} 0\right) = 1.$$

$\forall R \quad F_n(r) \to F(r)$ п.н. по УЗБЧ.
bf Доказательство: Также верно, что $\forall R \quad F_n(r_-) \to F(r_-)$ п.н.:

$$F_n(r_-) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{(X_i<r)} \xrightarrow[\text{УЗБЧ}]{\text{п.н.}} \mathbb{I}_{(X_i<r)} = \mathbb{P}(X_i < r) = F(r_-).$$

Теорема Гливенко-Кантелли

Зафиксируем $\epsilon > 0$. Построим последовательность $r_0 = -\infty < r_1 < \cdots < r_m = \infty$ со следующими свойствами:

$$F(r_{j+1-}) - F(r_j) \leq \epsilon.$$

Можно построить по индукции:

1. $r_0 = -\infty$;
2. $r_{j+1} = \inf\{r \colon F(r) \geq F(r_j) + \epsilon\}$.

Теорема Гливенко-Кантелли

Рассмотрим $t \in [r_j, r_{j+1})$:

$$F_n(t) \geq F_n(r_j) = F_n(r_j) - F(r_j) + F(r_j) \geq F_n(r_j) - F(r_j) + F(t) - \epsilon \Rightarrow$$

$$F_n(t) - F(t) \geq -|F_n(r_j) - F(r_j)| - \epsilon.$$

Аналогично:

$$F_n(t) - F(t) \leq |F_n(r_{j+1-}) - F(r_{j+1-})| + \epsilon.$$

Теорема Гливенко-Кантелли

Получаем:

$$\sup_t |F_n(t) - F(t)| \leq \max |F_n(r_j) - F(r_j)| +$$

$$+ \max |F_n(r_{j-}) - F(r_{j-})| + \epsilon \xrightarrow{\text{п.н.}} \epsilon.$$

Чтд.