Математическая статистика.

Андрей Тищенко @AndrewTGk

2024/2025

Лекция 10 января

Преамбула

Статистика. Мнения о появлении этого слова:

1. Статистиками в Германии назывались люди, собирающие данные о населении и передающие их государству.
2. В определённый день в Венеции народ выстраивался для выплаты налогов (строго фиксированных, в зависимости от рода действий). Государство собирало данные обо всём населении. Это происходило до появления статистиков в Германии, поэтому мы будем считать, что статистика пошла из Венеции.

Задача статистики — по результатам наблюдений построить вероятностную модель наблюдаемой случайной величины.

Основные определения

Определение

Однородной выборкой объёма n называется случайный вектор $X = (X_1, \ldots, X_n)$, компоненты которого являются независимыми и одинаково распределёнными. Элементы вектора X называются элементами выборки.

Определение

Если элементы выборки имеют распределение $F_\xi(x)$, то говорят, что выборка соответствует распределению $F_\xi(x)$ или порождена случайной величиной ξ с распределением $F_\xi(x)$.

Определение

Детерминированный вектор $x = (x_1, \ldots, x_n)$, компоненты которого x_i являются реализациями соответствующих случайных величин X_i $(i = \overline{1, n})$, называется реализацией выборки.

Уточнение

Если X — однородная выборка объёма n, то его реализацией будет вектор x, каждый элемент x_i которого является значением соответствующей ему случайной величины (элемента выборки) X_i.

Определение

Выборочным пространством называется множество всех возможных реализаций выборки

$$X = (X_1, \ldots, X_n)$$

Пример

У вектора $X = (X_1, \dots, X_{10})$ каждый элемент X_i которой порождён случайной величиной $\xi \sim U(0, 1)$, выборочным пространством является $\mathbb{R}^{10}$ (так как X_i может принять любое значение на $\mathbb{R}$)

Определение

Обозначим $x_{(i)}$ — i-ый по возрастанию элемент, тогда будет справедливо:

$$x_{(1)} \leqslant x_{(2)} \leqslant \dots \leqslant x_{(n)}$$

Обозначим $X_{(k)}$ случайную величину, реализация которой при каждой реализации x выборки X принимает значение $x_{(k)}$. Тогда последовательность $X_{(1)}, \dots, X_{(n)}$ называется вариационным рядом выборки.

Определение

Случайная величина $X_{(k)}$ называется k-ой порядковой статистикой выборки.

Определение

Случайные величины $X_{(1)}, X_{(n)}$ называются эстремальными порядковыми статистиками.

Определение

Порядковая статистика $X_{([n \cdot p])}$ называется выборочной квантилью уровня p, где $p \in [0, 1]$

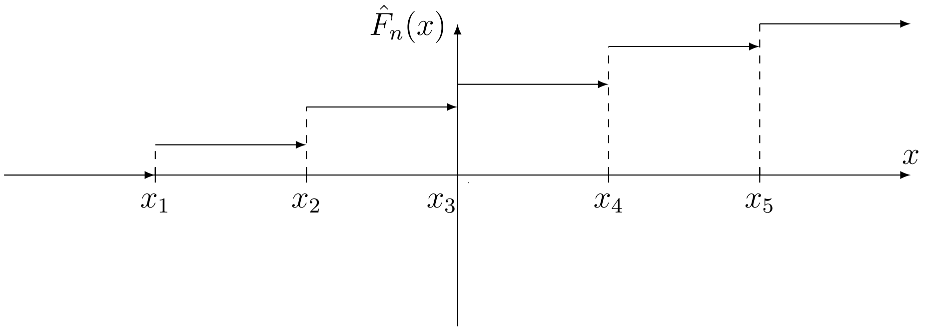
Определение

Пусть каждый элемент выборки X объёма n имеет распределение $F_\xi(x)$. Эмпирической функцией распределения такой выборки называется

$$\hat{F}_n(x) = \frac{1}{n} \sum_{k=1}^{n} I(X_k \leqslant x)$$

I — индикаторная функция. $I = \begin{cases} 1, & \text{если аргумент верен} \\ 0, & \text{иначе} \end{cases}$

Пусть $x_1, \dots, x_n$ — реализация выборки $X_1, \dots, X_n$



Свойства $\hat{F}_n(x)$

1. $\forall x \in \mathbb{R} \quad E\hat{F}_n(x) = E\left(\frac{1}{n} \sum_{k=1}^{n} I(X_k \leqslant x)\right) = \frac{1}{n} \sum_{k=1}^{n} EI(X_k \leqslant x) = P(X_1 \leqslant x) = F_\xi(x)$

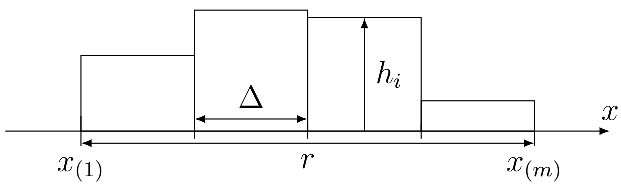
2. По усиленному закону больших чисел (УЗБЧ)

$$\forall x \in \mathbb{R} \quad \hat{F}_n(x) = \frac{1}{n} \sum_{k=1}^{n} I(X_k \leqslant x) \xrightarrow[n \to \infty]{\text{п. н.}} EI(X_k \leqslant x) = F_\xi(x)$$

Гистограмма

Разбить $\mathbb{R}$ на $(m+2)$ непересекающихся интервала. Рассматриваются $x_{(1)}, \ldots,\ x_{(m)}$

Название	Обозначение	Формула
Количество интервалов	m	—
Размах выборки	r	$r = x_{(m)} - x_{(1)}$
Ширина интервала	Δ	$\Delta = \frac{r}{m}$
Количество попаданий на i-ый интервал	ν_i	—
Частота попаданий на i-ый интервал	h_i	$h_i = \frac{\nu_i}{\Delta}$



Лекция 17 января

Определение

Пусть $X_1, \ldots,\ X_n \sim F(x,\ \theta)$. k-ым начальным выборочным моментом называется

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k,\ k \in \mathbb{N}$$

Выборочным средним называется:

$$\hat{\mu}_1 = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Определение

k-ым центральным выборочным моментом называется

$$\hat{\nu}_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^k,\ k = 2,\ 3, \ldots$$

$$\hat{\nu}_2 = S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \text{ называется выборочной дисперсией}$$

Пусть $(x_1,\ y_1), \ldots,\ (x_n,\ y_n)$ соответствует распределению $F(x,\ y,\ \theta)$

Определение

Выборочной ковариацией называется

$$\hat{K}_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

Определение

Выборочным коэффициентом корреляции называется

$$\hat{\rho}_{xy} = \frac{\hat{K}_{xy}}{\sqrt{S_x^2 S_y^2}}$$

Свойства выборочных моментов

1. $E\hat{\mu}_k = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^k\right) = \frac{1}{n}\sum_{i=1}^{n} EX_i^k = EX_1^k = \mu_k$
2. $E\overline{X} = m_x$
3. $\mathcal{D}\hat{\mu}_k = \mathcal{D}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^k\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathcal{D}X_i^k = \frac{1}{n}\mathcal{D}X_i^k = \frac{1}{n}\left(EX_1^{2k} - (EX_1^K)^2\right) = \frac{1}{n}(\mu_{2k} - \mu_k^2)$
4. $\mathcal{D}\overline{x} = \frac{\sigma_{x_1}^2}{n}$
5. По УЗБЧ
$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k \xrightarrow[n\to\infty]{\text{п. н.}} E\hat{\mu}_k = \mu_k$$
$$\hat{\nu}_k \xrightarrow[n\to\infty]{\text{п. н.}} \nu_k$$
6. По ЦПТ
$$\frac{\hat{\mu}_k - \mu_k}{\sqrt{\frac{\mu_{2k} - \mu_k^2}{n}}} \xrightarrow[d]{n\to\infty} U,\ U \sim N(0,\ 1)$$
$$\frac{\sqrt{n}(\overline{x} - m_{x_1})}{\sigma} \xrightarrow[n\to\infty]{d} U$$
7. $ES^2 = E\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right) = \frac{n-1}{n}\sigma^2$
8. $E\hat{K}_{xy} = \frac{n-1}{n}\operatorname{cov}(x,\ y)$

Определение

Оценкой $\hat{\theta}$ параметра θ, называется функция:

$$\hat{\theta} = T(x_1, \ldots,\ x_n),\ \text{не зависящая от } \theta$$

Например, отвратительная оценка среднего роста людей в аудитории.

$$\hat{m} = \frac{2x_2 + 5x_5 + 10x_{10}}{3}$$

Определение

Оценка $\hat{\theta}$ называется несмещённой, если $E\hat{\theta} = \theta$ для любых возможных значений этого параметра.

Определение

Оценка $\hat{\theta}(x_1, \ldots,\ x_n)$ называется асимптотически несмещённой оценкой θ, если

$$\lim_{n\to\infty} E\hat{\theta}(x_1, \ldots,\ x_n) = \theta$$

$$\lim_{n\to\infty} ES^2 = \lim_{n\to\infty}\frac{n-1}{n}\sigma^2 = \sigma^2$$

Определение

Несмещённой выборочной (или исправленной) выборочной дисперсией называется

$$\tilde{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Оценки

$$\hat{m}_1 = \frac{x_1 + x_2 + x_3}{3}$$
$$\hat{m}_2 = \frac{\sum_{i=1}^{10} x_i}{10}$$
$$\hat{m}_3 = \overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Являются несмещёнными.

Определение

Оценка $\hat{\theta}(x_1, \ldots,\ x_n)$ называется:
Состоятельной оценкой θ, если

$$\hat{\theta}(x_1, \ldots,\ x_n) \xrightarrow[n\to\infty]{p} \theta$$

Сильно состоятельной оценкой, если

$$\hat{\theta}(x_1, \ldots,\ x_n) \xrightarrow[n\to\infty]{\text{п. н.}} \theta$$

Определение

Пусть $\hat{\theta}$ — несмещённая оценка параметра θ. Если $\mathcal{D}\hat{\theta} \leqslant \mathcal{D}\theta^*$, где θ^* — любая несмещённая оценка параметра θ. Тогда $\hat{\theta}$ называется эффективной оценкой параметра θ.

R-эффективные оценки

Рассматриваем выборку $X_1, \ldots,\ X_n \sim f(x,\ \theta),\ \theta \in \Theta \subseteq \mathbb{R}^1$. Назовём модель $(S,\ f(x,\ \theta))$ регулярной, если она удовлетворяет следующим условиям:

1. $\forall x \in S \quad$ функция $f(x,\ \theta) = f(x_1, \ldots,\ x_n,\ \theta) > 0$ и дифференцируема по θ.
2. $\dfrac{\delta}{\delta\theta}\displaystyle\int_S f(x,\ \theta)\,dx = \int_S \frac{\delta}{\delta\theta} f(x,\ \theta)\,dx = \frac{\delta}{\delta\theta}\int_S T(x)f(x,\ \theta)\,dx = \int_S \frac{\delta}{\delta\theta}T(x)f(x,\ \theta)\,dx$

Пусть $\hat{\theta} = T(x) = T(x_1, \ldots,\ x_n)$ — несмещённая оценка параметра θ:

$$\int_S \frac{\delta}{\delta\theta} f(x,\ \theta)\,dx = 0,\ \text{так как не зависит от } \theta$$

$$\int_S \frac{\delta}{\delta\theta}T(x)f(x,\ \theta)\,dx = \frac{\delta}{\delta\theta}\theta = 1$$

Определение

Информацией Фишера о параметре θ, содержащейся в выборке $X_1, \ldots,\ X_n$ называется величина

$$I_n(\theta) = E\left(\frac{\delta \ln f(X,\ \theta)}{\delta\theta}\right)^2 = \int_S \left(\frac{\delta \ln f(x,\ \theta)}{\delta\theta}\right) f(x,\ \theta)\,dx$$

Неравенство Рао-Крамера

Если $S,\ f(x,\ \theta)$ — регулярная модель и $\hat{\theta}$ — несмещённая оценка θ, то

$$\mathcal{D}\hat{\theta} \geqslant \frac{1}{I_n(\theta)}$$

Докажем это неравенство.

Неравенство Коши-Буняковского

$$\left(\int \varphi_1(x)\varphi_2(x)\,dx\right)^2 \leqslant \int \varphi_1^2(x)\,dx \int \varphi_2^2(x)\,dx$$

Пользуясь этим:

$$\int\limits_S \frac{\delta}{\delta\theta} f(x,\ \theta)\,dx = \int\limits_S \frac{\delta f(x,\ \theta)}{\delta\theta}\frac{f(x,\ \theta)}{f(x,\ \theta)}\,dx = \int\limits_S \frac{\delta \ln f(x,\ \theta)}{\delta x} f(x,\ \theta)\,dx = 0$$

$$\int\limits_S \frac{\delta}{\delta\theta} T(x) f(x,\ \theta)\,dx = \int\limits_S T(x)\frac{\delta}{\delta\theta} f(x,\ \theta)\frac{f(x,\ \theta)}{f(x,\ \theta)}\,dx = \int\limits_S T(x)\frac{\delta \ln f(x,\ \theta)}{\delta x} f(x,\ \theta)\,dx = 1$$

Применяя неравенство Коши-Бунякоского:

$$1 = \int \big(T(x)-\theta\big)\frac{\delta \ln f(x,\ \theta)}{\delta\theta} f(x,\ \theta)\,dx \leqslant \underbrace{\int\limits_S \big(T(x)-\theta\big)^2 f(x,\ \theta)\,dx}_{=\mathcal{D}\hat{\theta}} \underbrace{\int\limits_S \left(\frac{\delta \ln f(x,\ \theta)}{\delta\theta}\right)^2 f(x,\ \theta)}_{I_n(\theta)}$$

Получили

$$1 \leqslant \mathcal{D}\theta I_n(\theta) \Rightarrow \mathcal{D}\theta \geqslant \frac{1}{I_n(\theta)}$$

Определение

Оценка $\hat{\theta}$ называется R-эффективной, если $E\hat{\theta} = \theta$ и $\mathcal{D}\hat{\theta} = \frac{1}{I_n(\theta)}$