Математическая Статистика

Лекции

Автор конспектов: Чубий Савва Андреевич

Преподаватель: Горяинова Елена Рудольфовна

2024–2025

2025-01-10

Введение

- Теория вероятностей: по известной модели ищем параметры
- Математическая статистика: по наблюдаемой величине строим модель

Определения

Опр. Однородной выборкой объема N называется случайный вектор $X = (X_1, ..., X_n)$, компоненты которого — независимые и одинокого распределенные.

Примечание: в ближайшее время будем обсуждать только **однородные** выборки.

Опр. Компоненты однородной выборки называются **элементами выборки**.

Опр. Если все элементы выборки $X_1, ..., X_n$ выборки имеют распределение $F_\xi(x)$, то говорят, что **выборка соответствует распределению $F_\xi(x)$ или выборка порождена случайной величиной ξ**.

Примечание: обычно $F_\xi(x)$ мы не знаем. Часто известно только семейство распределения: например, «гауссово» или «непрерывное».

Опр. Детерминированный вектор $x = (x_1, ..., x_n)$, где X_i есть реализация СВ $X_i, i = \overline{1...n}$ называется **реализацией выборки X**.

Опр. Выборочным пространством S называется множество всех возможных реализаций выборки.

Опр. Пара $(S, \mathcal{F})$, где $\mathcal{F}$ – семейство распределений, порождающих X, называется **моделью**.

Опр. Упорядочим реализацию $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$.

Пусть СВ X_k есть такой элемент выборки, реализация которого, который при любой реализации $X_1, ..., X_n$, принимает значение $X_{(k)}$.

Тогда пос-ть $X_{(1)}, ..., X_{(n)}$ называется **вариационным рядом** выборки, а $X_{(k)}$ — k-ой порядковой статистикой.

Опр. Порядковые статистики $X_{(1)}$ и $X_{(n)}$ называются **экстремальными**.

Когда мы зафиксируем конкретную реализацию, $X_{(i)}$ станет числом, до этого $X_{(i)}$ — СВ.

Эмпирическая функция распределения

Опр. Пусть $X_1, ..., X_n$ соответствует распределению $F_\xi(x)$, тогда функция:

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

называется **эмпирической функцией распределения**.

<u>Свойства</u>

- $E\hat{F}_n(x) = P(x_k \leq x) = F_\xi(x)$
- По УЗБЧ:

$$\frac{1}{n} \sum_{k=1}^{n} I(x_k \leq x) \xrightarrow[n \to \infty]{\text{п. н.}} F_\xi(x)$$

Гистограмма

$x_1, \ldots, x_n$ — реализация

Для построения гистограммы:
1. Разбить R^1 на $m+2$ не пересекающихся интервала
2. Обычно рассматривают **размах** выборки $r = x_{(n)} - x_{(1)}$
3. Первый и последний интервалы $(\left(-\infty, x_{(1)}\right)$ и $\left(x_{(n)}, +\infty\right))$ пустые, остальные — длинны равной $\Delta = \frac{r}{m}$
4. Для каждого интервала вычисляем частоту попаданию в него: ν_k
5. На каждом интервале строим прямоугольник высотой $h_k = \frac{\nu_k}{\Delta}$

При увеличении m гистограмма стремится к плотности.

Гистограмма ростов

$$x_{(1)} = 163$$
$$x_{(39)} = 203$$
$$m = 5$$
$$r = 203 - 163 = 40$$
$$\Delta = 8$$

Интервал	**Частота**	**Высота**
$[163, 171)$	$\frac{8}{39}$	…
$[171, 179)$	$\frac{11}{39}$	…
$[179, 187)$	$\frac{18}{39}$	…
$[187, 195)$	$\frac{1}{39}$	…
$[195, 203)$	$\frac{1}{39}$	…

100 наблюдений = 7 интервалов

2025-01-17

Выборочные моменты

$$X_1, \ldots, X_n \sim F(x, \theta)$$

Опр. k-ым начальным моментом называется:

$$\hat{\mu_k} = 1_n \sum_{i=1}^{n} x_i^k$$

Опр. Выборочное среднее:

$$\hat{\mu_1} = \overline{x}$$

Опр. k-ым центральным выборочным моментом называется

$$\hat{\nu_k} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^k$$

Опр. Выборочная дисперсия: $s^2 := \hat{\nu_2}$

Опр. Пусть $(x_1, y_1), ..., (x_n, y_n) \sim F(x, y, \theta)$

Выборочная ковариация:

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

Опр. Коэффициент выборочной ковариации:

$$\widehat{\rho_{xy}} = \frac{\widehat{K_{xy}}}{\sqrt{S_x^2 S_y^2}}$$

Свойства выборочных моментов

- $$E\widehat{\mu_k} = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i^k\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(x_i^k\right) = Ex_1^k = \mu_k$$
  $$E\overline{x} = m_{x_i}$$
- $$D\widehat{\mu_k} = \frac{1}{n^2}\sum_{i=1}^{n} D\left(x_i^k\right) = \frac{Dx_1^k}{n} = \frac{1}{n}\left(Ex_1^{2k} - \left(Ex_1^k\right)^2\right) = \frac{1}{n}(\mu_{2k} - \mu_k^2)$$
  Чем больше выборка, тем меньше дисперсия
- $$D\overline{x} = \frac{\sigma^2 x}{n}$$
- По УЗБЧ:
  $$\widehat{\mu_k} = \frac{1}{n}\sum_{i=1}^{n} x_i^k \underset{n\to\infty}{\overset{\text{п.н.}}{\to}} E\widehat{\mu_k} = \mu_k$$
- По УЗБЧ:
  $$\widehat{\nu_k} \overset{\text{п.н.}}{\to} \nu_k$$
- По ЦПТ:
  $$\frac{\widehat{\mu_k} - \mu_k}{\sqrt{\frac{\mu_{2k} - \mu_k^2}{n}}}) \underset{n\to\infty}{\overset{P}{\to}} U \sim N(0; 1)$$
  $$\frac{\sqrt{n}\left(\overline{x} - m_{x_1}\right)}{\sigma} \underset{n\to\infty}{\overset{d}{\to}} U$$
- $$Es^2 = E\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2\right) = \frac{n-1}{n}\sigma^2$$
- $$E\widehat{K_{xy}} = \frac{n-1}{n}\operatorname{cov}(x, y)$$

Оценка параметра

Опр. Оценкой $\widehat{\theta}$ параметра θ называется функция $\theta = T(x_1, ..., x_n)$, не зависящая от θ.

Опр. Оценка $\theta\hat{}$ называется **несмещенной**, если $E\theta\hat{} = \theta$, если $\forall\theta \in \Theta$.

Опр. Оценка $\theta\hat{}(x_1, ..., x_n)$ называется **асимптотически несмещенной**, если $\lim_{n\to\infty} E\theta\hat{}(x_1, ..., x_n) = \theta$.

Опр. Несмещенная (исправленная) выборочная дисперсия:

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

Опр. Оценка $\theta\hat{}$ называется **состоятельной**, если

$$\theta\hat{}(x_1, ..., x_n) \underset{n\to\infty}{\overset{P}{\to}} \theta$$

Опр. Оценка $\theta\hat{}$ называется **сильно состоятельной**, если

$$\theta\hat{}(x_1, ..., x_n) \underset{n\to\infty}{\overset{\text{п.н.}}{\to}} \theta$$

Опр. Эффективная оценка — оценка с минимальной дисперсией среди несмещенных.

R-эффективные оценки

$$x_1, ..., x_n \sim f(x, \theta), \theta \in \Theta \subset R^1$$

Опр. Назовем модель $(S, f(x, \theta))$ регулярной, если

1. функция $f(x, \theta) = f(x_1, ..., x_n, \theta) > 0 \forall x \in s$ и дифференцируема по θ
2. $$\frac{\partial}{\partial\theta}\int_\delta f(x, \theta)dx = \int_s \frac{\partial}{\partial\theta} f(x, \theta)dx$$

$$\frac{\partial}{\partial\theta}\int_s T(x)f(x, \theta)dx = \int_s \frac{\partial}{\partial\theta} T(x)f(x, \theta)dx$$

Пусть $T(x) = T(x_1, ..., x_n)$ — несмещенная оценка параметра θ

$$0 = \int_s \frac{\partial}{\partial\theta} f(x, \theta)dx$$

$$1 = \int_s \frac{\partial}{\partial\theta} T(x)f(x, \theta)dx$$

Опр. Информация Фишера о параметре θ, содержащейся в выборке, называет величина

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

Теорема Рао-Крамера

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

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

Опр. Несмещенная оценка называется **R-эффективной**, если её дисперсия равная $\frac{1}{I_n(\theta)}$.

Неравенство Коши-Буянковского

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

Док-во

Рао-Крамера

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

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

$$1 = \int (T(x)-\theta)\frac{\partial \ln f(x,\theta)}{\partial\theta} f(x,\theta)dx = \left\{\begin{matrix} \varphi_1(x) = (T(x)-\theta)\sqrt{f(x,\theta)} \\ \varphi_2(x) = \dfrac{\partial \ln f(x,\theta)}{\partial\theta}\sqrt{f(x,\theta)} \end{matrix}\right\} \leq$$

$$\leq \int (T(x)-\theta)^2 f(x,\theta)dx \int \left(\frac{\partial \ln f(x,\theta)}{\partial\theta}\right)^2 f(x,\theta) = D\theta\hat{}\, I_n(\theta)$$