Лекции по Математической Статистике 5 семестр

Ilya Yaroshevskiy

13 мая 2023 г.

Оглавление

Лекция 1

1.1 Введение

Определение. Генеральная совокупность — множество всех исходов определенного всех эксперимента

Определение. Выборочная совокупность — множество исходов наблюдаемых экспериментов

Примечание. Выборка *репрезентативная* если ее распределение совпадает с распределением генеральной совокупности

Определение. 1 Пусть проведено n наблюдаемых независимых экспериментов в которых наблюдаемые величины приняли значения: $X_1, X_2, \ldots, X_n$. Набор этих данных называется **выборкой объема** n

Определение. 2 **Выборкой объема** n называется набор из n независимых олинакого распределенных случайных величин

1.2 Выборочные характеристики

Пусть имеется выборка в смысле 1. Ее можно понимать как следующую дискретную случайную величину:

X_1	X_2	X_3	$\ldots$	X_n	$\sum$
p_i	$\frac{1}{n}$	$\frac{1}{n}$	$\ldots$	$\frac{1}{n}$	1

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

— **матожидание**?

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

— **дисперсия**

$$F_n^*(z) = \frac{1}{n} \sum_{i=1}^{n} I(X_i < z)$$

— **функция распределения**, где $I(X < z) = \begin{cases} 1 & , X < z \\ 0 & , X_i \geq z \end{cases}$ — индикатор

Теорема 1.2.1. $\forall z \in \mathbb{R}$

$$F_n^*(z) \xrightarrow[n\to\infty]{p} F(z)$$

Доказательство.

$$EI(X_1 < z) = 1 \cdot p(X < z) + 0 \cdot p(X_1 \geq z) = p(X_1 < z) = F(z)$$

— функция распределения X_1
По ЗБЧ Хнчина:

$$F_n^*(z) = \frac{\sum I(X_i < z)}{n} \xrightarrow{p} EI(X_1 < z) = F(z)$$

□

Примечание. На самом деле имеется даже равномерная сходимость по вероятности: *теорема Гливенко-Кантеля*:

$$\sup_{z\in\mathbb{R}} |F_n^*(z) - F(z)| \xrightarrow{p} [n \to \infty]0$$

1.3 Первоначальная обработка

Определение. Если упорядочить выборку по возрастанию(ранжирование) то получим **вариационный ряд**:

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$$

Если учесть повторяющиеся экземпляры, получим **частотный вариационный ряд**

$X_{(1)}$	$X_{(2)}$	$\dots$	$X_{(k)}$	$\sum$
n_1	n_2	$\dots$	n_k	n
p_1	$\frac{n_1}{n}$	$\dots$	$\frac{n_k}{n}$	1

- $h = X_{\max - X_{\min}}$ — **размах выборки**

Примечание. Допустим разбили интервал ($X_{\min, X_{\max})}$ на k интервалов, чаще всего одинаковой длинны. Тогда $h = \frac{h}{k}$. Тогда вариационный ряд можно заменить интервальным вариационным рядом.

Пример.

i	l_1	l_2	$\dots$	l_k	$\sum$
m_i	m_2	m_2	$\dots$	m_k	n
$\frac{m_i}{n}$	$\frac{m_1}{m}$	$\frac{m_2}{m}$	$\dots$	$\frac{m_k}{m}$	1

На координатной плоскости построим прямоугольники:

- l_i — основание прямоугольника соотвествующего интервала
- $\frac{m_i}{nl_i}$ — высота прямоугольника

Получаем ступенчатую фигуру площади 1, которая называется **гистограмма**.

Теорема 1.3.1. При

- $n \to \infty$

- $k(n) \to \infty$
- $\frac{k(n)}{n} \to 0$

Гистограмма по вероятности будет стремиться к плотности распределения:

$$\frac{m_i}{n} \xrightarrow{p} p(X_i \in l_i) = \int_{l_i} f(x)dx$$

Примечание. Чаще всего число интервалов берется по формулу Стержеса.

$$k \approx 1 + \log_k n$$

$$k \approx \sqrt[3]{n}$$

Примечание. Иногда выборка изображаеся в виде полигона

- l_i^m — середина i-го интервала

Лекция 2

$$\mathbb{X} = (\mathbb{X}_1, \ldots, \mathbb{X}_n)$$

Определение. Статистикой называется измеримая функция:

$$\Theta^* = \Theta^*(\mathbb{X}_1, \ldots, \mathbb{X}_n)$$

Примечание. Пусть требуется найти оценку параметра Θ случайную величины $\mathbb{X}$ по данной выборке. Оценку будет считать с помощьью некоторой статистики:

$$\Theta^* = \Theta^*(\mathbb{X}_1, \ldots, \mathbb{X}_n)$$

2.1 Свойства статистических оценок

1. Состоятельность

   Определение. Статистика Θ^* называется **состоятельной оценкой** параметра Θ, если

   $$\Theta^* \xrightarrow{p} \Theta \text{ при } n \to \infty$$

2. Несмещенность

   Определение. Статистика Θ^* называется несмещенной оценкой параметра Θ, если

   $$E\Theta^* = \Theta$$

   Примечание. С равной вероятностью можем ошибиться как в меньшую так и в большую сторону. Нет систематической ошибки.

   Определение. Статистика Θ^* называется **асимптотически несмещенной**, если

   $$\Theta^* \xrightarrow[n\to\infty]{} \Theta$$

3. Эффективность

   Определение. Оценка Θ_1^* **не хуже** Θ_2^*, если

   $$E(\Theta_1^* - \Theta) \leq E(\Theta_2^* - \Theta)$$

   или оценки несмещенные

   $$D\Theta_1^* \leq D\Theta_2^*$$

   Определение. Оценка Θ^* называется **эффективной** если она не хуже всех остальных оценок.

   Теорема 2.1.1. Не существует эффективной оценки в классе всех возможных оценок

   Теорема 2.1.2. В классе не смещенных оценок существует эффективная оценка

2.2 Точечные оценки моментов

Определение. Выборочным средним $\overline{\mathbb{X}_c}$ называется величина равная среднему арифметическому данных

$$\overline{\mathbb{X}_c} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{X}_i$$

Определение. Выборочная дисперсией D_c называется величина

$$D_c = \frac{1}{n}\sum_{i=1}^{n}(\mathbb{X}_1 - \mathbb{X}_c)^2$$

Определение. Исправленной выборочной дисперсией S^2 называется величина

$$S^2 = \frac{n}{n-1}D_c$$

Определение. Выборочным средним квадратическим отклонением σ_c называется велиична:

$$\sigma_c = \sqrt{D_c}$$

Определение. Исправленным выборочным средним квадратическим отклонением называется величина:

$$S = \sqrt{S^2}$$

Определение. Выборочным k-ым моментом называется величина

$$\overline{\mathbb{X}^k} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{X}_i^k$$

Определение. Модой M_0^* называется величина с наибольшей частотой

Определение. Медианой Me^* вариационного ряда называется значение варианты в середине ряда. Если число четное, то берем среднее от средних.

Теорема 2.2.1. $\overline{\mathbb{X}_c}$ является несмещенной состоятельной оценкой для математического ожидания

1. $E\overline{\mathbb{X}_c} = E\mathbb{X} = a$ — несмещенность
2. $\overline{\mathbb{X}_c} \xrightarrow{p} [n \to \infty] E\mathbb{X} = a$

Доказательство. Доделать □

Теорема 2.2.2. $\overline{\mathbb{X}^k}$ является несмещенной состоятельной оценкой для теоретического k-го момента

1. $E\overline{\mathbb{X}^k} = E\mathbb{X}^k = m_k$ — несмещенность
2. $\overline{\mathbb{X}^k} \xrightarrow{p} E\mathbb{X}^K = m_k$ — состоятельность

Теорема 2.2.3. D_c и S^2 являются состоятельными оценками для дисперсии

1. D_c — смещенная оценка
2. S^2 — несмещенная оценка

Доказательство. Доделать □

Примечание. Пусть имеется выборка $(\mathbb{X}_1, \dots, \mathbb{X}_n)$ неизвестного распределения, знаем тип распределения. Пусть этот тип определяется k неизвестными параметрами. Теоретическое распределение задает теоретический k-тый момент. *Например, если распределение непрерывное, то оно задается плотность*:

$$f(\mathbb{X}, \Theta_1, \dots, \Theta_k) \text{ и } m_k = \int_{-\infty}^{\infty} \mathbb{X}^k f(\mathbb{X}, \Theta_1, \dots, \Theta_k) d\mathbb{X} = f_k(\Theta_1, \dots, \Theta_k)$$

Метод моментов состоит в следующем: вычисляем выборочные моменты и подставляем их в уравнение вместо теоретичских. В результате получаем систему уравнений:

$$\begin{cases} \overline{\mathbb{X}} = f_1(\Theta_1, \dots, \Theta_k) \\ \overline{\mathbb{X}}^2 = f_2(\Theta_1, \dots, \Theta_k) \\ \vdots \\ \overline{\mathbb{X}}^k = f_k(\Theta_1, \dots, \Theta_k) \end{cases}$$

Решив эту систему получим оценки параметров $\Theta_1, \dots, \Theta_k$. При этом как правило получаем оценки состоятельные но смещенные

Пример. Пусть случайная величина $\mathbb{X} \in U(a, b)$, $a < b$. Обработав стат. данные получили оценки:

$$\overline{\mathbb{X}} = 2.25 \quad \overline{X}^2 = 6.75$$

Решение.

$$f(x) = \begin{cases} 0 & , x < a \\ \frac{1}{b-a} & , a \le x \le b \\ 0 & , x > b \end{cases}$$

$$E\mathbb{X} = \int_a^b \mathbb{X} f(\mathbb{X}) d\mathbb{X} = \dots = \frac{a+b}{2}$$

$$E\mathbb{X}^2 = \int_a^b \mathbb{X}^2 \frac{1}{b-a} d\mathbb{X} = \dots = \frac{a^2+ab+b^2}{3}$$

$$\begin{cases} 2.25 = \frac{a+b}{2} \\ 6.75 = \frac{a^2+ab+b^2}{3} \end{cases} \Leftrightarrow \dots \Leftrightarrow \begin{cases} b = 4.5 \\ a = 0 \end{cases}$$

Лекция 3

3.1 Метод масимального правдоподобия

Состоит в том, чтобы подобрать параметр таким образом, чтобы вероятность получения данной выборки была наибольшая. Если распределение дискретное, то вероятность выборки

$$P_\theta(\mathbb{X}_1 = x_1, \mathbb{X}_2 = x_2, \dots, \mathbb{X}_n = x_n) = P_\theta(\mathbb{X}_1 = x_1)P_\theta(\mathbb{X}_2 = x_2)\dots P_\theta(\mathbb{X}_n = x_n)$$

Определение. Функцией правдоподобия $L(\overline{X}, \theta)$ называется функция

$$P_\theta(\mathbb{X}_1 = x_1)\dots P_\theta(\mathbb{X}_n = x_n) \text{ — в случае дискретного распределения}$$

$$f_\theta(x_1)\dots f_\theta(x_n) = \prod_{i=1}^{n} f_\theta(x_n) \text{ — в сулчае абсолютно непрерывного распределения}$$

Определение. Логарифмической функцией правдоподобия $M(\overline{X})$ называется функция

$$\ln L(\overline{X})$$

Примечание. Так как логарифм – возрастающая функция, то экстремумы этих функций совпадают.

Определение. Оценкой максимального правдоподобия $\hat{\Theta}$ называется значение Θ при котором функция правдоподобия достигает наибольшего значения при фиксированных значениях выборки x_i

Пример. Пусть $\mathbb{X}_1, \dots, \mathbb{X}_n$ — выборка неизвестного распределения Пуассона с параметром λ.

$$P(\mathbb{X} = x_i) = \frac{\lambda^{x_i}}{x_i!} \cdot e^{-\lambda}$$

$$L(\overline{X}) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} \cdot e^{-\lambda} =$$

$$\frac{n \cdot \overline{X}}{\lambda} - n = 0 \implies \lambda = \overline{X}$$

$$\hat{\Theta} = \overline{X}$$

Пример. $f_{a,\sigma^2} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{\sigma^2}}$

$$a = \overline{X} \quad \sigma^2 = D_c$$

Пример. Пусть $\mathbb{X}_1, \dots, \mathbb{X}_n$ — выборка равномерного распределения вида. $X \in U(0, \Theta)$, $\Theta > 0$. Найти оценки Θ методами моментов и максимального правдоподобия и сравнить их.

$$EX = \frac{a+b}{2} = \frac{\Theta}{2} \implies \hat{\Theta} = 2\overline{X}$$

$$f_\Theta(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{\Theta} & 0 \le x \le \Theta \\ 0 & x > \Theta \end{cases}$$

$$L(\overline{X}, \Theta) = \prod_{i=1}^{n} f_\Theta(x) = \begin{cases} 0 & \Theta < \max(x_i) = \mathbb{X}_m \\ \frac{1}{\Theta^n} & \text{иначе} \end{cases}$$

Ясно, что функция $L(\overline{X}, \Theta)$ достигает наибольшего значения при $\Theta = \mathbb{X}_m$

Доделать

Примечание. ОМП часто эффективны, но могут быть смещенные.

3.2 Неравенство Раво-Крамера

Пусть известно что случайная величина $\mathbb{X} \in F_\Theta$ — семейство распределений с параметром Θ

Определение. Носителем семейства распределений F_Θ называется множество $C \in \mathbb{R}$ такое что $\forall \Theta: \ P(\mathbb{X} \in C) = 1$

Обозначение.

$$f_\Theta(x) = \begin{cases} f_\Theta(x) & \text{плотность, если распределение абсолютно непрерывное} \\ P_\Theta(\mathbb{X} = x) & \text{если распределние дискретное} \end{cases}$$

Определение. Информацией Фишера называется величина

$$I(\Theta) = E(\frac{\partial}{\partial \Theta} \ln f_\Theta(x))^2$$

при условии если она существует.

Определение. Семейство распределений F_Θ называется **регулярным**, если:

1. Существует носитель C семейства F_Θ такой, что $\forall x \in C$ функция $\ln f_\Theta(x)$ — непрерывно дифференцируема по Θ
2. Информация Фишера $I(\Theta)$ существует и непрерывна по Θ

Теорема 3.2.1 (Неравенство Раво-Крамера)**.** Пусть $(\mathbb{X}_1, \dots, \mathbb{X}_n)$ — выборка объема n из регулярного семейства распределений F_Θ, $\Theta^* = \Theta^*(\mathbb{X}_1, \dots, \mathbb{X}_n)$ — несмещенная оценка параметра Θ, дисперсия которой ограничена на любом компакте в области Θ. Тогда

$$D\Theta^* \ge \frac{1}{nI(\Theta)}$$

Пример. Пусть $(\mathbb{X}_1, \dots, \mathbb{X}_n)$ — $\mathbb{X} \in N(a, \sigma^2),\ a \in \mathbb{R}, \sigma^2 > 0$. Проверим эффективность оценки $a^* = \overline{X}$.

$$f(x) = \frac{\sigma\sqrt{2\pi}}{e}^{-\frac{(x-a)^2}{\sigma^2}}$$

Возьмем $C = (-\infty, +\infty)$.

$$\ln f(x) = \ln \sigma - \frac{1}{2} \ln - \frac{(x-a)^2}{\sigma^2}$$

$$\frac{\partial}{\partial a} \ln f(x) = \frac{x-a}{\sigma}$$

— непрерывна по a, $\forall a \in \mathbb{R}$.

$$I(a) = E(\frac{\partial}{\partial a} \ln f(x))' = \frac{1}{\sigma^2}$$

— непрерывна по a

$$Da^* = D\overline{X} = \frac{1}{a} D\mathbb{X} = \frac{\sigma^2}{a}$$

$$Da^* = \frac{\sigma^2}{a} = \frac{1}{nI(a)} = \frac{1}{n \cdot \frac{1}{\sigma^2}} = \frac{\sigma^2}{a}$$

$\Longrightarrow a^* = \overline{X}$

Примечание. Исправленная дисперсия S^2 также является эффективной оценкой для σ^2.
BLUE - оценки

$$\Theta^* = \alpha_1 X_1 + \alpha_2 X_2 + \dots + \alpha_n X_n$$

TODO Лекция 4

TODO Лекция 5

TODO Лекция 6

TODO Лекция 7

Лекция 9

8.1 Математическая регрессия

Определение. Регрессией X на Z называется функция $f(z) = E(X|Z = z)$.

$$x = f(z)$$

— **уравнение регрессии**, а ее график — **линия регрессии**

Пусть при n экспериментах, при значениях $z_1, \ldots, z_n$ случайной величины Z наблюдаемые значения $x_1, \ldots, x_n$ случайной величины X.

Обозначим ε разницу между наблюдаемыми и экспериментальными значениями X

$$\varepsilon_i = X_i - E(X|Z = z_i) = X_i - f(z_i)$$

Тогда $x_i = f(z_i) + \varepsilon_i$, $1 \le i \le n$, и ε_i — независимые случайные нормальные величины. $\varepsilon_i \in N(0, \sigma^2)$

$$a = 0, \text{ т.к. } E\varepsilon_i = E(X_i) - E(f(z_i)) = EX_i - E(X|Z = z) = 0$$

Цель: по данным значениям как можно более точно оценить функцию $f(z)$

Примечание. При этом предполагается(например из теории), что $f(z)$ — функция определенного типа, параметры которой не известны

8.1.1 Метод наименьшие квадратов

Метод наименьших квадратов состоит в выборе параметров функции $f(z)$, таким образом, чтобы минимизировать сумму квадратов ошибок

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (X - f(z))^2 \to \min$$

Определение. Пусть $\Theta = (\Theta_1, \ldots, \Theta_n)$ — набор параметров функции $f(z)$.

$$\hat{\Theta} \quad \min \sum_{i=1}^{n} \varepsilon_i^2$$

8.2 Линейная парная регрессия

Пусть имеется *линейная регрессия*:

$$f(z) = a + bz$$

Тогда $X_i = a + bz_i + \varepsilon_i$, $1 \le i \le n$. Найдем оценки неизвестных параметров a и b методом наименьших квадратов (МНК).

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (X_i - a - bz_i)^2 \to min$$

$$\frac{\partial}{\partial a} \sum_{i=1}^{n} \varepsilon_i^2 = 2\sum_{i=1}^{n} (X_i - a - bz_i) \cdot (-1) = -2\sum_{i=1}^{n} X_i + 2\sum_{i=1}^{n} a + 2b\sum_{i=1}^{n} z_i = -2(n\overline{X} - na - bn\overline{z}) = 0$$

$$\frac{\partial}{\partial b} \sum_{i=1}^{n} \varepsilon_i^2 = 2\sum_{i=1}^{n} (X_i - a - bz_i) \cdot (-z_i) = -2\left(\sum_{i=1}^{n} X_i z_i - a\sum_{i=1}^{n} z_i - b\sum_{i=1]^{n} z_i^2}\right) = -2(n\overline{Xz} - an\overline{z} - bn\overline{z}) = 0$$

$$\begin{cases} n\overline{X} - na - bn\overline{z} = 0 \\ n\overline{Xz} - an\overline{z} - bn\overline{z^2} \end{cases} \Leftrightarrow \begin{cases} \overline{X} - a - b\overline{z} = 0 \\ \overline{Xz} - z\overline{z} - b\overline{z^2} \end{cases} \Leftrightarrow \begin{cases} a + b\overline{z} = \overline{X} \\ a\overline{z} + b\overline{z^2} = \overline{Xz} \end{cases}$$

$$\begin{cases} a = \overline{X} - b\overline{z} \\ (\overline{Z} - b\overline{z})\overline{z} + b\overline{z^2} = \overline{Xz} \end{cases} \Leftrightarrow \begin{cases} a = \overline{x} - b\overline{z} \\ b(\overline{z^2} - \overline{z}^2) = \overline{Xz} - \overline{X} \cdot \overline{z} \end{cases} \Leftrightarrow \begin{cases} a = \overline{X} - b\overline{z} \\ b = \frac{\overline{Xz} - \overline{X} \cdot \overline{z}}{\overline{z^2} - \overline{z}^2} \end{cases}$$

$$\begin{aligned} \overline{X_z} &= E(\overline{X}|Z = z) = f(z) \\ \overline{X_z} &= a + bz \\ \overline{X_z} &= \overline{X} - b\overline{z} + b\overline{z} \\ \overline{X_z} - \overline{X} &= b(z - \overline{z}) \\ \overline{X_z} - \overline{X} &= \frac{\overline{Xz} - \overline{X} \cdot \overline{z}}{\hat{\sigma_z}^2} \cdot (z - \overline{z}) \\ \overline{X_z} - \overline{X} &= \frac{\overline{Zx} - \overline{X} \cdot \overline{z}}{\hat{\sigma_z} \cdot \hat{\sigma_x}} \cdot \frac{\hat{\sigma}_x}{\hat{\sigma_z}} \cdot (z - \overline{z}) \\ \frac{\overline{X_z} - \overline{X}}{\hat{\sigma_x}} &= \frac{\overline{Xz} - \overline{x} \cdot \overline{z}}{\hat{\sigma_z} \cdot \hat{\sigma_x}} \cdot \frac{z - \overline{z}}{\hat{\sigma_z}} \end{aligned}$$

$$\frac{\overline{X_z} - \overline{X}}{\hat{\sigma_x}} = r_B \cdot \frac{\overline{z} - z}{\hat{\sigma_z}}$$

, где $r_B = \frac{\overline{Xz} - \overline{X} \cdot \overline{z}}{\hat{\sigma_z} \cdot \hat{\sigma_x}}$

8.2.1 Геометрический смысл прямой линейной регрессии

Прямая регрессии строится таким образом, чтобы сумма квадратов длин отрезков была наименьшей.

Определение. Выборочный коэффициент линейной корреляции

$$r_B = \frac{\overline{Xz} - \overline{X} \cdot \overline{z}}{\hat{\sigma_x} \cdot \hat{\sigma_y}}$$

Примечание. Отсюда видим, что выборочный коэффициент линейной корреляции является оценкой теоретического коэффициента полученного по методу моментов. Поэтому выборный коэффициент корреляции характеризует силу линейной связи между двумя случайными величинами. Знак r_B показывается является ли корреляция прямой или обратной

8.2.2 Проверка гипотезы по значимости коэффициента линейной корреляции

Пусть двумерная случайная величины распределена нормально. По выборке объема n вычислен r_B, а r — теоретический коэффициент линейной корреляции.

- $H_0 : r = 0$
- против $H_1 : r \neq 0$, т.е. коэффициент r_B статистически значимый

Теорема 8.2.1. Если H_0 верна, то статистика

$$K = \frac{r_B\sqrt{n-z}}{\sqrt{q - r_0^2}} \in T_{n-1}$$

— распределение Стьюдента с $n-2$ степенями свободы
Критерий: Пусть $t_c r$ — квантиль T_{n-2} уравнения значимости α

8.2.3 Выборочное корреляционное отношение

Пусть имеется k выборок случайной величины X, полученных при уровнях фактора $z_1, \ldots, z_k$. Вычислены общая дисперсия D_O, D_M — межгрупповая дисперсия, D_B — внутригрупповая дисперсия. По теореме о разложении дисперсии $D_O = D_M + D_B$ Доделать