Прикладная статистика. Теория оценивания I

Леонид Иосипой

Программа «Математика для анализа данных»
Центр непрерывного образования, ВШЭ

10 ноября 2020

- Организационная информация
- Теория оценивания
- Сравнение оценок
- Методы построения оценок

Организационная информация

1. Время занятий/Перерывы
2. Теория и практика
3. Домашние задания
4. Сайт

Теория оценивания

Анализируемые данные обычно рассматриваются как реализация выборки из некоторого распределения, известного с точностью до параметра (или нескольких параметров).

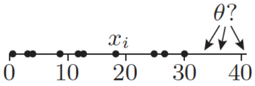
При таком подходе для определения распределения, наиболее подходящего для описания данных, достаточно уметь оценивать значение параметра по реализации выборки.

Теория оценивания

Эксперимент. Пусть θ — неизвестное положительное число. Ниже приведены координаты x_i десяти точек, взятых из равномерного распределения на отрезке $[0, \theta]$.

3.5 3.2 25.6 8.8 11.6 26.6 18.2 0.4 12.3 30.1

Попробуйте угадать значение параметра θ.



Теория оценивания

С формальной точки зрения мы имеем дело со следующей моделью: набор x_i — это реализация независимых и равномерно распределенных на отрезке $[0, \theta]$ случайных величин X_i с функцией распределения

$$F_\theta(x) = \begin{cases} 0, & \text{если } x \le 0, \\ x/\theta, & \text{если } 0 < x < \theta, \\ 1, & \text{если } x \ge \theta. \end{cases}$$

Здесь $\theta \in \Theta = (0, +\infty)$ — неизвестный параметр.

Теория оценивания

Постановка задачи.

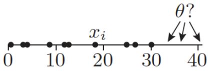
- В общем случае задается семейство функций распределения $\{F_\theta(x), \theta \in \Theta\}$, где Θ — множество возможных значений параметра.
- Данные $x_1, \ldots, x_n$ рассматриваются как реализация выборки $X_1, \ldots, X_n$, элементы которой имеют функцию распределения $F_{\theta_0}(x)$ при некотором неизвестном значении $\theta_0 \in \Theta$.
- Задача состоит в том, чтобы оценить (восстановить) θ_0 по выборке $x_1, \ldots, x_n$, по возможности, наиболее точно.

Теория оценивания

Будем оценивать θ_0 при помощи некоторых функций $\widehat{\theta}$ от n переменных $x_1, \ldots, x_n$, которые называются оценками или статистиками.

Подставляя в оценку $\widehat{\theta}$ реализацию выборки $x_1, \ldots, x_n$, мы получим число — оценку неизвестного параметра θ_0.

Теория оценивания



3.5 3.2 25.6 8.8 11.6 26.6 18.2 0.4 12.3 30.1

Для приведенного выше эксперимента в качестве оценок неизвестного параметра θ можно использовать:

1. $\widehat{\theta}_1(x_1, \ldots, x_n) = 35$;
2. $\widehat{\theta}_2(x_1, \ldots, x_n) = 2x_7$;
3. $\widehat{\theta}_3(x_1, \ldots, x_n) = \max\{x_1, \ldots, x_n\}$;
4. $\widehat{\theta}_4(x_1, \ldots, x_n) = 2\,\dfrac{x_1 + \ldots + x_n}{n}$.

Теория оценивания

- Какая из этих оценок точнее?
- Каким образом вообще можно сравнивать оценки?

Несмещенность

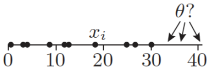
Оценка $\widehat{\theta}(x_1, \ldots, x_n)$ параметра θ называется несмещенной, если

$$\mathbb{E}_\theta \left[\widehat{\theta}(X_1, \ldots, X_n)\right] = \theta \text{ для всех } \theta \in \Theta.$$

Здесь индекс θ у математического ожидания $\mathbb{E}_\theta$ означает, что имеется в виду математическое ожидание случайной величины $\widehat{\theta}(X_1, \ldots, X_n)$, где X_i распределены с функцией распределения $F_\theta(x)$. В дальнейшем этот индекс будет опускаться, чтобы формулы не выглядели слишком громоздко.

Несмещенность означает, что при многократном вычислении оценки для разных данных среднее арифметическое полученных оценок будет стремится к истинному значению параметра θ.

Несмещенность



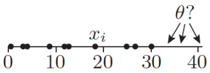
1. Является ли $\widehat{\theta}_1(x_1, \ldots, x_n) = 35$ несмещенной?

Нет, не является, так как

$$\mathbb{E}_\theta\left[\widehat{\theta}_1(X_1, \ldots, X_n)\right] = \mathbb{E}_\theta[35] = 35.$$

Но $35 \neq \theta$ для всех $\theta \in (0; +\infty)$.

Несмещенность

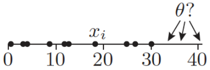


2. Является ли $\widehat{\theta}_2(x_1, \ldots, x_n) = 2x_7$ несмещенной?

Да, является, так как для произвольного $\theta \in (0; +\infty)$

$$\mathbb{E}_\theta \left[\widehat{\theta}_2(X_1, \ldots, X_n)\right] = \mathbb{E}_\theta[2X_7] = 2 \cdot \frac{\theta}{2} = \theta.$$

Несмещенность



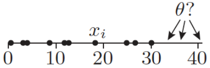
3. Является ли $\widehat{\theta}_3(x_1, \dots, x_n) = \max\{x_1, \dots, x_n\}$ несмещенной?

Нет, не является, так как, ввиду того, что $\mathbb{P}(X_i = \theta) = 0$ для всех $i = 1, \dots, n$, $\max\{X_1, \dots, X_n\} < \theta$ с вероятностью 1.

Следовательно,

$$\mathbb{E}_\theta\left[\widehat{\theta}_3(X_1, \dots, X_n)\right] = \mathbb{E}_\theta[\max\{X_1, \dots, X_n\}] < \mathbb{E}_\theta[\theta] = \theta.$$

Несмещенность



4. Является ли $\widehat{\theta}_4(x_1, \ldots, x_n) = 2\,\dfrac{x_1 + \ldots + x_n}{n}$ несмещенной?

Да, является, так как для произвольного $\theta \in (0; +\infty)$

$$\begin{aligned}
\mathbb{E}_\theta\left[\widehat{\theta}_4(X_1, \ldots, X_n)\right] &= \mathbb{E}_\theta\left[2\,\frac{X_1 + \ldots + X_n}{n}\right] \\
&= \frac{2}{n} \cdot \mathbb{E}_\theta[X_1 + \ldots + X_n] \\
&= \frac{2}{n} \cdot n \cdot \frac{\theta}{2} \\
&= \theta.
\end{aligned}$$

Несмещенность

Результаты проверки на несмещенность:

	$\widehat{\theta}_1$	$\widehat{\theta}_2$	$\widehat{\theta}_3$	$\widehat{\theta}_4$
Несмещенность	-	+	-	+

$$\widehat{\theta}_1(x_1, \ldots, x_n) = 35;$$
$$\widehat{\theta}_2(x_1, \ldots, x_n) = 2x_7;$$
$$\widehat{\theta}_3(x_1, \ldots, x_n) = \max\{x_1, \ldots, x_n\};$$
$$\widehat{\theta}_4(x_1, \ldots, x_n) = 2\,\frac{x_1 + \ldots + x_n}{n}.$$

Состоятельность

Оценка $\widehat{\theta}(x_1, \ldots, x_n)$ параметра θ называется состоятельной, если для всех $\theta \in \Theta$

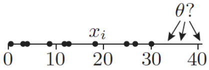
$$\widehat{\theta}(X_1, \ldots, X_n) \xrightarrow{\mathbb{P}_\theta} \theta \quad \text{при } n \to \infty.$$

Здесь $\xrightarrow{\mathbb{P}_\theta}$ обозначает «сходимость по вероятности»:

$$\text{для любого } \varepsilon > 0 \qquad \mathbb{P}_\theta\left(\left|\widehat{\theta}_n - \theta\right| > \varepsilon\right) \to 0 \text{ при } n \to \infty.$$

Состоятельность оценки означает концентрацию оценки около истинного значения параметра с ростом размера выборки n (что устремив $n \to \infty$, оценка сойдется к истинному значению параметра θ).

Состоятельность

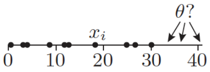


1. Является ли $\widehat{\theta}_1(x_1, \ldots, x_n) = 35$ состоятельной?

Нет, не является, так как, например, для $\theta = 36$ и $\varepsilon = 0.1$

$$\mathbb{P}\left(\left|\widehat{\theta}_1(X_1, \ldots, X_n) - \theta\right| > \varepsilon\right) = \mathbb{P}\left(\left|36 - 35\right| > 0.1\right) = 1 \not\to 0.$$

Состоятельность

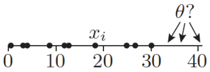


2. Является ли $\widehat{\theta}_2(x_1, \ldots, x_n) = 2x_7$ состоятельной?

Нет, не является, так как для произвольного $\varepsilon < \theta$

$$\begin{aligned}\mathbb{P}\left(\left|\widehat{\theta}_2(X_1, \ldots, X_n) - \theta\right| > \varepsilon\right) &= \mathbb{P}\left(\left|2X_7 - \theta\right| > \varepsilon\right) \\ &= \mathbb{P}\left(X_7 < \frac{\theta - \varepsilon}{2} \text{ или } X_7 > \frac{\theta + \varepsilon}{2}\right) \\ &= Const \not\to 0.\end{aligned}$$

Состоятельность



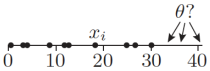
3. Является ли $\widehat{\theta}_3(x_1, \ldots, x_n) = \max\{x_1, \ldots, x_n\}$ состоятельной?

Да, является, так как

$$\begin{aligned}
\mathbb{P}\left(\left|\widehat{\theta}_3(X_1, \ldots, X_n) - \theta\right| > \varepsilon\right) &= \mathbb{P}\left(\left|\max\{X_1, \ldots, X_n\} - \theta\right| > \varepsilon\right) \\
&= \mathbb{P}\left(\max\{X_1, \ldots, X_n\} < \theta - \varepsilon\right) \\
&= \mathbb{P}\left(X_i < \theta - \varepsilon \text{ для всех } i = 1, \ldots, n\right) \\
&= \left(\frac{\theta - \varepsilon}{\theta}\right)^n \to 0,
\end{aligned}$$

так как число в скобке строго меньше 1.

Состоятельность



4. Является ли $\widehat{\theta}_4(x_1, \ldots, x_n) = 2\,\frac{x_1 + \ldots + x_n}{n}$ состоятельной?

Да, является, так как, согласно закону больших чисел,

$$\frac{X_1 + \ldots + X_n}{n} \xrightarrow{\mathbb{P}_\theta} \mathbb{E}_\theta[X] = \frac{\theta}{2}.$$

Следовательно,

$$2\,\frac{X_1 + \ldots + X_n}{n} \xrightarrow{\mathbb{P}_\theta} \theta.$$

Состоятельность

Результаты проверки на несмещенность и состоятельность:

	$\widehat{\theta}_1$	$\widehat{\theta}_2$	$\widehat{\theta}_3$	$\widehat{\theta}_4$
Несмещенность	-	+	-	+
Состоятельность	-	-	+	+

$$\widehat{\theta}_1(x_1, \ldots, x_n) = 35;$$
$$\widehat{\theta}_2(x_1, \ldots, x_n) = 2x_7;$$
$$\widehat{\theta}_3(x_1, \ldots, x_n) = \max\{x_1, \ldots, x_n\};$$
$$\widehat{\theta}_4(x_1, \ldots, x_n) = 2\,\frac{x_1 + \ldots + x_n}{n}.$$

Построение оценок

Основная идея: чтобы оценить d неизвестных параметров модели, нам необходимо составить d уравнений на них.

Чтобы упростить формулировки и обозначения, мы будем считать, что неизвестный параметр многомерный: $\theta \in \Theta \subset \mathbb{R}^d$

Метод моментов

Метод моментов: d уравнений на неизвестные параметры получаются приравниваем первых d теоретических моментов к их эмпирическим аналогам.

Метод моментов

Пусть дана реализация выборки $x_1, \dots, x_n$ из некоторого распределения X с неизвестным параметром θ.

(Теоретическим) моментом k-го порядка случайной величины X называется величина

$$A_k = \mathbb{E}X^k.$$

Выборочным моментом k-го порядка случайной величины X называется величина

$$a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k.$$

Метод моментов

В методе моментов в качестве уравнений на неизвестные параметры берутся следующие уравнения:

$$\begin{cases} A_1 = a_1, \\ A_2 = a_2, \\ \vdots \\ A_d = a_d. \end{cases} \iff \begin{cases} \mathbb{E}X = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i, \\ \mathbb{E}X^2 = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i^2, \\ \vdots \\ \mathbb{E}X^d = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i^d. \end{cases}$$

Теоретический момент зависит от неизвестных параметров модели, а выборочный момент — от известных нам данных.

Метод моментов

Эти уравнения имеют смысл, так как если моменты вплоть до порядка d существуют, то в силу закона больших чисел

$$a_k \xrightarrow{\mathbb{P}} A_k, \quad k = 1, \ldots, d.$$

Метод моментов

Задача. Пусть $x_1, \ldots, x_n$ – реализация из распределения Бернулли $\mathbf{B}_\theta$ с неизвестным параметром успеха $\theta \in [0, 1]$. Оценить θ с помощью метода моментов.

Решение. Найдем теоретический и эмпирический первые моменты

$$A_1 = \mathbb{E}[X] = 0 \cdot (1 - \theta) + 1 \cdot \theta = \theta,$$

$$a_1 = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Из уравнения $A_1 = a_1$ находим по методу моментов оценку

$$\widehat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Метод моментов

Задача. Пусть $x_1, \dots, x_n$ – реализация из модели сдвига показательного закона с плотностью распределения

$$f_\theta(u) = e^{-(u-\theta)} \cdot \mathbf{I}_{\{u \geq \theta\}} = \begin{cases} e^{-(u-\theta)}, & u \geq \theta, \\ 0, & u < \theta. \end{cases}$$

Оценить θ с помощью метода моментов.

Решение. Найдем теоретический и эмпирический первые моменты

$$A_1 = \int_\theta^{+\infty} u e^{-(u-\theta)} du = \left(-u e^{-(u-\theta)}\right)\Big|_\theta^\infty + \int_\theta^{+\infty} e^{-(u-\theta)} du$$
$$= \theta + 1,$$
$$a_1 = \frac{1}{n} \sum_{i=1}^n x_i.$$

Метод моментов

Из уравнения $A_1 = a_1$ находим по методу моментов оценку

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i - 1.$$

Метод максимального правдоподобия

Метод максимального правдоподобия: чтобы оценить d неизвестных параметров модели, нам необходимо найти максимум функции правдоподобия (то есть найти частные производные по d параметрам и приравнять их к нулю).

Метод максимального правдоподобия

Пусть дана реализация выборки $x_1, \ldots, x_n$ из некоторого распределения X с неизвестным параметром θ.

Введем величину:

$$f(u, \theta) = \begin{cases} \mathbb{P}_\theta(X = u) & \text{в дискретном случае,} \\ f_\theta(u) & \text{в непрерывном случае } (f_\theta \text{ — плотность}). \end{cases}$$

Метод максимального правдоподобия

Функцией правдоподобия называется величина:

$$L(\theta) = f(x_1, \theta) \cdot \ldots \cdot f(x_n, \theta).$$

В дискретном случае $L(\theta)$ равна вероятности получить реализацию $x_1, \ldots, x_n$ выборки при заданном θ.

В общем случае $L(\theta)$ характеризует вероятность получить реализацию $x_1, \ldots, x_n$ выборки при заданном θ.

Метод максимального правдоподобия

Представляется разумным в качестве оценки параметра θ взять наиболее правдоподобное значение, которое получается при максимизации функции $L(\theta)$.

Замечание. Часто проще искать точку максимума функции $\ln L(\theta)$, которая совпадает с максимумом $L(\theta)$ в силу монотонности логарифма.

Замечание. В случае, если функция $L(\theta)$ не является непрерывно дифференцируемой, необходимо дополнительно анализировать окрестности точек разрыва.

Метод максимального правдоподобия

Задача. Пусть $x_1, \ldots, x_n$ – реализация из распределения Бернулли $\mathbf{B}_\theta$ с неизвестным параметром успеха $\theta \in [0, 1]$. Оценить θ с помощью метода максимального правдоподобия.

Решение. Найдем сначала функцию правдоподобия:

$$f(u, \theta) = \mathbb{P}_\theta(X = u) = \begin{cases} 1 - \theta & \text{для } u = 0, \\ \theta & \text{для } u = 1, \end{cases} = (1-\theta)^{1-u}\theta^u,$$

$$L(\theta) = f(x_1, \theta) \cdot \ldots \cdot f(x_n, \theta) = (1-\theta)^{n - \sum_{i=1}^n x_i} \cdot \theta^{\sum_{i=1}^n x_i}.$$

Найдем логарифм функции правдоподобия:

$$\ln L(\theta) = \left(n - \sum_{i=1}^n x_i\right) \ln(1-\theta) + \sum_{i=1}^n x_i \ln \theta.$$

Метод максимального правдоподобия

Дифференцируя ее по θ, получаем

$$(\ln L(\theta))' = -\frac{n - \sum_{i=1}^n x_i}{1-\theta} + \frac{\sum_{i=1}^n x_i}{\theta}.$$

Приравняем производную к нулю:

$$\frac{n - \sum_{i=1}^n x_i}{1-\theta} = \frac{\sum_{i=1}^n x_i}{\theta}.$$

Откуда

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^n x_i.$$

Метод максимального правдоподобия

Задача. Пусть $x_1, \ldots, x_n$ – реализация из модели сдвига показательного закона с плотностью распределения

$$f_\theta(u) = e^{-(u-\theta)} \cdot \mathbf{I}_{\{u \geq \theta\}} = \begin{cases} e^{-(u-\theta)}, & u \geq \theta, \\ 0, & u < \theta. \end{cases}$$

Оценить θ с помощью метода максимального правдоподобия.

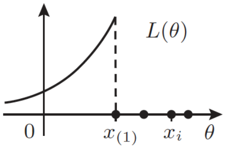
Решение. Найдем сначала функцию правдоподобия:

$$f(u, \theta) = f_\theta(u) = e^{-(u-\theta)} \cdot \mathbf{I}_{\{u \geq \theta\}},$$

$$L(\theta) = f(x_1, \theta) \cdot \ldots \cdot f(x_n, \theta) = e^{-\sum_{i=1}^n x_i + n\theta} \cdot \mathbf{I}_{\{\min_{i=1,\ldots,n} x_i \geq \theta\}}.$$

Метод максимального правдоподобия

Заметим, что здесь $L(\theta)$ не является гладкой функцией, и поэтому оценку нельзя вычислять, «глупо» приравнивая к нулю производную функции правдоподобия.



По графику получаем оценку

$$\widehat{\theta} = \min_{i=1,\ldots,n} x_i,$$

которая отлична от оценки метода моментов, найденной ранее.

Спасибо за внимание!