

Введение в математическую статистику. Оценивание параметров распределения.

Леонид Иосипой

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

2 июня 2021

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

Организационная информация

1. О курсе/о себе
2. Время занятий/Перерывы
3. Теория/Практика/Python
4. Домашние задания/Google Classroom
5. Сайт

Теория оценивания

В статистике данные часто рассматриваются как **реализация выборки из некоторого распределения, известного с точностью до одного или нескольких параметров.**

При таком подходе для определения распределения, наиболее подходящего для описания данных, достаточно уметь оценивать значения неизвестных параметров по реализации выборки.

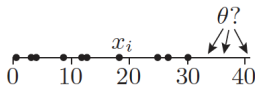
Теория оценивания

Пример

Пусть $\theta > 0$ — неизвестное число. Ниже даны координаты x_i десяти точек, взятых из равномерного распределения на $[0, \theta]$.

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 задачей:

- ▶ x_1, \dots, x_n — это реализация независимых равномерно распределенных на отрезке $[0, \theta]$ случайных величин X_i ;
- ▶ $\theta \in \Theta = (0, +\infty)$ — это неизвестный параметр;
- ▶ необходимо оценить θ по реализации выборки x_1, \dots, x_n наиболее точно.

Теория оценивания

Задача оценивания параметров в общем случае.

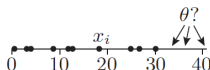
- ▶ x_1, \dots, x_n — это реализация выборки (независимых и одинаково распределенных случайных величин) из функции распределения $F_\theta(u)$;
- ▶ $\theta \in \Theta$ — это неизвестный параметр, а $\{F_\theta(u), \theta \in \Theta\}$ — некоторое семейство функций распределения;
- ▶ задача состоит в том, чтобы **оценить (восстановить) θ по реализации выборки x_1, \dots, x_n наиболее точно.**

Теория оценивания

Будем оценивать θ с помощью функций $\hat{\theta}(x_1, \dots, x_n)$ от n переменных x_1, \dots, x_n , которые мы будем называть **оценками** или **статистиками**.

Подставляя в оценку $\hat{\theta}$ реализацию выборки x_1, \dots, x_n , мы получим число — оценку неизвестного параметра θ .

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

Теория оценивания

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

Несмещенность

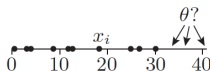
Оценка $\hat{\theta}(x_1, \dots, x_n)$ параметра θ называется **несмещенной**, если

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

Здесь индекс θ у математического ожидания \mathbb{E}_{θ} означает, что мы считаем математическое ожидание случайной величины $\hat{\theta}(X_1, \dots, X_n)$, где X_i распределены согласно $F_{\theta}(x)$.

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

Несмещенность



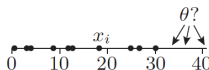
1. Является ли $\hat{\theta}_1(x_1, \dots, x_n) = 35$ несмещенной?

Нет, не является, так как

$$\mathbb{E}_\theta[\hat{\theta}_1(X_1, \dots, X_n)] = \mathbb{E}_\theta[35] = 35.$$

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

Несмещенность

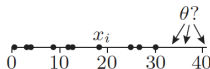


2. Является ли $\hat{\theta}_2(x_1, \dots, x_n) = 2x_7$ несмещенной?

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

$$\mathbb{E}_{\theta}[\hat{\theta}_2(X_1, \dots, X_n)] = \mathbb{E}_{\theta}[2X_7] = 2 \cdot \frac{\theta}{2} = \theta.$$

Несмещенность



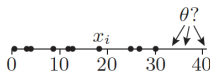
3. Является ли $\hat{\theta}_3(x_1, \dots, x_n) = \max\{x_1, \dots, x_n\}$ несмещенной?

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,

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

Несмещенность



4. Является ли $\hat{\theta}_4(x_1, \dots, x_n) = 2 \frac{x_1 + \dots + x_n}{n}$ несмещенной?

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

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

Несмещенность

Результаты проверки на несмещенность:

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

$$\hat{\theta}_1(x_1, \dots, x_n) = 35;$$

$$\hat{\theta}_2(x_1, \dots, x_n) = 2x_7;$$

$$\hat{\theta}_3(x_1, \dots, x_n) = \max\{x_1, \dots, x_n\};$$

$$\hat{\theta}_4(x_1, \dots, x_n) = 2 \frac{x_1 + \dots + x_n}{n}.$$

Состоятельность

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

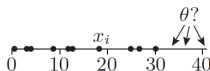
$$\hat{\theta}(X_1, \dots, X_n) \xrightarrow{\mathbb{P}_\theta} \theta \quad \text{при } n \rightarrow \infty.$$

Здесь $\xrightarrow{\mathbb{P}_\theta}$ обозначает «сходимость по вероятности»: для любого $\varepsilon > 0$

$$\mathbb{P}_\theta(|\hat{\theta}(X_1, \dots, X_n) - \theta| > \varepsilon) \rightarrow 0 \quad \text{при } n \rightarrow \infty.$$

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

Состоятельность

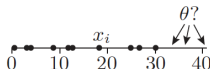


1. Является ли $\hat{\theta}_1(x_1, \dots, x_n) = 35$ состоятельной?

Нет, не является, так как, например, для $\theta = 34$ и $\varepsilon = 0.1$

$$\mathbb{P}(|\hat{\theta}_1(X_1, \dots, X_n) - \theta| > \varepsilon) = \mathbb{P}(|35 - 34| > 0.1) = 1 \not\rightarrow 0.$$

Состоятельность

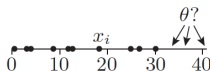


2. Является ли $\hat{\theta}_2(x_1, \dots, x_n) = 2x_7$ состоятельной?

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

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

Состоятельность



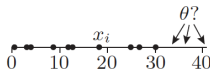
3. Является ли $\hat{\theta}_3(x_1, \dots, x_n) = \max\{x_1, \dots, x_n\}$ состоятельной?

Да, является, так как

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

так как число в скобке строго меньше 1.

Состоятельность



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

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

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

Следовательно,

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

Состоятельность

Результаты проверки на несмещенность и состоятельность:

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

$$\hat{\theta}_1(x_1, \dots, x_n) = 35;$$

$$\hat{\theta}_2(x_1, \dots, x_n) = 2x_7;$$

$$\hat{\theta}_3(x_1, \dots, x_n) = \max\{x_1, \dots, x_n\};$$

$$\hat{\theta}_4(x_1, \dots, x_n) = 2 \frac{x_1 + \dots + x_n}{n}.$$

Методы построения оценок

Перейдем теперь к методам построения оценок. Известно много таких методов: метод моментов, метод максимального правдоподобия, метод спейсингов и т.д.

Основная идея любого метода построения оценок: 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 многомерным 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:

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

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

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

Метод моментов

Задача

Пусть x_1, \dots, x_n — реализация выборки из распределения Бернулли $\text{Ber}(\theta)$ с неизвестным параметром успеха $\theta \in [0, 1]$.
Оценить θ с помощью метода моментов.

Метод 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

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

Метод моментов

Задача

Пусть x_1, \dots, x_n — реализация выборки из экспоненциального распределения $\text{Exp}(\theta)$ с неизвестным параметром $\theta > 0$.

Оценить θ с помощью метода моментов.

Метод моментов

Решение. Найдем теоретический и эмпирический первые моменты

$$A_1 = \int_0^{+\infty} u \cdot \theta e^{-\theta u} du = \left(-ue^{-\theta u} \right) \Big|_0^{\infty} + \int_0^{+\infty} e^{-\theta u} du = \frac{1}{\theta},$$

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

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

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

Спасибо за внимание!